Market Microstructure and Algorithmic Trading

University of Oxford

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

lecture notes

January 14, 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 very different: no economy of scale but high margins. Mass market means a lot of similar products, so logistics are optimized.

Why algorithmic trading ?

Three main reasons:

- Market fragmentation
- Intermediation
- Electronification of markets

price and speed has lowered the barriers to entry, since exchanges and market makers now compete on those two variables. New players have proliferated as a result, further fragmenting the US equity markets
near-instantaneous liquidity,
2005 Regulation National Market System (Reg NMS), which left exchanges competing mainly on price, may have unintentionally encouraged the fragmentation of execution venues and market makers while prioritizing speed over reliability and safety

Fragmentation

Slides 121 —>

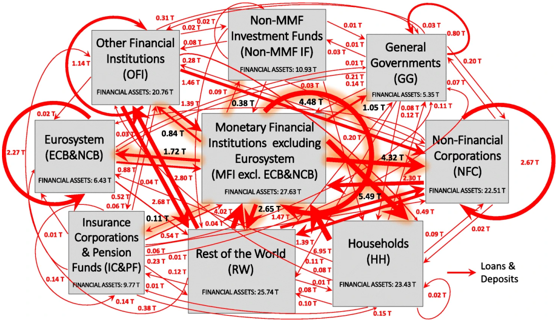
SLIDES 126 —-> CAL

Bid-ask spread drives the style of trading

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).

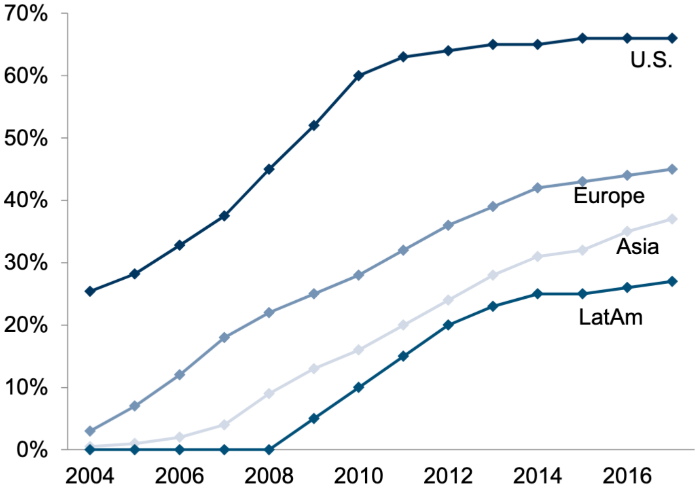
Banks: Match borrowers and lenders, match buyers and sellers, structured prod. / hedgeing.
Others: central banks, IBs, mutual funds, brokers, dealers, central counterparty clearing houses, insurances, etc.

Policy makers push towards electronification: improves traceability, and less information asymmetry. improve the process of price discover

increase in electronification and digitalisation in financial markets, which helps reduce the costs of trading
growth in electronic trading platforms (ETPs) and algorithmic (or automated) trading:
helped to pool liquidity more effectively than before by enabling the multilateral and cross- border interaction between buyers and sellers of financial assets

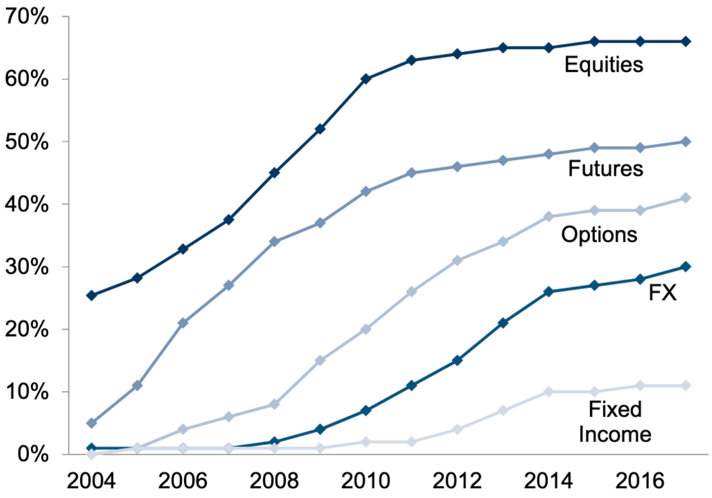
Main consequence of electronification: high-frequency trading (HFT) that reduce the delay, or latency, in execution have increased the speed at which market participants can acces markets

Market share of algorithmic trading by region, %



Source: Aite Group, Goldman Sachs Global Investment Research.

Market share of algorithmic trading by asset class, %



Source: Aite Group, Goldman Sachs Global Investment Research.

Limit Order Book

- Limit orders are price-contingent orders to buy or sell an asset.
- Limit orders follow price-time priority, and collectively form the book.

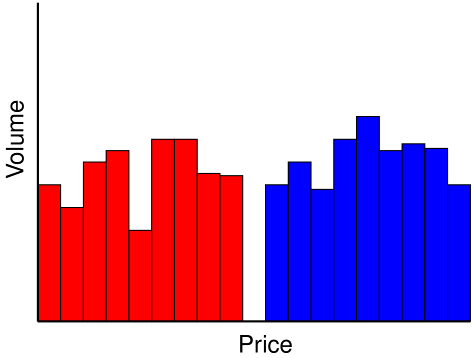
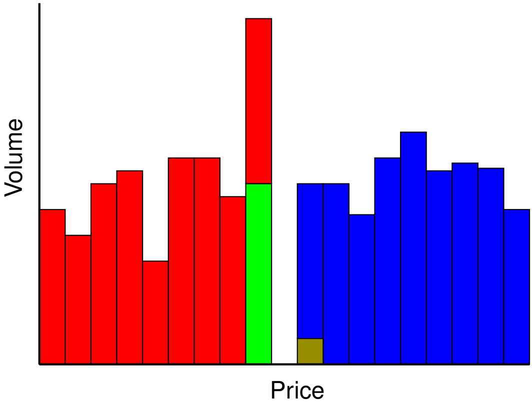


Figure: Limit order book.

Spoofing to sell the asset without crossing spread I

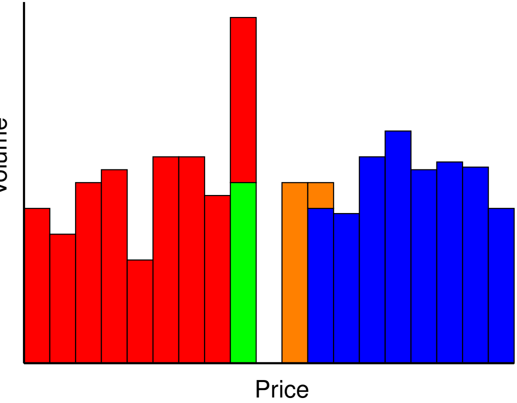


- Spoofing is a quote-based manipulative strategy.
- Limit orders are submitted to both sides of the book — intention is to trade only on one side.

For example, if objective is to sell an asset, then

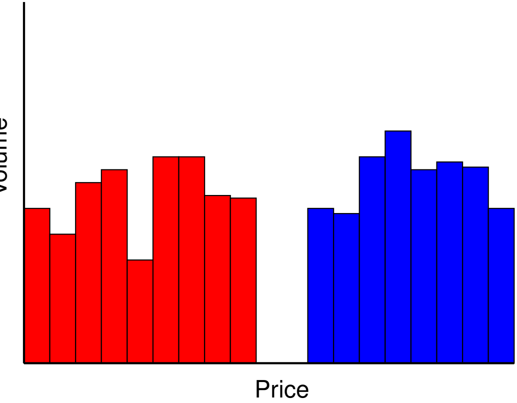
- submit a large buy limit order that will be cancelled, and
- submit a limit order on the ask that is intended to result in a transaction.

Spoofing to sell the asset without crossing spread II



- Increase in buy-pressure is interpreted as an expected increase in the price.
- A buy-heavy book is followed by an increase in the arrival rate of buy market orders that **cross the spread** in anticipation of a price increase.
- These market orders lift the limit sell order that is intended to result in a transaction.

Spoofing to sell the asset without crossing spread III



- The spoof order, i.e., the large limit buy,
  - is cancelled or expires, or
  - is inadvertently filled
- Spoofing allows one to buy or sell an asset at a more favorable price than was otherwise likely to occur, i.e., not cross the spread.

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.

Examples:

- Asset managers delegate to their dealing desk the process of buying or selling. They give to the trading desk specifications (speed, exposure, etc) to be fulfilled. For instance, some metaorders have to be executed fast, others more slowly.
- high frequency traders or fast hedge funds take decisions because of liquidity signals.
- Market Makers use optimal trading to provide liquidity to other market participants.
- Almost every agent solves an optimization problem.

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).
  - In smart routing of LOs: why ending as fast as possible is a good choice ?
- The **controls**:
  - What are the possible actions of the agent ?
  - In smart routing: why control the distribution of prices/quantities, and not fractions through a trading window ?

In **optimal trading** problems: the sequence of events is important because **liquidity has a memory**: if I go fast for the first order of my metaorder, I may impact the price such that I will pay more for next orders. $\Longrightarrow$ we need dynamic programming.

Optimal trading

- Multiple agents in a large fragmented market.
- Each agent has
  - A specific goal / objective.
  - An urgency 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.

Smart order routing of aggressive orders

- 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.
- 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 trading problem ?

Smart order routing of aggressive orders

- 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.
- 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 trading 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** $n \in \{1, \cdots, N\}$**.**

Smart order routing of aggressive orders

- Solution to the optimal trading problem ?

Smart order routing of aggressive orders

- Solution to the optimal trading problem ?
- 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

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

Smart order routing of aggressive orders

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..

Smart order routing of limit orders

- 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 ?

Smart order routing of limit orders

- 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 $Q^\star = \sum_n q_n$.

Smart order routing of limit orders

The **solution**:

- After our LOs, venue n has new queue quantity $Q_n + q_n$.
- Queue is consumed in t_n:

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

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

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

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

Problems: Latency & cancellations & iceberg orders

Optimal execution $\subset$ optimal trading.

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).
  - In smart routing of LOs: why ending as fast as possible is a good choice ?
- The **controls**:
  - What are the possible actions of the agent ?
  - In smart routing: why control the distribution of prices/quantities, and not fractions through a trading window ?

In **optimal trading** problems: the sequence of events is important because **liquidity has a memory**: if I go fast for the first order of my metaorder, I may impact the price such that I will pay more for next orders. $\Longrightarrow$ we need dynamic programming.

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.

Guéant, Olivier (Mar. 2016). **The Financial Mathematics of Market Liquidity: From Optimal Execution to Market Making**. ISBN: 9780429153778. DOI: 10.1201/b21350.

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.

Why AT? One example

- Institutional or large players need to trade (buy and sell) large amounts of securities. These quantities are too large for the market to process without prices moving in the ‘wrong direction’ (slippage).
- Thus, large orders are broken up in small ones and these are traded over time (minutes, hours, days, weeks, or even months) and across different venues.
- Deciding how to break up and execute a large order can mean saving millions of pounds for large players