Systematic Trading from First Principles

Oden Petersen

October 30, 2025

"$y = X\beta + \epsilon$, the rest is commentary."

About Me

Point of This Talk

Outline

Securities Markets

Spot Transactions

The point of trading is to obtain an asset by giving up money, or obtain money by giving up an asset.

If I give $q > 0$ units of some asset A, and you give me \$$pq$, then:

- I have **sold** q units of A to you at \$$p$
- You have **bought** q units of A from me for \$$p$

Buying and selling are collectively called 'trading'.

Suppose I own some amount of A and some amount of money. If we let s be $+1$ for buying and -1 for selling, then the result of any trade is to add qs to the amount of A I own, and add $-\$qps$ to the amount of money I have.

Securities Markets and Exchanges

The **market** is the collective activity of all traders. When we don't care who we trade with, we can just 'trade with the market'.

A **securities market** for some asset A, open at a time t, is any standardised way for traders to reach agreements to buy or sell A at a specified **settlement time** $T \geq t$.

Securities Markets and Exchanges

The **market** is the collective activity of all traders. When we don't care who we trade with, we can just 'trade with the market'.

A **securities market** for some asset A, open at a time t, is any standardised way for traders to reach agreements to buy or sell A at a specified **settlement time** $T \geq t$.

For example, $T = \ldots$

- t ('spot', e.g. blockchain)
- $t+1, t+2, \ldots$ ('clearing', e.g. equities)
- Last Thursday of month ('futures')

Securities Markets and Exchanges

The **market** is the collective activity of all traders. When we don't care who we trade with, we can just 'trade with the market'.

A **securities market** for some asset A, open at a time t, is any standardised way for traders to reach agreements to buy or sell A at a specified **settlement time** $T \geq t$.

For example, $T = \ldots$

- t ('spot', e.g. blockchain)
- $t+1, t+2, \ldots$ ('clearing', e.g. equities)
- Last Thursday of month ('futures')

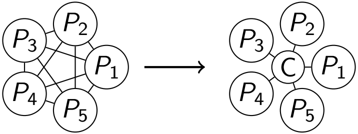
If you agree to give something to someone, you have an **obligation**. If someone agrees to give you something, you have a **right**.

Counterparty Risk

If I have an agreement with P_1 to buy 10 units for $\$p_1$ at T, and an agreement with P_2 to sell 10 units at $\$p_2$ at T, and no further rights/obligations, am I guaranteed to meet my obligations?

Centralisation

A **securities exchange** is a centralised venue serving a securities market for **exchange participants** (e.g. ASX, NYSE, TSE, HKEX, LME). Agreements not made through an exchange are often called OTC (over-the-counter).



Centralisation generally reduces **search costs** and **counterparty risk**.

Netting

Centralisation allows for **netting** of rights and obligations.
For any settlement time T, I only need to keep track of the difference between money owed to and by me, and units owed to and by me.
The quantity of A owned by me, plus the quantity owed to me, minus the quantity owed by me to others, is known as my **net position** in A.
If this is positive, I have a **long position**. If it is negative, I have a **short position**. If it is zero, I am **flat**.

Collateralisation

At certain intermediate times t' ($t \leq t' \leq T$), participants may be required to physically give ('post') something to the exchange to **collateralise** their obligations.

- Money ('margin')
- Assets ('locate'/'borrow')

If an agreement made on the exchange gives you rights to money or assets at T, this is typically as good as posting actual money or assets for an obligation at $T' \geq T$.
Some amount of interest may be charged to make up for the difference between the size of our obligations and the size of our collateral. For cash, this is according to an **interest rate**; for other assets, it is according to a **borrow rate**/**short rate**.

Summary

- **Trading** is swapping money and assets
- A **market** is whatever you use to trade
- A **securities market** is a standardised way to agree to trades
- Agreements consist of **rights** and **obligations**
- Finding a **counterparty** may involve **search cost**
- Agreements between two parties are subject to **counterparty risk**
- A **securities exchange** is a centralised trading venue
- After trades are agreed to on an exchange, they will be **settled** in some standardised way
- The net quantity of A that I have some claim to can either be positive (**long position**), negative (**short position**), or zero (**flat**).
- Traders may be obligated to post assets ('locate') or money ('margin')

Trading

Setup

A sequence of trades that collectively increases the amount of money you have and leaves the amount of each asset you have unchanged is clearly favourable.

Setup

A sequence of trades that collectively increases the amount of money you have and leaves the amount of each asset you have unchanged is clearly favourable.

Suppose that at each time t we have cash holdings of $\$c_t$ and net holdings of a_t units of some asset A.

Suppose also that trades $(s_t, q_t, \$p_t)$ take place at a finite set of distinct times

$$\tau = \{t_1, \ldots t_n\} \subset T = [t_-, t_+],$$

where $t_- < t_1 < \ldots < t_n < t_+$.

Setup

A sequence of trades that collectively increases the amount of money you have and leaves the amount of each asset you have unchanged is clearly favourable.

Suppose that at each time t we have cash holdings of $\$c_t$ and net holdings of a_t units of some asset A.

Suppose also that trades $(s_t, q_t, \$p_t)$ take place at a finite set of distinct times

$$\tau = \{t_1, \dots t_n\} \subset T = [t_-, t_+],$$

where $t_- < t_1 < \dots < t_n < t_+$.

Suppose further that p_t is a right-continuous function $\mathbb{R} \to \mathbb{R}$ with left-limits.

For instance, we could take $p_t = p_{\max(\tau \cap (-\infty, t])}$ for $t \geq \min \tau$ and $p_t = x$ otherwise for some arbitrary x. This is known as the last traded price.

Accounting

For any time-varying quantity x_t, let x_t^+ and x_t^- denote the right- and left-limits respectively.
Furthermore, define a signed measure x_ω such that for any interval $T' = [t'_-, t'_+]$ we have

$$x_{T'} := x^+_{t_+} - x^-_{t_-}.$$

Accounting

For any time-varying quantity x_t, let x_t^+ and x_t^- denote the right- and left-limits respectively.
Furthermore, define a signed measure x_ω such that for any interval $T' = [t'_-, t'_+]$ we have

$$x_{T'} := x_{t_+}^+ - x_{t_-}^-.$$

Then we have

$$\begin{aligned}
a_{T'} &= \sum_{t \in \tau} s_t q_t && = \int_{t \in T} da, \\
c_{T'} &= \sum_{t \in \tau} -p_t(s_t q_t) && = \int_{t \in T} -p_t da, \\
p_{T'} &= p_{t'_+} - p_{t'_-}^-. && = \int_{t \in T} dp.
\end{aligned}$$

Cash Holdings

It can be shown (see appendix) that the cashflow over the entire interval $T = [t_-, t_+]$ is

$$\$c_T = \int_{t \in T} -\$p_t da = \$p_{t_-} a_{t_-} - \$p_{t_+} a_{t_+} + \$ \int_{t \in T} a_t^- dp.$$

This is similar in spirit to integration by parts:

$$\int_a^b f \frac{dg}{dx} dx = f(b)g(b) - f(a)g(a) - \int_a^b g \frac{df}{dx} dx.$$

Then we have

$$\$(c_{t_+} + p_{t_+} a_{t_+}) - (c_{t_-} + p_{t_-} a_{t_-}) = \$ \int_{t \in T} a_t^- dp.$$

The quantity $\$v_t = \$p_t a_t$ is known as the **dollar value** of our A holdings **marked** to the price $\$p_t$.

Portfolio Valuation

Suppose now that we trade multiple assets, such that p_t, a_t and v_t are vector-valued, with v_t the elementwise product of p_t and a_t.
A collection of assets held in quantities a_t is known as a **portfolio**.

Portfolio Valuation

Suppose now that we trade multiple assets, such that p_t, a_t and v_t are vector-valued, with v_t the elementwise product of p_t and a_t.
A collection of assets held in quantities a_t is known as a **portfolio**.
We can write

$$(c_{t_+} + p_{t_+} \cdot a_{t_+}) - (c_{t_-} + p_{t_-} \cdot a_{t_-}) = \int_{t \in T} a_t^- \cdot dp,$$

where p_ω is now a vector-valued measure.

Portfolio Valuation

Suppose now that we trade multiple assets, such that p_t, a_t and v_t are vector-valued, with v_t the elementwise product of p_t and a_t.
A collection of assets held in quantities a_t is known as a **portfolio**.
We can write

$$(c_{t_+} + p_{t_+} \cdot a_{t_+}) - (c_{t_-} + p_{t_-} \cdot a_{t_-}) = \int_{t \in T} a_t^- \cdot dp,$$

where p_ω is now a vector-valued measure. Let

$$\Pi_t = c_t + p_t \cdot a_t = c_t + \sum v_t.$$

We call $\$\Pi_t$ the **value** of our portfolio **marked** to p_t.

Profitability

The quantity $\$\Pi_{t_+} - \$\Pi_{t_-}$ is our **net P&L** (profit and loss) over the interval T, marked to p_t. Then we have

$$\Pi_T = \int_{t \in T} a_t^- \cdot dp.$$

we can write

$$\Pi_{[t_i, t_{i+1}]} = \int_{[t_i, t_{i+1}]} a_t^- \cdot dp = \int_{t \in [t_i, t_{i+1}]} v_t^- \cdot \frac{dp}{p_t^-},$$

where the quotient $\frac{dp}{p_t^-}$ is computed elementwise.

Leverage

Suppose we can always make any trade we like at time t with price $\$p_t$.
Then we can freely convert a portfolio with value $\$\Pi_t$ to that much in cash.
Conversely, we can convert $\$\Pi_t$ worth of cash into any portfolio with that value.
In practice, there are limits on the trades we can make at a particular price and time.

Leverage

Suppose we can always make any trade we like at time t with price $\$p_t$.
Then we can freely convert a portfolio with value $\$\Pi_t$ to that much in cash.
Conversely, we can convert $\$\Pi_t$ worth of cash into any portfolio with that value.
In practice, there are limits on the trades we can make at a particular price and time.
Typically, $\$\Pi_t$ can change in two ways: trading assets, or transferring cash into and out of the portfolio. We will generally ignore the possibility of transfers.

Leverage

Suppose we can always make any trade we like at time t with price $\$p_t$. Then we can freely convert a portfolio with value $\$\Pi_t$ to that much in cash.

Conversely, we can convert $\$\Pi_t$ worth of cash into any portfolio with that value.

In practice, there are limits on the trades we can make at a particular price and time.

Typically, $\$\Pi_t$ can change in two ways: trading assets, or transferring cash into and out of the portfolio. We will generally ignore the possibility of transfers.

If we begin with a portfolio worth $\$\Pi_{t_1}$ and make a sequence of trades of the form (s_t, q_t, p_t) that result in a portfolio worth $\$\Pi_{t_n}$, then we could instead begin with a portfolio worth $L\$\Pi_{t_1}$ and make trades (s_t, Lq_t, p_t) to arrive at a portfolio worth $L\$\Pi_{t_n}$. The ratio L is known as the **leverage ratio**.

Return on Capital

Because of collateralisation requirements, portfolio management uses up cash.

Consider long-only spot-settled trading. If we were to turn our portfolio into cash, or convert cash into an identical portfolio, we would receive/require $\$\Pi_t$.[1]

[1] We generally want c_t to be small, but we may want some spare cash for trading, so it still forms part of the collateralisation requirement.

Return on Capital

Because of collateralisation requirements, portfolio management uses up cash.

Consider long-only spot-settled trading. If we were to turn our portfolio into cash, or convert cash into an identical portfolio, we would receive/require $\$\Pi_t$.[1]

If our initial portfolio value were $\$\Pi_{t_-} + \$N$ instead of Π_{t_-}, and we could simply scale up trade sizes at the same prices, then set

$$L = \frac{\Pi_{t_-} + N}{\Pi_{t_-}}.$$

[1] We generally want c_t to be small, but we may want some spare cash for trading, so it still forms part of the collateralisation requirement.

Return on Capital

Because of collateralisation requirements, portfolio management uses up cash.

Consider long-only spot-settled trading. If we were to turn our portfolio into cash, or convert cash into an identical portfolio, we would receive/require $\$\Pi_t$.[1]

If our initial portfolio value were $\$\Pi_{t_-} + \$N$ instead of Π_{t_-}, and we could simply scale up trade sizes at the same prices, then set

$$L = \frac{\Pi_{t_-} + N}{\Pi_{t_-}}.$$

The **return on capital** is defined as the increase in P&L per dollar added to initial portfolio value, i.e.

$$R_T = \frac{L \int_T d\Pi - \int_T d\Pi}{N} = \frac{\int_T d\Pi}{\Pi_{t_-}}.$$

[1]We generally want c_t to be small, but we may want some spare cash for trading, so it still forms part of the collateralisation requirement.

Log Returns

If we define

$$\ell_t = \log \Pi_t$$

for any t'_-, t'_+, then we have

$$R_T = \exp(\ell l_T) - 1,$$

and for any measurable set ω we can define

$$R_\omega = \exp(\ell_\omega) - 1 \approx \ell_\omega + O(\ell_\omega^2) \text{ (for small } \ell_\omega).$$

We call ℓ_T the **log-return** over the interval T.

Properties of Returns and Log-Returns

Let $w_t := \frac{1}{\Pi_t} v_t$ be the **weight vector**.

For an interval $T' = (t_i, t_{i+1}]$, we have a_t^- equal to a constant over T', and

$$R_{T'} = w_{t_i}^+ \cdot r_{T'},$$

where the elementwise quotient

$$r_{T'} = \frac{p_{t_{i+1}} - p_{t_i}}{p_{t_i}}$$

is known as the **asset returns** vector over T'. In contrast, $\ell_{T'}$ is not linear in $r_{T'}$.

For a disjoint collection of measurable sets $\omega_1, \ldots \omega_n$ whose union is Ω, we have

$$\ell_\Omega = \sum_{i=1}^n \ell_{\omega_i},$$

$$R_\Omega = \left(\prod_{i=1}^n (1 + R_{\omega_i}) \right) - 1 \approx \sum_{i=1}^n R_{\omega_i} + O\left(\sum_{i=1}^n \sum_{j=1}^n |R_{\omega_i} R_{\omega_j}| \right).$$

Summary

-

Market Microstructure

Trade Formation

In practice, the trades we can make at a time t and a price p_t are limited by our ability to find a willing counterparty.

Trade Formation

In practice, the trades we can make at a time t and a price p_t are limited by our ability to find a willing counterparty.

On an electronic exchange, trades are formed by interacting with the exchange's **matching engine**.

Trade Formation

In practice, the trades we can make at a time t and a price p_t are limited by our ability to find a willing counterparty.

On an electronic exchange, trades are formed by interacting with the exchange's **matching engine**.

For each trade $(s, q, \$p)$, the exchange will typically charge a fee proportional to the **dollar volume** $\$pq$ of the trade. Fee rates may vary depending on trade type and between participants in accordance with exchange policy.

Trade Formation

In practice, the trades we can make at a time t and a price p_t are limited by our ability to find a willing counterparty.

On an electronic exchange, trades are formed by interacting with the exchange's **matching engine**.

For each trade $(s, q, \$p)$, the exchange will typically charge a fee proportional to the **dollar volume** $\$pq$ of the trade. Fee rates may vary depending on trade type and between participants in accordance with exchange policy.

The most common type of matching engine design is a **limit-order book** (sometimes called a double auction), which can operate in either a **continuous** or **batched** fashion.

Limit Order Book

At any point in time, market participants can create a request (‘**limit order**’) of the form $(s, q, \$p)$ to trade up to q units in direction $s = \pm 1$ at any price $\$(p - sm), m \geq 0$.
The value $\$m$ is known as the **price improvement**.
They are then said to be “**bid** for $\$p$” $(s = +1)$ or “**ask**ing/**offer**ing at $\$p$” $(s = -1)$.

Limit Order Book

At any point in time, market participants can create a request ('**limit order**') of the form $(s, q, \$p)$ to trade up to q units in direction $s = \pm 1$ at any price $\$(p - sm), m \geq 0$.
The value $\$m$ is known as the **price improvement**.
They are then said to be "**bid** for $\$p$" $(s = +1)$ or "**ask**ing/**offer**ing at $\$p$" $(s = -1)$.
All limit orders active at time t are collected into a **limit-order book** $\mathcal{L}_t$.
By convention, $\mathcal{L}_t$ is right-continuous with left limits.
Users can add, cancel and modify orders, subject to exchange-specific rules.

Order Matching

Whenever $(+1, q_1, \$p_1), (-1, q_2, \$p_2) \in \mathcal{L}_t$ with $p_2 \leq p_1$, both orders could be at least partly satisfied by trading up to $q_{\max} = \min(q_1, q_2)$ units with one another at a price $\$p \in [\$p_2, \$p_1]$. If such a pair exists the book is said to be **in cross**.

Order Matching

Whenever $(+1, q_1, \$p_1), (-1, q_2, \$p_2) \in \mathcal{L}_t$ with $p_2 \leq p_1$, both orders could be at least partly satisfied by trading up to $q_{\max} = \min(q_1, q_2)$ units with one another at a price $\$p \in [\$p_2, \$p_1]$. If such a pair exists the book is said to be **in cross**.

If an order $(s, q, \$p)$ is in cross with another, it may be **matched** for q' units. The total matched quantity for all buy orders must equal the total matched quantity for all sell orders.

Order Matching

Whenever $(+1, q_1, \$p_1), (-1, q_2, \$p_2) \in \mathcal{L}_t$ with $p_2 \leq p_1$, both orders could be at least partly satisfied by trading up to $q_{\text{max}} = \min(q_1, q_2)$ units with one another at a price $\$p \in [\$p_2, \$p_1]$. If such a pair exists the book is said to be **in cross**.
If an order $(s, q, \$p)$ is in cross with another, it may be **matched** for q' units. The total matched quantity for all buy orders must equal the total matched quantity for all sell orders.
In this case, q' units will trade, and the order will become $(s, q - q', \$p)$.

Order Matching

Whenever $(+1, q_1, \$p_1), (-1, q_2, \$p_2) \in \mathcal{L}_t$ with $p_2 \leq p_1$, both orders could be at least partly satisfied by trading up to $q_{\max} = \min(q_1, q_2)$ units with one another at a price $\$p \in [\$p_2, \$p_1]$. If such a pair exists the book is said to be **in cross**.

If an order $(s, q, \$p)$ is in cross with another, it may be **matched** for q' units. The total matched quantity for all buy orders must equal the total matched quantity for all sell orders.

In this case, q' units will trade, and the order will become $(s, q - q', \$p)$.

If $q = q'$ the order is said to be **fully filled** and will be removed from the book. Otherwise, it is said to be **partially filled**.

Order Matching

Whenever $(+1, q_1, \$p_1), (-1, q_2, \$p_2) \in \mathcal{L}_t$ with $p_2 \leq p_1$, both orders could be at least partly satisfied by trading up to $q_{\max} = \min(q_1, q_2)$ units with one another at a price $\$p \in [\$p_2, \$p_1]$. If such a pair exists the book is said to be **in cross**.

If an order $(s, q, \$p)$ is in cross with another, it may be **matched** for q' units. The total matched quantity for all buy orders must equal the total matched quantity for all sell orders.

In this case, q' units will trade, and the order will become $(s, q - q', \$p)$. If $q = q'$ the order is said to be **fully filled** and will be removed from the book. Otherwise, it is said to be **partially filled**.

The ability to quickly find matches for a large number of units at a reasonable price is known as **liquidity**, and is another major benefit of centralisation.

Supply and Demand Curves

We can partition $\mathcal{L}_t$ into $\mathcal{L}_t = \mathcal{B}_t \cup \mathcal{A}_t$, with $\mathcal{B}_t$ the bid orders and $\mathcal{A}_t$ the ask orders.

Supply and Demand Curves

We can partition $\mathcal{L}_t$ into $\mathcal{L}_t = \mathcal{B}_t \cup \mathcal{A}_t$, with $\mathcal{B}_t$ the bid orders and $\mathcal{A}_t$ the ask orders.

Now define the functions

$$Q_t(+1, \$p) = \sum_{\substack{(+1,q',\$p') \in \mathcal{B}_t \\ \$p \leq \$p'}} q'$$

$$Q_t(-1, \$p) = \sum_{\substack{(-1,q',\$p') \in \mathcal{A}_t \\ \$p \geq \$p'}} q'$$

$$M_t(\$p) = \min(Q_t(+1, \$p), Q_t(-1, \$p))$$

Supply and Demand Curves

We can partition $\mathcal{L}_t$ into $\mathcal{L}_t = \mathcal{B}_t \cup \mathcal{A}_t$, with $\mathcal{B}_t$ the bid orders and $\mathcal{A}_t$ the ask orders.
Now define the functions

$$Q_t(+1, \$p) = \sum_{\substack{(+1, q', \$p') \in \mathcal{B}_t \\ \$p \leq \$p'}} q'$$

$$Q_t(-1, \$p) = \sum_{\substack{(-1, q', \$p') \in \mathcal{A}_t \\ \$p \geq \$p'}} q'$$

$$M_t(\$p) = \min(Q_t(+1, \$p), Q_t(-1, \$p))$$

The functions $Q_t(-1, \$p)$ and $Q_t(+1, \$p)$ are known as the **supply curve** and **demand curve** respectively. The function $M_t(\$p)$ represents the **matchable quantity** at $\$p$.

Supply and Demand Curves

We can partition $\mathcal{L}_t$ into $\mathcal{L}_t = \mathcal{B}_t \cup \mathcal{A}_t$, with $\mathcal{B}_t$ the bid orders and $\mathcal{A}_t$ the ask orders.
Now define the functions

$$Q_t(+1, \$p) = \sum_{\substack{(+1, q', \$p') \in \mathcal{B}_t \\ \$p \leq \$p'}} q'$$

$$Q_t(-1, \$p) = \sum_{\substack{(-1, q', \$p') \in \mathcal{A}_t \\ \$p \geq \$p'}} q'$$

$$M_t(\$p) = \min(Q_t(+1, \$p), Q_t(-1, \$p))$$

The functions $Q_t(-1, \$p)$ and $Q_t(+1, \$p)$ are known as the **supply curve** and **demand curve** respectively. The function $M_t(\$p)$ represents the **matchable quantity** at $\$p$. The book $\mathcal{L}_t$ is in cross if and only if there exists some $\$p$ with $M_t(\$p) > 0$.

Batch Matching

In **batch** or **auction** style matching, orders are matched with one another only at particular discrete times.

Batch Matching

In **batch** or **auction** style matching, orders are matched with one another only at particular discrete times.

1. Prior to the **match time** t^*, users can typically add, modify and cancel limit orders.
2. At each time $t \le t^*$, an **indicative price** $\$p_t^*$ will be selected such that $M_t(\$p_t^*)$ is maximal. Tiebreaking will depend on exchange rules.
3. Finally, at the match time t^*, some subset of the crossed limit orders will be matched at $\$p^*$ for a total quantity $M_{t^*}(\$p_{t^*}^*)$. After the match, the book will no longer be crossed.

Batch Matching

In **batch** or **auction** style matching, orders are matched with one another only at particular discrete times.

1. Prior to the **match time** t^*, users can typically add, modify and cancel limit orders.
2. At each time $t \leq t^*$, an **indicative price** $\$p_t^*$ will be selected such that $M_t(\$p_t^*)$ is maximal. Tiebreaking will depend on exchange rules.
3. Finally, at the match time t^*, some subset of the crossed limit orders will be matched at $\$p^*$ for a total quantity $M_{t^*}(\$p_{t^*}^*)$. After the match, the book will no longer be crossed.

Maximising $M_t(\$p_t^*)$ is equivalent to maximising the sum of qm across all orders, where q is the quantity filled and m is the price improvement.
It is common to use this matching style at the beginning or end of a trading day or lunch break, or when there is some kind of market instability such as following a large price move or company announcement.
Sometimes t^* is referred to as a **liquidity event** because of the large volume traded, and the relative insensitivity of $\$p_{t^*}^*$ to individual orders.

Batch Matching Properties

The following monotonicity properties typically hold:

- $\$p_t^*$ nondecreasing in $\mathcal{B}_t$ and nonincreasing in $\mathcal{A}_t$
- For each $\$p$, $M_t(\$p)$ nondecreasing in $\mathcal{L}_t$
- For individual orders $(s, q, \$p)$, we will have $\$p_t^*$ nondecreasing in $\$p$ and sq.
- For individual orders $(s, q, \$p)$ and each $\$p'$, we will have $M_t(\$p')$ nondecreasing in q and nondecreasing in $\$sp$.

Batch Matching Properties

The following monotonicity properties typically hold:

- $\$p_t^*$ nondecreasing in $\mathcal{B}_t$ and nonincreasing in $\mathcal{A}_t$
- For each $\$p$, $M_t(\$p)$ nondecreasing in $\mathcal{L}_t$
- For individual orders $(s, q, \$p)$, we will have $\$p_t^*$ nondecreasing in $\$p$ and sq.
- For individual orders $(s, q, \$p)$ and each $\$p'$, we will have $M_t(\$p')$ nondecreasing in q and nondecreasing in $\$sp$.

Price Priority

Because $M_t(\$p')$ is nondecreasing in sp, the matching will be designed to obey **price priority**.
If we have two orders $(s_1, q_1, \$p_1), (s_2, q_2, \$p_2)$ with $\$s_1p_1 > \$s_2p_2$, then the second order cannot be matched unless the first is completely filled.

Order Timing

The order book is often visible to all participants. Traders may be incentivised to wait until immediately before the match time to post orders.

Order Timing

The order book is often visible to all participants. Traders may be incentivised to wait until immediately before the match time to post orders. If everyone does this, the matching engine may be overloaded, and p_t^* will change very rapidly leading up to t^*.

Order Timing

The order book is often visible to all participants. Traders may be incentivised to wait until immediately before the match time to post orders. If everyone does this, the matching engine may be overloaded, and p_t^* will change very rapidly leading up to t^*.
The match time is typically chosen at random in some short interval in order to disincentivise this behaviour.

Order Timing

The order book is often visible to all participants. Traders may be incentivised to wait until immediately before the match time to post orders. If everyone does this, the matching engine may be overloaded, and p_t^* will change very rapidly leading up to t^*.

The match time is typically chosen at random in some short interval in order to disincentivise this behaviour. To further encourage early submission of orders, many exchanges also implement a time priority rule.

Time Priority

If two orders exist at the same price $\$p$, the one that reached the matching engine later cannot be matched unless the earlier order is completely filled.

Order Timing

The order book is often visible to all participants. Traders may be incentivised to wait until immediately before the match time to post orders. If everyone does this, the matching engine may be overloaded, and p_t^* will change very rapidly leading up to t^*.

The match time is typically chosen at random in some short interval in order to disincentivise this behaviour. To further encourage early submission of orders, many exchanges also implement a time priority rule.

Time Priority

If two orders exist at the same price $\$p$, the one that reached the matching engine later cannot be matched unless the earlier order is completely filled.

Tick Size

Time priority would not have much effect if we could just insert the later order at a price $\$p + s\epsilon$ for some very small $\epsilon > 0$.

To avoid this, prices must be integer multiples of some small increment $\$\delta$, known as the **tick size**.

Continuous Matching

In **continuous matching**, a match time is triggered every time a new limit order causes the book to become crossed.

Continuous Matching

In **continuous matching**, a match time is triggered every time a new limit order causes the book to become crossed.

If the matching engine receives an order at t, then immediately before and after t the book will be uncrossed, with $M_t(\$p) = 0$ at all $\$p$.

Continuous Matching

In **continuous matching**, a match time is triggered every time a new limit order causes the book to become crossed.

If the matching engine receives an order at t, then immediately before and after t the book will be uncrossed, with $M_t(\$p) = 0$ at all $\$p$.

The only orders involved in the match will be the arriving order and some set of orders $\mathcal{M}_t$ in the opposite direction.

The arriving order is known as the **active** or **aggressive** order, and the pre-existing orders are known as **passive**.

Typically we are not allowed to match with ourselves. Often the exchange will implement **self-trade protection** so that the quantity of our active and passive orders is simply cancelled out without recording a trade.

Price priority is still used, and time priority is usually used.

Continuous Matching

In **continuous matching**, a match time is triggered every time a new limit order causes the book to become crossed.

If the matching engine receives an order at t, then immediately before and after t the book will be uncrossed, with $M_t(\$p) = 0$ at all $\$p$.

The only orders involved in the match will be the arriving order and some set of orders $\mathcal{M}_t$ in the opposite direction.

The arriving order is known as the **active** or **aggressive** order, and the pre-existing orders are known as **passive**.

Typically we are not allowed to match with ourselves. Often the exchange will implement **self-trade protection** so that the quantity of our active and passive orders is simply cancelled out without recording a trade.

Price priority is still used, and time priority is usually used.

For each q' matched against a passive order $(s, q, \$p)$, the active order will trade q' units with the passive order at $\$p$.

Continuous Matching

In **continuous matching**, a match time is triggered every time a new limit order causes the book to become crossed.

If the matching engine receives an order at t, then immediately before and after t the book will be uncrossed, with $M_t(\$p) = 0$ at all $\$p$.

The only orders involved in the match will be the arriving order and some set of orders $\mathcal{M}_t$ in the opposite direction.

The arriving order is known as the **active** or **aggressive** order, and the pre-existing orders are known as **passive**.

Typically we are not allowed to match with ourselves. Often the exchange will implement **self-trade protection** so that the quantity of our active and passive orders is simply cancelled out without recording a trade.

Price priority is still used, and time priority is usually used.

For each q' matched against a passive order $(s, q, \$p)$, the active order will trade q' units with the passive order at $\$p$.

The per-unit price achieved by the active trader will be

$$\$p_t^* = \frac{\sum_{(s,q,\$p)\in\mathcal{M}_t} \$pq}{\sum \quad q}.$$

Instantaneous Price Impact

If we aggressively trade a very large quantity, we will exhaust all passive orders we would most prefer to trade with and $\mathcal{M}_t$ will need to include orders at worse price levels. This is sometimes known as **walking the book**.

Instantaneous Price Impact

If we aggressively trade a very large quantity, we will exhaust all passive orders we would most prefer to trade with and $\mathcal{M}_t$ will need to include orders at worse price levels. This is sometimes known as **walking the book**.

Assume continuous matching, and consider a market order of $q > 0$ units in direction s.

Instantaneous Price Impact

If we aggressively trade a very large quantity, we will exhaust all passive orders we would most prefer to trade with and $\mathcal{M}_t$ will need to include orders at worse price levels. This is sometimes known as **walking the book**.

Assume continuous matching, and consider a market order of $q > 0$ units in direction s.

The least favourable price in $\mathcal{M}_t$ will be given by

$$\$P_t(sq) = s \min_{\{\$p : Q_t(-s,p) \geq q\}} \$sp.$$

Instantaneous Price Impact

If we aggressively trade a very large quantity, we will exhaust all passive orders we would most prefer to trade with and $\mathcal{M}_t$ will need to include orders at worse price levels. This is sometimes known as **walking the book**.

Assume continuous matching, and consider a market order of $q > 0$ units in direction s.

The least favourable price in $\mathcal{M}_t$ will be given by

$$\$P_t(sq) = s \min_{\{\$p : Q_t(-s,p) \geq q\}} \$sp.$$

The unit price of the match will be given by

$$\$p_t^*(sq) = \frac{1}{q} \int_0^q \$P_t(sq')dq'.$$

We call the sensitivity of $\$p_t^*$ to sq the **instantaneous price impact**.

Bid-Ask Spread

We call the prices

$$\begin{aligned}
\$b_t &= \lim_{q \to 0^+} \$p_t^*(-q) && = \max_{(+1, q, \$p) \in \mathcal{B}_t} \$p \\
\$a_t &= \lim_{q \to 0^+} \$p_t^*(q) && = \min_{(-1, q, \$p) \in \mathcal{A}_t} \$p
\end{aligned}$$

the **bid price** and **ask price** respectively. All bid orders have price at most $\$b_t$ and all ask orders have price at least $\$a_t$.
The interval $[\$b_t, \$a_t]$ is known as the **spread**, and $\$a_t - \$b_t$ is the **width** of the spread. If $\$a_t - \$b_t = \$\delta$, we say that the market for the asset is **large-tick** or **tick-constrained**.

Price Proxies

Note that $\$p_t^*(0)$ is not yet defined. So long as we choose some price $\$m_t$ satisfying $\$m_t \in [\$b_t, \$a_t]$, setting $\$p_t^*(0) := \$m_t$ will make $\$p_t^*(\cdot)$ nondecreasing.

Price Proxies

Note that $\$p_t^*(0)$ is not yet defined. So long as we choose some price $\$m_t$ satisfying $\$m_t \in [\$b_t, \$a_t]$, setting $\$p_t^*(0) := \$m_t$ will make $\$p_t^*(\cdot)$ nondecreasing.

This is variously called the **theo**retical price, **microprice** or **price proxy** depending on context.

Price Proxies

Note that $\$p_t^*(0)$ is not yet defined. So long as we choose some price $\$m_t$ satisfying $\$m_t \in [\$b_t, \$a_t]$, setting $\$p_t^*(0) := \$m_t$ will make $\$p_t^*(\cdot)$ nondecreasing.

This is variously called the **theo**retical price, **microprice** or **price proxy** depending on context.

Some simple choices for m_t include:

- $\frac{1}{2}b_t + \$\frac{1}{2}a_t$ (arithmetic **midprice**)
- $\$\sqrt{b_t a_t}$ (**geometric midprice**)
- $\$(1 - I_t)b_t + \$I_t a_t$ (depth-$\$d$ **weighted midprice**)
- $\$b_t^{1-I_t} a_t^{I_t}$ (depth-$\$d$ **geometrically weighted midprice**)

We call I_t the **book imbalance**.

Price Proxies

Note that $\$p_t^*(0)$ is not yet defined. So long as we choose some price $\$m_t$ satisfying $\$m_t \in [\$b_t, \$a_t]$, setting $\$p_t^*(0) := \$m_t$ will make $\$p_t^*(\cdot)$ nondecreasing.

This is variously called the **theo**retical price, **microprice** or **price proxy** depending on context.

Some simple choices for m_t include:

- $\frac{1}{2}b_t + \$\frac{1}{2}a_t$ (arithmetic **midprice**)
- $\$\sqrt{b_t a_t}$ (**geometric midprice**)
- $\$(1 - I_t)b_t + \$I_t a_t$ (depth-$\$d$ **weighted midprice**)
- $\$b_t^{1-I_t} a_t^{I_t}$ (depth-$\$d$ **geometrically weighted midprice**)

We call I_t the **book imbalance**.

A popular choice for this is

$$I_t = \frac{Q_t(+1, \$b_t)}{Q_t(+1, \$b_t) + Q_t(-1, \$a_t)}.$$

Alternative price proxies are described in the appendix.

Persistent Price Impact

We call the difference $\$\lambda_t(sq) = \$p_t^*(sq) - \$m_t$ the **instantaneous price impact curve** of trading q units in direction s.
Buy orders will have nonnegative instantaneous price impact, while sell orders will have nonpositive instantaneous price impact.

Persistent Price Impact

We call the difference $\$\lambda_t(sq) = \$p_t^*(sq) - \$m_t$ the **instantaneous price impact curve** of trading q units in direction s.

Buy orders will have nonnegative instantaneous price impact, while sell orders will have nonpositive instantaneous price impact.

Because aggressive trades remove liquidity from one side of the book, there is also a persistent effect on $\mathcal{L}_t$ and consequently $\$m_t$. This is known as **persistent price impact**.

Persistent Price Impact

We call the difference $\$\lambda_t(sq) = \$p_t^*(sq) - \$m_t$ the **instantaneous price impact curve** of trading q units in direction s.
Buy orders will have nonnegative instantaneous price impact, while sell orders will have nonpositive instantaneous price impact.
Because aggressive trades remove liquidity from one side of the book, there is also a persistent effect on $\mathcal{L}_t$ and consequently $\$m_t$. This is known as **persistent price impact**.
The realised instantaneous price impact is given by

$$\$\lambda_t = \$p_t - \$m_t,$$

while the realised persistent price impact is given by

$$\$\nu_t = \$m_t - \$m_t^{\emptyset},$$

where $\$m_t^{\emptyset}$ is the path the microprice process would have taken had we not interacted at all with the matching engine.

P&L with transaction costs

We can write $\$p_t = \$m_t^{\emptyset} + \$\nu_t + \$\lambda_t$.

P&L with transaction costs

We can write $\$p_t = \$m_t^{\emptyset} + \$\nu_t + \$\lambda_t$.

$$
\begin{aligned}
\$\Pi_T &= \$\int_{t\in T} a_t^- (dm + dn + d\lambda) \\
&= \underbrace{\$\int_{t\in T} a_t^- dm}_{\text{Midprice P\&L}} - \underbrace{\$\int_{t\in T} \nu_t da}_{\text{PPI Penalty}} - \underbrace{\$\int_{t\in T} \lambda_t da}_{\text{IPI Penalty}} \\
&\quad + \underbrace{\$((\nu_{t_+} + \lambda_{t_+})a_{t_+} - (\nu_{t_-} + \lambda_{t_-})a_{t_-})}_{\$0 \text{ if } a_{t_+} = a_{t_-} = 0}.
\end{aligned}
$$

P&L with transaction costs

We can write $\$p_t = \$m_t^{\emptyset} + \$\nu_t + \$\lambda_t$.

$$\begin{aligned}\$\Pi_T &= \$\int_{t\in T} a_t^{-}(dm + dn + d\lambda) \\ &= \underbrace{\$\int_{t\in T} a_t^{-}\,dm}_{\text{Midprice P\&L}} - \underbrace{\$\int_{t\in T} \nu_t\,da}_{\text{PPI Penalty}} - \underbrace{\$\int_{t\in T} \lambda_t\,da}_{\text{IPI Penalty}} \\ &+ \underbrace{\$((\nu_{t_+} + \lambda_{t_+})a_{t_+} - (\nu_{t_-} + \lambda_{t_-})a_{t_-})}_{\$0 \text{ if } a_{t_+}=a_{t_-}=0}.\end{aligned}$$

Attempts to make ν_t and *da* covary negatively are very hard to pull off and usually considered manipulative.

But if we only use passive execution, it is guaranteed that λ_t and *da* will covary negatively, and this term will change from a loss to a profit. Trying to make money solely from this term is known as **market making** or **liquidity provision**.

Downsides of Market Making

Assuming $a_{t_-} = a_{t_+} = 0$,

$$\$\Pi_T = \underbrace{\$\int_{t \in T} a_t^- \, dm}_{\text{Midprice P\&L}} - \underbrace{\$\int_{t \in T} \nu_t \, da}_{\text{PPI Penalty}} - \underbrace{\$\int_{t \in T} \lambda_t \, da}_{\text{IPI Penalty}} .$$

Downsides of Market Making

Assuming $a_{t_-} = a_{t_+} = 0$,

$$\$\Pi_T = \underbrace{\$ \int_{t \in T} a_t^- \, dm}_{\text{Midprice P\&L}} - \underbrace{\$ \int_{t \in T} \nu_t \, da}_{\text{PPI Penalty}} - \underbrace{\$ \int_{t \in T} \lambda_t \, da}_{\text{IPI Penalty}} .$$

With a market-making strategy, we lose a lot of control over a_t^-.

Downsides of Market Making

Assuming $a_{t_-} = a_{t_+} = 0$,

$$\$\Pi_T = \underbrace{\$\int_{t\in T} a_t^- \, dm}_{\text{Midprice P\&L}} - \underbrace{\$\int_{t\in T} \nu_t \, da}_{\text{PPI Penalty}} - \underbrace{\$\int_{t\in T} \lambda_t \, da}_{\text{IPI Penalty}}.$$

With a market-making strategy, we lose a lot of control over a_t^-.
If market participants in general is making money on this term we will tend to lose money. This tendency is referred to as **adverse selection**.

Downsides of Market Making

Assuming $a_{t_-} = a_{t_+} = 0$,

$$\$\Pi_T = \underbrace{\$\int_{t \in T} a_t^- \, dm}_{\text{Midprice P\&L}} - \underbrace{\$\int_{t \in T} \nu_t \, da}_{\text{PPI Penalty}} - \underbrace{\$\int_{t \in T} \lambda_t \, da}_{\text{IPI Penalty}}.$$

With a market-making strategy, we lose a lot of control over a_t^-.
If market participants in general is making money on this term we will tend to lose money. This tendency is referred to as **adverse selection**.
In particular, if the priority of our orders are quite low, we will only trade against the aggressive orders with the largest quantity, which tend to be most predictive of midprice changes over a short time horizon. High order priority is therefore extremely valuable for a market making strategy.

Downsides of Market Making

Assuming $a_{t_-} = a_{t_+} = 0$,

$$\$\Pi_T = \underbrace{\$\int_{t \in T} a_t^- \, dm}_{\text{Midprice P\&L}} - \underbrace{\$\int_{t \in T} \nu_t \, da}_{\text{PPI Penalty}} - \underbrace{\$\int_{t \in T} \lambda_t \, da}_{\text{IPI Penalty}}.$$

With a market-making strategy, we lose a lot of control over a_t^-.
If market participants in general is making money on this term we will tend to lose money. This tendency is referred to as **adverse selection**.
In particular, if the priority of our orders are quite low, we will only trade against the aggressive orders with the largest quantity, which tend to be most predictive of midprice changes over a short time horizon. High order priority is therefore extremely valuable for a market making strategy.
However, it is still possible to make money on both terms. This is particularly true if a_t changes relatively slowly (**low-frequency trading**).
Whether this results in better performance overall is a different question.

Market Making Strategy

A highly simplified model of optimal market making is given by Avellaneda and Stoikov (2008). The market maker maintains two limit orders at any time in opposite directions, of the form

$$\left(s, 1, \$\left(m_t - \gamma q - \frac{1}{2} s \varsigma\right)\right),$$

where γ is a risk-aversion parameter[2] and ς is the difference between the two prices.

In general, to avoid large a_t, (which would make our P&L very sensitive to price changes), we want the amount we buy to match the amount we sell.

[2] Defined differently in the paper

Market Impact Modeling

A very simple model for ν_t takes the form

$$\nu_t = \int_0^t f(s_{t'} q_{t'}) G(t - t') dt',$$

for some $f, G : \mathbb{R} \to \mathbb{R}$. This is known as a **propagator model**.
Persistent impact is empirically seen to obey a **square-root law** (due to Grinold and Kahn, 1994),

$$\nu \propto \alpha\sigma \frac{\int da}{\sqrt{V|\int da|}},$$

where σ is the typical variance of returns and V is the typical volume over the period.
Gatheral (2016) derives this from a variety of different models of market impact.
Recall that our P&L from persistent price impact is given by

$$-\$ \int_{t \in T} \nu_t da.$$

Asset Pricing

Arrow-Debreu Securities

An **event** is a set of possible worlds distinguished from other possible worlds by some common feature.
For instance, there is a set of possible worlds where Australia wins its next game of cricket against the UK.
Suppose Ω is the set of all events.[3] At each time t', we can construct a set $\mathcal{F}_{t'} \subseteq \Omega$ containing all events whose outcome is known at time t'.[4]
An **Arrow-Debreu security** A_ω for $\omega \in \mathcal{F}_t$ is one which we can exchange at time t for \$1 if ω occurs and \$0 otherwise. Assume there exists t such that $\omega \in \mathcal{F}_t$.
Let $\$\mathbb{Q}_{t'}(\omega)$ be the market price of ω at time $t' \leq t$. Assume unlimited liquidity, no fees, and no collateral requirements or interest payments.

[3] We require that Ω be a σ-algebra.
[4] $\mathcal{F}_t$ is a filtration, i.e. a nondecreasing σ-algebra

Dutch Book Theorems

The market is **arbitrage-free** if there is no fixed set of trades that guarantees positive P&L in all possible worlds.
Assuming this is the case, we can deduce properties that $\mathbb{Q}_{t'}(\cdot)$ must satisfy.

$\mathbb{Q}_{t'}(\omega) \in [0, 1]$

If $\mathbb{Q}_{t'}(w) < 0$, we would actually receive money by buying A_ω, and we can sell it at t for at least \$0. This would guarantee positive P&L.
If $\mathbb{Q}_{t'}(w) > 1$, we could short A_ω, and buy it back at t for at most \$1. This would guarantee positive P&L.

If ω_1 and ω_2 are disjoint sets, we say they are **mutually exclusive events**.

$\mathbb{Q}_{t'}(\omega_1) + \mathbb{Q}_{t'}(\omega_2) = \mathbb{Q}_{t'}(\omega_1 \cup \omega_2)$

If the LHS exceeds the RHS, we buy $A_{\omega_1}, A_{\omega_2}$ and short $A_{\omega_1 \cup \omega_2}$; otherwise, we do the reverse.

An arbitrage strategy on Arrow-Debreu securities is sometimes called a

Fundamental Theorem of Asset Pricing

In general, we can consider securities that are redeemable at time t for some amount given by a $\mathcal{F}_t$-measurable function (**random variable**) S. A market is arbitrage-free at time t' if and only if the market price of each security S is given by

$$\mathbb{E}_{\mathbb{Q}_{t'}}[S] = \int_\Omega S d\mathbb{Q}_{t'},$$

where $\mathbb{Q}_{t'}$ is some probability measure on $\mathcal{F}_t$. This is known as the **fundamental theorem of asset pricing** (FTAP).

$\mathbb{Q}$-world vs $\mathbb{P}$-world

A set of securities M is said to form a **complete market** for Ω if and only if we can infer the arbitrage-free prices of every Arrow-Debreu security from the prices of M.

In practice, even if we erroneously assume identical bid and ask prices, markets are only ever complete for extremely trivial Ω. This means there are many different measures $\mathbb{Q}$ that are consistent with the prices of M.

The only way to see if asset prices are consistent with one another is by constraining them with some set of **model assumptions** to induce a particular probability measure $\mathbb{P}$.

This is one extremely general description of the family of strategies known as **statistical arbitrage**. Favourable outcomes are no longer guaranteed.

Portfolio Management

Utility

We now consider events relating to our P&L.
Suppose for simplicity that there are a finite number of possible P&L outcomes $\Pi^1, \Pi^2, \ldots \Pi^n$ between now and some later time t.
We can construct a $\mathcal{F}_t$-measurable function X that equals Π^i on a set of probability $\mathbb{P}(X = \Pi^i)$.
We might decide that for some pair of outcome distributions X, X' we prefer our P&L to follow X rather than X'.
A theorem of Von Neumann and Morgenstern (1947) shows that under quite reasonable assumptions about our preferences, there exists a function $u : \{\Pi^1, \Pi^2, \ldots \Pi^n\} \to \mathbb{R}$ such that we prefer X to X' if and only if

$$\mathbb{E}_{\mathbb{P}}[u(X)] > \mathbb{E}_{\mathbb{P}}[u(X')].$$

If different trading strategies produce different outcome distributions of P&L, we might therefore seek a **utility function** to compare them.

Compounding

Recall that

$$\ell^+_{t_+} = \ell^-_{t_-} + \ell_T.$$

If we have a series of intervals $T_1, T_2, \ldots$, we can define $T^n = \bigcup_{i=1}^n T_i$.
Then we have $\ell_{T^n} = \sum_{i=1}^n \ell_{T_i}$.
Suppose we have two strategies, one with log-returns measure ℓ and another with log-returns measure ℓ'.
If we assume for simplicity that $\ell_{T_i} - \ell'_{T_i}$ are independent and identically distributed, it follows from the law of large numbers that

$$\frac{\ell_{T^n} - \ell'_{T^n}}{n} \to \mathbb{E}[\ell_{T_i}] - \mathbb{E}[\ell'_{T_n}].$$

This means that in the long-run, ℓ will exceed ℓ' with probability approaching one if and only if

$$\mathbb{E}[\ell_{T_i}] > \mathbb{E}[\ell_{T_i}].$$

We call the stochastic process ℓ_t the **Kelly utility**, due to Kelly (1956).

Kelly Criterion

The Kelly utility of our P&L over some interval T is given by

$$u_T = \ell_T = \log(1 + R_T)$$

In practice, it may make sense to use the more conservative **fractional Kelly criterion**,

$$u_T = \log\left(1 + \frac{R_T}{L}\right),$$

where $L \in [0, 1]$ is a leverage parameter.
For small values of $\frac{R_T}{L}$, we can approximate this using a Taylor expansion:

$$\mathbb{E}[\log(1 + \frac{R_T}{L})] \approx \frac{1}{L}\mathbb{E}[R_T] - \frac{1}{2L^2}\mathbb{E}[R_T^2] + \mathbb{E}[O(\frac{R_T^3}{L^3})].$$

The quadratic approximation is known as the **quadratic utility** of the return on capital.

Quadratic Utility

If the central limit theorem applies to the u_{T_i}, then in the limit the distribution of u_{T^n} will be governed only by the first two moments. Recall that over certain intervals T', $R_{T'}$ is given by $w \cdot r_{T'}$, where w is the vector of portfolio weights. If w is deterministic, the quadratic utility over T' is given by

$$\frac{1}{L} w \cdot \mathbb{E}[r_{T'}] - \frac{1}{2L^2} w^\top \mathbb{E}[r_{T'} r_{T'}^\top] w,$$

and the optimal value of w is given by

$$w^* = L\mathbb{E}[r_{T'} r_{T'}^\top]^{-1} \mathbb{E}[r_{T'}] = L(\Sigma + \mu\mu^\top)^{-1}\mu,$$

where μ and Σ are the expectation and covariance of $r_{T'}$.

Markowitz Portfolio Optimisation

In practice, μ is relatively small compared to Σ, and so we have

$$w* = L(\Sigma + \mu\mu^\top)^{-1} \approx L\Sigma^{-1}\mu.$$

We can get this by maximising the quadratic loss function

$$Lw \cdot \mu - \frac{1}{2}w^\top \Sigma w.$$

This is equivalent to maximising the ratio

$$\frac{\mathbb{E}[R_{T'}]}{\sqrt{\mathrm{Var}[R_{T'}]}},$$

known as the **Sharpe ratio**, subject to $\mathrm{Var}[R_{T'}]$ having some desired positive value.
The problem of maximising expected returns subject to an upper bound on returns variance is known as **Markowitz portfolio optimisation**.

Estimation

If we have a number of returns vectors $r_1, r_2, \ldots r_n$, we might try to invoke the law of large numbers and write

$$\begin{aligned} \mu \approx \hat{\mu} &= \frac{1}{n}\sum_{i=1}^{n} r_n, \\ \Sigma \approx \hat{\Sigma} &= \frac{1}{n-1}\sum_{i=1}^{n} r_n r_n^\top. \end{aligned}$$

Unfortunately, n might be smaller than the number of assets p. In this case, the matrix $[r_1, r_2, \ldots r_n]^\top$ will have a nullspace containing portfolios whose estimated volatility is 0. Portfolios close to this nullspace will have arbitrarily large Sharpe ratios, and no optimal portfolio will exist, since we cannot solve $\Sigma w = L\mu$.

Furthermore, even if $n > p$, errors in the estimation may have a significant effect on the expected utility out-of-sample.

We can augment the loss function slightly by adding a **regularisation penalty**,

Factor Models

There are certain common factors that jointly influence all stocks.
In the Capital Asset Pricing Model (CAPM), due initially to Treynor (1961), the returns[5] on the jth asset over T_i are given by

$$r_i^j = \underbrace{\beta^j}_{\text{Asset Sensitivity}} \underbrace{f_i - \mathbb{E}[f_i]}_{\text{Market Return}} + \underbrace{\alpha^j + \sigma^j \epsilon_i^j}_{\text{Idiosyncratic Return}},$$

where ϵ_i^j have mean zero and unit variance, and are uncorrelated with the f_i and with each other.
We have

$$\Sigma = \beta \Sigma_f \beta^\top + \Sigma_\epsilon,$$
$$\mu = \alpha,$$

where α, β are populated by α^j, β^j, Σ_ϵ is a diagonal matrix with entries σ^{j2}, and Σ^f is the variance of f_i.
Let $H_\beta = \beta(\beta^\top \beta)^{-1} \beta^\top$ be the projection matrix onto the span of β. Then

$$\alpha = \underbrace{H_\beta \alpha} + \underbrace{(I - H_\beta)\alpha} = \beta \alpha^\beta + \alpha^\perp,$$

Factor Models

In general, we replace β with a matrix B and f_i becomes a vector.
Then we have

$$r_i = B(f_i - \mathbb{E}[f_i] + \alpha^B) + \alpha^\perp + \Sigma_\epsilon^{\frac{1}{2}} \epsilon_i.$$

Centering the factor returns f_i will not affect the model, so we can assume $\mathbb{E}[f_i] = 0$.

$$r_i = B(f_i + \alpha^B) + \alpha^\perp + \Sigma_\epsilon^{\frac{1}{2}} \epsilon_i.$$

Then we have

$$\Sigma = B\Sigma_f B^\top + \Sigma_\epsilon,$$
$$\mu = \alpha.$$

We say that the covariance matrix of r_i is **spiked**, meaning that it is the sum of a low-rank matrix and a diagonal matrix.
If we construct matrices $R, F, \alpha^\beta, \alpha^\perp, \varepsilon$ whose ith columns are $r_i, f_i, \alpha^\beta, \alpha^\perp, \epsilon_i$ respectively, then we have

$$R = B(F + \alpha^B) + \alpha^\perp + \Sigma_\epsilon^{\frac{1}{2}} \varepsilon = BF + \mathbb{E}[R] + \Sigma_\epsilon^{\frac{1}{2}} \varepsilon = BCC^{-1}F + \mathbb{E}[R] + \Sigma_\epsilon^{\frac{1}{2}} \varepsilon,$$

Fundamental Factor Models

We don't necessarily need to estimate B. We can use **characteristics** of the various assets. Some common choices:

- 1 if the asset belongs to a particular country or industry, 0 otherwise
- All 1s (equal weighting)

It is also possible to extend the model to include characteristics that vary day-to-day, such as

- Realised volatility
- Illiquidity (e.g. $\overline{\left(\frac{|\text{Daily Returns}|}{\text{Daily Volume}}\right)}$, due to Amihud (2002))
- Crowding
- Market capitalisation
- Recent returns (momentum)
- Numbers derived from financial reports, e.g. net asset value

It is usually best to ensure the characteristics are linearly independent.

Dynamic Portfolio Selection

Market Efficiency

If all we do is maintain a fixed portfolio, there is very little ongoing need for trading.[6]

If all investors do likewise, the total dollar value invested in each asset will be proportional to the optimal w.

However, most assets give their holder the right to potential future dividends. These dividends are affected by things in the real world, which we can improve our forecast of over time.

If market participants are generally good at forecasting dividends, prices will reflect dividend expectations.[7]

Making such forecasts is expensive. We could instead just look at the total dollar value invested in each asset and use this to form portfolio weights, known as **passive investing**.

If everybody did this, prices would no longer reflect changes in the expectations of cashflows. This is known as the **Grossman-Stiglitz paradox**.

[6] More precisely, w_t is fixed, but $a_t = \Pi_t \frac{w_t}{p_t}$ may need to change a little as p_t changes.

[7] The result of price impact is to make such forecasts a public good, in the microeconomic sense.

Edge

Price impact and fees make trading expensive. We should only do it if we think we can do better than passive investing.

Reasons we might be able to do better:

- Responding to publicly available information faster than prices can move to reflect it
- Access to multiple markets simultaneously, or ability to move information between markets faster than others
- Better understanding of exchange mechanics
- Access to better market data
- Knowledge of proprietary datasets other than market data ('alternative data')
- Access to more historical data for model estimation purposes
- Knowledge of typical market dynamics (from experience)
- More intelligent pricing and/or execution (e.g. intelligent placement of orders to get good queue position)
- Ability to provide liquidity without losing to adverse selection, e.g. PFOF

Feature Engineering

Variables that let us forecast the value of a **target variable** are called **features**.

As a general heuristic, we want to choose features such that the target variable will have a different mean for different values of the feature.

Especially useful are features that have nonnegligible correlation ρ with the target, since we can then use a simple linear model for forecasting.

However, we are not always so lucky.

Once we have raw numbers, we can create new features by applying various transformations.

- Applying a fixed nonlinear function like $\tan^{-1}$ or $x \mapsto x^2$
- Setting large values of the feature to some constant (winsorisation)
- Multiplying features together (interaction terms, used in Friedman's MARS)
- Returning 1 if the feature exceeds some value, 0 otherwise (as in decision trees)
- Computing a linear combination of features then applying a nonlinear transform (as in neural networks)

Signal Aggregation

In general, the best estimates for Σ and α should vary with time based on data relevant to the assets we are trading.
Consider a single asset A with returns $r_1^A, r_2^A, \ldots r_n^A$.
Suppose we have a vector x_i of many different model forecasts for r_i^A.
Construct a **feature matrix** $X = [x_1, x_2, \ldots x_n]^\top$.
We can hold asset j in proportion to some linear combination of x_i given by $x_i \cdot \beta$.
Let $F = [r_1^A x_1, r_2^A x_2, r_n^A x_n]^\top$ and $\overline{F} = \frac{1}{n}\sum_{i=1}^n r_i^A x_i$.
Assuming no transaction costs, our mean quadratic loss over the entire dataset will be given by

$$L\overline{F} \cdot \beta - \frac{1}{2n}\beta^\top F^\top F \beta,$$

and the optimal β is given by

$$L\left(\frac{1}{n}F^\top F\right)^{-1}\overline{F}.$$

The portfolio return for the ith time period will be $R_i^A = r_i^A x_i \cdot \beta$.

Transaction Cost Analysis

Recall that

$$\$\Pi_T = \underbrace{\$\int_{t\in T} a_t^- \, dm}_{\text{Midprice P\&L}} - \underbrace{\$\int_{t\in T} \nu_t \, da}_{\text{PPI Penalty}} - \underbrace{\$\int_{t\in T} \lambda_t \, da}_{\text{IPI Penalty}}.$$

We can use a modified propagator model for $\nu_t + \lambda_t$,

$$\nu_t + \lambda_t = p_t \int_{t_-}^{t} \lambda p_{t'} G(t-t') da(t'),$$

with $G(0) = 1$.

The P&L lost to price impact is

$$\$\int_{t\in T} p_t \int_{t_-}^{t} \lambda p_{t'} G(t-t') da(t') da(t) =$$

$\int_{t\in T} \Pi_t \int_{t_-}^{t} \lambda \Pi_{t'} G(t-t') dw(t') dw(t)$.

Let $dX = [0, x_2 - x_1, x_3 - x_2, \ldots x_n - x_{n-1}]^\top$. Furthermore, define an upper-triangular matrix Λ whose i,j entry is $\lambda G(t_j - t_i)$.

If we assume portfolio value $\$\Pi_t$ is roughly equal to a constant $\$\pi$, and

Statistical Arbitrage

Appendix

Proof Sketch for c_T Identity

$$
\begin{aligned}
c_T = \sum_{t\in\tau} -p_t(s_t q_t) \quad &= \sum_{i=1}^{n} -p_{t_i}(a_{t_i}^{+} - a_{t_i}^{-}) = -\sum_{i=1}^{n} p_{t_i} a_{t_i}^{+} + \sum_{i=1}^{n} p_{t_i} a_{t_i}^{-} \\
&= -\sum_{i=1}^{n-1} p_{t_i} a_{t_{i+1}}^{-} - p_{t_n} a_{t_n}^{+} + \sum_{i=1}^{n-1} p_{t_{i+1}} a_{t_{i+1}}^{-} + p_{t_1} a_{t_1}^{-} \\
&= p_{t_1} a_{t_1}^{-} + -p_{t_n} a_{t_n}^{+} + \sum_{i=2}^{n} (p_{t_i} - p_{t_{i-1}}) a_{t_i}^{-} \\
&= p_{t_1} a_{t_1}^{-} - p_{t_n} a_{t_n}^{+} + \int_{t\in[t_1,t_n]} a_t^{-}\, dp \\
&= p_{t_-} a_{t_-} - p_{t_+} a_{t_+} + \int_{t\in T} a_t^{-}\, dp.
\end{aligned}
$$

Annualised Returns

The **annualised log-return** over ω is $\ell_\omega \frac{1 \text{ year}}{\lambda_\omega}$, where λ_ω is the duration (lebesgue measure) of ω in units of time.
The **geometrically annualised return** over ω is
$(1 + R_\omega)^{\frac{1 \text{ year}}{\lambda_\omega}} - 1 = \exp\left(\ell_\omega \frac{1 \text{ year}}{\lambda_\omega}\right) - 1.$
The **arithmetically annualised return** over ω is $R_\omega \frac{1 \text{ year}}{\lambda_\omega}$.

Alternative Price Proxies

More generally, we can define the depth-$\$d$ imbalance,

$$I_t(\$d) = \frac{Q_t(+1, \$b_t - \$d)}{Q_t(+1, \$b_t - \$d) + Q_t(-1, \$a_t + \$d)}.$$

Alternative Price Proxies

More generally, we can define the depth-\$$d$ imbalance,

$$I_t(\$d) = \frac{Q_t(+1, \$b_t - \$d)}{Q_t(+1, \$b_t - \$d) + Q_t(-1, \$a_t + \$d)}.$$

We can also define an **exponentially weighted imbalance**,

$$I_t(\zeta) = \frac{\sum_{(+1,q,p)\in\mathcal{B}_t} q \exp(-\zeta|b_t - p|)}{\sum_{(+1,q,p)\in\mathcal{L}_t} q \exp(-\zeta \min(|b_t - p|, |a_t - p|))}$$

Stuff I missed

Bibliography