Maths Finance
Personal notes

Khanh-Hung TRAN
khanhhung92vt@gmail.com

January 16, 2024

This is my personal notes written in latex (**not finished yet**) for maths finance that I have started at the end of 2022. For organization :

- The first chapter represents basic concepts and knowledges in finance.

- The second chapter represents fondamentally what is probability and its related concepts, tools ... We find here also the convergence of random variables, uniform integrability, which are necessary in chapter stochastic process.

- The third chapter is about statistics which is mainly for hypothesis tests

- The fourth chapter is for stochastic process, which describes some famous processes (Gaussian, Poisson, Markov) and then martingale, cadlag process.

- The fifth chapter is for stochastic calculus with Ito lemma ...

- The sixth chapter is about quantitative (Black-Scholes model, binomial options pricing model,...

- The seventh chapter is for technical analysis method by well-known indicators.

- The eighth and ninth chapter is for optimal control, respectively deterministic and stochastic, which are important tools for portfolio optimization. We find here also the well-known Ramsey–Cass–Koopmans model.

- The tenth chapter is for Menton's portfolio problem.

- The eleventh is about finite difference method and finite element method.

Contents

Useful links & reference

https://www.investopedia.com/
https://www.quantstart.com/articles/
https://mhasoba.github.io/TheMulQuaBio/intro.html
https://tuanvanle.wordpress.com/
https://www.probabilitycourse.com/
https://osm.netlify.app/
https://almostsuremath.com/
https://uregina.ca/~kozdron/Teaching/Regina/862Winter06/Handouts/revised_
lecture1.pdf

https://www.randomservices.org/random/index.html
wikipédia
...

Chapter 1

General finance

1 Abbreviations and financial words

- NAV : Net Asset Value.

- Holding (company) : Company which acquires financial holdings in other companies and which directs or controls their activity.

- Security : a certificate attesting credit, the ownership of stocks or bonds, or the right to ownership connected with tradable derivatives.

- A money manager : a person or financial firm that manages the securities portfolio of individual or institutional investors.

2 ETF

ETF exchange traded fund a type of investment fund and exchange-traded product, i.e, they are traded on stock exchanges.

ETFs are consituted of basket of funds, bonds, commodities,...

Their holdings (nguoi nam giu quy hay chu cua quy) are completely public and transparent.

Typically, people investing in ETFs are more interested in a diversified portfolio and want to keep their investment in an ETF for a longer periode of time.

One of most common ETFs is SPDR (SPY) tracks S&P 500, which means SPY's evoluation is exactly similar to S&P 500.

Price of ETF is available in all trading time.

3 Mutual funds

A mutual fund is an investment vehicle that made up of a pool of funds collected from many investors.

Mutual fund is operated by money manager, who invest the **fund's capital** and attempt to produce gains and income for the **fund's inverstors.**

Mutual fund's portfolio is structured and maintained to match the investment objectives stated in its prospectus (but not always reached).

Mutual funds disclose their holdings typically once a quarter (but can vary by fund).

4 Hedge funds

Hedge funds are alternative investments using pooled funds that employ numerous different strategies to earn active return, called alpha, for their investors.

Hedge funds may be agressively managed or make use of derivatives (phai sinh) and leverage (don bay tai chinh) in both domestic and international markets with the goal of generating high returns (compared to a metric, which can be specified market benchmark such as S&P 500,...)

It is important to note that hedge funds are generally only accessible to **accredited investors** (nha dau tu duoc cong nhan) as they require less SEC (Securities and Exchange Commission) regulations than other funds.

An accredited investor is an individual or a business entity that is allowed to trade securities that may not be registered with financial authorities. They are entitled to this privileged access by satisfying at least one requirement regarding their income, net worth, asset size, governance status, or professional experience. In the U.S., the term accredited investor is used by the Securities and Exchange Commission (SEC) under Regulation D to refer to investors who are financially sophisticated and have a reduced need for the protection provided by regulatory disclosure filings. Accredited investors include high-net-worth individuals (HNWIs), banks, insurance companies, brokers, and trusts.

Hedge funds do not need disclose their strategies.

4.1 What hedge funds do ?

- Long/Short Equity: This strategy involves buying stocks that the hedge fund believes will increase in value (long positions) and selling stocks that are expected to decrease in value (short positions).

- Global Macro: This strategy involves making bets on global macroeconomic trends, such as interest rates, inflation, and currency fluctuations.

- Event-Driven: This strategy involves investing in companies that are going through major events, such as mergers, acquisitions, bankruptcies, or other corporate actions.

- Quantitative: This strategy involves using computer algorithms and mathematical models to analyze financial data and identify investment opportunities.

5 Funds comparison

ETFs are similar in many ways to mutual funds, except that ETFs are bought and sold from other owners throughout the day on stock exchanges whereas mutual funds are bought and sold between the issuer of fund and individials based on their price at day's end.

https://www.vietish.com/mutual-fund-vs-index-fund-vs-etf/#:˜:text=ETF% 20c%C5%A9ng%20nh%C6%B0%20mutual%20fund,v%E1%BB%9Bi%20nhau% 20%E1%BB%9F%20secondary%20market.

5.1 Fee

- ETF : Expense ratio 0.01 - 1 %

- Mutual funds : Expense ratio 0.5 - 3 %

- Hedge funds : 2% of fund and 20% of profit (2 and 20 rule). Succesful hedge funds can offer higher ratio, as they have no real rules to follows.

5.2 Liquidity

- ETF : as a stock

- Mutual funds : end of the day.

- Hedge funds : On agreement.

6 Margin & Leverage

Margin refers to the **amount of money** that a trader must put up to open and maintain a trading position.

$$margin = \frac{your\_money}{total\_value}$$

For example, if you have \$1000 and a broker requires a margin of 50% then you can borrow \$1000, which means \$2000 in total to buy stock. your \$1000 is also your deposited money.

6.1 Margin call

Suppose that your stock decreases from \$2000 to \$1500 and the safe margin is 40%. So your actual margin is $\frac{1000-(2000-1500)}{1500} = \frac{500}{1500} = 33.3\%$

Now, you need a margin call and you have two options to set margin at 40%:

- Put more amount x deposited money:

$$\frac{500 + x}{1500 + x} = 40\%$$

- Sell your stock of amount z, and this money is put directly in the deposit.

$$\frac{500 + z}{1500} = 40\%$$

6.2 Reverse margin call

As its name, it means that if your stock increases then you can have back you money.

6.3 Leverage

Leverage refers to the **amount of capital** that a trader can control with a smaller amount of money. Laverage notation : $1 : X$. For example, with laverage is $1 : 100$ which means you have \$1 but you can control \$100.

6.4 Leverage in balance sheet

Many types:

- **Accounting leverage** is total assets divided by the total assets minus total liabilities (no phai tra).

- **Notional leverage** (don bay danh nghia) is total notional amount of assets (tai san danh nghia) plus total notional amount of liabilities (no danh nghia) divided by equity (von chu so huu).

- Economic leverage is volatility of equity (bien dong von chu so huu) divided by volatility of an unlevered investment (bien dong mot khoan dau tu khong dung don bay) in the same assets.

7 Drawdown

The drawdown is the measure of the decline from a historical peak in some variable (typically the cumulative profit or total open equity of a financial trading strategy.

$$D(T) = \max_{t \in (0,T)} [X(t) - X(T), 0]$$

Average drawdown:

$$\frac{1}{T} \int_0^T D(t)dt$$

16

8 Difference between common stock and preferred stock

- Ownership and Voting Rights:

  - Common stockholders are the true owners of the company and have voting rights. They can participate in corporate decision-making by voting on important matters, such as the election of the board of directors and major company policy changes.

  - Preferred stockholders also have ownership in the company, but they typically **do not have voting rights** or **have limited voting rights**. Their ability to influence corporate decisions is generally restricted compared to common stockholders.

- Dividends

  - Common stockholders may receive dividends, but these dividends are typically not fixed and can vary based on the company's profitability and management decisions. Dividends for common stock are paid after all obligations to preferred stockholders have been met

  - Preferred stockholders have a higher claim on the company's assets and earnings compared to common stockholders. They usually receive fixed dividends at regular intervals, and these dividends are paid before any dividends are distributed to common stockholders. The fixed dividend rate may be expressed as a percentage of the stock's par value or as a fixed dollar amount

- Liquidation Preference

  - Common stockholders have the lowest priority in terms of receiving the remaining assets after all debts and obligations are settled. They are the last to be paid and may not receive anything if there are insufficient assets

  - Preferred stockholders typically have a higher priority compared to common stockholders in receiving the company's assets during liquidation. They have a preferential claim and are entitled to receive their investment back, along with any accrued dividends, before any distribution is made to common stockholders

- Capital Appreciation

  - If the company performs well and its share price increases, common stockholders can benefit from the price appreciation and potentially sell their shares at a higher price.

  - Preferred stockholders generally do not benefit directly from capital appreciation. Their returns primarily come from the fixed dividends they receive.

9 Derivative

3 groups of peopels deal with derivatives:

- Hedging

- Speculating

- Aribtrageurs

- Forwards : a simple contract between 2 parties at a moment in the future.

- Futures (highly standardized forward contract), having clearinghouse, counterparties do not know each other.

- Swaps : Two parties exchange cash flows based on an underlying asset at a future point of time. Underlying asset can be : Interest rate, stock price, bond price, commodity. Most of swap contract used are interest rate, e.g. exchange between a fixed interest rate and a variable one.

- Options : An option gives the holder the right, but not the obligation, to buy or sell the underlying asset at a specified price, called the strike price, within a specified time period. There are two main types of options: call options and put options.

The strike price, also known as the exercise price, is a key term in options trading. It refers to the price at which the underlying asset, such as a stock, can be bought or sold if the option is exercised.

For example, if an investor holds a *call* option with a strike price of $50 and the stock price is currently trading at $60, the investor has the right to buy the stock at $50, even though it is trading at a higher price in the market. If the investor chooses to exercise this right, they would buy the stock at $50 and sell it immediately for a profit of $10 per share. **But the investor, have to pay for the option at the beginning** and it is calculated by Black–Scholes model, in 14.

Options has a expiration date !!!

In the US, options typically have an American-style exercise, which means that the option holder can exercise the option at any time before the expiration date. In contrast, options in the EU have a European-style exercise, which means that the option holder can only exercise the option on the expiration date.

Option contracts traded on futures exchanges are mainly American-style, whereas those traded over-the-counter (co phieu chua len san hay chua niem yet) are mainly European.

9.1 Who emits or creates options ?

Traders write an option by creating a new option contract that sells someone the right to buy or sell a stock at a specific price (strike price) on a specific date (expiration date). In other words, the writer of the option can be forced to buy or sell a stock at the strike price.

9.2 Covered option

A covered option is a financial transaction in which the holder of securities sells (or "writes") a call or put option, against stock that they own or are shorting. The seller of a covered option receives compensation, or "premium", for this transaction, which can limit losses; however, the act of selling a covered option also limits their profit potential to the upside.

9.3 Put option

Previously, we have an example for a call option, now we go into an example for *put* option.

Here are the details of the put option:

- Stock: XYZ Corp.

- Strike Price: $50

- Expiration Date: 30 days from now

- Premium: $3 per share

- Contract Size: 100 shares (**typically one option contract represents 100 shares**)

Scenario 1: XYZ Corp.'s stock price decreases. Suppose the price of XYZ Corp.'s stock falls to $40 per share before the option's expiration date. In this case, you can exercise your put option, which can be understood as you buy (from market) 100 shares of XYZ Corp. at $40 per share, and then you use the right of your put option, resell (to the trader that you buy this put option) at the strike price of $50 per share. This allows you to profit from the difference between the strike price and the lower market price, which is $10 per share. Considering the contract size of 100 shares, your profit would be $10 per share multiplied by 100, minus the premium paid for the put option.

Scenario 2: XYZ Corp.'s stock price increases or remains above the strike price. If the stock price of XYZ Corp. remains above the strike price of $50, or even increases, it would not be advantageous for you to exercise the put option. In this case, the put option may expire worthless, and your loss would be limited to the premium paid for the put option, which is $3 per share multiplied by 100 shares.

9.4 Pay-off

The value of an option at maturity depends on the relationship between the delivery price (K) and the underlying price (S_T) at that time.

- For a long call this payoff is:

$$f_T = S_T - K$$

- For a long put :

$$f_T = K - S_T$$

9.5 Valuation of options

9.5.1 Intrinsic value

The intrinsic value is the difference between the underlying spot price and the strike price, in favor of the option holder.

For a call option, the option is in-the-money if the underlying spot price is higher than the strike price; then the intrinsic value is the underlying price minus the strike price. For a put option, the option is in-the-money if the strike price is higher than the underlying spot price; then the intrinsic value is the strike price minus the underlying spot price. Otherwise the intrinsic value is zero.

For example, when a call (bullish/long) option has the strike price \$18,000 and the underlying is priced at \$18,050 then there is a \$50 advantage even if the option were to expire today. This \$50 is the intrinsic value of the option.

In summary, intrinsic value

- = max(current stock price - strike price,0) (for call option)

- = max(strike price - current stock price,0) (for put option)

9.5.2 Extrinsic (Time) value

In finance, the time value (TV) (extrinsic or instrumental value) of an option is the premium a rational investor would pay over its current exercise value (intrinsic value), based on the probability it will increase or decrease in favor of the buyer before expiry. Hence the longer the expiration time, the higher time value. Time value can be obtained by

Time value = option premium - intrinsic value

Time value *decays* to zero at expiration, with a general rule that it will lose $\frac{1}{3}$ of its value during the first half of its life and $\frac{2}{3}$ in the second half. As an option moves closer to expiry, moving its price requires an increasingly larger move in the price of the underlying security.

9.6 Volatility and implied volatility

Implied volatility and volatility are two related but distinct concepts used in options trading and financial analysis.

- Historical Volatility: Historical volatility is calculated based on **past price** data and represents the actual observed volatility over a given period. It provides insight into how much the price of the underlying asset has moved in the past.

- Implied Volatility : Implied volatility, on the other hand, is derived from the prices of options (about somewhat in the future). Therefore, it represents the forecast of future volatility rather than being based on historical data.

9.6.1 How to get implied volatility ?

In financial mathematics, the implied volatility of an option contract is that value of the volatility of the underlying instrument which, when input in an option pricing model (such as Black–Scholes), will return a theoretical value equal to the current market price of said option. more precisely, let C is option price, f is a function for calclutating option price (e.g. Black–Scholes), σ volatility :

$$C = f(\sigma, \ldots)$$

Then given the market price of option $\bar{C}$, the implied volatility $\bar{\sigma}$ is calculated by :

$$\bar{\sigma} = f_\sigma^{-1}(\bar{C}, \ldots)$$

9.7 Put–call parity

This defines a relationship between the price of a *European* call option and *European* put option, both with the *identical strike price and expiry*, namely that a portfolio of a long call option and a short put option is equivalent to (and hence has the same value as) a single forward contract at this strike price and expiry.

This is because if the price at expiry is above the strike price, the call will be exercised, while if it is below, the put will be exercised, and thus in either case one unit of the asset will be purchased for the strike price, exactly as in a forward contract.

Small reminder for forward contract : it is a non-standardized contract between two parties to buy or sell an asset at a specified future time (delivery and paiement) while the price is agreed at the time that contract signed.

Note that, the validity of this relationship requires that certain assumptions be satisfied, e.g, transaction costs, financing costs must be negligible.

9.7.1 Statement

Put–call parity can be stated in the following equation:

$$C - P = D \cdot (F - K)$$

where C is the (current) value of a call, P is the (current) value of a put, D is the discount factor (to come back at present value), F is the forward price of the underlying asset, and K is the strike price.

Equivalently,

$$C - P = S - D \cdot K$$

where S is the spot price, which is obtained by by discounting the forward price F.

10   Spread trade

In finance, a spread trade (also known as relative value trade) is the simultaneous purchase of one security and sale of a related security. Each bought or sold security called leg. Spread trades are usually executed with *options* or *futures contracts* as the legs, but other securities are sometimes used.

These trades are executed to yield an overall net position whose value, called the spread, depends on the difference between the prices of the legs.

Common spreads are priced and traded as a unit on futures exchanges rather than as individual legs, thus ensuring simultaneous execution and eliminating the execution risk of one leg executing but the other failing.

Spread trades are executed to attempt to profit from the widening or narrowing of the spread, rather than from movement in the prices of the legs directly. Spreads are either "bought" or "sold" depending on whether the trade will profit from the widening or narrowing of the spread.

The volatility of the spread is typically *much lower* than the volatility of the individual legs, since a change in the market fundamentals of a commodity will tend to affect both legs similarly.

10.1   Example : Bull Call spread

Sometime called Call Debit Spread. Why named Bull Call spread ?

- Bull since the acquirer expect a bull trend

- Call since call option

- Spread, since it means strategy that use combination of two options

Theese two call options (one long and one short) are with different strike prices, but with the same expiration date, to bet on a rise of the underlying. The strike price in long position must be lower than one in short.

10.1.1   Example

- Long a call with strike \$53 expiring on day X for cost \$4.35

- Short a call with \$61 strike with the same expiration for \$1.39

So the total cost is 4.35 - 1.39 = 2.96, which is equivalent to \$296. Note that, this is the **loss maximal** for the spread acquirer. This happens when the underlying price is lower then strike in the long (\$53)

Gain maximal : When the underlying price is surpass \$61. Then the gain is : 6100-5300 - 296 = \$504. Hence, the Bull Call Spread is a strategy whose gains and losses maximal are known in advance.

So what happed if the underlying price is between the two strike values ? Let's call x the underlying price. Then the gain is : $(x - 53 - 2.96) \times 100$. We have a break even at underlying price of \$55.96, means that if x is greater than \$55.96 then we have gain and if x is lower than \$55.96 then we have loss.

When used ?By using a bull call spread, we are essentially betting that the underlying asset's price will increase moderately but not surpass the higher strike price.

We have in following several types of spread trades:

10.2 Intracommodity spreads

Also known as a calendar spread, horizontal spread or time spread, is an options trading strategy that involves simultaneously buying and selling options contracts with the same characteristics (underlying, strike price, free-rate...), the only difference is the expiration date. More precisely, we sell the nearby expiration and buy the further-out expiration (option). **Calendar spread has an expiration date and it is typically one of nearby (near-term) option**. If we perform inversly (buy the nearby expiration and sell the further-out expiration), it is known as a *reverse calendar spread*.

The key idea behind a calendar spread is to take advantage of time value which decays. As time passes, the value of the shorter-term option will decline at a **faster rate of decay**, potentially resulting in a profit. More precisely, let C_n and C_f be the price of near-term option and further-out option and suppose that the spread value has time value which decay exponentially :

$$(C_n - InVa)e^{r_n t} - (C_f - InVa)e^{r_f t}$$

$InVa$ means intrinsec value, which is the same for both near-term and further-out option. We have the rate of decay r_n is greater than r_f since **the rate of decay is proportional to implied volatility** and **the closer expiration date the higher implied volatility**. The latter is because near-term options are more sensitive to near-term market events and uncertainties, leading to higher implied volatility.

10.2.1 Trading strategies

Pick expiration months : When selecting the expiration date of the long option, it is wise to go at least two to three months out. This will depend largely on your forecast. However, when selecting the short strike, it is a good practice to always sell the shortest dated option available.

10.2.2 Notes

- Risk (loss) is limited to the net debit, means the difference between the premium of two options.

- Benefits from an increase in impiled volatility.

11 Pairs trading

The idea behind pairs trading is to identify two assets that have a strong historical correlation, meaning that their prices tend to move together over time, but may diverge temporarily due to market fluctuations or other factors.

When one asset's price rises and the other's falls, traders can buy the falling asset and short sell the rising one, anticipating that the spread between them will eventually narrow and they will revert to their historical correlation. Traders may use statistical tools to help identify when the spread between two assets deviates from its average, and then execute the trades to take advantage of the expected convergence.

As with any trading strategy, pairs trading involves risks and requires careful analysis and risk management.

11.1 Intercommodity spreads

Intercommodity spreads are formed from two distinct but related commodities, reflecting the economic relationship between them.

Common examples are:

- The *crack* (crarking) spread between crude oil and one of its byproducts, reflecting the premium inherent in refining (chi phi cua qua trinh loc dau) oil (from) into gasoline, gas oil, or heating oil

- The *spark* spread between natural gas and electricity, for gas-fired power stations

- The *crush* spread between soybeans and one of its byproducts, reflecting the premium inherent in processing soybeans into soy meal and soy oil

12 Long and short position

Key differences between long and short positions include:

- Directional Expectation: Long positions anticipate price appreciation (inscreasing), while short positions anticipate price depreciation (decreasing).

- Ownership: Long positions involve owning the asset, whereas short positions involve borrowing to sell (need to buy them back and return them to the lender)

- Profit/Loss Calculation: In a long position, the profit is realized if the asset's value increases, while the loss is incurred if the value decreases. In a short position, the profit is realized if the asset's value decreases, while the loss is incurred if the value increases.

- Risk Profile: Long positions typically have limited risk, as the maximum loss is generally limited to the amount invested. Short positions, especially when selling borrowed assets, have potentially unlimited risk if the asset's value rises significantly.

13 Straddle, Butterfly, Strangle

13.1 Straddle

In finance, a straddle strategy involves two transactions in options on the same underlying, with opposite positions (call and put). As a result, it involves the purchase or sale of particular option derivatives that allow the holder to profit based on how much the price of the underlying security moves, regardless of the direction of price movement.

For example, we buy a call and put with same strike price and expiration date. There are two considered cases:

- If the stock price is close to the strike price at expiration of the options, the straddle leads to a loss (the premiums).

- However, if there is a sufficiently large move in either direction, a significant profit will result (after compensating the premiums).

A straddle is appropriate when an investor is expecting a large move in a stock price but does not know in which direction the move will be.

A straddle made from the purchase of options is known as a *long straddle, bottom straddle, or straddle purchase*, while the reverse position, made from the sale of the options, is known as a *short straddle, top straddle, or straddle write.*

13.2 Long straddle

As the above example. This strategy is a limited risk, since the most a purchaser may lose at maximum is the cost of both options. At the same time, there is theoretically unlimited profit potential.

When to use ? For example, company XYZ is set to release its quarterly financial results in two weeks. A trader believes that the release of these results will cause a large movement in the price of XYZ's stock, but does not know whether the price will go up or down. He can enter into a long straddle, where he gets a profit no matter which way the price of XYZ stock moves, if the price changes enough either way.

If the options are American, the stock is sufficiently volatile, and option duration is long, the trader could profit from both options. This would require the stock to move both below the put option's strike price and above the call option's strike price at different times before the option expiration date.

Option payoff diagram for a short straddle struck at K where the total cost of the two constitutent options is V
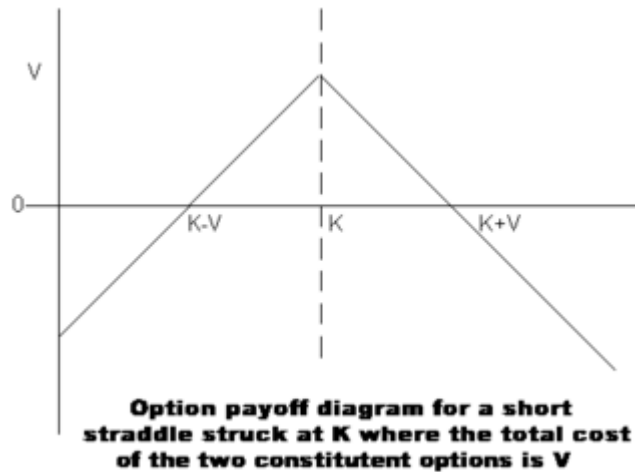
Figure 1.1: Short straddle, payoff in term of price evoluation.

13.3 Short straddle

A short straddle is a non-directional options trading strategy that involves simultaneously selling a put and a call of the same underlying security, strike price and expiration date.

The profit is limited to the premium received from the sale of put and call. The risk is virtually unlimited as large moves of the underlying security's price either up or down will cause losses proportional to the magnitude of the price move.

A maximum profit upon expiration is achieved if the underlying security trades **exactly** at the strike price of the straddle. In that case both puts and calls comprising the straddle expire worthless allowing straddle owner to keep full credit received as their profit.

This strategy is called "nondirectional" because the short straddle profits when the underlying security changes little in price before the expiration of the straddle.

The short straddle can also be classified as a *credit spread* because the sale of the short straddle results in a credit of the premiums of the put and call.

14 Contango et Backwardation

14.1 Contango

Contango is a situation where the futures price (or forward price) of a commodity is higher than the expected spot price of the contract at maturity (figure 1.2).

Here we note some informations about price :

- Spot price : The spot price refers to the current market price of a particular asset or commodity that is available for immediate delivery or settlement, as opposed to a future price that reflects the expected price at a future date

- Expected spot price at maturity : In general, there is no mathematical formula for expected spot price. It is more of an economic concept rather than a mathematical part. Keywords : Fundamental analysis, technical analysis, Monitor news and events, market sentiment, forecasting models, ...

- Future price : **In forward contracts or futures contracts**, where contract terms are agreed now but delivery and payment will occur at a future date.

- As definition of contango, **Future price and expected spot price are two thing differents**. The main difference is the two time instants : one at contract terms are agree and other at the payment. In the expected spot price, these two instants are the same in the future and in future price, one instant is now and other is in the future.

In a weaker definition, it's a situation where the cost of buying a commodity in the future is higher than the cost of buying it today.

Contango occurs when there is a higher demand for commodities in the future than there is currently, or when there is an expectation of a future shortage of the commodity. It can also occur due to the costs of storage and financing, as holding a commodity for delivery in the future involves additional costs such as storage fees and the cost of financing the purchase.

When a market is in contango, it creates an opportunity for traders to buy the commodity in the spot market and sell it in the futures market, with the expectation of making a profit when the futures contract expires. However, the potential profit from this strategy is limited by the costs of storage and financing, as well as any changes in the market conditions that may affect the future price of the commodity.

14.2 Backwardation

Backwardation is the opposition of contango. In other words, the spot price of a commodity is higher than the price of the futures contract for that same commodity.

This can occur when there is a shortage (thieu, pénurie) of the commodity, or when there is high demand for the commodity in the near term. In these situations, buyers are willing to pay more for the commodity right now, and sellers are willing to accept a lower price for delivery at a future date when they expect the supply to increase.
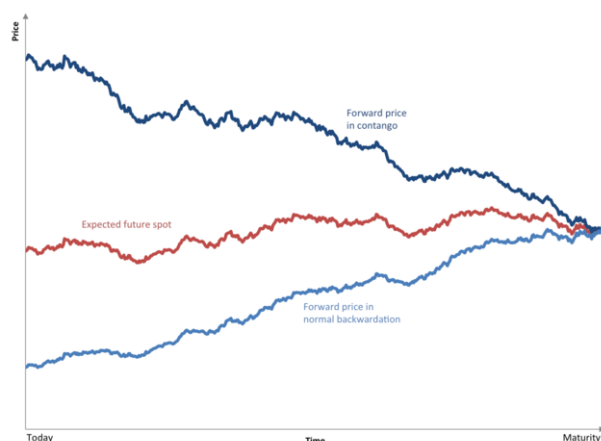
Figure 1.2: Contango and backwardation.

15   Making continuous price for future contracts

The expiration dates of the future contracts vary based on the underlying asset and range from monthly to quarterly, therefore there are no price of this asset type after these date. When expiration approaches, traders will roll their positions over to the next contract or let them expire. A basic approach to construct a continuous series would be to always use the front month contract's price and when the current front month contract expires, switch to the new front month contract.

However, there is one caveat, the price of the contracts when you rollover (reinvest) may not be the same, and in general, won't be the same. To properly evaluate the trading strategies with historical futures contract data, it is necessary to combine these contracts into a continuous price series. There are several methods:

15.1   Shift adjustment

Forward adjusting would shift the next contract to eliminate the gap by **subtracting** the adjustment factor from the next contract's price series. Backward adjusting would shift the previous contract to eliminate the gap by **adding** the adjustment factor to the previous contract's price series. Figure 1.3 below shows an example of these adjustments for an actual rollover. In this case, the size of this gap can be also called the adjustment factor.

15.2   Proportional adjustment

In this case, the gap is removed by using **multiplication**. Then the adjustment factor is :
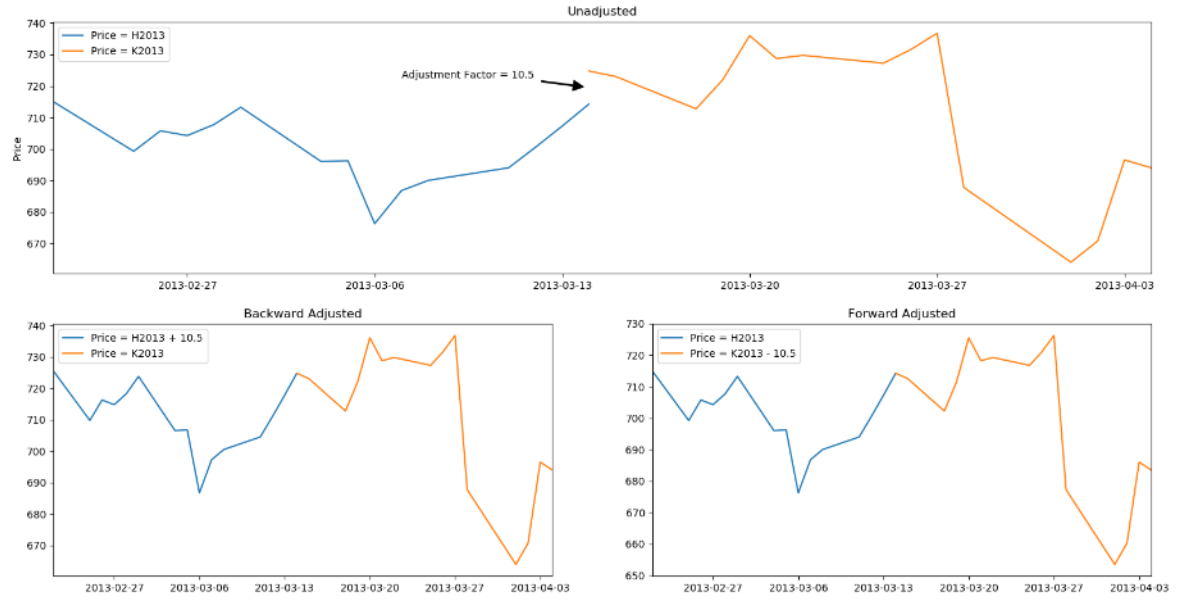
Figure 1.3: Shift adjustment, with backward and forward technical.

$$\frac{PriceAfter}{PriceActual}$$

This is calculated at the maturity of the price contract actual.

15.3 Perpetual method

Smooths the transitions between contracts by taking a weighted average of the contracts' prices during the transition period. The weight can be used based on :

- time left to expiration, with 5 rolling days: $p_1 = 0.8x_1 + 0.2x_2, p_2 = 0.6x_1 + 0.4x_2, ..., p5 = 0x_1 + 1x_2 = x_2$

- volume $w_1 = \frac{v1}{v1+v2}, w_2 = \frac{v2}{v1+v2}$

- open interest

- other properties of the contracts

15.4 Notes

- P&L : Preserverd in but not in Proportional adjustment

- Return rate : Preserverd in Proportional adjustment but not in Shift adjustment

The further we are from a contract's expiration; the more price speculation (dau co) is embedded into the price. The front or nearest month contract refers to the contract which has the soonest expiration date and thus has the least amount of speculation. Generally, front month contracts have the most trading activity.

16 Open interest

Open interest OI (Hop dong mo) is applicable for options contracts, also future contract in the derivatives market. It equals to the *number of (futures) contracts or positions (options) that are currently active.*

Examples:

Sharon, Cynthia and Kurt are trading on the same future contract. If Sharon buys a contract to enter a long-term trade, then OI will increase by 1. Cynthia also buys 6 contracts, then the OI will total 7. Now, if Kurt decides Going to short (ban khong) the market and sell three contracts, the open interest increased to 10 again. But if Sharon sell to Cynthia a contract, then the OI is still 10. (but the volume now is 11).

Rising open interest is often indicative of a strong developing price trend (bull market). This means that traders are supporting the trend by stepping up into the market. And vice versa, it will be a signal of a trend decline when traders lose confidence in the trend they leave the market (bear market).

Accordingly, Open Interest is an indicator that many traders in the futures market use in conjunction with other analysis to aid in trading decisions.

17 Price-earnings ratio

Called simply P/E :
$$\frac{Shareprice}{Earningpershare}$$

18 Market to book

Market-to-book (M/B) is a financial ratio that compares a company's market value to its book value (gia tri so sach). It is calculated by dividing the market capitalization of a company by its book value. The market value is the current stock price of a company's outstanding shares, while the book value is the value of a company's assets, minus its liabilities, as recorded on its balance sheet. M/B ratio is used to assess a company's financial performance and value, and it provides insight into the growth potential and future prospects of a company.

19   Year-to-date

Year-to-date (YTD) return is a financial metric that measures the performance of an investment or portfolio from the beginning of the current calendar year up to the present date. It provides a snapshot of the investment's performance over the specified period.

Here's an example to illustrate the calculation:

Let's say you invested $10,000 in a stock on January 1st (which is a tradable day) of the current year, and the current value of your investment is $12,000. The YTD return would be:

YTD Return = ($12,000 - $10,000) / $10,000 * 100 = 20%

20   Value stocks

Value stocks are stocks that are believed to be undervalued by the market, meaning that their price is lower than their intrinsic value. In other words, investors believe that the stock's current price does not reflect the company's actual worth, and that the stock is therefore a good investment opportunity. Value stocks are often found in industries that are currently out of favor or have low growth prospects. Examples of value stocks may include companies that have a strong balance sheet, generate steady cash flows, pay dividends, and have a low price-to-earnings (P/E) ratio.

21   Growth stocks

Growth stocks, on the other hand, are stocks of companies that are expected to grow at a faster rate than the overall market or their peers. These companies may be in industries that are experiencing high growth, have innovative products or services, or have a competitive advantage. Growth stocks are often characterized by a high P/E ratio, as investors are willing to pay a premium for the potential future growth of the company. Examples of growth stocks may include technology companies, biotech companies, and companies in emerging markets.

22   Cost of goods sold

The cost of goods sold (GOGS, gia von hang ban) is the sum of all direct cost associated with making a product. It appears on an income statement and typically includes money spent on raw materials and labour. It does not include coss associated with marketing, sales or distribution.

23 Operating expense

An operating expense, operating expenditure, operational expense, operational expenditure or opex (chi phi hoat dong) is an ongoing cost for running a product, business, or system.

Its counterpart, a capital expenditure (capex, chi phi von), is the cost of developing or providing non-consumable parts for the product or system.

For example, *the purchase of a photocopier involves capex*, on the other hand, *the annual paper, toner, power and maintenance costs represent opex.*

For larger systems like businesses, opex may also include the cost of workers and facility expenses such as rent and utilities.

24 Underlying instrument

In finance, the underlying instrument refers to the financial asset or security that serves as the basis for a derivative contract or other financial product. The value of the derivative product is derived from the underlying instrument, and its price movements are closely tied to the performance of the underlying asset.

Examples of underlying instruments include stocks, bonds, currencies, commodities, and indices. For example, a stock option is a derivative contract that gives the holder the right to buy or sell shares of a particular stock (the underlying instrument) at a predetermined price and time.

25 Income and outcome

Income refers to the **money** or other **financial gains** received by an individual or organization, typically from employment, business activities, investments, or other sources. Income can be earned in the form of wages, salaries, commissions, bonuses, rental income, investment income, or other forms of compensation.

Outcome, on the other hand, refers to the **results** or **consequences**, in a broad sense, of a particular action or decision.

Example : in business, we want launch a new product. The *ressources* is money for materials, the *action or activity* is the developpement... Then the *output* is the result (product). The *outcome* is the consequence of the output, e.g. the user of this product get better. Finally, the gain of selling this product is *income.*

25.1 Difference between output and outcome

To illustrate the difference between output and outcome, let's take an example of a non-profit organization providing food assistance to a community. The number of food packages distributed to the community would be an output, while the improvement in the nutritional status or food security of the recipients would be an outcome.

26 Arbitrage and Speculation

It is the practice of taking advantage of a difference (discrepancy) in prices in two or more markets. It strikes (makes) a combination of matching deals to capitalise (tan dung) on this difference. Then the profit being the difference between the market prices at which the unit is traded.

In simple terms, it is the possibility of a risk-free profit after transaction costs. Arbitrageurs buy an asset in one market where it is priced lower and sell it in another market where it is priced higher, profiting from the price difference. Arbitrageurs do not take on risk because they are guaranteed to make a profit as long as the price difference exists and they can execute the trades fast enough to capture the opportunity.

When used by academics, an arbitrage is a transaction that involves :

- No negative cash flow at any probabilistic or temporal state.

- At least one state, a positive cash flow.

In academic use, an arbitrage is risk-free

26.1 Arbitrage-free

Arbitrage-free refers to a condition in financial markets where there is no opportunity for riskless profit by exploiting price differences between similar financial instruments.

An arbitrage equilibrium is a precondition for a general economic equilibrium. The "no arbitrage" assumption is used in quantitative finance to calculate a unique risk neutral price for derivatives.

26.2 Speculation

Speculation involves buying or selling assets in the hope of making a profit from changes in their prices. Speculators take on risk and are exposed to the possibility of losing money if their predictions do not come true. Speculators may use a variety of methods to try to predict future price movements, such as technical analysis, fundamental analysis, or market sentiment.

In summary, speculation involves taking risks in the hope of making a profit, while arbitrage involves exploiting price discrepancies in different markets to make a guaranteed profit with no risk.

27 Risk neutral

Risk neutral is a term commonly used in finance and economics to describe a situation or perspective in which an individual or organization does not consider risk when making decisions.

In a risk-neutral environment, the value of an investment or financial asset is determined solely by its expected return, without taking into account any risks associated with that investment. This means that a risk-neutral person would be indifferent between two investments with the same expected return, even if one is riskier than the other.

The concept of risk neutrality is often used in financial models and pricing derivatives. It assumes that the market participants are willing to accept any level of risk, as long as they are adequately compensated for it through expected returns. This assumption allows for simplified calculations and easier comparison of different investment opportunities.

However, in reality, most individuals and organizations have varying levels of risk aversion and will take into account the potential risks and uncertainties associated with an investment before making a decision.

27.1   Diffence between risk-free and risk-neutral

Risk-free investments are those with no risk of loss, meaning that the return on the investment is guaranteed. Typically, these investments have a low rate of return, such as savings accounts or government bonds. Investors choose risk-free investments when they prioritize capital preservation over capital growth and want to minimize the risk of losing money.

On the other hand, risk-neutral investments do have some level of risk, but the expected return is equal to the risk assumed. In other words, investors are compensated for taking on risk. For example, stocks, mutual funds, and corporate bonds are considered risk-neutral investments because they offer the potential for higher returns but also come with the risk of losing money.

The key difference between the two is that risk-free investments have no risk of loss, while risk-neutral investments have some level of risk associated with them. Risk-neutral investors are willing to take on some risk for the potential of higher returns, while risk-averse investors prioritize the preservation of capital over the potential for growth.

27.2   P and Q measure

P measure and Q measure are probability measures used to model the behavior of financial instruments and markets.

P measure, also known as the physical measure, is a probability measure that describes the real-world probabilities of events. It is the measure that reflects the actual probability distribution of the returns of an underlying asset or portfolio.

Q measure, also known as the risk-neutral measure, is a probability measure that allows for a simpler pricing of derivatives and other financial instruments by removing the effects of risk aversion. It is the measure under which the expected value of a derivative is equal to its present value. The Q measure is constructed by adjusting the probabilities under the P measure so that the expected return on the derivative is the risk-free rate of interest.

In summary, the P measure is the actual probability measure of events, while the Q measure is a probability measure used for pricing derivatives and other financial instruments that accounts for the absence of risk premiums.

27.3   Risk premium

The risk premium is the extra return that an investor expects to receive for taking on additional risk compared to a "risk-free" investment such as a government bond. In finance, risk is typically defined as the variability of returns, and investors are generally risk-averse, meaning they prefer less risky investments.

The risk premium is calculated as the *difference between the expected return on a risky investment and the expected return on a risk-free investment with a similar maturity.* For example, if the expected return on a stock is 10%, and the risk-free rate is 3%, the risk premium would be 7%.

Investors demand a risk premium to compensate them for the potential losses that may arise from investing in a risky asset, such as a stock, bond, or mutual fund. The size of the risk premium will depend on the perceived risk of the investment, the investor's risk tolerance, and the prevailing market conditions.

27.4   Risk-neutral measure

Also called an equilibrium measure, or equivalent martingale measure. This means a **probability** measure such that *each share price is exactly equal to the discounted expectation of the share price* under this measure. This is heavily used in the pricing of financial derivatives due to the fundamental theorem of asset pricing, which implies that in a complete market (negligible transaction costs and every asset in every possible state of the world has a price), a derivative's price is the discounted expected value of the future payoff under the unique risk-neutral measure. Such a measure exists if and only if the market is arbitrage-free (khong co kinh doanh chenh lech gia).

27.5   Equilibrium pricing vs risk-neutral pricing

Equilibrium pricing is the pricing of financial assets based on the supply and demand of those assets in the market. In other words, the equilibrium price is the price at which the quantity of a financial asset demanded by investors equals the quantity of that asset supplied by issuers. The equilibrium price reflects the perceived value of the asset by market participants, taking into account various factors such as market trends, economic conditions, and investor sentiment.

On the other hand, risk-neutral pricing is a pricing method that assumes that investors are indifferent to risk when valuing financial assets. This means that the expected return on an asset is adjusted for the risk-free rate of return, and not for the risk associated with the asset. This method is commonly used in derivatives pricing, where the value of the derivative is based on the expected future value of the underlying asset, discounted at the risk-free rate.

Year	1	2	3	4	5
$(1 + \text{Discount Rate})^n$	$\frac{1}{(1+0.08)}$	$\frac{1}{(1+0.08)^2}$	$\frac{1}{(1+0.08)^3}$	$\frac{1}{(1+0.08)^4}$	$\frac{1}{(1+0.08)^5}$
Discounted value	925.93	857.34	793.83	735.03	680.58

Table 1.1: Discounted value per each year with payoff of $1000.

Therefore, while both equilibrium pricing and risk-neutral pricing are methods for pricing financial assets, they are based on different assumptions and take into account different factors.

28 Discounted rate

The discounted rate is used to determine the present value of future cash flows or investments. It takes into account the time value of money, which means that a dollar received or paid in the future is worth less than a dollar received or paid today.

28.1 Example

Suppose you have an investment opportunity that promises to pay you $1,000 per year for the next five years. However, you want to determine the present value of these cash flows today, considering a discount rate of 8%.

To calculate the present value for each year's cash flow, you can use the formula:

Present Value = Future Value / $(1 + \text{Discount Rate})^n$

where n is the number of years from the present.

Let's calculate the present value for each year's cash flow, in table 1.1.

Total Present Value = 925.93 + 857.34 + 793.83 + 735.03 + 680.58 = 3,992.71. This means this opportunity costs 3,992.71 at the present.

28.2 Estimation discounted rate

Usually, a discount rate value is the same as the rate of return, hence it can be estimated by :

- By using the risk premium, then added by risk free rate.

- By using return rate estimated by CAPM in section 8.

- By using return rate estimated by DDM in section 9.

- By using required return rate R_e infered from WACC in section 12, given $WACC, R_d, E, D, t$.

29 The Q world and the P world

The Q world and the P world are two different mathematical frameworks used in finance to model the behavior of financial assets and their prices.

The Q world, also known as the **risk-neutral** world, is a theoretical framework in which all investors are assumed to be risk-neutral, meaning that they do not care about the risk of an investment, only its expected return. In the Q world, the expected return on an investment is equal to the **risk-free rate of return**. This assumption allows for simpler calculations of the value of financial derivatives, such as options, and is widely used in financial engineering.

The P world, also known as the real-world or physical world, is the framework that reflects the actual behavior of financial markets and investors, who are **not necessarily risk-neutral**. In the P world, the expected return on an investment is equal to the **risk premium**, which is the additional return that investors demand to compensate for the risk they are taking. This framework is used to model the behavior of financial assets and their prices in the real world.

Both frameworks have their advantages and limitations, and their use depends on the specific application and the assumptions that are appropriate for the situation at hand.

30 Asian option

An Asian option is a type of option where the payoff depends on the average price of the underlying asset over a predetermined period of time. This is different from a US option and a European option, where the payoff depends solely on the price of the underlying asset at a specific point in time. For exemple, a call option payoff :

$$\max\left(A(0,T) - K, 0\right)$$

where $A(0,T)$ can be:

- In the continuous case, arithmetic average :

$$A(0,T) = \frac{1}{T}\int_0^T S(t)dt$$

- In the discrete case :

$$A(0,T) = \frac{1}{n}\sum_{i=1}^n S(i)$$

- Geometric average :

$$A(0,T) = \exp\left(\frac{1}{T}\int_0^T \log[S(t)]dt\right)$$

31 Moneyness

In finance, moneyness is the relative position of the current price (or future price) of an underlying asset (e.g., a stock) with respect to the strike price of a derivative. Moneyness is firstly a three-fold classification:

- If the derivative would have positive intrinsic value if it were to expire today, it is said to be in the money

- If the derivative would be worthless if expiring with the underlying at its current price, it is said to be out of the money

- And if the current underlying price and strike price are equal, the derivative is said to be at the money.

There are two slightly different definitions, according to whether one uses the current price (spot) or future price (forward), specified as "at the money spot" or "at the money forward".

32 Cash-flow

Cash flow refers to the movement of cash in and out of a business or individual's finances over a specific period of time. It is the difference between the amount of cash received and the amount of cash spent by a business or individual during a given time period. Positive cash flow means that more cash is coming in than going out, while negative cash flow means that more cash is going out than coming in.

Cash flow is an essential component of financial management because it provides insights into the financial health of a business or individual. It allows them to monitor their cash inflows and outflows, plan for future expenses and investments, and make informed decisions about borrowing, saving, and spending. Cash flow can also help a business or individual determine their ability to pay bills, meet financial obligations, and generate profits.

33 Order book and high frequency trading

33.1 Order book

An order book is a record of all buy and sell orders for a particular financial asset, such as stocks, bonds, or derivatives, that are currently open and waiting to be executed. It is essentially a list of all the current bids to buy and asks to sell a particular asset at various prices.

The order book provides traders and investors with important information about market activity, including the current supply and demand for a particular asset, as well as the prices that buyers and sellers are willing to accept. By

looking at the order book, traders can see where the market is headed and make informed decisions about when to buy or sell an asset.

In addition to displaying current buy and sell orders, the order book may also include information about recent trades that have taken place. This can provide valuable insights into the current state of the market and the behavior of other traders and investors.

Order books can be accessed by traders and investors through various platforms, including online trading platforms and trading terminals provided by financial institutions. They are a critical tool for anyone looking to trade in financial markets, as they provide valuable information about the supply and demand dynamics of the market.

33.2   High frequency trading

High frequency trading (HFT) is a type of automated trading that uses sophisticated algorithms and computer programs to execute large volumes of trades at extremely high speeds. HFT strategies typically involve analyzing market data in real-time and making split-second decisions to buy or sell securities.

The main goal of HFT is to profit from small price discrepancies that exist for only a fraction of a second. HFT firms use advanced technology and ultra-fast networks to ensure that they receive market data and execute trades before their competitors. This can result in HFT firms earning profits in a matter of microseconds.

While HFT can be highly profitable, it is also controversial. Some critics argue that HFT creates instability in the financial markets, and that it can lead to increased volatility and market crashes. Others argue that HFT provides liquidity to the market and helps to reduce bid-ask spreads, which can benefit all market participants.

Note that, HFT is only important if we trade a big volume of stock.

33.3   Bid-ask spread

The bid-ask spread is the difference between the highest price a buyer is willing to pay for a financial asset (the bid price) and the lowest price a seller is willing to accept for that same asset (the ask price). In other words, it is the difference between the highest price a buyer is willing to pay for an asset and the lowest price a seller is willing to accept.

For example, if the bid price for a stock is $100 and the ask price is $101, then the bid-ask spread is $1. The bid-ask spread represents the transaction costs of trading a particular asset and is an important factor to consider when buying or selling financial instruments.

The size of the bid-ask spread can vary depending on a number of factors, including the liquidity of the market, the volume of trading, and the volatility of the asset. In general, assets with higher trading volumes and greater liquidity

tend to have smaller bid-ask spreads, while assets with lower trading volumes and less liquidity tend to have larger bid-ask spreads.

Traders and investors must consider the bid-ask spread when buying or selling an asset, as it represents a cost that must be paid to complete the transaction. The larger the spread, the more expensive it is to trade an asset, which can impact overall profitability.

34   Survivorship bias

In finance, let's consider before and after dotcom bubble, the stocks in S&P 500 have changed a lot. Then algo made before this moment **may not be compatible** with things after this moment.

35   Efficient market hypothesis

The efficient-market hypothesis (EMH) is a hypothesis in financial economics that states that asset prices reflect all available information. A direct implication is that it is impossible to "beat the market" consistently on a risk-adjusted basis since market prices should only react to new information.

Hedge funds work with condition that this hypothesis is not always true. E,g in dotcom periode, stocks are overvalued.

36   Network effect

In economics, a network effect (also called network externality or demand-side economies of scale) is the phenomenon by which the value or utility a user derives from (given from) a good or service depends on the number of users of compatible products.

Network effects are typically *positive*, resulting in a given user deriving more value from (is given more value from) a product as more users join the same network.

The adoption of a product by an additional user can be broken into two effects:

- Total effect : An increase in the value to all other users.

- Marginal effect : Enhancement of other non-users' motivation for using the product

Network effects can be direct or indirect:

- Direct network effects arise when a given user's utility increases with the number of other users of the same product or technology, meaning that adoption of a product by different users is complementary. Direct network effects can be seen with social networking services, including Twitter, Facebook, Airbnb, Uber, and LinkedIn; telecommunications devices

like the telephone. E.g., utility of an user Airbnb increases when more people use this app, which means (maybe) having more reviews about an accommodation, or having more chooses if users are property... This effect is separate from effects related to price, such as a benefit to existing users resulting from price decreases as more users join.

- Indirect network effects (or cross-group) arise when there are "at least two different customer groups that are interdependent, and the utility of at least one group grows as the other group(s) grow". For example, the product is OS (Windows, Ubuntu), hence, hardware consumers may become more valuable with the growth of software consumers. Other ex, Airbnb, one cutomer group is landlord and other one is tenant.

Network effects are commonly mistaken for economies of scale, which describe decreasing average production costs in relation to the total volume of units produced. Economies of scale are a common phenomenon in traditional industries such as manufacturing, whereas network effects are most prevalent in new economy industries, particularly information and communication technologies. Network effects are the demand side counterpart of economies of scale, as they function by increasing a customer's willingness to pay due rather than decreasing the supplier's average cost.

Upon reaching critical mass, a bandwagon effect can result. As the network continues to become more valuable with each new adopter, more people are incentivised to adopt, resulting in a positive feedback loop. Multiple equilibria and a market monopoly are two key potential outcomes in markets that exhibit network effects. Consumer expectations are key in determining which outcomes will result.

37   Computational finance and financial engineering

Computational finance involves the use of *computer algorithms* and models to analyze financial data, prices, and risk. It focuses on *developing and implementing* numerical methods for pricing and hedging financial instruments, such as options, futures, and derivatives. This field encompasses a wide range of techniques, including Monte Carlo simulation, optimization, machine learning, and statistical inference.

Financial engineering, on the other hand, is the *application* of mathematical models and quantitative methods to design and create financial instruments and strategies. It involves the development and implementation of financial products, such as structured products, securitization, and risk management techniques. Financial engineers use mathematical models to understand the behavior of financial markets and create innovative financial solutions to meet the needs of investors and clients.

Both computational finance and financial engineering require a strong background in mathematics, statistics, and programming, as well as knowledge of

financial markets and instruments. These fields are used extensively in the financial industry for pricing and hedging complex financial instruments, risk management, portfolio optimization, and trading strategies. They also have applications in academia, where they are used to research and develop new financial theories and models.

38   Two-sided Market

A two-sided market (thi truong hai mat hay thi truong luong dien), also called a two-sided network, is an **intermediary economic platform** having **two distinct user groups** that provide each other with network benefits. The organization that creates value primarily by enabling direct interactions between two (or more) distinct types of affiliated customers is called a multi-sided platform. This concept of two-sided markets has been mainly theorised by the French economists Jean Tirole and Jean-Charles Rochet and Americans Geoffrey G Parker and Marshall Van Alstyne.

Two-sided networks can be found in many industries, sharing the space with traditional product and service offerings :

- Credit cards : Cardholders and Merchants

- Health maintenance organizations (patients and doctors)

- Operating systems (end-users and developers)

- Yellow pages (advertisers and consumers)

- Video-game consoles (gamers and game developers)

- Recruitment sites (job seekers and recruiters)

- Search engines (advertisers and users)

In France, for example, it's leboncoin.

39   Implemtation package on quant

- Quantopian

- Pyfolio : Providing performance and risk analysis of financial portfolio.

- Alphalens : Predive stock package (alpha)

- Numerai

- Quantiacs

- Zipline (Backtesting engine)

Chapter 2

Probability

1 Random variable

We discover first some concepts that is needed for the definition of random variable, notably the σ-algebra and the measure space.

1.1 σ-algebra

In mathematical analysis and in probability theory, a $\sigma-$algebra (also $\sigma-$field) on a set X, noted Σ, is a collection of subsets of X that satifies :

1. $\varnothing \in \Sigma$.

2. $A \in \Sigma \Rightarrow A^C (\text{ in } X) \in \Sigma$. (*closed under complementation*)

3. $A_1, A_2, \ldots, A_\infty \in \Sigma \Rightarrow \bigcup\limits_{i=1}^{\infty} A_i \in \Sigma$. (*closed under countable unions*)

From the second and the third property, we can infer the useful property :

$$A_1, A_2, \ldots \in \Sigma \Rightarrow A_1^C, A_2^C, \ldots \in \Sigma \Rightarrow \bigcup_i A_i^C \in \Sigma \Rightarrow \left(\bigcup_i A_i^C \right)^C \in \Sigma \text{ or } \bigcap_i A_i \in \Sigma$$

This is also called countable intersection.

If $X = \{a, b, c, d\}$ then one possible $\sigma-$algebra on X is

$$\Sigma = \{\varnothing, \{a, b\}, \{c, d\}, \{a, b, c, d\}\}$$

σ-algebra can be applied for subset or random variable:

- $\sigma(A)$, where A is a non-empty collection of subset in Ω, is the unique smallest $\sigma-$algebra that contains every elements of A.

- $\sigma(X)$, where X is a random variable (we will see this concept after in sec 1.6.4).

1.1.1 Example

What is the $\sigma(R, S)$ if $\Omega = \{1, 2, 3, 4, 5\}, R = \{1, 2, 3\}$ and $S = \{2, 4\}$?
We first find the $\sigma(R), \sigma(S)$:

- $\sigma(R) = \{\varnothing, \Omega, R, R^C\}$

- $\sigma(S) = \{\varnothing, \Omega, S, S^C\}$

Then $\sigma(R, S)$ must contain:

- R, R^C, S, S^C

- $R \cap S = \{2\}$

- $R \cap S^C = \{1, 3\}$

- $R^C \cap S = \{4\}$

- $R^C \cap S^C = \{5\}$

Since $\{2\}, \{1, 3\}, \{4\}, \{5\}$ generate R, R^C, S, S^C; then :

$$\sigma(R, S) = \sigma\left(\{\{2\}, \{1, 3\}, \{4\}, \{5\}\}\right)$$

Note that $\{2\}, \{1, 3\}, \{4\}, \{5\}$ *are disjoint sets that their union forms* Ω then each union between two sets, three sets, four sets, themself and the empty set form $\sigma(R, S)$, which has $2^4 = 16$ elements.

1.1.2 Algebra

In some lectures, we can find the concept named algebra. The definition of an algebra is the same as σ-algebra except the third condition :

1. $\varnothing \in \Sigma$.

2. $A \in \Sigma \Rightarrow A^C (\text{ in } X) \in \Sigma$. (*closed under complementation*)

3. $A_1, A_2, \ldots, A_n \in \Sigma \Rightarrow \bigcup_i^n A_i \in \Sigma$. (*closed under* **finite** *unions*)

Here is a simple example for the difference on the third condition between an algebra and a σ-algebra. Let's consider closed intervals $[\frac{1}{n}, 1 - \frac{1}{n}]$, then a **finite** union of these intervals is also a closed interval. However, a **countable** union can be an infinity union, then the result is (0,1), which is a open interval.

σ-algebra is well an algebra, thus it is an extension of algebra. Then the converse is not true, let's see the following example.

1.1.2.1   Example

Consider a set X of natural numbers, i.e. $X = N$ and the algebra Σ on it that consists of all subsets of X with the property that either the set or its complement has finitely many elements :

$$\Sigma = \{A \in \mathbb{N} \colon |A| \text{ or } |A^C| \text{ is finite}\}$$

Σ is well an algebra, but it is not a σ-algebra. This is because if we take $A_i = \{2i\}, \forall i \geq 0$, thus A_i is a singleton set that contains a even number, then the *countable* union condition of A_i infers that set E contains all even numbers must be in Σ. From here, by relative complement, set $O = E^C$ contains all odd numbers must be in Σ. However, $|E| = \infty$ and $|O| = \infty$, then E can not be in Σ.

1.1.3   Borel set

Given a space X, a Borel set is any set that can be formed from open sets (or, equivalently, from closed sets) through the operations of *countable* union, *countable* intersection, and relative complement.

The collection of all Borel sets in this space X forms a σ-algebra, known as the Borel algebra, denoted as $B(X)$. This Borel algebra on X is the smallest $\sigma-$algebra containing all open sets of X.

1.1.3.1   Example

Given the real number line ($X = \mathbb{R}$), then the collection of all Borel sets (or intervals), which are formed from open intervals in $\mathbb{R}$, is a Borel algebra on $\mathbb{R}$, denoted $B(\mathbb{R})$.

1.2 π-system

In mathematics, a π-system (or pi-system) on a set Ω is a collection P of certain subsets Ω, such that :

- P is non-empty

- $A, B \in P$ then $A \cap B \in P$

$\sigma-$algebra is $\pi-$system : since $A \cap B = (A^C \cup B^C)^C$ is an element in σ-algebra. However, $\pi-$system is not $\sigma-$algebra.

If Σ is a non-empty collection of subsets in Ω then $\mathcal{I}(\Sigma)$ or $\mathcal{I}_\Sigma$ is the unique smallest π-system that contains every elements of Σ. We have explicitly its formule :

$$\mathcal{I}_\Sigma = \{E_1 \cap \ldots \cap E_n \ : \ 1 \leq n \in \mathbb{N} \text{ and } E_1, \ldots, E_n \in \Sigma\}$$

1.3 Dynkin system

1.3.1 Definition

Let Ω be a non-empty set, and let D be a collection of subsets of Ω. Then D is a Dynkin system if

- $\Omega \in D$

- If $A \in D$ then $\Omega \setminus A \in D$

- If $A_1, A_2, A_3, \ldots$ is a sequence of pairwise disjoint sets in D ($A_i \cap A_j = \varnothing, \forall i \neq j$) then $\bigcup_{i=1}^{n} A_i \in D, \forall n = 1, \ldots, \infty$.

Dynkin systems are sometimes referred to as λ-systems or d-systems. Interesting properties:

- σ-algebra is λ-system

- If a set is both λ-system and π-system, then it is a σ-algebra. In general, λ-system is not π-system and vice versa.

1.3.2 Dynkin's π-λ theorem

$$\begin{cases} P \text{ is a } \pi\text{-system} \\ D \text{ is a } \lambda\text{-system} \\ P \subseteq D \end{cases} \Rightarrow \sigma(P) \subseteq D$$

Proof http://theanalysisofdata.com/probability/E_3.html

1.4 Measurable space

Consider a set X and a σ-algebra Σ on X. Then the tuple (X, Σ) is called a measurable space. No measure (or empty subset $\varnothing$) and X are needed in the σ-algebra Σ.

Examples Given the set: $X = \{1, 2\}$:

- $\Sigma_1 = \{X, \varnothing\}$. Then (X, Σ_1) is a measurable space.

- $\Sigma_2 = 2^X = \{X, \varnothing, \{1\}, \{2\}\}$ (powerset of X). Then (X, Σ_2) is also a measurable space.

DO NOT confuse **measurable space** with **measure space** in sec 1.8.

1.5 Measurable function

1.5.1 Definition

Let (Ω, Σ) and (Y, T) be measurable spaces. A function $f : \Omega \to Y$ is said to be measurable on (Σ, T) if for every $E \in T$, the pre-image of E under f is in Σ :

$$f^{-1}(E) := \{\omega \in \Omega \mid f(\omega) \in E\} \in \Sigma, \quad \forall E \in T.$$

Note that f^{-1} here take input value of a $\sigma-$algebra, which is a set of elements of Ω, which means : $f^{-1} : \sigma-$algebra of Y $\to \sigma-$algebra of ω.

1.5.2 Role of measurable function

A measurable function has the ability to "transport" a measure defined for the domain's measurable space (X, Σ) to the codomain's measurable space (Y, T).

1.5.3 $\mathcal{F}$-measurability

Let $\mathcal{F}$ be an arbitrary $\sigma-$algebra of Ω. $\mathcal{B}$ is Borel set (sec 1.1.3) on $\mathbb{R}$. A random variable or a map $X : \Omega \to \mathbb{R}$ is said to be $\mathcal{F}$-**measurable** for short or $(\mathcal{F}, \mathcal{B})$-measurable for long if for all $S \in \mathcal{B}$, the pre-image (i.e. the inverse image) under X belongs to $\mathcal{F}$, that is :

$$X^{-1}(S) \in \mathcal{F}, \quad \forall S \in \mathcal{B}$$

or

$$\{\omega \colon X(\omega) \in S\} \in \mathcal{F}, \quad \forall S \in \mathcal{B}$$

From the two above definitions ($\mathcal{F}$-measurability and σ-algebra generated), X is $\mathcal{F}$-measurable if and only if $\sigma(X) \subset \mathcal{F}$.

In short, X is $\mathcal{F}$-measurable if all pre-images of X are in $\mathcal{F}$.

1.5.3.1 Example 1

We use the example 1.6.4.2, if $\mathcal{F} = \{\Omega, \varnothing, \{1, 2, 3\}, \{4, 5, 6\}, \{1, 4\}\} \supset \sigma(X)$, then X is $\mathcal{F}$-measurable.

1.5.3.2 Example 2

Let $\Omega = \{-1, 0, 1\}$ and $\mathcal{F} = \{\varnothing, \Omega, \{-1, 1\}, \{0\}\}$:

- $X_1(\omega) := \omega$ is NOT $\mathcal{F}$-measurable. Since $X^{-1}(\{1\}) = \{1\} \notin \mathcal{F}$.

- $X_2(\omega) := \omega^2$ is $\mathcal{F}$-measurable.

1.5.3.3 Lebesgue measurable function

It means a measurable function $f\colon (\mathbb{R}, \mathcal{L}) \to (\mathbb{R}, B(\mathbb{R}))$, where $\mathcal{L}$ is the σ-algebra of **Lebesgue measurable sets** (sec 1.7.2.1 and 1.7.3.3) and $B(\mathbb{R})$ is the Borel algebra (1.1.3) on $\mathbb{R}$.

In the general case, the domain of f is not necessary $\mathbb{R}$ and we call X this domain, which means $f\colon X \to \mathbb{R}$. Then f is Lebesgue measurable function if and only if the set

$$\{f > \alpha\} = \{x \in X : f(x) > \alpha\}$$

is Lebesgue measurable for all $\alpha \in \mathbb{R}$. This is also equivalent to any of $\{f \geq \alpha\}, \{f < \alpha\}, \{f \leq \alpha\}$ being Lebesgue measurable for all $\alpha \in \mathbb{R}$, or the preimage of any open set being Lebesgue measurable. Continuous functions, monotone functions, step functions, semicontinuous functions, Riemann-integrable functions, and functions of bounded variation are all Lebesgue measurable.

Lebesgue measurable functions are of interest in mathematical analysis because they can be integrated, which means with Lebesgue measure L, then

$$\int_{[a,b]} f(x)dL(x) = \int_{[a,b]} f(x)dx$$

1.6 Random variable

A random variable is a **measurable function** that maps from the measurable space of sample space and event space $(\Omega, \mathcal{F})$ to a measurable space (Y, T).

1.6.1 Why measurable function ?

It is because we want that random variables are available to transport the **probability measure**.

1.6.2 Examples

Example 1 : Let's $\Omega = \{Head, Tail\}$, $Y = \{0, 1\}$ and a function $X : \Omega \to Y$ that:

$$\begin{cases} X(Head) = 0 \\ X(Tail) = 1 \end{cases}$$

In order to X be a random variable, we need to define two measurable spaces (domain and codomain) that X is measurable function :

- Domain : $(\Omega, \mathcal{F})$ where $\mathcal{F} = \{\varnothing, \Omega, \{Head\}, \{Tail\}\}$.

- Codemain : (Y, T) where $T = \{\varnothing, Y, \{0\}, \{1\}\}$.

Note that we precise $X^{-1}(\varnothing) = \varnothing$ and $X^{-1}(Y) = \{\omega \in \Omega \mid X(\omega) \in Y\} = \{Head, Tail\} = \Omega$.

Moreover, the two following measurable spaces can be used :

- Domain : $(\Omega, \mathcal{F})$ where $\mathcal{F} = \{\varnothing, \Omega\}$.

- Codemain : (Y, T) where $T = \{\varnothing, Y\}$.

It means that, given a function, it is not unique a couple of (domain, codomain) that makes this function a random variable.

Example 2 Using the above example, but with two measurable spaces :

- Domain $(\Omega, \mathcal{F})$ where $\mathcal{F} = \{\varnothing, \Omega\}$.

- Codemain : (Y, T) where $T = \{\varnothing, Y, \{0\}, \{1\}\}$.

This is non measurable function because $\{1\} \in T$ and $\{f^{-1}(\{1\})\} = \{Tail\} \notin \Sigma$.

Example 3 Uniforme distribution transform. Let's consider two measurable spaces : $([0,1], \sigma([0,1]))$ and $([c,d], \sigma([c,d]))$. Z is a function $[0,1] \to [c,d]$ such that :

$$Z(x) = (d-c)x + c, \forall x \in [0,1]$$

Then X is a random variable.

Example 4 Box–Muller transform. Let's consider two measurable spaces : $[0,1] \times [0,1], \sigma([0,1] \times [0,1])$ and $\mathbb{R}, \sigma(\mathbb{R})$. $Y_1 : [0,1] \times [0,1] \to \mathbb{R}$ and $Y_2 : [0,1] \times [0,1] \to \mathbb{R}$ such that :

$$Y_1 = R\cos(\Theta) = \sqrt{-2\ln U_1}\cos(2\pi U_2), \forall U_1 \in [0,1], U_2 \in [0,1]$$

$$Y_2 = R\sin(\Theta) = \sqrt{-2\ln U_1}\sin(2\pi U_2), \forall U_1 \in [0,1], U_2 \in [0,1]$$

Then Y_1, Y_2 are random variables.

1.6.3   Notes

- A random variable is just a measurable function and **it does not require a probability measure**. See figure 2.1.

- A measurable space consists of sample space and event space $(\Omega, \mathcal{F})$ and is equiped with a probability measure is a probability space.

1.6.4 σ-algebra generated by random variable

Suppose Ω is a sample space. Let $X : \Omega \to \mathbb{R}$ is a random variable and $\mathcal{B}(\mathbb{R})$ is the Borel σ-algebra on $\mathbb{R}$. Indeed $\mathcal{B}(\mathbb{R})$ contains all possible open interval in $\mathbb{R}$. The σ-algebra generated by X, noted $\sigma(X)$ is:

$$\sigma(X) = \left\{ X^{-1}(S) : \forall S \in \mathcal{B}(\mathbb{R}) \right\}$$

or

$$\sigma(X) = \{\{\omega \in \Omega | X(\omega) \in S\} : \forall S \in \mathcal{B}(\mathbb{R})\}$$

since $X^{-1}(S) := \{\omega \in \Omega | X(\omega) \in S\}$
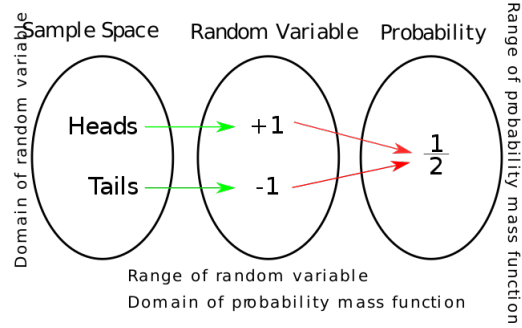
Figure 2.1: Ralation between sample space, random variable and probability measure (mass function).

1.6.4.1 Example 1

Let consider a random variable that always returns a constant, $X(\omega) = c$. Then :

- if $c \in S$ then $X^{-1}(S) = \Omega$

- if $c \notin S$ then $X^{-1}(S) = \varnothing$

Hence $\sigma(X) = \{\Omega, \varnothing\}$.

1.6.4.2 Example 2

Let consider a random variable X that $X(\omega) = -5$ if $\omega \in \{1, 2, 3\}$ and $X(\omega) = 5$ if $\omega \in \{4, 5, 6\}$. $\Omega = \{1, 2, 3, 4, 5, 6\}$. Then :

- If $-5 \notin S$ and $5 \notin S$ then $X^{-1}(S) = \varnothing$

- If $-5 \in S$ and $5 \notin S$ then $X^{-1}(S) = \{1, 2, 3\}$

- If $-5 \notin S$ and $5 \in S$ then $X^{-1}(S) = \{4, 5, 6\}$

- If $-5 \in S$ and $5 \in S$ then $X^{-1}(S) = \Omega$

Hence $\sigma(X) = \{\Omega, \varnothing, \{1, 2, 3\}, \{4, 5, 6\}\}$.

1.7 Measure

In mathematics, the concept of a **measure** can informally be :

- Geometrical measures (length, area, volume)

- Magnitude, mass

- Probability of events

1.7.1 Formal definition of measure

Let (X, Σ) be a **measurable space**. A *set function* μ from Σ to the *extended real number line* ($\mathbb{R}$ with two infinity elements $\pm\infty$) is called a **measure** if the following conditions hold:

1. *Non-negativity*: For all $E \in \Sigma, \mu(E) \geq 0$.

2. *Null empty set*: $\mu(\varnothing) = 0$.

3. *Countable additivity* : For all countable collections $\{E_k\}_{k=1}^{\infty}$ of pairwise **disjoint** sets in Σ, we have

$$\mu\left(\bigcup_{k=1}^{\infty} E_k\right) = \sum_{k=1}^{\infty} \mu(E_k)$$

Note that, the three above conditions do not require the existance of all $\mu(x)$ for $x \in X$, only $\mu(E)$ where $E \in \Sigma$ is required. The role of X here is just to say that Σ is a σ-algebra of X.

1.7.1.1 Signed measure

This is a flexible concept of a measure. μ is called signed measure if it satifies only the *Countable additivity* : For all countable collections $\{E_k\}_{k=1}^{\infty}$ of pairwise **disjoint** sets in Σ, we have

$$\mu\left(\bigcup_{k=1}^{\infty} E_k\right) = \sum_{k=1}^{\infty} \mu(E_k)$$

We call μ finite signed measure if $\mu(E) < \infty, \forall E \in \Sigma$.

1.7.1.2 σ-finite measure

- $\mu(E)$ is called finite measure (w.r.t set E) if $\mu(E) < \infty, \forall E \in \Sigma$.

- A measure μ is called σ-finite measure if sample space X is a countable union of measurable subsets E_i, which means $X = \bigcup_{i=1}^{n} E_i$ and each E_i has finite measure.

1.7.1.3 μ-almost everywhere

Given two functions f and g from X to Y (f and g are not necessary measurable function). We say that $f = g$ μ-almost everywhere if

$$\forall E \in 2^X, \text{ if } \mu(E) = 0 \text{ then } \begin{cases} f(x) = g(x), \forall x \in X/E \\ f(x) \neq g(x), \forall x \in E \end{cases}$$

1.7.2 Outer measure

Definition

Given a set X, let's 2^X denote the collection of all subsets of X, including the empty set $\varnothing$. An *outer measure* on X is a set function μ^* such that :

1. *Non-negativity*: For all $E \in 2^X, \mu^*(E) \geq 0$.

2. *Null empty set*: $\mu^*(\varnothing) = 0$.

3. *Countable subadditivity* : $\forall A, B_1, B_2, \ldots, B_\infty \in 2^X$,
$$\text{if } A \subseteq \bigcup_{i=1}^{\infty} B_i \text{ then } \mu^*(A) \leq \sum_{i=1}^{\infty} \mu^*(B_i)$$

Equivalent definition

1. *Non-negativity*: For all $E \in 2^X, \mu^*(E) \geq 0$.

2. *Null empty set*: $\mu^*(\varnothing) = 0$.

3. *Monotone*: $\forall A, B \in 2^X$ if $A \subseteq B$ then $\mu^*(A) \leq \mu^*(B)$

4. *Outer measure of countable sum is less than countable sum of outer measure*:
$$\mu^* \left(\sum_{i=1}^{\infty} B_i \right) \leq \sum_{i=1}^{\infty} \mu^*(B_i), \quad \forall B_1, B_2, \ldots, B_\infty \in 2^X$$

1.7.2.1 Measurability of sets relative to an outer measure

Let X be a set with an outer measure μ^*. One says that a set $E \in 2^X$ is μ^*-measurable (or Carathéodory-measurable relative to μ^*) if and only if it satifies the Carathéodory's criterion :

$$\begin{aligned} \mu^*(A) =& \mu^*(A \cap E) + \mu^*(A \setminus E) \\ =& \mu^*(A \cap E) + \mu^*(A \cap E^C), \quad \forall A \in 2^X \end{aligned}$$

Informally, this says that a μ^*-measurable set E is one which may be used as a building block (or basic block) such that any other set A can be decomposed into two parts, one is inside of the measurable set and the other is outside of the measurable set.

Properties

1. *Relative complement*: If set $E \in 2^X$ is a μ^*-measurable, then E^C (in X) is also μ^*-measurable.

2. *Countable additivity*: If $E_1, E_2, \ldots, E_\infty \in 2^X$ are μ^*-measurable pairwise **disjoint**, then
$$\mu^* \left(\bigcup_{k=1}^{\infty} E_k \right) = \sum_{k=1}^{\infty} \mu^*(E_k)$$

3. *Closed under countable unions*: If $E_1, E_2, \ldots, E_\infty \in 2^X$ are μ^*-measurable then $\bigcup\limits_{i=1}^{\infty} E_i$ is also μ^*-measurable.

4. *Completeness*: If $E \subseteq X$ such that $\mu^*(E) = 0$ then E is μ^*-measurable.

Because of complexity, we show only the proof for the last property :

- From the third characteristic (*monotone*) in the equivalent of outer measure, $\forall B \in 2^X$, if $B \subseteq E$ then $\mu^*(B) \le \mu^*(E) = 0$. However, from the first characteristic, $\mu^*(B) \ge 0$, then $\mu^*(B) = 0$.

- $\forall A \in 2^X$, then $A = (A \cap E) \cup (A \cap E^C)$.

- On one hand, from the third characteristic (*countable subadditivity*) in the definition of outer measure, we have :

$$\mu^*(A) \le \mu^*(A \cap E) + \mu^*(A \cap E^C) \le \mu^*(A \cap E^C),$$

since $A \cap E \subseteq E$, then $\mu^*(A \cap E) = 0$.

- On the one hand, $A \cap E^C \subseteq A$, then from the *monotone* characteristic, we have $\mu^*(A \cap E^C) \le \mu^*(A)$.

- Finally, $\mu^*(A) = \mu^*(A \cap E^C)$ or $\mu^*(A) = \mu^*(A \cap E) + \mu^*(A \cap E^C)$. we conclude that E is μ^*-measurable.

1.7.2.2 Purpose of outer measure

Given a set X, the purpose of constructing an outer measure on all subsets of X is to **pick out** a class (or set) of subsets, which is σ-algebra and the outer measure on this class is a measure.

Proof

The class that we mentioned above, denoted Σ, is thus the set of all μ^*-measurable subsets of X. We need to show that :

1. Σ is a σ-algebra. This is because :

   - $\varnothing \in \Sigma$: since from the definition of outer space $\mu^*(\varnothing) = 0$ and the *completeness*, property 4 says that $\varnothing \in \Sigma$.
   - *Relative complement*, property 1.
   - *Closed under countable unions*, property 3.

2. Outer measure μ^* on Σ is a measure. This is because:

   - *Non-negativity* and *null empty set* are herited from outer measure.
   - *Countable additivity* is the property 2.

Note that, a measure is outer measure but the converse is not true.

1.7.2.3 Valuation

Given a class $\mathcal{T}$ of open subsets of a topological space (which can be a set), a valuation is a set function such that :

1. *Non-negativity*: For all $E \in \mathcal{T}, v(E) \geq 0$.

2. *Null empty set*: $v(\varnothing) = 0$.

3. *Monotone*: $\forall A, B \in \mathcal{T}$ if $A \subseteq B$ then $v(A) \leq v(B)$.

4. *Modularity*:

$$v(A \cup B) + v(A \cap B) = v(A) + v(B), \quad \forall A, B \in \mathcal{T}$$

At a first look, a valuation is similar to a outer space, the only thing that we can distinguish between them is the last condition.

The purpose of valuation is a raw or basic set function to construct a outer measure, in the following text.

1.7.2.4 Construction of outer measure

We discover a manner to construct an outer space from a given set X and a given valuation v.

$$\mu^*(E) = \inf\left\{\sum_{i=0}^{\infty} p(A_i) \,\middle|\, E \subseteq \bigcup_{i=0}^{\infty} A_i, \forall i \in \mathbb{N}, A_i \in 2^X\right\}$$

We have a convention that the infimum is infinite if no such sequence of A_i exists to cover E.

1.7.3 Examples of measure

1.7.3.1 Counting measure

Given measurable space (X, Σ). The counting measure on Σ is defined by $\mu(E)$ = number of elements in E, for all $E \in \Sigma$. The three conditions *non-negetivity*, *null empty set* and *countable additivity* are trivially proved

1.7.3.2 Probability measure

P on measurable space $(\Omega, \mathcal{F})$ is measure since :

- *Non-negetivity*: $P(E) \geq 0, \forall E \in \mathcal{F}$.

- *Null empty set*: $P(\varnothing) = 0$, in addition $P(\Omega) = 1$.

- *Countable additivity*: $P\left(\bigcup_{k=1}^{\infty} E_k\right) = \sum_{k=1}^{\infty} P(E_k)$ where $\{E_k\}_{k=1,\ldots,\infty}$ are pairwise disjoint.

1.7.3.3 Lebesgue measure

The Lebesgue measure is the standard way of assigning a measure to subsets of Euclidean n-dimensional space. It coincides with the standard measure of length, area, or volume for $n = 1, 2, 3$. For $n > 3$, it is also called n-dimensional volume, n-volume, hypervolume, or simply volume. Let's start with the case $X = \mathbb{R}$ then we generalize for the case $X = \mathbb{R}^n$

Definition for $X = \mathbb{R}$

Given $X = \mathbb{R}$, we denote $2^{\mathbb{R}}$ a set that contains all open intervals in $\mathbb{R}$. Also, given a valuation (sec 1.7.2.3) l on $2^{\mathbb{R}}$ such that $l(I) = b - a$ with $I = (a, b)$. Then we construct an outer measure $\lambda^*(E)$, called **Lebsgue outer measure** as in sec 1.7.2.4, by :

$$\lambda^*(E) = \inf\left\{\sum_{i=0}^{\infty} l(A_i) \,\bigg|\, E \subseteq \bigcup_{i=0}^{\infty} A_i, \forall i \in \mathbb{N}, A_i \in 2^{\mathbb{R}}\right\}$$

Definition for $X = \mathbb{R}^n$

Given $X = \mathbb{R}^n$, we denote $2^{\mathbb{R}^n}$ a set that contains all open rectangular cuboids in $\mathbb{R}^n$, $C = I_1 \times I_2 \times \ldots \times I_n$. The valuation v on $2^{\mathbb{R}^n}$ is given $v(C) = l(I_1) \times l(I_2) \times \ldots \times l(I_n)$. Then Lebsgue outer measure is constructed by :

$$\lambda^*(E) = \inf\left\{\sum_{i=0}^{\infty} v(A_i) \,\bigg|\, E \subseteq \bigcup_{i=0}^{\infty} A_i, \forall i \in \mathbb{N}, A_i \in 2^{\mathbb{R}^n}\right\}$$

Finally, we just select all sets $E \in 2^{\mathbb{R}^n}$ such that each E is λ^*-measurable as in sec 1.7.2.1 and creat a σ-algebra Σ on these E. In Σ, the Lebsgue outer measure λ^* is **Lebesgue measure** λ.

To take away :

- A set E is said *Lebesgue measurable* if it is λ^*-measurable, where λ^* is Lebsgue outer measure.

- Non-measurable set $E \in 2^{\mathbb{R}^n}$ do exist, an example is the Vitali set (1.7.3.4).

- Trivially, we see that the outer Lebesgue measure λ^* of a countable set is 0. From properties of *completeness*, we say that the countable set is λ^* measurable or Lebesgue measure λ of a countable set is 0. For example $\mathbb{N}$ set of natural numbers and $\mathbb{Q}$ set of rational numbers are both countable, then $\lambda(\mathbb{N}) = 0$ and $\lambda(\mathbb{Q}) = 0$

1.7.3.4 Vitali set

This is an example of non-measurable set.

First, we define an *equivalence* on two real numbers $x, y \in \mathbb{R}$, that $x \sim y \Leftrightarrow x - y \in \mathbb{Q}$. For example, $0.5 + \pi \sim \pi$ and $\sqrt{3} \not\sim \pi$,...

Second, we put all real numbers that are equivalent (by above definition) to a group, we may have an infinit number of groups, no problem. Then by the axiom of choice, we can define a representative set A such that every real

number is equivalent to **exactly one element** of A. For example, in a group that contains $\sqrt{2}, \sqrt{2}+1, \sqrt{2}+2, \ldots$, we just take $\sqrt{2}$ and put it to A.

Third, instead of performing the second step in $\mathbb{R}$, we only do it in $[0,1]$ and this set A is called *Vitali set*.

Fourth, given a $q \in \mathbb{Q} \cap [0,1)$, we define a translation :

$$A_q = \{a + q \ (\text{mod } 1) \mid a \in A\}$$

Then we can see that

- A_q are pairwise disjoint, $(A = A_0)$.

- $\bigcup A_q, \forall q \in \mathbb{Q} \cap [0,1)$, is a *countable* union and it is equal to $[0,1]$.

Fifth, given a Lebesgue measure μ (sec 1.7.3.3) and suppose that A is measureable set, then the *countable additivity* property of a measure say that:

$$\mu(\bigcup A_q) = \sum \mu(A_q) = \mu([0,1])$$

Finally, by

- $\mu(A_q)$ is a constant since the Lebesgue measure is translation invariant

- There is infinit element $\mu(A_q)$ in the sum.

- $\mu([0,1])$ is also a fixed value that is strictly positive.

we have the contradiction.

1.8 Measure space

A measure space is generalized concept of probability space. A measure space is a triple $(\Omega, \mathcal{F}, \mu)$ where :

- Ω is a set

- $\mathcal{F}$ is is a σ-algebra on the set Ω

- μ is a measure on $(\Omega, \mathcal{F})$.

In other words, a measure space consists of a measurable space $(X, \mathcal{F})$ (see 1.4) together with a measure (see 1.7) on it.

If the measurable space consists of sample space and event space; and the measure μ is the probability measure P (probability function), then the measure space is probability space.

1.8.1 Probability space

If the measurable space consists of sample space and event space $(\Omega, \mathcal{F})$; and the measure μ is the probability measure P (probability function), then the measure space is probability space. More precise :

- A sample space Ω , which is the set of all possible **outcome**s.

- An event space $\mathcal{F}$, which is a set of events. An **event** is an subset of sample space. $\mathcal{F}$ is a σ-algebra of Ω.

- A probability measure P, which assigns each event in the event space a probability, which is a number between 0 and 1.

1.8.1.1 Example 1

The throw of a standard die :

- Ω to be $\{1, 2, 3, 4, 5, 6\}$.

- For the event space $\mathcal{F}$, we could simply use the set of all subsets of the sample space, but $\mathcal{F}$ must be a σ-algebra of Ω. E.g., we want that $\mathcal{F}$ contains events that die lands on $\{5\}$ and die lands on even numbers $\{2, 4, 6\}$. Then naively, $\mathcal{F} = \{\{5\}, \{2, 4, 6\}\}$. However, the latter is not a σ-algebra of Ω. What we need to do now is to make $\mathcal{F}$ a smallest σ-algebra of Ω that contains $\{5\}$ and $\{2, 4, 6\}$, by taking $\mathcal{F} = \{\Omega, \{5\}, \{2, 4, 6\}, \{1, 2, 3, 4, 6\}, \{1, 3, 5\}, \{2, 4, 5, 6\}, \{1, 3\}, \varnothing\}$.

- The probability function, that map each event in $\mathcal{F}$ to a probability, e.g event $\{5\}$ corresponds to $1/6$ and event $\{2, 4, 6\}$ corresponds to $1/2$, ...

1.8.1.2 Example 2

The fair coin is tossed three times.

- There are 8 possible outcomes: $\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$.

- We denote 2^Ω all possible subsets of Ω. Any subset $A \in 2^\Omega$ can make an event space by using $\mathcal{F} = \sigma(A)$, where $\sigma(A)$ denotes the smallest σ-algebra of Ω that contains A. Note that, two subsets can have the same $\mathcal{F}$ such as A and A^C, since $\sigma(A) = \sigma(A^C)$. The number of subsets is $2^{|\Omega|}$. In this case, we have $2^8 = 256$ subsets (or events). Let's take $\mathcal{F} = \{\varnothing, \{HHH, HHT, HTH, THH\}, \{HTT, THT, TTH, TTT\}, \Omega\}$, the event space with 2 events, the number of H is greater then one of T and vice-versa.

- The probability function : e.g., $P(\{HHH, HHT, HTH, THH\}) = 1/2$ and $P(\{HTT, THT, TTH, TTT\}) = 1/2$

1.8.1.3   Example 3

As the same of above example, the fair coin is tossed three times but we know only the total number of tails (we do not know the result of three time tossing).

- The sample space $\Omega = \{0, 1, 2, 3\}$.

- The event space that number of tails is unpair : $\mathcal{F} = \{\varnothing, \{1, 3\}, \{0, 2\}, \Omega\}$.

- The probability function : e.g. $P(\{1, 3\}) = P(\{HHT, HTH, THH, TTT\}) = 1/2$.

1.8.1.4   Example 4

A number between 0 and 1 is chosen at random (**uniformly**).

- $\Omega = [0, 1]$

- E.g. the event that $x < 0.1$ or $x > 0.9$. $\mathcal{F} = \{\varnothing, \Omega, (0.9, 1], [0, 0.1), [0, 0.9], [0.1, 1]\}$.

- The probability function : $P((a, b)) = (b - a)$, which generates the *Lebesgue measure* on [0,1].

1.9   Lebesgue integral

We first reminding the concept of Riemann integral

1.9.1   Riemann integral

Let's $f : [a, b] \to \mathbb{R}$. Then f is said *Riemann integrable* over $[a, b]$ if

$$\lim_{\|P\| \to 0} \sum_{i=1}^{\infty} (x_i - x_{i-1}) f(t_i) \text{ converges,}$$

where P is a partition over $[a, b]$, $\|P\| = \max_i (x_i - x_{i-1})$, $t_i \in [x_i, x_{i-1}]$. In this case, the Riemann integral is equal to the limit and denoted $\int_a^b f(x) dx$.

Properties

- If f is Riemann integrable, then it is bounded.

- If f is bounded and piecewise continuous, then it is Riemann integrable

1.9.2   Intuitive idea for Riemann integral and Lebesgue integral

We discover two main ideas for the intuitive of Lebesgue integral, one from sum of discontinuable horizontal rectangles and one from discontinuable vertical rectangles.
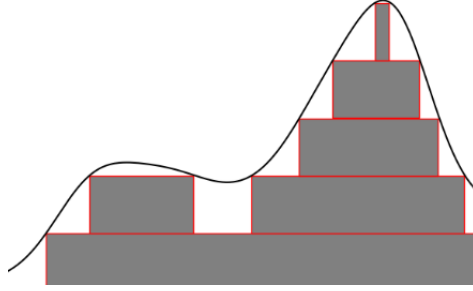
Figure 2.2: Lebesgue integral illustrated by summing up of discontinuable horizontal rectangles.

1.9.2.1 Sum of discontinuable horizontal rectangles

We see the idea in figure 2.2. A little bit formally, let (X, Σ, μ) be a measure space and f is a non-negativity measurable function from (X, Σ) to $(\mathbb{R}, B(\mathbb{R}))$, then the Lebesgue integral can be written :

$$\lim_{\|P\| \to 0} \sum_{k=1}^{\infty} (y_k - y_{k-1}) \mu(\{x \mid f(x) > y_k\})$$

where P is a partition on the ordinate over interval $[0, \max_x(f(x))]$.

1.9.2.2 Sum of discontinuable vertical rectangles
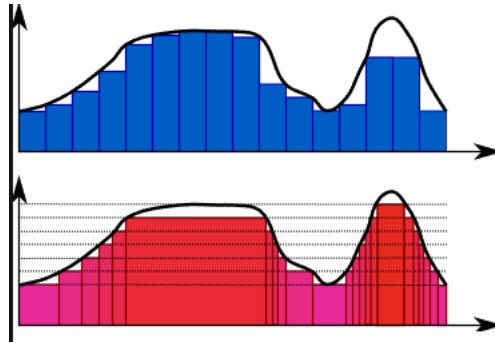
Figure 2.3: Intuitive idea between Riemann and Lebesgue integral by discontinuable vertical rectangles.

In figure 2.3, we see the difference between calculation of integral, proposed by Riemann and Lebesgue. Visually,

- In Riemann integral, we sum progessively small vertical rectangles.

- In Lebesgue integral, we group first small vertical rectangles that have the same height and let's call this group "*discontinuable vertical rectangle*". Then we sum these discontinuable vertical rectangles. In figure 2.3, the color changes from rose to red for different discontinuable vertical rectangles, characterized by the height or output value of function.

Formally, the Lebesgue integral can be written:

$$\lim_{\|P\| \to 0} \sum_{k=1}^{\infty} \xi_i \mu(\{x \mid y_{k-1} \le f(x) < y_k\})$$

where $\xi_i \in [y_{k-1}, y_k]$ and P is a partition on the ordinate over interval $[0, \max_x(f(x))]$.

This concept, which is considered as a generalization of Riemann integral, have a formal definition by using simple function in sec 1.9.3 that we discover right after.

1.9.2.3 Example

Riemann integral is invented at the mid-1800s and Lebesgue is invented at 1902 and it has several advantages compared to Riemann integral. The classical one is the following example. Let $\mathbb{1}_{\mathbb{Q}}(x)$ be indicator function, which is equal to 1 if $x \in \mathbb{Q}$, 0 otherwise. Then it is show that over interval $[0, 1]$, $\mathbb{1}_{\mathbb{Q}}(x)$ is not Riemann integrable, but we still calculate integral of this function by using Lebesgue integral. The idea is simple, we can write the integral by

$$1 \times l_{[0,1]}(\mathbb{Q}) + 0 \times l_{[0,1]}(\mathbb{R}/\mathbb{Q})$$

where $l_{[0,1]}$ is Lebesgue measure on measurable space $([0,1], B([0,1]))$ or l means the length of interval. We see that in sec 1.7.3.3, Lebesgue measure of countable set $\mathbb{Q}$ is 0 or $l_{[0,1]}(\mathbb{Q}) = 0$. Then by the countable additivity, $l_{[0,1]}(\mathbb{R}/\mathbb{Q}) = 1$.

1.9.3 Simple function

Given measure space (X, Σ, μ) and non-negativity measurable function f is a non-negativity measurable function from (X, Σ) to $(\mathbb{R}, B(\mathbb{R}))$.

An **indicator function** $\mathbb{1}_A(x)$, is defined equal to 1 if $x \in A$ and 0 otherwise, where $A \in \Sigma$.

Then a **simple function** s is a finite linear combination of indicator functions of μ-measurable sets A_k :

$$s(x) = \sum_{k=1}^{\infty} a_k \mathbb{1}_{A_k}(x),$$

where sets A_k are pairwise disjoint.

Given a simple function $s(x)$, among of its all possible decompositions (into indicator functions), it exists only one called canonical decomposition such that

- a_k are pairwise disjoint and non null.

- $s(x) = 0, \forall x$ if and only if $n = 0$.

1.9.4 Definition of Lebesgue integral

Given a measure space (X, Σ, μ) and a non-negative measurable function f from (X, Σ) to $(\mathbb{R}, B(\mathbb{R}))$, we define

$$\int_X f \, d\mu = \sup_s \left\{ \int_X s \, d\mu \colon 0 \le s(x) \le f(x) \text{ for all } x \in X, \ s \text{ is simple function} \right\}$$

It is proved that Lebesgue integral give always the same result by Riemann integral.

1.9.4.1 Integral of simple function

To calculate $\int_X s \, d\mu$, we use the **canonical decomposition** of s :

$$\int_X s \, d\mu = \int_X \sum_{k=1}^{\infty} a_k \mathbb{1}_{A_k} \, d\mu$$

$$= \sum_{k=1}^{\infty} a_k \int_X \mathbb{1}_{A_k} \, d\mu$$

To **assign a value** to the integral of the indicator function $\mathbb{1}_A$ of a measurable set A consistent with the given measure μ, the only reasonable choice is to **set**:

$$\int_X \mathbb{1}_A \, d\mu = \mu(A)$$

Finally,

$$\int_X s \, d\mu = \sum_{k=1}^{\infty} a_k \mu(A_k)$$

1.9.4.2 Emphasizing

Note that, to emphasize that f and μ are dependent on $x \in X$, we can write

$$\int_X f \, d\mu = \int_X f(x) \, \mu(dx) = \int_X f(x) \, d\mu(x)$$

In short $d\mu, \mu(dx)$ or $d\mu(x)$ means the measure of infinitesimal set around x.

1.9.4.3 In case of random variable

This is a particular case, where the measure space is probability space, the measurable function f is a random variable X and the measure μ is probability measure P.

$$\int f \, \mathrm{d}\mu = \int X(\omega) \, \mathrm{d}P(\omega)$$

$$= \int X(\omega)p(\omega) \, \mathrm{d}\omega$$

$$= \mathbb{E}[X]$$

where p is density function. **Do not confuse** the measure P with density function p. For example, if $X \sim \mathcal{N}(0,1)$, $P(a \leq X \leq b) = \int_a^b p(\omega)d\omega$.

1.9.5 When Lebesgue integral is defined and integrable

To handle any measurable function f that take value in $\mathbb{R}$, we use the concept of positive par and negative part :

$$f = f^+ - f^-$$

where f^+ and f^- are positive part and negative part of f, they are always non-negativity. Moreover, since f is measurable function then both f^+ and f^- are measurable functions and we can apply Lebesgue integral on them. Then

$$\int f \mathrm{d}\mu = \int f^+ \mathrm{d}\mu - \int f^- \mathrm{d}\mu$$

We say that the $\int f \mathrm{d}\mu$ is defined if and only if

$$\min(\int f^+ \mathrm{d}\mu, \int f^- \mathrm{d}\mu) < \infty$$

Note that if $\int f \mathrm{d}\mu = \pm\infty$, it is also definded. The state of $\int f \mathrm{d}\mu$ is given by :

	$\int f^+ \mathrm{d}\mu < \infty$	$\int f^+ \mathrm{d}\mu = \infty$
$\int f^- \mathrm{d}\mu < \infty$	finite	$+\infty$
$\int f^- \mathrm{d}\mu = \infty$	$-\infty$	**not defined**

The measurable function f is said **Lebesgue integrable** if and only if

$$\int |f| \mathrm{d}\mu < +\infty$$

1.9.5.1 In case of random variable

We have seen that in sec 1.9.4.3, $\int f \, \mathrm{d}\mu$ is $\mathbb{E}[X]$, where $X = f$.
We say that the $\mathbb{E}[X]$ is defined if and only if

$$\min(\mathbb{E}[X^+], \mathbb{E}[X^-]) < \infty$$

and X is said integrable if and only if

- In continuous case :

$$\mathbb{E}[|X|] = \int |x|p(x)\mathrm{d}x < +\infty$$

where p here is density function.

- In discret case :
$$\mathbb{E}[|X|] = \sum |x|P(X = x) < +\infty$$

1.9.6 Properties of Lebesgue integral

As Riemann integral, Lebesgue integral has the same properties for *linearity* and *monotonicity*

- If f and g are Lebesgue integrable functions, then $af + bg$ is Lebesgue integrable and

$$\int (af + bg)\, d\mu = a \int f\, d\mu + b \int g\, d\mu$$

- If $f \leq g$, then
$$\int f\, d\mu \leq \int g\, d\mu$$

1.10 Independence

Some time called independence σ-algebra. Let $(\Omega, \mathcal{F}, \mu)$ be a measure space, which means there exists $\mu(E), \forall E \in \mathcal{F}$.

Two events $A, B \in \mathcal{F}$ are said independent, noted $A \perp B$, if

$$\mu(A \cap B) = \mu(A)\mu(B)$$

In general, the events $A_1, \ldots, A_n \in \mathcal{F}$ are independent, if :

$$\mu(A_1 \cap \ldots \cap A_n) = \mu(A_1) \ldots \mu(A_n)$$

1.10.1 Illustration example

1.10.1.1 Non independence case

Supppose that we have one dice with probability space $(\omega, \mathcal{F}, P)$ and two random variable A and B :

ω	1	2	3	4	5	6
$A(\omega)$	0	-1	0	-1	0	-1
$B(\omega)$	0	-1	-1	0	0	-1

The sample space $\Omega = \{1, 2, 3, 4, 5, 6\}$ and let's take $\mathcal{F} = \sigma(\Omega)$.

- $\{\omega|A(\omega) = -1\} = \{2,4,6\}$ then $P(A = -1) = \frac{1}{2}$

- $\{\omega|B(\omega) = -1\} = \{2,3,6\}$ then $P(B = -1) = \frac{1}{2}$

- $\{\omega|A(\omega) = -1 \cap B(\omega) = -1\} = \{2,6\}$ then $P(A = -1 \cap B = -1) = \frac{1}{3}$

Then $P(A = -1)P(B = -1) = \frac{1}{4} \neq P(A = -1 \cap B = -1) = \frac{1}{3}$

1.10.1.2 Independence case

Now suppose that we have two dices, one for A and the other for B. In order to distinguish better between two dices, we label six faces of one dice by a, b, c, d, e, f. Let's define two random variables A and B by :

ω_1	1	2	3	4	5	6
$A(\omega_1)$	0	-1	0	-1	0	-1
ω_2	a	b	c	d	e	f
$B(\omega_2)$	0	-1	-1	0	0	-1

The sample space $\Omega = \{1,2,3,4,5,6\} \times \{a,b,c,d,e,f\} = \{(1,a),\ldots,(6,f)\}$ and let's take $\mathcal{F} = \sigma(\Omega)$. In this case each outcome, .e.g., $(2,C)$ has probability $\frac{1}{36}$.

- $\{\omega|A(\omega) = -1\} = \{(2,.),(4,.),(6,.)\}$ then $P(A = -1) = \frac{1}{2}$

- $\{\omega|B(\omega) = -1\} = \{(.,b),(.,c),(.,f)\}$ then $P(B = -1) = \frac{1}{2}$

- $\{\omega|A(\omega) = -1 \cap B(\omega) = -1\} = \{2,4,6\} \times \{b,c,f\}$ then $P(A = -1 \cap B = -1) = \frac{9}{36} = \frac{1}{4}$

Then $P(A = -1)P(B = -1) = \frac{1}{4} = P(A = -1 \cap B = -1) = \frac{1}{4}$. Then we can show that $P(A)P(B) = P(AB)$.

1.10.2 Between subsets

Subsets $\mathcal{A}_1,\ldots,\mathcal{A}_n$ of $\mathcal{F}$ are independent if :

$$A_1, A_2, \ldots, A_n \text{ are independent}, \forall A_1 \in \mathcal{A}_1, A_2 \in \mathcal{A}_2, \ldots, A_n \in \mathcal{A}_n$$

1.10.3 Independence of $\pi-$systems implies independence of σ-algebras

If $\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_n$ are $\pi-$system and independent, then $\sigma(\mathcal{A}_1), \sigma(\mathcal{A}_2), \ldots, \sigma(\mathcal{A}_n)$ are independent. (See here for demonstration http://theanalysisofdata.com/probability/E_4.html)

1.10.4   Conditional independence

Let A, B, C be events. A and B are said to be conditionally independent given C if and only if $P(C) > 0$ and :

$$P(A \mid B, C) = P(A \mid C)$$

This property is often written: $A \perp\!\!\!\perp B \mid C$ Equivalently, conditional independence may be stated as:

$$P(A, B|C) = P(A|C)P(B|C)$$

It demonstrates that $(A \perp\!\!\!\perp B \mid C)$ is equivalent to $(B \perp\!\!\!\perp A \mid C)$.

1.10.5   Uncorrelation and independence

Uncorrelation does not necessarily mean independence :

- Uncorrelated variables have a correlation coefficient of zero, which means there is **no linear relationship** between the variables. However, there may still have **a nonlinear relationship** between the variables, or they may be related in some other ways that is not captured by the correlation coefficient.

- Independent variables, on the other hand, are variables that have no relationship with each other, whether linear or nonlinear. So, all independent variables are uncorrelated, but not all uncorrelated variables are independent.

In summary, if two variables are independent, they are uncorrelated, but the inverse is not necessarily true.

1.10.6   Note 1

Let's consider random variable X, Y, Z. If X is independent to Y and X is independent to Z, it **do not** imply that X is independent to $aY + bZ$.

Example Let Y and Z are two Bernoulli(0.5) variables and the variable X such that :

Y	0	1	0	1
Z	0	0	1	1
X	0	1	1	0

It means that if $Y = 0$ and $Z = 0$ then $X = 0$ ... We see that :

- $P(X = 0|Y = 1) = P(X = 0|Y = 0) = P(X = 0) = 0.5$ (same as $X = 1$), then $P(X|Y) = P(X)$, then $X \perp Y$.

- $P(X = 0|Z = 1) = P(X = 0|Z = 0) = P(X = 0) = 0.5$ (same as $X = 1$), then $P(X|Z) = P(X)$, then $X \perp Z$.

But in case $X = 0$:

- $P(Y + Z = 0 | X = 0) = 0.5 \neq P(Y + Z = 0) = 0.25$

- $P(Y + Z = 2 | X = 0) = 0.5 \neq P(Y + Z = 2) = 0.25$

- $P(Y + Z = 1 | X = 0) = 0 \neq P(Y + Z = 1) = 0.5$

and in case $X = 1$

- $P(Y + Z = 0 | X = 1) = 0 \neq P(Y + Z = 0) = 0.25$

- $P(Y + Z = 2 | X = 1) = 0 \neq P(Y + Z = 2) = 0.25$

- $P(Y + Z = 1 | X = 1) = 1 \neq P(Y + Z = 1) = 0.5$

Then X and $Y + Z$ are not independent.

1.10.7 Note 2

However, if X is independent to (Y, Z), it **implies** that X is independent to $aY + bZ$.

Proof : Using the property that if $X \perp (Y, Z)$ then $f_1(X) \perp f_2(Y, Z), \forall f_1, f_2$. Then taking $f_1(X) = X$ and $f_2(Y, Z) = Y + Z$.

1.10.8 Note 3

If X is independent to Y and Z, it **does not** imply that X is independent to YZ.

Example

Let's say we choose a number randomly from: $\{1, 2, 3, 4, 6, 7, 8, 9\}$. Let X be:

- $X = 1$ if (Chosen number is even AND less than five) OR (Chosen number is odd AND greater than 5)

- $X = 0$ otherwise

Let Y be:

- $Y = 1$ if chosen number is even

- $Y = 0$ otherwise

Let Z be:

- $Z = 1$ if your chosen number is greater than 5

- 0 otherwise

Now we know :

- $X = 1$ with probability 0.5 (2,4,7,9)

- $X = 0$ with probability 0.5 (1,3,6,8)

If we know :

- $Y = 1$, then X is still 1 with p = 0.5 (2,4) and 0 with p = 0.5 (6,8)

- $Y = 0$, then X is still 1 with p = 0.5 (7,9) and 0 with p = 0.5 (1,3)

If we know:

- $Z = 1$, then X is still 1 with p = 0.5 (7,9) and 0 with p = 0.5 (6,8)

- $Z = 0$, then X is still 1 with p = 0.5 (2,4) and 0 with p = 0.5 (1,3)

So X is independent of Y and X is independent of Z.

But knowing if a number is even AND knowing if it's greater than 5 ($Y\&Z$), makes us know X with certainty. e.g. $Y = 1, Z = 1$, then $YZ = 1$, X has to be 0 with probability 1 (as X is 0 if the number is an even number > 5)

In short :

- $P(X = 0|Y) = P(X = 0) = 0.5$ (same for $X = 1$) then $X \perp Y$

- $P(X = 0|Z) = P(X = 0) = 0.5$ (same for $X = 1$) then $X \perp Z$

But $P(X = 0|YZ = 1) = 1 \neq P(X = 0) = 0.5$ (similar for $X = 1$).

2 Useful tools

2.1 Notations

Usually used, but not conventionning :

- Capital X is for random variable.

- Lower case x is for a particular value of X.

- Capital P represents the probability of an event occurring. For example, if we toss a fair coin, the probability of getting heads is denoted by $P(X = heads) = 0.5$, and the probability of getting tails is denoted by $P(X = tails) = 0.5$.

- Lowercase p represents a probability density function (PDF) or probability mass function (PMF), depending on whether the variable is continuous or discrete. A PDF gives the probability of a continuous random variable taking on a particular value or falling within a certain range of values, while a PMF gives the probability of a discrete random variable taking on a particular value.

Example of using : In discrete case, for a given dice, we can write :

$$P(X = 1) = p(1) = \frac{1}{6}$$

where p is PMF.

But in continuous case, given $X \sim \mathcal{N}(0,1)$,

$$P(X = x) \approx P(x - \frac{dx}{2} \leq X \leq x + \frac{dx}{2}) = p(x)dx$$

where p is PDF.

2.2   Law of the unconscious statistician

LOTUS, This is a name for a method that one (may) uses to calculate instinctly the expected value but do not know its name. Let X that one knows its probability distribution and a function $g(X)$ that one does not know its distribution.

In discret case, let P_X is the probability mass associated to X:

$$\mathbb{E}[g(X)] = \sum_x g(x)P_X(X = x)$$

Let $f_X(x)$ be the density function associated to X, then in continuous case we have:

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x)f_X(x)dx$$

If we know the cumulative probability distribution function of X, means F_X:

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x)dF_X(x)$$

2.3   Expectation with exponentiation

Let W and Z be non-negative random variables on probability space $(\Omega, \mathcal{F}, P)$. Then for any $r > 0$,

$$\mathbb{E}[WZ^r] = r \int_0^{+\infty} t^{r-1}\mathbb{E}[W\mathbb{1}_{\{Z>t\}}]\,dt$$

Proof

$$r \int_0^{+\infty} t^{r-1} \mathbb{E}[W \mathbb{1}_{\{Z>t\}}] dt$$

$$= r \int_0^{+\infty} t^{r-1} \int_{Z>t} W \, dP \, dt$$

$$= \int_0^{+\infty} \int_{Z>t} rt^{r-1} W \, dP \, dt$$

$$= \int_\Omega \int_0^Z rt^{r-1} W \, dt \, dP \quad \text{(bound changing)}$$

$$= \int_\Omega W \int_0^Z rt^{r-1} \, dt \, dP$$

$$= \int_\Omega W Z^r \, dP$$

$$= \mathbb{E}[W Z^r]$$

2.4 Expected value of nonnegetive random variable

Given X is a nonnegetive random variable, then

$$\mathbb{E}[X] = \int_0^{+\infty} P(X \ge x) \, dx$$

Proof

Let $f_X(x)$ be the density function of X, then

$$\int_0^{+\infty} P(X \ge x) \, dx = \int_0^{+\infty} \int_x^{+\infty} f_X(z) \, dz \, dx$$

$$= \int_0^{+\infty} \int_0^z f_X(z) \, dx \, dz$$

$$= \int_0^{+\infty} f_X(z) \int_0^z dx \, dz$$

$$= \int_0^{+\infty} f_X(z) z \, dz$$

$$= \mathbb{E}[X]$$

The second equality is that given $0 \le x \le z \le +\infty$, if we fix x first and if we fix z first, we must have the same thing. The last equality is because X is nonnegative.

An other proof which is faster by using 2.3 with $W = 1, r = 1$ and note that $\mathbb{E}[\mathbb{1}_{\{Z>t\}}] = P(Z > t)$.

2.4.1 Related inequality

Given X is a nonnegetive random variable, then

$$\mathbb{E}[X] = \int_0^{+\infty} P(X \geq x)\,\mathrm{d}x \qquad = \sum_{i=0}^{+\infty} \int_i^{i+1} P(X \geq x)\,\mathrm{d}x$$

$$\geq \sum_{i=0}^{+\infty} \int_i^{i+1} P(X \geq i+1)\,\mathrm{d}x \qquad = \sum_{i=0}^{+\infty} P(X \geq i+1) \int_i^{i+1} \mathrm{d}x$$

$$= \sum_{i=0}^{+\infty} P(X \geq i+1) \qquad = \sum_{i=1}^{+\infty} P(X \geq i)$$

The greater or equal because if x in interval $[i, i+1]$, $P(X \geq x) \geq P(X \geq i+1)$.

2.4.2 More general case

Given X is a nonnegetive random variable and $a \geq 0$, then

$$\mathbb{E}[(X-a)^+] = \int_a^{+\infty} P(X \geq x)\,\mathrm{d}x$$

Proof
By using 2.3 with $W = \mathbb{1}_{\{X \geq a\}}, Z = X - a, r = 1$:

$$\mathbb{E}[(X-a)^+] = \mathbb{E}[(X-a)\mathbb{1}_{\{X \geq a\}}]$$

$$\leq \int_0^{+\infty} \mathbb{E}[\mathbb{1}_{\{X \geq a\}}\mathbb{1}_{\{X-a \geq x\}}]dx$$

$$= \int_0^{+\infty} \mathbb{E}[\mathbb{1}_{\{X-a \geq x\}}]dx$$

$$= \int_0^{+\infty} \mathbb{E}[\mathbb{1}_{\{X \geq a+x\}}]dx$$

$$= \int_a^{+\infty} \mathbb{E}[\mathbb{1}_{\{X \geq x\}}]dx$$

where the last equality is by variable change from x to $x + a$.

2.5 Equality by expectation

Given X and Y two random variables in probability space $(\Omega, \mathcal{F}, P)$, if

$$\mathbb{E}[X.\mathbb{1}_A] = E[Y.\mathbb{1}_A], \forall A \in \mathcal{F}$$

Then $X = Y$ almost surely, which means $P(A = B) = 1$

Proof

Let $B = \omega\colon X > Y$, then $B \in \mathcal{F}$ and by the hypothese, we must have $\mathbb{E}[(X - Y).\mathbb{1}_B] = 0$. Note that $X - Y$ is a strictly positive, then $P(B) = 0$. By analogy with $C = \omega\colon X < Y$, we must have $P(C) = 0$. Finally $P(X = Y) = 1$.

2.6 Layer cake representation

Given a measure space $(\Omega, \mathcal{F}, \mu)$ and function $f : \Omega \to \mathbb{R}^+$ which is measurable (1.5), from $(\Omega, \mathcal{F})$ to $(\mathbb{R}^+, B(\mathbb{R}^+))$. Then we have :

$$f(x) = \int_0^\infty \mathbb{1}_{L(f,t)}(x)\,\mathrm{d}t \quad , \forall x \in \Omega$$

where $L(f, t)$ denotes the super-level set

$$L(f, t) = \{\omega \in \Omega \mid f(\omega) \geq t\}$$

We call layer cake representation because when t increases, we have a new level layer.

Before proof, we consider the following lemma

2.6.1 Lemma

$$\mathbb{1}_{L(f,t)}(x) = \mathbb{1}_{[0,f(x)]}(t) \quad , \forall x \in \Omega$$

Thus :

- If $x \in L(f, t)$, it means $f(x) \geq t$, then $\mathbb{1}_{L(f,t)}(x) = \mathbb{1}_{[0,f(x)]}(t) = 1$.

- If $x \notin L(f, t)$, it means $f(x) < t$, then $\mathbb{1}_{L(f,t)}(x) = \mathbb{1}_{[0,f(x)]}(t) = 0$

2.6.2 Proof

We use the above lemma, then :

$$\int_0^\infty \mathbb{1}_{L(f,t)}(x)\,\mathrm{d}t = \int_0^\infty \mathbb{1}_{[0,f(x)]}(t)\,\mathrm{d}t$$

$$= \int_0^{f(x)} \mathrm{d}t$$

$$= f(x)$$

2.7 Inequalities on probability

2.7.1 Markov

Given X a non-negative random variable and $a > 0$:

$$P(X \geq a) \leq \frac{\mathbb{E}[X]}{a}$$

2.7.2 Chebyshev

Given X a random variable admits a second moment, then $\forall \varepsilon > 0$:

$$P(\mid X - \mathbb{E}[X] \mid > \varepsilon) \leq \frac{Var(X)}{\varepsilon^2}$$

2.7.3 Jensen

Given X a real-valued random variable and ϕ a convex function.

$$\phi(\mathbb{E}[X]) \leq \mathbb{E}[\phi(X)]$$

2.7.3.1 Conditional Jensen's Inequality

Given X an real-value integrable random variable and let $f : \mathbb{R} \to \mathbb{R}$ be a convex function such that $f(X)$ is also integrable, then

$$f(\mathbb{E}[X \mid \mathcal{G}]) \leq \mathbb{E}[f(X) \mid \mathcal{G}]$$

Note that, the inequality changes direction if f is concave. This is trivally proved by using $-f$, which is convex function.

2.7.4 Product with indicator function

If X and Y two random variables such that $|X| < |Y|$, then

$$|X|.\mathbb{1}_{\{|X|>K\}} \leq |Y|.\mathbb{1}_{\{|Y|>K\}}$$

since $\{|X| > K\} \subseteq \{|Y| > K\}$. Thus, $\forall \omega, |X(\omega)| > K \to |Y(\omega)| > K$.

2.7.5 Absolute sum decomposition

Let X and Y two random variables, then

$$|X + Y|.\mathbb{1}_{|X+Y| \geq 2K} \leq 2|X|.\mathbb{1}_{\{|X| \geq K\}} + 2|Y|.\mathbb{1}_{\{|Y| \geq K\}}$$

Proof

- By inequality 2.7.4, We have :

$$|X + Y|.\mathbb{1}_{\{|X+Y| \geq 2K\}} \leq (|X| + |Y|).\mathbb{1}_{\{|X|+|Y| \geq 2K\}}$$

- Now, for all ω such that $|X(\omega)| \leq |Y(\omega)|$, then $|X| + |Y| \leq 2|Y|$, by inequality 2.7.4 :

$$(|X| + |Y|).\mathbb{1}_{\{|X|+|Y| \geq 2K\}} \leq 2|Y|.\mathbb{1}_{\{2|Y| \geq 2K\}} = 2|Y|.\mathbb{1}_{\{|Y| \geq K\}}$$

- By anology, for all ω such that $|Y(\omega)| \geq |X(\omega)|$ then we have

$$(|X| + |Y|).\mathbb{1}_{\{|X|+|Y| \geq 2K\}} \leq 2|X|.\mathbb{1}_{\{|X| \geq K\}}$$

Finally, for all ω, we have :

$$(|X + Y|).\mathbb{1}_{\{|X+Y| \geq 2K\}} \leq 2|X|.\mathbb{1}_{\{|X| \geq K\}} + 2|Y|.\mathbb{1}_{\{|Y| \geq K\}}$$

2.7.6 Holder

Given $p, q \in (1, \infty)$, with $1/p + 1/q = 1$.

$$\mathbb{E}[|XY|] \leq \mathbb{E}[|X|^p]^{\frac{1}{p}} \mathbb{E}[|Y|^q]^{\frac{1}{q}}$$

In some case, Jensen inequality is used with $X = |X|^r, Y = 1$ and $p = \frac{s}{r}$, with $0 < r < s$:

$$\mathbb{E}[|X|^r] \leq E[|X|^s]^{\frac{r}{s}}$$

2.7.7 Boole's inequality

Also called *union bound* for **countable sets** $A_1, A_2, \ldots$ we have

$$P(\bigcup_{i=1}^{\infty} A_i) \leq \sum_{i=1}^{\infty} P(A_i)$$

Moreover, we have Bonferroni inequalities :

$$P(\bigcup_{i=1}^{\infty} A_i) \geq \sum_{i=1}^{\infty} P(A_i) - \sum_{i<j} P(A_i \cap A_j)$$

$$P(\bigcup_{i=1}^{\infty} A_i) \leq \sum_{i=1}^{\infty} P(A_i) - \sum_{i<j} P(A_i \cap A_j) + \sum_{i<j<k} P(A_i \cap A_j \cap A_k)$$

and ect ...

2.7.8 Expected value with negative and positive part

Let's remind the definition for the negative and positive part of number x :

$$x^- = -\min(x, 0)$$
$$x^+ = \max(x, 0)$$

Then given a random variable X, we have

$$\mathbb{E}[(X - a)^-] - \mathbb{E}[X^-] \leq a^+ \quad \forall a \in \mathbb{R}$$

Proof

$$\mathbb{E}[(X-a)^-] - \mathbb{E}[X^-]$$

$$= \int_{-\infty}^{a} (a-x)p(x)dx + \int_{a}^{+\infty} 0p(x)dx$$

$$- \int_{-\infty}^{0} -xp(x)dx - \int_{0}^{+\infty} 0p(x)dx$$

$$= \int_{-\infty}^{a} ap(x)dx + \begin{cases} \int_{0}^{a} -xp(x)dx \text{ if } a \geq 0 \\ -\int_{a}^{0} -xp(x)dx \text{ if } a < 0 \end{cases}$$

$$\leq \int_{-\infty}^{a} ap(x)dx$$

$$\leq \max(a, 0)$$

$$= a^+$$

In the same way, we have

$$\mathbb{E}[(X-a)^+] - \mathbb{E}[X^+] \leq a^- \quad \forall a \in \mathbb{R}$$

Proof

Let's remind that $x = x^+ - x^-$, we have :

$$\mathbb{E}[(X-a)] = \mathbb{E}[(X-a)^+] - \mathbb{E}[(X-a)^-]$$
$$\mathbb{E}[X] = \mathbb{E}[X^+] - \mathbb{E}[X^-]$$

Then

$$\mathbb{E}[(X-a)^+] - \mathbb{E}[X^+]$$
$$= \mathbb{E}[(X-a)] + \mathbb{E}[(X-a)^-] - \mathbb{E}[X] - \mathbb{E}[X^-]$$
$$= \mathbb{E}[(X-a)^-] - \mathbb{E}[X^-] - a$$
$$\leq a^+ - a$$
$$= a^-$$

Finally, note that $x^- = (-x)^+$ and $x^+ = (-x)^-$, then we can write :

$$\mathbb{E}[(a-X)^+] - \mathbb{E}[(-X)^+] \leq a^+ \quad \forall a \in \mathbb{R}$$

and

$$\mathbb{E}[(a-X)^-] - \mathbb{E}[(-X)^-] \leq a^- \quad \forall a \in \mathbb{R}$$

2.7.9 Divers

2.7.9.1 Inequality 1

Given random variable X and two functions f_1, f_2, with $f_1(X) \leq f_2(X)$, then :

$$P_X(f_1(X) \geq a) \leq P_X(f_2(X) \geq a)$$

2.7.9.2 Inequality 2

Given probability space $(\Omega, \mathcal{F}, P)$, then for all $A \in \mathcal{F}$, we have :

$$X \mathbb{1}_A \leq \max(X, 0)$$

Proof

Given $\omega \in \Omega$, then

- $\omega \in A$, then we have $X \leq \max(X, 0)$

- $\omega \notin A$, then we have $0 \leq \max(X, 0)$

2.8 Derivability and differentiability

- A function f is said **derivable** at x_0 if there exists

$$\lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0}$$

- A function f is said **differentiable** at x_0 if there exists a such that

$$f(x_0 + \varepsilon) \approx f(x_0) + a\varepsilon, \quad \forall |\varepsilon| \ll 1$$

2.8.0.1 Properties

- A function is differentiable then it is derivable but the converse is not true, e.g.

$$f(x) = \begin{cases} x + 1 \text{ if } x \geq 0 \\ x - 1 \text{ if } x < 0 \end{cases}$$

This function is derivable at $x_0 = 0$. But it is not differentiable since it is not continuous.

- A function f is differentiable but its derivative is does not have to be continuous, e.g.

$$f(x) = \begin{cases} 0 & \text{if } x = 0 \\ x^2 \sin\left(\frac{1}{x}\right) & \text{otherwise} \end{cases}$$

is differentiable, but its derivative

$$f'(x) = \begin{cases} 0 & \text{if } x = 0 \\ 2x \sin\left(\frac{1}{x}\right) - \cos\left(\frac{1}{x}\right) & \text{otherwise} \end{cases}$$

is not continuous at $x = 0$.

- A function f is differentiable but its derivative does not have to be Riemann integrable, e.g.

$$f(x) = \begin{cases} 0 & \text{if } x = 0 \\ x^2 \sin\left(\frac{1}{x^2}\right) & \text{otherwise} \end{cases}$$

is differentiable on $[-1, 1]$ but its derivative is unbounded on $[-1, 1]$, then it is not Riemann integrable.

2.9 Radon–Nikodym theorem

The Radon–Nikodym theorem is a result in measure theory that shows under certain conditions, **there exists a relationship between two measures** defined on the same measurable space. Before going to a formal description, let's remind the definition for (oridnary) continuity, uniform continuity, absolute continuity for a function and absolute continuity for measure.

2.9.1 Oridnary continuity, uniform continuity and absolute continuity

For a function $f \colon X \to Y$ with metric spaces (X, d_1) and (Y, d_2) :

- The function f is said to be **continuous** if $\forall x \in X \ \forall \varepsilon > 0 \ \exists \delta > 0 \ \forall y \in X :$ $d_1(x, y) < \delta \implies d_2(f(x), f(y)) < \varepsilon$.

- The function f is said to be **uniformly continuous** if $\forall \varepsilon > 0 \ \exists \delta > 0 \ \forall x \in X \ \forall y \in X : d_1(x, y) < \delta \implies d_2(f(x), f(y)) < \varepsilon$.

At a quick look, these two defintions seem to be the same. However, let's focus on the position of $(\forall x \in X)$, which make the difference between these two concepts. In short, in the definition of continuity, we first fix x then ε, but in the definition of uniform continuity, we first fix ε then x.

2.9.1.1 Example

Given two metric space $(\mathbb{R}, |.|)$ and $(\mathbb{R}, |.|)$, where $|.|$ is for L_1 distance. The function $f(x) = \frac{1}{x}$ on $(0, 1]$ is continuous but not uniformly continous. This is because we fix first ε, then suppose that exists δ such that forall x, y if $|x - y| < \delta$ then $\left| \frac{1}{x} - \frac{1}{y} \right| < \varepsilon$. The latter is not true since :

$$\left| \frac{1}{x} - \frac{1}{y} \right| = \frac{|x - y|}{|xy|} > +\infty$$

if $y = x + \frac{\delta}{2}$ and $x, y \to 0^+$.

By this example, we see that continuous functions can *fail* to be uniformly continuous if they are unbounded on a closed interval.

2.9.1.2 Absolute continuity

For a function $f \colon X \to Y$ with metric spaces (X, d_1) and (Y, d_2). The function f is **absolutely continuous** if $\forall \varepsilon > 0$, $\exists \delta > 0$ such that whenever a finite sequence of pairwise disjoint sub-intervals (x_k, y_k) in X satisfies

$$\sum_k d_1(y_k, x_k) < \delta$$

then

$$\sum_k d_2(f(y_k), f(x_k)) < \varepsilon$$

Equivalent definitions

Over interval $[a, b]$, the following statement are equivalent :

1. f is absolutely continuous.

2. f has a derivative f' almost everywhere (analogous to the notion of almost surely in probability theory) and the derivative f' is Lebesgue integrable (sec 1.9.5).

3. There exists a Lebesgue integrable function g on $[a, b]$ such that

$$f(x) = f(a) + \int_a^x g(t)\, dt$$

for all $x \in [a, b]$.

2.9.2 Absolute continuity of measure

Given :

- A measure μ (sec 1.7) on measurable space $(\mathbb{R}, B(\mathbb{R}))$, where $B(\mathbb{R})$ denotes Borel algebra (in sec 1.1.3) on $\mathbb{R}$.

- A Lebesgue measure λ (in sec 1.7.3.3) on the same measurable space $(\mathbb{R}, B(\mathbb{R}))$.

Then the measure μ is said absolutely continuous with respect to the Lebesgue measure λ or absolutely continuous (in short) if μ is dominated by λ ($\mu \ll \lambda$), which mean :

$$\lambda(E) = 0 \Rightarrow \mu(E) = 0, \quad \forall \lambda\text{-measurable } E \in B(\mathbb{R})$$

equivalently,

$$\mu(E) > 0 \Rightarrow \lambda(E) > 0, \quad \forall \lambda\text{-measurable } E \in B(\mathbb{R})$$

The same principle holds for measures on Borel subsets of $\mathbb{R}^n, n \geq 2$.

2.9.2.1 Equivalent definitions

The following statements are equivalent :

1. μ is absolutely continuous.

2. $\forall \varepsilon > 0, \exists \delta > 0$ such that for all λ-measurable E if $\lambda(E) < \delta$, then

$$\mu(E) < \varepsilon$$

3. There exists a Lebesgue integrable function g on the real line such that

$$\mu(E) = \int_E g\, d\lambda$$

2.9.3 Illustrative case

Given a measure μ on a measurable space (X, Σ) and a density function f, then for each $A \in \Sigma$, $\nu(A)$ is defined by Lebesgue integral (sec 1.9) of f over A with measure μ,

$$\nu(A) = \int_A f \, d\mu,$$

is also a measure on (X, Σ).

Example 1

If f represents the mass density and μ is the Lebesgue measure in three-dimensional space $\mathbb{R}^3$, which means volume. Then $\mu(A)$ is the volume of A and $\nu(A)$ is a new measure, which is thus the mass of A.

Example 2

If f represents the probability density function and μ is the Lebesgue measure in one-dimensional space $\mathbb{R}$, which means length. Then $\mu(A)$ is the length of A and $\nu(A)$ is a new measure, which is the probability that A happens.

2.9.4 Statement

The Radon–Nikodym theorem involves a measurable space (X, Σ) on which two σ-finite measures (sec 1.7.1.2) are defined, μ and ν. It states that, if $\nu \ll \mu$ (that is, if ν is absolutely continuous with respect to μ, sec 2.9.2), **then there exists** a Σ-measurable function $f \colon X \to [0, \infty)$, such that for any measurable set $A \in \Sigma$:

$$\nu(A) = \int_A f \, d\mu,$$

- The function f is μ-almost everywhere unique, that is, if g is an another function which satisfies the same property, then $f = g$ μ-almost everywhere (in sec 1.7.1.3).

- The function f is commonly written $\frac{d\nu}{d\mu}$ and is called the **Radon–Nikodym derivative**.

2.9.4.1 Extension to signed or complex measures

The Radon–Nikodym theorem is still true if μ is a nonnegative σ-finite measure, and ν is a finite signed measure (sec 1.7.1.1).

2.9.4.2 Corollary

If f and g are both $\mathcal{H}$-measurable, where $\mathcal{H} \subseteq \Sigma$ and

$$\int_A f \, d\mu = \int_A g \, d\mu,$$

then $f = g$ μ-almost everywhere.

If g and f are not both $\mathcal{H}$-measurable, a counter example is the Conditional expectation with respect to a sub sigma algebra (in sec 2.10)

2.10 Conditional expectation with respect to a (sub) σ-algebra

Consider the following:

- $(\Omega, \mathcal{F}, P)$ is a probability space.

- $X \colon \Omega \to \mathbb{R}^n$ is a random variable on that probability space with **defined expectation** (sec 1.9.5.1).

- $\mathcal{H} \subseteq \mathcal{F}$, which means $\mathcal{H}$ is a sub σ-algebra of $\mathcal{F}$.

A conditional expectation of X given $\mathcal{H}$, denoted as $\mathbb{E}[X \mid \mathcal{H}]$, is any $\mathcal{H}$-**measurable function** or **random variable** $H \colon \Omega \to \mathbb{R}^n$ which satisfies:

$$\int_A \mathbb{E}[X \mid \mathcal{H}](A)\, \mathrm{d}P = \int_A H(A)\, \mathrm{d}P = \int_A X\, \mathrm{d}P$$

for all events $A \in \mathcal{H}$. This implies

$$\mathbb{E}[X \mid \mathcal{H}](A) = \frac{\int_A X\, \mathrm{d}P}{\int_A \mathrm{d}P} = \frac{\int_A X\, \mathrm{d}P}{P(A)}$$

since $\mathbb{E}[X \mid \mathcal{H}](A)$ is a constant with given A.

2.10.0.1 Proof for existence

We can prove that $\mathbb{E}[X \mid \mathcal{H}]$ is always existing. Since X has a defined expection then $\mathbb{E}[|X|] < \infty$ or $[E] = \pm\infty$. In case that $\mathbb{E}[X] = \pm\infty$, we just take $\mathbb{E}[X \mid \mathcal{H}](A) = \pm\infty$ if $\int_A X\, \mathrm{d}P = \pm\infty$. In case that $\mathbb{E}[|X|] < \infty$, or X has finite expectation, we show in the following the existence of $\mathbb{E}[X \mid \mathcal{H}]$.

Thus, let's μ^X is defined by :

$$\mu^X(A) = \int_A X\, \mathrm{d}P, \forall A \in \mathcal{F}$$

- Since X is not non-negativity then we can only say that μ^X is signed measure (sec 1.7.1.1) and from that X has fnite expectation then μ^X is **finite signed measure**.

- If $P(A) = 0$ then $P(A_i) = 0$ for all $\bigcup_{i=0}^{\infty} = A$, then $\mu^X(A) = 0$ or μ^X is absolutely continuous with respect to P or $\mu^X \ll P$.

Let $P|_{\mathcal{H}}$ and $\mu^X|_{\mathcal{H}}$ be respectively measure P and μ_X restricted to $\mathcal{H}$, which is thus simply means $P|_{\mathcal{H}}$ and $\mu^X|_{\mathcal{H}}$ are measure on $(\omega, \mathcal{H})$. Since $\mu^X \ll P$ in $\mathcal{F}$ and $\mathcal{H} \subseteq \mathcal{F}$ then $\mu^X|_{\mathcal{H}} \ll P|_{\mathcal{H}}$. By Radon–Nikodym theorem (sec 2.9), there exists

$$\mathbb{E}(X \mid \mathcal{H}) = \frac{\mathrm{d}\mu^X|_{\mathcal{H}}}{\mathrm{d}P|_{\mathcal{H}}}$$

2.10.0.2 Important note

Do not confuse $\mathbb{E}[X \mid \mathcal{H}]$ with X. Note that

- X is $\mathcal{F}$-measurable and **is not** $\mathcal{H}$-measurable.

- $\mathbb{E}[X \mid \mathcal{H}]$ is $\mathcal{H}$-measurable then it **is also** $\mathcal{F}$-measurable.

We remind the set measurability in sec 1.5.3). See the following example for being clear.

2.10.1 Example

Consider the a probability space $(\Omega, \mathcal{F}, P)$ related to a dice and a random variable X on this space :

ω	I	II	III	IV	V	VI
X	1	2	3	4	5	6

Given the sub $\sigma-$algebra (of $\mathcal{F}$) $\mathcal{H} = \{\Omega, \{\mathrm{I}, \mathrm{II}\}, \{\mathrm{I}, \mathrm{II}, \mathrm{III}, \mathrm{IV}\}, \varnothing\}$, then

- If $A = \{\mathrm{I}, \mathrm{II}\}$, the $\mathbb{E}[X \mid \mathcal{H}](A)$ is inferred by :

$$\int_A \mathbb{E}[X \mid \mathcal{H}](A) \, \mathrm{d}P = \int_A X \, \mathrm{d}P$$
$$\Rightarrow \mathbb{E}[X \mid \mathcal{H}](A)(\frac{1}{6} + \frac{1}{6}) = (1 \times \frac{1}{6} + 2 \times \frac{1}{6})$$
$$\Rightarrow \mathbb{E}[X \mid \mathcal{H}](A) = \frac{3}{2}$$

- If $A = \{\mathrm{I}, \mathrm{II}, \mathrm{III}, \mathrm{IV}\}$, the $\mathbb{E}[X \mid \mathcal{H}](A)$ is inferred by :

$$\int_A \mathbb{E}[X \mid \mathcal{H}](A) \, \mathrm{d}P = \int_A X \, \mathrm{d}P$$
$$\Rightarrow \mathbb{E}[X \mid \mathcal{H}](A)(\frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6}) = (3 \times \frac{1}{6} + 4 \times \frac{1}{6} + 5 \times \frac{1}{6} + 6 \times \frac{1}{6})$$
$$\Rightarrow \mathbb{E}[X \mid \mathcal{H}](A) = 3$$

2.11 Law of iterated expectation

2.11.1 With one time conditioned

Also called law of total expectation, the tower rule or the smoothing theorem :

$$\mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[X]$$

Proof

$$\mathbb{E}[\mathbb{E}[X|Y]] = \int_y \mathbb{E}[X|Y=y]P(Y=y)dy$$

$$= \int_y \int_x xP(X=x|Y=y)dxP(Y=y)dy$$

$$= \int_y \int_x xP(X=x,Y=y)dxdy$$

$$= \int_x \int_y xP(X=x,Y=y)dydx \quad \text{(by Fubini's theorem)}$$

$$= \int_x xP(X=x)dx$$

$$= \mathbb{E}[X]$$

2.11.1.1 Conditional by sigma algebra

Let $(\Omega, \mathcal{F}, P)$ be a probability space and $\mathcal{H} \subseteq \mathcal{F}$, then

$$\mathbb{E}[\mathbb{E}[X \mid \mathcal{H}]] = \mathbb{E}[X]$$

Proof

We can understand that $\mathbb{E}[X]$ is thus a random variable, which means $\mathbb{E}[X] = \mathbb{E}[X \mid \mathcal{F}]$ can be evaluated at every $A \in \mathcal{F}$, denoted $\mathbb{E}[X \mid \mathcal{F}](A)$. Conventionally, wit $\mathbb{E}[X] = \mathbb{E}[X](\Omega)$.

From sec 3.4.2,

$$\int_A \mathbb{E}[X \mid \mathcal{H}] \, dP = \int_A X \, dP, \quad \forall A \in \mathcal{H}$$

then taking $A = \Omega \in \mathcal{H}$, we have

$$\mathbb{E}[\mathbb{E}[X \mid \mathcal{H}]] = \int_\Omega \mathbb{E}[X \mid \mathcal{H}] \, dP$$

$$= \int_\Omega X \, dP$$

$$= \mathbb{E}[X]$$

Note that to use $\mathcal{H} = \sigma(Y)$, then we refind $\mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[X]$.

2.11.2 With two time conditioned

$$\mathbb{E}[\mathbb{E}[X \mid Y, Z] \mid Z] = \mathbb{E}[X \mid Z]$$

ω	a	b	c	d	e	f
$X(\omega)$	1	2	3	4	5	6
$Y(\omega)$	0	-1	-1	0	0	-1
$Z(\omega)$	-1	0	-1	0	0	-1

2.11.2.1 Illustrative example

Let's consider a fair dice with 6 faces labelled by $a, \ldots, f$ and three random variables X, Y, Z such that :

In case $Z = 0$

- On one hand

  - $\mathbb{E}[X|Y=0, Z=0] = \frac{4+5}{2} = \frac{9}{2}$.
  - In the same manner, $\mathbb{E}[X|Y=1, Z=0] = 2$.
  - Then $\mathbb{E}[\mathbb{E}[X|Y, Z=0]|Z=0] = \mathbb{E}[X|Y=0, Z=0]P(Y=0, Z=0) + \mathbb{E}[X|Y=1, Z=0]P(Y=1, Z=0) = \frac{9}{2} \times \frac{2}{3} + 2 \times \frac{1}{3} = \frac{11}{3}$.

- On the other hand $\mathbb{E}[X|Z=0] = \frac{11}{3}$

Formal proof

Note that $\mathbb{E}[X \mid Y, Z]$ is a function of random variables $g(Y, Z)$ (see clearly in sec 3.4.2). Then $\mathbb{E}[\mathbb{E}[X \mid Y, Z] \mid Z]$ means the conditional expectation this function $g(Y, Z)$, **given** Z. We have

$$\mathbb{E}[\mathbb{E}[X \mid Y, Z] \mid Z]$$
$$= \int_y \mathbb{E}[X \mid Y=y, Z=z]P(Y=y \mid Z=z)dy$$
$$= \int_y \int_x XP(X=x \mid Y=y, Z=z)dx \frac{P(Y=y, Z=z)}{P(Z=z)}dy$$
$$= \int_x \int_y \frac{P(X=x, Y=y, Z=z)}{P(Z=z)}dydx$$
$$= \int_x \frac{P(X=x, Z=z)}{P(Z=z)}dx$$
$$= \int_x P(X=x|Z=z)dx$$
$$= \mathbb{E}[X|Z]$$

2.11.3 General case

Let $(\Omega, \mathcal{F}, P)$ be a probability space. Two sub σ-algebras $\mathcal{H}_1$ and $\mathcal{H}_2$ such that $\mathcal{H}_1 \subseteq \mathcal{H}_2 \subseteq \mathcal{F}$. For a random variable X on $(\Omega, \mathcal{F}, P)$, if $\mathbb{E}[X]$ is defined, which means $\min(\mathbb{E}[X^-], \mathbb{E}[X^+]) < \infty$ (in sec 1.9.5.1), then

$$\mathbb{E}[\mathbb{E}[X \mid \mathcal{H}_2] \mid \mathcal{H}_1] = \mathbb{E}[X \mid \mathcal{H}_1]$$

2.11.3.1 Proof

In general, we need to prove :

1. The existence of $\mathbb{E}[\mathbb{E}[X \mid \mathcal{H}_2] \mid \mathcal{H}_1]$ and $\mathbb{E}[X \mid \mathcal{H}_1]$.

2. $\mathbb{E}[\mathbb{E}[X \mid \mathcal{H}_2] \mid \mathcal{H}_1]$ and $\mathbb{E}[X \mid \mathcal{H}_1]$ are both $\mathcal{H}_1$ measurable or equivalently, they are both measurable function on $(\Omega, \mathcal{H}_1)$.

3. $\int_A \mathbb{E}[\mathbb{E}[X \mid \mathcal{H}_2] \mid \mathcal{H}_1] \, \mathrm{d}P = \int_A \mathbb{E}[X \mid \mathcal{H}_1] \, \mathrm{d}P$ for all $A \in \mathcal{H}_1$.

The first requirement :

- Since X has a defined expectation then $\mathbb{E}[X \mid \mathcal{H}_1]$ exists (sec 2.10).

- Let's $Y = \mathbb{E}[X \mid \mathcal{H}_2]$, since $\mathbb{E}[Y] = \mathbb{E}[X]$ (sec 2.11.1.1) is defined, then $\mathbb{E}[Y \mid \mathcal{H}_1]$ exists or $\mathbb{E}[\mathbb{E}[X \mid \mathcal{H}_2] \mid \mathcal{H}_1]$ exists.

The second requirement is trivial by the definition of conditional expectation (sec 2.10).

The third requirement. From the definition of conditional expectation (sec 2.10) :

- $\int_A \mathbb{E}[X \mid \mathcal{H}_1] \, \mathrm{d}P = \int_A X \, \mathrm{d}P$ for all $A \in \mathcal{H}_1$

- $\int_A \mathbb{E}[\mathbb{E}[X \mid \mathcal{H}_2] \mid \mathcal{H}_1] \, \mathrm{d}P = \int_A \mathbb{E}[X \mid \mathcal{H}_2] \, \mathrm{d}P$ for all $A \in \mathcal{H}_1$. Since $\mathcal{H}_1 \subseteq \mathcal{H}_2$, then $A \in \mathcal{H}_2$. This infers $\int_A \mathbb{E}[X \mid \mathcal{H}_2] \, \mathrm{d}P = \int_A X \, \mathrm{d}P$ for all $A \in \mathcal{H}_1$.

We use the corollary of Radon–Nikodym theorem 2.9.4.2 for the second and the third to have q.e.d.

2.11.3.2 Corollary

- If $\mathcal{H}_1 = \sigma(Z)$ and $\mathcal{H}_2 = \sigma(Y, Z)$, we refind $\mathbb{E}[\mathbb{E}[X \mid Y, Z] \mid Z] = \mathbb{E}[X \mid Z]$.

- If $\sigma(Z) \subseteq \sigma(Y)$, we can write $\mathbb{E}[\mathbb{E}[X \mid Y] \mid Z] = \mathbb{E}[X \mid Z]$.

2.12 Fatou's lemma

Given a measure space $(\Omega, \mathcal{F}, \mu)$ and a set $X \in \mathcal{F}$. $\mathcal{B}_{\mathbb{R}_{\geq 0}}$ denotes the Borel set on real nonnegative number. Let f_n be measurable functions on $(\mathcal{F}, \mathcal{B}_{\mathbb{R}_{\geq 0}})$, $f_n : \Omega \to \mathbb{R}_{\geq 0}$. Define the function $f : \Omega \to \mathbb{R}_{\geq 0}$ by setting

$$f(\omega) = \lim_{n \to +\infty} \inf_{k \geq n} f_k(\omega[0 : k]) \quad \forall \omega \in X$$

where ω here is an outcome related to a **realization** and $\omega[0 : k]$ is information of ω until instant k. In this case, we can say f_k must be **almost surely convergent** to f (sec 4.6). Then f is also $(\mathcal{F}, \mathcal{B}_{\mathbb{R}_{\geq 0}})$-measurable and

$$\int_X f(\omega) \, d\mu(\omega) \leq \lim_{n \to +\infty} \inf_{k \geq n} \int_X f_k(\omega[0 : k]) \, d\mu(\omega[0 : k])$$

or

$$\int_X \varliminf_{n \to +\infty} \inf_{k \geq n} f_k(\omega[0:k]) \, d\mu(\omega[0:k]) \leq \varliminf_{n \to +\infty} \inf_{k \geq n} \int_X f_k(\omega[0:k]) \, d\mu(\omega[0:k])$$

If in case that $X = \Omega$ then :

$$\mathbb{E}\left[\lim_{n \to +\infty} \inf_{k \geq n} f_k\right] \leq \lim_{n \to +\infty} \inf_{k \geq n} \mathbb{E}\left[f_k\right]$$

A literature explanation for this lemma is that integral of limit inferior is less than limit inferior of integral. In case of mean, taking mean first then taking limit inferior is greater than taking limit inferior first then taking mean. The notation $\lim_{n \to +\infty} \inf_{k \geq n}$ can be noted $\liminf_{n \to +\infty}$ for short.

2.12.1 Reverse Fatou lemma

If there exists a non-negative integrable function g on Ω such that $f_n(\omega) \leq g(\omega)$ for all ω, n (statewise dominance 7.1), then

$$\int_X \limsup_{n \to +\infty} f_n \, d\mu \geq \limsup_{n \to +\infty} \int_X f_n \, d\mu$$

Here g integrable means that g is $(\mathcal{F}, \mathcal{B}_{\mathbb{R}_{\geq 0}})-$measurable and that $\int_X g(\omega) \, d\mu(\omega) < \infty$. The role of g here is to make an upper bound for f_n and make exist $\int_X \limsup_{n \to +\infty} f_n \, d\mu$.

2.13 Fubini–Tonelli theorem

Given measure space $(X, \mathcal{A}, \mu)$ and $(Y, \mathcal{B}, \nu)$ with two measures μ and ν are σ-finite measure (1.7.1.2). Then the $(X \times Y, \mathcal{A} \otimes \mathcal{B}, \mu \times \nu)$ is also a measure space with a product measure. If

$$f : X \times Y \to \mathbb{R}$$

is a function that is $\mathcal{A} \otimes \mathcal{B}$-measurable, then two functions

$$g_y(x) := x \mapsto \int_Y |f(x, y)| \, \mathrm{d}\nu(y)$$

and

$$g_x(y) := y \mapsto \int_X |f(x, y)| \, \mathrm{d}\mu(x)$$

are respectively $\mathcal{A}$-measurable and $\mathcal{B}$-measurable.

Besides, if

$$\int_X \int_Y |f(x, y)| \, \mathrm{d}\nu(y)\mu(x)$$

is finite then

$$x \mapsto \int_Y f(x,y) \, \mathrm{d}\nu(y)$$

$$y \mapsto \int_X f(x,y) \, \mathrm{d}\nu(x)$$

are respectively $\mathcal{A}$-measurable and $\mathcal{B}$-measurable.

2.14 Convergence monotone theorem

Also called Beppo Levi theorem. If $f_n \colon E \to [0, +\infty)$ is a sequence of measurable functions on a measurable set E such that $f_n \to f$ pointwise almost everywhere and f_n is increasing ($f_1 \leq f_2 \leq \ldots$), then :

$$\lim_{n \to \infty} \int_E f_n = \int_E f$$

The Monotone Convergence Theorem (MCT), the Dominated Convergence Theorem (DCT 2.15), and Fatou's Lemma (2.12) are three major results in the theory of Lebesgue integration that answer the question : When we have that $\lim_{n \to \infty}$ and $\int$ can commute.

2.15 Dominated convergence theorem

2.15.1 Statement

Let $(f_n)_{n \in \mathbb{N}}$ be a sequence of measurable function (sec 1.5) on a measure space (S, Σ, μ). f_n is from (S, Σ) to $(\mathbb{C}, B(\mathbb{C}))$. Suppose that

1. f_n is pointwise convergent (sec 4.1.1) to a function f. In general case, it is *converge in measure*.

2. f_n dominated by an integrable function g in the sense that

$$|f_n(x)| \leq g(x), \forall x \in S \text{ and } \forall n \in \mathbb{N}$$

Then f is integrable and

$$\lim_{n \to \infty} \int |f_n - f| \, \mathrm{d}\mu = 0,$$

which implies

$$\lim_{n \to \infty} \int f_n \, \mathrm{d}\mu = \int \lim_{n \to \infty} f_n \, \mathrm{d}\mu = \int f \, \mathrm{d}\mu$$

2.15.2 Corollary

If in case of probability space, suppose that :

1. $X_n \xrightarrow{p} X$ (converge in probability 4.3).

2. $|X_n| \leq Y$, where Y is integrable, which means $\int |Y| \, dP < +\infty$.

Then X is integrable and

$$\lim_{n \to \infty} \mathbb{E}[X_n] = \mathbb{E}[X]$$

Note that this corrollary is weaker than the lemma 4.5.4, since in proposition 3.3.5 :

- $|X_n| \leq Y$, where Y is integrable $\Rightarrow X_n$ is uniformly integrable.

- X_n is uniformly integrable $\nRightarrow$ there exists integrable such that $Y \, |X_n| \leq Y$.

2.15.3 Example

Finding

$$\lim_{n \to \infty} \int_0^1 f_n(x) \, dx$$

where $f_n(x) = \frac{nx^n}{1+x}, \forall x \in [0, 1]$.

At a first look, $f_n(x)$ is not be pointwise convergent on $[0, 1]$, in particular at $x = 1$. Let's use the integration by part :

$$\int_0^1 \frac{nx^n}{1+x} \, dx = \frac{n}{2(n+1)} + \int_0^1 \frac{n}{(n+1)} \frac{x^{n+1}}{(1+x)^2} \, dx$$

$$= \frac{n}{2(n+1)} + \int_0^1 g_n(x) \, dx$$

Now use dominated convergence theorem to find :

$$\lim_{n \to \infty} \int_0^1 g_n(x) \, dx$$

1. $g_n(x)$ is pointwise convergent to $g(x) = \begin{cases} 0 \text{ if } x \in [0, 1) \\ 1/4 \text{ if } x = 1 \end{cases}$

2. $g_n(x)$ is dominated by the integrable function $\frac{1}{(1+x)^2}$ on $[0, 1]$.

Then by the dominated convergence theorem :

$$\lim_{n \to \infty} \int_0^1 g_n(x) \, dx = \int_0^1 0 \, dx = 0$$

Finally,

$$\lim_{n \to \infty} \int_0^1 \frac{nx^n}{1+x} \, dx = \frac{n}{2(n+1)} = \frac{1}{2}$$

3 Uniform integrability

In mathematics, uniform integrability is an important concept in measure theory. It plays a vital role in stochastic process, the theory of martingales (in sec 9.1),...

3.1 Uniform integrability in measure theory

Let $(\Omega, \mathcal{F}, \mu)$ be a positive measure space, which means $\mu(\omega) \geq 0, \forall \omega \in \Omega$. $\mathcal{S}$ is a **family of mesurable functions** f (in sec 1.5), from $(\Omega, \mathcal{F})$ to $(\mathbb{R}, B(\mathbb{R}))$. We say that $\mathcal{S}$ is uniformly integrable (abbreviated as UI) if the following conditions hold :

1. $\sup\limits_{f \in \mathcal{S}} \int |f| d\mu < +\infty$.

2. For all $\varepsilon > 0$, there exists a $\delta > 0$ such that for all $A \in \mathcal{F}$ that satify $\mu(A) < \delta$, we have :

$$\int_{\omega \in A} |f(\omega)| d\mu(\omega) < \varepsilon, \quad \forall f \in \mathcal{S}$$

Important note : As mentionned in the definition, the uniformly ingerability is eligible only for the **set (or family) of measurable functions** and **not for** a single measurable function.

3.2 Uniform integrability for set of random variable

Given probability space $(\Omega, \mathcal{F}, P)$ and $\mathcal{S}$ is a family of random variable X, from $(\Omega, \mathcal{F})$ to $(\mathbb{R}, B(\mathbb{R}))$. We say that $\mathcal{S}$ is uniformly integrable if the following conditions hold :

1. $\sup\limits_{X \in \mathcal{S}} \mathbb{E}[|X|] < +\infty$.

2. Absolutely continuous : For all $\varepsilon > 0$, there exists a $\delta > 0$ such that for all $A \in \mathcal{F}$ that satify $P(A) < \delta$, we have :

$$\int_{\omega \in A} |X(\omega)| \, dP(\omega) < \varepsilon, \quad \forall X \in \mathcal{S}$$

or

$$\mathbb{E}[|X|.\mathbb{1}_A] < \varepsilon, \quad \forall X \in \mathcal{S}$$

where $\mathbb{1}_A(\omega) = 1$ if $\omega \in A$, 0 otherwise.

Important note : Again, the uniformly ingerability **is eligible only for the set (or family) of random variables** and **not for** a single random variable.

3.2.1 Conditional expectation and expectation of product with indicator function

Do not confuse between two following concepts :

- $\mathbb{E}[X \mid \mathcal{H}](A)$, sometime shortly noted $\mathbb{E}[X \mid A]$, the conditional expectation (in sec 2.10).

- $\mathbb{E}[X.\mathbb{1}_A]$, the expectation of product between X and the indicator function $\mathbb{1}_A$.

Thus,

$$\mathbb{E}[X \mid A] = \frac{\int_A X \, \mathrm{d}P}{P(A)} = \frac{\mathbb{E}[X.\mathbb{1}_A]}{P(A)}$$

Note that, $X.\mathbb{1}_A$ is a new random variable whom distribution is modified from X (set $X(\omega) = 0$ for all $\omega \in A^C$).

3.2.2 Equivalent definition

The two previous conditions are equivalent to the following : For all $\varepsilon > 0$, there exists $K > 0$ such that

$$\mathbb{E}[|X|.\mathbb{1}_{\{|X|>K\}}] < \varepsilon, \quad \forall X \in \mathcal{S}$$

The set $\{|X| > K\}$ is short notation for $\{\omega : |X(\omega)| > K\}$. Thus the above condition means that the integral of X at two extreme sides is very small. Then it is also equivalent to

$$\lim_{K \to \infty} \sup_{X \in \mathcal{S}} \mathbb{E}[|X|.\mathbb{1}_{\{|X|>K\}}] = 0$$

Since the uniform integrability is eligible for set of random variables, we have also the definition of uniform integrability for stochastic process (which is also set of random variables with indexation) in the following section.

3.3 Uniform integrability of stochastic process

From the uniform integrability of set of random variables (in sec 3.2), we can expand this concept for stochastic process. A SP X_n is called uniformly integrable if the following condition holds :

$$\lim_{K \to \infty} \sup_{n \in \mathbb{N}} \mathbb{E}[|X_n|.\mathbb{1}_{\{|X_n|>K\}}] = 0$$

3.3.1 Examples

- **Uniform ingerable** stochastic process : Assume our probability space concerns a fair dice. X_n is the result of nth rolling. Then

  - with $K = 0$: $\displaystyle\sup_{n \in \mathbb{N}} \mathbb{E}[|X_n|.\mathbb{1}_{\{|X_n|>0\}}] = 3.5$

– with $K = 1 : \sup\limits_{n \in \mathbb{N}} \mathbb{E}[|X_n|.\mathbb{1}_{\{|X_n|>1\}}] = 4$

– ...

– with $K = 6 : \sup\limits_{n \in \mathbb{N}} \mathbb{E}[|X_n|.\mathbb{1}_{\{|X_n|>1\}}] = 1$

– with $K > 6 : \sup\limits_{n \in \mathbb{N}} \mathbb{E}[|X_n|.\mathbb{1}_{\{|X_n|>6\}}] = 0$

Then
$$\lim_{K \to \infty} \sup_{n \in \mathbb{N}} \mathbb{E}[|X_n|.\mathbb{1}_{\{|X_n|>K\}}] = 0$$

- **Non uniform ingerable** stochastic process : Assume our probability space is $([0,1], \mathcal{F}, P)$ and outcome ω is uniformly distributed in $[0,1]$. Let's take

$$X_n(\omega) = \begin{cases} n \text{ if } \omega \in [0, \frac{1}{n}] \\ 0 \text{ otherwise} \end{cases}$$

Then $\mathbb{E}[X_n] = 1$ or X_n is integrable with all $n \in \mathbb{N}$, even we have $\sup\limits_{n \in \mathbb{N}}(\mathbb{E}[|X_n|]) < \infty$. However, the process (X_n) is not uniformly integrable since if $n \leq K$, then

$$\mathbb{E}[|X_n|.\mathbb{1}_{\{|X_n|>K\}}] = \begin{cases} 0 \text{ if } n \leq K \\ 1 \text{ if } n > K \end{cases}$$

This means
$$\lim_{K \to \infty} \sup_{n \in \mathbb{N}} \mathbb{E}[|X_n|.\mathbb{1}_{\{|X_n|>K\}}] = 1 \neq 0$$

3.3.2 Equivalence between defintions

This is a proof for the equivalence between defintion in 3.2 and in 3.2.2. The statement is the following.

$$\lim_{K \to \infty} \sup_{n \in \mathbb{N}} \mathbb{E}[|X_n|.\mathbb{1}_{\{|X_n|>K\}}] = 0$$

if and only if we have two following conditions :

1. (X_n) is L^1 bounded :
$$\sup_{n \in \mathbb{N}} \mathbb{E}[|X_n|] < +\infty$$

2. (X_n) is absolutely continous :
$$\sup_{n \in \mathbb{N}} \mathbb{E}[|X_n|.\mathbb{1}_A] \to 0 \text{ as } P(A) \to 0$$

In the direct sense, let's $h(K) = \sup_{n \in \mathbb{N}} \mathbb{E}[|X_n|.\mathbb{1}_{\{|X_n| > K\}}]$. First, we have

$$\sup_{n \in \mathbb{N}} \mathbb{E}[|X_n|]$$
$$= \sup_{n \in \mathbb{N}} \left(\mathbb{E}[|X_n|.\mathbb{1}_{\{|X_n| > K\}}] + \mathbb{E}[|X_n|.\mathbb{1}_{\{|X_n| \le K\}}] \right)$$
$$\le \sup_{n \in \mathbb{N}} \mathbb{E}[|X_n|.\mathbb{1}_{\{|X_n| > K\}}] + \sup_{n \in \mathbb{N}} \mathbb{E}[|X_n|.\mathbb{1}_{\{|X_n| \le K\}}]$$
$$\le h(K) + KP(|X_n| \le K)$$
$$\le h(K) + K, \quad \forall K > 0$$

Since $\lim_{K \to \infty} h(K) = 0$, then there exists K such that $h(K) < \infty$. Then X_n is L^1 bounded. Second,

$$\sup_{n \in \mathbb{N}} \mathbb{E}[|X_n|.\mathbb{1}_A]$$
$$\le \sup_{n \in \mathbb{N}} \mathbb{E}[|X_n|.\mathbb{1}_{A \cup \{|X_n| > K\}}]$$
$$= \sup_{n \in \mathbb{N}} \mathbb{E}[|X_n|.\mathbb{1}_{(A \cap \{|X_n| \le K\}) \cup \{|X_n| > K\}}]$$
$$= \sup_{n \in \mathbb{N}} \left(\mathbb{E}[|X_n|.\mathbb{1}_{A \cap \{|X_n| \le K\}}] + \mathbb{E}[|X_n|.\mathbb{1}_{\{|X_n| > K\}}] \right)$$
$$\le \sup_{n \in \mathbb{N}} \mathbb{E}[|X_n|.\mathbb{1}_{A \cap \{|X_n| \le K\}}] + \sup_{n \in \mathbb{N}} \mathbb{E}[|X_n|.\mathbb{1}_{\{|X_n| > K\}}]$$
$$\le KP(A) + h(K)$$

Then if $K \to \infty$, we fix K first and then take $P(A) \to 0$, we have $\sup_{n \in \mathbb{N}} \mathbb{E}[|X_n|.\mathbb{1}_A] \to 0$.

In the converse sense, for a given $\varepsilon > 0$ suppose that there exists k such that

$$\mathbb{E}[|X_k|.\mathbb{1}_{\{|X_k| > K\}}] > \varepsilon$$

By Markov's inequality 2.7.1

$$P(|X_k| > K) \le \frac{\mathbb{E}[X_k]}{K}$$

Then if $K \to \infty$, then $P(|X_k| > K) \to 0$. By the property of absolutely continuous, if $P(|X_k| > K) \to 0$ then $\mathbb{E}[|X_k|.\mathbb{1}_{\{|X_k| > K\}}] \to 0$ (contradiction). Then X_n is uniformly integrable.

3.3.3 Linear sum of two uniform integrable processes

The linear sum of two uniform integrable processes is also uniform integrable.
Proof
Let's consider X_n and Y_n are both uniform integrable. Then we show $X_n + Y_n$ is uniformly integrable and αX_n is uniformly integrable.
First, by Absolute sum decomposition inequality 2.7.5

$$\lim_{K\to\infty} \sup_{n\in\mathbb{N}} \mathbb{E}[|X_n + Y_n|.\mathbb{1}_{\{|X_n+Y_n|>K\}}]$$

$$\leq \lim_{K\to\infty} \sup_{n\in\mathbb{N}} \left(2\mathbb{E}[|X_n|.\mathbb{1}_{\{|X_n|>\frac{K}{2}\}}] + 2\mathbb{E}[|X_n|.\mathbb{1}_{\{|X_n|>\frac{K}{2}\}}] \right)$$

$$=0$$

Second,

$$\lim_{K\to\infty} \sup_{n\in\mathbb{N}} \mathbb{E}[|\alpha X_n|.\mathbb{1}_{\{|\alpha X_n|>K\}}]$$

$$= \lim_{K\to\infty} \sup_{n\in\mathbb{N}} \alpha\mathbb{E}[|X_n|.\mathbb{1}_{\{|X_n|>\frac{K}{|\alpha|}\}}]$$

$$=0$$

3.3.4 Lemma

Given process Y_n which is uniformly integrable. If the process X_n satifies that $\exists K_0, \forall |X_n| \geq K_0$:

$$|X_n| \leq |Y_n|$$

Then X_n is uniformly integrable.

Proof

For all $K \geq K_0$:

$$\lim_{K\to\infty} \sup_{n\in\mathbb{N}} \mathbb{E}[|X_n|.\mathbb{1}_{\{|X_n|>K\}}]$$

$$\leq \lim_{K\to\infty} \sup_{n\in\mathbb{N}} \mathbb{E}[|Y_n|.\mathbb{1}_{\{|X_n|>K\}}]$$

$$\leq \lim_{K\to\infty} \sup_{n\in\mathbb{N}} \mathbb{E}[|Y_n|.\mathbb{1}_{\{|Y_n|>K\}}]$$

$$=0$$

where the last less or equal is by inequality 2.7.4.

3.3.5 Proposition

Let $(X_n)_{n\in\mathbb{N}}$ be a stochastic process, and let Y be an integrable random variable, such that $|X_n| \leq Y$ for all $n \in \mathbb{N}$ (dominated). Prove that (X_n) is uniformly integrable.

Before the main proof, we prove the following lemma

3.3.5.1 Lemma

If Y is integrable $(\mathbb{E}[|Y|] < \infty)$, then

$$\lim_{K\to\infty} \mathbb{E}[|Y|.\mathbb{1}_{\{|Y|>K\}}] = \lim_{K\to\infty} \int |Y|.\mathbb{1}_{\{Y>K\}} dP = 0$$

Proof

Let's set the sequence (f_K) where $f_K = |Y|.\mathbb{1}_{\{|Y|>K\}}$. Then :

1. (f_K) is pointwise convergent (4.1.1) to 0 (more precise, to $f(\omega) = 0, \forall \omega$), as $K \to +\infty$.

2. (f_K) is dominated by $|Y|$, which means $f_K(\omega) \leq |Y(\omega)|, \forall \omega$ and $|Y|$ is integrable.

By the dominated convergence theorem in sec2.15, we have

$$\lim_{K \to \infty} \int |Y|.\mathbb{1}_{\{|Y|>K\}}dP = \int \lim_{K \to \infty} |Y|.\mathbb{1}_{\{|Y|>K\}}dP = \int 0 dP = 0$$

3.3.5.2 Proof

First, since $|X_n| \leq Y$, we have

$$\{|X_n| > K\} \subseteq \{Y > K\}.$$

Then

$$\int |X_n|.\mathbb{1}_{\{|X_n|>K\}}dP = \int_{\{|X_n|>K\}} |X_n|dP \leq \int_{\{Y>K\}} Y dP = \int Y.\mathbb{1}_{\{Y>K\}}dP$$

for all $n \in \mathbb{N}$, or

$$\sup_{n \in \mathbb{N}} \mathbb{E}[|X_n|.\mathbb{1}_{\{|X_n|>K\}}] \leq \int Y.\mathbb{1}_{\{Y>K\}}dP$$

Second, by lemma in sec 3.3.5.1, Y is integrable and $Y = |Y|$ (since $Y > |X_n| > 0$), we have :

$$\lim_{K \to \infty} \int Y.\mathbb{1}_{\{Y>K\}}dP = 0$$

From the first and the second, we have :

$$\lim_{K \to \infty} \sup_{n \in \mathbb{N}} \mathbb{E}[|X_n|.\mathbb{1}_{\{|X_n|>K\}}] = 0$$

3.3.5.3 Converse statement

However, conversely, if $(X_n)_{n \in \mathbb{N}}$ is uniformly integrable, it does not guarantee that it exists a **integrable** random variable Y that $|X_n| \leq Y$ for $n \in \mathbb{N}$.

Counter example

Let $X_n = n$ with probability a_n, where $a_n \in (0, 1)$ and $X_n = 0$ otherwise. Assume all the X_n are independent. By definition of X_n, the expectation of the tail is

$$E[|X_n|.\mathbb{1}_{\{|X_n|\geq K\}}] = \begin{cases} na_n \text{ if } n \geq K \\ 0 \text{ if } n < K \end{cases}$$

Set $a_n = 1/(n \log n)$ for $n \geq 2$, then

$$E[|X_n|.\mathbb{1}_{\{|X_n| \geq K\}}] \leq n\frac{1}{n \log n} = \frac{1}{\log n} = \frac{1}{\log K}$$

or

$$\lim_{K \to \infty} \sup_{n \in \mathbb{N}} \mathbb{E}[|X_n|.\mathbb{1}_{\{|X_n| > K\}}] = 0$$

Thus the family (X_n) is uniformly integrable. However, if there were to exist a random variable Z such that $Z \geq |X_n|$ for all n, then we must have $P(Z \geq n) \geq P(X_n \geq n) = a_n$. From the inequality in sec 2.4.1, we have

$$\mathbb{E}[Z] \geq \sum_{i=1}^{+\infty} P(Z \geq i) \geq \sum_{i=2}^{+\infty} a_i = \sum_{i=2}^{+\infty} \frac{1}{i \log i}$$

However

$$\sum_{i=2}^{+\infty} \frac{1}{i \log i} \geq \frac{1}{2 \log 2} + \frac{2}{4 \log 4} + \frac{4}{8 \log 8} + \ldots + \frac{2^{k-1}}{2^k \log 2^k} \quad (k \to \infty)$$

$$= \frac{1}{2 \log 2}\left(1 + \frac{1}{2} + \frac{1}{3} + \ldots + \frac{1}{k}\right) \quad (k \to \infty)$$

$$= +\infty$$

Then we say that, it does not alway exist an integrable random variable Z which dominate (X_n).

3.3.6   Lemma

Random variable X is integrable if and only if X is absolutely continuous.
In the direct sense, since X is integrable then by lemma 3.3.5.1 :

$$\lim_{K \to \infty} \mathbb{E}[|X|.\mathbb{1}_{\{|X| > K\}}] = 0$$

Then $\forall \varepsilon > 0, \exists K_0, \forall K \geq K_0, \mathbb{E}[|X|.\mathbb{1}_{\{|X| > K_0\}}] < \varepsilon$.
Then the same as in the proof 3.3.2:

$$\mathbb{E}[|X|.\mathbb{1}_A]$$
$$\leq \mathbb{E}[|X|.\mathbb{1}_{A \cup \{|X| > K_0\}}]$$
$$= \mathbb{E}[|X|.\mathbb{1}_{A \cap \{|X| \leq K_0\}}] + \mathbb{E}[|X|.\mathbb{1}_{\{|X| > K_0\}}]$$
$$\leq K_0 P(A) + \mathbb{E}[|X|.\mathbb{1}_{\{|X| > K_0\}}]$$

Then $\forall 2\varepsilon > 0$, let's $\delta = \frac{\varepsilon}{K_0}$. Then if $P(A) < \delta$, we have $\mathbb{E}[|X|.\mathbb{1}_A] < 2\varepsilon$ or X is absolutely continuous.
In the converse sense, since X is absolutely continuous, then given a $\varepsilon > 0, \exists \delta$ such that if $P(A) < \delta$, we have $\mathbb{E}[|X|.\mathbb{1}_A] < \varepsilon$. Moreover, $P(|X| > K) \to 0$ as $K \to \infty$ then $\exists K_0, \forall K \geq K_0, P(|X| > K_0) < \delta$. Finally,

$$\mathbb{E}[|X|] = \mathbb{E}[|X|.\mathbb{1}_{\{|X| \leq K_0\}}] + \mathbb{E}[|X|.\mathbb{1}_{\{|X| > K_0\}}] \leq K_0 + \varepsilon < \infty$$

or X is integrable.

4 Convergence of Random Variables

In this section, we want to see a sequence of random variables $X_1, X_2, \ldots, X_n$ converges to a random variable X. There are four types of convergeces :

- Convergence in distribution
- Convergence in probability
- Convergence in mean
- Almost sure convergence.

All of them have the objective to prove that $X_n \to X$, but they are different in the method (by definition) and the intensity (from weak to strong) of convergence.
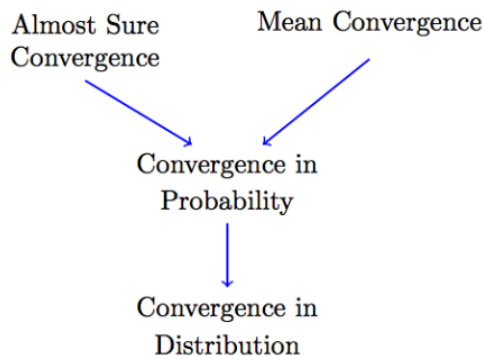
Figure 2.4: Relations between different types of convergence

Before diving into the convergence of random variable, it may be necessary remind basic definitions on convergence of function.

4.1 Convergence of function

4.1.1 Pointwise convergence

Let $\{f_n\}$ be a sequence of function, $f_n : \mathbb{R} \to \mathbb{R}$. $\{f_n\}$ is called pointwise convergent to f if $\forall x_0$ (x_0 is given first),

$$|f_n(x_0) - f(x_0)| < \varepsilon, \forall \varepsilon > 0,$$

when $n \to \infty$

4.1.2 Uniform convergence

Let $\{f_n\}$ be a sequence of function, $f_n : \mathbb{R} \to \mathbb{R}$. $\{f_n\}$ is called uniform convergent to f if $\forall \varepsilon > 0$ (ε is given first),

$$|f_n(x) - f(x)| < \varepsilon, \forall x,$$

when $n \to \infty$

4.1.3 Cauchy's criterion

The Cauchy's criterion is equivalent to uniform convergence. Let $\{f_n\}$ be a sequence of function, $f_n : \mathbb{R} \to \mathbb{R}$. $\{f_n\}$ is uniform convergent if $\forall \varepsilon > 0$ (ε is given first),

$$\exists n_0, \forall p, q > n_0, |f_p(x) - f_q(x)| < \varepsilon$$

4.1.4 Notes and example

- In pointwise convergence, ε can depend on x_0, but in uniform convergence, ε is independent to x_0.

- The uniform convergence is stronger than pointwise convergence, it means that if we have uniform convergence then we have also pointwise convergence but conversely, it is not true.

- In pointwise and uniform uniform convergence, f must be given, but in Cauchy's criterion, f is no needed.

Let's take $f_n(x) = x^n$ and consider two interval $[0, r]$ where $r < 1$ and $[0, 1)$.

- On $[0, r]$, f_n is both poinwise convergent and uniform convergent to $f = 0$.

- On $[0, 1)$, f_n is only poinwise convergent to $f = 0$, but not uniform convergent to $f = 0$. The problem for uniform convergence is if we fix first ε closed to 0, then $|x^n - 0|$ can be greater ε, by taking $x = 0.5^{1/n}$ then $x^n = 0.5$ or by taking $x = 1 - \frac{1}{n}$ then $x^n = \frac{1}{e}$.

4.2 Convergence in distribution

A sequence of random variables $X_1, X_2, X_3, \ldots, X_n$ converges in distribution (or converge weakly) to a random variable X, shown by $X_n \xrightarrow{d} X$, if :

$$\lim_{n \to \infty} F_{X_n}(x) = F_X(x)$$

for all x at which $F_X(x)$ is continuous. F is cumulative distribution.

4.2.1 Equivalent definition

We have the convergence in distribution if any of the following statements are true:

- $\lim\limits_{n\to\infty} P(X_n \leq x) \to P(X \leq x), \forall x.$

- $\lim\limits_{n\to\infty} \mathbb{E}[f(X_n)] \to \mathbb{E}[f(X)]$ for all bounded and continuous functions f.

4.2.2 Example

Given a sequence of random variables $X_1, X_2, X_3, \ldots, X_n$ such that:

$$F_{X_n}(x) = \begin{cases} 1 - \left(1 - \frac{1}{n}\right)^{nx} & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

Show that X_n converges in distribution to $Exponential(1)$.

Remind that CDF of $Exponential(\lambda)$ is :

$$F_X(x) = 1 - e^{-\lambda x}$$

Then

$$\lim_{n\to\infty} F_{X_n}(x) = \lim_{n\to\infty} 1 - \left(1 - \frac{1}{n}\right)^{nx}$$
$$= 1 - e^{-x}, \forall x > 0$$

Then $X_n \xrightarrow{d} X$, where $X \sim Exponential(1)$.

4.2.3 Integer-valued cases

Given sequence of random variables $X_1, X_2, X_3, \ldots, X_n$ and random variable X. Assume that X and X_n (for all n) are non-negative and integer-valued, i.e.,

$$R_{X_i} \subset 0, 1, 2, \ldots, i = 1, \ldots, n$$
$$R_X \subset 0, 1, 2, \ldots$$

where R_X means the set of possible value of X. Then $X_n \xrightarrow{d} X$ if and only if:

$$\lim_{n\to\infty} P_{X_n}(k) = P_X(k), \forall k \in R_X$$

4.2.3.1 Application

Given a sequence of random variables $X_1, X_2, X_3, \ldots, X_n$ such that:

$$X_n \sim Binomial(n, \frac{\lambda}{n}), \quad \text{for } n \in \mathbb{N}, n > \lambda$$

where $\lambda > 0$ is a constant. Show that X_n converges in distribution to $Poisson(\lambda)$.

$$
\begin{aligned}
\lim_{n \to \infty} P_{X_n}(k) &= \lim_{n \to \infty} \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} \\
&= \frac{n(n-1)\ldots(n-k+1)}{k!} \frac{\lambda^k}{n^k} \left(1 - \frac{\lambda}{n}\right)^{\frac{n}{\lambda} \frac{\lambda(n-k)}{n}} \\
&= \frac{\lambda^k}{k!} e^{-\lambda}
\end{aligned}
$$

4.2.4 Lemma

If :

- $X_n \xrightarrow{d} X$

- X_n are uniformly integrable (3.3).

- X is integrable.

Then $\lim_{n \to \infty} \mathbb{E}[X_n] = \mathbb{E}[X]$ (convergence in mean).

Proof

Let $f_K(x) = \begin{cases} x \text{ if } |x| \leq K \\ 0 \text{ otherwise} \end{cases}$, then f_K is a bounded and continuous function. By the equivalent definition of convergence in distribution 4.2.1, we have $\lim_{n \to \infty} \mathbb{E}[f_K(X_n)] \to \mathbb{E}[f_K(X)]$.

With triangle inequality :

$$
\begin{aligned}
&|\mathbb{E}[X_n] - \mathbb{E}[X]| \\
\leq &|\mathbb{E}[X_n] - \mathbb{E}[f_K(X_n)]| + |\mathbb{E}[f_K(X_n)] - \mathbb{E}[f_K(X)]| + |\mathbb{E}[f_K(X)] - \mathbb{E}[X]| \\
= &|\mathbb{E}[X_n . \mathbb{1}_{\{|X_n| > K\}}]| + |\mathbb{E}[f_K(X_n)] - \mathbb{E}[f_K(X)]| + |\mathbb{E}[X . \mathbb{1}_{\{|X| > K\}}]| \\
\leq &\mathbb{E}[|X_n| . \mathbb{1}_{\{|X_n| > K\}}] + |\mathbb{E}[f_K(X_n)] - \mathbb{E}[f_K(X)]| + \mathbb{E}[|X| . \mathbb{1}_{\{|X| > K\}}], \forall K
\end{aligned}
$$

Then

$$
\begin{aligned}
&\lim_{n \to \infty} |\mathbb{E}[X_n] - \mathbb{E}[X]| \\
\leq &\lim_{n \to \infty} \mathbb{E}[|X_n| . \mathbb{1}_{\{|X_n| > K\}}] + \lim_{n \to \infty} |\mathbb{E}[f_K(X_n)] - \mathbb{E}[f_K(X)]| + \lim_{n \to \infty} \mathbb{E}[|X| . \mathbb{1}_{\{|X| > K\}}] \\
\leq &\lim_{K \to \infty} \sup_{n \to \mathbb{N}} \mathbb{E}[|X_n| . \mathbb{1}_{\{|X_n| > K\}}] + 0 + \lim_{K \to \infty} \sup_{n \to \mathbb{N}} \mathbb{E}[|X| . \mathbb{1}_{\{|X| > K\}}] = 0
\end{aligned}
$$

This is because X_n are uniformly integrable then $\lim_{K\to\infty} \sup_{n\to\mathbb{N}} \mathbb{E}[|X_n|.\mathbb{1}_{\{|X_n|>K\}}]$ (sec 3.3) and X is integrable then $\lim_{K\to\infty} \sup_{n\to\mathbb{N}} \mathbb{E}[|X|.\mathbb{1}_{\{|X|>K\}}] = 0$ (lemma 3.3.5.1).

4.3   Convergence in probability

A sequence of random variables $X_1, X_2, X_3, \ldots, X_n$ converges in probability to a random variable X, shown by $X_n \xrightarrow{p} X$, if :

$$\lim_{n\to\infty} P(|X_n - X| \geq \epsilon) = 0, \quad \text{for all } \epsilon > 0.$$

or in more details :

$$\lim_{n\to\infty} P(\{\omega \colon |X_n(\omega) - X(\omega)| \geq \epsilon\}) = 0, \quad \text{for all } \epsilon > 0.$$

4.3.1   Law of large numbers

It states that if you repeat an experiment independently a large number of times and average the result, what you obtain should be close to the expected value. There are two main versions of the law of large numbers. They are called the **weak** (convergence in probability) and **strong** laws of the large numbers (Almost sure convergence, in sec 4.6.9).

Given i.i.d. random variables $X_1, X_2, \ldots, X_n$. The sample mean, denoted by $\bar{X}$, is defined as:

$$\bar{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

$$\mathbb{E}[\bar{X}] = \frac{\mathbb{E}[X_1] + \mathbb{E}[X_2] + \ldots + \mathbb{E}[X_n]}{n} = \mathbb{E}[X_i] = \mu$$

4.3.1.1   Weak version

In the weak version, we assume that X_i admit a second moment, it means:

$$Var[\bar{X}] = \frac{1}{n^2} \sum_{i=1}^{n} Var[X_i] = \frac{Var[X]}{n}$$

Then, $\bar{X} \xrightarrow{p} \mu$ for any $\epsilon > 0$:

$$\lim_{n\to\infty} P(|\bar{X} - \mu| \geq \epsilon) = 0$$

An simple proof can be done by using Chebyshev's inequality :

$$P(|\bar{X} - \mathbb{E}[\bar{X}]| \geq \epsilon) = P(|\bar{X} - \mu| \geq \epsilon) \leq \frac{Var[X]}{n\epsilon^2}$$

4.3.2 Example

Let $X_n \sim Exponential(n)$, show that $X_n \xrightarrow{p} 0$.

$$
\begin{aligned}
\lim_{n\to\infty} P(|X_n - 0| \geq \epsilon) &= \lim_{n\to\infty} P(X_n \geq \epsilon) \quad \text{since } P(X_n < 0) = 0 \\
&= \lim_{n\to\infty} 1 - (1 - e^{-n\epsilon}) \\
&= \lim_{n\to\infty} e^{-n\epsilon} \\
&= 0
\end{aligned}
$$

4.3.3 Relation between convergence in probability and in distribution

As we mentioned previously, convergence in probability is stronger than convergence in distribution. That is, if $X_n \xrightarrow{p} X$, then $X_n \xrightarrow{d} X$. The converse is not necessarily true. For example, let $X_1, X_2, X_3, \ldots, X_n$ be a sequence of i.i.d. $Bernoulli(\frac{1}{2})$. Let also $X \sim Bernoulli(\frac{1}{2})$ be independent from the X_i. Then, $X_n \xrightarrow{d} X$. However, X_n does not converge in probability to X, since $|X_n - X|$ is in fact also a $Bernoulli(\frac{1}{2})$ random variable and

$$
P(|X_n - X| \geq \epsilon) = \frac{1}{2} \neq 0, \text{for } 0 < \epsilon < 1.
$$

4.4 Convergence in probability implies convergence in distribution

We prove first the lemma

$$
P(Y \leq a) \leq P(X \leq a + \varepsilon) + P(|Y - X| > \varepsilon), \forall \varepsilon > 0
$$

By using this lemma, we have $\forall \varepsilon > 0$:

$$
P(X_n \leq a) \leq P(X \leq a + \varepsilon) + P(|X_n - X| > \varepsilon)
$$

and

$$
P(X \leq a - \varepsilon) \leq P(X_n \leq a) + P(|X_n - X| > \varepsilon)
$$

This leads to

$$
F_X(a - \varepsilon) \leq \lim_{n\to\infty} P(X_n \leq a) \leq F_X(a + \varepsilon), \, for all \varepsilon > 0
$$

Then we have Q.E.D.

4.5 Convergence in mean

Let $r \geq 1$ be a fixed number. A sequence of random variables $X_1, X_2, \ldots, X_n$ converges in the rth mean or in the L^r norm to a random variable X, shown by $X_n \xrightarrow{L_r} X$, if we have:

1. X is L^r integrable.

2. $X_1, X_2, \ldots, X_n$ are L^r integrable.

3.
$$\lim_{n \to \infty} \mathbb{E}[|X_n - X|^r] = 0.$$

- If $r = 1$, it is called by defaut convergence in mean or L^1 convergence.

- If $r = 2$, it is called the mean-square convergence, and it is shown by $X_n \xrightarrow{m.s.} X$

4.5.1 Example

Let $X_n \sim Uniform(0, \frac{1}{n})$. Show that $X_n \xrightarrow{L_r} 0$.

$$
\begin{aligned}
\lim_{n \to \infty} \mathbb{E}[|X_n - X|^r] &= \lim_{n \to \infty} \mathbb{E}[X_n^r] \\
&= \lim_{n \to \infty} \int_0^{\frac{1}{n}} x^r n \, dx \\
&= \lim_{n \to \infty} n \frac{1}{r+1} \left(\frac{1}{n}\right)^{r+1} \\
&= 0
\end{aligned}
$$

4.5.2 Relation between convergence in mean and in probability

As we mentioned previously, convergence in mean is stronger than convergence in probability. That is, if $X_n \xrightarrow{L_r} X$, then $X_n \xrightarrow{p} X$. The converse is not necessarily true. For example, consider a sequence $\{X_n, n = 1, 2, 3, \ldots\}$ such that:
$$
X_n = \begin{cases} n^2 & \text{with probability } \frac{1}{n} \\ 0 & \text{with probability } 1 - \frac{1}{n} \end{cases}
$$

We show that $X_n \xrightarrow{p} 0$, thus :

$$
\begin{aligned}
\lim_{n \to \infty} P(|X_n - 0| \geq \epsilon) &= \lim_{n \to \infty} P(X_n \geq \epsilon) \\
&= \lim_{n \to \infty} P(X_n = n^2) \\
&= \lim_{n \to \infty} \frac{1}{n} \\
&= 0
\end{aligned}
$$

But

$$\lim_{n\to\infty} \mathbb{E}[|X_n - 0|^r] = \lim_{n\to\infty} \mathbb{E}[X_n^r]$$

$$= \lim_{n\to\infty} n^{2r}\frac{1}{n} + 0^r\left(1 - \frac{1}{n}\right)$$

$$= \lim_{n\to\infty} n^{2r-1}$$

$$= \infty$$

where $r \geq 1$, then we do not have $X_n \xrightarrow{L_r} 0$.

4.5.3 Convergence in mean implies convergence in probability

The demonstration is trivial by using Markov's inequality (in sec 2.7.1)

$$P(|X_n - X| \geq \epsilon) \leq \frac{\mathbb{E}[|X_n - X|]}{\epsilon}, \forall n \in \mathbb{N}$$

We first fix ϵ and take $n \to \infty$, then $\mathbb{E}[|X_n - X|] \to 0$ or $P(|X_n - X| \geq \epsilon) \to 0$.

4.5.4 Lemma

$X_n \xrightarrow{L_1} X$ if and only if $X_n \xrightarrow{p} X$ and X_n is uniformly integrable (sec 3.3).
Proof

In the direct sense, we see that the convergence in mean implies the convergence in probability in sec 4.5.3. On one hand, from the definition of convergence in mean (sec 4.5), $\mathbb{E}[|X|] < +\infty$ and since $\lim_{n\to\infty} \mathbb{E}[|X_n - X|] \to 0$, with the triangle inequality, we have :

$$\sup_{n\in\mathbb{N}} \mathbb{E}[|X_n|] < \mathbb{E}[|X|] + \sup_{n\in\mathbb{N}} \mathbb{E}[|X_n - X|] < +\infty$$

On the other hand, always with the triangle inequality :

$$\sup_{n\in\mathbb{N}} \mathbb{E}[|X_n|.\mathbb{1}_A]$$

$$\leq \mathbb{E}[|X|.\mathbb{1}_A] + \sup_{n\in\mathbb{N}} \mathbb{E}[|X_n - X|.\mathbb{1}_A]$$

$$\leq KP(A) + \mathbb{E}[|X|.\mathbb{1}_{\{|X|>K\}}] + \max\left(\sup_{n\leq n_0} \mathbb{E}[|X_n - X|.\mathbb{1}_A], \sup_{n>n_0} \mathbb{E}[|X_n - X|.\mathbb{1}_A]\right)$$

$$= KP(A) + \mathbb{E}[|X|.\mathbb{1}_{\{|X|>K\}}] + \max\left(KP(A) + \sup_{n\leq n_0} \mathbb{E}[|X_n - X|.\mathbb{1}_{\{|X_n-X|>K\}}], \sup_{n>n_0} \mathbb{E}[|X_n - X|.\mathbb{1}_A]\right)$$

We have, given a $\varepsilon > 0$:

- X is integrable and X_n are integrable (by def of converge in mean or by $\sup_{n\in\mathbb{N}} \mathbb{E}[|X_n|] < +\infty$ that we've just proved). Then $E[|X - X_n|] < E[|X|] + E[|X_n|] < +\infty$ or $X - X_n$ are integrable.

101

- Since X and $X_n - X, \forall n \leq n_0$ are integrable, set $\{X, X_1 - X, \ldots, X_{n_0} - X\}$ has a finite number of elements, thne we can choose K_0 such that $\forall K > K_0$, $\mathbb{E}[|X|.\mathbb{1}_{\{|X|>K\}}] < \varepsilon$ and $\sup_{n \leq n_0} \mathbb{E}[|X_n - X|.\mathbb{1}_{\{|X_n-X|>K\}}] < \varepsilon$ (sec 3.3.6).

- We can choose also n_0 such that $\forall n > n_0, \mathbb{E}[|X_n - X|] < \varepsilon$, this infers $\mathbb{E}[|X_n - X|.\mathbb{1}_A] < \varepsilon$.

- Take $\delta = \frac{\varepsilon}{K_0}$, then if $P(A) < \delta$, then $K_0 P(A) < \varepsilon$.

Finally,

$$\sup_{n \in \mathbb{N}} \mathbb{E}[|X_n|.\mathbb{1}_A] \leq \epsilon + \epsilon + \max\left(\epsilon + \epsilon, \epsilon\right) = 4\epsilon, \text{ if } P(A) < \delta$$

or X_n are uniformly integrable.

Conversely, we show that if $X_n \xrightarrow{p} X$ and X_n is uniformly integrable then $X_n \xrightarrow{L_1} X$.

First, we show that X is integrable. Since $X_n \xrightarrow{p} X$ then there exists a subsequence (n_k) such that $X_{n_k} \xrightarrow{a.s} X$ (sec 4.6.1. By Fatou's lemma (sec 2.12), we have :

$$\mathbb{E}[|X|] = \mathbb{E}[\liminf_{k \to \infty} |X_{n_k}|] \leq \liminf_{k \to \infty} \mathbb{E}[|X_{n_k}|] \leq \sup_{n \in \mathbb{N}} \mathbb{E}[|X_n|] < +\infty$$

Second, by triangle inequality:

$$\mathbb{E}[|X_n - X|] \leq \mathbb{E}[|X_n| + |X|] = \mathbb{E}[|X_n|] + \mathbb{E}[|X|] < +\infty$$

Then we have

- $Y = |X_n| + |X|$ is integrable.

- $|X_n - X| \leq Y = |X_n| + |X|$.

By proposition 3.3.5, $(X_n - X)$ is uniformly integrable.

Finally,

$$\mathbb{E}[|X_n - X|]$$
$$= \mathbb{E}[|X_n - X|.\mathbb{1}_{|X_n-X|>\varepsilon}] + \mathbb{E}[|X_n - X|.\mathbb{1}_{|X_n-X|\leq\varepsilon}]$$
$$\leq \mathbb{E}[|X_n - X|.\mathbb{1}_{|X_n-X|>\varepsilon}] + \varepsilon, \quad \forall \varepsilon > 0$$

By two following statements :

- $X_n \xrightarrow{p} X$, which means $|X_n - X| \xrightarrow{p} 0$ or $\lim_{n \to \infty} P(|X_n - X| > \varepsilon) = 0$

- The absolutely continuous property in the original definition of uniformly integrable (sec 3.2)

Then $\lim_{n \to \infty} \mathbb{E}[|X_n - X|.\mathbb{1}_{|X_n-X|>\varepsilon}] < \varepsilon_1, \forall \varepsilon_1 > 0$, this implies :

$$\lim_{n \to \infty} \mathbb{E}[|X_n - X|] < \varepsilon + \varepsilon_1, \forall \varepsilon, \varepsilon_1 > 0$$

or $X_n \xrightarrow{L_1} X$.

4.6 Almost sure convergence

A sequence of random variables $X_1, X_2, X_3, \ldots, X_n$ converges almost surely to a random variable X, shown by $X_n \xrightarrow{a.s.} X$ if:

$$A = \{\omega \in \Omega : \lim_{n \to \infty} X_n(\omega) = X(\omega)\}$$

$$P(A) = 1$$

where Ω is the sample space, which is the set of all possible outcomes ω. $X_i(\omega)$ here is a mapping function from Ω to the set of real numbers.

Each time ω is fixed ($= \omega_0$), it means while solving for A, we check if

$$X_1(\omega_0), X_2(\omega_0), X_3(\omega_0), \ldots, X_n(\omega_0) \to X(\omega_0)$$

4.6.1 Properties

- If $X_n \xrightarrow{a.s} X$ then $X_n \xrightarrow{p} X$.

- If $X_n \xrightarrow{p} X$ then there exists a sub-sequence (n_k) such that $X_{n_k} \xrightarrow{a.s} X$.

- The dominated convergence theorem (sec 2.15.2) gives sufficient conditions for almost sure convergence to imply convergence in mean :

$$\left.\begin{array}{l} X_n \xrightarrow{a.s} X (\Rightarrow X_n \xrightarrow{p} X) \\ |X_n| < Y \\ E[|Y|] < +\infty \end{array}\right\} \Rightarrow X_n \xrightarrow{L_1} X$$

4.6.2 Example

Consider the following random experiment: A fair coin is tossed once. Here, the sample space has only two elements $S = \{H, T\}$. We define a sequence of random variables $X_1, X_2, X_3, \ldots$ on this sample space as follows:

$$X_n(s) = \begin{cases} \frac{n}{n+1} & \text{if } \omega = H \\ (-1)^n & \text{if } \omega = T \end{cases}$$

We define also :

$$X(\omega) = 1, \forall \omega \in \{S, T\}$$

Is $X_n \xrightarrow{a.s.} X$?

In the first step, we need to find A. Thus $A = \{H\}$.

Since $P(A) = \frac{1}{2}$. Then X_n is not almost sure convergence.

4.6.3 Example in a continuous case

Consider the sample space $\Omega = [0,1]$ with a probability measure that is uniform on this space, i.e.,

$$P([a,b]) = b - a, \forall a, b \text{ with } 0 \le a \le b \le 1.$$

Define the sequence $X_n, n = 1, 2, \ldots$ as follows:

$$X_n(\omega) = \begin{cases} 1 & \text{if } 0 \le \omega < \frac{n+1}{2n} \\ 0 & \text{otherwise} \end{cases}$$

Also, define the random variable X on this sample space as follows

$$X(\omega) = \begin{cases} 1 & \text{if } 0 \le \omega < \frac{1}{2} \\ 0 & \text{otherwise} \end{cases}$$

In the first step, we need to find A. Thus $A = [0,1] - \{\frac{1}{2}\}$.

However, we still have $P(A) = 1$. Then $X_n \xrightarrow{a.s.} X$.

4.6.4 Remarks

We note that in **the continous case** if there are single values (or infinite countable set) ω such that $X_n(\omega)$ do not converge, we still have the convergence (a.s.). The convergence is not true if there are intervals (even small) such that $\forall \omega \in (a,b), X_n(\omega)$ do not converge.

Conversely, in **discret case** (first example), if there is only one single values ω such that $X_n(\omega)$ do not converge, we do not have the a.s. convergence.

4.6.5 Relation between convergence in probability and almost sure convergence

Let's consider space sample $\Omega = [0,1]$, $X(\omega) = 0, \forall \omega \in \Omega$ and $\forall n \ge 1$:

$$X_n(\omega) = \begin{cases} 1 & \text{if } \omega \in \left[\frac{k}{2^m}, \frac{k+1}{2^m} \right] \\ 0 & \text{otherwise.} \end{cases}$$

where $m = \lfloor \log_2(n) \rfloor$ (floor part of $\log_2(n)$) and $k = n - 2^m$. Here are some $\left[\frac{k}{2^m}, \frac{k+1}{2^m} \right]$ for $n = 1, 2, \ldots$:

$$\left[\frac{0}{1}, \frac{1}{1} \right], \left[\frac{0}{2}, \frac{1}{2} \right], \left[\frac{1}{2}, \frac{2}{2} \right], \left[\frac{0}{4}, \frac{1}{4} \right], \left[\frac{1}{4}, \frac{2}{4} \right], \left[\frac{2}{4}, \frac{3}{4} \right], \left[\frac{3}{4}, \frac{4}{4} \right], \ldots$$

First we show that X_n converges in probability :

$$\begin{aligned}
\lim_{n \to \infty} P(|X_n - X| \ge \epsilon) &= \lim_{n \to \infty} P(|X_n - 0| \ge \epsilon) \\
&= \lim_{n \to \infty} P(X_n = 1) \\
&= \lim_{n \to \infty} \frac{1}{2^{\lfloor \log_2(n) \rfloor}} \\
&= 0
\end{aligned}$$

Second, we show that X_n does converge a.s. Even, we can say

$$P(\{\omega \in \Omega \colon \lim_{n \to \infty} X_n(\omega) = X(\omega)\}) = P(\{\omega \in \Omega \colon \lim_{n \to \infty} X_n(\omega) = 0\}) = 0$$

since for each $\omega \in \Omega$, $X_n(\omega)$ outputs infinitely 1 as $n \to \infty$. This is because ω will be found in a interval $\left[\frac{k}{2^m}, \frac{k+1}{2^m}\right]$.

Then convergence in probability do not imply convergence a.s.

Illustratively, we can see some first random variables of this example in figure 2.5.
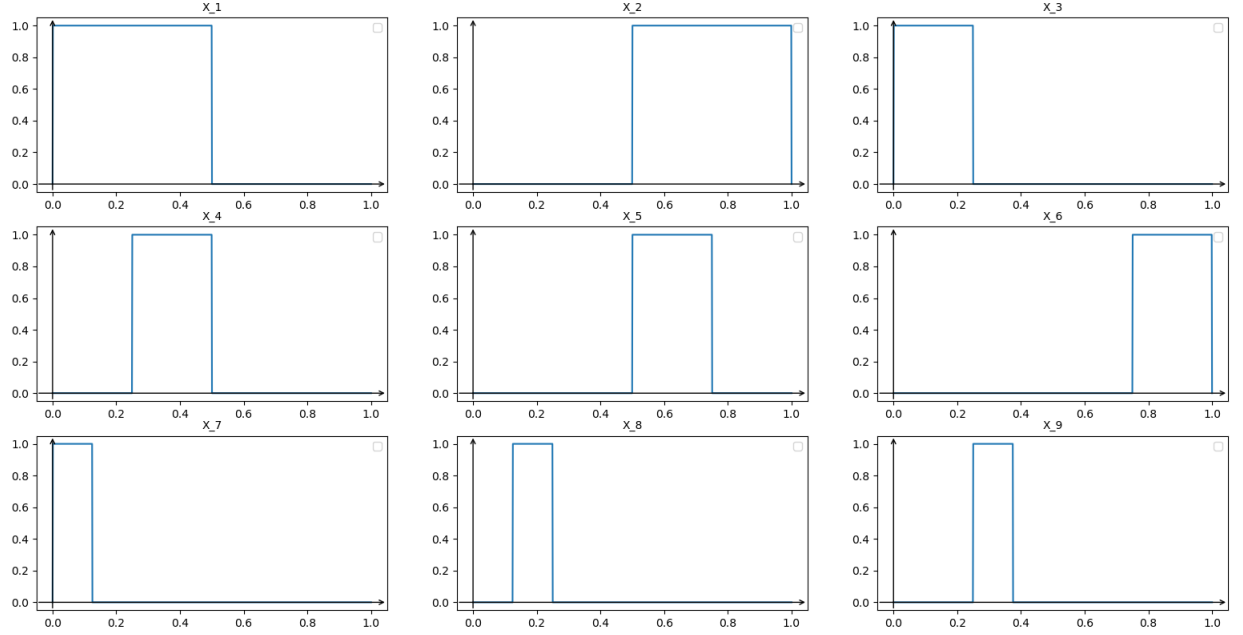
Figure 2.5: Distinction between convergence in probability and almost sure. Here $\Omega = [0, 1]$ and ω follows uniform distribution. ω is represented by the horizontal axe. We present only the first 9 random variable $X_1, \ldots, X_9$ but it can be continued. The mainpoint is that with a fixed ω, as n increases, we can refind 1.

4.6.5.1 Note

- For convergence in probability : ω is not necessarily fixed. For a given n, then we find set of ω and move $n \to \infty$:

$$\lim_{n \to \infty} P(\{\omega \colon |X_n(\omega) - X(\omega)| \geq \epsilon\}) = 0, \quad \text{for all } \epsilon > 0.$$

In figure 2.5, given n first and we find ω, then it is ok for convergence in probability.

- For convergence a.s. : ω is fixed first and then move $n \to \infty$:

$$P(\{\omega \in \Omega : \lim_{n \to \infty} X_n(\omega) = X(\omega)\}) = 1$$

In figure 2.5, we fix ω first then when $n \to \infty$, we refind always 1.

With a close attention, the position of $\lim_{n \to \infty}$ is different in two above definitions.

4.6.5.2 Convergence in mean and convergence almost surely

Moreover, we can show that X_n converges in mean :

$$
\begin{aligned}
\lim_{n \to \infty} \mathbb{E}[|X_n - X|^r] &= \lim_{n \to \infty} \mathbb{E}[|X_n - 0|^r] \\
&= \lim_{n \to \infty} \mathbb{E}[X_n^r] \\
&= \lim_{n \to \infty} 1^r \frac{1}{2^{\lfloor \log_2(n) \rfloor}} + 0^r \left(1 - \frac{1}{2^{\lfloor \log_2(n) \rfloor}}\right) \\
&= \lim_{n \to \infty} \frac{1}{2^{\lfloor \log_2(n) \rfloor}} \\
&= 0
\end{aligned}
$$

Then convergence in mean does not imply convergence a.s.. Note that convergence a.s. also so does not imply convergence in mean.

4.6.6 Useful method in sufficient condition

In some problems, proving almost sure convergence directly can be difficult. Thus, it is desirable to know some **sufficient** conditions for almost sure convergence. Here is a result that is sometimes useful when we would like to prove almost sure convergence.

Consider the sequence $X_1, X_2, X_3, \ldots$. If for all $\epsilon > 0$, we have :

$$\sum_{n=1}^{\infty} P(|X_n - X| > \epsilon) < \infty,$$

then $X_n \xrightarrow{a.s.} X$.

Example Consider the sequence $X_1, X_2, X_3, \ldots$ such that:

$$X_n = \begin{cases} -\frac{1}{n} & \text{with probability } \frac{1}{2} \\ \frac{1}{n} & \text{with probability } \frac{1}{2} \end{cases}$$

Let $X = 0$, then

$$\sum_{n=1}^{\infty} P(|X_n - X| > \epsilon) = \sum_{n=1}^{\infty} P(\frac{1}{n} > \epsilon)$$

$$= \sum_{n=1}^{\infty} P(\frac{1}{\epsilon} > n)$$

$$\leq \sum_{n=1}^{\lfloor \frac{1}{\epsilon} \rfloor} P(\frac{1}{\epsilon} > n)$$

$$= \lfloor \frac{1}{\epsilon} \rfloor$$

$$< \infty$$

Then $X_n \xrightarrow{a.s.} 0$

4.6.7 Useful method in sufficient and necessary

The above condition provides only a sufficient direction for almost sure convergence. In particular, if we obtain

$$\sum_{n=1}^{\infty} P(|X_n - X| > \epsilon) = \infty$$

then we still don't know whether the X_n converges to X almost surely or not. Here, we provide a condition that is both necessary and sufficient:

Consider the sequence $X_1, X_2, X_3, \ldots$, for any $\epsilon \to 0$, define the set :

$$A_m = \{\omega \mid |X_n(\omega) - X(\omega)| < \epsilon, \quad \text{for all } n \geq m\}.$$

Then $X_n \xrightarrow{a.s.} X$ if and only if for any $\epsilon > 0$, we have :

$$\lim_{n \to \infty} P(A_m) = 1$$

Example

Let $X_1, X_2, X_3, \ldots$ be independent random variables, where $X_n \sim Bernoulli(\frac{1}{n})$ for $n = 2, 3, \ldots$ The goal here is to check whether $X_n \xrightarrow{a.s.} 0$.

First,

$$\sum_{n=1}^{\infty} P(|X_n - 0| > \epsilon) = \sum_{n=1}^{\infty} P(X_n > \epsilon)$$

$$= \sum_{n=1}^{\infty} P(X_n = 1)$$

$$= \sum_{n=1}^{\infty} \frac{1}{n}$$

$$= \infty$$

107

Up to here, we are not sure that X_n converges. For any $\epsilon > 0$, we define A_m such that :
$$A_m = \{\omega \mid |X_n(\omega)| < \epsilon, \quad \text{for all } n \geq m\}.$$
Then
$$A_m = \{\omega \mid X_n(\omega) = 0, \quad \text{for all } n \geq m\}.$$

$$
\begin{aligned}
\lim_{m \to \infty} P(A_m) &= \lim_{m \to \infty} P(X_m = 0, X_{m+1} = 0, \dots, X_\infty = 0) \\
&\leq \lim_{m \to \infty} P(X_m = 0, X_{m+1} = 0, \dots, X_{2m} = 0) \\
&= \lim_{m \to \infty} P(X_m = 0) P(X_{m+1} = 0) \dots P(X_{2m} = 0) \quad \text{(independent)} \\
&= \lim_{m \to \infty} \frac{m-1}{m} \frac{m}{m+1} \dots \frac{2m-1}{2m} \\
&= \lim_{m \to \infty} \frac{m-1}{2m} \\
&= \frac{1}{2} < 1
\end{aligned}
$$

Therefore, we conclude that X_n do not converge (a.s.) to 0. Note that X_n converges in probability. This example shows again convergence in probability does not imply convergence in a.s.

4.6.8 Borel-Cantelli lemma

4.6.8.1 Statement

Let $E_1, E_2, \dots$ be a sequence of events (set of ω) in some probability space. The Borel–Cantelli lemma states that, if the sum of the probabilities of the events E_n is finite :

$$\sum_{n=1}^{\infty} P(E_n) < \infty$$

then the probability that infinitely many of them occur is 0, that is,

$$P\left(\limsup_{n \to \infty} E_n\right) = 0$$

where

$$\limsup_{n \to \infty} E_n = \bigcap_{n}^{\infty} \bigcup_{k=n}^{\infty} E_k$$

We call $\limsup_{n \to \infty} E_n$ set-theoretic limit supermum, where we perform union first then intersection. For example of set-theoretic limit supermum, let $E_n = \left(-\frac{1}{n}, 1 - \frac{1}{n}\right]$. Then

$$\limsup_{n \to \infty} E_n = \bigcap_{n} \bigcup_{j \geq n} \left(-\frac{1}{j}, 1 - \frac{1}{j}\right] = \bigcap_{n} \left(-\frac{1}{n}, 1\right) = [0, 1)$$

Example

Suppose X_n is a sequence of random variables with $P(X_n = 0) = \frac{1}{n^2}$. Let $E_n = \{\omega \colon X_n(\omega) = 0\}$. Then we have

$$\sum_{n=1}^{\infty} P(E_n) = \sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6} < \infty$$

where the sum of $\frac{1}{n^2}$ is Basel problem.

$$
\begin{aligned}
\limsup_{n\to\infty} E_n &= \bigcap_{n} \bigcup_{k=n}^{\infty} E_k \\
&= \{\omega \colon \bigcap_{n}^{\infty} (X_n(\omega) = 0 \text{ or } X_{n+1}(\omega) = 0 \text{ or } \ldots)\} \\
&= \{\omega \colon \bigcap_{n}^{\infty} \text{ At least one of } X_n, \ldots, X_\infty \text{ is } 0\} \\
&= \{\omega \colon \lim_{n\to\infty} \text{ At least one of } X_n, \ldots, X_\infty \text{ is } 0\}
\end{aligned}
$$

By Borel Cantelli lemma, then

$$P\left(\limsup_{n\to\infty} E_n\right) = 0$$

It means the probability that $X_n = 0$ occurs for infinite set of n is 0.

4.6.8.2 Converse result

Also called second Borel Cantelli lemma. If $\sum_{n=1}^{\infty} P(E_n) = \infty$ and the events $(E_n)_{n=1}^{\infty}$ are independent, then

$$P(\limsup_{n\to\infty} E_n) = 1$$

Example

Let's reuse the example in sec 4.6.7. Let $X_1, X_2, X_3, \ldots$ be independent random variables, where $X_n \sim Bernoulli(\frac{1}{n})$ for $n = 2, 3, \ldots$

First,

$$\sum_{n=1}^{\infty} P(X_n = 1) = \sum_{n=1}^{\infty} \frac{1}{n} = \infty$$

and X_n are independent, then we have

$$P(\omega \colon \lim_{n\to\infty} \text{ At least one of } X_n, \ldots, X_\infty \text{ is } 1) = 1$$

It means the probability that $X_n = 1$ occurs for infinite set of n is 1. Therefore, $P(\omega \colon \lim_{n\to\infty} X_n \to 0) = 0$, then X_n do not converge almost surely to $X = 0$.

4.6.9 Strong version of law of large numbers

This is a strong version, in constrast with the one (weak) in subsection 4.3.1.1.

Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables with a finite expected value $\mathbb{E}[X_i] = \mu < \infty$. Then

$$\bar{X} \xrightarrow{a.s.} \mu$$

Note that in this strong version, we do not need to assume that X_i has the second moment.

5 Signatures of a distribution

5.1 Moment generating function

Given a random variable X, its Moment generating function (MGF) is definded by :

$$M_X(t) = \mathbb{E}\left[e^{tX}\right]$$

MGF of several popular distributions :

- Normal $\mathcal{N}(\mu, \sigma^2)$: $\exp\left(t\mu + \frac{1}{2}\sigma^2 t^2\right)$

- Multivariate normal $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$: $\exp\left(t^T\boldsymbol{\mu} + \frac{1}{2}t^T\boldsymbol{\Sigma}t\right)$

5.2 Characteristic function

Given a random variable X, its Characteristic function (CF) is definded by :

$$\varphi_X(t) = \mathbb{E}\left[e^{itX}\right]$$

CF of several popular distributions :

- Normal $\mathcal{N}(\mu, \sigma^2)$: $\exp\left(it\mu - \frac{1}{2}\sigma^2 t^2\right)$

- Multivariate normal $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$: $\exp\left(it^T\boldsymbol{\mu} - \frac{1}{2}t^T\boldsymbol{\Sigma}t\right)$

5.3 Remarks

While both MGF and CF are useful for describing the properties of random variables, they have some important differences :

- The MGF of a real-valued distribution does not always exist, indeed MGF only exists for random variables with finite moments. While the CF always exists.

- Additionally, the MGF is typically used for calculating moments and derivatives, while the CF is used for calculating convolutions and characterizing the shape of the distribution.

5.4 Probability generating function

Given X is a *discrete* random variable that takes value in $\mathbb{N}$. Then the probability generating function of X, noted G_X is defined by :

$$G_X(t) = \mathbb{E}(t^X) = \sum_{n=0}^{+\infty} P(X = n)t^n$$

We will show that $G_X(t)$ converges with $t \in [-1, 1]$. Thus :

$$\sum_{n=0}^{+\infty} |P(X = n)t^n| = \sum_{n=0}^{+\infty} |P(X = n)||t^n|$$
$$\leq \sum_{n=0}^{+\infty} P(X = n)$$
$$= 1$$

Remind the absolute convergence implies the (oridnary) convergence. The converse is not true. In short, if $\sum_{n=0}^{+\infty} |a_n|$ converges then $\sum_{n=0}^{+\infty} a_n$ converges. We infer from here that $G_X(t) = \sum_{n=0}^{+\infty} P(X = n)t^n$ converges, since $\sum_{n=0}^{+\infty} |P(X = n)t^n|$ converges.

Here, we note that the radius of convergence is surely the disk $|t| \leq 1$ for all probability generating functions. Futhermore, in many examples, the radius of convergence is maybe larger than 1.

5.4.1 Properties

- $G_X = G_Y$ (with $t \in [-1, 1]$) if and only if X and Y are the same distribution.

- $G_{X+Y}(t) = G_X(t)G_Y(t)$ with X, Y are two independent random variables.

- $G'_X(1) = \mathbb{E}[X]$

- $E[X(X - 1)] = G''_X(1)$

- $Var[X] = G''_X(1) + G'_X(1) - G'_X(1)^2$

5.4.2 PGF of some usual disitrubtions

- $X \sim \mathcal{U}([1, N])$ then $G_X(t) = \frac{1}{N} \sum_{n=1}^{N} t^n$

- $X \sim \mathcal{B}(p)$ with $p \in [0, 1]$ then $G_X(t) = 1 - p + pt$

- $X \sim \mathcal{B}(n, p)$ with $p \in [0, 1]$ then $G_X(t) = (1 - p + pt)^n$

- $X \sim \mathcal{G}(p)$ with $p \in [0, 1]$ then with $t \in \left]-\frac{1}{1-p}, \frac{1}{1-p}\right[$, $G_X(t) = \frac{pt}{1-(1-p)t}$

- $X \sim \mathcal{P}(\lambda)$ with $\lambda > 0$, then $G_X(t) = e^{\lambda(t-1)}$

5.5 Expected value at any event

Given a probability space $(\Omega, \mathcal{F}, P)$ and a random variable X the expect value of X at an event $A \in \mathcal{F}$ is denoted by :

$$\mathbb{E}[X\mathbb{1}_A]$$

5.5.1 Lemma

Given two random variables X and Y on the same probability space $(\Omega, \mathcal{F}, P)$, then $X = Y$ almost surely if

$$\mathbb{E}[X\mathbb{1}_A] = \mathbb{E}[Y\mathbb{1}_A] \quad , \forall A \in \mathcal{F}$$

Proof

We consider three types of A for any $A \in \mathcal{F}$:

1. $A = \{X = Y\} = \{\omega | X(\omega) = Y(\omega)\}$

2. $A = \{X < Y\} = \{\omega | X(\omega) < Y(\omega)\}$

3. $A = \{X > Y\} = \{\omega | X(\omega) > Y(\omega)\}$

Then we need to show only for the second type and the third type that $X = Y$.

For the second type, consider the random variable $X - Y$, which the expected value is

$$\mathbb{E}[(X - Y)\mathbb{1}_A] = \mathbb{E}[X\mathbb{1}_A] - \mathbb{E}[Y\mathbb{1}_A] = 0$$

In this case, $(X - Y)\mathbb{1}_A$ is strictly negative, then $\mathbb{E}[(X - Y)\mathbb{1}_A] = 0$ can happen in two cases :

- There is no such as A as an interval.

- A contains only a single outcome.

For the third type, we do the same thing. Finally, we can say that $X = Y$ almost surely.

6 Multivariate normal distribution

6.1 The Method of Transformations

Starting example : Let X be a $Uniform(0, 1)$ random variable, and let $Y = e^X$. What is the pdf of Y ?

In general, one calculates the cdf and then taking its derivative. As $R_X = [0, 1]$, $R_Y = [1, e]$. The cdf of X is:

$$F_X(x) = x(0 \leq x \leq 1)$$

Then the cdf of Y is:

$$\begin{aligned}
F_Y(y) &= P(Y \le y) \\
&= P(e^X \le y) \\
&= P(X \le ln(y)) \\
&= F_X(ln(y)) \\
&= ln(y)
\end{aligned}$$

Finally, the pdf of Y:

$$\begin{aligned}
f_Y(y) &= F'_Y(y) \\
&= \frac{1}{y}(1 \le y \le e)
\end{aligned}$$

6.1.1 Theorem 1

Suppose that X is a continuous random variable and $g : \mathbb{R} \to \mathbb{R}$ is a strictly monotonic differentiable function. Let $Y = g(X)$. Then the PDF of Y is given by:

$$f_Y(y) = \begin{cases} \frac{f_X(x)}{|g'(x)|} = f_X(x)\left|\frac{dx}{dy}\right| = f_X(x)\left|\frac{dx}{dg(x)}\right| \\ 0 \text{ if } y = g(x) \text{ does not have a solution.} \end{cases}$$

Application for the above example $y = g(x) = e^x$:

$$\begin{aligned}
f_Y(y) &= f_X(x)\left|\frac{dx}{dg(x)}\right| \\
&= 1\left|\frac{dx}{e^x dx}\right| \\
&= \frac{1}{y}
\end{aligned}$$

6.1.2 Theorem 2

Consider a continuous random variable X with domain R_X, and let $Y = g(X)$. Suppose that we can partition R_X into a finite number n of intervals such that $g(x)$ is strictly monotone and differentiable on each partition. Then the PDF of Y is given by:

$$f_Y(y) = \sum_i^n \frac{f_X(x_i)}{|g'(x_i)|} = \sum_i^n f_X(x_i)\left|\frac{dx_i}{dy}\right|$$

where, $x_1, ..., x_n$ are real solution for $g(x) = y$

Application : Let X be a continuous random variable (standard normal distribution) with PDF :

$$f_X(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}$$

and let $Y = X^2$. Find $f_Y(y)$.

We note that the function $g(x) = x^2$ is strictly decreasing on the interval $(-\infty, 0)$, strictly increasing on the interval $(0, \infty)$, and differentiable on both intervals, $g'(x) = 2x$. Given an $y > 0$, we have two solutions for $y = g(x)$, in particular :

$$x_1 = \sqrt{y}, x_2 = -\sqrt{y}$$

Then :

$$\begin{aligned}
f_Y(y) &= \frac{f_X(x_1)}{|g'(x_1)|} + \frac{f_X(x_2)}{|g'(x_2)|} \\
&= \frac{f_X(\sqrt{y})}{2\sqrt{y}} + \frac{f_X(-\sqrt{y})}{2\sqrt{y}} \\
&= \frac{1}{2\sqrt{2\pi y}} e^{-\frac{1}{2}y} + \frac{1}{2\sqrt{2\pi y}} e^{-\frac{1}{2}y} \\
&= \frac{1}{\sqrt{2\pi y}} e^{-\frac{1}{2}y}
\end{aligned}$$

where $y > 0$

At y = 0:

$$f_Y(0) = \frac{1}{\sqrt{2\pi}}$$

Then f_Y is not continue at 0.

6.1.3 Theorem 3, for bivariate case

Let X and Y be two jointly continuous random variables. Let $(Z, W) = g(X, Y) = (g_1(X, Y), g_2(X, Y))$, where $g : \mathbb{R}^2 \to \mathbb{R}^2$ is a continuous one-to-one (invertible) function with continuous partial derivatives. Let $h = g^{-1}$, i.e., $(X, Y) = h(Z, W) = (h_1(Z, W), h_2(Z, W))$. Then Z and W are jointly continuous and their joint PDF, $f_{ZW}(z, w), \forall (z, w) \in R_{ZW}$ is given by :

$$f_{ZW}(z, w) = f_{XY}(h_1(z, w), h_2(z, w))|J|$$

where J is Jacobian matrix :

$$J = \det \begin{bmatrix} \frac{\partial h_1}{\partial z} & \frac{\partial h_1}{\partial w} \\ \frac{\partial h_2}{\partial z} & \frac{\partial h_2}{\partial w} \end{bmatrix}$$

Application : Let X and Y be two independent standard normal random variables. Let also

$$\begin{cases} Z = 2X - Y \\ W = -X + Y \end{cases}$$

Find $f_{ZW}(z, w)$

One needs to find h_1 and h_2 : From :

$$\begin{cases} X = Z + W \\ Y = Z + 2W \end{cases}$$

Then :

$$\begin{cases} h_1(z, w) = z + w \\ h_2(z, w) = z + 2w \end{cases}$$

$$J = \det \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix} = 1$$

Finally,

$$\begin{aligned} f_{ZW}(z, w) &= f_{XY}(h_1(z, w), h_2(z, w)) |J| \\ &= f_{XY}(z + w, 2w + z) 1 \\ &= \frac{1}{2\pi} e^{-\frac{(z+w)^2 + (z+2w)^2}{2}} \end{aligned}$$

6.2 Bivariate case

In this case, the number of dimension $k = 2$.
Let X and Y two random variables, then

$$\Sigma = \begin{bmatrix} Var(X) & Cov(X, Y) \\ Cov(X, Y) & Var(Y) \end{bmatrix} = \begin{bmatrix} \sigma_X^2 & \rho \sigma_X \sigma_Y \\ \rho \sigma_X \sigma_Y & \sigma_Y^2 \end{bmatrix}$$

where ρ is the correlation between X and Y. Remind the relation between covariance and correlation :

$$\rho(X, Y) = \frac{E[(X - E(X))(Y - E(Y))]}{\sqrt{Var(X)Var(Y)}} = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}}$$

Then $Cov(X, Y) = \rho \sigma_X \sigma_Y$. Remind that the density for multi-dimensions $\mathcal{N}(\mu, \Sigma)$:

$$\frac{1}{(2\pi)^{\frac{k}{2}} \det(\Sigma)^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}$$

Then the expression in the exponentiel :

$$\begin{aligned} &(x - \mu)^T \Sigma^{-1} (x - \mu) \\ &= \begin{bmatrix} x - \mu_X & y - \mu_Y \end{bmatrix} \frac{1}{\sigma_X^2 \sigma_Y^2 (1 - \rho^2)} \begin{bmatrix} \sigma_Y^2 & -\rho \sigma_X \sigma_Y \\ -\rho \sigma_X \sigma_Y & \sigma_X^2 \end{bmatrix} \begin{bmatrix} x - \mu_X \\ y - \mu_Y \end{bmatrix} \\ &= \frac{\sigma_Y^2 (x - \mu_X)^2 - 2\rho \sigma_X \sigma_Y (x - \mu_X)(y - \mu_Y) + \sigma_X^2 (y - \mu_Y)^2}{\sigma_X^2 \sigma_Y^2 (1 - \rho^2)} \\ &= \frac{1}{1 - \rho^2} \left((\frac{x - \mu_X}{\sigma_X})^2 - \frac{2\rho(x - \mu_X)(y - \mu_Y)}{\sigma_X \sigma_Y} + (\frac{y - \mu_Y}{\sigma_Y})^2 \right) \end{aligned}$$

Then we have the first defition of two random variables are jointly normal distribution, with the density function :

6.2.1 Definition 1

$$f_{XY}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}\exp\left(-\frac{1}{2(1-\rho^2)}\left((\frac{x-\mu_X}{\sigma_X})^2 - \frac{2\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + (\frac{y-\mu_Y}{\sigma_Y})^2)\right)\right)$$

In some literature, we can have the second definition:

6.2.2 Definition 2

Two random variables are jointly normal distribution if any linear combination of them is univariate normal distribution.

6.2.3 Decomposition of bivariate normal distribution

Let X and Y be two bivariate normal random variables, i.e., their joint PDF is given by difinition 1. Then there exist independent standard normal random variables Z_1 and Z_2 such that :

$$\begin{cases} X = \sigma_X Z_1 + \mu_X \\ Y = \sigma_Y(\rho Z_1 + \sqrt{1-\rho^2}Z_2) + \mu_y \end{cases}$$

where ρ is $Corr(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$

Proof:

Thus, the idea is that taking Z_1 by standard normalizing X and then taking Z_2 by removing correlated part with Z_1. Note that the denumerator $\sqrt{1-\rho^2}$ help Z_2 to be standard normal.

$$\begin{cases} Z_1 = \frac{X-\sigma_X}{\sigma_X} \\ Z_2 = \frac{1}{\sqrt{1-\rho^2}}\frac{Y-\sigma_Y}{\sigma_Y} - \frac{\rho}{\sqrt{1-\rho^2}}\frac{X-\sigma_X}{\sigma_X} \end{cases}$$

We need to prove :

- 1, Z_1 and Z_2 are standard normal distribution.

- 2, Z_1 and Z_2 are independent.

First, one calculates the density function of joint distribution $f_{Z_1 Z_2}(z_1, z_2)$, using the theorem in sec 6.1.3 :

$$\begin{cases} h_1(z_1, z_2) = \sigma_X z_1 + \mu_X \\ h_2(z_1, z_2) = \sigma_Y \rho z_1 + \sigma_Y\sqrt{1-\rho^2}z_2 + \mu_y \end{cases}$$

$$\det(J) = \det\begin{bmatrix} \sigma_X & 0 \\ \sigma_Y\rho Z_1 & \sigma_Y\sqrt{1-\rho^2} \end{bmatrix} = \sigma_X\sigma_Y\sqrt{1-\rho^2}$$

116

$$f_{Z_1 Z_2}(z_1, z_2) = f_{XY}(\sigma_X z_1 + \mu_X, \sigma_Y \rho z_1 + \sigma_Y \sqrt{1 - \rho^2} z_2 + \mu_y)|J|$$

$$= \frac{1}{2\pi} \exp\left(-\frac{1}{2(1-\rho^2)} \left(z_1^2 - 2\rho z_1(\rho z_1 + \sqrt{1-\rho^2} z_2) + (\rho z_1 + \sqrt{1-\rho^2} z_2)^2)\right)\right)$$

$$= \frac{1}{2\pi} \exp\left(-\frac{z_1^2 + z_2^2}{2}\right)$$

This mean that the two marginal probability density function $f_{Z_1}(z_1), f_{Z_2}(z_2)$ are standard normal distribution :

$$f_{Z_1}(z_1) = \int_{-\infty}^{+\infty} f_{z_1 z_2}(z_1, z_2) dz_2$$

$$= \frac{1}{2\pi} \exp\left(-\frac{z_1^2}{2}\right) \int_{-\infty}^{+\infty} \frac{1}{2\pi} \exp\left(-\frac{z_2^2}{2}\right) dz_2$$

$$= \frac{1}{2\pi} \exp\left(-\frac{z_1^2}{2}\right)$$

Z_1 and Z_2 are independent since :

$$f_{Z_1 Z_2}(z_1, z_2) = f_{Z_1}(z_1) f_{Z_2}(z_2)$$

In addition, recheck the covariance between Z_1 and Z_2 :

$$Cov(Z_1, Z_2) = E[Z_1 Z_2] - E[Z_1]E[Z_2]$$

$$= E\left[(\frac{X - \sigma_X}{\sigma_X})(-\frac{\rho}{\sqrt{1-\rho^2}} \frac{X - \sigma_X}{\sigma_X} + \frac{1}{\sqrt{1-\rho^2}} \frac{Y - \sigma_Y}{\sigma_Y})\right] - 0$$

$$= E\left[-\frac{\rho}{\sqrt{1-\rho^2}}(\frac{X - \sigma_X}{\sigma_X})^2 + \frac{1}{\sqrt{1-\rho^2}}(\frac{X - \sigma_X}{\sigma_X})(\frac{Y - \sigma_Y}{\sigma_Y})\right]$$

$$= -\frac{\rho}{\sqrt{1-\rho^2}} 1 + \frac{1}{\sqrt{1-\rho^2}} E\left[(\frac{X - \sigma_X}{\sigma_X})(\frac{Y - \sigma_Y}{\sigma_Y})\right]$$

$$= -\frac{\rho}{\sqrt{1-\rho^2}} + \frac{1}{\sqrt{1-\rho^2}} Cov\left[\frac{X - \sigma_X}{\sigma_X}, \frac{Y - \sigma_Y}{\sigma_Y}\right]$$

$$= -\frac{\rho}{\sqrt{1-\rho^2}} + \frac{1}{\sqrt{1-\rho^2}} Cov\left[\frac{X}{\sigma_X}, \frac{Y}{\sigma_Y}\right]$$

$$= -\frac{\rho}{\sqrt{1-\rho^2}} + \frac{1}{\sqrt{1-\rho^2}} \frac{Cov[X, Y]}{\sigma_X \sigma_Y}$$

$$= -\frac{\rho}{\sqrt{1-\rho^2}} + \frac{1}{\sqrt{1-\rho^2}} \rho$$

$$= 0$$

6.2.4 Proof for the equivalence between the two definitions

In the direction from definition 1 to definition 2 : Using the decomposition (sec 6.2.3).

In the direction from definition 2 to definition 1 : Given that $aX + bY$ is normal distribution, for all $a, b \in \mathbb{R}^2$. Then prove that X and Y form a bivariate normal.

Let ρ the correlation of X and Y and let Z_1, Z_2 be :

$$\begin{cases} Z_1 = \frac{X - \sigma_X}{\sigma_X} \\ Z_2 = -\frac{\rho}{\sqrt{1 - \rho^2}} \frac{X - \sigma_X}{\sigma_X} + \frac{1}{\sqrt{1 - \rho^2}} \frac{Y - \sigma_Y}{\sigma_Y} \end{cases}$$

Then Z_1, Z_2 are both standard normal distribution. Now, we use MGF (Moment generating function, sec 5.1) to show that Z_1, Z_2 follows bivariate (standard) normal distribution.

First, we remind the MGF for vector-valued random variables. Given a vector-valued random variable $\mathbf{A}$ (dimension $d > 1$), its MGF is :

$$M_{\mathbf{A}}(\mathbf{t}) = \mathbb{E}\left[e^{\langle \mathbf{t}, \mathbf{A} \rangle}\right]$$

where $\langle ., . \rangle$ is the dot product. We infer from this the MFG of $[Z_1, Z_2]$:

$$M_{Z_1, Z_2}(s, t) = \mathbb{E}[e^{sZ_1 + tZ_2}]$$

Since $aX + bY$ is normal distribution then $sZ_1 + tZ_2$ is also normal distribution, then $e^{sZ_1 + tZ_2} \sim Lognormal(\mu_L, \sigma_L^2)$ (log-normal distribution), where $\mu_L = \mathbb{E}[sZ_1 + tZ_2] = 0$ and $\sigma_L^2 = Var(sZ_1 + tZ_2) = s^2 + t^2$. Then the mean of this log-normal distribution is

$$\mathbb{E}[e^{sZ_1 + tZ_2}] = \mu_L + \frac{1}{2}\sigma_L^2 = \frac{1}{2}s^2 + \frac{1}{2}t^2$$

We found that $M_{Z_1, Z_2}(s, t)$ has the same MGF as standard bivariate normal distribution. Remind that the MGF of standard bivariate normal distribution $[A, B]$:

$$M_{A,B}(s,t) = \mathbb{E}\left(\exp([s \quad t]\begin{bmatrix} A \\ B \end{bmatrix})\right)$$

$$= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \frac{1}{2\pi}\exp\left(-\frac{1}{2}(a^2 + b^2)\right)\exp(sa + tb)da db$$

$$= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \frac{1}{2\pi}\exp\left(-\frac{1}{2}(a^2 + b^2) + sa + sb - \frac{1}{2}(s^2 + t^2) + \frac{1}{2}(s^2 + t^2)\right)da db$$

$$= \int_{-\infty}^{+\infty} \frac{1}{2\pi}\exp\left(-\frac{1}{2}(a - s)^2\right)da \int_{-\infty}^{+\infty}\exp\left(-\frac{1}{2}(b - t)^2\right)db \exp\left(\frac{1}{2}(s^2 + t^2)\right)$$

$$= \exp\left(\frac{1}{2}(s^2 + t^2)\right)$$

Then Z_1, Z_2 form standard bivariate normal distribution, so X, Y has density function as in definition 1.

6.2.5 Conditional distribution in bivariate normal

Let X and Y be jointly normal random variables with parameters $\mu_X, \sigma_X^2, \mu_Y, \sigma_Y^2$ and ρ. Find the conditional distribution $p(Y|X = x)$.

1, One way to solve this problem is by using the joint PDF formula :

$$f_{Y|X}(y|x) = \frac{f_{Y,X}(y, x)}{f_X(x)}$$

2, Or we can use the decomposition of bivariate normal distribution :

$$\begin{cases} X = \sigma_X Z_1 + \mu_X \\ Y = \sigma_Y(\rho Z_1 + \sqrt{1 - \rho^2} Z_2) + \mu_y \end{cases}$$

With Z_1, Z_2 are independent and standard normal distribution. Then :

$$f_{Y|X}(Y|X = x) = \sigma_Y(\rho \frac{x - \mu_X}{\sigma_X} + \sqrt{1 - \rho^2} Z_2) + \mu_y$$

is also a normal distribution and :

$$E(Y|X = x) = \sigma_Y \rho \frac{x - \mu_X}{\sigma_X} + \mu_Y$$

$$Var(Y|X = x) = (1 - \rho^2)\sigma_Y^2$$

6.2.6 Remarks

The decomposition X, Y into two independent random variable Z_1 and Z_2 has three folds advantage :

- First, it is more convenient and insightful than the joint PDF formula.

- Second, sometimes the construction using Z_1 and Z_2 can be used to solve problems regarding bivariate normal distributions.

- Third, this method gives us a way to generate samples from the bivariate normal distribution, by using independent random variables.

6.2.7 Two non-independent normal distributions can form a non joint normal distribution

The fact that two random variables X and Y both have a normal distribution **does not imply** that the pair (X, Y) has a joint normal distribution.
Here is an famous counter example :

- X is a standard normal distribution.

- Y is defined by $Y = X$ if $|X| > a$ and $Y = -X$ otherwise. As the sysmetric of X, then Y is also normal distribution.

Then $Z = X + Y = 2X$ if $|X| > a$ and $Z = X + Y = 0$ otherwise. One observes that Z **is not normal distribution**. The shape of Z is one peak at 0, then 0 extended to two sides util $2a$ and $-2a$. After that $f_Z(2x) = f_X(x)$, where $|x| > a$.

Remark : By controling a, we can set any value $Cov(X, Y)$.

By using the definition 2, then X, Y are not jointly normal distribution.

6.3 Multivariate case

For three or more random variables, the joint PDF, joint PMF, and joint CDF are defined in a similar way to what we have already seen for the case of two random variables (bivariate). Suppose that $X_1, X_2, ..., X_n$ are random variables. Then we have two equivalent definitions of Multivariate normal distribution :

6.3.1 Definition 1

If the joint distribution of $\mathbf{x} = x_1, x_2, ..., x_n$

$$f_{X_1, X_2, ..., X_n}(x_1, x_2, ..., x_n) = \frac{1}{\sqrt{(2\pi)^k |\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\mathbf{x} - \mu)^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \mu)\right)$$

6.3.2 Definition 2

If any linear combination of k components is univariate normal distribution.

6.3.3 Generating a multivariate normal distribution

Given $\mu, \boldsymbol{\Sigma}$, where $\boldsymbol{\Sigma}$ is Hermitian (sysmmetric in real case), positive-definite matrix. We want to generate $\mathcal{N}(\mu, \boldsymbol{\Sigma})$:

- Generating $\mathbf{z} = Z_1, Z_2, ..., Z_n$ standard normal distribution and independent.

- With given $\boldsymbol{\Sigma}$, we can apply the Cholesky decomposition $\boldsymbol{\Sigma} = LL^*$, where L is a lower triangular matrix.

- Then $L\mathbf{z} + \mu \sim \mathcal{N}(\mu, \boldsymbol{\Sigma})$

6.3.4 Decomposition of multivariate normal distribution

Given $\mathbf{x} = X_1, X_2, ..., X_n$ are multivariate normal distribution $\mathcal{N}(\mu, \boldsymbol{\Sigma})$, then there exists the decomposition of $\mathbf{x}$ into $\mathbf{z} = Z_1, Z_2, ..., Z_n$:

$$\mathbf{x} = L\mathbf{z} + \mu$$

Or
$$\mathbf{z} = L^{-1}(\mathbf{x} - \mu)$$

where $Z_1, Z_2, ..., Z_n$ are all standard normal and independent.

Check in case of bivariate normal :

$$\Sigma = \begin{bmatrix} Var(X) & Cov(X,Y) \\ Cov(X,Y) & Var(Y) \end{bmatrix} = \begin{bmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{bmatrix}$$

Then :

$$L = \begin{bmatrix} \sigma_X & 0 \\ \rho\sigma_Y & \sqrt{1-\rho^2}\sigma_Y \end{bmatrix}$$

6.3.5 Conditional distributions

Given that $\mathbf{x} = X_1, X_2, ..., X_n$ and $\mathbf{x} \sim \mathcal{N}(\mu, \Sigma)$. With a partition $\mathbf{x} = [\mathbf{x_1}, \mathbf{x_2}] = [[X_1, ..., X_q], [X_{q+1}, ..., X_n]]$, we have

$$p(\mathbf{x_2}|\mathbf{x_1}) \sim \mathcal{N}(\mu_{2|1}, \Sigma_{2|1})$$

To prove that $p(\mathbf{x_2}|\mathbf{x_1})$ is normal, as in bivariate case, we can use the decompostion of $\mathbf{x}$ into independent $\mathbf{z}$: $[\mathbf{x_1}, \mathbf{x_2}] = L[\mathbf{z_1}, \mathbf{z_2}] + [\mu_\mathbf{1}, \mu_\mathbf{2}]$, where $L = \begin{bmatrix} L_{11} & 0 \\ L_{21} & L_{22} \end{bmatrix}$. Suppose that we know $\mathbf{x_1}$, hence $\mathbf{z_1}$. So

$$\mathbf{x_2}|\mathbf{x_1} = ([L_{21} \quad L_{22}]\mathbf{z} + \mu_\mathbf{2})|(\mathbf{z_1}, \mu_1)$$

Since $\mathbf{z_2}$ contains components that are independent, hence any linear combination of them is univariate normal distribution, so $\mathbf{x_2}|\mathbf{x_1}$ is multivariate normal.

Expected value

$$\begin{aligned}
\mu_{2|1} &= E\left[[L_{21} \quad L_{22}]\mathbf{z} + \mu_\mathbf{2}\right] \\
&= [L_{21} \quad L_{22}]E[\mathbf{z}] + \mu_\mathbf{2} \\
&= L_{21}\mathbf{z_1} + \mu_\mathbf{2} \\
&= L_{21}L_{11}^{-1}(\mathbf{x_1} - \mu_\mathbf{1}) + \mu_\mathbf{2} \\
&= \Sigma_{21}\Sigma_{11}^{-1}(\mathbf{x_1} - \mu_\mathbf{1}) + \mu_\mathbf{2}
\end{aligned}$$

Note that :

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} = LL^* = \begin{bmatrix} L_{11}L_{11} & L_{11}L_{21} \\ L_{21}L_{11} & L_{21}L_{21} + L_{22}L_{22} \end{bmatrix}$$

Then

$$\Sigma_{21}\Sigma_{11}^{-1} = L_{21}L_{11}(L_{11}L_{11})^{-1} = L_{21}L_{11}^{-1}$$

Covariance

$$Cov(\mathbf{x}_2|\mathbf{x}_1, \mathbf{x}_2|\mathbf{x}_1) = Cov\left[\begin{bmatrix} L_{21} & L_{22} \end{bmatrix}\mathbf{z} + \mu_2, \begin{bmatrix} L_{21} & L_{22} \end{bmatrix}\mathbf{z} + \mu_2\right]$$
$$= Cov\left[\begin{bmatrix} L_{21} & L_{22} \end{bmatrix}\mathbf{z}, \begin{bmatrix} L_{21} & L_{22} \end{bmatrix}\mathbf{z}\right]$$
$$= Cov\left[L_{21}\mathbf{z}_1 + L_{22}\mathbf{z}_2, L_{21}\mathbf{z}_1 + L_{22}\mathbf{z}_2\right]$$
$$= Cov\left[L_{22}\mathbf{z}_2, L_{22}\mathbf{z}_2\right]$$
$$= L_{22}L_{22}Cov(\mathbf{z}_2, \mathbf{z}_2)$$
$$= L_{22}L_{22}I$$
$$= L_{22}L_{22}$$
$$= \mathbf{\Sigma}_{22} - \mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11}^{-1}\mathbf{\Sigma}_{12}$$

Note that :

$$\mathbf{\Sigma}_{22} - \mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11}^{-1}\mathbf{\Sigma}_{12} = L_{21}L_{21} + L_{22}L_{22} - L_{21}L_{11}(L_{11}L_{11})^{-1}L_{11}L_{21}$$
$$= L_{22}L_{22}$$

6.4   Elliptical distribution

An elliptical distribution is a type of multivariate probability distribution that generalizes the concept of the normal distribution, which means that elliptical distributions is a family that contains the normal distribution and its other related distributions.

It is called "elliptical" because the contours (lines) of constant probability density are elliptical in shape when plotted on a graph. Elliptical distributions are widely used in multivariate statistics and finance for modeling joint distributions of random variables.

6.4.1   Definition 1

Elliptical distributions are defined in terms of the *characteristic function* of probability theory. A random vector X on a Euclidean space has an elliptical distribution if its characteristic function ϕ satisfies the following functional equation (for every column-vector.

$$\phi_{X-\mu}(t) = \psi(t'\Sigma t)$$

for some location parameter μ , some nonnegative-definite matrix Σ and some scalar function (mapping n dimensions to 1 dimension) ψ.

6.4.2   Definition 2

Some elliptical distributions are alternatively defined in terms of their density functions. An elliptical distribution with a density function f has the form:

$$f(x) = k \cdot g((x - \mu)'\Sigma^{-1}(x - \mu))$$

where k is the normalizing constant, x is an n-dimensional random vector with median vector μ (which is also the mean vector if the latter exists), and Σ is the covariance matrix if the latter exists.

6.4.3 Examples

Examples include the following multivariate probability distributions:

- Multivariate normal distribution

- Multivariate t-distribution (see its density function and the definition 2).

7 Stochastic dominance

Stochastic dominance is a concept to compare and rank probability distributions or random variables.

7.1 Statewise dominance

The simplest case of stochastic dominance is statewise dominance (also known as state-by-state dominance), defined as follows: Random variable A is statewise dominant over random variable B if A gives at least as good a result in every state (every possible set of outcomes), and a strictly better result in at least one state.

7.1.0.1 Example

State	a	b	c	d
Player 1's reward	$1	$2	$2	$2
Player 2's reward	$1	$1	$2	$2

Table 2.1: Player 1 is statewise dominant player 2.

7.2 First-order stochastic dominance

Random variable A has first-order stochastic dominance over random variable B if for any value x (in codomain of A and B), A gives at least as high a probability of receiving at least x as does B, and for some x, A gives a higher probability of receiving at least x. In notation form :

$$P[A \geq x] \geq P[B \geq x], \quad \forall x,$$
$$P[A \geq x] > P[B \geq x], \quad \text{and for some } x$$

In terms of the cumulative distribution functions of the two random variables, A dominating B means that :

$$F_A(x) \leq F_B(x), \quad \forall x,$$
$$F_A(x) < F_B(x), \quad \text{and for some } x$$

since $F_A(x) = 1 - P[A \geq x]$ and $F_B(x) = 1 - P[B \geq x]$

Note that **Statewise dominance implies first-order stochastic dominance**.

7.2.0.1 Example

. Let continue with the above example, and now with an hypothesis that the probability of apprearance of state a, b, c, d are the same.

State	a	b	c	d
Player 1's reward	$1	$2	$2	$2
Player 2's reward	$1	$1	$2	$2
Player 3's reward	$3	$3	$1	$1

- Player 1 statewise dominates player 2, then we can say that Player 1 first-order stochastic dominantes player 2 (as remark above). Indeed, $P(P1 \geq 1) = (P2 \geq 1) = 1$ and $P(P1 \geq 2) = 3/4 > P(P2 \geq 2) = 1/2$.

- Player 3 first-order stochastic dominantes player 2 since $P(P3 \geq 1) = (P2 \geq 1) = 1, P(P3 \geq 2) = P(P2 \geq 2) = 1/2$ and $P(P3 \geq 3) = 1/2 > P(P2 \geq 3) = 0$

- Player 1 and Player 3 cannot be ordered relative to each other

7.3 Second-order

Given two distributions ρ and ν, in terms of cumulative distribution functions F_ρ and F_ν, ρ is second-order stochastically dominant over ν if and only if

$$\int_{-\infty}^{x} (F_\rho(t) - F_\nu(t)) \, dt \leq 0, \forall x$$

and with strict inequality at some x.

7.3.1 Equivalent definitions

There are two equivalent definition for the above one :

- ρ dominates ν in the second order if and only if $\mathbb{E}_{X \sim \rho}[u(X)] \geq \mathbb{E}_{X \sim \nu}[u(X)]$ for nondecreasing and concave functions $u(x)$.

- **Decomposition** : ρ second-order stochastically dominates ν if and only if there exist distributions y and z such that $x_\rho \overset{d}{=} (x_\nu + y + z)$, with $y \leq 0$ and $\mathbb{E}(z \mid x_\rho + y) = 0$.

These are analogous with the equivalent definitions of first-order stochastic dominance, given above. Also, **first-order stochastic dominance implies for second-order dominance**, since $F_\rho(t) \leq F_\nu(t)$ implies $\int_{-\infty}^{x} (F_\rho(t) - F_\nu(t))\, dt \leq 0$

7.3.2 Mean-preserving spread

Mean-preserving spread (MPS) is a change from one probability distribution A to another probability distribution B, where B is formed by spreading out one or more portions of A's probability density function or probability mass function while leaving the mean (the expected value) unchanged.

Mathematical definitions Let x_A and x_B be the random variables associated with gambles A and B. Then B is a mean-preserving spread of A if and only if $x_B \stackrel{d}{=} (x_A + z)$ for some random variable z having $E(z \mid x_A) = 0$ for all values of x_A. Here $\stackrel{d}{=}$ means "is equal in distribution to" (that is, "has the same distribution as").

Example This example shows MPS between A and B

Density	0.2	0.2	0.2	0.2	0.2
x_A	3	4	5	6	7
x_B	2	4	5	6	8

B is obtained by moving 20% of A from 3 to 2 and 20% of A from 7 to 8.

7.4 Third-order

Let F_ρ and F_ν be the cumulative distribution functions. ρ dominates ν in the third order if and only if both

$$\int_{-\infty}^{x} \left(\int_{-\infty}^{z} [F_\rho(t) - F_\nu(t)]\, dt \right) dz \leq 0, \forall x$$
$$\mathbb{E}_\rho(x) \geq \mathbb{E}_\nu(x)$$

7.4.1 Properties

- If ρ dominates ν in the third order, then $\mathbb{E}_\rho(\log(x)) \geq \mathbb{E}_\nu(\log(x))$. This means that the geometric mean of ρ must be greater than or equal to the geometric mean of ν.

- If ρ dominates ν in the third order, then $\min_\rho(x) \geq \min_\nu(x)$. This means that the left tail of F_ν must be thicker than the left tail of F_ρ.

- As above observations, second-order dominance implies third-order dominance (trivial demonstration).

Chapter 3

Statistics

1 Theorems

1.1 Central limit theorem

If $X_1, X_2, \ldots, X_n \in \mathbb{R}$ are random samples drawn from a population with overall mean μ and finite variance σ^2, and if $\bar{X}_n$ is the sample mean of the first n samples, then the limiting form of the distribution,

$$Z = \lim_{n \to \infty} \left(\frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \right)$$

is a standard normal distribution.

1.1.1 Multidimensional

Now consider $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n \in \mathbb{R}^k$ are random vectors drawn from a population with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, then :

$$\sqrt{n}(\bar{\mathbf{X}}_n - \boldsymbol{\mu}) \xrightarrow{n \to \infty} \mathcal{N}_k(0, \boldsymbol{\Sigma})$$

1.2 Cochran's theorem

Simplified version. Let $\mathbf{X} \in \mathbb{R}^n$ be a standard normal random vector, sampled from $\mathcal{N}(0_{\mathbb{R}^n}, I_n)$ and F a subspace (that has dimension d) of $\mathbb{R}^n$. $F^\perp$ is its coressponding orthogonal subspace. $P_F, P_{F^\perp}$ are matrix related to *orthogonal projections* on $F, F^\perp$; then :

- Vectors $P_F \mathbf{X}$ and $P_{F^\perp} \mathbf{X}$ are independent, which means each component in $P_F \mathbf{X}$ is independent with each component in $P_{F^\perp} \mathbf{X}$.

- $P_F \mathbf{X} \sim \mathcal{N}(0_{\mathbb{R}^n}, P_F)$ and $P_{F^\perp} \mathbf{X} \sim \mathcal{N}(0_{\mathbb{R}^n}, P_{F^\perp})$.

- $\sum_{i=1}^n P_F \mathbf{X}[i]^2 \sim \chi^2(d)$ and $\sum_{i=1}^n P_{F^\perp} \mathbf{X}[i]^2 \sim \chi^2(n - d)$

1.2.1 Example

Let's F be a subspace represented by the plan $x + y + z = 0$ with normalized normale vector $N = \begin{bmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \end{bmatrix}^T$. Then the matrix P_F to **project orthogonally** any point of $\mathbb{R}^3$ into F is :

$$P_F = I - NN^T = \frac{1}{3} \begin{bmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{bmatrix}$$

and $dim(P_F) = 2$. The matrix $P_{F^\perp}$ to **project orthogonally** any point of $\mathbb{R}^3$ into $F^\perp$ is

$$P_{F^\perp} = NN^T = \frac{1}{3} \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

and $dim(P_{F^\perp}) = 1$.

Then given a vector $X \sim \mathcal{N}(0_{\mathbb{R}^3}, I_3)$; vectors $P_F \mathbf{X}$ and $P_{F^\perp} \mathbf{X}$ are independent.

1.2.2 Note 1 : Linear transform of random variable vector

If $\mathbf{X} \sim \mathcal{N}(0_{\mathbb{R}^n}, \Sigma)$, then $P\mathbf{X} \sim \mathcal{N}(0_{\mathbb{R}^n}, P\Sigma P^T)$, where P is a matrix to represent a linear transform. This is because :

$$Cov(PX) = E[(PX - P0_{\mathbb{R}^n})(PX - P0_{\mathbb{R}^n})^T] = P\Sigma P^T$$

Then we can demonstrate the second statement in Cochran's theorem :

$$P_F \mathbf{X} \sim \mathcal{N}(0_{\mathbb{R}^n}, P_F I_n P_F^T) = \mathcal{N}(0_{\mathbb{R}^n}, P_F),$$

since P_F is sysmetric ($P_F = P_F^T$) and P_F is idempotent ($P_F P_F = P_F$). To prove P_F is sysmetric, we can rely on :

$$P_F = A(A^T A)^{-1} A^T$$

where A is the matrix whose columns form an orthonormal basis for the subspace F.

1.2.3 Note 2 : Concatenation of two random variable vectors

If $\mathbf{X} \sim \mathcal{N}(0_{\mathbb{R}^n}, \Sigma_X)$ and $\mathbf{Y} \sim \mathcal{N}(0_{\mathbb{R}^m}, \Sigma_Y)$, then :

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim \mathcal{N}(0_{\mathbb{R}^{m+n}}, \begin{bmatrix} \Sigma_X \\ \Sigma_Y \end{bmatrix} \begin{bmatrix} \Sigma_X^T & \Sigma_Y^T \end{bmatrix})$$

Then we can demonstrate the first statement in Cochran's theorem :

$$\begin{bmatrix} P_F X \\ P_{F^\perp} X \end{bmatrix} \sim \mathcal{N}(0_{\mathbb{R}^{2n}}, \begin{bmatrix} P_F \\ P_{F^\perp} \end{bmatrix} \begin{bmatrix} P_F^T & P_{F^\perp}^T \end{bmatrix})$$

$$\sim \mathcal{N}(0_{\mathbb{R}^{2n}}, \begin{bmatrix} P_F & 0 \\ 0 & P_{F^\perp} \end{bmatrix})$$

Then $P_F X$ and $P_{F^\perp} X$ are independent.

1.2.4 Note 3 : Why orthogonal projection ?

The orthogonal projection P_F and $P_{F\perp}$ help for the second statement and the third statement in Cochran's theorem, notably the fact that P_F and $P_{F\perp}$ are idempotent.

Note that, if P_F and $P_{F\perp}$ are not orthogonal projection to F and $F^\perp$ respectly, e.g. :

$$P_F = \begin{bmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{bmatrix}$$

$$P_{F\perp} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

, the first statement in Cochran's theorem is still true.

1.3 Basu's theorem

Before diving into this theorem, we discover some necessary notions.

1.3.1 Statistic

Sample statistic is any quantity computed from values in a sample, e.g.

- Sample mean

- Sample median

- Sample variance

- . . .

1.3.2 Completeness

Completeness is a property of a statistic between a *parameterized model* and *set of observations*.

1.3.2.1 Definiton

Consider :

- Random variable X whose distribution is parametrized by θ.

- Parameter space Θ ($\theta \in \Theta$).

- A set of n observations $X_1, \ldots, X_n$ of the random variable X.

- A statistic T from the set of observation. Note T_θ statistic that has distribution parametrized by θ.

Then the statistic T is said to be complete for the distribution of X if for all function g :

$$\text{If } E_T[g(T_\theta)] = 0, \forall \theta \in \Theta \text{ implies } P_T(g(T) = 0) = 1$$

1.3.2.2 Example : Binomial model

Let $(X_i)_{i=1,\dots,n}$ be n observations sampled from Bernoulli distribution with parameter p. Let's $T = \sum_{i=1}^n X_i$ then T is a statistic of the set of observations. Since X_i is sampled from Bernoulli distribution then T follows the binomial distribution : $T \sim B(n,p)$.

We show that T is a complete statistic, with a parameter space $(0,1)$.

$$E_T[g(T_p)] = \sum_{t=0}^n g(t) P(T = t)$$

$$= \sum_{t=0}^n g(t) \binom{n}{t} p^t (1-p)^{n-t}$$

$$= (1-p)^n \sum_{t=0}^n g(t) \binom{n}{t} \left(\frac{p}{1-p}\right)^t$$

since $p \in (0,1)$ then if $E_T[g(T_p)] = 0, \forall p$, then :

$$\sum_{t=0}^n g(t) \binom{n}{t} r^t = 0, \forall r$$

where $r = \frac{p}{1-p}, r \in (0, +\infty)$. Note that the above expression is a polynomial in r since t is integer, this infers that $g(t) = 0, \forall t$ or $P_T(g(T) = 0) = 1$.

1.3.2.3 Continuation of above example

- If $\Theta = \{0.5\}$ and $n = 1$, then $E_T[g(T_p)] = 0$ implies :

$$g(0) + g(1) = 0$$

  All function g such that $g(0) = \alpha$ and $g(1) = -\alpha$ can be solution for the above equation. As α can be different to 0, then $P_T(g(T) = 0) \neq 1$ or T is not complete.

- If $\Theta = \{0.25, 0.75\}$, then $E_T[g(T_p)] = 0, p \in \Theta$ implies :

$$\begin{cases} g(0) + \frac{1}{3}g(1) = 0 \\ g(0) + 3g(1) = 0 \end{cases}$$

  Then $g(0) = 0$ and $g(1) = 0$ or $P_T(g(T) = 0) = 1$. T is now complete.

- If Θ contains at least $n + 1$ distinct values, then $n + 1$ equations. Note that $t = 0, \ldots, n$ makes $n + 1$ unknowns. The coefficient vectors given by $\left[1, \left(\frac{p}{1-p} \right), \left(\frac{p}{1-p} \right)^2, \ldots, \left(\frac{p}{1-p} \right)^n \right], \forall p \in \Theta$ are independent since $\frac{p}{1-p}$ is monotone (ascending). In this case T is complete.

1.3.2.4 Example of not complete statistic

Let $(X_i)_{i=1,\ldots,n}$ be n observations sampled from $\mathcal{N}(\theta, 1)$. Let's $T = (X_1, X_2)$ is a statistic and let's take $g(T) = X_1 - X_2$. Then $E_T[g(T_p)] = E[X_1] - E[X_2] = \theta - \theta = 0, \forall \theta$. However, this do not implies that $X_1 = X_2$.

1.3.2.5 Analogy to vector space

There is an analogy between statistic complete and vector space.

Suppose that we have a vector space of dimension n. Then a set of vectors $v_1, \ldots, v_n$ is complete if they span the whole space. That is, any v can be written as a linear combination $v = \sum_i a_i v_i$ of these vectors. The latter is equivalent to the fact that if w is orthogonal to all v_i, then $w = 0$.

Now if we consider $P(T = t)$ as a vector, then these vectors (t varies) are complete if $g(t)$ is orthogonal to all of them.

1.3.3 Sufficient statistic

Two definitions :

- A statistic $t = T(X)$ is sufficient for underlying parameter θ if the conditional probability distribution of the data X, given the statistic $t = T(X)$, does not depend on the parameter θ:

$$\Pr(X = x | T(X) = t, \theta) = \Pr(X = x | T(X) = t)$$

- A statistic T is sufficient if any two prior distributions will yield different distributions on T.

1.3.4 Note

Consider the map $f : p_\theta \mapsto p_{T|\theta}$, where :

- p_θ : space of distributions (model) parameterized by θ.

- $p_{T|\theta}$: space of distribution of statistic T. Since T is obtained from sampled random variables that have distribution in p_θ.

The statistic T is said :

- complete when f is surjective

- sufficient when f is injective.

1.3.5 Ancillary statistic

An ancillary statistic is a statistic whose distribution does not depend on the parameters of the model.

1.3.5.1 Example

Suppose $X_1, \ldots, X_n$ are independently sampled from $\mathcal{N}(\mu, 1)$. Then the following statistic is ancillary:

- $T = max(X_1, \ldots, X_n) - min(X_1, \ldots, X_n)$

- $T = \frac{\sum (X_i - \overline{X})^2}{n}$ estimation of variance, since the true variance is 1.

1.3.6 Basu's theorem

Let $(P_\theta; \theta \in \Theta)$ be a family of distributions on a measurable space $(X, \mathcal{A})$ and statistics T, A map from $(X, \mathcal{A})$ to some measurable space $(Y, \mathcal{B})$. If :

- If T is a boundedly complete sufficient statistic for θ.

- A is ancillary to θ.

Then conditional on θ, T is independent of $A : T \perp\!\!\!\perp A | \theta$.

1.3.6.1 Example : Independence of sample mean and sample variance of a normal distribution

Let $(X_i)_{i=1,\ldots,n}$ be iid sampled from $\mathcal{N}(\mu, \sigma^2)$. Then with respect to the parameter μ, one can show that :

- Sample mean $\widehat{\mu} = \dfrac{\sum X_i}{n}$ a is complete and sufficient statistic.

- Sample variance $\widehat{\sigma}^2 = \dfrac{\sum \left(X_i - \bar{X} \right)^2}{n - 1}$ is an ancillary statistic with respect to μ.

Therefore, from Basu's theorem it follows that these statistics are independent, conditional on σ^2.

2 Statistical inference : Classical method

2.1 Random sample

The collection of random variables $X_1, X_2, X_3, \ldots, X_n$ is said to be a random sample of size n if they are independent and identically distributed (i.i.d.), i.e. :

- $X_1, X_2, X_3, \ldots, X_n$ are independent.

- $F_{X_1}(x) = F_{X_2}(x) = ... = F_{X_n}(x)$. The equality of cumulative distribution is equivalent to the equality of density distribution.

2.1.1 Sampling with replacement

If the population is large, then the probability of choosing one person twice is extremely low, and it can be shown that the results obtained from sampling with replacement are very close to the results obtained using sampling without replacement. The big advantage of sampling with replacement is that X_i will be independent and this makes the analysis much simpler.

2.1.2 Bootstrap

Bootstrap data is a statistical technique used to estimate the sampling distribution of a statistic by resampling with replacement from a single data sample. In other words, bootstrap data involves creating multiple "bootstrapped" samples from a single dataset, with each bootstrapped sample being the same size as the original dataset.

$$(5, 4, 2, 1) \rightarrow (5, 5, 2, 2)$$
$$\rightarrow (5, 1, 1, 1)$$
$$\rightarrow ...$$

Suppose we have a dataset of 50 observations on a variable of interest. We want to estimate the mean of the population from which these observations were drawn, but we are uncertain about the sampling distribution of the mean.

To estimate the sampling distribution of the mean, we can use bootstrap data as follows:

Randomly sample 50 observations with replacement from the original dataset to create a bootstrap sample. This bootstrap sample may contain some observations multiple times and may omit some observations altogether.

Calculate the mean of the bootstrap sample.

Repeat steps 1 and 2 a large number of times (e.g., 1000 times) to create a distribution of bootstrapped means.

Use the bootstrapped mean distribution to estimate the standard error of the mean and construct a confidence interval for the population mean.

2.2 Order Statistics

Given a random sample, we might be interested in quantities such as the largest, the smallest, or the middle value in the sample. Thus, we often order the observed data from the smallest to the largest. We call the resulting ordered random variables *order statistics*. More specifically, let $X_1, X_2, X_3, \ldots, X_n$ be a random sample from a *continuous* distribution with CDF $F_X(x)$ and PDF

$f_X(x)$. Let us order X_i's from the smallest to the largest and denote the resulting sequence of random variables as:

$$X_{(1)}, X_{(2)}, \ldots, X_{(n)}.$$

Thus, we have:

$$X_{(1)} = \min(X_1, X_2, \ldots, X_n)$$

and

$$X_{(n)} = \max(X_1, X_2, \ldots, X_n)$$

We call $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ the order statistics. Then the CDF and PDF of $X_{(i)}$ are given by :

$$
\begin{aligned}
P(X_{(i)} = x) &= P(X = x)P(x \text{ is } i^{th} \text{ smallest}) \\
\Leftrightarrow f_{X_{(i)}}(x)dx &= f_X(x)dx P(x \text{ is } i^{th} \text{ smallest}) \\
\Leftrightarrow f_{X_{(i)}}(x) &= f_X(x)P(x \text{ is } i^{th} \text{ smallest}) \\
\Leftrightarrow f_{X_{(i)}}(x) &= f_X(x) \left[F_X(x)\right]^{i-1} \left[1 - F_X(x)\right]^{n-i} n\binom{n-1}{i-1}
\end{aligned}
$$

$$
\begin{aligned}
F_{X_{(i)}}(x) &= \int f_{X_{(i)}}(x)dx \\
&= n\binom{n-1}{i-1} \int f_X(x) \left[F_X(x)\right]^{i-1} \left[1 - F_X(x)\right]^{n-i} dx \\
&= n\binom{n-1}{i-1}\frac{1}{i} \int \left[1 - F_X(x)\right]^{n-i} d\left[F_X(x)\right]^{i} \\
&= \binom{n}{i} \left[\left[1 - F_X(x)\right]^{n-i} \left[F_X(x)\right]^{i} \Big|_0^x - \int \left[F_X(x)\right]^{i} d\left[1 - F_X(x)\right]^{n-i} \right] \\
&= \binom{n}{i} \left[1 - F_X(x)\right]^{n-i} \left[F_X(x)\right]^{i} + \binom{n}{i}(n-i) \int \left[F_X(x)\right]^{i} \left[1 - F_X(x)\right]^{n-i-1} f_X(x)dx \\
&= \binom{n}{i} \left[1 - F_X(x)\right]^{n-i} \left[F_X(x)\right]^{i} + n\binom{n-1}{i} \int f_X(x) \left[F_X(x)\right]^{i} \left[1 - F_X(x)\right]^{n-i-1} dx \\
&= \binom{n}{i} \left[1 - F_X(x)\right]^{n-i} \left[F_X(x)\right]^{i} + f_{X_{(i+1)}}(x) \\
&= \sum_{k=i}^{n} \binom{n}{k} \left[1 - F_X(x)\right]^{n-k} \left[F_X(x)\right]^{k}
\end{aligned}
$$

Also, the joint PDF of $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ is given by

$$f_{X_{(1)}, X_{(2)}, \ldots, X_{(n)}}(x_1, x_2, \ldots, x_n) = n! f_X(x_1) f_X(x_2) \ldots f_X(x_n)$$

where $x_1 \leq x_2 \leq \ldots \leq x_n$.

2.3 Estimator's bias

Let $\hat{\Theta} = h(X_1, X_2, \ldots, X_n)$ be a point estimator for θ (true value). The **bias** of point estimator $\hat{\Theta}$ is defined by:

$$B(\hat{\Theta}) = \mathbb{E}(\hat{\Theta}) - \theta$$

In case that $B(\hat{\Theta}) = 0$, it means that $\hat{\Theta}$ is an unbiased estimator.

Example. Given X_i are iid. sample mean is unbiased and variance is biased :

$$
\begin{aligned}
B(\hat{\Theta}) = B(\bar{X}) &= B(\frac{X_1 + \ldots + X_n}{n}) \\
&= \mathbb{E}(\frac{X_1 + \ldots + X_n}{n}) - \mu \\
&= \frac{n\mathbb{E}(X_i)}{n} - \mu \\
&= 0
\end{aligned}
$$

$$
\begin{aligned}
B(\hat{\Theta}) = B(\hat{\sigma}^2) &= B\left(\frac{1}{n}\sum_{k=1}^{n}(X_k - \bar{X})^2\right) \\
&= B\left(\frac{1}{n}\sum_{k=1}^{n}(X_k^2 - n\bar{X}^2)\right) \\
&= \mathbb{E}\left(\frac{1}{n}\sum_{k=1}^{n}(X_k^2 - n\bar{X}^2)\right) - \sigma^2 \\
&= \frac{1}{n}\sum_{k=1}^{n}\mathbb{E}(X_k^2) - \mathbb{E}(\bar{X}^2) - \sigma^2 \\
&= \frac{1}{n}\sum_{k=1}^{n}\left(Var(X_k) + \mathbb{E}(X_k)^2\right) - Var(\bar{X}) - \mathbb{E}(\bar{X})^2 - \sigma^2 \\
&= \sigma^2 + \mu^2 - \frac{\sigma^2}{n} - \mu^2 - \sigma^2 \\
&= -\frac{\sigma^2}{n}
\end{aligned}
$$

It is worth to note that $\hat{\sigma}^2 = \frac{1}{n-1}\sum_{k=1}^{n}(X_k - \bar{X})^2$ is an unbiased estimator for variance.

2.4 Mean squared error of estimator

This quantity is used to evaluate the performance of an estimator. The mean squared error (MSE) of a point estimator $\hat{\Theta}$, shown by $MSE(\hat{\Theta})$, is defined as:

$$MSE(\hat{\Theta}) = \mathbb{E}[(\hat{\Theta} - \theta)^2]$$

Example What is the better estimator for mean between X_1 and $\bar{X}$?

$$MSE(X_1) = \mathbb{E}[(X_1 - \mu)^2]$$
$$= Var(X_1)$$
$$= \sigma^2$$

$$MSE(\bar{X}) = \mathbb{E}[(\bar{X} - \mu)^2]$$
$$= Var(\bar{X} - \mu) + \mathbb{E}[(\bar{X} - \mu)]^2$$
$$= \frac{\sigma^2}{n} + (0 - 0)^2$$
$$= \frac{\sigma^2}{n}$$

Hence, we take the estimator $\bar{X}$ since it gives better error.

Relation between MSE and Bias :

$$MSE(\hat{\Theta}) = \mathbb{E}[(\hat{\Theta} - \theta)^2]$$
$$= Var(\hat{\Theta} - \theta) + \mathbb{E}[(\hat{\Theta} - \theta)]^2$$
$$= Var(\hat{\Theta}) + B(\hat{\Theta})^2$$

2.5   Consistency

Loosely speaking, we say that an estimator is consistent if as the sample size n gets larger, $\hat{\Theta}$ converges (in probability, sec 4.3) to the real value of θ. More precisely, we have the following definition:

Let $\hat{\Theta}_1, \hat{\Theta}_2, \ldots, \hat{\Theta}_n$ be a sequence of point estimators of θ. We say that $\hat{\Theta}^n$ is a consistent estimator of θ, if

$$\lim_{n \to \infty} P(|\hat{\Theta}_n - \theta| \geq \epsilon) = 0, \quad \text{for all } \epsilon > 0.$$

Note that if we have $\lim_{n \to \infty} MSE(\hat{\Theta}_n) = 0$, then $\hat{\Theta}_n$ is a consistent estimator. We note that :

$$P(|\hat{\Theta}_n - \theta| \geq \epsilon) = P((\hat{\Theta}_n - \theta)^2 \geq \epsilon^2)$$
$$\leq \frac{\mathbb{E}[(\hat{\Theta}_n - \theta)^2]}{\epsilon^2} \quad \text{(inequality Markov)}$$
$$= \frac{MSE(\hat{\Theta}_n)}{\epsilon^2}$$

An example for consistency is the sample mean estimator (sec 2.7)

2.5.1   Asymptotic efficiency

An estimator is said to be asymptotically efficient if it achieves the smallest possible variance among all consistent estimators as the sample size increases.

More formally, let θ be an unknown parameter that we wish to estimate based on a random sample $X_1, X_2, \ldots, X_n$. Let $Y_n(\theta)$ be an estimator of θ. The estimator Y_n is said to be asymptotically efficient if, as n approaches infinity, the variance of Y_n approaches the Cramer-Rao lower bound, which is the smallest possible variance among all unbiased estimators of θ.

Intuitively, an asymptotically efficient estimator is the best possible estimator of θ that we can construct using the available information from the sample. However, it is important to note that asymptotic efficiency only holds in the limit as the sample size approaches infinity, and may not hold for finite sample sizes.

2.6   Asymptotic normality

In simple terms, an estimator $\hat{\Theta}$ is said to be asymptotically normal if its sampling distribution approaches a normal distribution as the sample size becomes large, which means :

$$\lim_{n \to \infty} f_{\hat{\Theta}}(x) \to f(x)$$

where $f(x)$ is density function of a normal distribution $\mathcal{N}(a, b)$ or

$$\lim_{n \to \infty} P(\hat{\Theta} \leq x) = \Phi(x)$$

where $\Phi(x)$ is CDF of normal distribution.

An example for asymptotic normality is the sample mean estimator (sec 2.7).Note that other assumptions and conditions may need to be met for asymptotic normality to hold.

2.7   Sample mean estimator

Let $X_1, X_2, X_3, \ldots, X_n$ be a random sample from a distribution with mean μ and std σ.

Sample mean estimator :

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

1. $E(\hat{\mu}) = \mu$

2. Variance of sample mean estimator

$$Var[\hat{\mu}] = Var[\frac{1}{n} \sum_{i=1}^{n} x_i] = \frac{1}{n^2} \left(\sum_{i=1}^{n} Var[x_i] + \sum_{i}^{n} \sum_{i \neq j}^{n} Cov(x_i, x_j) \right) = \frac{\sigma^2}{n}$$

3. Consistency : $\hat{\mu} \to \mu$, by strong version of law of large numbers (sec 4.6.9)

4. Asymptotic normality : By the Central Limit Theorem (sec 1.1)

$$\lim_{n \to \infty} \left(\frac{\hat{\mu} - \mu}{\frac{\sigma}{\sqrt{n}}} \right)$$

or distribution of $\hat{\mu}$ converges to $\mathcal{N}(\mu, \frac{\sigma^2}{n})$. Then $\hat{\mu}$ is asymptotically normal.

2.8 Maximum likelihood estimation

In statistics, maximum likelihood estimation (MLE) is a method of estimating the parameters of an assumed probability distribution, given some observed data. This is achieved by maximizing a likelihood function so that, under the assumed statistical model, the observed data is most probable.

The point in the *parameter space* that maximizes the likelihood function is called the maximum likelihood estimate.

Fomally, the problem for maximum likelihood estimation :

$$\hat{\theta}_{MLE}(\mathbf{X}) = \underset{\theta}{\operatorname{argmax}} \, \mathcal{L}(\mathbf{X}|\theta)$$

where $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ a random sample.

2.8.1 Likelihood function

Likelihood function $\mathcal{L}$ is defined as :

$$\mathcal{L}(\mathbf{X} = \mathbf{x}|\theta) = P_{X_1 X_2 \ldots X_n}(x_1, x_2, \ldots, x_n|\theta),$$

if X_i are discrete value and

$$\mathcal{L}(\mathbf{X} = \mathbf{x}|\theta) = f_{X_1 X_2 \ldots X_n}(x_1, x_2, \ldots, x_n|\theta)$$

if X_i are continuous value. P and f here are respectively mass probability function and density function.

If X_i are independent

$$\mathcal{L}(\mathbf{X} = \mathbf{x}|\theta) = \Pi_{i=1}^n f_{X_i}(x_i|\theta)$$

If X_i are independent and identically distributed

$$\mathcal{L}(\mathbf{X} = \mathbf{x}|\theta) = \Pi_{i=1}^n f(x_i|\theta)$$

In practice, it is often convenient to work with the natural logarithm of the likelihood function, called the Log-Likelihood l function :

$$l(\mathbf{X} = \mathbf{x}|\theta) = \ln(\mathcal{L}(\mathbf{X} = \mathbf{x}|\theta)) = \sum_{i=1}^n \ln(f(x_i|\theta))$$

2.8.2 Asymptotic Properties

- Consistency (subsection 2.5). Thus we have the convergence in probability if $n \to \infty$:
$$\lim_{n \to \infty} P(|\hat{\theta}_{MLE} - \theta| > \epsilon) = 0$$

- $\hat{\theta}_{MLE}$ is asymptotically unbiased :
$$\lim_{n \to \infty} \mathbb{E}[\hat{\theta}_{MLE}] = \theta$$

- $\hat{\theta}_{MLE}$ is asymptotically normal (subsection 2.6)

$$\lim_{n \to \infty} \frac{\hat{\theta}_{MLE} - \theta}{\sqrt{Var(\hat{\theta}_{MLE})}} \to \mathcal{N}(0, 1).$$

- Asymptotically efficient (subsection 2.5.1), it reaches the Cramér-Rao bound. It means that $Var(\hat{\theta}_{MLE}$ is smallest among all estimators.

2.8.3 Example

Given a coin that is unbiased, we toss this coin and have a random sample :

$$X_i = \begin{cases} 1 \text{ if face} \\ 0 \text{ if tail} \end{cases}$$

What is the parameter p that tossed coin is face ?

We know that density function of Bernoulli distribution with a parameter p need to be estimated is:
$$f(x|p) = p^x (1-p)^{1-x}$$

Log-Likelihood :

$$l(\mathbf{x}|\theta) = \sum_{i=1}^{n} \ln(p^{x_i}(1-p)^{1-x_i})$$

$$= \sum_{i=1}^{n} x_i \ln(p) + \sum_{i=1}^{n} (1-x_i) \ln(1-p)$$

$$\frac{\partial l(\mathbf{x}|\theta)}{\partial \theta} = \frac{\sum_{i=1}^{n} x_i}{p} - \frac{\sum_{i=1}^{n}(1-x_i)}{1-p}$$

$$= \frac{\sum_{i=1}^{n} x_i - np}{p(1-p)}$$

Hence :

$$\hat{p} = \frac{\sum_{i=1}^{n} x_i}{n}$$

138

However, in the more complicated model, finding and solving for the derivative of log likelihood is not always trivial. Let's take an example, that we want to estimate parameters of a GMM (Gaussian mixture model). This GMM is composed of two kernels, with $\theta = (w_1, w_2, \mu_1, \sigma_1, \mu_2, \sigma_2)$:

$$f(x|\theta) = \frac{w_1}{w_1 + w_2} \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left(-\frac{1}{2}(\frac{x - \mu_1}{\sigma_1})^2\right) + \frac{w_2}{w_1 + w_2} \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left(-\frac{1}{2}(\frac{x - \mu_2}{\sigma_2})^2\right)$$

Log-likelihood function :

$$l(\mathbf{x}|\theta) = \sum_{i=1}^{n} \ln f(x_i|\theta)$$

is hard to get the derivative for each parameter and also solve for them.

2.9 Confidence Intervals

Let $X_1, X_2, \ldots, X_n$ be a random sample (observation) from a distribution with a parameter θ which is to be estimated. Suppose that we have observed $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$ and the point estimator $\hat{\theta}$ is a function of $X_1, X_2, \ldots, X_n$. The problem is raised here, we do not know how close $\hat{\theta}$ is to the real θ. Hence, we have concept of **interval estimation**.

Instead of finding out only an estimation $\hat{\theta}$, we provide in addition :

- An *interval* $[\hat{\theta}_l, \hat{\theta}_h]$, with high and low estimate

- A *confidence level* $(1 - \alpha)$. We would like α to be small. Common values for α are $0.1, 0.05, 0.01$, which correspond to confidence levels 90%, 95%, 99% respectively.

In summarizing, we need to find $\hat{\theta}_l$ and $\hat{\theta}_h$ such that the probability that the interval $[\hat{\theta}_l, \hat{\theta}_h]$ includes θ is larger than $1 - \alpha$:

$$P(\hat{\theta}_l \leq \theta \leq \hat{\theta}_h) \geq 1 - \alpha$$

2.9.1 Finding Interval Estimator

Let X be a continuous random variable with CDF $F_X(x) = P(X \leq x)$. Suppose that we are interested in finding two values x_h and x_l such that

$$P(x_l \leq X \leq x_h) = 1 - \alpha$$

One way to do this, is to chose x_l and x_h such that

$$P(X \leq x_l) = \frac{\alpha}{2} \quad \text{and} \quad P(X \geq x_h) = \frac{\alpha}{2}$$

Equivalently

$$F_X(x_l) = \frac{\alpha}{2} \quad \text{and} \quad F_X(x_h) = 1 - \frac{\alpha}{2}$$

Then
$$x_l = F_X^{-1}(\frac{\alpha}{2}) \quad \text{and} \quad x_h = F_X^{-1}(1 - \frac{\alpha}{2})$$

Example

Let $X_1, X_2, \ldots, X_n$ be a random sample from a normal distribution $N(\theta, 1)$. Find a 95% confidence interval for θ. In subsection 2.7, we have that the sample mean is asymptotically normal :

$$\frac{\bar{X} - \theta}{\sqrt{Var(\bar{X})}} = \frac{\bar{X} - \theta}{\sqrt{\frac{1}{n}Var(X_i)}} = \frac{\bar{X} - \theta}{\sqrt{\frac{1}{n}}} \xrightarrow{n \to \infty} \mathcal{N}(0, 1)$$

$$P(-x_{\frac{\alpha}{2}} \leq \sqrt{n}(\bar{X} - \theta) \leq x_{\frac{\alpha}{2}}) = 1 - \alpha$$

where $\alpha = 0.05$ and $x_{\frac{\alpha}{2}} = 1.96$. To find the interval :

$$P(\bar{X} + \frac{x_{\frac{\alpha}{2}}}{\sqrt{n}} \geq \theta \geq \bar{X} - \frac{x_{\frac{\alpha}{2}}}{\sqrt{n}}) = 1 - \alpha$$

Therefore, we can report the interval $[\bar{X} - \frac{x_{\frac{\alpha}{2}}}{\sqrt{n}}, \bar{X} + \frac{x_{\frac{\alpha}{2}}}{\sqrt{n}}]$ as our $(1 - \alpha)100\%$ confidence interval for θ, which means that $(1 - \alpha)100\%$ we have the real θ in this interval.

2.9.2   Pivotal Quantity

Let $X_1, X_2, X_3, \ldots, X_n$ be a random sample from a distribution with a parameter θ that is to be estimated. The random variable Q is said to be a pivot or a pivotal quantity, if it has the following properties:

- Q depends only on $X_1, X_2, X_3, \ldots, X_n$ and parameter θ.

- The probability distribution of Q does not depend on θ.

2.9.2.1   Example

Let X_1, X_2, X_3, X_4, X_5 be a random sample from $\mathcal{N}(\theta, 1)$. Then

$$Q = \bar{X} - \theta$$

where $\bar{X}$ is the sample mean, is a pivotal quantity since :

- Q depends only on $X_1, X_2, X_3, \ldots, X_n$ and parameter θ.

- The probability distribution of Q is $\mathcal{N}(0, \frac{1}{5})$.

2.10 Hypothesis testing

Often, we need to test whether a hypothesis is true or false :

- H_0 (the null hypothesis): $\theta \in S_0$.

- H_1 (the alternative hypothesis): $\theta \in S_1$.

We distinct between simple and composite:

- Simple hypothesis : only one hypothesis, e.g. $\theta = \frac{1}{3}$

- Composite hypothesis : more than one hypothesis, e.g. $\theta = \frac{1}{3}$ ou $\theta = \frac{2}{3}$

Example 1, consider a radar system that uses radio waves to detect aircraft. The system receives a signal and, based on the received signal, it needs to decide whether an aircraft is present or not. Here, there are again two opposing hypotheses:

- H_0 : No aircraft is present. (null hypothesis)

- H_1: An aircraft is present. (alternative hypothesis)

Example 2

We have a coin and after a number of observations n, we would like to check whether it is fair or not. More specifically, let θ be the probability of heads, $\theta = P(H)$, we have two hypotheses:

- H_0 : Coin is fair $\theta = \frac{1}{2}$. (null hypothesis)

- H_1: The coin is not fair. (alternative hypothesis)

2.10.1 Type I error

In order to find T, we need define the first notion called *Type I error* as the event that we reject H_0 when H_0 is true. :

$$P(H_0 \text{ is rejected } | H_0 \text{ is true})$$

equivalently

$$P(H_0 \text{ is rejected } | \theta), \quad \forall \theta \in S_0$$

If the probability of type I error satisfies

$$P(H_0 \text{ is rejected } | \theta) \leq \alpha, \quad \forall \theta \in S_0,$$

then we say the test has significance level α or simply the test is a level α test.

2.10.1.1 Example

Let $X = X_1 + X_2 + ...X_n$, where X_i is an observation. Logically, if we have threshold T, we can say that:

- If $|X - \frac{n}{2}| \leq T : H_0$ is accepted

- If $|X - \frac{n}{2}| > T : H_0$ is rejected

Then the Type I error is :

$$P(|X - \frac{n}{2}| > T | H_0 \text{ is true})$$

As in the estimation of confidence intervals (sec 2.9), usually we take $\alpha = 0.05$

Now, as X_i follows Bernouilli distribution then X follows Binominal distribution. If H_0 is true, $X \sim Binominal(n, \frac{1}{2})$. Since $\alpha = 0.05$ and Binominal distribution is symetric (only in this case $\theta = 0.5$) at $\frac{n}{2}$, then T can be found such as :

$$\sum_{k=50+T}^{n} 0.5^k (1 - 0.5)^{n-k} = \frac{\alpha}{2} = 0.025$$

If n is large enough. H_0 is true means that $X_i \sim Bernouill(0.5)$ with mean 0.5 and variance 0.5(1-0.5). Applying CLT (sec 1.1)

$$Y = \frac{\bar{X} - 0.5}{\frac{\sqrt{0.5(1-0.5)}}{\sqrt{n}}} = \frac{\sqrt{n}(\frac{X}{n} - 0.5)}{0.5} \sim \mathcal{N}(0, 1)$$

T can be found as:
$$\frac{\sqrt{n}(X - \frac{n}{2})}{0.5n} = x_{\frac{\alpha}{2}}$$

$$T = \frac{1}{2}\sqrt{n} x_{\frac{\alpha}{2}} = 9.8 \approx 10$$

2.10.1.2 Remarks

Suppose that we toss the coin 100 times and observe 55 heads. Based on the above discussion we should accept H_0 (if $\alpha = 0.05$. However, it is often recommended to say "we failed to reject H_0" instead of saying "we are accepting H_0". The reason is that we have not really proved that H_0 is true.

In fact, all we know is that the result of our experiment was not statistically contradictory to H_0.

2.10.2 Type II error

This is the second possible error that we can make is to accept H_0 when H_0 is false, equivalently when H_1 is true.

$$P(H_0 \text{ is accepted} | H_1 \text{ is accepted})$$

Since the alternative hypothesis, H_1, is usually a composite hypothesis (so it includes more than one value of θ), the probability of type II error is usually a function of θ :

$$\beta(\theta) = P(H_0 \text{ is accepted } | \theta)$$

or

$$\beta(\theta) = P(H_1 \text{ is rejected } | \theta)$$

for all $\theta \in S_1$

2.10.2.1 Example

Consider a radar system that uses radio waves to detect aircraft. Let X be the received signal. If $X = 0$ there is no aircraft and if $X = 1$, there is an aircraft. Suppose that we know only noised signal, such that :

- $X = W$, if no aircraft is present.

- $X = 1 + W$, if an aircraft is present.

where $W \sim \mathcal{N}(0, \frac{1}{9})$.
    Then the hypotheses :

- H_0 (null hypothesis): No aircraft is present.

- H_1 (alternative hypothesis): An aircraft is present.

What is the threshold c if we would like the probability of missing a present aircraft to be less than 1% ?
    Solution In this problem, the type II error is considered ;

$$P(H_0 \text{ is accepted } | H_1 \text{ is accepted })$$

H_1 is accecpted means that $X = 1 + W$ or $X \sim \mathcal{N}(1, \frac{1}{9})$. As $\beta = 0.01$, we must find c such that :
$$P(X < c) = 0.01$$

or
$$P(3(X - 1) < 3(c - 1)) = 0.01$$

As $3(X - 1) \sim \mathcal{N}(0, 1)$, then

$$3(c - 1) = \Phi^{-1}(0.01)$$

or

$$c = \frac{1}{3}\Phi^{-1}(0.01) + 1$$

where Φ is cumulative distribution of the standard normal.

2.10.2.2   Note

We have the following explanation:

- Type I error : this is the probability of detecting (true) when there is no aircraft (false alarm, false positive). If given $alpha = 0.05$, it means that a false alarm is less than 5%.

- Type II error : this is the probability of missing a present aircraft (false negative). If given $\beta = 0.01$ it means that a false negative is less than 1%.

2.10.3   Hypothesis Testing for the Mean

We assume that we have a random sample $X_1, X_2, \ldots, X_n$ from a distribution and our goal is to make inference about **the mean of the distribution** μ. We consider two tests : two-sided and one-sided. Hyphotheses in two-sided test:

- $H_0 : \mu = \mu_0$

- $H_1 : \mu \neq \mu_0$

and hyphotheses in one-sided test :

- $H_0 : \mu \leq \mu_0$

- $H_1 : \mu > \mu_0$

where μ_0 is given.

In both two tests, we use the sample mean :

$$\bar{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

Let's define :

$$Q(X_1, X_2, \ldots, X_n) = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

if the true std σ of the distribution is given or

$$Q(X_1, X_2, \ldots, X_n) = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

where S is sample standard deviation. Under hypothesis H_0, in the first case, Q follows standard normal distribution and in the second case, Q follows $T(n-1)$ distribution (sec 4.2).

2.10.3.1   Two-sided Tests for the Mean

- $H_0 : \mu = \mu_0$

- $H_1 : \mu \neq \mu_0$

In this case, the null hypothesis is a simple hypothesis and the alternative hypothesis is a **two-sided** hypothesis ($\mu < \mu_0$ or $\mu > \mu_0$). Then we need to find c such that :

$$
\begin{aligned}
P(\text{type I error}) &= P(H_0 \text{ is rejected } | H_0 \text{ is true}) \\
&= P(|Q| > c | \mu = \mu_0) \\
&\leq \alpha
\end{aligned}
$$

or

$$
\begin{aligned}
P(\text{type II error}) &= P(H_0 \text{ is accepted } | H_1 \text{ is true}) \\
&= P(|Q| < c | \mu \neq \mu_0) \\
&\leq \beta
\end{aligned}
$$

Example Let $X_1, X_2, ..., X_n$ be a random sample from a $\mathcal{N}(\mu, \sigma^2)$ distribution, where μ is unknown but σ is known. Design tests (type I error and type II error) to choose between the hypotheses :

- $H_0 : \mu = \mu_0$

- $H_1 : \mu \neq \mu_0$

For type I error, under the H_0 hyphothesis, the mean of the distribution μ is equal to μ_0, hence $Q \sim \mathcal{N}(0,1)$. We need to choose threshold c such that :

$$
P(|Q| > c) = \alpha
$$

or since $\mathcal{N}(0,1)$ is sysmetric :

$$
P(Q > c) = \frac{\alpha}{2}
$$

we infer that $c = \Phi^{-1}(1 - \frac{\alpha}{2})$, where Φ is CDF of the standard normal. Then H_0 is rejected if :

$$
\left| \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \right| > \Phi^{-1}(1 - \frac{\alpha}{2})
$$

For type II error, under the H_1 hypothesis, the mean of the distribution μ is not equal to μ_0. Now μ can take a range of value (composite hypothesis), therefore, we study for each case of μ.

Since $X_i \sim \mathcal{N}(\mu, \sigma^2)$, then $\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$, then $Q = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim \mathcal{N}(\frac{\mu - \mu_0}{\sigma/\sqrt{n}}, 1)$.

Let β is a acceptable false negative rate. **For each given value** μ we find $c(\mu) > 0$ such that :

$$
\begin{aligned}
&P(|Q| < c(\mu) | \mu) = \beta \\
&P(-c(\mu) < Q < c(\mu) | \mu) = \beta \\
&P\left(-c(\mu) - \frac{\mu - \mu_0}{\sigma/\sqrt{n}} < Q - \frac{\mu - \mu_0}{\sigma/\sqrt{n}} < c(\mu) - \frac{\mu - \mu_0}{\sigma/\sqrt{n}} \Big| \mu \right) = \beta \\
&\Phi\left(c(\mu) - \frac{\mu - \mu_0}{\sigma/\sqrt{n}} \right) - \Phi\left(-c(\mu) - \frac{\mu - \mu_0}{\sigma/\sqrt{n}} \right) = \beta
\end{aligned}
$$

This is thus a hard problem to obtain $c(\mu)$. However, we can use this equation in the reverse direction, if $c(\mu)$ is known, we can infer easily β, the false negative rate.

2.10.3.2   One-sided Tests for the Mean

- $H_0 : \mu \leq \mu_0$

- $H_1: \mu > \mu_0$

In this case, we have one-sided test and note the follwoing inquality:

$$P(\text{type I error}) = P(H_0 \text{ is rejected } | H_0 \text{ is true})$$
$$= P(Q > c | \mu \leq \mu_0)$$
$$= \alpha$$

Example Let $X_1, X_2, \ldots, X_n$ be a random sample from a $\mathcal{N}(\mu, \sigma^2)$ distribution, where μ is unknown and σ is known. Design a level α test to choose between

- $H_0 : \mu \leq \mu_0$

- $H_1 : \mu > \mu 0$

We study for each given value $\mu \leq \mu_0$:

$$P(Q > c | \mu \leq \mu_0)$$
$$= P(Q - \frac{\mu - \mu_0}{\sigma/\sqrt{n}} > c - \frac{\mu - \mu_0}{\sigma/\sqrt{n}} | \mu \leq \mu_0)$$
$$\leq P(Q - \frac{\mu - \mu_0}{\sigma/\sqrt{n}} > c | \mu \leq \mu_0), \quad \text{since } \mu \leq \mu_0$$
$$= 1 - \Phi(c)$$

Then we can choose $c = \Phi^{-1}(1 - \alpha)$

2.10.4   P-Values

P-value is the lowest significance level α that results in rejecting the null hypothesis.

Example Suppose that we have a two-sided test and the pivotal quantity $Q \sim \mathcal{N}(0, 1)$. The observation given that $Q = 2$. Then the p-value, which corresponds to $c = 2$ (for a two-sided test) is:

$$2(1 - \Phi(2)) = 0.046$$

2.10.5   Likelihood Ratio Tests

2.10.5.1   Likelihood Ratio Test for Simple Hypotheses

Let $X_1, X_2, \ldots, X_n$ be a random sample (observation) from a distribution with a parameter θ.

- $H_0 : \theta = \theta_0$

- $H_1 : \theta = \theta_1$

we define

$$\lambda(x_1, x_2, \ldots, x_n) = \frac{\mathcal{L}(x_1, x_2, \ldots, x_n | \theta = \theta_0)}{\mathcal{L}(x_1, x_2, \ldots, x_n | \theta = \theta_1)}$$

To perform a likelihood ratio test (LRT), we choose a constant c. We reject H_0 if $\lambda < c$ and accept it if $\lambda \geq c$. The value of c can be chosen based on the desired α (type I) or β (type II).

Example

Here, we look at the radar problem. More specifically, we observe the random variable $X : X = \theta + W$, where $W \sim \mathcal{N}(0, \sigma^2 = \frac{1}{9})$. We need to decide between

- $H_0 : \theta = \theta_0 = 0$

- $H_1 : \theta = \theta_1 = 1$

Let $X = x$ an observation. Design a level $\alpha = 0.05$ to decide between H_0 and H_1.

$$L(x | \theta = \theta_0) = \frac{1}{\sqrt{2\pi \frac{1}{3}}} \exp\left(-\frac{1}{2} 9 x^2\right)$$

$$L(x | \theta = \theta_1) = \frac{1}{\sqrt{2\pi \frac{1}{3}}} \exp\left(-\frac{1}{2} 9 (x - 1)^2\right)$$

$$\lambda(x) = \exp\left(\frac{9(1 - 2x)}{2}\right)$$

Equivalently, we accept H_0 if

$$\lambda(x) \geq c \Leftrightarrow x \leq \frac{1}{2}\left(1 - \frac{2}{9} lnc\right).$$

Let $c' = \frac{1}{2}(1 - \frac{2}{9} lnc)$. Given α, it means type I error, such that :

$$P(\text{type I error}) = P(H_0 \text{ is rejected } | H_0 \text{ is true})$$
$$= P(X > c' | \theta = \theta_0) \leq \alpha$$
$$= P(X > c') \leq \alpha \quad \left(X \sim \mathcal{N}(0, \frac{1}{9})\right)$$

Then we can find the value c' and then c.

2.10.5.2 Likelihood Ratio Test for Composite Hypotheses

Let $X_1, X_2, \ldots, X_n$ be a random sample (observation) from a distribution with a parameter θ.

- $H_0 : \theta \in S_0$

- $H_1 : \theta \in S_1$

we define

$$\lambda(x_1, x_2, \ldots, x_n) = \frac{\max_\theta L(x_1, x_2, \ldots, x_n | \theta \in S_0)}{\max_\theta L(x_1, x_2, \ldots, x_n | \theta \in S)}$$

where $S = S_0 \cup S_1$. To perform a likelihood ratio test (LRT), we choose a constant $c \in [0, 1]$. We reject H_0 if $\lambda < c$ and accept it if $\lambda \geq c$.

2.11 Linear Regression

This model tries to find the linear relationship (finding β_i) between variables between y and x_i :

$$y = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

There are different types of linear regression such as :

- (Linear) Least square

$$\min_\beta (y - (\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p))^2$$

- Polynomial (linear) regression

$$\min_\beta (y - (\beta_0 + \beta_1 x^1 + \beta_2 x^2 + \ldots + \beta_k x^k))^2$$

- Ridge (linear) regression

$$\min_\beta (y - (\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p))^2 + \alpha \, \|\beta\|_2^2$$

- Lasso (linear) regression

$$\min_\beta (y - (\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p))^2 + \alpha \, \|\beta\|_1$$

Note that **Least Squares** and **Linear model** are totally two different concepts:

- **Least Squares** means minimizing the square distance between true value and its predicted value $(y - \hat{y})^2$

- **Linear model** : only in case that $\hat{y}$ is modeled by a linear relation.

We will discover several method to resolve a *simple linear regression model* (finding β_0 and β_1). Our model is

$$Y = \beta_0 + \beta_1 X + \epsilon$$

with n observation $(y_i, x_i), i = 1 \ldots n$.

2.11.1 First method

In this first method, we must have the two following hypotheses :

- $\epsilon \sim \mathcal{N}(0, \sigma^2)$

- ϵ and X are independent.

First, we take expectation from both sides :

$$\mathbb{E}[Y] = \beta_0 + \beta_1 \mathbb{E}[X]$$

Second, we take the covariance between Y and X:

$$\begin{aligned}
Cov(Y, X) &= Cov(\beta_0 + \beta_1 X + \epsilon, X) \\
&= \beta_1 Cov(X, X) + Cov(\epsilon, X) \\
&= \beta_1 Var(X)
\end{aligned}$$

Finally,

$$\beta_1 = \frac{Cov(Y, X)}{Var(X)}$$

$$\beta_0 = \mathbb{E}[Y] - \frac{Cov(Y, X)}{Var(X)} \mathbb{E}[X]$$

where : $\mathbb{E}[X] = \frac{\sum x_i}{n}$, $\mathbb{E}[Y] = \frac{\sum y_i}{n}$, $Var(X) = \frac{\sum (x_i - \mathbb{E}[X])^2}{n}$, $Cov(X, Y) = \frac{\sum (x_i - \mathbb{E}[X])(y_i - \mathbb{E}[Y])}{n}$

2.11.2 Second method

In this method, no hypothesis is needed. Let g be the sum of square error :

$$g(\beta_0, \beta_1) = \sum_i (y_i - \beta_0 - \beta_1 x_i)^2$$

First order condition :

$$\frac{\partial g}{\partial \beta_0} = 0$$

$$\Leftrightarrow -2 \sum (y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\Leftrightarrow \beta_0 + \beta_1 \frac{\sum x_i}{n} = \frac{\sum y_i}{n}$$

$$\frac{\partial g}{\partial \beta_1} = 0$$

$$\Leftrightarrow \sum -2x_i (y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\Leftrightarrow \beta_0 \sum x_i + \beta_1 \sum x_i^2 = \sum x_i y_i$$

Hence:

$$\beta_1 = \frac{(x_i - \bar{x})^2}{(x_i - \bar{x})(y_i - \bar{y})}$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

2.11.3 Third method

In this method by MLE (sec 2.8), we must have the the following hypothesis :

- $\epsilon \sim \mathcal{N}(0, \sigma^2)$

Hence $Q = Y - \beta_0 - \beta_1 X$ follows standard normal distribution.

$$L(x_1, \ldots, x_n, y_1, \ldots, y_n | \beta_0, \beta_1) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(y_i - \beta_0 - \beta_1 x_i)^2\right)$$

Maximizing $ln(L)$ leads us to the second method.

2.11.4 Ordinary Least Squares

Ordinary Least Squares (OLS) is a common statistical method used for estimating the relationship between output y (also called dependent variable or reponse) and p inputs $(x_i)_{i=1,\ldots,p}$ (also called explanatory variables). Under the additional assumption that the noises (errors) *are normally distributed with zero mean*, OLS is the maximum likelihood estimator that outperforms any non-linear unbiased estimator. A remind about the difference between OLS and LS, mainly by **noise** :

- Ordinary Least Squares (OLS) - Noise is white noise, i.e., i.i.d. from the normal distribution with zero mean and finite variance.

- Linear Least Squares (LLS) - Noise with *different parameters* per sample or *correlated noise.* Hence OLS is a particular case of LLS

Problem:
$$X\beta \approx y$$

$$\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \times \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} \approx \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

- n observations $\mathbf{x}_i, y_i, i = 1, \ldots, n$.

- $\mathbf{x}_i = [x_{i1}, ..., x_{ip}]^T$.

- row i of X is $\mathbf{x_i}^T$ is i-th observations on all the explanatory variables.

- X is called matrix of regressors and $X[:, j]$ is called a regressor or all observations of explanatory variable j.

- Usually $x_{i1} = 1, \forall i$ and β_1 is called intercept.

Regressors do not have to be **independent**: there can be any desired relationship between the regressors (so long as it is not a **linear relationship**)
  Solution :

$$\beta = (X^T X)^{-1} X^T y$$

- $X^T X$ is called Gram matrix

- $X^T y$ is called moment matrix

- $(X^T X)^{-1}$ is the cofactor matrix of β, closely related to its covariance matrix, C_β.

- $(X^T X)^{-1} X^T$ is Moore–Penrose pseudo-inverse of X.

We find that OLS (with linear model) is something equivalent linea MMSE in sec 3.5.1

2.11.4.1 Important notes

- After estimating β by $\hat{\beta}$, the fitted values (or predicted values) from the regression will be:

$$\hat{y} = X\hat{\beta} = Py = X(X^T X)^{-1} X^T y$$

  P is the projection matrix which maps y to $\hat{y}$. We can say $P : R^n \to V$, where V spanned by the columns of X.

- Another matrix, closely related to P is the *annihilator* matrix $M = I_n - P$; this is a projection matrix onto the subspace orthogonal to V.

- Both matrices P and M are symmetric and idempotent (meaning that $P^2 = P$ and $M^2 = M$), and relate to the data matrix X via identities $PX = X$ and $MX = 0$.

- Let $\varepsilon = y - X\beta$ be the true noise (error, residual) then the estimated noise can be expressed as :

$$\hat{\varepsilon} = y - \hat{y} = y - X\hat{\beta} = (I_n - P)y = My = M(X\beta + \varepsilon) = M\varepsilon$$

2.11.4.2 Noise variance estimation

There are two estimators for the variance of estimated noise $\hat{\varepsilon}$, one is biased but has but has a smaller mean squared error and the other is unbiased.

- **Biased estimator**, by MLE :

$$\hat{\sigma}_1^2 = \frac{\hat{\varepsilon}^T \hat{\varepsilon}}{n}$$

- **Unbiased estimator**:

$$\hat{\sigma}_2^2 = \frac{\hat{\varepsilon}^T \hat{\varepsilon}}{n - p}$$

  This is a minimum-variance unbiased estimator, which has lower variance than any other unbiased estimator. Also it is called the OLS estimator for noise variance.

Always with the unbiased estimator, let's set :

$$s^2 = \hat{\sigma}_2^2 = \frac{(y - X\hat{\beta})^T (y - X\hat{\beta})}{n - p} = \frac{(My)^T My}{n - p} = \frac{y^T M^T My}{n - p} = \frac{y^T My}{n - p}$$

where s (without square) is also called **standard error of regression**. Furthermore, we can show that s^2 follows reduced chi-square statistic (sec 4.1.6).

2.11.5 Coefficient of determination

R^2 is a measure of the goodness of fit of a model. Given

- $y_1, y_2, ..., y_n$ observations
- $f_1, f_2, ..., f_n$ predicted value of y_i
- $SS_{res} = \sum_i (f_i - y_i)^2$
- $SS_{tot} = \sum_i (y_i - \bar{y})^2$

Then R^2 is defined by :

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{\sum_i (f_i - y_i)^2}{\sum_i (y_i - \bar{y})^2}$$

SS means sum of square. In a good fitting, we must have $R^2 = 1$ or closed to 1. A trivaial fitting by taking $f_i = \bar{y}$ gives $R^2 = 0$. Note that a bad fitting can make $R^2 < 0$. In practice : R^2 less than 10% is bad, R^2 greater than 30% is good.

2.11.5.1 Other explanation

We continue the mothod described in subsection 2.11.1

$$
\begin{aligned}
R^2 &= 1 - \frac{\sum_i (\hat{y}_i - y_i)^2}{\sum_i (y_i - \bar{y})^2} \\
&= 1 - \frac{\sum \epsilon_i^2}{n Var(Y)} \\
&= 1 - \frac{Var(\epsilon)}{Var(Y)} \quad \left(\frac{\sum \epsilon_i^2}{n} = \mathbb{E}[\epsilon^2] = Var(\epsilon) \right) \\
&= \frac{Var(Y) - Var(\epsilon)}{Var(Y)} \\
&= \frac{\beta_1^2 Var(X)}{Var(Y)} \quad \left(Var(Y) = \beta_1^2 Var(X) + Var(\epsilon) \right) \\
&= \frac{Cov(X,Y)^2}{Var(X)Var(Y)} \\
&= Cor(X,Y)^2
\end{aligned}
$$

In this view, R^2 is large if X and Y are highly correlated or highly uncorrelated.

2.11.5.2 Coefficient of determination in OLS

$$
R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{\sum_i (\hat{y}_i - y_i)^2}{\sum_i (y_i - \bar{y})^2} = 1 - \frac{y^T M y}{y^T L y} = \frac{y^T P^T L P y}{y^T L y} = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2}
$$

where L is centering matrix $L = I - \frac{1}{n} J_n$. J_n is all one matrix.

2.11.6 Adjusted R-squared

$$
1 - \frac{(1 - R^2)(N - 1)}{(N - p - 1)}
$$

where p number of independent variables (explanatory variables or predictors) and N number of used observations.

Problematic : Mesuring the performance of model (fitting) in taking into account the number of independent variables. E.g, given 2 independent variables (inputs) a and b, y is output, then in model 1, we fit y with only a and in model 2, we fit y with a and b. Then the R^2 of model 2 is equal or greater than model 1. But model 2 has advantage that it uses 2 independent variables, then in order to have a fair comparison, p is introduced in the adjusted R-squared.

3 Statistical inference : Bayesian Inference

In classical or frequentist approach. The unknown quantity θ is assumed to be a **fixed** (non-random) quantity that is to be estimated by the observed data.

In this section, we would like to discuss a different framework for inference, namely the Bayesian approach. In the Bayesian framework, we treat the unknown quantity, θ, as a random variable.

3.1 Bayes' theorem

With A, B are two random variables that have the same probability space $(\Omega, \mathcal{F}, P)$, then

$$P(A = a | B = b) = \frac{P(A = a \cap B = b)}{P(B = b)} = \frac{P(A = a | B = b)P(A = a)}{P(B = b)}$$

for all $a \in \mathcal{F}_A$ and $b \in \mathcal{F}_B$. $a \cap b \neq \emptyset$ indicates that two events a and b have common elements in sample space Ω.

Note that do not confuse $P(A \cap B)$ and $P(A, B)$:

- In $P(A \cap B)$: A, B take value from the same event space $\mathcal{F}$ and the function probability $P : \Omega \to [0, 1]$.

- In $P(A, B)$: A, B take value respectively from $\mathcal{F}_A$ and $\mathcal{F}_B$. The function probability $P : \Omega_A \times \Omega_B \to [0, 1]$.

Or in some probability literature, with $z \to x$:

- z latent variable

- x observation

- $p(x)$ marginal or evidence

- $p(z|x)$ posterior probability

- $p(x|z)$ likelihood

- $p(z)$ prior probability

Then :

$$p(z|x) = \frac{p(x|z)p(z)}{p(x)}$$

3.1.1 Example 1

Let $X \sim Uniform(0, 1)$. Suppose that we know $Y|X = x \sim Geometric(x)$. Find the posterior density of X given $Y = 2$, $f_{X|Y}(x|2)$.

Solution:

- X (prior) is uniform with density function :

$$P(X = x) = 1$$

- $Y|X$ (likelihood) follows geometric distribution $\mathcal{G}(x)$, with density function :

$$P(Y|X = x) = (1 - x)^{y-1}x$$

Then posterior :

$$
\begin{aligned}
f_{X|Y}(x|2) &= P(X|Y = 2) \\
&= \frac{P(Y = 2|X = x)P(X = x)}{P(Y = 2)} \\
&= \frac{(1 - x)x}{\int_0^1 (1 - x)x\,dx} \\
&= \frac{(1 - x)x}{\frac{1}{6}} \\
&= 6(1 - x)x
\end{aligned}
$$

3.1.2 Example 2

You have a coin and your prior assumption is that its probability of heads θ is chosen from a uniform distribution on $[0, 1]$. You toss the coin 10 times and get 6 heads. What is the estimate of θ?

Solution:

- θ (prior) is uniform with density function :

$$P(\Theta = \theta) = 1$$

- $X|\Theta$ (likelihood) follows binominal $\mathcal{B}(10, \theta)$, with density function :

$$P(X = 6|\Theta = \theta) = \binom{6}{10}\theta^6(1 - \theta)^{10-6}$$

Then posterior :

$$P(\Theta|X = 6) = \frac{P(X = 6|\Theta)P(\Theta)}{P(X = 6)}$$

We find θ in this case, which is the expected value :

$$\mathbb{E}_{\Theta|X}[\theta|6] = \int_0^1 \theta P(\theta|6)d\theta$$

$$= \frac{1}{P(X=6)} \int_0^1 \binom{6}{10}\theta^7(1-\theta)^4 d\theta$$

$$= \frac{1}{\sum_\theta P(X=6|\Theta=\theta)P(\Theta=\theta)} \int_0^1 \binom{6}{10}\theta^7(1-\theta)^4 d\theta$$

$$= \frac{\int_0^1 \binom{6}{10}\theta^7(1-\theta)^4 d\theta}{\int_0^1 \binom{6}{10}\theta^6(1-\theta)^4 d\theta}$$

$$= \frac{0.0530}{0.0909}$$

$$= 0.58333$$

The integral can be calculated by a calculatrice or using e.g. python.

Other solution:

We remind that the density function of *beta*-distribution, depends on a and b. This density function is denoted by :

$$\beta(a,b)(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}x^{(a-1)}(1-x)^{(b-1)}$$

where Γ is *gamma*-function, $\Gamma(n) = (n-1)!$. Then :

- θ (prior) is uniform and it is *beta*-distribution. Density function :

$$f(\theta) = \beta(1,1)(\theta)$$

- $(X=6|\Theta=\theta)$. Its density function is also *beta*-distribution :

$$P(X=6|\theta) = \binom{6}{10}\theta^6(1-\theta)^4 = \beta(7,5)(\theta)\frac{\binom{6}{10}}{\frac{\Gamma(12)}{\Gamma(5)\Gamma(7)}}$$

- $P(X=6)$ is a constant.

Then posterior distribution is (by Baye) :

$$f_{\theta|X}(\theta) = \frac{\binom{6}{10}\theta^6(1-\theta)^4}{P(X=6)} = \beta(7,5)(\theta)\frac{\binom{6}{10}}{P(X=6)\frac{\Gamma(12)}{\Gamma(5)\Gamma(7)}}$$

Here we can say that posterior distributio $\theta|X$ follows also posterior distribution $\beta(7,5)$, since $\frac{\Gamma(12)}{\Gamma(5)\Gamma(7)}$ and $\frac{\binom{6}{10}}{P(X=6)}$ is just for normalization. Simple check if these two quantities are equal :

$$\frac{\Gamma(12)}{\Gamma(5)\Gamma(7)} = 2310$$

$$\frac{\binom{6}{10}}{P(X=6)} = \frac{210}{0.0909} = 2310.23$$

As the expected value of $\beta(a,b)$ distribution is $\frac{a}{a+b}$, hence $\mathbb{E}(\theta|X) = \frac{7}{7+5} = \frac{7}{12} = 0.58333$.

3.2 Conditional probability

The probability of Y conditional on $X = x$ is denoted :

$$P(Y|X = x)$$

3.2.1 Example

Suppose that the signal $X \sim \mathcal{N}(0, \sigma_X^2)$ is transmitted over a communication channel. Assume that the received signal is given by $Y = X + W$, where $W \sim \mathcal{N}(0, \sigma_W^2)$ is independent of X. Show that $X|Y$ is normal distribution.

False reasoning:
From $Y = X + W$, we have:

$$(X|Y = y) = y - W$$

This is wrong since $X|Y$ depends on W, violated the fact that X and W are independent.

Good reasoning:
Since :

- $Y = X + W$, then $\alpha Y + \beta X = (\alpha + \beta)X + \alpha W$.

- In addition, X and W are independent normal distribution, then $(\alpha + \beta)X + \alpha W$ is also normal distribution for all α, β.

. The two above observations infer that $\alpha Y + \beta X$ is normal distribution for all α, β or X, Y are jointly normal. Then from density function of bivariate normal $f_{X,Y}$, we infer that $f_{X|Y}$ is normal distribution.

3.3 Maximum a posteriori estimation MAP

Maximum a posteriori (MAP) is a statistical method used to estimate the most likely value of a parameter in a probability distribution, given some observed data. In other words, it is a method for finding the point in the parameter space that maximizes the posterior probability, which is the probability of the parameter given the observed data and any prior knowledge we may have about the parameter.

Mathematically, the MAP estimate of a parameter θ can be written as:

$$\hat{\theta}_{MAP}(\mathbf{X} = \mathbf{x}) = \operatorname*{argmax}_{\theta} f(\theta|\mathbf{x})$$

By Baye's theorem, the MAP estimate takes into account both the observed data and any prior knowledge we may have about the parameter :

$$\hat{\theta}_{MAP}(\mathbf{X} = \mathbf{x}) = \operatorname*{argmax}_{\theta} f(\mathbf{x}|\theta)f(\theta)$$

Compared to MLE in sec 2.8, we have in addition prior distribution $f(\theta)$.

3.3.1 Example

Let X be a continuous random variable with the following PDF:

$$f_X(x) = \begin{cases} 2x & \text{if } 0 \leq x \leq 1 \\ 0 & \text{overwise} \end{cases}$$

Also, suppose that

$$Y|X = x \sim Geometric(x)$$

Find the MAP estimate of X given $Y = 3$.
Solution We need to maximize $g(x)$:

$$g(x) = P_{Y|X}(3|x)f_X(x) = (1-x)^2 x 2x$$

Then MAP estimation is at x such that $g'(x) = 0$, equivalently $x = \frac{1}{2}$.

3.4 Minimum Mean Squared Estimator

3.4.1 Definition

Given a random variable X that follows a certain distribution. Then estimator MMSE of X is a such that:

$$\min_{a} \mathbb{E}[(X - a)^2]$$

Using first order condition w.r.t a, we have $a = \mathbb{E}[X]$. And the minimum of $\mathbb{E}[(X - a)^2]$ is thus $Var(X)$.

3.4.2 Conditional expectation

Given a posterior distribution, $f_{X|Y}(x|y)$ contains all the knowledge that we have about the unknown quantity X. Therefore, to find a point estimate of X, we can just choose a summary statistic of the posterior such as its mean, median, or mode. If we choose the mode (the value of x that maximizes $f_{X|Y}(x|y)$), we obtain the MAP estimate of X. Another option would be to choose the expected value of posterior :

$$\hat{x} = \mathbb{E}[X|Y = y]$$

We aslo call $\mathbb{E}[X|Y=y]$ is conditional expectation.

This is a particular case of conditional expectation with respect to a sub sigma algebra in sec 2.10, with

$$\mathbb{E}[X|Y] := \mathbb{E}[X|\sigma(Y)]$$

3.4.3   Example

Let X be a continuous random variable with the following PDF

$$f_X(x) = \begin{cases} 2x & \text{if } 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$$

We also know that

$$f_{Y|X}(y|x) = \begin{cases} 2xy - x + 1 & \text{if } 0 \le y \le 1 \\ 0 & \text{otherwise} \end{cases}$$

Find the MMSE estimate of X, given $Y = y$ is observed.

Solution

We need to find $\mathbb{E}[X|Y]$ with the density function :

$$f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x)f_X(x)}{f_Y(y)}$$

Finding $f_Y(y)$:

$$f_Y(y) = \int_0^1 f_{Y|X}(y|x)f_X(x)dx$$

$$= \int_0^1 (2xy - x + 1)2xdx$$

$$= \int_0^1 (4x^2y - 2x^2 + 2x)dx$$

$$= y\frac{4x^3}{3} - \frac{2x^3}{3} + x^2\Big|_0^1 = \frac{4y}{3} + \frac{1}{3}$$

Hence:

$$f_{X|Y}(x|y) = \frac{(2xy - x + 1)2x}{\frac{4y}{3} + \frac{1}{3}}$$

MMSE :

$$\mathbb{E}[X|Y=y] = \int_0^1 \frac{(2xy - x + 1)2x}{\frac{4y}{3} + \frac{1}{3}}xdx$$

$$= \frac{3y + \frac{1}{2}}{4y + 1}$$

3.4.4 Mean of MMSE estimator

Let $\hat{X}_M$ is MMSE estimator of poterior $X|Y$, which means:

$$\hat{X}_M = \mathbb{E}[X|Y]$$

Then by the law of iterated expectation (in sec 2.11), we have :

$$\mathbb{E}[\hat{X}_M] = \mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[X]$$

3.4.5 Lemma

Let $\hat{X}_M = \mathbb{E}[X|Y]$ be the MMSE estimator of X given Y and let $\tilde{X} = X - \hat{X}_M$ be the estimation error, then we have :

1. $W = \mathbb{E}[\tilde{X}|Y] = 0$

2. For any function $g(Y)$, we have $\mathbb{E}[\tilde{X}g(Y)] = 0$

3. $Cov(\tilde{X}, \hat{X}_M) = 0$

4. $Var(X) = Var(\hat{X}_M) + Var(\tilde{X})$

$$\begin{aligned}
\mathbb{E}[\tilde{X}|Y] &= \mathbb{E}[X - \hat{X}_M|Y] \\
&= \mathbb{E}[X|Y] - \mathbb{E}[\hat{X}_M|Y] \\
&= \hat{X}_M - \mathbb{E}[\hat{X}_M] \\
&= \hat{X}_M - \hat{X}_M \\
&= 0
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}[\tilde{X}g(Y)] &= \mathbb{E}[(X - \hat{X}_M)g(Y)] \\
&= \mathbb{E}[Xg(Y)] - \mathbb{E}[\hat{X}_Mg(Y)] \\
&= g(Y)\mathbb{E}[X] - \mathbb{E}[\hat{X}_M]g(Y) \quad \mathbb{E} \text{ in term of } X \\
&= 0 \quad \text{Since } \mathbb{E}[\hat{X}_M] = \mathbb{E}[X]
\end{aligned}$$

$$\begin{aligned}
Cov(\tilde{X}, \hat{X}_M) &= \mathbb{E}[\tilde{X}\hat{X}_M] - \mathbb{E}[\tilde{X}]\mathbb{E}[\hat{X}_M] \\
&= \mathbb{E}[\tilde{X}\hat{X}_M] \quad (\mathbb{E}[\tilde{X}] = 0) \\
&= \mathbb{E}[\tilde{X}g(y)] \quad (\text{As } \hat{X}_M \text{ is function of } y) \\
&= 0
\end{aligned}$$

$$Var(X) = Var(\tilde{X} + \hat{X}_M)$$
$$= Var(\tilde{X}) + Var(\hat{X}_M) + 2Cov(\tilde{X}, \hat{X}_M)$$
$$= Var(\tilde{X}) + Var(\hat{X}_M)$$

This infers that:

$$\mathbb{E}[X^2] - \mathbb{E}[X]^2 = \mathbb{E}[\tilde{X}^2] - \mathbb{E}[\tilde{X}]^2 + \mathbb{E}[\hat{X}_M^2] - \mathbb{E}[\hat{X}_M]^2$$
$$\Leftrightarrow \mathbb{E}[X^2] = \mathbb{E}[\tilde{X}^2] + \mathbb{E}[\hat{X}_M^2] \quad \text{Since } \mathbb{E}[X]^2 = \mathbb{E}[\hat{X}_M]^2 \text{ and } \mathbb{E}[\tilde{X}] = 0$$

3.5   Linear MMSE Estimation of Random Variables

We might face some difficulties if we want to use the MMSE in practice.

- $f_{X|Y}(x|y)$ might not be easy to find in some problems

- computing $\mathbb{E}[X|Y = y]$ might not be easy

- the function $g(y) = \mathbb{E}[X|Y = y]$ might have a complicated form

Hence, an alternative solution is to suppose that the estimator of X is a linear function of Y:

$$\hat{X}_L = aY + b$$

Then we find a, b such that :

$$a^*, b^* = \operatorname*{argmin}_{a,b} h(a, b) = \operatorname*{argmin}_{a,b} \mathbb{E}[(X - \hat{X}_L)^2]$$

The solution (a^*, b^*) satisfy these following things :

1. $a^* = \frac{Cov(X,Y)}{Var(Y)}, b^* = \mathbb{E}[X] - a^*\mathbb{E}[Y]$

2. $h(a^*, b^*) = (1 - \rho^2)Var(X)$

3. $\mathbb{E}[(X - a^*Y - b)Y] = 0$ or $\mathbb{E}[\tilde{X}Y] = 0$ (orthogonality principle)

The first thing is obtained by taking the derivative of $h(a, b)$, (as in section 2.11.2) :

$$h(a, b) = \mathbb{E}[(X - aY - b)^2]$$
$$= \mathbb{E}[(X^2 + a^2Y^2 + b^2 - 2aXY - 2bX + 2abY]$$
$$= \mathbb{E}[X^2] + a^2\mathbb{E}[Y^2] + b^2 - 2a\mathbb{E}[XY] - 2b\mathbb{E}[X] + 2ab\mathbb{E}[Y]$$

$$\begin{cases} 2a\mathbb{E}[Y^2] - 2\mathbb{E}[XY] + 2b\mathbb{E}[Y] = 0 \\ 2b - 2\mathbb{E}[X] + 2a\mathbb{E}[Y] = 0 \end{cases} \tag{3.1}$$

$$\begin{cases} a = \frac{Cov(X,Y)}{Var(Y)} \\ b = \mathbb{E}[X] - a\mathbb{E}[Y] \end{cases}$$

Here, we find the same result as in sec 2.11.1.

The second thing:

$$\begin{aligned} h(a^*, b^*) &= \mathbb{E}[(X - a^*Y - b^*)^2] \\ &= Var(X - a^*Y - b^*) \quad \text{Since } \mathbb{E}[X - a^*Y - b^*] = 0 \\ &= Var(X) + (a^*)^2 Var(Y) - 2a^* Cov(X,Y) \\ &= Var(X) + \frac{Cov(X,Y)^2}{Var(Y)^2} Var(Y) - 2\frac{Cov(X,Y)}{Var(Y)} Cov(X,Y) \\ &= Var(X) - \frac{Cov(X,Y)^2}{Var(Y)} \\ &= (1 - \rho^2) Var(X) \end{aligned}$$

The third thing :

$$\mathbb{E}[(X - a^*Y - b^*)Y] = \mathbb{E}[XY] - a^*\mathbb{E}[Y^2] - b^*\mathbb{E}[Y]$$
$$= 0 \quad \text{as 3.1}$$

3.5.1 Random Vector

This is an extension for random vector instead of random variable. Supppose that we have a random vector $\mathbf{X} = [X_1, \ldots, X_m]^T$ that we want to model by a linear relation of $\mathbf{Y} = [Y_1, \ldots, Y_n]^T$ by :

$$\mathbf{X} = \mathbf{AY} + \mathbf{b}$$

where $\mathbf{A} \in \mathbb{R}^{m \times n}, \mathbf{b} \in \mathbb{R}^n$.

Remind that in the case of random variable X and Y, the solution is:

$$\hat{X}_L = \frac{Cov(X,Y)}{Var(Y)}(Y - \mathbb{E}[Y]) + \mathbb{E}[X]$$

By analogy, in the case of random vector:

$$\hat{\mathbf{X}}_L = Cov(\mathbf{X}, \mathbf{Y})Cov(\mathbf{Y}, \mathbf{Y})^{-1}(\mathbf{Y} - \mathbb{E}[\mathbf{Y}]) + \mathbb{E}[\mathbf{X}]$$

and

$$\mathbb{E}[\tilde{\mathbf{X}}] = 0$$
$$Cov(\tilde{\mathbf{X}}, Y_j) = \mathbb{E}[\tilde{\mathbf{X}}Y_j] = 0$$

where $\tilde{\mathbf{X}} = \mathbf{X} - \hat{\mathbf{X}}_L$.

The first condition gives us m equations and the second give us $m \times n$ equations, as the same as the number of variable in $\mathbf{A}$ and $\mathbf{b}$, which is $m \times (n+1)$. Linear MMSE for random vector is equivalent to OLS in sec 2.11.4.

3.5.1.1   Example

Let X be an unobserved random variable with $\mathbb{E}[X] = 0, Var(X) = 4$. Assume that we have observed Y_1 and Y_2 given by

$$Y_1 = X + W_1$$
$$Y_2 = X + W_2$$

where $\mathbb{E}[W_1] = \mathbb{E}[W_2] = 0, Var(W_1) = 1$, and $Var(W_2) = 4$. Assume that W_1, W_2 , and X are independent random variables. Find the linear MMSE estimator of X, given Y_1 and Y_2.

Solution 1

We will find directly $\mathbf{A}$ and $\mathbf{b}$

$$\mathbf{A} = Cov(\mathbf{X}, \mathbf{Y})Cov(\mathbf{Y}, \mathbf{Y})^{-1}$$
$$= \begin{bmatrix} Cov(X, Y_1) & Cov(X, Y_2) \end{bmatrix} \begin{bmatrix} Cov(Y_1, Y_1) & Cov(Y_1, Y_2) \\ Cov(Y_2, Y_1) & Cov(Y_2, Y_2) \end{bmatrix}$$

$$Cov(X, Y_1) = Cov(X, X + W_1) = Cov(X, X) = Var(X) = 4$$

$$Cov(X, Y_2) = Cov(X, X + W_2) = Cov(X, X) = Var(X) = 4$$

$$Cov(Y_1, Y_1) = Cov(X+W_1, X+W_1) = Cov(X, X)+Cov(W_1, W_1) = Var(X)+Var(W_1) = 5$$

$$Cov(Y_1, Y_2) = Cov(X + W_1, X + W_2) = Cov(X, X) = Var(X) = 4$$

$$Cov(Y_2, Y_2) = Cov(X+W_2, X+W_2) = Cov(X, X)+Cov(W_2, W_2) = Var(X)+Var(W_2) = 8$$

$$\mathbf{A} = \begin{bmatrix} 4 & 4 \end{bmatrix} \begin{bmatrix} 5 & 4 \\ 4 & 8 \end{bmatrix}^{-1} = \begin{bmatrix} 4 & 4 \end{bmatrix} \frac{1}{24} \begin{bmatrix} 8 & -4 \\ -4 & 5 \end{bmatrix}^{-1} = \begin{bmatrix} \frac{2}{3} & \frac{1}{6} \end{bmatrix}$$

$$\mathbf{b} = \mathbb{E}[\mathbf{X}] - \mathbf{A}\mathbb{E}[\mathbf{Y}] = 0$$

Solution 2

By using orthogonality principle. Let $\tilde{X} = aY_1 + bY_2 + c$:

The condition $\mathbb{E}[\tilde{\mathbf{X}}] = 0$ gives us:

$$\mathbb{E}[X - (aY_1 + bY_2 + c)] = 0$$
$$\Leftrightarrow \mathbb{E}[X] - a\mathbb{E}[Y_1] - b\mathbb{E}[Y_2] - c = 0$$
$$\Leftrightarrow c = 0$$

The condition $\mathbb{E}[\tilde{\mathbf{X}}Y_1] = 0$ gives us:

$$\mathbb{E}[(X - aY_1 - bY_2 - c)Y_1] = 0$$
$$\Leftrightarrow \mathbb{E}[(X - a(X + W_1) - b(X + W_2) - c)(X + W_1)] = 0$$
$$\Leftrightarrow \mathbb{E}[X^2] - a\mathbb{E}[X^2] - b\mathbb{E}[X^2] - c\mathbb{E}[X] - a\mathbb{E}[W_1^2] - c\mathbb{E}[W_1] = 0$$
$$\Leftrightarrow 4 - 4a - 4b - a = 0$$
$$\Leftrightarrow 5a + 4b = 4$$

The condition $\mathbb{E}[\tilde{\mathbf{X}}Y_2] = 0$ gives us:

$$\mathbb{E}[(X - aY_1 - bY_2 - c)Y_2] = 0$$
$$\Leftrightarrow \mathbb{E}[(X - a(X + W_1) - b(X + W_2) - c)(X + W_2)] = 0$$
$$\Leftrightarrow \mathbb{E}[X^2] - a\mathbb{E}[X^2] - b\mathbb{E}[X^2] - c\mathbb{E}[X] - b\mathbb{E}[W_2^2] - c\mathbb{E}[W_2] = 0$$
$$\Leftrightarrow 4 - 4a - 4b - 4b = 0$$
$$\Leftrightarrow 4a + 8b = 4$$

Then we can have $a = \frac{2}{3}$ and $b = \frac{1}{6}$

3.6 Bayesian Hypothesis Testing

Suppose that we need to decide between two hypotheses H_0 and H_1. In the Bayesian setting, we assume that we know prior probabilities of H_0 and H_1. That is, we know $P(H_0)$ and $P(H_1)$. We observe the random variable (or the random vector) $Y = y$. Then we compare these two following posteriors $P(H_0|Y = y)$ and $P(H_1|Y = y)$ and select the hypothesis with higher posterior.

The average error probability for a hypothesis test can be written as :

$$P_e = P(choose H_1|H_0)P(H_0) + P(choose H_0|H_1)P(H_1).$$

3.6.1 Example

Suppose that the random variable X is transmitted over a communication channel. Assume that the received signal is given by

$$Y = X + W$$

, where $W \sim \mathcal{N}(0, \sigma^2)$ is independent of X. Suppose that $X = 1$ with probability p, and $X = -1$ with probability $1 - p$. The goal is to decide between $X = 1$

and $X = -1$ by observing the random variable Y, which means to find higher posterior for this problem.

Solution

$$P(H_0|Y = y) = P(X = 1|Y = y)$$
$$= \frac{P(Y = y|X = 1)P(X = 1)}{P(Y = y)}$$
$$= \frac{\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y-1)^2}{2\sigma^2}\right)p}{\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y-1)^2}{2\sigma^2}\right)p + \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y+1)^2}{2\sigma^2}\right)(1-p)}$$

$$P(H_1|Y = y) = P(X = -1|Y = y)$$
$$= \frac{P(Y = y|X = -1)P(X = -1)}{P(Y = y)}$$
$$= \frac{\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y+1)^2}{2\sigma^2}\right)(1-p)}{\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y-1)^2}{2\sigma^2}\right)p + \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y+1)^2}{2\sigma^2}\right)(1-p)}$$

$$P(H_0|Y = y) > P(H_1|Y = y)$$
$$\Leftrightarrow \exp\left(-\frac{(y-1)^2}{2\sigma^2}\right)p > \exp\left(-\frac{(y+1)^2}{2\sigma^2}\right)(1-p)$$
$$\Leftrightarrow \exp\left(\frac{2y}{\sigma^2}\right) > \frac{1-p}{p}$$
$$\Leftrightarrow y > \frac{\sigma^2}{2}\ln\left(\frac{1-p}{p}\right)$$

3.7 Bayesian Interval Estimation

Given a posterior, it's logic that we can have an interval estimation of X, given $Y = y$.

$$P(a \le X \le b|Y = y) = 1 - \alpha$$

where interval $[a, b]$ is said to be a $(1 - \alpha)100\%$ credible interval.

3.7.1 Example

Let X and Y be jointly normal and $X \sim \mathcal{N}(0, 1)$, $Y \sim \mathcal{N}(1, 4)$, and $\rho(X, Y) = 0.5$. Find a 95% credible interval for X, given $Y = 2$ is observed.

Solution

From sec 6.2.1,

$$f_{XY}(x,y)$$
$$=\frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}\exp\left(-\frac{1}{2(1-\rho^2)}\left((\frac{x-\mu_X}{\sigma_X})^2-\frac{2\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y}+(\frac{y-\mu_Y}{\sigma_Y})^2)\right)\right)$$
$$=\frac{1}{4\pi\sqrt{1-0.5^2}}\exp\left(-\frac{1}{2(1-0.5^2)}\left(x^2-\frac{xy}{2}+\frac{y^2}{4}\right)\right)$$

$$f_{X|Y}(x|2)$$
$$=\frac{f_{XY}(x,y)}{f_Y(2)}$$
$$=\frac{\frac{1}{4\pi\sqrt{1-0.5^2}}\exp\left(-\frac{1}{2(1-0.5^2)}\left(x-1\right)^2\right)}{\frac{1}{\sqrt{2\pi}2}\exp\left(-\frac{1}{2}\frac{2^2}{2^2}\right)}$$
$$=\frac{1}{\sqrt{2\pi}\sqrt{1-0.5^2}}\exp\left(-\frac{2}{3}\left(x-1\right)^2+\frac{1}{2}\right)$$
$$=\frac{1}{\sqrt{2\pi}\sqrt{3/4}}\exp\left(-\frac{1}{2}\frac{(x-1/4)^2}{3/4}\right)$$

The above density function is of $\mathcal{N}(,\frac{3}{4}$

Or a faster way is to use directly conditional distribution in sec 6.2.5. It means $X|Y=y$ follows normal distribution with :

$$E(X|Y=y)=\sigma_X\rho\frac{y-\mu_Y}{\sigma_Y}+\mu_X=0.5\frac{2-1}{2}=0.25$$

$$Var(X|Y=y)=(1-\rho^2)\sigma_X^2=\frac{3}{4}$$

Now the problem is just to find the credible (or confidence) interval , such as in sec 2.9.

4 Distribution and hypothesis testing

4.1 Chi-squared distribution

Chi-squared distribution (noted χ^2) with k degrees of freedom is the distribution of a sum of the squares of k independent standard normal random variables.

If $Z_1,...,Z_J$ are independent, standard normal random variables, then the sum of their squares Q is distributed according to the chi-squared distribution with k degrees of freedom.

$$Q = \sum_{i=1}^{J} Z_i^2$$

$$Q \sim \chi^2(J)$$

It can be shown that the Chi-squared distribution, in fact, is a particular case of gamma distribution :

$$Q \sim \Gamma(\frac{n}{2}, \frac{1}{2})$$

4.1.1 Test of adequacy

In this case, we test the adequacy for a **multinomial** distribution. Remind some properties of the multinomial distribution :

- n number of trials

- k number of events

- $p_1, p_2, ..., p_k$ event probabilities and $\sum_{i=1}^{k} p_i = 1$.

- X_i be the random variable that count the number of the event i which happens.

- Support : $X_i \in \{0, 1, ..., n\}$ with $1 \le i \le k$

- PMF : $\frac{n!}{x_1!...x_k!} p_1^{x_1}...p_k^{x_k}$, where $\sum_{i=1}^{k} x_i = n$.

- Mean $\mathbb{E}(X_i) = np_i$

- Variance $Var(X_i) = np_i(1 - p_i)$

- Covariance $Cov(X_i, X_j) = -np_ip_j$

- Let Y_i^l be a indicator random variable :

$$\begin{cases} Y_i^l = 1 \text{ if event } i \text{ happens at trial } l \\ Y_i^l = 0 \text{ if not} \end{cases}$$

Y_i^l follows Bernoulli distribution. We can express X_i as:

$$X_i = Y_i^1 + Y_i^2 + ...Y_i^n$$

On the sidelines : Proof for covariance

$$
\begin{aligned}
Cov(X_i, X_j) &= \mathbb{E}(X_i X_j) - \mathbb{E}(X_i)\mathbb{E}(X_j) \\
&= \mathbb{E}(\sum_k Y_i^l \sum_l Y_j^h) - n^2 p_i p_j \\
&= \mathbb{E}(\sum_{l=h} Y_i^l Y_j^h + \sum_l \sum_{h\neq l} Y_i^l Y_j^h) - n^2 p_i p_j \\
&= \mathbb{E}(\sum_l \sum_{h\neq l} Y_i^l Y_j^h) - n^2 p_i p_j, \quad \text{since } i \neq j \text{ then } Y_i^l Y_j^l = 0 \\
&= \sum_l \mathbb{E}(Y_i^l) \sum_{h\neq l} \mathbb{E}(Y_j^h) - n^2 p_i p_j \\
&= n p_i (n-1) p_j - n^2 p_i p_j \\
&= - n p_i p_j
\end{aligned}
$$

Let's come back the main subject :

Hypothesis null : $(X_i)_{i=1,\ldots,k}$ follow the multinominal distribution. We show that under the hypothesis null, we have :

$$
\sum_{i=1}^{k} \frac{(\hat{N}_i - np_i)^2}{np_i} \sim \chi^2(k-1)
$$

by two methods.

4.1.1.1 First method

From the above hypothesis, we have that Y_i^l is sampled from the iid distribution (Bernouilli) with mean $\mu_{Y_i} = p_i$ and std $\sigma_{Y_i} = p_i(1-p_i)$. If n is sufficiently big, then the theorem central limite says:

$$
Z_i = \frac{\frac{X_i}{n} - \mu_{Y_i}}{\frac{\sigma_{Y_i}}{\sqrt{n}}} = \frac{X_i - n\mu_{Y_i}}{\sqrt{n}\sigma_{Y_i}} = \frac{n\hat{p}_i - np_i}{\sqrt{np_i(1-p_i)}}
$$

Z_i is a standard normal distribution. Since $\sum_{i=1}^{k} X_i = n$, we take only $k-1$ first Z_i to guarantee that $(Z_i)_{i=1,\ldots,k-1}$ are independent. Then

$$
T = \sum_{i=1}^{k-1} Z_i^2 = \sum_{i=1}^{k-1} \frac{(n\hat{p}_i - np_i)^2}{np_i(1-p_i)} = \sum_{i=1}^{k-1} \frac{(\hat{N}_i - np_i)^2}{np_i(1-p_i)} \sim \chi^2(k-1)
$$

Note that we have the following equality. With $\sum_{i=1}^{k} \hat{N}_i = n = n(\sum_{i=1}^{k} p_i)$, then we can rewrite T :

$$
T = \sum_{i=1}^{k} \frac{(\hat{N}_i - np_i)^2}{np_i}
$$

.

Demonstration for $k = 2$:

$$
\begin{aligned}
T &= \frac{(\hat{N}_1 - np_1)^2}{np_1(1 - p_1)} \\
&= \frac{(\hat{N}_1 - np_1)^2}{np_1} + \frac{(\hat{N}_1 - np_1)^2}{n(1 - p_1)} \\
&= \frac{(\hat{N}_1 - np_1)^2}{np_1} + \frac{(\hat{N}_2 - np_2)^2}{np_2}
\end{aligned}
$$

Since $\hat{N}_1 + \hat{N}_2 = np_1 + np_2$.

4.1.1.2 Second method

Let Y be a random variable such that $P(Y = i) = p_i$ and consider the following random vector Z:

$$
\begin{aligned}
Z &= [Z_1, \ldots, Z_k]^T \\
&= \left[\frac{[Y = 1] - p_1}{\sqrt{p_1}}, \frac{[Y = 2] - p_2}{\sqrt{p_2}}, \ldots, \frac{[Y = k] - p_k}{\sqrt{p_k}} \right]^T
\end{aligned}
$$

where:

$$
[Y = j] = \begin{cases} 1 & \text{if } Y = j \\ 0 & \text{otherwise} \end{cases}
$$

We have :

- $\mathbb{E}(Z_i) = 0$

- $\mathrm{Var}(Z_i) = 1 - p_i$

- $\mathrm{Cov}(Z_i, Z_j) = -\sqrt{p_i p_j}, i \neq j$

- Covariance matrix $\boldsymbol{\Sigma} = I_k - \begin{bmatrix} \sqrt{p_1} \\ \vdots \\ \sqrt{p_k} \end{bmatrix} [\sqrt{p_1}, \ldots, \sqrt{p_k}] = I_k - \sqrt{p}\sqrt{p}^T$

Then by applying the central limit theorem (1.1) in the multidimensinal case:

$$
\tilde{Z} = \sqrt{n} \frac{Z^{(i)} + \ldots + Z^{(n)}}{n} \to \mathcal{N}(0, \boldsymbol{\Sigma})
$$

Here, we have two ways to show that $\tilde{Z} \sim \chi^2(k - 1)$

First method : Using Cochran theorem in section 1.2, let's suppose that random vertor $\mathbf{A} \sim \mathcal{N}(0, I_k)$. Let F be the subspace that attachs to $\sqrt{p}$, then $P_F = \sqrt{p}\sqrt{p}^T$. Let $F^\perp$ be the subspace that perpendicular to $\sqrt{p}$, then

$P_{F^\perp} = I_k - \sqrt{p}\sqrt{p}^T$. This is because I_k corresponds to matrix for projecting from $\mathbb{R}^k$ to itself $\mathbb{R}^k$; F and $F^\perp$ are orthogonal, then for getting $P_{F^\perp}$ we just take the subtraction. From $dim(\sqrt{p}\sqrt{p}^T) = 1$, then $dim(P_F) = 1$, this infers $dim(P_{F^\perp}) = k-1$. Hence, we get $P_{F^\perp}\mathbf{A} \sim \mathcal{N}(0, I_k - \sqrt{p}\sqrt{p}^T)$ and $\sum_i P_{F^\perp}\mathbf{A}[i] \sim \chi^2(k-1)$.

Remind that $\tilde{Z}$ has the same distribution as $P_{F^\perp}\mathbf{A}$, then we must have $\sum_i \tilde{Z}[i] \sim \chi^2(k-1)$.

Second method : $\boldsymbol{\Sigma} = I_k - \sqrt{p}\sqrt{p}^T$ has eigenvalues 0 (with eigenvector $V_1 = \sqrt{p}$) and 1 (with $k-1$ eigenvectors $V_2, \ldots, V_k$). We see that V_1 corresponds to subspace F and $V_2, \ldots, V_k$ coresponds to subspace $F^\perp$. Also, since $\boldsymbol{\Sigma}$ is symmetric, we have $V_1 \perp [V_2, \ldots, V_k]$.

Reminding a short proof, given A is symmetric and two different couples (eigenvalue, eigenvector) (λ, x) and (μ, y):

$$\lambda <x, y> = <\lambda x, y> = <Ax, y> = <x, A^T y>$$
$$= <x, Ay> = <x, \mu y> = \mu <x, y>$$

therefore $\lambda <x, y> = \mu <x, y>$, so $<x, y> = 0$

Suppose that $Q = [V_2, \ldots, V_k]$ are orthonormalized (if not can use e.g. Gram–Schmidt). Let's set $\mathbf{B} = Q^T \tilde{Z}$, hence after this linear transformation $\mathbf{B} \sim \mathcal{N}(0, Q^T \boldsymbol{\Sigma} Q)$. This linear transformation of normal distribution can be proved by Moment generating function (sec 5.1). Note that :

$$Q^T \boldsymbol{\Sigma} Q = Q^T(I_k - \sqrt{p}\sqrt{p}^T)Q = Q^T Q - Q^T \sqrt{p}\sqrt{p}^T Q = I_{k-1}$$

Then $\mathbf{B} \sim \mathcal{N}(0, I_{k-1})$. Also,

$$\sum_i \tilde{Z}[i]^2 = \tilde{Z}^T \tilde{Z} = \mathbf{B}^T Q^T Q \mathbf{B} = \sum_i \mathbf{B}[i]^2 \sim \chi^2(k-1)$$

Let's resume to the main, after the analyse covariance for matrix, from $\sum_i \tilde{Z}[i]^2 \sim \chi^2(k-1)$, it means:

$$\sum_{i=1}^k \tilde{Z}[i]^2 = \sum_{i=1}^k \left(\frac{1}{\sqrt{n}} \sum_{j=1}^n \frac{[Y^{(j)} = i] - p_i}{\sqrt{p_i}} \right)^2$$
$$= \sum_{i=1}^k \frac{(n\hat{p}_i - np_i)^2}{np_i}$$
$$= \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}$$
$$\sim \chi^2(k-1)$$

where O_i means the observation for event i and E_i means the true expected value for event i. Note that E_i is not always available, and we maybe need to estimate these values.

4.1.2 Degree of freedom

We need to determine the degree of freedom, in order to use correctly Chi squared distribution. Indeed, it is not necessary to analyse the covariance matrix as before, just using the following :

- Number of normal distributions : J.

- With N_c number of constraints,(e.g. $\sum_i O_i = n$), the degree of freedom : $q = J - N_c$

- In addition, if there are s number unknown parameters (of a distribution to test) need to be estimated : $q = J - N_c - s$

4.1.3 Application numeric

Ex : Check the balance of a dice with 600 throws:

number	1	2	3	4	5	6
effectifs	88	109	107	94	105	97

$$T = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$
$$= \frac{(88 - 100)^2}{100} + \frac{(109 - 100)^2}{100} + \frac{(107 - 100)^2}{100} + \frac{(94 - 100)^2}{100} + \frac{(105 - 100)^2}{100} + \frac{(97 - 100)^2}{100}$$
$$= 3.44$$

$T \sim \chi^2(6 - 1)$ (since there is only one constraint $\sum_i O_i = n$), risk $\alpha = 0.05$: $P(T < 11.07) = 0.95$. Since $3.44 < 11.07$ then we can say the dice is unbiased.

Ex : Check adequacy with a distribution, e.g. Poisson.

number	0	1	2	3	4
effectifs	31	45	16	7	1

4.1.4 Test homogenity

The question here is to ask whether two lists of numbers with the same total effectif n can derive from the same law of probability.

4.1.5 Test independency

This test makes it possible to verify the absence of a statistical link between two variables X and Y. The two are said to be independent when there is no statistical link between them, in other words, knowledge of X does not in any way make it possible to pronounce on Y.

Reminding the independence :

$$P(X = x, Y = y) = P(X = x)P(Y = y)$$

Need to check the property above.

Example : check the independency between X and Y :

X / Y	1	2	3	4	Total
A	50	70	110	60	290
B	60	75	100	50	285
Total	110	145	210	110	575

- Observation $O_{ij} = nP(X = x_i, Y = y_j)$

- Estimated value $E_{ij} = nP(X = x_i)P(Y = y_j)$

$$T = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Therefore T follows χ^2, we need to find out the degree of freedom q for T. X and Y take a fini number of values, I pour X, J pour Y.

- Number of values : IJ

- One constraint : $\sum_{i,j} O_{ij} = 1$.

- For X, there are $(I - 1)$ probabilities $P(X = x_i)$ need to be estimated and the same for Y, there are $(J - 1)$ probabilities $P(Y = y_i)$ need to be estimated.

- Hence, $q = IJ - 1 - (I - 1) - (J - 1) = (I - 1)(J - 1) = 3$.

$$T = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$
$$= \sum \frac{(50 - 575\frac{290}{575}\frac{110}{575})^2}{575\frac{290}{575}\frac{110}{575}}$$
$$= 2.4235$$

4.1.6 Reduced chi-squared statistic

In statistics, the reduced chi-square statistic is used extensively in goodness of fit testing. The reduced chi-squared χ^2_ν is defined as chi-squared χ^2 per degree of freedom:

$$\chi^2_\nu = \frac{\chi^2}{\nu},$$

where chi-squared is a weighted sum of squared deviations:

$$\chi^2 = \sum_i \frac{(O_i - C_i)^2}{\sigma_i^2}$$

with O_i is observation, C_i is reference or mean (true value), σ_i is std of distribution i. The degree of freedom $\nu = n - m$, equals the number of observations n minus the number of fitted parameters m.

Goodness of Fit:

- A reduced chi-squared value close to 1 indicates a good fit.

- A reduced chi-squared value significantly greater than 1 suggests an underfitting model

- A reduced chi-squared value significantly smaller than 1 suggests an overfitting model.

In ordinary least squares, the definition simplifies to:

$$\chi_\nu^2 = \frac{RSS}{\nu}$$

where $RSS = \sum r^2$ is residual sum of squares.

4.2 Student distribution

Let Z random variable of normal distribution, centered and reduced. U is an independent variable compared to Z and follows Chi-squared distribution of k degree of freedom.

$$T = \frac{Z}{\sqrt{U/k}}$$

follows Student distribution with degree of freedom k.

4.2.1 Student distribution in sampling

Let $X_1, ..., X_n$ n random variables multually independent and distributed from normal distribution $\mathcal{N}(\mu, \sigma^2)$.

Empirical mean:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Variance estimated without bias :

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

Consider $\bar{X}$ as a random variable, hence $\bar{X}$ is a normal distribution, $mean(\bar{X}) = \mu$. Then by CTL,

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

is a normal standard distribution.

Replacing σ (true value) by S (estimated):

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} = \frac{\sigma}{S} \frac{(\bar{X} - \mu)}{\sigma/\sqrt{n}}$$

T follows Student distibution of $(n-1)$ degree liberty, because we set :

$$\frac{U}{n-1} = \frac{S^2}{\sigma^2}$$

where

$$U = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\sigma^2}$$

Then

$$T = \frac{Z}{\sqrt{U/(n-1)}}$$

We need to prove :

- Z (sample mean) and U (sample variance) are independent. One can use Basu's theorem (sec 1.3.6)

- $U \sim \chi^2(n-1)$. This can be done by using theorem of Cochran (sec 1.2).

4.2.1.1 Application

To get the bilateral confidence interval for μ, in case of unknown variance. Then with risk α or with confidance $(1-\alpha)$, this bilateral confidence interval is given by :

$$\left[\bar{X} - t_{\alpha/2}^{n-1} \frac{S}{\sqrt{n}} ; \bar{X} + t_{\alpha/2}^{n-1} \frac{S}{\sqrt{n}} \right]$$

where t^{n-1} follow Student distribution with degree of freedom $n-1$.

Ex: Given the table of height for 8 persons, estimate the confident interval lied to mean of a normal distribution which has unknown variance.

i	1	2	3	4	5	6	7	8
x_i	155	160	161	167	171	177	180	181

Mean estimated:

$$\bar{X} = \frac{1}{8}\sum_{i=1}^{8} X_i = 169$$

Variance estimated without bias :

$$S = \frac{1}{7}\sum_{i=1}^{8}(X_i - \bar{X})^2 = 96.85$$

Then:

$$T = \frac{\bar{X} - \mu}{S/\sqrt{8}}$$

follows Student distribution, of degree 7.

4.2.2 Standard error

The standard error (SE) of a statistic (usually an estimate of a parameter) is the standard deviation (std) of its sampling distribution or an estimate of that standard deviation.

4.2.2.1 Example

Let $X_1, \ldots, X_n$ be a random sample from a distribution that has mean μ and std σ. Then sample mean is used to estimate μ by :

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Then the standard error of $\hat{\mu}$, denoted $s.e.(\hat{\mu})$ is $\frac{\sigma}{\sqrt{n}}$.

4.2.3 t-statistic

Let $\hat{\beta}$ be an estimator of parameter β in some statistical model. Then a t-statistic for this parameter is any quantity of the form :

$$t = \frac{\hat{\beta} - \beta_0}{s.e.(\hat{\beta})}$$

where :

- β_0 is a constant, using in hypothesis $H_0 : \beta = \beta_0$.

- where $s.e.$ means standard error (sec 4.2.2).

If $\hat{\beta}$ is an estimator by using ordinary least squares with the classical linear regression model, and if the true value of the parameter β is equal to β_0. Then t follows the Student distribution with $(n - k)$ degrees of freedom, where n is the number of observations, and k is the number of regressors (including the intercept).

In the majority of models, the estimator $\hat{\beta}$ is consistent (sec 2.5) for β and is distributed asymptotically normally (sec 2.6). If the true value of the parameter β is equal to β_0, and the quantity $s.e.(\hat{\beta})$ correctly estimates the asymptotic variance of this estimator, then the t-statistic will asymptotically have the standard normal distribution.

t-statistic is very similar to the z-score ($z = \frac{x-\mu}{\sigma}$) but with the difference that t-statistic is used when the sample size is small or the population standard deviation is unknown. It is also used along with p-value when running hypothesis tests where the p-value tells us what the odds are of the results to have happened.

4.3 Prediction interval

In statistical, specifically predictive inference, a prediction interval is an estimate of an interval in which a **future observation** will fall, with a certain probability, given what has already been observed. Prediction intervals are often used in regression analysis.

4.3.1 Unknown mean, known variance

- Given a normal distribution with unknown mean μ but known variance (normalized to 1). Then the sample mean $\bar{X}$ of the observations $X_1, X_2, ..., X_n$ has distribution $\mathcal{N}(\mu, 1/n)$ (Central limit theorem).

- While the future observation or observation to test X_t has distribution $\mathcal{N}(\mu, 1)$

- We remind that sum of two independent normal distribution follows $\mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$. Then taking the difference of these ($\bar{X}$ and X_t) cancels the μ and yields a normal distribution of variance $1 + 1/n$.

In short :
$$X_t - \bar{X} \sim \mathcal{N}(0, 1 + 1/n)$$

Then the prediction interval for X_t can be infered from $\mathcal{N}(\bar{X}, 1 + 1/n)$.

Notice that this prediction distribution is more conservative (careful) than using the estimated mean $\bar{X}$ and known variance 1, as this uses variance $1 + \frac{1}{n}$, hence yields wider intervals. This is necessary for the desired confidence interval property to hold.

4.3.2 Known mean, unknown variance

- Given a normal distribution with known mean 0 but unknown variance σ^2.

- $\bar{X}$ is the mean of the observations $X_1, X_2, ..., X_n$ follows normal distribution $\mathcal{N}(0, \sigma^2/n)$.

- Future observation or observation to test X_t that follows $\mathcal{N}(0, \sigma^2)$.

- Since $X_1, \ldots, X_n$ are i.i.d., then $\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\sigma^2} \sim \chi_{n-1}^2$.

- Then $\frac{X_t/\sigma}{\sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\sigma^2(n-1)}}} \sim T^{(n-1)}$ or $\frac{X_t}{s_n} \sim T^{(n-1)}$, where $s_n^2 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1}$ (sample variance without bias).

Then the prediction interval for X_t can be infered from $s_n T^{(n-1)}$.

4.3.3 Unknown mean, unknown variance

From previous results :

- $(X_t - \bar{X}) \sim \mathcal{N}\left(0, \sigma^2(1 + \frac{1}{n})\right)$.

- $\sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} - \frac{\bar{X} - \mu}{\sigma}\right)^2 \sim \chi_{n-1}^2$ or $\frac{(n-1)s_n^2}{\sigma^2} \sim \chi_{n-1}^2$

Then :

$$\frac{\frac{X_t - \bar{X}_n}{\sigma\sqrt{1 + \frac{1}{n}}}}{\sqrt{\frac{(n-1)s_n^2}{\sigma^2}/(n-1)}} \sim T^{(n-1)}$$

Or :

$$\frac{X_t - \bar{X}_n}{s_n\sqrt{1 + \frac{1}{n}}} \sim T^{(n-1)}$$

Then the prediction interval for X_t can be infered from $\bar{X}_n + s_n\sqrt{1 + \frac{1}{n}}T^{(n-1)}$.

4.4 F-distribution

Let S_1 and S_2 are independent random variables with chi-square distributions with respective degrees of freedom d_1 and d_2. Then

$$X = \frac{S_1/d_1}{S_2/d_2}$$

follows F-distribution (Fisher–Snedecor distribution), noted $F(d_1, d_2)$.

4.4.1 F-test

1, The hypothesis that the **means** of a given set of normally distributed populations, all having the same standard deviation, are equal. This is perhaps the best-known F-test, and plays an important role in the ANalysis Of VAriance (ANOVA)

2, In OLS model

Several notations :

- SSE : Sum of Squares residual Error

- SSR : Sum of Squares Regression

- SST : Sum of Squares Total

f-statistique

$$f = \frac{SSR/d(SSR)}{SSE/d(SSE)}$$

where $d(SSR) = p$, $d(SSE) = N - p - 1$, with p number of independent variables and N number of observations. Under the hypothesis H0 that model
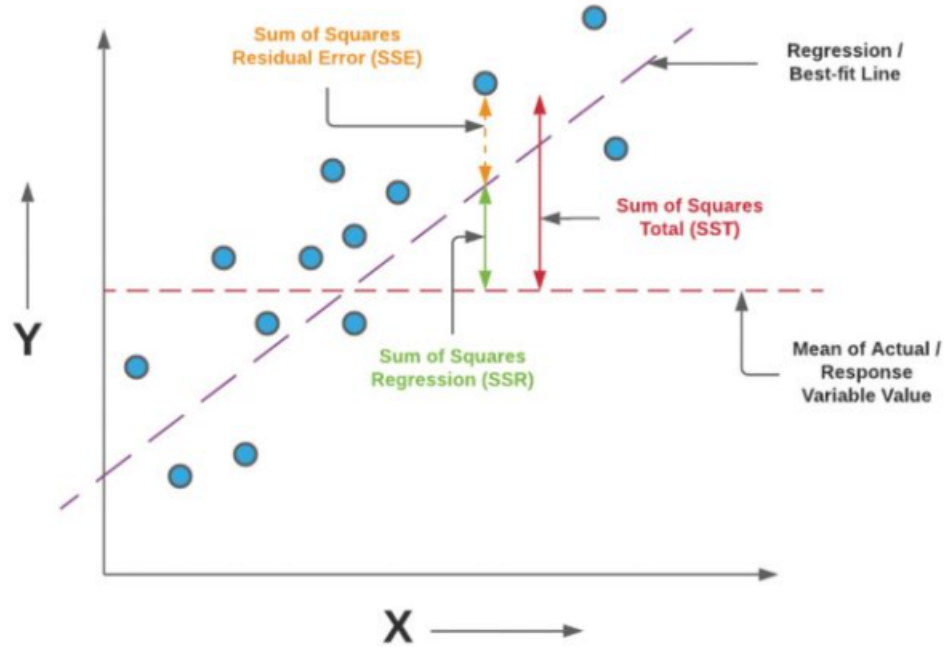
Figure 3.1: Several sum of square.

2 (with intercept and other p independent variables) does not provide a significantly better fit than model 1 (with only intercept), f-statistique follows F-distribution.

Proof : https://stats.stackexchange.com/questions/258461/proof-that-f-statistic-follows-f-distribution
In general case with 2 models :

$$f = \frac{\frac{SSR_1 - SSR_2}{p_2 - p_1}}{\frac{SSR_2}{N - p_2}}$$

4.5   Analysis of variance (ANOVA)

ANOVA, or Analysis of Variance, is a statistical method used to analyze the differences between group means. It is commonly used to determine if there are significant differences between the means of three or more groups.

4.5.1   Example

Let's take an example to illustrate the method. Imagine a breeder who wants to buy new cows for his milk production. He has three different breeds of cows and therefore wonders whether the breed is important for his choice. He has
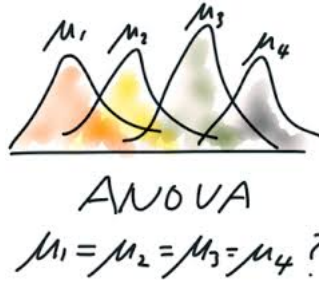
Figure 3.2: Illustration for ANOVA

as information the breed of each of his animals (this is the discrete explanatory variable or variability factor, which can take 3 different values A,B,C) and their daily milk production (this is the continuous variable to be explained, which corresponds to the volume of milk in liters).

In this example, the null hypothesis is that all the cows produce the same daily quantity of milk (to the nearest random factor) whatever the breed. The alternative hypothesis is that one of the breeds produces significantly more or less milk than the others.

A	20.1	19.8	21.3	20.7			
B	22.6	24.1	23.8	22.5	23.4	24.5	22.9
C	31.2	31.6	31.0	32.1	31.4		

$\mu_A = 20.475, \mu_B = 23.4, \mu_C = 31.46, \mu = 25.1875$

The sum of square for inter-classes:

$$SS_{inter} = 4(20.475-25.1875)^2 + 7(23.4-25.1875)^2 + 5(31.46-25.1875)^2 = 307.918$$

The sum of square for intra-class:

$$SS_{intra} = \sum_A (20.1 - 20.475)^2 + \sum_B (22.6 - 23.4)^2 + \sum_C (31.2 - 31.46)^2 = 5.6$$

$$F = \frac{\frac{SS_{inter}}{3-1}}{\frac{SS_{intra}}{4+7+5-3}} = 357.44$$

This F corresponds to p-value 4.338×10^{-12} and if we consider $\alpha = 0.05$, then the null hypothesis is accepted.

4.6 Dickey–Fuller test

4.6.1 Unit root

Given the time serie signal (stochastic process) y_t :

$$y_t = a_1 y_{t-1} + a_2 y_{t-2} + \cdots + a_p y_{t-p} + \varepsilon_t.$$

Then the carateristic polynomial of y_t is given by :

$$x^p - x^{p-1}a_1 - x^{p-2}a_2 - \cdots - a_p = 0$$

We say that this signal has an unit root if $x = 1$ is a root of the carateristic polynomial. A unit root can have a multiplicity $r > 1$.

4.6.1.1 Example

Given :

$$y_t = \alpha y_{t-1} + \varepsilon_t$$

There cases need to be considered:

- $\alpha > 1$: explosive, can check visually by observing t tends to infini. Non-stationary

- $\alpha = 1$: Non-stationary

- $\alpha < 1$: Stationary from an instant t_0.

- $\alpha < 0$ possible but unrealistic in real scenario.

The most subtle case is to check if there is an unit root in the AR model.

4.6.2 Order of integration

The order of integration of a time series is a summary statistic, which reports the minimum number of differences required to obtain a covariance-stationary series.

A time series is integrated of order d, denoted $I(d)$, if

$$(1 - L)^d X_t$$

is a stationary process, where L is the lag operator and $1 - L$ is the difference operator, i.e.

$$(1 - L)X_t = X_t - X_{t-1} = \Delta X$$

In other words, a process is integrated to order d if taking d times difference operator yields a stationary process.

4.6.3 Dickey–Fuller test

In statistics, the Dickey–Fuller test tests the null hypothesis that a unit root (sec 4.6.1) is present in an autoregressive time series model. The alternative hypothesis is different depending on which version of the test is used, but is usually stationarity or trend-stationarity. The test is named after the statisticians David Dickey and Wayne Fuller, who developed it in 1979.

The first model is Test for an unit root :

$$y_t = \rho y_{t-1} + u_t$$

To find ρ, the regression model can be written as :

$$\Delta y_t = (\rho - 1)y_{t-1} + u_t = \delta y_{t-1} + u_t$$

Test if $\delta = 0$ or $\rho = 1$ (unit root). We can use OLS (Ordinary Least Squares), for each time instant t to get δ, then check if mean of δ is 0. Using OLS because it works with hypothese that u_t is white noise.

To check if mean of δ is 0, we use t-statistic. Since the test is done over the residual term $(y_t - y_{t-1})$ rather than raw data, it is not possible to use standard t-distribution to provide critical values, since $(y_t - y_{t-1})$ and $(y_{t-1} - y_{t-2})$ are not independent. Therefore, this t-statistic has a specific distribution simply known as the Dickey–Fuller table.

The two other models in Dickey–Fuller test :

Model 2, Test for an unit root with drift :

$$\Delta y_t = a_0 + (\rho - 1)y_{t-1} + u_t = \delta y_{t-1} + u_t$$

Model 3, Test for an unit root with drift and deterministic time trend :

$$\Delta y_t = a_0 + a_1 t + (\rho - 1)y_{t-1} + u_t = \delta y_{t-1} + u_t$$

4.7 Augmented Dickey–Fuller test

In statistics, an augmented Dickey–Fuller test (ADF) tests the null hypothesis that a unit root is present in a time series sample. The alternative hypothesis is different depending on which version of the test is used, but is usually stationarity or trend-stationarity. It is an augmented version of the Dickey–Fuller test for a larger and more complicated set of time series models.

$$\Delta y_t = \alpha + \beta t + \delta y_{t-1} + \delta_1 \Delta y_{t-1} + ... + \delta_{p-1} \Delta y_{t-p+1} + u_t$$

Return values by funtion adfuller in statsmodels :

- ADF test statistic : $\frac{\hat{\delta}}{s.e.(\hat{\delta})}$

- pvalue (need to check if greater or less than 0.05)

- Number of used lags

- ...

4.8 Kolmogorov–Smirnov test

The Kolmogorov-Smirnov (K-S) test is a statistical test used to determine whether a sample or dataset follows a specific distribution. Its primary role is to assess the goodness of fit between the observed data and a theoretical distribution.

4.8.1 Between dataset and known distribution

Suppose that :

- F is the CDF of the known distribution.

- F_n is the EDF (empirical distribution function), commonly also called an empirical Cumulative Distribution Function (eCDF) from the dataset. Given n is the number of observations of dataset, the EDF is calculated by :
$$F_n = \frac{\text{number of observations less or equal to x}}{n}$$

- The K-S test statistic:
$$D_n = \sup_x |F(x) - F_n(x)|$$

Then with given α and n, we look in K-S test table to get the value $K_{n,\alpha}$ to compare with D_n. If $D_n < K_{n,\alpha}$, we accept the hypothesis H_0 that the dataset follows the known distribution, if not, we reject H_0.

Example Test if the following dataset follows the normality.

68, 72, 75, 77, 79, 81, 82, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 122, 124, 126, 128, 130

From the dataset and the requirement, we infer that the known distribution is normal distribution with $\mu = 100.76$ and $\sigma = 15.39$. Then we calculate F_n, D_n and ....

4.8.2 Between two dataset

In this case the K-S statistic is :
$$D_{n,m} = \sup_x |F_{1,n}(x) - F_{2,m}(x)|$$

and the critic value $K_{m,n,\alpha}$
$$K_{m,n,\alpha} = c(\alpha)\sqrt{\frac{n+m}{mn}}$$

where $c(\alpha)$ is obtained from table :

α	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001
$c(\alpha)$	1.073	1.138	1.224	1.358	1.48	1.628	1.731	1.949

4.9 Jarque–Bera test

In statistics, the Jarque–Bera test is a goodness-of-fit test of whether sample data have the skewness and kurtosis matching a normal distribution.

The test statistic JB is defined as

$$JB = \frac{n}{6}\left(S^2 + \frac{1}{4}(K-3)^2\right)$$

where n is the number of observations (or degrees of freedom in general); S is the sample skewness, K is the sample kurtosis :

$$S = \frac{\hat{\mu}_3}{\hat{\sigma}^3} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{3/2}}$$

$$K = \frac{\hat{\mu}_4}{\hat{\sigma}^4} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2}$$

If the data comes from a normal distribution, the JB statistic asymptotically has a chi-squared (χ_2) distribution with **two** degrees of freedom, so the statistic can be used to test the hypothesis that the data are from a normal distribution.

Chapter 4

Stochastic process

1 Stochastic process

1.1 Definition

A stochastic process is defined as a collection (family) of random variables defined on a common probability space $(\Omega, \mathcal{F}, P)$, indexed par T, output values in state space S, where S is the sample space of measurable space (S, Σ). Reminding that the definition of random variable in sec 1.6, random variable is a measurable function that maps from a measurable space to an other measurable space. Here, it means from $(\Omega, \mathcal{F})$ to (S, Σ).

In other words, for a given probability space $(\Omega, \mathcal{F}, P)$ and a measurable space (S, Σ), a stochastic process is a collection of S-valued random variables, which can be written as

$$\{X(t) : t \in T\}$$

A stochastic process can also be written as a *function* X in term of $t \in T$ and $\omega \in \Omega : X(t, \omega)$.

1.2 Examples

1.2.1 Bernoulli process

One of the simplest stochastic processes, which is a sequence of **independent and identically distributed** random variables $(X_t)_{t=1,\dots,n}$ that follow Bernoulli distribution $Bernoulli(p)$.

Since random variables X_i are independent, we can take the probability space $(\Omega, \mathcal{F}, P)$ where $\Omega = \{H, T\}, \mathcal{F} = \{\varnothing, H, T, \Omega\}, P(\omega = H) = p$ and $P(\omega = T) = 1 - p$. Here we assume that the domain of X is measurable space $(\Omega, \mathcal{F})$ which is related to outcomes of a coin. The state space $S = \{0, 1\}$.

1.2.2 Coin and dice

Let's suppose that we have a coin and a dice :

- Coin $C(\omega) = \begin{cases} 0 \text{ if } \omega = H \\ 1 \text{ if } \omega = T \end{cases}$, where $\Omega_C = \{H, T\}$, $P_C(H) = P_C(T) = 0.5$

- Dice $D(\omega) = \begin{cases} 1 \text{ if } \omega = a \\ 2 \text{ if } \omega = b \\ 3 \text{ if } \omega = c \\ 4 \text{ if } \omega = d \\ 5 \text{ if } \omega = e \\ 6 \text{ if } \omega = f \end{cases}$, where $\Omega_D = \{a, b, c, d, e, f\}$, $P_D(a) = \ldots =$

$P_D(f) = 1/6$.

Let $(C_t)_{t=1,\ldots,n}$ be sequence of iid random variables (of coin) or stochastic process of coin and let $(D_t)_{t=1,\ldots,n}$ be sequence of iid random variables (of dice) or stochastic process of dice. We set up a stochastic process $(X_t)_{t=1,\ldots,n}$:

$$X_t = C_t + D_t$$

The main thing here is that C_t and D_t **do not have the same probability space**. In order to make sense for X_t, it requires that, for a given ω, $C_t(\omega) + D_t(\omega)$ can give a output value. We can deal with this by adapting probability space for C_t and D_t, with a new probability space :

- $\Omega = \Omega_C \times \Omega_D = \{(H, a), \ldots, (T, f)\}$

- $\mathcal{F} = \sigma(\Omega)$, which is thus a smallest σ-algebra on Ω that contains all element in Ω. In this case, this is also the biggest σ-algebra on Ω.

- $P((H, a)) = P_C(H) \times P_D(a)$ ...

Note that, this is also the probability space for X_t. The new random variable C and D are :

- Coin $C(\omega) = \begin{cases} 0 \text{ if } \omega = (H, .) \\ 1 \text{ if } \omega = (T, .) \end{cases}$

- Dice $D(\omega) = \begin{cases} 1 \text{ if } \omega = (., a) \\ 2 \text{ if } \omega = (., b) \\ 3 \text{ if } \omega = (., c) \\ 4 \text{ if } \omega = (., d) \\ 5 \text{ if } \omega = (., e) \\ 6 \text{ if } \omega = (., f) \end{cases}$

where '.' means that it is not independent to value.

The state space of X_t is $\{1, 2, 3, 4, 5, 6, 7\}$.

1.3 Filtration

1.3.1 Starting example

Again, let's suppose that we have a coin and a dice as in sec 1.2.2 :

- Coin $C(\omega) = \begin{cases} 0 \text{ if } \omega = H \\ 1 \text{ if } \omega = T \end{cases}$, where $\Omega_C = \{H, T\}$, $P_C(H) = P_C(T) = 0.5$

- Dice $D(\omega) = \begin{cases} 1 \text{ if } \omega = a \\ 2 \text{ if } \omega = b \\ 3 \text{ if } \omega = c \\ 4 \text{ if } \omega = d \\ 5 \text{ if } \omega = e \\ 6 \text{ if } \omega = f \end{cases}$, where $\Omega_D = \{a, b, c, d, e, f\}$, $P_D(a) = \ldots =$
$P_D(f) = 1/6$.

Now, we want to set up a stochastic process $(X_t)_{t=1,2}$ (only two time index 1 and 2) :

$$\begin{cases} X_1 = C \\ X_2 = X_1 + D \end{cases}$$

This example differs from one in sec 1.2.2 as we must toss the coin before throwing the dice, where in sec 1.2.2, the coin and the dice are executed at the same time. In other words, X_2 is depends on X_1 and in sec 1.2.2, X_t are independent.

From the definition of stochastic (sec 1.1), X_1 and X_2 must be defined on the common probability space. Then the probability space $(\Omega, \mathcal{F}, P)$ for X_1 and X_2 is the same as example in sec 1.2.2.

Let $\omega = (\omega_C, \omega_D)$ be an outcome. The keypoint here is that **at the moment we toss the coin**, we know that the first compenent of outcome ω_C can be only H or T. For the second compenent ω_D, we still have no information about it. In other words, for the coin, we can split (event) $\{H, T\}$ into H and T but still not for the dice. Hence for X_1, its event space can only be :

$$\mathcal{F}_1 = \{\varnothing, \{(H, a), (H, b), (H, c), (H, d), (H, e), (H, f)\},$$
$$\{(T, a), (T, b), (T, c), (T, d), (T, e), (T, f)\}, \Omega\}$$

Note that, $\mathcal{F}_1$ is a σ-algebra and is truely a subset of $\mathcal{F}$.

At the moment we throw the dice, now we can split $\{a, b, c, d, e, f\}$ into each each single element. Hence $\mathcal{F}_2$ is just $\mathcal{F}$ that is equal to $\sigma(\Omega)$.

We call $(\mathcal{F}_t)_{t=1,2}$ is a filtration. In some litterature, we can find that the probability space of a stochastic process is expressed by using filtration instead of event space, such as $(\Omega, \mathcal{F}_n, P)$.

1.3.2 Definition

A family $(\mathcal{F}_t)_{0 \leq t \leq T}$ of $\sigma-$algebra on Ω is called a filtration if $\mathcal{F}_s \subset \mathcal{F}_t$ whenever $s \leq t$. For brevity, we denote $\mathcal{F} = (\mathcal{F}_t)_{0 \leq t \leq T}$.

The name "filtration" is ralated to the fact that the greater t, the more refined event space $\mathcal{F}_t$.

1.3.3 Construction of filtration

Let $(X_n)_{n \in \mathbb{N}}$ be a stochastic process on the probability space $(\Omega, \mathcal{A}, P)$. Then $\mathcal{F} = (F_n)_{n \in \mathbb{N}}$ is a filtration, where :

$$F_n := \sigma(X_k \mid k \leq n) = \sigma(X_1, X_2, \ldots, X_n),$$

which denotes the σ-algebra generated by the **sequence** of random variables $X_1, X_2, \ldots, X_n$. More precisely,

$$\sigma(X_1, X_2, \ldots, X_n) = \bigvee_{i=1}^{n} \sigma(X_i) = \sigma\left(\bigcup_{i=1}^{n} \sigma(X_i)\right),$$

where $a \vee b$ ($\vee$ pronounced vee) is an operator ouputs the smallest σ-algebra that contains a and b. In this document $a \vee b = \sigma(a \cup b)$.

The function $\sigma(X_i)$ is the σ-algebra gerenated by random variable X_i (sec 1.6.4). Generally, $\bigcup_{i=1}^{n} \sigma(X_i)$ is not a σ-algebra. We will see a method to get $\sigma\left(\bigcup_{i=1}^{n} \sigma(X_i)\right)$, let's define Π (π-system in sec 1.2), where :

$$\Pi = \left\{\bigcap_{i=1}^{k} A_i : A_i \in \sigma(X_{j_i}); j_i \in \{1, \ldots, n\}; \forall 1 \leq k \leq n\right\},$$

this notation for Π is may quite complicated, let's explain it (from right to left):

1. Take any number k between 1 and n.

2. Take any k σ-algebra $\sigma(X_{j_1}), \ldots, \sigma(X_{j_k})$ from $\sigma(X_1), \ldots, \sigma(X_n)$.

3. From each chosen σ-algebra $\sigma(X_{j_i})$, select A_i.

4. An element of Π is $\bigcap_{i=1}^{k} A_i$.

Then we have that:

$$\sigma\left(\bigcup_{i=1}^{n} \sigma(X_i)\right) = \sigma(\Pi)$$

Proof

On one hand, take $k = 1$, then $\forall A \in \sigma(X_i)$, we have $A \in \Pi$. This implies $\sigma(X_i) \subset \Pi, \forall i$. The latter implies:

$$\bigcup_{i=1}^{n} \sigma(X_i) \subset \Pi$$

This implies:

$$\sigma\left(\bigcup_{i=1}^{n}\sigma(X_i)\right) \subseteq \sigma(\Pi)$$

On the other hand, given k σ-algebra $\sigma(X_{j_1}),\ldots,\sigma(X_{j_k})$. Let's consider an element in $\Pi : \bigcap_{i=1}^{k} A_i$, where $A_i \in \sigma(X_{j_i})$. We show that this is also an element of $\sigma\left(\bigcup_{i=1}^{n}\sigma(X_i)\right)$.

Before going into details, let's see some common mistakes that we usually fail into. Given that $A_1 \in \Sigma_1$ and $A_2 \in \Sigma_2$, then :

- $(A_1 \cap A_2) \subset (\Sigma_1 \cup \Sigma_2)$: **This is wrong**

- $(A_1 \cup A_2) \subset (\Sigma_1 \cup \Sigma_2)$: **This is wrong**

For example : Given $\Sigma_1 = \{\{a,c\},\{b,d\}\}$ and $\Sigma_2 = \{\{a,d\},\{b,c\}\}$. Then, $\Sigma_1 \cup \Sigma_2 = \{\{a,c\},\{b,d\},\{a,d\},\{b,c\}\}$. In addition, $\Sigma_1 \cap \Sigma_2 = \varnothing$. Let's take $A_1 = \{a,c\}$ and $A_2 = \{b,c\}$, then $A_1 \cap A_2 = \{c\} \not\subset \Sigma_1 \cup \Sigma_2$. Also, $A_1 \cup A_2 = \{a,b,c\} \not\subset \Sigma_1 \cup \Sigma_2$.

Let's return to the proof. With the properties of σ-algebra in sec 1.1 :

$$A_i \in \sigma(X_{j_i})$$
$$A_i^C \in \sigma(X_{j_i})$$
$$A_i^C \in \bigcup_{l=1}^{n}\sigma(X_l)$$
$$A_i^C \in \sigma\left(\bigcup_{l=1}^{n}\sigma(X_l)\right)$$
$$\bigcup_{i=1}^{k} A_i^C \in \sigma\left(\bigcup_{l=1}^{n}\sigma(X_l)\right)$$
$$\left(\bigcup_{i=1}^{k} A_i^C\right)^C \in \sigma\left(\bigcup_{l=1}^{n}\sigma(X_l)\right)$$
$$\bigcap_{i=1}^{k} A_i \in \sigma\left(\bigcup_{l=1}^{n}\sigma(X_l)\right)$$

This means :

$$\Pi \subseteq \sigma\left(\bigcup_{i=1}^{n}\sigma(X_i)\right)$$

This implies:

$$\sigma(\Pi) \subset \sigma\left(\sigma\left(\bigcup_{i=1}^{n}\sigma(X_i)\right)\right) = \sigma\left(\bigcup_{i=1}^{n}\sigma(X_i)\right)$$

188

Finally,

$$\sigma\left(\bigcup_{i=1}^{n}\sigma(X_i)\right) = \sigma(\Pi)$$

1.3.4 Example : coin, cards, dice

This is an extension of example in sec 1.3.1

Let consider the case that we first toss a coin, then we draw a card from a hand $\{J, Q, K\}$, finally we throw a dice. Let

- X_1 be the outcome of the tossed coin

- X_2 be the outcome of the drawed card

- X_3 be the outcome of the threw dice

For particular sample space :

- For the coin $\Omega_C = \{H, T\}$

- For the hand $\Omega_H = \{J, Q, K\}$

- For the dice $\Omega_D = \{1, 2, 3, 4, 5, 6\}$

The sample space $\Omega = \Omega_C \times \Omega_H \times \Omega_D = \{(H, J, 1), \ldots, (H, J, 6), \ldots, (H, K, 6), \ldots, (T, K, 6)\}$ which has 36 elements. **Note that the sample space Ω must be the same for** X_1, X_2, X_3. With $\omega \in \Omega$, where $\omega = (x, y, z)$, we can define :

- $X_1(\omega) = X_1(x, y, z) = x, x \in \{H, T\}$

- $X_2(\omega) = X_2(x, y, z) = y, y \in \{J, Q, K\}$

- $X_3(\omega) = X_3(x, y, z) = z, z \in \{1, 2, 3, 4, 5, 6\}$

Here we use a kind of identify mapping, but it's not mandatory, we can use other mappings. Then the state space $S = \{H, T, J, Q, K, 1, 2, 3, 4, 5, 6\} = \Omega_C \cup \Omega_H \cup \Omega_D$.

In the next step, we discover $\sigma(X_1), \sigma(X_2), \sigma(X_3)$:

$$\begin{aligned}
\sigma(X_1) &= \{A \times \Omega_H \times \Omega_D | A \in \mathcal{P}(\Omega_C)\} \\
&= \{\varnothing \times \{J, Q, K\} \times \{1, 2, 3, 4, 5, 6\}, \\
&\quad \{H\} \times \{J, Q, K\} \times \{1, 2, 3, 4, 5, 6\}, \\
&\quad \{T\} \times \{J, Q, K\} \times \{1, 2, 3, 4, 5, 6\}, \\
&\quad \{H, T\} \times \{J, Q, K\} \times \{1, 2, 3, 4, 5, 6\}\}
\end{aligned}$$

which has 2^2 elements. $\mathcal{P}$ denotes the powerset. More precisely:

- $\varnothing \times \{J, Q, K\} \times \{1, 2, 3, 4, 5, 6\} = \varnothing$. The Cartesian product of the empty set is also the empty set.

- $\{H, T\} \times \{J, Q, K\} \times \{1, 2, 3, 4, 5, 6\} = \Omega$

- $\{H\} \times \{J, Q, K\} \times \{1, 2, 3, 4, 5, 6\}$ is the complement of $\{T\} \times \{J, Q, K\} \times \{1, 2, 3, 4, 5, 6\}$ in Ω and vice-versa.

$$
\begin{aligned}
\sigma(X_2) &= \{\Omega_C \times A \times \Omega_D | A \in \mathcal{P}(\Omega_H)\} \\
&= \{\{H, T\} \times \varnothing \times \{1, 2, 3, 4, 5, 6\}, \\
&\quad \{H, T\} \times \{J\} \times \{1, 2, 3, 4, 5, 6\}, \\
&\quad \ldots \\
&\quad \{H, T\} \times \{J, Q\} \times \{1, 2, 3, 4, 5, 6\}, \\
&\quad \ldots \\
&\quad \{H, T\} \times \{J, Q, K\} \times \{1, 2, 3, 4, 5, 6\}\}
\end{aligned}
$$

which has 2^3 elements.

$$
\begin{aligned}
\sigma(X_3) &= \{\Omega_C \times \Omega_H \times A | A \in \mathcal{P}(\Omega_D)\} \\
&= \{\{H, T\} \times \{J, Q, K\} \times \varnothing, \\
&\quad \{H, T\} \times \{J, Q, K\} \times \{1\}, \\
&\quad \ldots \\
&\quad \{H, T\} \times \{J, Q, K\} \times \{1, 2\}, \\
&\quad \ldots \\
&\quad \{H, T\} \times \{J, Q, K\} \times \{1, 2, 3\}, \\
&\quad \ldots \\
&\quad \{H, T\} \times \{J, Q, K\} \times \{1, 2, 3, 4, 5, 6\}\}
\end{aligned}
$$

which has 2^6 elements.

Before throwing the dice.

$$
F_0 = \{\varnothing, \Omega\}
$$

which has **2** elements (events). Here we only have two possibilites, the event Ω definitely occurs ($P(\Omega) = 1$) or $\varnothing$ definitely does not occur ($P(\varnothing) = 0$). For any other event, we can not measure it.

After tossing the coin and before drawing a card

$$
F_1 = \sigma(X_1)
$$

We see $F_0 \subset F_1$.

After drawing a card and before throwing the dice

$$F_2 = \sigma(X_1, X_2)$$
$$= \sigma\{\sigma(X_1) \cup \sigma(X_2)\}$$
$$= \sigma\{\{A \times \Omega_H \times \Omega_D | A \in \mathcal{P}(\Omega_C)\} \cup \{\Omega_C \times B \times \Omega_D | B \in \mathcal{P}(\Omega_H)\}\}$$
$$= \sigma\{\{A \times \Omega_H | A \in \mathcal{P}(\Omega_C)\} \cup \{\Omega_C \times B | B \in \mathcal{P}(\Omega_H)\}\} \times \Omega_D$$

In the first method, let's extract $U = \{A \times \Omega_H | A \in \mathcal{P}(\Omega_C)\} \cup \{\Omega_C \times B | B \in \mathcal{P}(\Omega_H)\} =$

$$\big\{\varnothing, \{(H,J),(H,Q),(H,K)\}, \{(T,J),(T,Q),(T,K)\}, \Omega_C \times \Omega_H\big\}$$
$$\cup \big\{\varnothing, \{(H,J),(T,J)\}, \{(H,Q),(T,Q)\}, \{(H,K),(T,K)\}, \{(H,J),(T,J),(H,Q),(T,Q)\},$$
$$\{(H,J),(T,J),(H,K),(T,K)\}, \{(H,Q),(T,Q),(H,K),(T,K)\}, \Omega_C \times \Omega_H\big\}$$
$$= \big\{\varnothing, \{(H,J),(H,Q),(H,K)\}, \{(T,J),(T,Q),(T,K)\},$$
$$\{(H,J),(T,J)\}, \{(H,Q),(T,Q)\}, \{(H,K),(T,K)\}, \{(H,J),(T,J),(H,Q),(T,Q)\},$$
$$\{(H,J),(T,J),(H,K),(T,K)\}, \{(H,Q),(T,Q),(H,K),(T,K)\}, \Omega_C \times \Omega_H\big\}$$

Now we need to find $\sigma(U)$, where U is collection of subsets. It also means that find the smallest σ-algebra that contains all events in U as in general, U itself is not a σ-algebra. Then, to get $\sigma(U)$, for each $A, B \in U$, we add $A \cup B$ and $(A \cup B)^C$ to U until having (smallest) $\sigma-$algebra, after that we can remove the symbole σ.

$$F_2 = \sigma(X_1, X_2)$$

$$= \sigma\Big(\{\varnothing, \{(H,J),(H,Q),(H,K)\}, \{(T,J),(T,Q),(T,K)\},$$

$$\{(H,J),(T,J)\}, \{(H,Q),(T,Q)\}, \{(H,K),(T,K)\}, \{(H,J),(T,J),(H,Q),(T,Q)\},$$

$$\{(H,J),(T,J),(H,K),(T,K)\}, \{(H,Q),(T,Q),(H,K),(T,K)\}, \Omega_C \times \Omega_H, \quad \text{(new events start from here)}$$

$$\{(H,J),(H,Q),(H,K),(T,J)\}, \{(T,Q),(T,K)\}, \quad \text{(tuple (4,2))}$$

$$\{(H,J),(T,J),(T,Q),(T,K)\}, \{(H,Q),(H,K)\},$$

$$\{(H,J),(H,Q),(H,K),(T,Q)\}, \{(T,J),(T,K)\},$$

$$\{(H,Q),(T,J),(T,Q),(T,K)\}, \{(H,J),(H,K)\}$$

$$\{(H,J),(H,Q),(H,K),(T,K)\}, \{(T,J),(T,Q)\},$$

$$\{(H,K),(T,J),(T,Q),(T,K)\}, \{(H,J),(H,Q)\},$$

$$\{(H,J),(H,Q),(H,K),(T,J),(T,Q)\}, \{(T,K)\} \quad \text{(tuple (5,1))}$$

$$\{(H,J),(H,Q),(H,K),(T,J),(T,K)\}, \{(T,Q)\}$$

$$\{(H,J),(H,Q),(H,K),(T,Q),(T,K)\}, \{(T,J)\}$$

$$\{(H,Q),(H,K),(T,J),(T,Q),(T,K)\}, \{(H,K)\}$$

$$\{(H,J),(H,K),(T,J),(T,Q),(T,K)\}, \{(H,Q)\}$$

$$\{(H,J),(H,Q),(T,J),(T,Q),(T,K)\}, \{(H,J)\}\}\Big) \times \Omega_D$$

Here we see that there is the elementary set $\{\{(T,K)\}, \{(T,Q)\}, \{(T,J)\}, \{(H,K)\}, \{(H,Q)\}, \{(H,J)\}\}$ is a subset of above set. Then $\sigma(U)$ is just $\sigma(\Omega_C \times \Omega_H)$ (which has 2^6 elements)

In the second method, we use $\sigma(\Pi)$ (in sec 1.3.3), where:

$$\Pi = \{A \cap B | A \in \sigma(X_1), B \in \sigma(X_2)\}$$

Also, we see that there are all elementary set $\{(T,K)\}, \{(T,Q)\}, \{(T,J)\}, \{(H,K)\}, \{(H,Q)\}, \{(H,J)\}$ in Π, then $\sigma(\Pi)$ must be also $\sigma(\Omega_C \times \Omega_H)$. The latter is :

$$F_2 = \sigma(\Omega_C \times \Omega_H)$$

$$= \{A \times \Omega_D | A \in \mathcal{P}(\Omega_C \times \Omega_H)\}$$

$$= \{\varnothing \times \{1,2,3,4,5,6\},$$

$$\{(H,J)\} \times \{1,2,3,4,5,6\},$$

$$\dots$$

$$\{(H,J),(H,Q)\} \times \{1,2,3,4,5,6\},$$

$$\dots$$

$$\{(H,J),(H,Q),\dots,(T,K)\} \times \{1,2,3,4,5,6\}\}$$

We see that $F_1 \subset F_2$.

After throwing the dice

$$F_3 = \sigma(X_1, X_2, X_3) = \sigma(\sigma(X_1) \cup \sigma(X_2) \cup \sigma(X_3))$$

which is the same as

$$F_3 = \sigma(\omega) = \mathcal{P}(\Omega) = \mathcal{P}(\Omega_C \times \Omega_H \times \Omega_D)$$

which has $2^{36} = 68719476736$ elements (events). We see that $F_2 \subset F_3$.

1.3.5   Random walk

In case of 1-dimensional, we start at 0 at the beginning. At each time i, we toss a fair coin, if head (H), we go left one unit and if tail (T), we go right one unit. Let's note $(X_i)_{i=1,\dots,n}$ this stochastic process and we will discover its probability space and state space. By setting Z_i is random variable for tossing coin at time i, then :
$$X_i = Z_1 + \dots + Z_i$$

Note that i varies from 0 to fixed value n.

- Probability space $\Omega = \{H, T\} \times \{H, T\} \times \dots \times \{H, T\} = \{H, T\}^n$

- Event space $\mathcal{F} = \mathcal{P}(\Omega)$ ($\mathcal{P}$ is power set).

- Filtration $F_i = \mathcal{P}(\{H, T\}^i) \times \{H, T\}^{(n-i)}$

- Probability measure, given $\omega \in \Omega$, $P(\omega) = \frac{1}{2^n}$

State space is $\{-n, -(n-1), \dots, -1, 0, 1, \dots, n-1, n\}$

1.4   Stochastic process adpated to a filtration

A stochastic process $(X_t)_{\{0 \le t \le T\}}$ is said to be $\mathbb{F} - adapted$ if the random variable X_t is $\mathcal{F}_t$-measurable (see 1.5.3), with $0 \le t \le T$.

1.5   Generated Filtration

Generated filtration or natural filtration associated to a stochastic process is a filtration which records the "past behaviour" of this stochastic process at each time.

More formally, let Ω be the sample space and let (S, T) be a measurable space for the codomain of stochastic process W. The generated filtration associated to stochastic process W up to instant n is $(\mathcal{F}_i^W)_{0 \le i \le n}$, where :

$$\mathcal{F}_i^W = \sigma\left(W_j^{-1}(A), \forall j \le i, \forall A \in T\right)$$

It also means the smallest σ-algebra on demain Ω that contains all pre-images of T. Sometime, we use the notation :

$$\mathcal{F}_i^W = \sigma\left(W_0, \dots, W_i\right)$$

1.6 Stochastic process adpated to a generated filtration

Let $\mathcal{F}_t^W$ the generated filtration associated to stochastic process W_t at time t. Then a stochastic process X_t is adapted to $\mathcal{F}_t^W$ if X_l is $\mathcal{F}_l^W$-measurable ($\sigma(X_l) \subset \mathcal{F}_l^W$), for $0 \leq l \leq t$. This point is crutial for Itô calculus or stochastic integrals.

1.7 Continuity

1.7.1 Stochastically continuous

A continuous time stochastic process $(X_t)_{\{t \geq 0\}}$ is said to be stochastically continuous if $\forall t \geq 0, \forall \varepsilon > 0$:

$$\lim_{s \to t} P(|X_t - X_s| > \varepsilon) = 0$$

Then stochastically continuous is also call *continuous in probability*.

1.7.2 Continuous stochastic process

A continuous time stochastic process $(X_t)_{\{t \geq 0\}}$ is said to be *continuous* or to have *continuous sample paths* if for every given $\omega \in \mathcal{F}_\infty$, $X(t, \omega)$ is continuous with respect to t.

The we have the same thing for *right-continuous* and *left-continuous*.

1.8 The sameness between two stochastic process

Given a probability space $(\Omega, \mathcal{F}, P)$. Let X_α and Y_α ($\alpha \in I$) be two stochastic processes in this space. We use notation α instead of i or t, in order to not be biased at first glance to discret or continuous case.

1.8.1 Modification

We say that Y is a *modification*, a *version* of X or is M-equivalent to X if for every $\alpha \in I$, we have

$$P(X_\alpha = Y_\alpha) = P(\{\omega \in \Omega : X_\alpha(\omega) = Y_\alpha(\omega)\}) = 1$$

1.8.2 Indistinguishable

We say that X and Y are indistinguishable or D-equivalent if

$$P(X_\alpha = Y_\alpha, \forall \alpha \in I) = P(\{\omega \in \Omega : X_\alpha(\omega) = Y_\alpha(\omega), \forall \alpha \in I\}) = 1$$

1.8.3 Remark

If X_α and Y_α are D-equivalent, then they are M-equivalent. This is trivial by using the above definitions. However, the converse statement is not true. Let's discover by the following classical example.

Suppose that $\Omega = [0,1]$, $\mathcal{F} = \mathcal{B}([0,1])$, which means the Borel sets (sec 1.1.3) of $[0,1]$, P is the uniform probability (i.e., Lebesgue measure, sec 1.7.3.3) on $[0,1]$. I here is also $[0,1]$, we define two continuous stochastic processes X_t and Y_t by :

$$X(t,\omega) = 0$$

$$Y(t,\omega) = \mathbb{1}_{t=\omega}$$

These two processes are M-equivalent but not D-equivalent, in interval I

Proof

First, these two processes are M-equivalent because given $t \in [0,1]$ then :

$$\begin{cases} X(t,\omega) = Y(t,\omega) = 0 & \text{if } \omega \neq t \\ X(t,\omega) \neq Y(t,\omega) & \text{if } \omega = t \end{cases}$$

Therefore on interval $[0,1]$ there is only one point $\omega = t$ that $X(t,\omega) \neq Y(t,\omega)$. This means $P(X(t,.) = Y(t,.)) = 1, \forall t \in [0,1]$.

Second, we show that in general, these two processes are not D-equivalent on any interval $I = [a,b]$, where $0 \leq a < b \leq 1$. Thus

$$P(X(t,.) \neq Y(t,.), \forall t \in I) = P(\{\omega \colon \omega \in [a,b]\}) = b - a > 0$$

This means

$$P(X(t,.) = Y(t,.), \forall t \in I) < 1$$

1.8.4 Lemma

If X_α and Y_α are M-equivalent and I is countable, then X and Y are D-equivalent.

Proof

By using the *countable additivity* property of probability measure.

1.8.5 Lemma

Let X_t and Y_t be right-continuous processes (resp. left-continuous processes) and they are M-equivalent, at each time $t \geq 0$. Then, they are D-equivalent.

Proof

First, by lemma 1.8.4, X_t and Y_t are D-equivalent with $t \in I = Q$, the set of rationals, which is countable.

Then using a partition on R such as $\left\{0, \frac{1}{n}, \ldots, \frac{k}{n}, \ldots, \frac{\infty}{n}\right\}$. We see that all elements are rational. Suppose that X_t and Y_t are right-continuous, then they can be approximated by :

$$X(t,\omega) = \lim_{n\to\infty} X^n(t,\omega) = \lim_{n\to\infty} \sum_{k=0}^{\infty} X(\frac{k}{n},\omega)\mathbb{1}_{[\frac{k}{n},\frac{k+1}{n})]}(t)$$

$$Y(t,\omega) = \lim_{n\to\infty} Y^n(t,\omega) = \lim_{n\to\infty} \sum_{k=0}^{\infty} Y(\frac{k}{n},\omega)\mathbb{1}_{[\frac{k}{n},\frac{k+1}{n})]}(t)$$

Since

- $\frac{k+1}{n} \to \frac{k}{n}$ as $n \to \infty$

- each $t > 0$ can be squeezed by a couple $\frac{k}{n}$ and $\frac{k+1}{n}$

- X_t and Y_t are right-continuous

then $X^n(t,\omega) = Y^n(t,\omega)$ as $n \to \infty$ or $X(t,\omega) = Y(t,\omega)$ for all $\omega \in \Omega$ or they are D-equivalent.

1.9 Jointly measurable process

Given probability space $(\Omega, \mathcal{F}, P)$. The stochastic process $X(\alpha, \omega)$ where $\alpha \in I$, defined by $X\colon I \times \Omega \to \mathbb{R}$. Then X_α is said to be **jointly measurable** if for each $\omega \in \Omega$, $X(.,\omega)\colon I \to \mathbb{R}$ is a measurable function (sec 1.5).

In other words, a stochastic process is measurable if the function

$$X\colon (I \times \Omega, B(I) \otimes \mathcal{F}) \to (\mathbb{R}, B(\mathbb{R}))$$

is measurable. $\otimes$ means the Kronecker product. Thus, we call *jointly measurable* to emphasize that X is with two dependants α and ω.

1.9.1 Progressively measurable process

This is the case where the event space $\mathcal{F}$ is a filtration $F_t, t \geq 0$ (sec 1.3). The process X_t is said to be progressively measurable if at each time t the function

$$X(s,\omega)\colon [0,t] \times \Omega \to \mathbb{R}$$

is $B([0,t]) \otimes F_t$-measurable. This implies that X is F_t-adapted.

1.9.2 Example of non measurable stochastic process

Suppose that for each $t \in [0,1]$, the random variable X_t is uniformly distributed on $[-1,1]$. Suppose further that the collection of random variables $(X_t)_{\{t\geq 0\}}$ is pairwise independent. Then $(X_t)_{\{t\geq 0\}}$ is not jointly measurable process.

Proof

The proof is conducted by contradiction. Formally, X is reprensented by

$$X\colon [0,1] \times \Omega \to [-1,1]$$

Here we define two measure spaces :

- $(\Omega, \mathcal{F}, P)$, we can take, e.g, $\Omega = [0,2], \mathcal{F} = B([0,2])$ and probability measure P corresponds an uniform distribution on $[0,2]$. In this case $X(.,\omega) = \omega - 1$.

- $([0,t], B([0,t]), L)$ is thus a measure space with Lebesgue measure L (sec 1.7.3.3).

Suppose that X is a jointly measurable process, which mean X is $B([0,1]) \otimes \mathcal{F}$-measurable. Then X is also $B([0,t]) \otimes \mathcal{F}$-measurable since $B([0,t]) \subseteq B([0,1]), \forall t \in [0,1]$. By Fubini–Tonelli theorem (sec 2.13), and note that $X(.,.) \in [-1,1]$ then

$$\int_{[0,t]} \int_\Omega |X(s,\omega)| dP(\omega) dL(s)$$

is finite, then

$$Y_t(\omega) = \int_{[0,t]} X_s(\omega) dL(s)$$

$$= \int_{[0,t]} X_s(\omega) ds \quad \text{(Lebesgue measurable function 1.5.3.3)}$$

is $\mathcal{F}$-measurable or $Y_t, \forall t \in [0,1]$ is also a random variable on $(\Omega, \mathcal{F}, P)$. In addition, $Y(t,\omega)$ is continuous with respect to t for a given ω.

For a given t, we have :

$$\mathbb{E}[Y_t^2] = \mathbb{E}\left[\left(\int_0^t X(s,\omega)\, ds\right)^2\right]$$

$$= \mathbb{E}\left[\int_0^t X(s,\omega)\, ds \int_0^t X(r,\omega)\, dr\right]$$

$$= \mathbb{E}\left[\int_0^t \int_0^t X(s,\omega)X(r,\omega)\, dsdr\right]$$

$$= \int_0^t \int_0^t \mathbb{E}[X(s,\omega)X(r,\omega)]\, dsdr$$

$$= \int_0^t \int_0^t \mathbb{E}[X(s,\omega)X(r=s,\omega)]\, dsdr + \int_0^t \int_0^t \mathbb{E}[X(s,\omega)]\mathbb{E}[X(r \neq s,\omega)]\, dsdr$$

$$= \int_0^t \int_0^t \mathbb{E}[X_s^2]\mathbb{1}_{r=s}\, dsdr \quad \text{(since i.i.d of } X_s\text{)}$$

$$= \mathbb{E}[X_0^2] \int_0^t \int_0^t \mathbb{1}_{r=s}\, dsdr \quad \text{(since } \mathbb{E}[X_s^2] \text{ is a constant)}$$

$$= 0 \quad \text{(since area of diagonal line)}$$

From $\mathbb{E}[Y_t^2] = 0$, we infer that $P(Y_t = 0) = 1$ or process Y_t and process constant 0 are M-equivalent. In addition, Y_t is continuous, then by lemma 1.8.5, Y_t and process constant 0 are D-equivalent or $Y_t = 0, \forall t \in [0,1]$ almost surely. Remind that $Y_t(\omega) = \int_0^t X(s,\omega)\, ds$, then $X_t = Y_{t+\Delta} - Y_t = 0$ almost surely. This is contradictory with the fact that X_t is uniformly in $[-1,1]$.

1.9.3 Lemma

If process X is stochastically continuous (in 1.7.1), then X has a jointly measurable version.

1.9.4 Lemma

All right-continuous (left-continuous, 1.7.2) processes are jointly measurable. Example 1.9.2 show that the role of continuity.

1.10 Independent increments

Let $X = (X_t)_{t \in T}$ be a stochastic process with $X_0 = 0$. $(F_t^X)_{t \in T}$ is the natural filtration of X. Then the two following statements are equivalent :

1. X has independent increments, i.e., for all $0 \le s_1 < t_1 \le s_2 < t_2$ we have $(X_{t_1} - X_{s_1}), \ldots, (X_{t_2} - X_{s_2})$ are independent.

2. For every $s < t \in T$, we have $X_t - X_s$ is independent of F_s^X.

From $1 \Rightarrow 2$: We need to prove that $\sigma(X_{t_2} - X_{s_2}) \perp F_{s_2}^X, \forall t_2 > s_2$. We have

$$\sigma(X_{t_2} - X_{s_2}) \perp \sigma(X_{t_1} - X_{s_1}), \forall 0 \le s_1 < t_1 \le s_2 < t_2$$

Let's take $s_1 = 0, t_1 = t$ and note that $X_0 = 0$, hence :

$$\sigma(X_{t_2} - X_{s_2}) \perp \sigma(X_t), \forall 0 < t \le s_2$$

In addtion, we have also $\sigma(X_{t_2} - X_{s_2}) \perp \sigma(X_0)$ since X_0 is a constant. Then

$$\sigma(X_{t_2} - X_{s_2}) \perp \sigma(X_t), \forall 0 \le t \le s_2$$

Then $\sigma(X_{t_2} - X_{s_2}) \perp \sigma\left(\cup_i^n \sigma(X_{a_i})\right), n \to \infty$ and $0 = a_1 < a_2 < \ldots < a_n = s_2$. Note that $\sigma(X_{t_2} - X_{s_2})$ is a σ-algebra then it is a π-system. Also, $\sigma\left(\cup_i^n \sigma(X_{a_i})\right)$ is a π-system. From the property that if π-systems are independent, then their corresponding σ-algebras are also independent (see 1.10.3), we have :

$$\sigma\left(\sigma(X_{t_2} - X_{s_2})\right) \perp \sigma\left(\cup_i^n \sigma(X_{a_i})\right)$$

or

$$\sigma(X_{t_2} - X_{s_2}) \perp \sigma\left(\cup_i^n \sigma(X_{a_i})\right) = \sigma\left((X_t)_{0 \le t \le s_2}\right) = F_{s_2}^X$$

From $2 \Rightarrow 1$: We need to prove that $\sigma(X_{t_2} - X_{s_2}) \perp \sigma(X_{t_1} - X_{s_1})$. We have :

$$\sigma(X_{t_2} - X_{s_2}) \perp F_{s_2}^X$$
$$\Leftrightarrow \sigma(X_{t_2} - X_{s_2}) \perp \sigma\left((X_t)_{0 \le t \le s_2}\right)$$
$$\Rightarrow \sigma(X_{t_2} - X_{s_2}) \perp \sigma(X_{t_1}, X_{s_1}), \forall 0 \le s_1 < t_1 \le s_2 \quad (\text{Since } \sigma(X_{t_1}, X_{s_1}) \subset \sigma\left((X_t)_{0 \le t \le s_2}\right))$$
$$\Rightarrow \sigma(X_{t_2} - X_{s_2}) \perp \sigma(X_{t_1} - X_{s_1}), \forall 0 \le s_1 < t_1 \le s_2 \quad (\text{Since } \sigma(X_{t_1} - X_{s_1}) \subset \sigma(X_{t_1}, X_{s_1}))$$

This means $(X_{t_2} - X_{s_2})$ and $(X_{t_1} - X_{s_1})$ are independent, for all $0 \le s_1 < t_1 \le s_2$.

1.11   Stopping time

A stopping time is a *random variable* whose value is interpreted as the time at which a given stochastic process stops. A stopping time is often defined by a stopping rule, a mechanism for deciding whether to continue or stop a process on the basis of the *present and past events*.

1.11.1   Definition

1.11.1.1   Discrete time

Let τ be a random variable, which is defined on the filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_n)_{n \in \mathbb{N}}, P)$ with values in $\mathbb{N} \cup \{+\infty\}$. Then τ is called a stopping time if the following condition holds:

$$\{\omega | \tau(\omega) \leq n\} \in \mathcal{F}_n, \forall n$$

Intuitively, this condition means that the "decision" of whether to stop at time n must be based only on the information present at time n, not on any future information.

1.11.1.2   General case

Let τ be a random variable, which is defined on the filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t \in T}, P)$ with values in T. In most cases, $T = [0, +\infty)$. Then τ is called a stopping time if the following condition holds:

$$\{\omega | \tau(\omega) \leq t\} \in \mathcal{F}_t^X, \forall t \in T$$

1.11.2   Example

Let's consider you are in a casino and play a game:

- Playing exactly five games corresponds to the stopping time $\tau = 5$, and *it is* a stopping rule.

- Playing until you have the maximum amount ahead you will ever be *is not* a stopping rule and does not provide a stopping time, as it requires information about the future as well as the present and past.

- Playing until you double your money (borrowing if necessary) *is not* a stopping rule, as there is a positive probability that you will never double your money.

- Playing until you either double you money or run out of money *is* a stopping rule, even though there is potentially no limit to the number of games you play, since the probability that you stop in a finite time is 1.

1.11.3 Example

Let's SP X in the above definition be a Brownian motion B :

- Let $a \in \mathbb{R}$. Then $\tau := \inf\{t \geq 0 \mid B_t = a\}$ is a stopping time for Brownian motion, corresponding to the stopping rule: "stop as soon as the Brownian motion hits the value a."

- Another stopping time is given by $\tau := \inf\{t \geq 1 \mid B_s > 0, \forall s \in [t-1, t]\}$. It corresponds to the stopping rule "stop as soon as the Brownian motion has been positive over a contiguous stretch of length 1 time unit."

- In general, if τ_1 and τ_2 are stopping times on the same filtration, then their minimum $\tau_1 \wedge \tau_2$, their maximum $\tau_1 \vee \tau_2$, and their sum $\tau_1 \tau_2$ are also stopping times. (This is not true for differences and products, because these may require "looking into the future" to determine when to stop.)

1.11.4 Stopped process

A stopped process is a stochastic process that is forced to assume the same value after a prescribed time.

1.11.4.1 Definition

Let

- $(\Omega, \mathcal{F}, P)$ be a probability space.

- (S, Σ) be a measurable space.

- $X : T \times \Omega \to S$ be a stochastic process, where T can be discret ($T = \{1, 2, \ldots, +\infty\}$) or continuous ($T = [0, +\infty)$).

- $\tau : \Omega \to T$ be a stopping time, which is adapted to filtration $\{\mathcal{F}_t | t \geq 0\}$ of $\mathcal{F}$.

Then the stopped process X^τ is defined for $t \geq 0$ and $\omega \in \Omega$ by

$$X_t^\tau(\omega) := X_{\min(t, \tau(\omega))}(\omega)$$

1.11.4.2 Example

Let's resume the example in sec 1.3.4. We have that $(\Omega, \mathcal{F}, F_t, P), X_t, T$, and S are defined or determined. Remind that $T = \{1, 2, 3\}$. We define the stopping time τ

$$\begin{cases} \tau(\omega) = 2 \text{ if } \omega = (H, Q, .) \\ \tau(\omega) = 3 \text{ otherwise} \end{cases}$$

Suppose that an outcome is $\omega = (H, Q, 3)$, then $X_1 = H, X_2 = Q, X_3 = Q$, where $X_3 = X_2$.

2 Properties of stochastic process

2.1 Stationary

In probability theory, a stochastic process is said to be (strict-sense) stationary if its statistical properties, such as its mean, variance, and autocorrelation, remain constant over a time length. In other words, a stationary process has the same statistical properties at any point in time, regardless of when those properties are measured.

More formally, a stochastic process $X(t)$ is said to be stationary if, for any set of time instants $\{t_1, t_2, \ldots, t_n\}$ (length n) and any integer k, the joint probability distribution of $X(t_1), X(t_2), \ldots, X(t_n)$ is the same as the joint probability distribution of $X(t_1 + k), X(t_2 + k), \ldots, X(t_n + k)$. This means that the statistical properties of the process do not change over time, and are invariant under time-shifts.

Stationarity is an important concept in probability theory, as it simplifies the analysis of stochastic processes by reducing the number of parameters that need to be estimated.

2.1.1 Weakly stationary

A stochastic process is said to be weakly stationary (also known as weak-sense stationary or wide-sense stationary or WSS) if its mean and autocorrelation are stationary :

$$\mathbb{E}[X(t_1)] = \mathbb{E}[X(t_2)]$$

$$R_X(t_1, t_2) = \mathbb{E}[X(t_1)X(t_2)] = \mathbb{E}[X(t_1 + \Delta)X(t_2 + \Delta)]$$

Sometimes we write :

$$\mathbb{E}[X(t)X(t + \tau)] = R_X(\tau) = \mathbb{E}[X(t + \tau)X(t)] = R_X(-\tau)$$

Hence R_X is an even function.

Example

Consider the random process $\{X(t), t \in \mathbb{R}\}$ defined as $X(t) = cos(t + U)$, where $U \sim Uniform(0, 2\pi)$. Show that $X(t)$ is a weak-sense stationary process.

2.1.2 Jointly Wide-Sense Stationary

Two random processes $\{X(t), t \in \mathbb{R}\}$ and $\{Y(t), t \in \mathbb{R}\}$ are said to be jointly wide-sense stationary if

- $X(t)$ and $Y(t)$ are each weak-sense stationary.

- $R_{XY}(t_1, t_2) = R_{XY}(t_1 - t_2)$

R_{XY} is not an even function as $R_X(t)$ or $R_Y(t)$

2.1.3 Cyclostationary processes

A continuous-time random process $\{X(t), t \in \mathbb{R}\}$ is cyclostationary if there exists a positive real number T such that, for all $t_1, t_2, \ldots, t_r \in \mathbb{R}$, the joint CDF of

$$X(t_1), X(t_2), \ldots, X(t_r)$$

is the same as the joint CDF of

$$X(t_1 + T), X(t_2 + T), \ldots, X(t_r + T).$$

Example

The random process $\{X(t), t \in \mathbb{R}\}$ defined as $X(t) = A\cos(\omega t)$, where A is a random variable, is cyclostationary.

2.1.4 Weak-sense cyclostationary

A continuous-time random process $\{X(t), t \in \mathbb{R}\}$ is weak-sense cyclostationary or wide-sense cyclostationary if there exists a positive real number T such that :

- $\mu_X(t + T) = \mu_X(t)$, for all $t \in \mathbb{R}$

- $R_X(t_1 + T, t_2 + T) = R_X(t_1, t_2)$, for all $t_1, t_2 \in \mathbb{R}$

2.2 Ergodicity

In physics, statistics, econometrics and signal processing, a stochastic process is said to be in an ergodic regime if **the time-averages converge to the ensemble averages**. In this regime, any collection of random samples from a process must represent the average statistical properties of the entire regime. Conversely, a process that is not in ergodic regime is said to be in non-ergodic regime.

In ergodicity, we may need to consider the concepts of *statistics along time* and *statistics along realizations*. In the figure 4.1, related with a continuous SP, the time variable is t (horizontal axis) and the realizations are indexed with A, B, C, D (vertical axis) and X is denoted for SP. This illustrates the meaning of the two axes associated with a SP.

Then **the time-averages converge to the ensemble averages** can be written as:

$$\lim_{T \to +\infty} \frac{1}{T} \int_0^T X_A(s)\mathrm{d}s = E[X(t)]$$

where A is indexed for a realization.
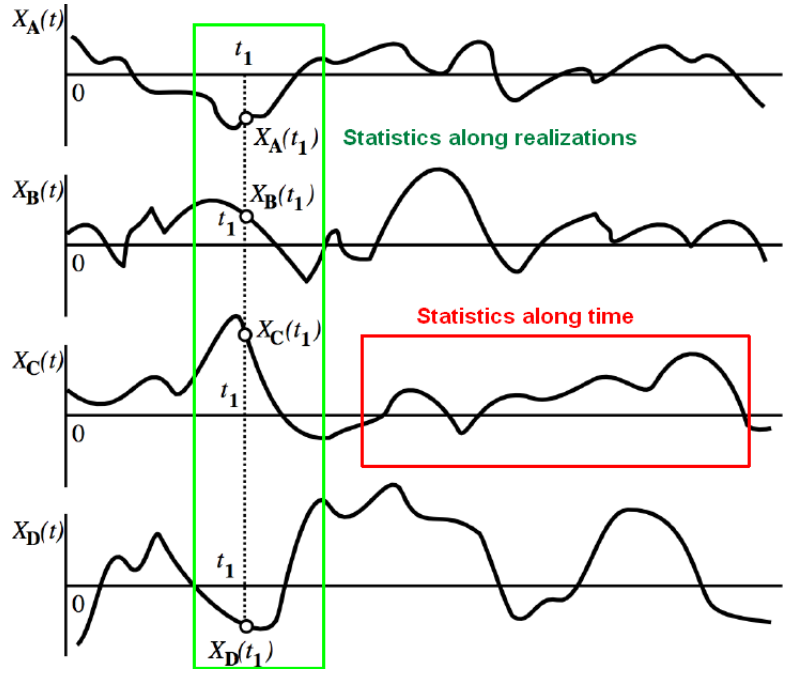
Figure 4.1: Statistics along realization and along time.

2.2.1 Difference with stationary

There are processes that are stationary but not ergodic. Let consider these two following examples:

1, Suppose that we have two coins: one coin has head and tail and the other has two heads. We choose (at random) one of the coins first, and then perform a sequence of independent tosses of our selected coin, with output 1 for heads and 0 for tails. Here are some realizations:

- (Fair) : 1,1,0,1,0,0...

- (Unfair) : 1,1,1,1,1,1,...

- (Fair) : 0,1,1,1,0,0...

On one hand, the long term time-average is either $\frac{1}{2}$ if fair or 1 if unfair. On the other hand, the ensemble average is $\frac{1}{2} \times \frac{1}{2} + \frac{1}{2} \times 1 = \frac{3}{4}$. Hence, this random process is not ergodic in mean, but it is stationary.

2, A little bit the same as the example before, but with 2 dices, one has 6 faces from 1 to 6 and the other has 4 faces from 1 to 4. We choose (at random) one of two dices first, and then perform a sequence of independent throw of our selected dice.

Here are some realizations:

- (D4) : 1,2,4,3,2,1,4...

- (D6) : 2,5,6,3,4,1,...

- ...

On one hand, the long term time-average is either $\frac{5}{2}$ if (D4) or $\frac{7}{2}$ if (D6). On the other hand, the ensemble average is $\frac{1}{2} \times \frac{5}{2} + \frac{1}{2} \times \frac{7}{2} = 3$. In the same manner, this random process is not ergodic in mean, but it is stationary.

2.2.2 Ergodicity implies mean stationarity

From the definition of ergodicity and given two time instant t_1, t_2 and a realization A, we have :

$$\lim_{T \to +\infty} \frac{1}{T} \int_0^T X_A(s)\mathrm{d}s = E[X(t_1)]$$

$$\lim_{T \to +\infty} \frac{1}{T} \int_0^T X_A(s)\mathrm{d}s = E[X(t_2)]$$

This implies:

$$E[X(t_1)] = E[X(t_2)]$$

3 Processing on random variable

3.1 Power Spectral Density

Consider a wide-sense stationary (WSS) random process $X(t)$. Since $\mathbb{E}[X(t)]$ is constant for all instants t, then we exploit rather the autocorrelation $R_X(\tau)$ to know the properties of random process. We define the *Power Spectral Density* (PSD) by the Fourier transform of $R_X(\tau)$:

$$S_X(f) = \mathcal{F}\{R_X(\tau)\} = \int_{-\infty}^{\infty} R_X(\tau)e^{-2j\pi f\tau}d\tau$$

May a question is raised here is that why we use the Fourier transform ? One of possible answer is that we want to capture the periodic of a random process. Measuring directly the periodic (if existing) on one or more realizations of random process may be a hard task, specially in case of low number of realizations. An alternative solution is to use the autocorrelation. Then the Fourier transform of the autocorrelation is for getting the periodic or other processings.

If $X(t)$ is a real-valued random process, then $R_X(\tau)$ is an even, real-valued function of τ, hence:

- $S_X(f) = S_X(-f)$

- $S_X(f) \geq 0$, which means to reconstruct $R_X(\tau)$, there are only positive coefficient in the sum. (Proof in 3.3)

Expected power

$$E[X(t)^2] = R_X(0) = \int_{-\infty}^{\infty} S_X(f)e^{2j\pi f0}df = \int_{-\infty}^{\infty} S_X(f)df$$

Cross Spectral Density (between two SP X and Y:

$$S_{XY}(f) = \mathcal{F}\{R_{XY}(\tau)\} = \int_{-\infty}^{\infty} R_{XY}(\tau)e^{-2j\pi f\tau}d\tau$$

3.2 Linear Time-Invariant (LTI) Systems

We remind the impulse response in signal processing. Denote $X(t), h(t), Y(t)$ repectively input, filter and output :

$$Y(t) = \int_{-\infty}^{\infty} h(\alpha)X(t-\alpha)d\alpha = \int_{-\infty}^{\infty} h(t-\alpha)X(\alpha)d\alpha$$

Image that $h(t)$ is a blackbox and we do not know it. A simple method to $h(t)$ is to apply at input a Dirac pulse :

$$X(t) = \begin{cases} 1 & \text{if } t = 0 \\ 0 & \text{otherwise} \end{cases}$$

Then at output we have $Y(t) = h(t)$, or simply the impulse response.

Now in case of random process, consider an LTI system with impulse response $h(t)$. Let $X(t)$ be a WSS random process. If $X(t)$ is the input of the system, then the output, $Y(t)$, is also a random process. More specifically, we can write:

$$Y(t) = h(t) * X(t) = \int_{-\infty}^{\infty} h(\alpha)X(t-\alpha)d\alpha.$$

Then we have the following properties:

- $Y(t)$ is also WSS.

- $X(t)$ and $Y(t)$ are jointly WSS.

- $\mu_Y(t) = \mu_Y = \mu_X \int_{-\infty}^{\infty} h(\alpha)d\alpha$

- $R_{XY}(\tau) = h(-\tau) * R_X(\tau) = \int_{-\infty}^{\infty} h(-\alpha)R_X(t-\alpha)d\alpha$

- $R_Y(\tau) = h(\tau) * h(-\tau) * R_X(\tau)$

(*)$Y(t)$ is WSS :

First,

$$\mathbb{E}[Y(t)] = \mathbb{E} \int_{-\infty}^{\infty} h(\alpha)X(t-\alpha)d\alpha$$

$$= \int_{-\infty}^{\infty} \mathbb{E}[h(\alpha)X(t-\alpha)]d\alpha$$

$$= \int_{-\infty}^{\infty} h(\alpha)\mathbb{E}[X(t-\alpha)]d\alpha$$

$$= \mu_X \int_{-\infty}^{\infty} h(\alpha)d\alpha$$

$$= \mu_X H(0), \quad H \text{ denotes Fourier tf}$$

$$= \mu_Y$$

Second,

$$R_Y(\tau) = \mathbb{E}[Y(t)Y(t+\tau)]$$

$$= Cov[Y(t), Y(t+\tau)] + \mathbb{E}[Y(t)]\mathbb{E}[Y(t+\tau)]$$

$$= Cov[\int_{-\infty}^{\infty} h(\alpha)X(t-\alpha)d\alpha, \int_{-\infty}^{\infty} h(\alpha)X(t+\tau-\alpha)d\alpha] + \mu_Y^2$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(\alpha)h(\beta)Cov[X(t-\alpha), X(t+\tau-\beta)]d\alpha d\beta + \mu_Y^2$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(\alpha)h(\beta)\left(\mathbb{E}[X(t-\alpha), X(t+\tau-\beta)] - \mathbb{E}[X(t-\alpha)]\mathbb{E}[X(t+\tau-\beta)]\right)d\alpha d\beta + \mu_Y^2$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(\alpha)h(\beta)\left(R_X(\tau+\alpha-\beta) - \mu_X^2\right)d\alpha d\beta + \mu_Y^2$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(\alpha)h(\beta)R_X(\tau+\alpha-\beta)d\alpha d\beta - \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(\alpha)h(\beta)\mu_X^2 d\alpha d\beta + \mu_Y^2$$

$$= \int_{-\infty}^{\infty} h(\alpha)h(\tau+\alpha) * R_X(\tau+\alpha)d\alpha - \int_{-\infty}^{\infty} h(\alpha)d\alpha\mu_X \int_{-\infty}^{\infty} h(\beta)d\beta\mu_X + \mu_Y^2$$

$$= \int_{-\infty}^{\infty} h(\alpha)h(\tau+\alpha) * R_X(\tau+\alpha)d\alpha$$

$$= h(-\tau) * [h(\tau) * R_X(\tau)] \quad \text{Since } h(-t) * g(t) = \int_{-\infty}^{\infty} h(\alpha)g(t+\alpha)d\alpha$$

Hence $Y(t)$ is WSS, Moreover:

$$S_Y(f) = \mathcal{F}\{R_Y(\tau)\}$$

$$= \mathcal{F}\{h(-\tau) * h(\tau) * R_X(\tau)\}$$

$$= H(-f)H(f)S_X(f)$$

$$= \bar{H}(f)H(f)S_X(f)$$

$$= |H(f)|^2 S_X(f)$$

$H(-f) = \overline{H(f)}$ since $h(t)$ is assumed to be a real signal.

(**) $X(t)$ and $Y(t)$ are jointly WSS

$$
\begin{aligned}
R_{XY}(\tau) &= \mathbb{E}[X(t)Y(t-\tau)] \quad \text{(by definition)} \\
&= \mathbb{E}[X(t)\int_{-\infty}^{\infty} h(\alpha)X(t-\tau-\alpha)d\alpha] \\
&= \mathbb{E}[\int_{-\infty}^{\infty} h(\alpha)X(t)X(t-\tau-\alpha)d\alpha] \\
&= \int_{-\infty}^{\infty} h(\alpha)\mathbb{E}[X(t)X(t-\tau-\alpha)]d\alpha \\
&= \int_{-\infty}^{\infty} h(\alpha)R_X(-\tau-\alpha)d\alpha \\
&= h(\tau) * R_X(-\tau) \\
&= h(\tau) * R_X(\tau)
\end{aligned}
$$

Hence $X(t)$ and $Y(t)$ are jointly WSS. Moreover:

$$
\begin{aligned}
S_{XY}(f) &= \mathcal{F}\{R_{XY}(\tau)\} \\
&= \mathcal{F}\{h(\tau) * R_X(\tau)\} \\
&= H(f)S_X(f)
\end{aligned}
$$

3.3 Power in a Frequency Band

Consider the following transfer function H, which is the Fourier transform of a filter h (in sec 3.2) :

$$
H(f) = \begin{cases} 1 & f_1 \leq |f| \leq f_2 \\ 0 & \text{otherwise} \end{cases}
$$

Then the power spectral density of output is:

$$
S_Y(f) = |H(f)|^2 S_X(f) = \begin{cases} S_X(f) & f_1 \leq |f| \leq f_2 \\ 0 & \text{otherwise} \end{cases}
$$

An the expected power :

$$
\begin{aligned}
\mathbb{E}[Y(t)^2] &= \int_{-\infty}^{\infty} S_Y(f) \\
&= \int_{-\infty}^{\infty} S_X(f) \\
&= 2\int_{f_1}^{f_2} S_X(f) \quad \text{Since } S_X(-f) = S_X(f)
\end{aligned}
$$

207

Equivalently, $\mathbb{E}[Y(t)^2]$ here corresponds to the expected power of $X(t)$, taking into account frequencies over the range $f_1 \leq |f| \leq f_2$.

Here note that if $f_1 \rightarrow f_2$, then we must have $S_X(f)$ non-negative since $\mathbb{E}[Y(t)^2] \geq 0$.

3.4 White Noise

The random process $X(t)$ is called a white noise process if its Power Spectral Density function is constant (or flat), denoted by $\frac{N_0}{2}$ by convention :

$$S_X(f) = \frac{N_0}{2}$$

The autocorrelation of white noise is :

$$R_X(\tau) = \mathcal{F}^{-1}\{\frac{N_0}{2}\} = \frac{N_0}{2}\delta(\tau)$$

where δ is a Dirac function.

3.4.1 White Gaussian noise

The random process $X(t)$ is called a **white Gaussian noise** if :

- $X(t)$ is white noise.

- $X(t)$ is Gaussian process (in sec 4), with zero mean $\mu_X = 0$.

- $X(t)$ is a stationary.

3.4.2 Difference between white noise and white Gaussian noise

In case of white noise, we have $R_X(\tau) = \frac{N_0}{2}\delta(\tau)$. This means that $X(t_1)$ and $X(t_2)$ are uncorrelated or white noise is uncorrelated. In case of white Gaussian noise, we have the stronger thing that $X(t_1)$ and $X(t_2)$ are independent (if two gaussian distributions are uncorrelated then they are independent).

4 Gaussian Process

In probability theory and statistics, a Gaussian process X_t is a stochastic process if every finite collection of k random variables

$$X_{t_1}, X_{t_2}, \ldots, X_{t_k}$$

has a multivariate normal distribution, i.e. every finite linear combination of them is normally distributed (sec 6.3).

Note that do not confuse between *stochastic process* and *multivariate random variable* :

- First, in stochastic process, variables must be indexed (ordered).

- Second, in stochastic process, the number of index can tend to infinity, which differs from multivariate random variable that has the fixed number of random variable.

4.1 LTI of a Gaussian Process

Let $X(t)$ be a Gaussian process. If $X(t)$ is the input to an LTI system (in sec 3.2), then the output random process, $Y(t)$, is also a Gaussian process.

$$Y(t) = \int_{-\infty}^{\infty} h(\alpha)X(t - \alpha)d\alpha$$

4.1.1 Illustration

Let's :

$$h(\alpha) = \begin{cases} a \text{ if } \alpha = 0 \\ b \text{ if } \alpha = 1 \\ 0 \text{ otherwise} \end{cases}$$

Then

$$Y(t) = aX(t) + bX(t - 1)$$

Consider $[Y(t_1), \ldots, Y(t_k)]$ that equals to $aX(t_1) + bX(t_1 - 1), \ldots, aX(t_k) + bX(t_k - 1)$. Then any linear combination :

$$\sum_{i=1}^{k} c_i Y(t_i) = \sum_{i=1}^{k} c_i a X(t_i) + \sum_{i=1}^{k} c_i b X(t_i - 1)$$

On the left hand side, we have a normal distribution since it is a linear combination of $2k$ jointly normal random variable. This implies that $Y(t)$ is Gaussian process

4.1.2 LTI of stationary Gaussian process

Moreover, if $X(t)$ is *stationary* Gaussian process, then $Y(t)$ is also a stationary Gaussian process. Note that, for Gaussian processes, wide-sense stationarity implies (strict) stationarity.

4.2 Gaussian Process Regression

Given fixed n, suppose that $X_{t_i}, t_1 < \ldots < t_n$ is a Gaussian process or $(X_{t_1}, \ldots, X_{t_n})$ is standard multivariate normal and :

- $X = (X_{i_1}, \ldots, X_{i_q})$, known variables, where $i_1, \ldots, i_q \in \{t_1, \ldots, t_n\}$. This means we know $[X_{i_1}, \ldots, X_{i_q}] = [y_{i_1}, \ldots, y_{i_q}]$.

- X^* : resting $n - q$ variables need to be predicted (or inference).

- $[X, X^*] \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$

- Covariance matrix $\mathbf{\Sigma} = \begin{bmatrix} \mathbf{K} & \mathbf{K}_* \\ \mathbf{K}_*^T & \mathbf{K}_{**} \end{bmatrix}$ is unknown and need to be estimated before doing inference. In real case, if we need taking into account for noise, which means :

$$y = X + \varepsilon$$

where $\varepsilon \sim \mathcal{N}(\mathbf{0}, \sigma_n^2 I_q)$. Then $[X, X^*] \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \mathbf{K} + \sigma_n^2 I_q & \mathbf{K}_* \\ \mathbf{K}_*^T & \mathbf{K}_{**} \end{bmatrix}\right)$

4.2.1 Hyper-parameter estimation

Kernel or covariance functions, usually, we use the Squared exponential, combined with constant C :

$$K_{SE}(x, x') = C \exp\left(-\frac{(x - x')^2}{2l^2}\right)$$

where C, l are two hyper-parameters need to be estimated. We can suppose that the covariance between two variables X_{t_a} and X_{t_b} is:

$$\mathbf{\Sigma}(X_{t_a}, X_{t_b}) = K_{SE}(t_a, t_b)$$

Considering now only known variable X, as X follows multivariate normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{K} + \sigma_n^2 I_q)$:

$$f_{\mathbf{X}|\Theta}(\mathbf{y}|\theta) = \frac{1}{\sqrt{(2\pi)^q |\mathbf{K} + \sigma_n^2 I_q|}} \exp\left(-\frac{1}{2}\mathbf{y}^T(\mathbf{K} + \sigma_n^2 I_q)^{-1}\mathbf{y}\right)$$

where Θ contains hyper-parameters of kernel C, l in $\mathbf{K}$ and also σ_n. Then the log likehood :

$$\log f_{\mathbf{X}|\Theta}(y|\theta) = -\frac{1}{2}\mathbf{y}^T(\mathbf{K} + \sigma_n^2 I_q)^{-1}\mathbf{y} - \frac{1}{2}\log(|\mathbf{K} + \sigma_n^2 I_q|) - \frac{q}{2}\log(2\pi)$$

As $\log f_{\mathbf{X}|\Theta}(y|\theta)$ is differential, we can use for example the bfgs method to minimize $\log f_{\mathbf{X}|\Theta}(y|\theta)$ and get θ.

4.2.2 Inference

After estimateing θ, we have the explicit form of $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$. In the inference, we use the **conditional distribution** as described in the sec 6.3.5, then *with each test point or prediction X_i^**:

$$f_{X_i^*|\mathbf{X}}(x_i^*|\mathbf{x}) \sim \mathcal{N}(\mu_i^*, \sigma_i^{*2})$$

where :

$$\mu_i^* = \mathbf{k}_i^{*T}[\mathbf{K} + \sigma_n^2 I_q]^{-1}\mathbf{y}$$

210

$$\sigma_i^{*2} = Cov(X_i^*, X_i^*) - \mathbf{k}_i^{*T}[\mathbf{K} + \sigma_n^2 I_q]^{-1}\mathbf{k}_i^*$$

with $\mathbf{k}_i^* = Cov(X_i^*, \mathbf{X})$.

As previously mentioned, the inference is performed for each point X_i^* instead of all points X^* because we want to obtain the variance σ_i^{*2} for each point. The latter value is important for giving a vision on confidence interval of estimated value.

4.2.2.1 Cholesky decomposition trick

Since the complexity for matrix inversion $[\mathbf{K} + \sigma_n^2 I_q]^{-1}$ is $\mathcal{O}(q^3)$. Then in practice, we use the Cholesky decomposition which have $\mathcal{O}(q^2)$ for matrix inversion. Remind that given A symmetric semi positive (Then $A = LL^*$ or LL^T in real case):

$$A\backslash b = A^{-1}b = L^T\backslash(L\backslash b)$$

where the notation $A\backslash b$ denote the vector x that solves $Ax = b$. Then we get two new formulas for mean μ_i^* and variance σ_i^*:

$$\mu_i^* = \mathbf{k}_i^{*T}\alpha$$

$$\sigma_i^* = Cov(X_i^*, X_i^*) - \mathbf{v}^T\mathbf{v}$$

where $\mathbf{K} + \sigma_n^2 I_q = LL^T, \alpha = L^T\backslash(L\backslash\mathbf{y})$ and $\mathbf{v} = L\backslash\mathbf{k}_i^*$.

4.2.3 Multi-features

Now in case that $X_{t_i} = [a_1, \ldots, a_d]$, where d is number of features and a_i is an features. Then we only need to use the new kernel for multi dimension, such as :

$$K_{SE}(x, x') = C\exp\left(-\frac{||x - x'||_2^2}{2l^2}\right)$$

4.2.4 Standardization

It is desirable to standardize the data while learning Gaussian processes regression. There are a number of reasons:

- In common Gaussian processes regression model we suppose that output y has zero mean, so we should mean shift y to match our assumption.

- In case of multi-features, we have different *std* (scale). So, we should standardize features to be in the same scale, then better for estimation θ, if not, for each feature, we can put its scale as a new hyper-parameter.

- Gaussian processes regression is prone to numerical problems as we have to inverse ill-conditioned covariance matrix. To make this problem less severe, you should standardize your data.

4.3 Bayesian optimization

Bayesian optimization (BO) is a powerful method for optimizing hyperparameters in machine learning models. It is particularly useful when the objective function that we are trying to optimize is expensive to evaluate, meaning that each evaluation takes a significant amount of time or computational resources.

The basic idea behind Bayesian optimization is to build a probabilistic model of the objective function, and use this model to decide which hyperparameters to try next. The model is updated as new evaluations are made, so that it becomes increasingly accurate over time.

Here, we can use Gaussian Processes Regression as model, with input is hyperparameters and output is the score of performance.

At each step of the optimization process, Bayesian optimization chooses the next set of hyperparameters to evaluate based on a trade-off between exploration and exploitation. On one hand, it tries to explore new regions of the hyperparameter space where the objective function might be better. On the other hand, it tries to exploit regions that are known to be good based on previous evaluations.

Paper : "A Tutorial on Bayesian Optimization" by Peter I. Frazier

4.3.1 Acquisition function

It means a mathematical function used to guide the search for the optimal solution in an efficient manner. The primary goal of an acquisition function is to balance the trade-off between exploration and exploitation in the optimization process. In other words, given expected value (mean) and uncertainty (variance) at different points in the search space, the acquisition function estimate a potential point that have best performance.

There are several common acquisition functions used in Bayesian optimization:

- Upper confidence bound

- Expected Improvement

- Probability of Improvement

4.3.1.1 Upper confidence bound

Assuming that you want to solve a maximization problem, the upper confidence bound (UCB) can be writed as :

$$a(x) = \mu(x) + \kappa\sigma(x)$$

with κ is a constant for the trade-off. Comparing two points x_1 and x_2:

- If their means μ are the same, then BO will pick the one that has larger σ. This is called **exploration**.

- If their std σ are the same, then BO will pick the one that has larger μ. This is called **exploitation**.

4.3.1.2 Expected improvement

Expected improvement (EI) is a short for expected value of improvement. Let's suppose that after n evaluations at $x_1, \ldots, x_n$, we have $f(x_1) = y_1, \ldots, f(x_n) = y_n$. Let $y^* = \max(y_1, \ldots, y_n)$ Then the objective is to find x_{n+1}, such as:

$$x_{n+1} = \underset{x}{\operatorname{argmax}} \, \mathbb{E}\left[[f(x) - f^*]^+\right],$$

where $[x]^+ = max(0, x)$. This means for a given $f(x)$ with mean and uncertainy, we consider **only the part that is greater** f^*.

Since the set of $f(x)$ and $f(x_1), \ldots, f(x_n)$ forms a multivariate normal distribution, then the conditional distribution

$$f(x)|f(x_1), \ldots, f(x_n)\mathcal{N}(\mu(x), \sigma(x)^2)$$

where mean $\mu(x)$ and std $\sigma(x)$ are calculated as in sec 6.3.5. For each x, let $z_0 = \frac{y^* - \mu(x)}{\sigma(x)}$ be the separated point (in the following integral), we have :

$$\mathbb{E}\left[f(x) - f^*\right]^+\right]$$
$$= \int_{z_0}^{\infty} (\mu(x) + \sigma(x)z - y^*)\phi(z)dz$$
$$= (\mu(x) - y^*)\Phi\left(\frac{\mu(x) - y^*}{\sigma(x)}\right) + \sigma(x)\phi\left(\frac{\mu(x) - y^*}{\sigma(x)}\right)$$

where $\Phi(\cdot)$ and $\phi(\cdot)$ are the CDF and PDF of standard normal distribution $\mathcal{N}(0, 1)$. This calculation is more or less similar to the one in Black-Scholes model (sec 14). Then we need to maximize the above function to find x_{n+1}.

5 Poisson Processes

5.1 Preliminaries

5.1.1 Counting processes

A random process $\{N(t), t \in [0, \infty)\}$ is said to be a counting process if $N(t)$ is the number of events occurred from time 0 up to and including time t. For a counting process, we assume

1. $N(0) = 0$

2. $N(t) \in \{0, 1, 2, \ldots\}$

3. for $0 \leq s < t, N(t) - N(s)$ shows the number of events that occur in the interval $(s, t]$

5.1.2 Independent increments

Let $\{X(t), t \in [0, \infty)\}$ be a continuous-time random process. We say that $X(t)$ has independent increments if, for all $0 \leq t_1 < t_2 < t_3 < \ldots < t_n$, the random variables

$$X(t_2) - X(t_1), X(t_3) - X(t_2), \ldots, X(t_n) - X(t_n - 1)$$

are independent. We have seen the independent increments in sec 1.10.

A counting process has independent increments if the numbers of arrivals in non-overlapping (disjoint) intervals are independent.

5.1.3 Stationary increments

We say that $X(t)$ has stationary increments if, for all $t_2 > t_1 \geq 0$, and all $r > 0$, the two random variables

$$X(t_2) - X(t_1) \text{ and } X(t_2 + r) - X(t_1 + r)$$

have the same distributions, Equivalently, $N(t_2) - N(t_1)$ has the same distribution as $N(t_2 - t_1) - N(0) = N(t_2 - t_1)$.

In other words, the distribution of the difference depends only on the length of the interval $(t_1, t_2]$, and not on the exact location of the interval on the real line.

5.1.4 Poisson distribution

$X \sim Poisson(\mu)$ if its range is $R_X = \{0, 1, 2, 3, \ldots\}$ and its PMF is given by:

$$p(X = k) = e^{-\mu} \frac{\mu^k}{k!}, k \in R_X$$

Here are some useful facts:

1. If $X \sim Poisson(\mu)$, then $\mathbb{E}[X] = \mu$ and $Var[X] = \mu$

2. If $X_i \sim Poisson(\mu_i)$ and X_i are independent, then

$$X_1 + X_2 + \ldots + X_n \sim Poisson(\mu_1 + \mu_2 + \ldots + \mu_n)$$

3. The Poisson distribution can be viewed as the limit of binomial distribution : Let $\mu > 0$ be a fixed real number, $Y_n \sim Binomial\left(n, p = \frac{\mu}{n}\right)$ then the PMF of Y_n converges to a Poisson PMF when $n \to \infty$.

5.2 Definition of the Poisson Process

5.2.1 First definition

Let $\lambda > 0$ be fixed. The counting process $\{N(t), t \in [0, \infty)\}$ is called a Poisson process with rates λ if all the following conditions hold :

1. $N(0) = 0$.

2. $N(t)$ has independent increments.

3. $N(\tau) \sim Poisson(\lambda\tau)$ distribution, hence $N(t)$ has stationary increments.

5.2.2 Second definition

Let $\lambda > 0$ be fixed. The counting process $\{N(t), t \in [0, \infty)\}$ is called a Poisson process with rate λ if all the following conditions hold:

1. $N(0) = 0$

2. $N(t)$ has independent and stationary increments

3. We have if $\Delta \to 0$ that

   - $P(N(\Delta) = 0) = 1 - \lambda\Delta + o(\Delta)$
   - $P(N(\Delta) = 1) = \lambda\Delta + o(\Delta)$
   - $P(N(\Delta) \geq 2) = o(\Delta)$

5.2.3 Equivalence between these two definitions

From the first definition to the second one, using Taylor Series.
Inversely, from the second definition to the first one,
We define:
$$g_k(t) = P(N(t) = k)$$

Then using the property *independent increment* and with $\Delta \to 0$

$g_k(t + \Delta)$
$= P(N(t + \Delta) = k)$

$= P(N(t) = k)P(N(\Delta) = 0) + P(N(t) = k - 1)P(N(\Delta) = 1) + \sum_{i=2}^{k} P(N(t) = k - i)P(N(\Delta) = i)$

$= g_k(t)(1 - \lambda\Delta + o(\Delta)) + g_{k-1}(t)(\lambda\Delta + o(\Delta)) + \sum_{i=2}^{k} g_{k-i}(t)o(\Delta)$

$= g_k(t)(1 - \lambda\Delta) + g_{k-1}(t)\lambda\Delta + o(\Delta)$

Or,

$$g_k(t + \Delta) - g_k(t) = -\lambda\Delta g_k(t) + \lambda\Delta g_{k-1}(t) + o(\Delta)$$

Or,

$$g_k'(t) = -\lambda g_k(t) + \lambda g_{k-1}(t)$$

$$e^{\lambda t} g_k'(t) + e^{\lambda t} \lambda g_k(t) = e^{\lambda t} \lambda g_{k-1}(t)$$

$$\frac{d}{dt}[e^{\lambda t} g_k(t)] = \lambda [e^{\lambda t} g_{k-1}(t)]$$

Or,

$$\frac{d^k}{dt^k}[e^{\lambda t} g_k(t)] = \lambda^k [e^{\lambda t} g_0(t)] \quad (k \geq 1)$$

Now we try to find out $g_0(t)$, indeed :

$$g_0(t) = P(N(t) = 0)$$

$$= \lim_{n \to \infty} \prod_{i=1}^{n} P(N(\frac{t}{n}) = 0) \quad \text{(independent increments)}$$

$$= \lim_{n \to \infty} \prod_{i=1}^{n} \left(1 - \lambda \frac{t}{n} + o(\frac{t}{n}) \right) \quad \text{(Third property of } 2^{nd} \text{ definition)}$$

$$= \lim_{n \to \infty} \left(1 - \lambda \frac{t}{n} \right)^n$$

$$= e^{-\lambda t} \quad \text{(Definition of } e)$$

Hence,

$$\frac{d^k}{dt}[e^{\lambda t} g_k(t)] = \lambda^k \quad (k \geq 1)$$

This implies that :

$$e^{\lambda t} g_k(t) = \lambda^k \left(\frac{t^k}{k!} + \sum_{h=0}^{k-1} c_h \frac{t^h}{h!} \right) \quad (k \geq 1)$$

Note that we have $g_i(0) = P(N(0) = i) = 0$ for $i = 1, \ldots, k$, these k conditions imply that $c_h = 0$ for $h = 0, \ldots, k - 1$. Finally,

$$g_k(t) = e^{-\lambda t} \frac{(\lambda t)^k}{k!}$$

or $N(t) \sim Poisson(\lambda \tau)$

5.3   Arrivals

5.3.1   First arrival and Interarrival Times

The first arrival time, noted X_1, is the time from 0 to the first occurence (arrival).

The interarrival times, noted $X_i, (i > 1)$ is the time between the $(i-1)^{th}$ occurence and i^{th} occurence.

Let $N(t)$ be a Poisson process with rate λ, $\forall i$:

$$P(X_i > t) = P(N(t) = 0)$$
$$= e^{-\lambda t} \frac{(\lambda t)^0}{0!}$$
$$= e^{-\lambda t}$$

We try to find the distribution of X_i, thus the CDF :

$$F_{X_i}(t) = P(X_i \leq t)$$
$$= 1 - P(X_i > t)$$
$$= 1 - e^{-\lambda t}$$

Hence $X_i \sim Exponential(\lambda)$.

5.3.2   Arrival's distribution

Now that we know the distribution of the interarrival times, we can find the distribution of arrival times:

- $T_1 = X_1$

- $T_2 = X_1 + X_2$

- $T_3 = X_1 + X_2 + X_3$

- $\vdots$

More specifically, T_n is the sum of n independent $Exponential(\lambda)$ random variables. If $T_n = X_1 + X_2 + \ldots + X_n$ where the X_i are independent $Exponential(\lambda)$ random variables, then $T_n \sim Gamma(n, \lambda)$ or Erlang distribution. The PDF of T_n :

$$f_{T_n}(t) = \frac{\lambda^n t^{n-1} e^{-\lambda t}}{(n-1)!}$$

5.4   Merging and splitting

5.4.1   Merging Independent Poisson Processes

Let $N_1(t), N_2(t), \ldots, N_m(t)$ be m independent Poisson processes with rates $\lambda_1, \lambda_2, \ldots, \lambda_m$. Let's also

$$N(t) = N_1(t) + N_2(t) + \ldots + N_m(t)$$

Then, $N(t)$ is a Poisson process with rate $\lambda_1 + \lambda_2 + \ldots + \lambda_m$.

5.4.2 Splitting (Thinning) of Poisson Processes

Let $N(t)$ be a Poisson process with rate λ. Here, we divide $N(t)$ to two processes $N_1(t)$ and $N_2(t)$ in the following way: For each arrival, a coin with $P(H) = p$ is tossed. If the coin lands heads up, the arrival is counted for the first process $(N_1(t))$, otherwise it is counted for the second process. The coin tosses are independent and are independent of $N(t)$. Then,

- $N_1(t)$ is a Poisson process with rate λp

- $N_2(t)$ is a Poisson process with rate $\lambda(1-p)$

- $N_1(t)$ and $N_2(t)$ are independent.

5.5 Nonhomogeneous Poisson Processes

Let $N(t)$ be the number of customers arriving at a fast food restaurant by time t. We think that the customers arrive somewhat randomly, so we might want to model $N(t)$ as a Poisson process. However, we notice that this process does not have *stationary increments*. For example, we note that the arrival rate of customers is larger during lunch time compared to, say, 4 p.m. In such scenarios, we might model $N(t)$ as a nonhomogeneous Poisson process. Such a process has all the properties of a Poisson process, except for the fact that **its rate is a function of time**, i.e., $\lambda = \lambda(t)$. We have the following definition :

Let $\lambda(t) : [0, \infty) \to [0, \infty)$ be an integrable function. The counting process $\{N(t), t \in [0, \infty)\}$ is called a nonhomogeneous Poisson process with rate $\lambda(t)$ if all the following conditions hold.

- $N(0) = 0$

- $N(t)$ has independent increments

- for any $t \in [0, \infty)$:

  - $P(N(t + \Delta) - N(t) = 0) = 1 - \lambda(t)\Delta + o(\Delta)$
  - $P(N(t + \Delta) - N(t) = 1) = \lambda(t)\Delta + o(\Delta)$
  - $P(N(t + \Delta) - N(t) \geq 2) = o(\Delta)$

More specifically, we can write

$$N(t + s) - N(t) \sim Poisson\left(\int_t^{s+t} \lambda(\alpha)d\alpha\right)$$

6 Discrete-Time Markov Chains

6.1 Introduction

Consider :

- The discrete random process $\{X_n, n = 0, 1, 2, \ldots\}$

- State space $S \subset \{1, 2, \ldots\}$ and $R_{X_i} \in S$

We say that this process is a Markov chain if

$$P(X_{m+1} = j | X_m = i, X_{m-1} = i_{m-1}, \ldots, X_0 = i_0) = P(X_{m+1} = j | X_m = i) = p_{ij}$$

which means the X_{m+1} depends only on X_m and not other previous random variables. If the number of states is finite, e.g., $|S| = r$, we call it a finite Markov chain. From here, we consider that $S = \{1, 2, \ldots, r\}$.

6.1.1 Transition probability matrix

The probability to pass from state i to j is noted as p_{ij} and they forms a transition probability matrix :

$$P = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1r} \\ p_{21} & p_{22} & \cdots & p_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ p_{r1} & p_{r2} & \cdots & p_{rr} \end{bmatrix}$$

Given that we are in state i, then the next state must be one of the possible states, hence :

$$\sum_{j=1}^{r} p_{ij} = 1$$

or, the sum of each row in the ransition probability matrix must be 1.

6.1.2 Probability distributions

Suppose that we know the probability distribution of X_0 :

$$\pi^{(0)} = [P(X_0 = 1) P(X_0 = 2) \ldots P(X_0 = r)]$$

Then the distribution of X_n is :

$$\pi^{(n)} = [P(X_n = 1) P(X_n = 1) \ldots P(X_n = r)] = \pi^{(0)} P^n$$

Note that it is $\pi^{(0)} P^n$ and **not** $P^n \pi^{(0)T}$.

6.1.3 n-Step Transition Probabilities

It means the probability that after n step, we get the state j from i.

$$p_{ij}^{(n)} = P(X_{n_0+n} = j | X_{n_0} = i)$$

The Chapman-Kolmogorov equation can be written as

$$p_{ij}^{(m+n)} = P(X_{m+n} = j | X_0 = i) = \sum_{k \in S} p_{ik}^{(m)} p_{kj}^{(n)}.$$

6.2 Clustering of state

- We say that state j is **accessible** from state i, written as written $i \rightarrow j$, if $\exists n, p_{ij}^{(n)} > 0$. We assume every state is accessible from itself since $p_{ii}^{(0)} = 1$.

- Two states i and j are said to **communicate**, written as $i \leftrightarrow j$, if they are accessible from each other.

- Two states i and j **belong to the same class** (or group) if and only if $i \leftrightarrow j$.

- Given a Markov chain, it is said to be **irreducible** if all states communicate with each other.

Given a state i in a Markov chain, it is said to be :

- **recurrent** if $f_{ii} = P(\exists n_i \geq 1, X_{n_i} = i | X_0 = i) = 1$. It means just that if we go start from state i, then there is at least one path to return back to i.

- **transient** if $f_{ii} = P(\forall n \geq 1, X_n = i | X_0 = i) < 1$. It means just that, from a certain moment, if we go start from state i, we have no path to return to i.

To bring away :

- Every state is either recurrent or transient.

- If two states are in the same class, either both of them are recurrent, or both of them are transient.

- A class is said to be recurrent (transient) if the states in that class are recurrent (transient).

- If we can show there is a state recurrent (transient) in class, then this class is recurrent (transient)

6.3 Periodicity

The period of a state i is the largest integer d satisfying the following property:

$$d = \begin{cases} p_{ii}^{(n)} > 0 & \text{if } n \equiv 0 \mod d \\ p_{ii}^{(n)} = 0 & \text{if } n \not\equiv 0 \mod d \end{cases}$$

The period of state i is shown by $d(i)$.

- If $d(i) > 1$, we say that state i is periodic.

- If $d(i) = 1$, we say that state i is not periodic or aperiodic.

- If $i \leftrightarrow j$, then $d(i) = d(j)$

Consider a finite irreducible Markov chain X_n, it is said to be aperiodic if:

2/3

1/3

1/2

1/2

1

1/2

1
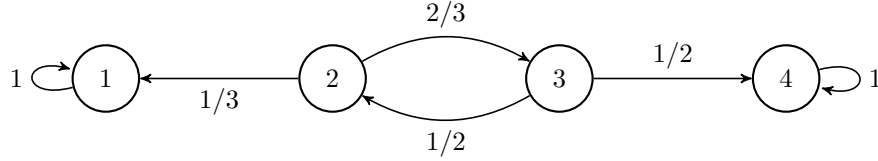
Figure 4.2: A state transition diagram.

- There is a self-transition state in the chain, $p_{ii} > 0$ for some i.

- We can start from state i and come back to state i by two paths, with length repectively m and l and $gcd(m, l) = 1$ (*gcd* means greatest common divisor).

- There exists a positive integer n such that all elements of the matrix P^n are strictly positive, which means $p_{ij}^{(n)} > 0, \forall i, j$.

6.4 Using the law of Total Probability with Recursion

A very useful technique in the analysis of Markov chains is using law of total probability. Here, we will use this technique to find absorption probabilities, mean hitting times, and mean return times.

6.4.1 Absorption Probabilities

The *absorbing* states means that once we enter those states, we never leave them : $p_{ii} = 1$. Note that absorbing state is recurrent state but the inserve is not true. Consider a finite Markov chain $\{X_n, n = 0, 1, 2, \ldots\}$ with state space $S = \{1, 2, \ldots, r\}$. Suppose that all states are either absorbing or transient. Let $l \in S$ be **an absorbing state**. Define :

$$a_{il} = P(\text{absorption in } l | X_0 = i), \quad \text{for all } i \in S$$

By the above definition, we have $a_{ll} = 1$, and $a_{lj} = 0$ if j is any other absorbing state.

To find the unknown values of a_{il}, we can use the following equations (or **law of total probability with recursion**:

$$a_{il} = \sum_k p_{ik} a_{kl}, \quad \text{for } i \in S$$

6.4.1.1 Example

Given the diagram in figure 4.2. At a first look, we can say that $a_{11} = 1, a_{41} = 0$. We will find a_{21} and a_{31} by law of total probability with recursion :

$$\begin{cases} a_{21} = p_{21}a_{11} + p_{22}a_{21} + p_{23}a_{31} + p_{24}a_{41} = \frac{1}{3} + \frac{2}{3}a_{31} \\ a_{31} = p_{31}a_{11} + p_{32}a_{21} + p_{33}a_{31} + p_{34}a_{41} = \frac{1}{2}a_{21} \end{cases}$$

221

Then $a_{21} = \frac{1}{2}$ and $a_{31} = \frac{1}{4}$.

6.4.1.2 Note

If the PDF of $X_0, \pi^{(0)} = [P(X_0 = 1)P(X_0 = 2)\ldots P(X_0 = r)]$ is given, then absorption probabilities of state l compared to other states j, without knowing state at X_0 is:

$$\frac{\sum_i a_{il}P(X_0 = i)}{\sum_j \sum_i a_{ij}P(X_0 = i)}$$

6.4.2 Mean First Hitting Times

Here, we would like to study the expected time until the process hits a state or a set of state for the first time.

Consider a finite Markov chain $\{X_n, n = 0, 1, 2, \ldots\}$ with state space $S = \{1, 2, \ldots, r\}$. Given a state l in S, random variable T_{il} denotes the time that first hits state l or T_{iA} denotes the time that first hits set of state A, both from state i, then the mean first hitting times :

$$t_{il} = \mathbb{E}[T_{il}] \text{ and } t_{iA} = \mathbb{E}[T_{iA}]$$

By the above definition, we have $t_{ll} = 0$ or $t_{lA} = 0$ if $l \in A$. To find the unknown values of t_{il} or t_{iA}, we can use the law of total probability with recursion :

$$t_{il} = 1 + \sum_k p_{ik}t_{kl}, \quad \text{for } i \neq l$$

and

$$t_{iA} = 1 + \sum_k p_{ik}t_{kA}, \quad \text{for } i \notin A,$$

here the unit of time is 1.

6.4.2.1 Example

Let's resume the figure 4.2, we want to the mean first hitting time for the absorbing class $A = \{1, 4\}$. At a first look, we can say that $t_{1A} = 0, t_{4A} = 0$. We will find t_{2A} and t_{3A} by law of total probability with recursion :

$$\begin{cases} t_{2A} = 1 + p_{21}t_{1A} + p_{22}t_{2A} + p_{23}t_{3A} + p_{24}t_{4A} = 1 + \frac{2}{3}t_{3A} \\ t_{3A} = 1 + p_{31}t_{1A} + p_{32}t_{2A} + p_{33}t_{3A} + p_{34}t_{4A} = 1 + \frac{1}{2}t_{2A} \end{cases}$$

Then $t_{2A} = \frac{5}{2}$ and $t_{3A} = \frac{9}{4}$

6.4.2.2 Note

If the PDF of $X_0, \pi^{(0)} = [P(X_0 = 1)P(X_0 = 2)\ldots P(X_0 = r)]$ is given. Then the mean first hitting times A without knowing state at X_0 is:

$$\sum_i t_{iA}P(X_0 = i)$$

6.4.3 Mean First Return Times

Another interesting thing is the first return time. In particular, assuming the chain is in state l, we consider the expected time (number of steps) needed until the chain returns to state l.

Consider a finite Markov chain $\{X_n, n = 0, 1, 2, \ldots\}$ with state space $S = \{1, 2, \ldots, r\}$. Let r_l be the mean return time to state l, then

$$r_l = 1 + \sum_k p_{lk} t_{kl}$$

where t_{kl} is the expected time until the chain hits state l from state k, calculated as in section 6.4.2. Note that $r_l \geq 1, \forall l$ and to have the first return time, usually, we need to find first the first hitting time.

6.4.3.1 Example

If resuming the figure 4.2 :

$$r_2 = 1 + p_{21} t_{12} + p_{22} t_{22} + p_{23} t_{32} + p_{24} t_{42} = 1 + p_{21} t_{12} + p_{23} t_{32} = +\infty$$

since $p_{21} = \frac{1}{3}$ and $t_{12} = +\infty$.

6.5 Limiting and stationary distributions

6.5.1 Limiting Distributions

We would like to discuss long-term behavior of Markov chains. In particular, we would like to know the distribution of X_n as n becomes large. Let P be the transition matrix. More specifically :

$$\lim_{n \to +\infty} [P(X_n = 1) P(X_n = 2) \ldots P(X_n = r)] = \lim_{n \to +\infty} \pi^{(n)} = \lim_{n \to +\infty} \pi^{(0)} P^n$$

When the limiting distribution exists, it does not depend on the initial state X_0.

6.5.1.1 Example

Consider a Markov chain X_n with two possible states, $S = \{0, 1\}$. In particular, suppose that the transition matrix is given by :

$$P = \begin{bmatrix} 1 - a & b \\ a & 1 - b \end{bmatrix},$$

where $0 < a, b < 1$. Find the limiting distribution of X.

Solution

We find first P^n. The matrix P have two eigenvalues, the first $\lambda = 1$ with eigenvector $[1,1]^T$ and $\lambda = 1 - a - b$ with eigenvector $[a, -b]^T$. Then P can be decomposed by :

$$P = \begin{bmatrix} 1 & a \\ 1 & -b \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1-a-b \end{bmatrix} \begin{bmatrix} 1 & a \\ 1 & -b \end{bmatrix}^{-1}$$

Then

$$\begin{aligned}
P^n &= \begin{bmatrix} 1 & a \\ 1 & -b \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & (1-a-b)^n \end{bmatrix} \begin{bmatrix} 1 & a \\ 1 & -b \end{bmatrix}^{-1} \\
&= \begin{bmatrix} 1 & a \\ 1 & -b \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & (1-a-b)^n \end{bmatrix} \begin{bmatrix} b & a \\ 1 & -1 \end{bmatrix} \frac{1}{a+b} \\
&= \frac{1}{a+b} \begin{bmatrix} b + a(1-a-b)^n & a - a(1-a-b)^n \\ b - b(1-a-b)^n & a + b(1-a-b)^n \end{bmatrix}
\end{aligned}$$

As $\lim\limits_{n \to +\infty} (1-a-b)^n = 0$, then $\lim\limits_{n \to +\infty} P^n = \frac{1}{a+b} \begin{bmatrix} b & a \\ b & a \end{bmatrix}$

Finally,

$$\lim_{n \to +\infty} \pi^{(0)} P^n = \begin{bmatrix} \alpha & 1-\alpha \end{bmatrix} \frac{1}{a+b} \begin{bmatrix} b & a \\ b & a \end{bmatrix} = \begin{bmatrix} \frac{b}{a+b} & \frac{a}{a+b} \end{bmatrix}$$

6.5.2 Stationary distribution

We consider Markov chains with a finite number of states. In general, a finite Markov chain can consist of several transient as well as recurrent states. As n becomes large the chain will enter a recurrent class and it will stay there forever. Therefore, when studying long-run behaviors we focus only on the recurrent classes. If a finite Markov chain has more than one recurrent class, then the chain will get absorbed in one of the recurrent classes.

Thus, we can limit our attention to the case where our Markov chain consists of one recurrent class. Consider a finite Markov chain $\{X_n, n = 0, 1, 2, \ldots\}$ with state space $S = \{1, 2, \ldots, r\}$. Assume that the chain is irreducible and aperiodic. Then,

- The set of equations

$$\pi = \pi P$$

$$s.t. \quad \sum_{j \in S} \pi_j = 1$$

has a unique solution, that we called **stationary distribution**, which equals to limiting distribution $\lim\limits_{n \to \infty} \pi^{(n)}$ in sec 6.5.1.

- r_j is the mean return time to state j and it equals to

$$r_j = \frac{1}{\pi_j}, \quad \text{for all } j \in S$$

6.6 Countably Infinite Markov Chains

When a Markov chain has an infinite (but countable) number of states, we need to distinguish between two types of recurrent states: *positive* recurrent and *null* recurrent states.

Let i be a recurrent state. Assuming $X_0 = i$, let R_i be the number of transitions needed to return to state i **in a given path**, i.e.,

$$R_i = \min\{n \geq 1 : X_n = i\}$$
$$r_i = \mathbb{E}[R_i]$$

State i is said to be :

- positive recurrent if $r_i < \infty$

- null recurrent if $r_i = \infty$

6.6.1 Properties

Consider a *infinite* Markov chain $\{X_n, n = 0, 1, 2, \ldots\}$ with state space $S = \{1, 2, \ldots, \infty\}$. Assume that the chain is irreducible and aperiodic. Then, one of the following cases can occur:

1. All states are transient, and

$$\lim_{n \to \infty} P(X_n = j | X_0 = i) = 0, \quad \forall i, j$$

2. All states are null recurrent, and

$$\lim_{n \to \infty} P(X_n = j | X_0 = i) = 0, \quad \forall i, j$$

3. All states are positive recurrent. In this case, there exists a limiting distribution, $\pi = [\pi_1, \ldots, \pi_n]$, where

$$\pi_j = \lim_{n \to \infty} P(X_n = j | X_0 = i) > 0, \quad \forall i, j$$

π is calculated as in 6.5.2.

Note that, prove that all states are positive recurrent can be a hard task, then how do we use the above theorem ? We can try to find a stationary distribution π by solving the equations :

$$\pi = \pi P$$
$$s.t. \quad \sum_{j \in S} \pi_j = 1$$

If the above equations have a unique solution, we conclude that the chain is positive recurrent and the stationary distribution is the limiting distribution of this chain. On the other hand, if no stationary solution exists, we conclude that the chain is either transient or null recurrent.

6.6.2 Example

Consider the following Markov chain :
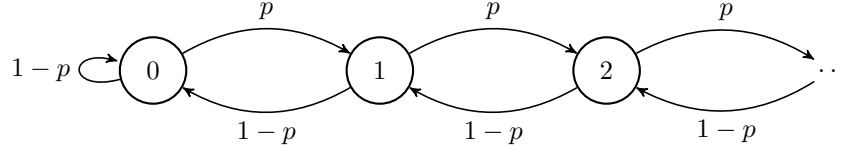
Assume that $0 < p < \frac{1}{2}$. Does this chain have a limiting distribution?

Solution

This chain is irreducible since all states communicate with each other. It is also aperiodic since it includes a self-transition, $P_{00} > 0$. Let's write the equations for a stationary distribution. For state 0, we can write :

$$\pi_0 = (1 - p)\pi_0 + (1 - p)\pi_1,$$

or

$$\pi_1 = \frac{p}{1 - p}\pi_0,$$

For state 1, we can write

$$\pi_1 = p\pi_0 + (1 - p)\pi_2$$
$$= (1 - p)\pi_1 + (1 - p)\pi_2,$$

which results in

$$\pi_2 = \frac{p}{1 - p}\pi_1$$

Similarly, we have

$$\pi_i = \alpha\pi_{i-1} = \ldots = \alpha^i\pi_0$$

where $\alpha = \frac{p}{1-p}$. Since $0 < p < \frac{1}{2}$ then $0 < \alpha < 1$.

Finally, we must have

$$1 = \sum_{i=0}^{+\infty} \pi_i = \sum_{i=0}^{+\infty} \alpha^i\pi_0 = \frac{1}{1 - \alpha}\pi_0$$

This implies:

$$\begin{cases} \pi_0 = 1 - \alpha \\ \pi_i = \alpha^i\pi_0 \end{cases}$$

Since we have found a stationary distribution, we conclude that all states are positive recurrent.

7 Continuous-Time Markov Chains

7.1 Preliminary question

In Discrete-Time Markov Chains, after a unit time $\Delta = 1$ ($X_{n\Delta}$ to $X_{(n+1)\Delta}$), we stay in the same state or jump (move) to other state. Consider T_i the time that we stay in the same state i:

- If $p_{ii} = 1$ then $T_i = \infty$

- If $p_{ii} = 0$ then $T_i = \Delta$

- If $1 > p_{ii} > 0$ then probability of success for a jump is $(1 - p_{ii})$ then $T_i = k\Delta$, where $k \sim Geometric(1 - p_{ii})$.

Remarks important in In Discrete-Time Markov Chains :

- After a constant time Δ, we have the right to jump eventually to other state.

- T_i the time that we stay in the same state, is infered from p_{ii} or is dependent of p_{ii}.

Hence raised a question, what if we want to model T_i that takes a continous value instead of discret values ($k\Delta$) ? Then Continuous-Time Markov Chains comes for this issue.

7.2 Definition

More specifically, we will consider a random process $\{X(t), t \in [0, \infty)\}$. Again, we assume that we have a countable state space $S \subset \{1, 2, \ldots\}$. If $X(0) = i$, then X stays in state i for a random amount of time, say T_1, where T_1 is a continuous random variable. At time T_1, the process jumps to a new state j and will spend a random amount of time T_2 in that state and so on. We have the following definition:

A continuous-time Markov chain $X(t)$ is defined by two components:

- A *jump chain* that consists transition probabilities p_{ij} .

- A set of *holding time* parameters λ_i of **exponential distribution**, where $i \in S$. **Do not confuse** this parameter with the mean time that we stay in state i, which is thus $\frac{1}{\lambda_i}$.

We assume $p_{ii} = 0$ and there is no absorbing states $i \in S$. Then the mechanisme for jumping from a state to other state is :

- If $X(t) = i$, the time until the state changes has $Exponential(\lambda_i)$ distribution.

- If $X(t) = i$, the next state will be $j \neq i$ with probability p_{ij}.

The process satisfies the Markov property. That is, for all times where state changes $0 < t_1 < t_2 < \ldots < t_n < t_{n+1}$, we have :

$$P(X(t_{n+1}) = j | X(t_n) = i_n, X(t_{n-1}) = i_{n-1}, \ldots, X(t_1) = i_1)$$
$$= P(X(t_{n+1}) = j | X(t_n) = i_n),$$

where $j \neq i_n$ and $i_k \neq i_{k-1}, \forall k = n-1, \ldots, 2$.

7.2.1 Second definition

A continuous Markov chain can be sufficiently defined by the *continuous transition matrix* $P(t)$:

$$P(t) = \begin{bmatrix} p_{11}(t) & p_{12}(t) & \ldots & p_{1r}(t) \\ p_{21}(t) & p_{22}(t) & \ldots & p_{2r}(t) \\ \vdots & \vdots & \ddots & \vdots \\ p_{r1}(t) & p_{r2}(t) & \ldots & p_{rr}(t) \end{bmatrix}$$

where $p_{ij}(t) = P(X(t+s) = j | X(s) = i)$.

We must distinguish the jump chain (as transition matrix P in discrete Markov Chain case) and this transition matrix $P(t)$ in the continous case, which is dependent on t.

7.2.2 Why exponential distribution ?

Here raised a question, why exponential distribution is taken for modeling time until changing state ?

It is because its property "Memorylessness" :

$$P(X > t + s \mid X > t) = P(X > s),$$

where X means the time that state changes or event occurs. This means that the probability of an event occurring in the future is independent of how much time has already passed.

In propability and statistic domain, we have only the Geometric distribution (discret case) and Exponential distribution (continous case) that have the Memorylessness property.

And why Memorylessness property ?

This is because Markov property implies the memorylessness. To make any prediction about the future, it should not matter how long the process has been in a state. Thus, the time that the process spends in each state must have a "memoryless" property. Note that, it's the same as First arrival and Interarrivals in Poisson process in sec 5.3.1.

7.2.3 Example

Given continuous-time Markov chain $X(t)$ with the jump chain:
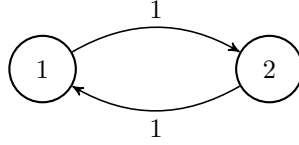
Assume the holding time parameters are given by $\lambda_0 = \lambda_1 = \lambda > 0$, and the jump chain is $p_{12} = p_{21} = 1$. Find the transition matrix $P(t)$.

Solution

$$p_{11}(t) = P(X(t) = 1|X(0) = 1)$$
$$= P(\text{even number of change for state})$$
$$= \sum_{n=0}^{\infty} e^{-\lambda t} \frac{(\lambda t)^{2n}}{(2n)!}$$
$$= e^{-\lambda t} \sum_{n=0}^{\infty} \left[\frac{e^{\lambda t} + e^{-\lambda t}}{2}\right] \quad \text{Taylor serie}$$
$$= \frac{1}{2} + \frac{1}{2}e^{-2\lambda t}$$

$$p_{12}(t) = 1 - p_{11}(t) = \frac{1}{2} - \frac{1}{2}e^{-2\lambda t}$$

Finally,

$$P(t) = \begin{bmatrix} \frac{1}{2} + \frac{1}{2}e^{-2\lambda t} & \frac{1}{2} - \frac{1}{2}e^{-2\lambda t} \\ \frac{1}{2} - \frac{1}{2}e^{-2\lambda t} & \frac{1}{2} + \frac{1}{2}e^{-2\lambda t} \end{bmatrix}$$

As mentionned in sec 7.2.1, it is sufficiently to use this transition matrix $P(t)$ to describe the continuous-time Markov chain $X(t)$.

7.3 Stationary and limiting distributions

7.3.1 Stationary distribution

As in the case of discrete-time Markov chains, for "nice" chains, a unique stationary distribution exists and it is equal to the limiting distribution.

Let $X(t)$ be a continuous-time Markov chain with transition matrix $P(t)$ and state space $S = \{1, 2, \ldots\}$. A probability distribution π on S, i.e, a vector $\pi = [\pi_1, \pi_2, \ldots]$, where $\pi_i \in [0, 1]$ and $\sum_{i \in S} \pi_i = 1$ is said to be a **stationary distribution** for $X(t)$ if

$$\pi = \pi P(t), \quad \text{for all } t \geq 0.$$

This implies that if the probability distribution of $X(0)$ is π, then the distribution of $X(t)$ is also given by π, for any $t \geq 0$.

7.3.1.1 Example

Let's come back at the example in 7.2.3. We try to solve for :

$$\begin{bmatrix} \pi_0 & \pi_1 \end{bmatrix} = \begin{bmatrix} \pi_0 & \pi_1 \end{bmatrix} \begin{bmatrix} \frac{1}{2} + \frac{1}{2}e^{-2\lambda t} & \frac{1}{2} - \frac{1}{2}e^{-2\lambda t} \\ \frac{1}{2} - \frac{1}{2}e^{-2\lambda t} & \frac{1}{2} + \frac{1}{2}e^{-2\lambda t} \end{bmatrix}$$

$$\text{s.t} \quad \pi_0 + \pi_1 = 1$$

Then we get $\pi_0 = \pi_1 = \frac{1}{2}$

7.3.2 Limiting distributions

The probability distribution $\pi = [\pi_1, \pi_2, \ldots]$ is called the limiting distribution of the continuous-time Markov chain $X(t)$ if

$$\pi_i = \lim_{t \to \infty} P(X(t) = i)$$

and

$$\sum_{i \in S} \pi_i = 1$$

Note that, the limiting distribution equals to the stationary distribution. The latter can be checked by using the example 7.2.3.

7.3.3 A method for stationary distribution

In theory, we can find the stationary (and limiting) distribution by solving $\pi = \pi P(t)$ or by finding $\lim_{t \to \infty} P(X(t) = i)$. However, in practice, finding $P(t)$ itself is usually very difficult.

Let $X(t), t \geq 0$ be a continuous-time Markov chain with an irreducible positive recurrent jump chain. Suppose that the unique stationary distribution of the *jump chain* (only discret transition probability p_{ij}) is given by

$$\tilde{\pi} = [\tilde{\pi}_1, \tilde{\pi}_2, \tilde{\pi}_3, \ldots]$$

and λ_i is the parameter in Exponential distribution for the time until the state changes from state i. Then

$$\pi_i = \lim_{t \to \infty} P(X(t) = i) = \frac{\frac{\tilde{\pi}_i}{\lambda_j}}{\sum_{k \in S} \frac{\tilde{\pi}_k}{\lambda_k}},$$

if $0 < \sum_{k \in S} \frac{\tilde{\pi}_k}{\lambda_k} < \infty$.

An intuitive for this formula is that, one we have $\tilde{\pi}_i$, we go further to get π_i by multiplying with the average time $\frac{1}{\lambda_i}$ that we stay in state i.

7.4 The generator matrix

The generator matrix, usually shown by G, gives us a way of analyzing continuous-time Markov chains, specially for getting *stationary distribution* or *limiting distribution*. It can be considered as an **operator** to get the derivative of continuous transition matrix $P(t)$.

7.4.1 Definition

Let's define matrix G as :

$$\begin{cases} g_{ii} = -\lambda_i \\ g_{ij} = \lambda_i p_{ij} \text{ with } j \neq i \end{cases}$$

where p_{ij} are discret transition probabilities in jump chaine. Note that in $p_{ii} = 0$ by assumption in defintion of continuous Markov process. From $\sum\limits_{j \neq i} p_{ij} = 1$ then $\sum\limits_{j} g_{ij} = 0$.

Then we have that $P'(t)$ is expressed by

- "forward equations" :
$$P'(t) = P(t)G$$

- "backward equations":
$$P'(t) = GP(t)$$

Proof

$$\begin{aligned} P'_{ij}(t) &= \lim_{\Delta \to 0} \frac{P_{ij}(t+\Delta) - P_{ij}(t)}{\Delta} \\ &= \lim_{\Delta \to 0} \frac{\sum_k P_{ik}(t)P_{kj}(\Delta) - P_{ij}(t)}{\Delta} \quad \text{(Chapman-Kolmogorov)} \\ &= \lim_{\Delta \to 0} \frac{\sum_{k \neq j} P_{ik}(t)P_{kj}(\Delta) + P_{ij}(t)P_{jj}(\Delta) - P_{ij}(t)}{\Delta} \\ &= \frac{\sum_{k \neq j} P_{ik}(t)[\lambda_k \Delta p_{kj}] + P_{ij}(t)[1 - \lambda_j \Delta] - P_{ij}(t)}{\Delta} \quad \text{(showed in next paragraph)} \\ &= \sum_{k \neq j} P_{ik}(t)\lambda_k p_{kj} - P_{ij}(t)\lambda_j \\ &= \sum_{k \neq j} P_{ik}(t)g_{kj} + P_{ij}(t)g_{jj} \\ &= \sum_{k} P_{ik}(t)g_{kj} \end{aligned}$$

We need to prove:

- $\lim_{\Delta \to 0} P_{kj}(\Delta) = \lambda_k \Delta p_{kj}, \forall k \neq j$, which means that state changes from k to j after exactly time length Δ.

- $\lim_{\Delta \to 0} P_{jj}(\Delta) = 1 - \lambda_j \Delta$, which means that state rests in state j after exactly time length Δ.

$$
\begin{aligned}
\lim_{\Delta \to 0} P_{kj}(\Delta) &= \lim_{\Delta \to 0} P(X(\Delta) = j | X(0) = k) p_{kj} \\
&= \lim_{\Delta \to 0} P(X(\Delta) \neq k | X(0) = k) p_{kj} \\
&= \lim_{\Delta \to 0} P(T < \Delta) p_{kj} \\
&= \lim_{\Delta \to 0} (1 - e^{-\lambda_k \Delta}) p_{kj} \\
&= \lambda_k \Delta p_{kj}
\end{aligned}
$$

$$
\begin{aligned}
\lim_{\Delta \to 0} P_{jj}(\Delta) &= \lim_{\Delta \to 0} P(X(\Delta) = j | X(0) = j) \\
&= \lim_{\Delta \to 0} [1 - P(X(\Delta) \neq j | X(0) = j)] \\
&= 1 - \lambda_j \Delta
\end{aligned}
$$

7.4.2 Example

Let's check the property $P'(t) = P(t)G = GP(t)$ in example 7.2.3. The generator matrix G is

$$
G = \begin{bmatrix} -\lambda & \lambda \\ \lambda & -\lambda \end{bmatrix}
$$

Then

$$
P'(t) = \begin{bmatrix} -\lambda e^{-2\lambda t} & \lambda e^{-2\lambda t} \\ \lambda e^{-2\lambda t} & -\lambda e^{-2\lambda t} \end{bmatrix}
$$

$$
\begin{aligned}
P(t)G &= \begin{bmatrix} \frac{1}{2} + \frac{1}{2} e^{-2\lambda t} & \frac{1}{2} - \frac{1}{2} e^{-2\lambda t} \\ \frac{1}{2} - \frac{1}{2} e^{-2\lambda t} & \frac{1}{2} + \frac{1}{2} e^{-2\lambda t} \end{bmatrix} \begin{bmatrix} -\lambda & \lambda \\ \lambda & -\lambda \end{bmatrix} \\
&= \begin{bmatrix} -\lambda e^{-2\lambda t} & \lambda e^{-2\lambda t} \\ \lambda e^{-2\lambda t} & -\lambda e^{-2\lambda t} \end{bmatrix}
\end{aligned}
$$

$$
\begin{aligned}
GP(t) &= \begin{bmatrix} -\lambda & \lambda \\ \lambda & -\lambda \end{bmatrix} \begin{bmatrix} \frac{1}{2} + \frac{1}{2} e^{-2\lambda t} & \frac{1}{2} - \frac{1}{2} e^{-2\lambda t} \\ \frac{1}{2} - \frac{1}{2} e^{-2\lambda t} & \frac{1}{2} + \frac{1}{2} e^{-2\lambda t} \end{bmatrix} \\
&= \begin{bmatrix} -\lambda e^{-2\lambda t} & \lambda e^{-2\lambda t} \\ \lambda e^{-2\lambda t} & -\lambda e^{-2\lambda t} \end{bmatrix}
\end{aligned}
$$

7.4.3 Corollary

Consider a continuous Markov chain $X(t)$ with the state space S and the generator Matrix G. The probability distribution π on S is a *stationary distribution* for $X(t)$ if and only if it satisfies:

$$\pi G = 0$$

Suppose that we have : $\pi = \pi P(t), \forall t$

$$\frac{d}{dt}[\pi] = \frac{d}{dt}[\pi P(t)], \quad \forall t$$
$$\Leftrightarrow 0 = \pi P'(t), \quad \forall t$$
$$\Leftrightarrow 0 = \pi G P(t), \quad \forall t$$

As $P(0) = P(t = 0) = I$ the identity matrix, this implies $\pi G = 0$
Inversely, suppose that $\pi G = 0$, then

$$\pi P'(t) = \pi G P(t) = 0, \forall t$$

Then $\pi P(t) = C \in \mathbb{R}^r, \forall t$. With $P(0) = I$, this implies $C = \pi$. Finally, $\pi P(t) = \pi$

7.4.3.1 Example

Let's resum in example 7.2.3. The generator matrix G is

$$G = \begin{bmatrix} -\lambda & \lambda \\ \lambda & -\lambda \end{bmatrix}$$

Then to find the stationary distribution, we need to solve :

$$\pi G = 0 = \begin{bmatrix} \pi_0 & \pi_1 \end{bmatrix} \begin{bmatrix} -\lambda & \lambda \\ \lambda & -\lambda \end{bmatrix} = 0$$
$$\text{s.t} \quad \pi_0 + \pi_1 = 1$$

Then $\pi_0 = \pi_1 = \frac{1}{2}$.

7.4.4 Why named generator matrix ?

Consider a continuous-time Markov chain $X(t)$. Assume $X(t_0) = i$. The chain will jump to the next state at time $t_0 + T$, where $T \sim Exponential(\lambda_i)$. We have seen that

$$\lim_{\Delta \to 0} P(T < \Delta) = \lambda_i \Delta$$

Thus, in a short interval of length Δ, the probability of leaving state i is approximately $\lambda_i \Delta$. For this reason, λ_i is often called the transition rate out of

state i. Furthermore, since we go from state i to state j with probability p_{ij}, then $g_{ij} = \lambda_i p_{ij}$ is the transition rate from state i to state j.

Therefore, we call "generator matrix" G because it is something characterizing the dynamics for a continuous Markov process X_t, which means they contain all sufficient information to describe X_t or to **generate realizations**, as the continuous transition matrix $P(t)$ does. There is an equivalence between two concepts, which means once we know one of them, we can infer the other :

- If we know $P(t)$, then $G = P(t)^{-1} P'(t)$. Since the latter is trur for all t, we take $t = 0$. Then $G = P(0)^{-1} P'(0) = P'(0)$ (since $P(0) = I$).

- If we know G, then we get $P(t)$ by solving the system of differential equations :
$$P'(t) = GP(t),$$
  where $P(0) = I$.

This generator matrix G is also known as the "infinitesimal generator" or the "Q-matrix."

8   Brownian Motion

In mathematics, this is also described by the *Wiener process*, a continuous-time stochastic process named in honor of Norbert Wiener. The Wiener process Wt is characterized by four facts :

- $W_0 = 0$

- W_t is almost surely (sec 4.6) continuous, $P(\lim_{t \to t_0} W_t = W_{t_0}) = 1$.

- W_t has independent increments : for all $0 \leq s_1 < t_1 \leq s_2 < t_2$, increments $W_{t_1} - W_{s_1}$ and $W_{t_2} - W_{s_2}$ are independent. By 1.10, the alternative of this condition is $W_t - W_s$ and W_s are independent, $\forall s < t$.

- $W_t - W_s \sim \mathcal{N}(0, |t - s|)$

8.1   Note

Let's call $\Delta W = W_t - W_{t+\Delta_t}$, then :

- $\mathbb{E}[\Delta W] = 0$

- $\mathbb{E}[(\Delta W)^2] = Var[\Delta W] = \Delta_t$

- $Var[(\Delta W)^2] = \mathbb{E}[(\Delta W)^4] - \mathbb{E}[(\Delta W)^2]^2 = 3\Delta_t^2 - \Delta_t^2 = 2\Delta_t^2$. Remind that $\mathbb{E}[(\Delta W)^{2n}] = (\Delta_t)^n (n2 - 1)!!$, !! means double factorial (e.g. $5!! = 1.3.5$) and here $n = 2$.

- From the second and third observations, if $\Delta_t \to 0$, then we can consider that $\Delta_t^2 = 0$, hence $Var[(\Delta W)^2] = 0$, hence $\mathbb{E}[(\Delta W)^2] = (\Delta W)^2$ or $(\Delta W)^2 = \Delta_t$.

- A Brownian motion is known that *everywhere continous* but *nowhere differentiable* (the slope at each side is stochastic so they are not the same). Then dW_t exists only in a stochastic definition but not deterministic definition.

Sometime, we use the notation B_t instead of W_t.

8.2 Quadratic variation

In mathematics, quadratic variation is used in the analysis of stochastic processes such as Brownian motion and other martingales. Quadratic variation is something like the *variation* of a process.

Suppose that X_t is a real-valued stochastic process defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and with time index t ranging over the non-negative real numbers. Its quadratic variation is the process, written as $[X]_t$, defined as :

$$[X]_t = \lim_{\|P\| \to 0} \sum_{k=1}^{n} (X_{t_k} - X_{t_{k-1}})^2$$

where P ranges over partitions of the interval $[0, t]$ and the norm of the partition P is the mesh, which is the $\max_k |t_k - t_{k-1}|$.

8.2.1 Covariation of two processes

Given two processes X and Y, this is also called cross-variance and denotes :

$$[X, Y]_t = \lim_{\|P\| \to 0} \sum_{k=1}^{n} \left(X_{t_k} - X_{t_{k-1}}\right)\left(Y_{t_k} - Y_{t_{k-1}}\right).$$

The covariation may be written in terms of the quadratic variation by the polarization identity:

$$[X, Y]_t = \tfrac{1}{2}([X + Y]_t - [X]_t - [Y]_t).$$

Notation: the quadratic variation is also notated as $\langle X \rangle_t$ or $\langle X, X \rangle_t$. We can say that this covariation is something that corresponds to covariance of two random variables.

8.2.2 Exercice

Show that with a partition Π, given $0 = t_0 \leq t_1 \leq ... \leq t_k \leq ... \leq t_n = T$, then the quadratic variation of Wiener process W_t is:

$$[W]_T = \lim_{\|\Pi\| \to 0} \sum_{k=1}^{n} (W_{t_k} - W_{t_{k-1}})^2 = T$$

Solution. Observation, if $t_k - t_{k-1} = \frac{T}{n}, \forall k$, then $[W]_T = \sum \frac{T}{n}\varepsilon^2 = T\mathbb{E}[\varepsilon^2] = T$, where $\varepsilon \leftarrow \mathcal{N}(0,1)$

We prove that:

$$\mathbb{E}\left[([W]_T - T)^2\right] = 0$$

$$\mathbb{E}\left[([W]_T - T)^2\right] = Var([W]_T - T) + \mathbb{E}^2\left[([W]_T - T)\right]$$

$$= Var([W]_T) + \left(\sum_{k=1}^{n} \mathbb{E}\left[(W_{t_k} - W_{t_{k-1}})^2\right] - T\right)^2$$

$$= \sum_{k=1}^{n} Var\left[(W_{t_k} - W_{t_{k-1}})^2\right] + \left(\sum_{k=1}^{n}(t_k - t_{t-1}) - T\right)^2$$

$$= \sum_{k=1}^{n} \mathbb{E}\left[(W_{t_k} - W_{t_{k-1}})^4\right] - \sum_{k=1}^{n}\mathbb{E}^2\left[(W_{t_k} - W_{t_{k-1}})^2\right] + (t_n - t_0 - T)^2$$

$$= \sum_{k=1}^{n} 3(t_k - t_{t-1})^2 - \sum_{k=1}^{n}(t_k - t_{t-1})^2 + (T - T)^2$$

$$= \sum_{k=1}^{n} 2(t_k - t_{t-1})^2$$

$$\leq 2\max_k(t_k - t_{t-1})\left(\sum_{k=1}^{n}(t_k - t_{t-1})\right)$$

$$= 2\max_k(t_k - t_{t-1})T$$

Note that, if $X \sim \mathcal{N}(0, \sigma^2)$, then $\mathbb{E}(X^{2n}) = (\sigma^2)^n(n2 - 1)!!$, where double factorial is, e.g. $7!! = 7 \times 5 \times 3 \times 1$. This can be show by using partial integral on $\int x^{2n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{x^2}{\sigma^2}} dx$

We observe that when $\max_k(t_k - t_{t-1}) \to 0$ then $\mathbb{E}\left[([W]_T - T)^2\right] \to 0$, hence $[W]_T = T$.

9 Martingale

9.1 Martingale

A martingale is a stochastic process which stays the same, on average. That is, the expected future value conditional on the present is equal to the current value.

9.1.1 Definition

Given probability space Ω, F_t, P, a basic definition of martingale is a stochastic process that satisfies,

- in discret case:

  - $\mathbb{E}(|X_n|) < \infty$ (integrable)
  - $\mathbb{E}(X_{n+1}|X_n, ..., X_1) = X_n$

- in continuous case:

  - $\mathbb{E}(|X_t|) < \infty$
  - $\mathbb{E}(X_t \mid \{X_\tau, \tau \leq s\}) = X_s$ or $\mathbb{E}(X_t|F_s) = X_s \quad \forall s < t$

where the second condition (conditional expectation 2.10) is equivalent to, $\forall s < t$,

$$\mathbb{E}[(X_t - X_s)\mathbb{1}_A] = 0, \quad \forall A \in F_s$$

9.1.2 Properties

9.1.2.1 Constant expectation

Given X_t is a martingale, then

$$\mathbb{E}[X_t] = \mathbb{E}[X_0]$$

Proof
By using law of iterated expectations in sec 2.11.1.1 :

$$\mathbb{E}_s(\mathbb{E}_t(X_t|F_s)) = \mathbb{E}_t(X_t)$$

Then from the definition of martingale :

$$\mathbb{E}_t(X_t|F_s) = X_s$$
$$\Rightarrow \mathbb{E}_s(\mathbb{E}_t(X_t|F_s)) = \mathbb{E}_s(X_s)$$
$$\Rightarrow \mathbb{E}_t(X_t) = \mathbb{E}_s(X_s) \quad s \leq t$$

By analogy, in discret case:

$$\mathbb{E}(X_{n+1}) = \mathbb{E}(X_n) = \ldots = \mathbb{E}(X_0)$$

9.1.2.2 Linear combination of two martingales

Given two martingales X_t and Y_t are both adapted to filtration $\mathcal{F}_t$, then

$$aX_t + bY_t$$

is also a martingale, $\forall a, b \in \mathbb{R}$.

Proof

$$\mathbb{E}[aX_t + bY_t \mid F_s] = a\mathbb{E}[X_t \mid F_s] + b\mathbb{E}[Y_t \mid F_s]$$
$$= aX_s + bY_s$$

In addition $\mathbb{E}[|aX_t + bY_t|] < \infty$ since X_t and Y_t are both integrable. Then we have Q.E.D

9.1.3 Examples

Examples include the wealth of a gambler as a function of time, assuming that he is playing a fair game. The canonical example of a continuous time martingale is Brownian motion and, in discrete time, a symmetric random walk is a martingale.

- A gambler's fortune (capital) is a martingale if all the betting games which the gambler plays are fair. To be more specific: suppose X_n is a gambler's fortune after n tosses of a fair coin, where the gambler wins \$1 if the coin comes up heads and loses \$1 if it comes up tails. The gambler's conditional expected fortune after the next trial, given the history, is equal to their present fortune. This sequence is thus a martingale.

- $X_n^2 - n$, where X_n in the above exemple is also martingale. In general, if $X_{n+1} = X_n + \epsilon$, $\mathbb{E}(\epsilon) = 0$, $Var(\epsilon) = 1$, ϵ and X_n are independent, hence $X_n^2 - n$ is martingale.

$$\mathbb{E}(X_{n+1}^2 - n - 1) = \mathbb{E}(X_n^2 + 2X_n\epsilon + \epsilon^2) - n - 1$$
$$= \mathbb{E}(X_n^2) - 2\mathbb{E}(X_n)\mathbb{E}(\epsilon) + \mathbb{E}(\epsilon^2) - n - 1$$
$$= \mathbb{E}(X_n^2) - n$$

- **Example of not martingale**. Let's consider an biased random walk. We start at position 0. At each time step, flip a biased coin: If it's heads with probability 0.75, take a step to the right (increase by 1). If it's tails with probability 0.25, take a step to the left (decrease by 1).

Let X_i be our position at time instant i. We check if X_i is a martingale:

- First, $\mathbb{E}[X_1] = 0.75 - 0.25 = 0.5$ and $\mathbb{E}[X_i] = \frac{1}{2}i$, so $\mathbb{E}[X_i] < \infty$ (ok).

- Second, $\mathbb{E}(X_{n+1}|X_n, ..., X_1) = X_n + 0.5$, which is not equals to X_n (not ok).

Hence, X_i is not martingale.

9.1.4 Submartingale

Submartingale is a stochastic process that is integrable and

$$\mathbb{E}[X_t \mid \mathcal{F}_s] \geq X_s, \quad \forall s < t,$$

which means that the process increases on average.

Similar to martingale, a property of submartingale is :

$$\mathbb{E}[X_t] \geq \mathbb{E}[X_s], \quad \forall s < t$$

9.1.4.1 Lemma

If X is a martingale and $f \colon \mathbb{R} \to \mathbb{R}$ is a convex function such that $f(X)$ is integrable, then, $f(X)$ is a submartingale. This is a direct consequence of conditional Jensen's inequality (sec 2.7.3.1) :

$$\mathbb{E}[f(X_t) \mid \mathcal{F}_s] \geq f(\mathbb{E}[X_t \mid \mathcal{F}_s]) = f(X_s)$$

9.1.4.2 Lemma

In the case that X is a submartingale, beside the condition that f is convex, we need a supplement condition that f is **increasing** function, which implies:

$$\mathbb{E}[f(X_t) \mid \mathcal{F}_s] \geq f(\mathbb{E}[X_t \mid \mathcal{F}_s]) \geq f(X_s)$$

then $f(X)$ is submartigale.

9.1.4.3 Lemma

If X is a submartingale and $f \colon \mathbb{R} \to \mathbb{R}$ is a **decreasing** and **concave** function such that $f(X)$ is integrable, then, $f(X)$ is a **supermartingale**. By Conditional Jensen's Inequality 2.7.3.1 in case of concave function and f is decreasing function, we have:

$$\mathbb{E}[f(X_t) \mid \mathcal{F}_s] \leq f(\mathbb{E}[X_t \mid \mathcal{F}_s]) \leq f(X_s)$$

9.1.4.4 Inequality

If X is a submartingale and $\mathbb{1}_A$ is an indicator function $(A \in \mathcal{F}_s)$, then

$$\mathbb{E}[X_s \mathbb{1}_A] \leq \mathbb{E}[X_t \mathbb{1}_A], \quad s < t$$

Proof

$$\begin{aligned}
\frac{\mathbb{E}[X_s \mathbb{1}_A]}{P(A)} &= \mathbb{E}[X_s \mid A] \\
&\leq \mathbb{E}[\mathbb{E}[X_t \mid \mathcal{F}_s] \mid A] \\
&= \mathbb{E}[X_t \mid A] \\
&= \frac{\mathbb{E}[X_t \mathbb{1}_A]}{P(A)}
\end{aligned}$$

where the one before last equality is by the general case law of iterated expectation 2.11.3. Here $\mathcal{H}_2 = \mathcal{F}_s$ and $\mathcal{H}_1 = \{\Omega, \varnothing, A, A^C\}$. We see that $\mathcal{H}_1 \subseteq \mathcal{H}_2$ since $A \in \mathcal{F}_n$.

9.1.5 Supermartingale

Supermartingale is a stochastic process that is integrable and

$$\mathbb{E}[X_t \mid \mathcal{F}_s] \leq X_s, \quad \forall s < t,$$

which means that the process decreases on average. As in case of submartingale in 9.1.4, we have also a respective property and lemma for supermartingale.

9.1.6 Doob decomposition theorem

In discrete time stochastic process, the Doob decomposition theorem gives a *unique* decomposition of every *adapted* and *integrable* stochastic process as the sum of a martingale and a *predictable* process (or drift) starting at zero. The analogous theorem in the continuous-time case is the Doob–Meyer decomposition theorem.

9.1.6.1 Statement

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $I = \{0, 1, 2, \ldots, N\}$, $(\mathcal{F}_n)_{n \in I}$ a filtration and $(X_n)_{n \in I}$ an adapted stochastic process with $\mathbb{E}[|X_n|] < \infty$ for all $n \in I$. Then there exists a martingale $M = (M_n)_{n \in I}$ and an integrable predictable process $A = (A_n)_{n \in I}$ starting with $A_0 = 0$ such that $X_n = M_n + A_n$ for every $n \in I$. Here predictable means that A_n is $\mathcal{F}_{n-1}$-measurable for every $n \in I \setminus \{0\}$ or informally, $A_n(\omega)$ is known with $\omega \in \mathcal{F}_{n-1}$ or $\mathbb{E}[A_n \mid \mathcal{F}_{n-1}] = A_n$.

This decomposition is almost surely unique.

9.1.6.2 Proof

Existence

Using conditional expectations (sec 2.11.1.1), define the processes A and M, explicitly by

$$A_n = \begin{cases} \sum_{k=1}^n \left(\mathbb{E}[X_k \mid \mathcal{F}_{k-1}] - X_{k-1} \right) & \text{if } n \geq 1 \\ 0 & \text{if } n = 0 \end{cases}$$

$$M_n = \begin{cases} X_0 + \sum_{k=1}^n \left(X_k - \mathbb{E}[X_k \mid \mathcal{F}_{k-1}] \right) & \text{if } n \geq 1 \\ X_0 & \text{if } n = 0 \end{cases}$$

Then we clearly see that $X_n = M_n + A_n$. By definition of A_n, we see that it is $\mathcal{F}_{n-1}$-measurable. First, $\mathbb{E}[|A_n|] < \infty$ and $\mathbb{E}[|M_n|] < \infty$ since $\mathbb{E}[|X_n|] < \infty, \forall n$.

Second,

$$\mathbb{E}[M_n \mid \mathcal{F}_{n-1}]$$

$$=\mathbb{E}\left[X_0 + \sum_{k=1}^{n}\big(X_k - \mathbb{E}[X_k \mid \mathcal{F}_{k-1}]\big) \mid \mathcal{F}_{n-1}\right]$$

$$=\mathbb{E}[X_n \mid \mathcal{F}_{n-1}] - \mathbb{E}[\mathbb{E}[X_n \mid \mathcal{F}_{n-1}] \mid \mathcal{F}_{n-1}] + \mathbb{E}[X_0 \mid \mathcal{F}_{n-1}] + \sum_{k=1}^{n-1}\big(\mathbb{E}[X_k \mid \mathcal{F}_{n-1}] - \mathbb{E}[\mathbb{E}[X_k \mid \mathcal{F}_{k-1}] \mid \mathcal{F}_{n-1}]\big)$$

$$=\mathbb{E}[X_0 \mid \mathcal{F}_{n-1}] + \sum_{k=1}^{n-1}\big(\mathbb{E}[X_k \mid \mathcal{F}_{n-1}] - \mathbb{E}[\mathbb{E}[X_k \mid \mathcal{F}_{k-1}] \mid \mathcal{F}_{n-1}]\big)$$

$$=X_0 + \sum_{k=1}^{n}\big(X_k - \mathbb{E}[X_k \mid \mathcal{F}_{k-1}]\big)$$

$$=M_{n-1}$$

where $\mathbb{E}[X_n \mid \mathcal{F}_{n-1}] - \mathbb{E}[\mathbb{E}[X_n \mid \mathcal{F}_{n-1}] \mid \mathcal{F}_{n-1}] = 0$ by general law of total expectation in sec 2.11.3 and $\mathbb{E}[X_k \mid \mathcal{F}_{n-1}]$ with $k \leq n-1$ is thus X_k since we know value of process until instant $n-1$. Note that, the law of total expectation is not applicable for $\mathbb{E}[\mathbb{E}[X_k \mid \mathcal{F}_{k-1}] \mid \mathcal{F}_{n-1}]$ since $\mathbb{E}[X_k \mid \mathcal{F}_{k-1}]$ is already knwon if given $\mathcal{F}_{n-1}$.

Uniqueness

Now suppose that we can decompose X_n by $X_n = A_n + M_n$ and $X_n = A'_n + M'_n$, let's consider $Y_n = M_n - M'_n = A'_n - A_n$. Since M_n and M'_n are martingale adapted to $\mathcal{F}_n$ then $M_n - M'_n$ is also martingale (sec 9.1.2.2). In addition, A_n and A'_n are predictable process then $A'_n - A_n$ is also predictable process (trivial demonstration). Then Y_n is martingale :

$$\mathbb{E}[Y_n \mid \mathcal{F}_{n-1}] = Y_{n-1}$$

and predictable process :

$$\mathbb{E}[Y_n \mid \mathcal{F}_{n-1}] = Y_n$$

This means $Y_n = Y_{n-1} = \ldots = Y_0 = 0$, since we have $Y_0 = A'_0 - A_0 = 0$ by defnition of A. This means the decomposition is almost surely unique.

9.1.6.3 Corollary

X is a (real-valued) submartingale if and only if it has a Doob decomposition into a martingale M and an integrable predictable process A that is almost surely increasing. If it is a supermartingale, if and only if A is almost surely decreasing.

9.1.7 Lemma

If X_n is a martingale adaptad to $\mathcal{F}_n$ and there exists a random variable Y such that $X_n = E(Y|\mathcal{F}_n), \forall n \in I$ where I is index set. Then the martingale X_n is called **right-closable** and X_n is uniformly integrable.

Proof

$$\frac{\mathbb{E}[|X_n|.\mathbb{1}_{\{|X_n|>M\}}]}{P(|X_n|>M)} = \mathbb{E}[|X_n| \mid \{|X_n|>M\}]$$

$$= \mathbb{E}[|\mathbb{E}[Y \mid \mathcal{F}_n]| \mid \{|X_n|>M\}]$$

$$\leq \mathbb{E}[\mathbb{E}[|Y| \mid \mathcal{F}_n] \mid \{|X_n|>M\}]$$

$$= \mathbb{E}[|Y| \mid \{|X_n|>M\}]$$

$$= \frac{\mathbb{E}[|Y|.\mathbb{1}_{\{|X_n|>M\}}]}{P(|X_n|>M)}, \forall n \in I$$

$$\Leftrightarrow \mathbb{E}[|X_n|.\mathbb{1}_{\{|X_n|>M\}}] \leq \mathbb{E}[|Y|.\mathbb{1}_{\{|X_n|>M\}}]$$

where

- The inequality is by Conditional Jensen's Inequality 2.7.3.1 with convex function $f(x) = |x|$.

- The one before last equality is by the general case law of iterated expectation 2.11.3, here $\mathcal{H}_2 = \mathcal{F}_n$ and $\mathcal{H}_1 = \{\Omega, \varnothing, \{|X_n|>K\}, \{|X_n| \leq K\}\}$. We see that $\mathcal{H}_1 \subseteq \mathcal{H}_2$ since $\{|X_n|>K\} \in \mathcal{F}_n$

The above equations are written by a lazy manner, for the correct notation, we need to write $\mathbb{E}[\mathbb{E}[Y \mid \mathcal{F}_n] \mid \{|X_n|>K\}]$ by $\mathbb{E}[\mathbb{E}[Y \mid \mathcal{F}_n] \mid \mathcal{H}_1](\{|X_n|>K\})$. If not using and by the same manner, we have $\mathbb{E}[X_n.\mathbb{1}_{\{|X_n|>M\}}] = \mathbb{E}[Y.\mathbb{1}_{\{|X_n|>M\}}]$.

Next, we use the similar method as in 3.3.2.

$$\mathbb{E}[|Y|.\mathbb{1}_{\{|X_n|>M\}}]$$

$$\leq \mathbb{E}[|Y|.\mathbb{1}_{\{|X_n|>M\}\cup\{|Y|>K\}}]$$

$$= \mathbb{E}[|Y|.\mathbb{1}_{\{|Y|\leq K\}\cap\{|X_n|>M\}}] + \mathbb{E}[|Y|.\mathbb{1}_{\{|Y|>K\}}]$$

$$\leq KP(|X_n|>M) + \mathbb{E}[|Y|.\mathbb{1}_{\{|Y|>K\}}]$$

$$\leq K\frac{\mathbb{E}[|X_n|]}{M} + \mathbb{E}[|Y|.\mathbb{1}_{\{|Y|>K\}}] \quad \text{(Markov inequality 2.7.1)}$$

$$\leq K\frac{\mathbb{E}[|Y|]}{M} + \mathbb{E}[|Y|.\mathbb{1}_{\{|Y|>K\}}]$$

where the last inequality is by appling Conditional Jensen's Inequality 2.7.3.1 with convex function $f(x) = |x|$:

$$|X_n| = |\mathbb{E}[Y \mid \mathcal{F}_n]| \leq \mathbb{E}[|Y| \mid \mathcal{F}_n]$$

This infers $\mathbb{E}[|X_n|] \leq \mathbb{E}[|Y|]$ by law of iterated expectation.

Finally,

$$\lim_{M\to\infty} \sup_{n\in I} \mathbb{E}[|X_n|.\mathbb{1}_{\{|X_n|>M\}}] \leq \lim_{M\to\infty} \sup_{n\in I} \left(\lim_{K\to\infty} K\frac{\mathbb{E}[|Y|]}{M} + \lim_{M\to\infty} \mathbb{E}[|Y|.\mathbb{1}_{\{|Y|>K\}}] \right) = 0$$

Here we note that K converges first to infinity and M can be chosen after K. In addition, $\lim_{M \to \infty} \mathbb{E}[|Y|.\mathbb{1}_{\{|Y|>K\}}] \to 0$ by 3.3.5.1. Hence X_n is uniformly integrable.

9.1.8 Uniform integrability of a backward submartingale

Let $\{X_t\}_{t \in T}$ be a submartingale with respect to a filtered probability space $(\Omega, \mathcal{F}_t, P)$. Then X_{t_n} is uniformly integrable for any decreasing sequence t_n (backward) which is lower bounded in T.

9.1.8.1 Proof

Thus, t_n decreases and converges to $t_\infty = a$. By the corollary of Doob decomposition theorem (sec 9.1.6.3), we can decompose X_{t_n} into a martingale M_{t_n} and a integrable predictable process A_{t_n}. Note that, t_n is decreasing sequence, then A_{t_n} is decreasing instead of increasing (backward). By definition, we have $A_0 = 0$ and since $t_n \geq 0, \forall n$ then $A_{t_n} \geq A_0 = 0$.

First, $0 \leq A_{t_n} \leq A_{t_0}$ or $|A_{t_n}| \leq A_{t_0}$, which is integrable then A_{t_n} is uniformly integrable by proposition 3.3.5.

Second, we show that M_{t_n} is also uniformly integrable. **Note that a martingale is by default not uniformly integrable**, in this case M_{t_n} is right-closable (sec 9.1.7) because

$$M_{t_n} = \mathbb{E}[M_{t_0} \mid \mathcal{F}_{t_n}], \forall n$$

Hence, M_{t_n} is uniformly integrable.

Finally, X_{t_n} is uniformly integrable by lemma 3.3.3.

9.1.9 Corollary

Let $\{X_t\}_{t \in T}$ be a submartingale with respect to a filtered probability space $(\Omega, \mathcal{F}_t, P)$. Then given $a \in \mathbb{R}$, we define $X_{t_n}^{+a} = \max(X_{t_n}, a)$. Then $X_{t_n}^{+a}$ is uniformly integrable for any decreasing sequence t_n which is lower bounded in T.

Proof

This is thus a direct consequence of lemma 9.1.8, which say that X_{t_n} is uniformly integrable. Then we note that if taking $K_0 = |a|$, then for all $|X_{t_n}^{+a}| \geq K_0$:

$$|X_{t_n}^{+a}| \leq |X_{t_n}|$$

Then by lemma 3.3.4, $X_{t_n}^{+a}$ is uniformly integrable.

9.2 Elementary process

Given $T = [0, L]$ and an partition of T with $0 = t_0 < t_1 < \ldots < t_N = L$. Let (Ω, F_{t_k}, P) be probability space. Let (Z_k) be random variable which is F_{t_k}-measurable, then Z is a stochastic process. An **elementary process** ξ_t (of Z) is a stochastic process defined by:

$$\xi(t,\omega) := Z_0(\omega)\mathbb{1}_{\{t=0\}} + \sum_{k=1}^{N} Z_k(\omega)\mathbb{1}_{(t_{k-1},t_k]}(t), \quad \forall(t,\omega) \in T \times \Omega$$

9.2.1   Stochastic integral

The stochastic integral is defined by an elementary process as integrand and stochastic process X_t as infinitesimal.

$$\int_0^L \xi(t,\omega)\,\mathrm{d}X(t,\omega) = \sum_{k=1}^{N} Z_k(\omega)(X(t_k,\omega) - X(t_{k-1},\omega))$$

Note that, for this expression to make sense, $X(t_k)$ must be also F_{t_k}-measurable function. The result of a stochastic integral is a **random variable**.

9.2.2   Integral process

In the case that the upper integral bound is $l \in [0,L]$, then

$$Y_l = \int_0^l \xi(t,\omega)\,\mathrm{d}X(t,\omega) = \sum_{k=1}^{N} Z_k(\omega)(X(t_k \wedge t,\omega) - X(t_{k-1} \wedge t,\omega)),$$

is a **stochastic process**, called integral process. The notaion $t_k \wedge t$ is just $\min(t_k, t)$. If bounds of integral are not presiced, we can denote $Y = \int \xi \mathrm{d}X$ for short. In the differential form, the equivalent equation can be written :

$$dY = \xi\,\mathrm{d}X$$

9.2.3   Theorem

An adapted (to F_{t_k}) integrable process X is

1. a submartingale if and only if

$$\mathbb{E}\left[\int_0^{+\infty} \xi\,dX\right] \geq 0$$

for all nonnegative bounded elementary processes ξ.

2. a supermartingale if and only if

$$\mathbb{E}\left[\int_0^{+\infty} \xi\,dX\right] \leq 0$$

for all nonnegative bounded elementary processes ξ.

3. a martingale if and only if

$$\mathbb{E}\left[\int_0^{+\infty} \xi \, dX\right] = 0$$

for all bounded elementary processes ξ.

Before going details for each statement, we show a common result :

$$
\begin{aligned}
\mathbb{E}\left[\int_0^{+\infty} \xi \, dX\right] &= \mathbb{E}\left[\sum_{k=1}^n Z_k(\omega)(X(t_k,\omega) - X(t_{k-1},\omega))\right] \quad (t_n = +\infty)\\
&= \sum_{k=1}^n \left(\mathbb{E}_{t_k}\left[Z_k X(t_k)\right] - \mathbb{E}_{t_{k-1}}\left[Z_k X(t_{k-1})\right]\right)\\
&= \sum_{k=1}^n \left(\mathbb{E}_{t_{k-1}}\left[\mathbb{E}_{t_k}\left[Z_k X(t_k)|F_{t_{k-1}}\right]\right] - \mathbb{E}_{t_{k-1}}\left[Z_k X(t_{k-1})\right]\right)\\
&= \sum_{k=1}^n \left(\mathbb{E}_{t_{k-1}}\left[\mathbb{E}_{t_k}\left[Z_k X(t_k)|F_{t_{k-1}}\right] - Z_k X(t_{k-1})\right]\right)
\end{aligned}
$$

Proof for 1

(*) We show that X is submartingale $\Rightarrow \mathbb{E}\left[\int_0^{+\infty} \xi \, dX\right] \geq 0$, for all nonnegative bounded elementary processes ξ.

We remind that with two random variable A and B, if $\mathbb{E}[A] \geq B$ and α is a bounded non-negative random variable, then $\mathbb{E}[\alpha A] \geq \alpha B$.

Since X is a submartingale, then $\mathbb{E}_{t_k}\left[X(t_k)|F_{t_{k-1}}\right] \geq X(t_{k-1})$. Apply the above reminding, with Z_k is a non-negative random variable, then $\mathbb{E}_{t_k}\left[Z_k X(t_k)|F_{t_{k-1}}\right] \geq Z_k X(t_{k-1})$. So we have the Q.E.D.

(**) Conversely, we show that $\mathbb{E}\left[\int_0^{+\infty} \xi \, dX\right] \geq 0$ for all nonnegative bounded elementary processes $\xi \Rightarrow X$ is submartingale.

We show that $\mathbb{E}\left[X(t)|F_s\right] \geq X(s), \forall s < t$ by contradiction.

On one hand, suppose that there exists $s < t$ that $\mathbb{E}[X(t)|F_s] < X(s)$. This implies $\mathbb{E}_s[\mathbb{E}_t[X(t)|F_s]] < \mathbb{E}_s[X(s)]$ or $\mathbb{E}[X(t)] < \mathbb{E}[X(s)]$.

On other hand, we take the elementary processes $\xi = \mathbb{1}_{(s,t]}$, which is thus 1 in $(s,t]$ and 0 otherwise. From $\mathbb{E}\left[\int_0^{+\infty} \xi \, dX\right] \geq 0$, we have $\mathbb{E}[X(t)] - \mathbb{E}[X(s)] \geq 0$ or $\mathbb{E}[X(t)] \geq \mathbb{E}[X(s)]$.

Then we have the contradiction. This contradiction show that $\mathbb{E}\left[X(t)|F_s\right] \geq X(s), \forall s < t$ or X is a submartingale.

Proof for 2

The same way as proof for 1

Proof for 3

(*) We show that X is martingale $\Rightarrow \mathbb{E}\left[\int_0^{+\infty} \xi \, dX\right] = 0$, for all bounded elementary processes ξ.

This is by using the common result and the remark that if $\mathbb{E}[A] = B$ and α is a bounded random variable, then $\mathbb{E}[\alpha A] = \alpha B$.

(**) Conversely, we show that $\mathbb{E}\left[\int_0^{+\infty} \xi\, dX\right] = 0$, for all bounded elementary processes $\xi \Rightarrow X$ is martingale. Thus, $\mathbb{E}\left[\int_0^{+\infty} \xi\, dX\right] = 0$ with all nonnegative bounded elementary processes infers that X is both submartingale and supermartingale, then X is a martingale.

9.2.4 Lemma

Let X be a process and ξ be a bounded elementary process. Define $Y_t = \int_0^t \xi\, dX$. Then

1. If X is a submartingale then

$$\begin{cases} Y \text{ is a submartingale if } \xi \text{ is non-negative.} \\ Y \text{ is a supermartingale if } \xi \text{ is non-positive.} \end{cases}$$

2. If X is a supermartingale then

$$\begin{cases} Y \text{ is a supermartingale if } \xi \text{ is non-negative.} \\ Y \text{ is a submartingale if } \xi \text{ is non-positive.} \end{cases}$$

3. If X is a martingale then so is Y.

Proof for 1
Given $s < t$, on one hand :

$$\mathbb{E}_s[\mathbb{E}_t[Y_t|F_s]] = \mathbb{E}[Y_t]$$
$$= \mathbb{E}\left[\int_0^t \xi\, dX\right]$$
$$= \mathbb{E}[\sum_{k=1}^n Z_k(X(t_k \wedge t) - X(t_{k-1} \wedge t))]$$
$$= \sum_{k=1}^n \mathbb{E}[Z_k(X(t_k \wedge t) - X(t_{k-1} \wedge t))]$$

On the other hand :

$$\mathbb{E}[Y_s] = \mathbb{E}\left[\int_0^s \xi\, dX\right]$$
$$= \sum_{k=1}^n \mathbb{E}\left[Z_k(X(t_k \wedge s) - X(t_{k-1} \wedge s))\right]$$

(*) *In case that ξ is non-negative* :

Suppose that $\mathbb{E}_t[Y_t|F_s] < Y_s$, then $\mathbb{E}_s[\mathbb{E}_t[Y_t|F_s]] < \mathbb{E}[Y_s]$ or

$$\sum_{k=1}^{n} \mathbb{E}[Z_k(X(t_k \wedge t) - X(t_{k-1} \wedge t))] < \sum_{k=1}^{n} \mathbb{E}\left[Z_k(X(t_k \wedge s) - X(t_{k-1} \wedge s))\right]$$

This is equivalent to

$$\mathbb{E}[Z_l(X(s) - X(t_l)) + \sum_{k=l+1}^{h-1} \mathbb{E}[Z_k(X(t_k) - X(t_{k-1}))] + \mathbb{E}[Z_h(X(t) - X(t_{h-1}))] < 0$$

where $s \in (t_{l-1}, t_l]$ and $t \in (t_{h-1}, t_h]$. By analogy to the proof for theorem 9.2.3, we have if X is a submartingale and ξ is non-negative, then $\mathbb{E}[Z_k(X(a) - X(b))] \geq 0, \forall a > b$ and $\forall k$. Then

$$\mathbb{E}[Z_l(X(s) - X(t_{l-1})) + \sum_{k=l+1}^{h-1} \mathbb{E}[Z_k(X(t_k) - X(t_{k-1}))] + \mathbb{E}[Z_h(X(t) - X(t_{h-1}))] \geq 0$$

This latter is equivalent to $\mathbb{E}_s[\mathbb{E}_t[Y_t|F_s]] \geq \mathbb{E}_s[Y_s]$ (contradiction). Then we must have $\mathbb{E}_t[Y_t|F_s] \geq Y_s$ or Y_t is submartingale.

(**) *In case that ξ is non-positive* :

By the same manner in case that ξ is non-negative (contradiction), with remark that if X is a submartingale and ξ is non-positive, then $\mathbb{E}[Z_k(X(a) - X(b))] \leq 0, \forall a > b$.

Proof for 2

By the same manner as 1.

Proof for 3

Given $s < t$, on one hand :

$$
\begin{aligned}
\mathbb{E}[Y_t|F_s] &= \mathbb{E}\left[\int_0^t \xi \, dX\right] \\
&= \mathbb{E}[\sum_{k=1}^{n} Z_k(X(t_k \wedge t) - X(t_{k-1} \wedge t)|F_s)] \\
&= \sum_{k=1}^{n} \left(\mathbb{E}[Z_k X(t_k \wedge t)|F_s] - \mathbb{E}[X(t_{k-1} \wedge t)|F_s]\right) \\
&= \sum_{k=1}^{l-1} \left(Z_k X(t_k) - Z_k X(t_{k-1})\right) + \mathbb{E}[Z_l X(t_l)|F_s] - Z_l X(t_{l-1}) \\
&\quad + \sum_{k=l+1}^{h-1} \left(\mathbb{E}[Z_k X(t_k)|F_s] - \mathbb{E}[X(t_{k-1})|F_s]\right) + \mathbb{E}[Z_h X(t)|F_s] - \mathbb{E}[Z_h X(t_{h-1})|F_s]
\end{aligned}
$$

On the other hand :

$$Y_s = \int_0^s \xi \, \mathrm{d}X$$

$$= \sum_{k=1}^n Z_k(X(t_k \wedge s) - X(t_{k-1} \wedge s))$$

$$= \sum_{k=1}^{l-1}(Z_k X(t_k) - Z_k X(t_{k-1})) + Z_l X(s) - Z_l X(t_{l-1})$$

where $s \in (t_{l-1}, t_l]$ and $t \in (t_{h-1}, t_h]$

$$\mathbb{E}[Y_t|F_s] - Y_s$$

$$=\mathbb{E}[Z_h X(t)|F_s] - \mathbb{E}[Z_h X(t_{h-1})|F_s] + \sum_{k=l+1}^{h-1}(\mathbb{E}[Z_k X(t_k)|F_s] - \mathbb{E}[X(t_{k-1})|F_s])$$

$$+\mathbb{E}[Z_l X(t_l)|F_s] - Z_l X(s)$$

Since X is a martingale then $\mathbb{E}[Z_k X(a)|F_s] = X(s)$ with $\forall a \geq s$, then $\mathbb{E}[Y_t|F_s] - Y_s = 0$ or Y_t is a martingale.

9.2.5 Lemma

Let X be a martingale (resp. submartingale, supermartingale) and τ be a simple stopping time. Then, the stopped process X^τ is also a martingale (resp. submartingale, supermartingale).

This is inferred from the above lemma :

$$X^\tau = \int_0^\tau \xi \mathrm{d}X$$

where $\xi = \mathbb{1}_{[0,\tau]}$.

9.3 Upcrossings, downcrossings and martingale convergence

9.3.1 Upcrossings and downcrossings

Consider a process X_t whose time index t runs through an index set $T \subseteq \mathbb{R}$. Given real numbers $a < b$ and $t_0 = 0$. Let's define a sequence (can be infinite) s_k, t_k by

$$\begin{cases} s_k = \inf\{m \geq t_{k-1} : x_m \leq a\} \\ t_k = \inf\{m \geq s_k : x_m \geq b\} \end{cases}$$

Figure 4.3 show an illustration for upcrossings and downcrossings of a process. We denote $U[a,b]$ number of upcrossings and $D[a,b]$ number of downcrossings.

Note that between any two upcrossings there is a downcrossing and, similarly, between any two downcrossings there is an upcrossing. It follows that $U[a,b]$
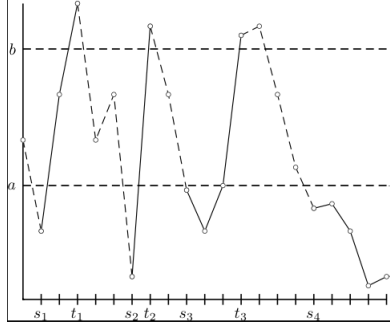
Figure 4.3: Upcrossing downcrossing for given interval $[a, b]$. Thus a, b can be consider like thresholding. This process has 3 upcrossings and 4 downcrossings.

and $D[a, b]$ **can differ by at most 1**, and they are either both finite or both infinite.

9.3.2 Lemma

Let's consider a probability space $(\Omega, \mathcal{F}, P)$ and a filtration of σ-algebras $(F_t)_{t \in T}$. Let X_t be a stochastic process with time t running over the **finite index** set $T \subseteq \mathbb{R}$, and set $t_M = \max(T)$. Let the elementary process ξ_A be defined as :

$$\xi_A(t, \omega) = \sum_{k=1}^{n} \mathbb{1}_{(s_k, t_k]}(t)$$

where s_k, t_k are downcrossing and upcrossing instant concerns to X_t defined by

$$\begin{cases} s_k = \inf\left(\{m \geq t_{k-1} : x_m \leq a\}\right) \\ t_k = \inf\left(\{m \geq s_k : x_m \geq b\}\right) \end{cases}$$

Then

$$\int \xi_A \, dX = \sum_{k=1}^{\infty} (X_{t_k \wedge t_M} - X_{s_k \wedge t_M})$$

Since T is a finite set, then there exists t_N the last upcrossing instant. There are two cases need to be considered:

- After t_N, X_t is always strictly greater than a. In this case :

$$\int \xi_A \, dX = \sum_{k=1}^{N} (X_{t_k \wedge t_M} - X_{s_k \wedge t_M}) \geq (b - a)U[a, b]$$

- After t_N, X_t has one more downcrossing at s_{N+1} and then X is always strictly less than b.

$$\int \xi_A \, dX = \sum_{k=1}^{N} (X_{t_k \wedge t_M} - X_{s_k \wedge t_M}) + X_{t_M} - X_{s_{N+1}} \geq (b-a)U[a, b] - \max(a - X_{t_M}, 0),$$

249

by supplementally conditioning whether $X_{t_M} < a$ or $X_{t_M} \geq a$.

In both cases, we have

$$\int \xi_A \, \mathrm{d}X \geq (b-a)U[a,b] - \max(a - X_{t_M}, 0)$$

9.3.3 Doob's upcrossing lemma

It is also called Doob's upcrossing inequality. Let X_t be a supermartingale with time t running through a **countable index set** T. Then the expected number of upcrossings of any $a < b$ satisfies

$$(b-a)\mathbb{E}\left[U[a,b]\right] \leq \mathbb{E}\left[(a - X_{t_M}) \vee 0\right]$$

Proof

Let's denote J_n **finite index set** that is increasting sequence $(J_{n-1} \subseteq J_n)$ and converges to **countable index set** T ($\lim\limits_{n\to\infty} J_n \to T$). Given a and b, let $f_n : J_n \to \mathbb{R}^+$ are measurable function that

$$f_n(x) = \begin{cases} 1 & \text{if } x \text{ is a upcrossing point } t_k \\ 0 & \text{otherwise} \end{cases}$$

Then f_n converges pointwise to $f : T \to \mathbb{R}^+$. By convergence monotone theorem 2.14, we have

$$(U[a,b], J_n) = \lim_{n\to\infty} \int_T f_n = \int_T \lim_{n\to\infty} f_n = \int_T f = (U[a,b], T)$$

where $(U[a,b], T)$ is the number of upcrossing in set T.

Come back to Doob's upcrossing lemma, it is enough to prove this for **finite index set** J.

From the lemma 9.3.2, we have

$$\begin{aligned} (b-a)\mathbb{E}\left[U[a,b]\right] &\leq \mathbb{E}\left[\int \xi_A \, \mathrm{d}X\right] + \mathbb{E}\left[(a - X_{t_M}) \vee 0\right] \\ &\leq \mathbb{E}\left[(a - X_{t_M}) \vee 0\right] \\ &= \mathbb{E}\left[(a - X_{t_M})^+\right] = \mathbb{E}\left[(X_{t_M} - a)^-\right] \end{aligned}$$

We have $\mathbb{E}\left[\int \xi_A \, \mathrm{d}X\right] \leq 0$ since ξ_A is nonnegative bound evementary process and X is supermartingale (theorem 9.2.3). The notation $x^+ = \max(x,0)$ means the positive part of x and $x^- = -\min(x,0)$ means the negative part of x. Both positive part et negative part are nonnegative.

9.3.4 Lemma

Given a stochastic process X and let t_n be a monotone sequence. Note that by indexing with n, then the $\{t_n \colon n \in \mathbb{N}\}$ is countable set. If $U[a,b](\omega)$ of $X_{t_n}(\omega)$ is finite for all $a, b \in \mathbb{R}$, then X_{t_n} almost surely converges.

Proof

Let's consider

$$
\begin{aligned}
\Lambda &= \{\omega \colon \lim_{n \to \infty} X_{t_n}(\omega) \text{ does not converge}\} \\
&= \{\omega \colon \liminf_{n \to \infty} X_{t_n}(\omega) < \limsup_{n \to \infty} X_{t_n}(\omega)\} \\
&= \bigcup_{a < b; a, b \in \mathbb{Q}} \{\omega \colon \liminf_{n \to \infty} X_{t_n}(\omega) < a < b < \limsup_{n \to \infty} X_{t_n}(\omega)\} \\
&= \bigcup_{a < b; a, b \in \mathbb{Q}} \{\omega \colon U[a,b](\omega) = +\infty\}
\end{aligned}
$$

We use $a, b \in \mathbb{Q}$ instead of $a, b \in \mathbb{R}$ to benefit the countable additivity of a measure (1.7.1). As $\mathbb{Q}$ is countable set, then $\mathbb{Q}^2$ is also countable set then we can apply the countable additivity:

$$
P(\Lambda) = \sum_{a < b; a, b \in \mathbb{Q}} P(\{\omega \colon U[a,b](\omega) = +\infty\}) = 0
$$

Then X_n converges almost surely (4.6).

9.3.5 Doob's Forward Convergence Theorem

Also called Doob's martingale convergence theorem.

9.3.5.1 Statement

Let $(X_n)_{n=1,2,\dots}$ be a supermartingale with the filtration F_n such that

$$
\sup_{n \in \mathbb{N}} \mathbb{E}[X_n^-] < +\infty
$$

Then, with probability one, the limit

$$
\lim_{n \to +\infty} X_n(\omega[\colon n]) = X_\infty(\omega), \quad \forall \omega \in F_\infty
$$

exists (where X_∞ is a random variable) and $E[X_\infty]$ is finite (or X_∞ is integrable). In other words, the random variables X_n converge **almost surely** (sec 4.6) to X_∞.

9.3.5.2 Important note

There are two concepts that must be distinguished, given the process X_n :

1. $\mathbb{E}[|X_n|] < \infty, \forall n \in N$, which means that X_n is bounded in L_1 for all $n \in \mathbb{N}$.

2. $\sup_{n \in \mathbb{N}} \mathbb{E}[|X_n|] < +\infty$, which means the process is (bounded) in L_1.

The second is stronger than the first. If we have the first, it does not mean we have the second. For example, if $\mathbb{E}[|X_n|] = n$, then each X_n is bounded in L_1 for all $n \in N$, but the process (X_n) is not bounded in L_1.

9.3.5.3 Proof

1. **Given an (fixed)** ω, then for all $a < b$, we have :

$$(b-a)\mathbb{E}\left[U[a,b]\right] \leq \mathbb{E}\left[\left(\lim_{n \to +\infty} X_n - a\right)^-\right]$$
$$\leq \mathbb{E}\left[\left(\lim_{n \to +\infty} X_n\right)^-\right] + a^+$$

where

- The first inequality is from Doob's upcrossing lemma (sec 9.3.3) with X_n is a supermartingale and index $n \in \mathbb{N}$, which is a countable index set.

- The second inequality is from sec 2.7.8.

In addition, we have $\sup_{n \in \mathbb{N}} \mathbb{E}[X_n^-] < +\infty$ then $\exists M$ that $\mathbb{E}[X_n^-] < M, \forall n$ then :

$$\mathbb{E}\left[U[a,b]\right] \leq \frac{M + a^+}{b - a} < +\infty$$

2. By lemma 9.3.4, as $n \in \mathbb{N}$ is a countable set and $U[a,b]$ of X_n is finite, then X_n converges almost surely.

3. We show that $\mathbb{E}[X_\infty]$ is finite. Since X_n is a supermartingale then from definition in sec 9.1.5, we have

$$\mathbb{E}[X_n] < X_1$$

Then $\mathbb{E}[X_\infty] < +\infty$. Now suppose that $\mathbb{E}[X_\infty] = -\infty$. From $\mathbb{E}[X_\infty] = \mathbb{E}[X_\infty^+] - \mathbb{E}[X_\infty^-]$ and $\mathbb{E}[X_\infty^+]$ is something positive, we infer that $\mathbb{E}[X_\infty^-] = +\infty$. But we have that $\mathbb{E}[X_n^-] < M, \forall n$, then we have the contracdiction. This infers that $\mathbb{E}[X_\infty] < -\infty$ then $\mathbb{E}[X_\infty]$ is finite.

9.3.5.4 Analogy for submartingale

If $(X_n)_{n=1,2,\dots}$ is a submartingale such that $\sup_{n \in \mathbb{N}} \mathbb{E}[X_n^+]$ is (upper) bounded $(< +\infty)$, then $X_\infty = \lim_{n \to +\infty} X_n$ exists and $E[X_\infty]$ is finite.

9.3.5.5 Analogy for martingale

If $(X_n)_{n=1,2,\dots}$ is a martingale such that $\sup\limits_{n\in\mathbb{N}} \mathbb{E}[X_n^+]$ and $\sup\limits_{n\in\mathbb{N}} \mathbb{E}[X_n^-]$ are both (upper) bounded or equivalently $\sup\limits_{n\in\mathbb{N}} \mathbb{E}[|X_n|]$ is bounded, then $X_\infty = \lim\limits_{n\to+\infty} X_n$ exists and $E[X_\infty]$ is finite.

The condition that the martingale (X_n) is bounded is essential; for example, an unbiased ± 1 random walk is a martingale but does not converge.

9.3.6 Extended of Doob's Forward Convergence Theorem

Let $(X_n)_{n=1,2,\dots}$ be a martingale with filtration F_n, where X_n is uniformly integrable (in sec 3.3). Then it exists an random variable X_∞, which is F_∞-measurable and also uniformly integrable such that

$$X_n \xrightarrow{L_1} X_\infty$$

or X_n converges in mean (or in L^1, in sec 4.5).

Proof

Since X_n is uniformly integrable, this implies that $\sup\limits_{n\in\mathbb{N}} \mathbb{E}[|X_n|] < \infty$ (sec 3.3.2). Then Doob's Forward Convergence Theorem in sec 9.3.5 says that X_n converges almost surely to X_∞. Since converge almost surely is stronger than converge in probability (in figure 2.4), then X_n converges also in probability to X_∞.

Moreover, X_n is uniformly integrable, then X_n converges in mean (or in L^1) to X_∞ (lemma 4.5.4).

In continuous case, we have theorem 11.1.3

9.3.7 Exercice

Let $(X_n)_{n\in\mathbb{N}}$ be a standard random walk. That is, $X_1 = 0$ and

$$P(X_{n+1} = X_n + 1 \mid \mathcal{F}_n) = P(X_{n+1} = X_n - 1 \mid \mathcal{F}_n) = \frac{1}{2}$$

Then, for every integer a, we have $X_n = a$ for some n with probability one.

Solution

We aim to solve this exercice by contradiction. Without loss of generality, let's take $a < 0$. Let T be the first time n that $X_n = a$. If we never have n such that $X_n = a$, then $T = +\infty$ and $X_n > a, \forall n$. The latter means that $\sup\limits_{n\in\mathbb{N}} \mathbb{E}[X_n^-] < +\infty$. Since X_n is martingale then X_n is also supermartingale. Then Doob's Forward Convergence Theorem (in sec 9.3.5) says that X_n converge almost surely to X_∞, or for all $\varepsilon > 0$, it exists n_0 such that $\forall m, n > n_0$, $|X_m - X_n| < \varepsilon$. However, the property of random walk is that $|X_{n+1} - X_n| = 1$. Here, we have the contradiction.

9.3.8 Levy's Upwards and Downwards Theorems

A very useful application of martingale convergence is to the following results concerning conditional expectations by σ-algebra (sec 2.11.1.1), for a limit of increasing or decreasing σ-algebras.

9.3.8.1 Upward

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $(\mathcal{F}_n)_{n \in \mathbb{N}}$ be an increasing sequence of sub σ-algebras of $\mathcal{F}$, which means $\mathcal{F}_k \subseteq \mathcal{F}_{k+1}, \forall k$. Set $\mathcal{F}_\infty \equiv \sigma\left(\bigcup_n \mathcal{F}_n\right)$. For any random variable X in L^1 (integrable). Then

$$\lim_{n \to \infty} \mathbb{E}[X \mid \mathcal{F}_n] \xrightarrow{a.s.} \mathbb{E}[X \mid \mathcal{F}_\infty]$$

Proof

Let's $Y_n = \mathbb{E}[X \mid \mathcal{F}_n]$. First, by law of iterated expectation in sec 2.11.3 :

$$\mathbb{E}[Y_{n+1} \mid \mathcal{F}_n] = \mathbb{E}[\mathbb{E}[X \mid \mathcal{F}_{n+1}] \mid \mathcal{F}_n] = \mathbb{E}[X \mid \mathcal{F}_n] = Y_n,$$

then Y_n is a martingale.

Second, we show that $\sup_{n \in \mathbb{N}} \mathbb{E}[|Y_n|] < +\infty$. Thus

$$\sup_{n \in \mathbb{N}} \mathbb{E}[|Y_n|] = \sup_{n \in \mathbb{N}} \mathbb{E}[|\mathbb{E}[X \mid \mathcal{F}_n]|] \leq \sup_{n \in \mathbb{N}} \mathbb{E}[\mathbb{E}[|X| \mid \mathcal{F}_n]] = \mathbb{E}[|X|] < +\infty$$

where the inequality is by the conditional Jensen's Inequality in 2.7.3.1, with $f(x) = |x|$ is convex.

Finally, from the first and the second, with Doob's martingale convergence theorem (sec 9.3.5.5), we have Q.E.D

9.3.8.2 Downward

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $(\mathcal{F}_n)_{n \in \mathbb{N}}$ be an decreasing sequence of sub σ-algebras of $\mathcal{F}$, which means $\mathcal{F}_k \supseteq \mathcal{F}_{k+1}, \forall k$. Set $\mathcal{F}_\infty \equiv \bigcap_n \mathcal{F}_n$. For any random variable X in L^1 (integrable). Then

$$\lim_{n \to \infty} \mathbb{E}[X \mid \mathcal{F}_n] \xrightarrow{a.s.} \mathbb{E}[X \mid \mathcal{F}_\infty]$$

10 cadlag

10.1 Definition

It means *continue à droite, limite à gauche* in french, which means *right continuous with left limits*. Figure 4.4 show an example of cadlag function. The reference for left ou right is an abscisse point x, **not to confuse with the left or the right of function**. In this figure :

- **Continue à droite** : $\lim\limits_{x \to a^+} f(x) = f(a)$.

- **Limite à gauche** : $\lim\limits_{x \to a^-} f(x)$ has a limit.
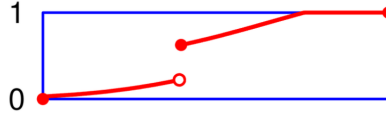
Figure 4.4: cadlag function.

10.1.1 Remark

Only right continuous does not imply left limit, e.g. the following function

$$f(x) = \begin{cases} \sin(\frac{1}{x}) & \text{if } x < 0 \\ 0 & \text{if } x \geq 0 \end{cases}$$

is right continuous but has not left limit at $x = 0$.

10.2 cadlag process

A cadlag process X_t is a type of stochastic process that is, for all $\omega \in \Omega$, then $X(t, \omega)$ or X_t is a cadlag function. Cadlag processes allow understanding and modeling real-world phenomena subject to random fluctuations.

10.3 Theorem

Let X be an adapted stochastic process with index $t \in \mathbb{R}^+$. If either of the following conditions holds :

- For all t, $\mathbb{E}[|X_t|] < \infty$ (X is integrable) and $\mathbb{E}[\int_0^t \mathbb{1}_A \, dX]$ is bounded, for all $A \subseteq \mathbb{R}^+$.

- For all t, $\int_0^t \mathbb{1}_A \, dX$ is bounded in probability, for all $A \subseteq \mathbb{R}^+$.

Then, X has a version Y (ou modification, sec 1.8.1) which has left and right limit for all t, and there is a countable subset $S \subset \mathbb{R}_+$ for which Y_t is right-continuous at every $t \notin S$.

10.3.1 Proof

- X has a version Y which has left and right limit for all $t \in \mathbb{R}^+$. Let's a sequence $t_n \in \mathbb{R}^+$ converges monotonically to t, we show that X_{t_n} almost surely converges.

By lemma 9.3.2 :

$$(b - a)U_{[a,b]}^{X_{t_n}} \leq \int \mathbb{1}_B \, \mathrm{d}X + (a - X_{t_n})^+$$

where B is indicator set, which equals to 1 between a downcrossing and its nextby upcrossing, 0 otherwise.

– If for all t, $\mathbb{E}[|X_t|] < \infty$ (X is integrable) and $\mathbb{E}[\int_0^t \mathbb{1}_A \, \mathrm{d}X]$ is bounded, for all $A \subseteq \mathbb{R}^+$. Then

$$\begin{aligned}
\mathbb{E}[U_{[a,b]}^{X_{t_n}}] &\leq \frac{\mathbb{E}[\int \mathbb{1}_B \, \mathrm{d}X] + \mathbb{E}[(a - X_{t_n})^+]}{(b-a)} \\
&< \frac{M_1 + a^- + \mathbb{E}[X_{t_n}^+]}{(b-a)} \\
&< \frac{M_1 + a^- + M_2}{(b-a)}
\end{aligned}$$

where M_1 is a bound for $\mathbb{E}[\int \mathbb{1}_B \, \mathrm{d}X]$ (since $B \subset \mathbb{R}^+$), inequality $\mathbb{E}[(a - X_{t_n})^+] \leq a^- + \mathbb{E}[X_{t_n}^+]$ by 2.7.8 and M_2 is a bound for $\mathbb{E}[X_{t_n}^+]$. Therefore, $U_{[a,b]}^{X_{t_n}}$ is finite for all n.

– If for all t, $\int_0^t \mathbb{1}_A \, \mathrm{d}X$ is bounded in probability, for all $A \subseteq \mathbb{R}^+$. Then there exists a function $f \colon \mathbb{R}^+ \to \mathbb{R}^+$ such that $P(|\int_0^t \mathbb{1}_A \, \mathrm{d}X| > K) < f(K)$ and $\lim_{K \to \infty} f(K) \to 0$. Hence

$$\begin{aligned}
P((b-a)U_{[a,b]}^{X_{t_n}} > K) &\leq P(\int \mathbb{1}_B \, \mathrm{d}X + (a - X_{t_n})^+ > K) \\
&= P(|\int \mathbb{1}_B \, \mathrm{d}X| + (a - X_{t_n})^+ > K) \\
&\leq P(|\int \mathbb{1}_B \, \mathrm{d}X| > \frac{K}{2}) + P((a - X_{t_n})^+ > \frac{K}{2}) \\
&\leq f(\frac{K}{2}) + P((a - X_{t_n})^+ > \frac{K}{2})
\end{aligned}$$

Note that $\int \mathbb{1}_B \, \mathrm{d}X \leq 0$ and **for a given** t_n, we move $K \to \infty$ after, then $\lim_{K \to \infty} P((b-a)U_{[a,b]}^{X_{t_n}} > K) = 0$ or $P(U_{[a,b]}^{X_{t_n}} = \infty) = 0$ or $U_{[a,b]}^{X_{t_n}}$ is finite for all n.

Finally, from lemma 9.3.4, X_{t_n} almost surely converges.

• There is a countable subset $S \subset \mathbb{R}_+$ for which Y_t is right-continuous at every $t \notin S$. https://almostsuremath.com/2009/12/18/cadlag-modifications/

10.3.2 Corollary

If now X is right continuous in probability, then X has a version Y_t which is cadlag for all t.

10.4 Cadlag martingales

10.4.1 Lemma

Let X be a submartingale (9.1.4). Then, it has a version Y which has left and right limit for all t, and there is a countable subset $S \subset \mathbb{R}_+$ for which Y_t is right-continuous at every $t \notin S$.

Proof

We see that X is integrable since X is submartingale, moreover a set $A \subseteq [0, t]$ can be expressed by $A = \bigcup_{k=1}^{N} [l_k, u_k]$, N can be infinity. Then on one hand, $\forall A$:

$$\mathbb{E}[\int_0^t \mathbb{1}_A \, dX] = \mathbb{E}[\sum_{k=1}^{N} (X_{u_k} - X_{l_k})] = \sum_{k=1}^{N} (\mathbb{E}[X_{u_k}] - \mathbb{E}[X_{l_k}]) \geq 0$$

since X_t is submartingale implies $\mathbb{E}[X_{u_k}] \geq \mathbb{E}[X_{l_k}]$. From this we infer that $\mathbb{E}[\int_0^t \mathbb{1}_{A^C} \, dX]$ is positive since A^C is also a subset of $[0, t]$.

On the other hand,

$$\mathbb{E}[\int_0^t \mathbb{1}_A \, dX] = \mathbb{E}[\int_0^t \mathbb{1}_{[0,t]} \, dX] - \mathbb{E}[\int_0^t \mathbb{1}_{A^C} \, dX] \leq \mathbb{E}[X_t] - \mathbb{E}[X_0] < \infty$$

since X is integrable.

Finally, $\mathbb{E}[\int_0^t \mathbb{1}_A \, dX]$ is bounded, for all $A \subseteq [0, t]$ then by theorem 10.3, we have Q.E.D.

Remark This result is also valable for supermartingle (just replacing X by $-X$) and martingale.

10.4.2 Lemma

Suppose that a submartingale X is adapted to a right-continuous filtration : $\mathcal{F}_t = \bigcap_{\varepsilon > 0} \mathcal{F}_{t+\varepsilon}$ or $\mathcal{F}_t = \mathcal{F}_{t+}$, which means that no information is added with infinitesimal step.

Then X has a cadlag version if and only if $t \mapsto \mathbb{E}[X_t]$ is right-continuous.

Proof

By lemma 10.4.1, since X_t is a submartingale, it has a version Y_t which has left and right limit for all t, such that there is a countable subset $S \subset \mathbb{R}_+$ for which Y_t is right-continuous at every $t \notin S$.

In direct sense, we show that if $t \mapsto \mathbb{E}[X_t]$ is right-continuous then X has a cadlag version. Thus we need to prove that :

$$\lim_{t \to a^+} Y_t = Y_a, \quad \forall a \in S$$

or equivalently, if t_n is a decreasing sequence to a:

$$\lim_{n \to \infty} Y_{t_n} = Y_a, \quad \forall a \in S$$

Here is a sketch of proof

1. $Y_{t_n} \xrightarrow[n \to \infty]{a.s} Y_{a+}$ since Y_t has right limit.

2. Y_{t_n} is uniformly integrable, by lemma 9.1.8.

3. $\lim\limits_{n \to \infty} \mathbb{E}[Y_{t_n}] = \mathbb{E}[Y_{a+}]$. This is because $\lim\limits_{n \to \infty} \mathbb{E}[|Y_{t_n} - Y_{a+}|] = 0$ (convergence in mean) by lemma 4.5.4

4. $\lim\limits_{n \to \infty} \mathbb{E}[Y_{t_n}] = \mathbb{E}[Y_a]$. This is because $\mathbb{E}[X_t]$ is right-continuous, then $\mathbb{E}[Y_t]$ is also right continuous.

5. $\mathbb{E}[Y_a] = \mathbb{E}[Y_{a+}]$, then $\mathbb{E}[Y_a \mid \mathcal{F}_a] = \mathbb{E}[Y_{a+} \mid \mathcal{F}_a]$, then $Y_a = \mathbb{E}[Y_{a+} \mid \mathcal{F}_a]$ (since $Y_a = \mathbb{E}[Y_a \mid \mathcal{F}_a]$), then $Y_a = \mathbb{E}[Y_{a+} \mid \mathcal{F}_{a+}]$ (since $\mathcal{F}_a = \mathcal{F}_{a+}$), then $Y_a = Y_{a+}$ (since $Y_{a+} = \mathbb{E}[Y_{a+} \mid \mathcal{F}_{a+}]$). We have Q.E.D.

In converse sense, it is trivial by using the two last points in above proof.

11 Martingale convergence with continuous index

In sec 9.3.5, we've seen the convergence for martingale with discret index. In this section, we work with continuous index. We start first with stochastic processes which are only right-continous (e.g. cadlag), then continuous martingale stochastic processes.

11.1 Only right-continuous

$$\lim_{t \to a+} X_t = X_a$$

11.1.1 Doob's first martingale convergence theorem

Let's X satify three folllowing conditions:

1. right continuous

2. L_1-bounded $(\sup\limits_{t \geq 0} \mathbb{E}[|X_t|] < C)$

3. submartingale or supermartingale or martingale.

Then, the limit $\lim\limits_{t \to \infty} X_t = X_\infty$ exists with probability one (a.s. convergence) and $\mathbb{E}[X_\infty]$ is finite.

Proof We prove for case that X is submartingale. The case of supermartingale can be similar with $-X$. Let's F_t the filtration of X. We define the stochastic process Y by

$$Y_t = \begin{cases} X_{\frac{t}{1-t}} & \text{if } t < 1 \\ 0 & \text{if } t \geq 1 \end{cases}$$

with $G_t = F_{\frac{t}{1-t}}$ when $t < 1$ and $G_t = F_\infty$ otherwise. By consequence, Y_t is right continuous, L_1-bounded with all $t \geq 0$ and Y_t is submartingale with $t < 1$.

Now the main problem is to show that $\lim_{t \to 1^-} Y_t$ converges a.s. We will show that Y_t is a cadlag by 10.3 (with the first condition). Note that Y_t is already right continuous and L_1-bounded, then we only need to show that, $\forall t$, $\mathbb{E}[\int_0^t \mathbb{1}_A \, dY]$ is bounded for all $A \subseteq \mathbb{R}^+$.

Let's $A = \bigcup_{k=1}^N [l_k, u_k]$, N can be infinity. Suppose that l_m is the first time moment that is greater than 1. We have

$$\mathbb{E}[\int_0^t \mathbb{1}_A \, dY] = \mathbb{E}[\sum_{k=1}^{m-1} (Y_{u_k} - Y_{l_k})]$$

$$= \mathbb{E}[Y_{u_{m-1}}] - \mathbb{E}[Y_{l_{m-1}}] + \sum_{k=1}^{m-2} (\mathbb{E}[Y_{u_k}] - \mathbb{E}[Y_{l_k}])$$

$$\geq \max(0, C) + 0$$

and

$$\mathbb{E}[\int_0^t \mathbb{1}_A \, dY] = \mathbb{E}[\sum_{k=1}^{m-1} (Y_{u_k} - Y_{l_k})]$$

$$= \mathbb{E}[Y_{u_{m-1}}] - \sum_{k=m-1}^{2} (\mathbb{E}[Y_{l_k}] - \mathbb{E}[Y_{u_{k-1}}]) - \mathbb{E}[Y_{l_1}]$$

$$\leq 2C$$

Then $\mathbb{E}[\int_0^t \mathbb{1}_A \, dY]$ is bounded. Therefore, we have Y is cadlag, it means it has left limit or $\lim_{t \to 1^-} Y_t$ converges a.s. or $\lim_{t \to \infty} X_t \to X_\infty$.

11.1.2 Corollary

If X is a nonnegative right-continuous martingale (or supermartingale) then the limit $X_\infty = \lim_{t \to \infty} X_t$ exists and is finite, with probability one.

Proof

The supermartingale property gives $0 \leq \mathbb{E}[|X_t|] = \mathbb{E}[X_t] \leq \mathbb{E}[X_0]$ showing that X is L_1 bounded. So the previous theorem 11.1.1 applies.

11.1.3 Theorem

Let X be

1. right-continuous

2. uniformly integrable

3. martingale

Then, the limit $X_\infty = \lim_{t \to \infty} X_t$ exists with probability one. Furthermore, this is the unique (up to a zero probability set) integrable and $\mathcal{F}_\infty$-measurable random variable such that $X_t = \mathbb{E}[X_\infty \mid \mathcal{F}_t]$ for each $t \in \mathbb{R}_+$.

Proof

- As uniformly integrable process X infers that X is L_1-bounded, then theorem 11.1.1 shows that the limit $X_\infty = \lim_{t \to \infty} X_t$ exists and is almost surely finite.

- The uniformly integrable allows us to appy the limit for expectation : $\lim_{s \to \infty} \mathbb{E}[X_s \mid \mathcal{F}_t] = \mathbb{E}[X_\infty \mid \mathcal{F}_t] = X_t$. This is by using lemma 4.5.4 with $\lim_{t \to \infty} X_t = X_\infty$ almost surely then we infer that X_t converges in probability to X_∞.

- If there is Z such that $\mathbb{E}[Z \mid \mathcal{F}_t] = X_t, \forall t$. Then

$$
\begin{aligned}
\mathbb{E}[\mathbb{1}_A(X_\infty - Z)] &= \mathbb{E}[X_\infty . \mathbb{1}_A] - \mathbb{E}[Z.\mathbb{1}_A] \\
&= \mathbb{E}[X_\infty \mid \mathcal{F}_t](A)P(A) - \mathbb{E}[Z \mid \mathcal{F}_t](A)P(A) \\
&= X_t(A)P(A) - X_t(A)P(A) \\
&= 0 \quad \forall A \in \mathcal{F}_t
\end{aligned}
$$

  Then by the equality by expectation 2.5, we have $X_\infty = Z$.

  This is the continous version of 9.3.6

11.2 Continuous

$$\lim_{t \to a} X_t = X_a$$

11.2.1 Continuous martingale convergence

Let X be a **continuous martingale**. Then, almost surely, one of the following is satisfied :

- $X_\infty = \lim_{t \to \infty} X_t$ exists and is finite.

- $\limsup_{t \to \infty} X_t = +\infty$ and $\liminf_{t \to \infty} X_t = -\infty$. In this case, the process hits every value in $\mathbb{R}$ at arbitrarily large times.

We can separate into two cases :

- $\sup_{t \geq 0} \mathbb{E}[|X_t|] < C$. We use the theorem 11.1.1 and we have the first statement.

- $\sup\limits_{t\geq 0} \mathbb{E}[|X_t|] = \infty$. In this case, we can consider the continuous Brownian motion, which is a continuous martingale. Thus continuous Brownian motion is not L_1-bounded and X_t can reach $+\infty$ or $-\infty$ if $t \to \infty$.

11.2.2 Continuous submartingale convergence

Let X be a **continuous submartingale**. Then, almost surely, one of the following is satisfied.

- $X_\infty = \lim\limits_{t\to\infty} X_t$ exists and is finite.

- $\limsup\limits_{t\to\infty} X_t = +\infty$.

The same as in 11.2.1, except that this is submartingale, then $\liminf\limits_{t\to\infty} X_t = -\infty$ is not possible.

12 Doob's martingale inequality

12.1 Statement of the inequality

12.1.1 Discret case

Let $X_1, \ldots, X_n$ be a discrete-time submartingale relative to a filtration $\mathcal{F}_1, \ldots, \mathcal{F}_n$ and $C > 0$. Then

$$P(\max_{1\leq i\leq n} X_i \geq C) \leq \frac{\mathbb{E}[X_n \mathbb{1}_{\{\max\limits_{1\leq i\leq n} X_i \geq C\}}]}{C} \leq \frac{\mathbb{E}[\max(X_n, 0)]}{C} = \frac{\mathbb{E}[X_n^+]}{C}$$

Proof
Let define sets:

$$W_i = \{\omega \colon X_i \geq C \text{ and } X_j < C, \forall j < i\}$$

therefore, we have that W_i are pairwise disjoint and

$$P(\max_{1\leq i\leq n} X_i \geq C) = P(\bigcup_{i=1}^{n} W_i) = P(W)$$

where $W = \bigcup\limits_{i=1}^{n} W_i = \{\omega \colon \max\limits_{1\leq i\leq n} X_i(\omega) \geq C\}$.

$$CP(W_i) = C \int_{W_i} \mathrm{d}P$$

$$= \int_{W_i} C \, \mathrm{d}P$$

$$\leq \int_{W_i} X_i \, \mathrm{d}P$$

$$\leq \int_{W_i} \mathbb{E}[X_n \mid \mathcal{F}_i](W_i) \, \mathrm{d}P$$

$$= \int_{W_i} X_n \, \mathrm{d}P$$

where the last equality is by 2.10. Since W_i are pairwise disjoint sets, then

$$CP(W) = CP(\bigcup_{i=1}^{n} W_i)$$

$$= \sum_{i=1}^{n} CP(W_i)$$

$$\leq \sum_{i=1}^{n} \int_{W_i} X_n \, \mathrm{d}P$$

$$= \int_{W} X_n \, \mathrm{d}P \quad \text{(by disjoint sets)}$$

$$= \int_{\Omega} X_n \mathbb{1}_W \, \mathrm{d}P$$

$$= \mathbb{E}[X_n \mathbb{1}_W]$$

$$\leq \mathbb{E}[\max(X_n, 0)]$$

where we have the last inequality because 2.7.9.2

Note that, if the submartingale is restricted to a random variable and non-negetive then we have Markov inequality 2.7.1.

12.1.2 Continuous case

Let X_t be a right-continuous submartingale then

$$P(\sup_{0 \leq t \leq T} X_t \geq C) \leq \frac{\mathbb{E}[X_T \mathbb{1}_{\{\sup_{0 \leq t \leq T} X_t \geq C\}}]}{C} \leq \frac{\mathbb{E}[\max(X_T, 0)]}{C} = \frac{\mathbb{E}[X_T^+]}{C}$$

Proof

The proof can be inspired from discret case 12.1.1. Given a fixed ω, we consider the first time at which the process reaches the value C (first hitting):

$$\tau(\omega) = \inf_{0 \leq t \leq T} X_t(\omega) \geq C$$

where τ now is also a random variable. From this defintion, we can trivially show that :

$$\{\omega : \sup_{0 \leq t \leq T} X_t \geq C\} = \{\omega : \tau(\omega) \leq T\} = W$$

Hence,

$$C\mathbb{1}_W \leq X_\tau \mathbb{1}_W$$
$$\Leftrightarrow C\mathbb{E}[\mathbb{1}_W] \leq \mathbb{E}[X_\tau \mathbb{1}_W]$$
$$\Leftrightarrow CP(W) \leq \mathbb{E}[X_\tau \mathbb{1}_W]$$
$$\leq \mathbb{E}[X_T \mathbb{1}_W]$$
$$\leq \mathbb{E}[X_T^+]$$

where the before last inequality is by 9.1.4.4 and the last inequality is by 2.7.9.2.

12.2 Further inequalities

12.2.1 Inequality I

Let X_t be a right-continuous martingale or right-continuous positive submartingale. Then with $p \geq 1$:

$$P(\sup_{0 \leq t \leq T} |X_t| \geq C) \leq \frac{\mathbb{E}[|X_T|^p]}{C^p}$$

Proof

- In case that X_t is martingale, then with $f(x) = |x|^p$ where $p \geq 1$ is a convex function and by lemma 9.1.4.1, we have $|X_t|^p$ is submartingale.

- In case that X_t is positive submartingale, then with $f(x) = |x|^p$ where $p \geq 1$ is a convex function and f is increasing for $x \leq 0$, by lemma 9.1.4.2, we have $|X_t|^p$ is submartingale.

Therefore, in both cases, $|X_t|^p$ is submartingale, then by inequality 12.1.2, we have :

$$P(\sup_{0 \leq t \leq T} |X_t|^p \geq C) \leq \frac{\mathbb{E}[|X_T|^p \mathbb{1}_{\{\sup_{0 \leq t \leq T} |X_T|^p \geq C\}}]}{C} \leq \frac{\mathbb{E}[\max(|X_T|^p, 0)]}{C} = \frac{\mathbb{E}[|X_T|^p]}{C}$$

Now we replace the constant C by C^p, then

$$P(\sup_{0\le t\le T}|X_t|^p\ge C^p)\le\frac{\mathbb{E}[|X_T|^p\mathbb{1}_{\{\sup_{0\le t\le T}|X_T|^p\ge C^p\}}]}{C^p}\le\frac{\mathbb{E}[|X_T|^p]}{C^p}$$

$$\Leftrightarrow P(\sup_{0\le t\le T}|X_t|\ge C)\le\frac{\mathbb{E}[|X_T|^p\mathbb{1}_{\{\sup_{0\le t\le T}|X_T|\ge C\}}]}{C^p}\le\frac{\mathbb{E}[|X_T|^p]}{C^p}$$

12.2.2 Inequality II

Let X_t is a non-negative submartingale and $p>1$, then

$$\mathbb{E}[|X_T|^p]\le\mathbb{E}\left[\sup_{0\le t\le T}|X_t|^p\right]\le\left(\frac{p}{p-1}\right)^p\mathbb{E}[|X_T|^p]$$

Sometimes known as Doob's maximal inequality. The first inequality is evident since $\sup_{0\le t\le T}|X_t|^p\ge|X_T|^p$. For short, we denote $X_T^*=\sup_{0\le t\le T}X_t$.

Proof The non-negative is because we are in case that x^p and it requires that $x\ge 0$. By using (twice) the expectation with exponentiation 2.3 :

$$\mathbb{E}[(X_T^*)^p]=p\int_0^\infty t^{p-1}\mathbb{E}[\mathbb{1}_{\{X_T^*>t\}}]\,\mathrm{d}t\quad\text{(first use)}$$

$$=p\int_0^\infty t^{p-1}P(X_T^*>t)\,\mathrm{d}t$$

$$=p\int_0^\infty t^{p-2}tP(X_T^*>t)\,\mathrm{d}t$$

$$\le p\int_0^\infty t^{p-2}\mathbb{E}[X_T\mathbb{1}_{\{X_T^*>t\}}]\,\mathrm{d}t\quad\text{inequality 12.1.2}$$

$$=\frac{p}{p-1}\mathbb{E}[X_T(X_T^*)^{p-1}]\quad\text{(second use)}$$

$$=\frac{p}{p-1}\mathbb{E}[(X_T)^p]^{\frac{1}{p}}\mathbb{E}[(X_T^*)^{(p-1)q}]^{\frac{1}{q}}\quad\text{(Holder inequality 2.7.6)}$$

$$=\frac{p}{p-1}\mathbb{E}[(X_T)^p]^{\frac{1}{p}}\mathbb{E}[(X_T^*)^p]^{\frac{p-1}{p}}\quad\text{by taking }q=\frac{p}{p-1}$$

We infer that :

$$\mathbb{E}[(X_T^*)^p]^{\frac{1}{p}}\le\frac{p}{p-1}\mathbb{E}[(X_T)^p]^{\frac{1}{p}}$$

$$\Leftrightarrow\mathbb{E}[(X_T^*)^p]\le\left(\frac{p}{p-1}\right)^p\mathbb{E}[(X_T)^p]$$

12.2.3 Inequality III

Let X_t is a non-negative submartingale, then

$$\mathbb{E}[X_T^*] \leq \frac{e}{e-1}\left(1 + \mathbb{E}[X_T \ln X_T]\right)$$

12.2.3.1 Lemma

$$x \ln\left(\max(1, y)\right) + \min(1, y) \leq x \ln x + e^{-1}y + 1$$

Let's consider :

$$f(x) = x \ln x - x \ln\left(\max(1, y)\right) + e^{-1}y + 1 - \min(1, y)$$

and its differential :

$$f'(x) = \ln x + 1 - \ln\left(\max(1, y)\right)$$

Then the first-order optimality condition give :

$$x_0 = \exp\left(\ln\left(\max(1, y)\right)\right) - 1 = e^{-1}\left(\max(1, y)\right)$$

Replace in f :

$$f(x_0) = e^{-1}(\max(1, y))(\ln\left(\max(1, y)\right) - 1) - e(\max(1, y))\ln\left(\max(1, y)\right) + e^{-1}y + 1 - \min(1, y)$$

$$= -e^{-1}(\max(1, y)) + e^{-1}y + 1 - \min(1, y)$$

$$= e^{-1}(y - \max(1, y)) + 1 - \min(1, y)$$

$$= e^{-1}(\min(1, y) - 1) + 1 - \min(1, y)$$

$$= (1 - \min(1, y))(1 - e^{-1})$$

$$\geq 0$$

12.2.3.2 Proof

Since X_T^* is non-negative, then by 2.4.2 :

$$\mathbb{E}[(X_T^* - 1)^+] = \int_1^\infty P(X_T^* \geq x)\,\mathrm{d}x$$

$$\leq \int_1^\infty \frac{\mathbb{E}[X_T \mathbb{1}_{\{X_T^* \geq x\}}]}{x}\,\mathrm{d}x$$

$$= \int_1^\infty \frac{1}{x}\int_\Omega X_T \mathbb{1}_{\{X_T^* \geq x\}}\,\mathrm{d}P\,\mathrm{d}x$$

$$= \int_\Omega X_T \int_1^{X_T^*} \frac{1}{x}\,\mathrm{d}x\,\mathrm{d}P$$

$$= \int_\Omega X_T \ln\left(\max(1, X_T^*)\right)\,\mathrm{d}P$$

$$= \mathbb{E}[X_T \ln\left(\max(1, X_T^*)\right)]$$

Finally,

$$
\begin{aligned}
\mathbb{E}[X_T^*] &= \mathbb{E}[(X_T^* - 1)^+] + \mathbb{E}[(X_T^* - 1)^-] \\
&\leq \mathbb{E}[X_T \ln\left(\max(1, X_T^*)\right)] + \mathbb{E}[(X_T^* - 1)^-] \\
&= \mathbb{E}[X_T \ln\left(\max(1, X_T^*)\right)] + \mathbb{E}[\min(1, X_T^*)] \\
&\leq \mathbb{E}[X_T \ln X_T + e^{-1} X_T^* + 1] \\
&= \mathbb{E}[X_T \ln X_T] + 1 + e^{-1}\mathbb{E}[X_T^*]
\end{aligned}
$$

or

$$
\mathbb{E}[X_T^*] \leq \frac{e}{e-1}\left(1 + \mathbb{E}[X_T \ln X_T]\right)
$$

13 Local martingale

In a simple intuition, a local martingale is a stochastic process which is locally a martingale.

13.1 Definition

A process X is a local martingale if there exists an increasing sequence of stopping times τ_k where $\lim\limits_{k \to +\infty} \tau_k \to +\infty$, such that the stopped process (in sec 1.11.4)

$$
X_t^{\tau_k}
$$

is a martingale for each k.

13.1.1 Properties

- Every martingale is a local martingale but the inverse is not true.

- Every bounded local martingale is a martingale

- Local martingale that is bounded from below is a supermartingale

- Local martingale that is bounded from above is a submartingale

14 Semimartingale

A real valued stochastic process X defined on the filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t \geq 0}, P)$ is called a semimartingale if it can be decomposed as :

$$
X_t = M_t + A_t
$$

where M is a local martingale and A is a cadlag (continue à droite, limite à gauche) adapted process of locally bounded variation.

14.0.1 Properties

1. The semimartingales **form the largest class** of **stochastic processes** for which the Itô integral can be defined. This means that if a stochastic process is semimartingale, then its Itô integral exists.

2. Linear combinations of semimartingales are semimartingales.

3. Products of semimartingales are semimartingales, which is a consequence of the integration by parts formula for the Itô integral.

15 Feller process

Before giving the definition of Feller process, let's define the Feller semigroup.

15.0.1 Feller semigroup

Let's define space $C_0(E)$ in the following manner:

- E is a locally compact Hausdorff space with a countable base, (a space that contains particular functions).

- $C_0(E) = \{f \in E | f \text{ is continous} \cap \lim_{x \to \pm\infty} f(x) = 0 \cap \exists \|f\|_\infty\}$ (continous, vanishing at infinity and quipped with the sup-norm).

Let $T_t(t \geq 0)$ be a *positive linear map* from $C_0(E)$ to itself, which means with $f \in C_0(E)$ then $T_t f := T_t(f) \in C_0(E)$, note that the input of T_t is a function, not a value.

A Feller semigroup on $C_0(E)$ is a collection of T_t such that :

1. Given $f \in C_0(E) : \|T_t f\| \leq \|f\|, \forall t \geq 0$ (contraction)

2. $T_{t+s} = T_t \circ T_s, \forall s, t \geq 0$ (the semigroup property)

3. $\lim_{t \to 0^+} \|T_t f - f\| = 0, \forall f \in C_0(E)$

From the second property (semigroup) and the third property :

$$\lim_{t \to t_0^+} T_t f = \lim_{s \to 0^+} T_{t_0+s} f = \lim_{s \to 0^+} T_{t_0} \circ T_s f = T_{t_0} \lim_{s \to 0^+} T_s f = T_{t_0} f$$

This means that $T_t f$ is right continuous with respect to t.

15.0.2 Definition

A Feller process is a Markov process (sec 7) with a Feller transition function.

16 Stock simulation

16.1 Monte Carlo simulation

Why do we usually use returns and not prices to model financial data in time series analysis ? Because prices usually have a unit root, while returns can be assumed to be stationary and it is eventually considered to be normal distribution.

Then for return rates of given protfolio, with arithmetic return $r_n = \frac{S_{n+1} - S_n}{S_n}$, its μ and σ are estimated by past data :

$$\begin{cases} \mu = \mathbb{E}\left[r_n\right] \\ \sigma = \sqrt{Var\left[r_n\right]} \end{cases}$$

where S_n is the price of stock at instant n. μ and σ are the mean and std of return from the time n to $n+1$. Then, for a future price simulation, we can sample :

$$\frac{S_{n+1}}{S_n} - 1 \leftarrow \mathcal{N}(\mu, \sigma^2)$$

With a given S_0, we can have a determinist evoluation of price from S_0 to S_N, notably :

$$E[S_n] = (1+\mu)^n$$

$$Var[S_n] = (\sigma^2 + (1+\mu)^2)^n - (1+\mu)^{2n}$$

16.2 By stochastic differential equation

Suppose that the time from instant n to $(n+1)$ is Δ. Then we can express r is a function of time t:

$$r(t) = \mu\frac{t}{\Delta} + \sigma B_{\frac{t}{\Delta}}$$

Note that if $t = \Delta$, we get $r \leftarrow \mathcal{N}(\mu, \sigma^2)$. For the simplicity, we set Δ as a unit as 1.

$$r(t) = \mu t + \sigma B_t$$

where $r_n = \log(\frac{S_{n+1}}{S_n})$ and

$$\begin{cases} \mu = \mathbb{E}\left[r_n\right] \\ \sigma = \sqrt{Var\left[r_n\right]} \end{cases}$$

Taking the derivative in term t :

$$dr(t) = \mu dt + \sigma dB_t$$

Remind that $r(t)$ is the return rate after a **time length** t, then we have:

$$r(t) = \log\left(\frac{S_{t_0+t}}{S_{t_0}}\right)$$

For simplicity, set $t_0 = 0$. Taking the derivative in term t :

$$dr(t) = \frac{dS_t}{S_t}$$

Hence, by matching $\frac{dS_t}{S_t}$ and $\mu dt + \sigma dB_t$ via $dr(t)$:

$$\frac{dS_t}{S_t} = \mu dt + \sigma dB_t$$

$$dS_t = \mu S_t dt + \sigma S_t dB_t$$

This leads to the origine for the Geometric Brownian Motion (see 2.4). By applying the result of GBM :

$$S_t = S_0 \exp\left((\mu - \frac{1}{2}\sigma^2)t + \sigma B_t\right)$$

or if $t = \Delta \to 0$:

$$\frac{S_\Delta - S_0}{S_0} = \log\frac{S_\Delta}{S_0} = (\mu - \frac{1}{2}\sigma^2)\Delta + \sigma B_\Delta$$

Simulation. For a future price simulation, we can sample :

$$\frac{S_{n+1}}{S_n} = \exp\left((\mu - \frac{1}{2}\sigma^2) + \sigma x\right)$$

where $x \leftarrow \mathcal{N}(0,1)$ or taking directly :

$$S_{n+1} = S_0 \exp\left((\mu - \frac{1}{2}\sigma^2)n + \sigma\sqrt{n}x\right)$$

where $x \leftarrow \mathcal{N}(0,1)$.

16.3   The difference between two methods

The key difference between Monte Carlo simulation method and the stochastic differential equation (SDE) approach is the **arithmetic return and the logarithmic return**. Consequently, this is why we have two distinct means of return :

- $\mu_a + 1$: a means arithmetic
- $\mu_l - \frac{1}{2}\sigma^2$: l means logarithmic

17   Time series

Given a time serie signal (or process) y_t, it can contain the main following components :

- T_t the trend component at time t, which reflects the long-term progression of the series (secular variation). A trend exists when there is a persistent increasing or decreasing direction in the data. The trend component does not have to be linear.

- C_t the cyclical component at time t, which reflects repeated but non-periodic fluctuations. The duration of these fluctuations depend on the nature of the time series.

- S_t the seasonal component at time t, reflecting seasonality (seasonal variation). A seasonal pattern exists when a time series is influenced by seasonal factors. Seasonality occurs over a fixed and known period (e.g., the quarter of the year, the month, or day of the week).

- I_t, the irregular component (or "noise") at time t, which describes random, irregular influences. It represents the residuals or remainder of the time series after the other components have been removed.

If y_t is additive model:

$$y_t = T_t + C_t + S_t + I_t$$

If y_t is multiplicative model:

$$y_t = T_t \times C_t \times S_t \times I_t$$

An additive model would be used when the variations around the trend do not vary with the level of the time t, whereas a multiplicative model would be appropriate if the trend is proportional to the level of the time series.

17.1   ETS Error-Trend-Seasonality

TO DO

17.2   Hodrick–Prescott filter

y_t is decomposed into $y_t = \tau_t + c_t$, where τ_t trend component or smoothed version of y_t and c_t is cyclical component.

In two-sided version (with respect to τ_t), then trend component is determined by :

$$\min_{\tau} \left(\sum_{t=1}^{T} (y_t - \tau_t)^2 + \lambda \sum_{t=2}^{T-1} [(\tau_{t+1} - \tau_t) - (\tau_t - \tau_{t-1})]^2 \right)$$

The first term, try to make y_t closed to τ_t and the second term penalizes the second order, which prevents τ to varies to fast.

The Hodrick–Prescott filter will only be optimal when:

- Data exists in a I(2) trend (I(2) is order of integration sec 4.6.2). If one-time permanent shocks or split growth rates occur, the filter will generate shifts in the trend that do not actually exist.

- Noise in data is approximately normally distributed.

- Contain all histotical data (from the beginning), constrast to moving average.

One-sided version ?

$$\min_{\tau} \left(\sum_{t=1}^{T} (y_t - \tau_t)^2 + \lambda \sum_{t=3}^{T} [(\tau_t - \tau_{t-1}) - (\tau_{t-1} - \tau_{t-2})]^2 \right)$$

17.3 Autoregressive model

AR(p) :

$$X_t = \sum_{i=1}^{p} \phi_i X_{t-i} + \varepsilon_t$$

Fitting methods :

- Ordinary Least Squares (OLS)

- Maximum Likelihood Estimation (MLE)

- Yule-Walker with autocovariance or autocorrelation.

MLE is very general and can apply to various model fitting, in this case, we have that ε_t follows normal distribution, hence :

$$f(X_t; \phi_1, \cdots, \phi_p) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(X_t - \sum_{i=1}^{p} \phi_i X_{t-i})^2}{2\sigma^2} \right)$$

Note that, before using the AR model, we need to check if there is a unit root (sec 4.6.1) in signal by Dickey–Fuller test (sec 4.6) or augmented Dickey–Fuller test (sec 4.7).

17.3.1 Yule-Walker

AR(p) :

$$X_t = \sum_{i=1}^{p} \phi_i X_{t-i} + \varepsilon_t$$

$$\gamma_m = Cov(X_t, X_t - m)$$

$$= Cov(\sum_{i=1}^{p} \phi_i X_{t-i} + \varepsilon_t, X_t - m)$$

$$= \sum_{i=1}^{p} \phi_i \gamma_{m-i} + Cov(\varepsilon_t, X_t - m)$$

where
$$Cov(\varepsilon_t, X_t - m) = \begin{cases} \sigma_\varepsilon^2 \text{ if } m = 0 \\ 0 \text{ otherwise} \end{cases}$$

Taking al least p equations $m = 1, \cdots, p$ to sovle for p value $\phi_1, \cdots, \phi_p$ and taking $m = 0$ for calculating σ_ε.

17.4 Moving Average model

MA(q):
$$X_t = \mu + w_t + \theta_1 w_{t-1} + \ldots + \theta_q w_{t-q}$$

where $w_t, \ldots, w_{t-q} \overset{iid}{\sim} \mathcal{N}(0, \sigma_w^2)$

17.4.1 MA(1)

In case $q = 1$ or MA(1):

$$X_t = \mu + w_t + \theta_1 w_{t-1}$$

- $E(X_t) = \mu$
- $Var(X_t) = (1 + \theta_1^2)\sigma_w^2$
- $ACF(X_t, X_{t-1}) = \frac{\theta_1}{1+\theta_1^2}$
- $ACF(X_t, X_{t-h}) = 0$, if $h > 1$.

Remind a way to calculate ACF (Autocorrelation function):

$$ACF(X_t, X_{t-h}) = \frac{Cov(X_t, X_{t-h})}{\sigma_X^2}$$

There is no unique connection between values of θ_1 and $ACF(h = 1)$ in MA(1). It's means for a given value of $ACF(h = 1) = 0.4$, it may happen that $\theta_1 = 0.5$ or $\theta_1 = 2$.

17.4.2 Invertibility

Thus, AR model and MA can be transformed, one to other if several contraints are satisfied.

From AR to MA :

$$
\begin{aligned}
x_t &= \phi_1 x_{t-1} + \varepsilon_t \\
&= \phi_1(\phi_1 x_{t-2} + \varepsilon_{t-1}) + \varepsilon_t \\
&= \phi_1^2 x_{t-2} + \phi_1 \varepsilon_{t-1} + \varepsilon_t \\
&= \ldots \\
&= \phi_1^h x_{t-h} + \phi_1^{h-1} \varepsilon_{t-(h-1)} + \ldots + \phi_1^3 \varepsilon_{t-3} + \phi_1^2 \varepsilon_{t-2} + \phi_1 \varepsilon_{t-1} + \varepsilon_t
\end{aligned}
$$

If $|\phi_1| < 1$ and h is sufficiently big, we can remove the term $\phi_1^h y_{t-h}$ and it becomes MA model.

From MA to AR :

$$
\begin{aligned}
x_t &= w_t + \theta_1 w_{t-1} \\
&= w_t + \theta_1 x_{t-1} - \theta_1^2 w_{t-2} \\
&= \ldots \\
&= w_t + \theta_1 x_{t-1} - \theta_1^2 x_{t-2} + \theta_1^3 x_{t-3} - \theta_1^4 x_{t-4} + \ldots + \theta_1^{2h-1} x_{t-(2h-1)} - \theta_1^{2h} x_{t-2h}
\end{aligned}
$$

If $|\theta_1| < 1$ then from a sufficient big h, we can cut off small signals.

17.4.3 Fiting

For a MA(2) : $y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2}$. Given y as observations, we need to find θ and ε.

$$
\begin{bmatrix} y_2 \\ y_3 \\ \vdots \\ y_{t+1} \end{bmatrix}
\approx
\begin{bmatrix} \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_{t+1} \end{bmatrix}
+ \theta_1
\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_t \end{bmatrix}
+ \theta_2
\begin{bmatrix} \varepsilon_0 \\ \varepsilon_1 \\ \vdots \\ \varepsilon_{t-1} \end{bmatrix}
=
\begin{bmatrix} \varepsilon_2 & \varepsilon_1 & \varepsilon_0 \\ \varepsilon_3 & \varepsilon_2 & \varepsilon_1 \\ \vdots & \vdots & \vdots \\ \varepsilon_{t+1} & \varepsilon_t & \varepsilon_{t-1} \end{bmatrix}
\times
\begin{bmatrix} 1 \\ \theta_1 \\ \theta_2 \end{bmatrix}
$$

By using iterative procedure, we fit alternatively ε and θ :

1. Initializing θ_1, θ_2, ε_0, ε_1

2. Calculating for $\varepsilon_2, \cdots, \varepsilon_{t+1}$

3. Consider that $[\varepsilon_2, \cdots, \varepsilon_{t+1}]^T$ is noise, then fitting for only θ by minimizing

   with LS as :
   $$
   \begin{bmatrix} y_2 \\ y_3 \\ \vdots \\ y_{t+1} \end{bmatrix}
   \approx
   \begin{bmatrix} \varepsilon_1 & \varepsilon_0 \\ \varepsilon_2 & \varepsilon_1 \\ \vdots & \vdots \\ \varepsilon_t & \varepsilon_{t-1} \end{bmatrix}
   \times
   \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}
   $$

4. Repeat from step 2.

A python for this is found in sec **??**.

17.5 Akaike Information Criterion (AIC)

The Akaike information criterion (AIC) is an estimator of prediction error and thereby relative quality of statistical models for a given set of data. Given a collection of models for the data, AIC estimates the quality of each model, relative to each of the other models. Thus, AIC provides a means for model selection. With k number of parameters and $\hat{\mathcal{L}}$ is maximum likehood of fitted model :

$$AIC = -2ln(\hat{\mathcal{L}}) + 2k$$

The preferred model, from a selection of models, has the minium AIC of the group. You can see that the AIC grows as the number of parameters, k, increases, but is reduced if the negative log-likelihood increases. Essentially it penalises models that are overfit.

17.6 Bayesian Information Criterion (BIC)

Essentially it has similar behaviour to the AIC in that it penalises models for having too many parameters. This may lead to overfitting. The difference between the BIC and AIC is that the BIC is more stringent with its penalisation of additional parameters.

$$BIC = -2ln(\hat{\mathcal{L}}) + kln(n)$$

where n is the number of data points in the time series.

17.7 Ljung-Box Test

Ljung-Box Test is a type of statistical test of whether any of a group of autocorrelations of a time series are different from zero. Instead of testing randomness at each distinct lag, it tests the "overall" randomness based on a number of lags.

- H_0: The data are independently distributed (i.e. the correlations in the population from which the sample is taken are 0, so that any observed correlations in the data result from randomness of the sampling process).

- H_a: The data are not independently distributed; they exhibit serial correlation.

$$Q = n(n+2) \sum_{k=1}^{h} \frac{\hat{p}_k^2}{n-h}$$

where n is the sample size, $\hat{p}_k$ is the sample autocorrelation at lag k, and h is the number of lags being tested. Under H_0 the statistic Q asymptotically follows a $\chi^2_{(h)}$.

To evaluate performance or to decide whether an ARIMA model was a good fit to a time series, Ljung-Box test is more appropriate than AIC and BIC. Note that it is applied to the **residuals** of a fitted ARIMA model, not the original series, and in such applications the hypothesis actually being tested is that the residuals from the ARIMA model have no autocorrelation. When testing the residuals of an estimated ARIMA model, the degrees of freedom need to be adjusted to reflect the parameter estimation. For example, for an $ARIMA(p, 0, q)$ model, the degrees of freedom should be set to $h - p - q$.

17.8 Partial correlation

Given a vector space $\mathbb{R}^n$. Let $\mathcal{Z}$ be a subspace of $\mathbb{R}^n$ and $X, Y \in \mathbb{R}^n$ two vectors. Then the partial correlation between X and Y with repect to $\mathcal{Z}$, denoted $\rho_{XY.\mathcal{Z}}$, is the correlation between the residual r_X and r_Y. These residuals r_X, r_Y are obtained by removing component of X and Y in $\mathcal{Z}$.

Suppose that $\mathcal{Z}$ is of dimension $k \leq n$, with a base matrix $Z \in \mathbb{R}^{n \times k}$:

$$Z = [z_1, z_2, ..., z_k], z_i \in \mathbb{R}^n$$

Then the residuals r_X, r_Y are defined by :

- $r_X = X - ZW_X$, where $W_X = \underset{W \in \mathbb{R}^k}{\text{argmin}} \ \|X - ZW\|_2^2$

- $r_Y = Y - ZW_X$, where $W_Y = \underset{W \in \mathbb{R}^k}{\text{argmin}} \ \|Y - ZW\|_2^2$

Finally, the partial correlation between X and Y is :

$$\rho_{XY.\mathcal{Z}} = corr(r_X, r_y)$$

An geometrical interpretation for partial correlation is showed by figure 4.5.

17.9 ARMA

This is a model that combines AR and MA models:

$$X_t = \varepsilon_t + \sum_{i=1}^{p} \phi_i X_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}$$

Finding appropriate values of p and q in the $ARMA(p, q)$ model can be facilitated by plotting the *partial autocorrelation* functions for an estimate of p, and likewise using the *autocorrelation* functions for an estimate of q.

It is generally considered good practice to find the smallest values of p and q which provide an acceptable fit to the data.
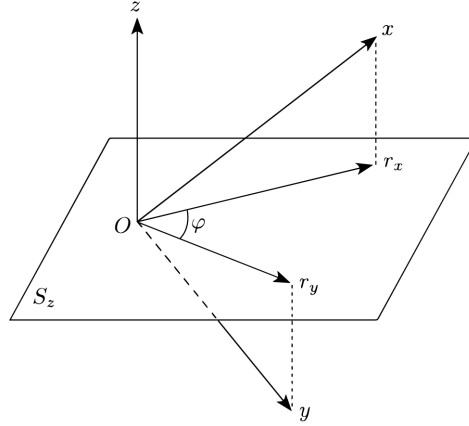
Figure 4.5: In space $\mathbb{R}^3$. $\mathcal{Z} = span(\{(0,0,1)\})$ or the space is generated by only the axe Oz. Two signals x et y are first projected into Oz, then we calculate the correlation between their residuals (after removing component projected in Oz).

17.9.1 Using ACF and PACF to estimate p and q

ACF means Autocorrelation Function and PACF means Partial Autocorrelation Function.

We plot two corrgrams (correlogram), one for ACF and one for PACF :

- $ACF(X_t, X_{t-h})$.

- $PACF(X_t, X_{t-h})$, with $\mathcal{Z} = [X_{t-1}, ..., X_{t-h+1}]$.

There are in general two types of corrgram :

- Sharp Drop-off after k time units.

- Gradual Decline.

Identification of an AR model is often best done with the PACF. For an AR model, the theoretical PACF "shuts off" shows the order of the model. The phrase "shuts off" means that in theory the partial autocorrelations are equal to 0 beyond that point. Put another way, the number of non-zero partial autocorrelations gives the order of the AR model.

Identification of an MA model is often best done with the ACF rather than the PACF. For an MA model, the theoretical PACF does not shut off, but instead tapers toward 0 in some manner. A clearer pattern for an MA model is in the ACF. The number of non-zero autocorrelations gives the order of the MA model.

276

17.9.2 Fitting

ARMA models in general can be, after choosing p and q, fitted by least squares regression to find the values of the parameters which minimize the error term.

For a pure AR model, the Yule-Walker equations (sec 17.3.1) may be used to provide a fit.

For $ARMA(1,2)$, with only X are observation.

$$
\begin{bmatrix} X_2 \\ X_3 \\ \vdots \\ X_{t+1} \end{bmatrix} \approx \phi_1 \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_t \end{bmatrix} + \begin{bmatrix} \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_{t+1} \end{bmatrix} + \theta_1 \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_t \end{bmatrix} + \theta_2 \begin{bmatrix} \varepsilon_0 \\ \varepsilon_1 \\ \vdots \\ \varepsilon_{t-1} \end{bmatrix}
$$

$$
= \begin{bmatrix} X_1 & \varepsilon_2 & \varepsilon_1 & \varepsilon_0 \\ X_2 & \varepsilon_3 & \varepsilon_2 & \varepsilon_1 \\ \vdots & \vdots & \vdots & \vdots \\ X_t & \varepsilon_{t+1} & \varepsilon_t & \varepsilon_{t-1} \end{bmatrix} \times \begin{bmatrix} \phi_1 \\ 1 \\ \theta_1 \\ \theta_2 \end{bmatrix}
$$

- As in MA, we can use iterative procedure, fitting for $\varepsilon, \phi, \theta$ (now we have additionial variable ϕ). Iterative approaches are also used in practice. Moreover, Generalized method of moments (GMM, sec **??**) can be used.

- Second method (by using MLE as in AR but now with back shift operator): https://stats.stackexchange.com/questions/314536/fitting-maq-and-arimaq-model

Choosing the best $ARMA(p,q)$ model *after fitting* : In order to determine which order (p,q) of the ARMA model is appropriate for a series, we need to use the AIC (or BIC) across a subset of values for (p,q), and then apply the Ljung-Box test to determine if a good fit has been achieved, for particular values of (p,q).

ARMA model, generally known,is to describe a *regular wide-sense stationary* or *weakly stationary* time series. A numerical application of ARMA https://www.quantstart.com/articles/Autoregressive-Moving-Average-ARMA-p-q-Models-for-Time-Series-Analysis-P

17.10   ARIMA

- The AR part of ARIMA indicates AR model (sec 17.3). In short, the variable of interest is regressed on its own lagged (i.e., prior) values.

- MA here is a model (sec 17.4). Not to be confused this model with Moving average filter.

- The I (for "integrated") indicates that the data values have been replaced with the difference between their values and the previous values (and this differencing process may have been performed more than once). This is for objective to remove unit root.

Autoregressive integrated moving average is extended version of autoregressive moving average (ARMA) model :

- ARIMA models are applied in some cases where data show evidence of non-stationarity in the sense of mean (but not variance/autocovariance), where an initial differencing step (corresponding to the "integrated" part of the model) can be applied one or more times to eliminate the non-stationarity of the mean function (i.e., the trend).

- When the seasonality shows in a time series, the seasonal-differencing could be applied to eliminate the seasonal component.

Non-seasonal ARIMA models are generally denoted $ARIMA(p, d, q)$ where parameters p, d, and q are non-negative integers :

- p is the order (number of time lags) of the autoregressive model (AR).

- d is the degree of differencing (the number of times the data have had past values subtracted).

- q is the order of the moving-average model (MA)

Seasonal ARIMA models are usually denoted $ARIMA(p, d, q)(P, D, Q)_m$, where m refers to the number of periods in each season, and the uppercase P, D, Q refer to the autoregressive, differencing, and moving average terms for the seasonal part of the ARIMA model.

ARIMA models can be estimated following the Box–Jenkins approach.

Chapter 5

Stochastic calculus

1   Derivatives and Integrals of RP

We have seen concepts such as continuity, differentiability, and integrability in calculus for deterministic signals (deterministic functions). Here, we need to extend those concepts to random processes.

Let $X(t)$ be a continuous-time random process. We say that $X(t)$ is **mean-square continuous** at time t if

$$\lim_{\delta \to 0} \mathbb{E}\left[|X(t+\delta) - X(t)|^2\right] = 0.$$

It is important to note that the mean-square continuity does not imply that every possible realization of $X(t)$ is a continuous function. It roughly means that $\mathbb{E}[X(t+\delta) - X(t)]$ is small.

1.0.1   Derivative and integral

Let $X(t) = A + Bt + Ct^2, \forall t \in [0, \infty)$ where A, B, C are random variables (which does not depends on t), then $X'(t) = B + C2t$ and $\int_t X(t) = At + B\frac{t^2}{2} + C\frac{t^3}{3} + C_0$

A other informal definition of derivative :

$$\lim_{t \to t_0} \frac{X(t) - X(t_0)}{t - t_0} = A,$$

where A is a random variable.

1.0.2   Interchange

A key point to note is that differentiation and integration are linear operations. This, for example, means that you can often interchange integration and

expectation. More specifically, we can write :

$$\mathbb{E}\left[\int_0^t X(u)du\right] = \int_0^t \mathbb{E}[X(u)]du$$

$$\mathbb{E}\left[\frac{d}{dt}X(t)\right] = \frac{d}{dt}\mathbb{E}[X(t)]$$

2 Stochastic Integrals

Given the stochastic differential equation (SDE) :

$$dY_t = X_t dB_t$$

where X_s is a stochastic process (SP) and B_s is Wiener process. Then the solution for this SDE is a stochastic integral (SI) :

$$Y_t = Y_0 + \int_0^t X_s dB_s,$$

We discover in this section:

- When the SI is defined or integrable.

- Itô lemma

2.1 Integrability of SI

Let $\mathcal{L}^2 = \mathcal{L}^2(S,T)$ (S,T are two bounds) be the class that contains SP g such that :

- $g(t,\omega) : [0,\infty] \times \Omega \to \mathbb{R}$

- $g(t,\omega)$ is $\mathcal{F}_t^B$-adapted, where B is Wiener process (see subsubsection 1.6)

- $\mathbb{E}\left[\int_S^T g(t,\omega)^2 dt\right] < \infty$

If a SP $h \in \mathcal{L}(S,T)^2$, then h is integrable on $[S,T]$, or

$$\int_S^T g(t,\omega)dB_t(\omega) = \lim_{n\to\infty}\int_S^T \phi_n(t,\omega)dB_t(\omega)$$

is defined or integrable. $\phi_n(t,\omega)$ is call the elementary function of $g(t,\omega)$ and is definded by :

$$\phi_n(t,\omega) = \sum_{i=0}^n g(t_i,\omega)\mathbb{1}_{[t_i,t_{i+1})}(t)$$

2.1.1 Notes

The second condition, $g(t, \omega)$ is $\mathcal{F}_t^B$-adapted, $\mathcal{F}_t^g \subseteq \mathcal{F}_t^B$ is for the objective to benefit the properties of Weiner process B:

- $g(., \omega)$ and $B(., \omega)$ have the same sample space. Furthermore, note that, when evaluating, ω is the same for $g(., \omega)$ and $B(., \omega)$.

- $\mathcal{F}_t^B$ represents only the information available up to time t based on the observed behavior of the Wiener process B up to that point. g is adapted to $\mathcal{F}_t^B$ guarantees that g only uses information from the past and present, not the future.

- $g(t_i, \omega)$ and $B(t_i) - B(t_{i+1})$ are independent, since $\mathcal{F}_{t_i}^g \subset \mathcal{F}_{t_i}^B$ and $\mathcal{F}_{t_i}^B$, $B(t_i) - B(t_{i+1})$ are independent (see subsec 1.10).

2.1.2 Properties of integrable SI

If $g \in \mathcal{L}^2(S, T)$:

- $\mathbb{E}\left[\int_S^T g_t dW_t\right] = 0$

- $\mathbb{E}\left[\left(\int_S^T g_t dW_t\right)^2\right] = \int_S^T E[g_t^2] dW_t$

- $\int_S^T g_t dW_t$ is $\mathcal{F}_T^B$-measurable.

2.2 Itô lemma

2.2.1 Motivation

Suppose X_t is an Itô drift-diffusion process that satisfies the stochastic differential equation :

$$dX_t = \mu(t)dt + \sigma(t)dB_t$$

where B_t is a Wiener process (stochastic), $\mu(t)$ and $\sigma(t)$ are function *deterministic* of t (not stochastic), independent with B_t. In general, it's not possible to write a solution X_t directly in terms of B_t. However, we can have a stochastic solution. First, formally write an integral solution :

$$X_t = X_0 + \int_0^t \mu(s)ds + \int_0^t \sigma(s)dB_s$$

It is proved that X_t has the same moments as the normal distribution $\mathcal{N}\left(\int_0^t \mu(s)ds, \int_0^t \sigma(s)^2 ds\right)$:

$$\mathbb{E}(X_t) = \mathbb{E}\left[\int_0^t \mu(s)ds + \int_0^t \sigma(s)dB_s\right]$$

$$= \int_0^t \mathbb{E}[\mu(s)ds] + \int_0^t \mathbb{E}[\sigma(s)dB_s]$$

$$= \int_0^t \mu(s)ds + \int_0^t \mathbb{E}[\sigma(s)]\mathbb{E}[dB_s]$$

$$= \int_0^t \mu(s)ds + 0$$

$$Var(X_t) = Var\left[\int_0^t \mu(s)ds + \int_0^t \sigma(s)dB_s\right]$$

$$= Var\left[\int_0^t \sigma(s)dB_s\right], \quad \int_0^t \mu(s)ds \text{ is deterministic}$$

$$= \mathbb{E}\left[\left(\int_0^t \sigma(s)dB_s\right)^2\right] - \mathbb{E}\left[\int_0^t \sigma(s)dB_s\right]^2$$

$$= \mathbb{E}\left[\left(\sum_i e_i\Delta B_i\right)^2\right] - \mathbb{E}\left[\sum_i e_i\Delta B_i\right]^2$$

$$= \mathbb{E}\left[\sum_i\sum_j e_ie_j\Delta B_i\Delta B_j\right] - \left(\sum_i e_i\mathbb{E}[\Delta B_i]\right)^2$$

$$= \sum_i\sum_j e_ie_j\mathbb{E}[\Delta B_i\Delta B_j]$$

$$= \left(\sum_{i=j} + 2\sum_{i\neq j}\right)e_ie_j\mathbb{E}[\Delta B_i\Delta B_j]$$

$$= \sum_{i=j} e_i^2\mathbb{E}[\Delta B_i^2]$$

$$= \int_0^t \sigma(s)^2 ds$$

$$\mathbb{E}\left[\left(\frac{X_t - E(X_t)}{\sqrt{Var(X_t)}}\right)^3\right] = \frac{1}{\sqrt{Var(X_t)}^3}\mathbb{E}\left[\left(\int_0^t \sigma(s)dB_s\right)^3\right]$$

$$= \frac{1}{\sqrt{Var(X_t)}^3}\mathbb{E}\left[\left(\sum_i e_i\Delta B_i\right)^3\right]$$

$$= \frac{1}{\sqrt{Var(X_t)}^3}\mathbb{E}\left[\sum_i\sum_j\sum_k e_ie_je_k\Delta B_i\Delta B_j\Delta B_k\right]$$

$$= \frac{1}{\sqrt{Var(X_t)}^3}\sum_i\sum_j\sum_k e_ie_je_k\mathbb{E}\left[\Delta B_i\Delta B_j\Delta B_k\right]$$

$$= \frac{1}{\sqrt{Var(X_t)}^3}\left(\sum_{i=j=k} + 3\sum_{i=j\neq k} + 6\sum_{i\neq j\neq k}\right) e_ie_je_k\mathbb{E}\left[\Delta B_i\Delta B_j\Delta B_k\right]$$

$$= 0$$

$$\mathbb{E}\left[\left(\frac{X_t - E(X_t)}{\sqrt{Var(X_t)}}\right)^4\right] = \frac{1}{\sqrt{Var(X_t)}^4}\mathbb{E}\left[\left(\int_0^t \sigma(s)dB_s\right)^4\right]$$

$$= \frac{1}{\sqrt{Var(X_t)}^4}\mathbb{E}\left[\left(\sum_i e_i\Delta B_i\right)^4\right]$$

$$= \frac{1}{\sqrt{Var(X_t)}^4}\mathbb{E}\left[\sum_i\sum_j\sum_k\sum_l e_ie_je_ke_l\Delta B_i\Delta B_j\Delta B_k\Delta B_l\right]$$

$$= \frac{1}{\sqrt{Var(X_t)}^4}\sum_i\sum_j\sum_k\sum_l e_ie_je_ke_l\mathbb{E}\left[\Delta B_i\Delta B_j\Delta B_k\Delta B_l\right]$$

$$= \frac{1}{\sqrt{Var(X_t)}^4}\left(\sum_{i=j=k=l}+4\sum_{i=j=k\neq l}+6\sum_{i=j\neq k=l}+12\sum_{i=j\neq k\neq l}+24\sum_{i\neq j\neq k\neq l}\right)\times$$
$$e_ie_je_ke_l\mathbb{E}\left[\Delta B_i\Delta B_j\Delta B_k\Delta B_l\right]$$

$$= \frac{1}{\sqrt{Var(X_t)}^4}\left(\sum_{i=j=k=l}+6\sum_{i=j\neq k=l}\right)e_ie_je_ke_l\mathbb{E}\left[\Delta B_i\Delta B_j\Delta B_k\Delta B_l\right]$$

$$= \frac{1}{\sqrt{Var(X_t)}^4}\left(\sum_i e_i^4\mathbb{E}[\Delta B_i^4] + 6\sum_i\sum_{j\neq i}e_i^2e_j^2\mathbb{E}[\Delta B_i^2]\mathbb{E}[\Delta B_j^2]\right)$$

$$= \frac{1}{\sqrt{Var(X_t)}^4}\left(\sum_i e_i^4 3\mathbb{E}[\Delta B_i^2]^2 + 6\sum_i\sum_{j\neq i}e_i^2e_j^2\mathbb{E}[\Delta B_i^2]\mathbb{E}[\Delta B_j^2]\right)$$

$$= \frac{1}{\sqrt{Var(X_t)}^4}3\int_0^t\sigma(s)^2ds\int_0^t\sigma(s)^2ds$$

$$= \frac{1}{\sqrt{Var(X_t)}^4}3Var(X_t)Var(X_t)$$

$$= 3$$

where $e_i = \sigma(s_i)$, $\Delta B_i = B_{s_i} - B_{s_{i+1}}$. $E[(\Delta B_i)^{2n+1}] = 0$. $\Delta B_i, \Delta B_j, (i \neq j)$ are independent. Suppose that X_t follows the normal distribution, then we can sample :

$$X_t \sim \mathcal{N}\left(\int_0^t \mu(s)ds, \int_0^t \sigma(s)^2ds\right)$$

or

$$X_t = \int_0^t \mu(s)ds + \epsilon \leftarrow \mathcal{N}(0, \int_0^t \sigma(s)^2ds)$$

However, sometimes we are faced with a stochastic differential equation for a more complex process Y_t, in which the process appears on the deterministic functions and this becomes stochastic. That is, say:

$$dY_t = a_1(Y_t, t)dt + a_2(Y_t, t)dB_t$$

for some functions a_1 and a_2. In this case, we cannot immediately write a formal solution as we did for the simpler case above. Instead, we hope to write the process Y_t as a function of a simpler process X_t taking the form above. That is, we want to identify three functions $f(t, x), \mu_t$ and σ_t such that $Y_t = f(t, X_t)$ and $dX_t = \mu_t dt + \sigma_t dB_t$. In practice, Ito's lemma is used in order to find this transformation.

2.2.2 Informal derivation

Supposons we have already X_t satisfies:

$$dX_t = \mu(t)dt + \sigma(t)dB_t$$

where $\mu(t)$ and $\sigma(t)$ are deterministic.

If $f(t, X_t)$ is a twice-differentiable scalar function, its expansion in a Taylor series is :

$$df = \frac{\partial f}{\partial t}dt + \frac{\partial f}{\partial X_t}dX_t + \frac{1}{2}\frac{\partial^2 f}{\partial X_t^2}dX_t^2 + ...$$

Substituting $dX_t = \mu_t dt + \sigma_t dB_t$:

$$df = \frac{\partial f}{\partial t}dt + \frac{\partial f}{\partial X_t}(\mu_t dt + \sigma_t dB_t) + \frac{1}{2}\frac{\partial^2 f}{\partial X_t^2}(\mu_t dt + \sigma_t dB_t)^2 + ...$$

Because in the limit $dt \to 0$, the terms dt^2 and $dtdB_t$ tend to zero faster than dB_t^2, which is thus dt. Setting the dt^2 and $dtdB_t$ terms to zero, substituting dt for dB_t^2 (see subsection 8.1 of Brownian motion), and collecting the dt and dB_t terms, we obtain :

$$df = \left(\frac{\partial f}{\partial t} + \frac{\partial f}{\partial X_t}\mu_t + \frac{1}{2}\frac{\partial^2 f}{\partial X_t^2}\sigma_t^2\right)dt + \left(\frac{\partial f}{\partial X_t}\sigma_t\right)dB_t$$

By identifing :

$$\begin{cases} a_1(f(t, X_t), t) = \frac{\partial f}{\partial t} + \frac{\partial f}{\partial X_t}\mu_t + \frac{1}{2}\frac{\partial^2 f}{\partial X_t^2}\sigma_t^2 \\ a_2(f(t, X_t), t) = \frac{\partial f}{\partial X_t}\sigma_t \end{cases}$$

Then we need to sovle the above system of DE for getting f hence Y_t.

2.2.3 Important result

If we the function f is only depends on X_t, where :

$$dX_t = \mu(t)dt + \sigma(t)dB_t$$

$\mu(t)$ and $\sigma(t)$ are deterministic. This means $f = f(X_t)$ and from section 2.2.2 :

$$
\begin{aligned}
df(X_t) &= \left(\frac{\partial f}{\partial X_t}\mu_t + \frac{1}{2}\frac{\partial^2 f}{\partial X_t^2}\sigma_t^2 \right)dt + \left(\frac{\partial f}{\partial X_t}\sigma_t \right)dB_t \\
&= \left(\frac{\partial f}{\partial X_t}\mu_t dt + \frac{\partial f}{\partial X_t}\sigma_t dB_t \right) + \frac{1}{2}\frac{\partial^2 f}{\partial X_t^2}\sigma_t^2 dt \\
&= f'(X_t)dX_t + \frac{1}{2}\frac{\partial^2 f}{\partial X_t^2}\sigma_t^2 dt
\end{aligned}
$$

2.2.4 Example

Solve the following Linear SDE :

$$
\begin{aligned}
dY_t &= a_1(Y_t, t)dt + a_2(Y_t, t)dB_t \\
&= \left(c(t)Y_t + d(t) \right)dt + \left(p(t)Y_t + q(t) \right)dB_t
\end{aligned}
$$

From the above, we get the system of DE :

$$
\begin{cases}
c(t)f(t, X_t) + d(t) = \frac{\partial f}{\partial t} + \frac{\partial f}{\partial X_t}\mu_t + \frac{1}{2}\frac{\partial^2 f}{\partial X_t^2}\sigma_t^2 \\
p(t)f(t, X_t) + q(t) = \frac{\partial f}{\partial X_t}\sigma_t
\end{cases}
$$

We discover next the Geometric Brownian Motion, which is a special case of Linear SDE.

2.3 Notes

Not every $a_1(Y_t, t)$ and $a_2(Y_t, t)$ can make the stochastic differential equation

$$dY_t = a_1(Y_t, t)dt + a_2(Y_t, t)dB_t$$

has a solution. To check that (if there are solutions), we can see if :

- (By Itô lemma)

$$
\begin{cases}
a_1(f(t, X_t), t) = \frac{\partial f}{\partial t} + \frac{\partial f}{\partial X_t}\mu_t + \frac{1}{2}\frac{\partial^2 f}{\partial X_t^2}\sigma_t^2 \\
a_2(f(t, X_t), t) = \frac{\partial f}{\partial X_t}\sigma_t
\end{cases}
$$

  has solution.

- $a_2(Y_t, t)$ satifies conditions in subsection 2.1.

2.4 Geometric Brownian Motion

A stochastic process Y_t is said to follow a GBM if it satisfies the following stochastic differential equation (SDE):

$$dY_t = aY_t dt + bY_t dB_t$$

Let $Y_t = f(t, x)$, apply Itô's lemma by identifing coefs of dt and dB_t, the objective then to find $f(t, x), \mu_t, \sigma_t$ satisfy:

$$\begin{cases} \frac{\partial f}{\partial x} \sigma_t = bf(t, x) \\ \frac{\partial f}{\partial t} + \frac{\partial f}{\partial x} \mu_t + \frac{1}{2} \frac{\partial^2 f}{\partial x^2} \sigma_t^2 = af(t, x) \end{cases}$$

From the first sde (stochastic differential equation), we have :

$$f(t, x) = C(t) \exp\left(\frac{b}{\sigma_t} x\right)$$

Then replace this into the second sde :

$$C'(t) \exp\left(\frac{b}{\sigma_t} x\right) - C(t) \exp\left(\frac{b}{\sigma_t} x\right) \frac{bx}{\sigma_t^2} \sigma_t' + C(t) \exp\left(\frac{b}{\sigma_t} x\right) \frac{b}{\sigma_t} \mu_t + \frac{1}{2} C(t) \exp\left(\frac{b}{\sigma_t} x\right) \left(\frac{b}{\sigma_t}\right)^2 \sigma_t^2$$

$$= aC(t) \exp\left(\frac{b}{\sigma_t} x\right)$$

or :

$$C'(t) - C(t) \frac{bx}{\sigma_t^2} \sigma_t' + C(t) \frac{b}{\sigma_t} \mu_t + \frac{1}{2} C(t) b^2 = aC(t)$$

where $C(t), \mu_t, \sigma_t$ need to be identified. First, we see that there are only x in the left side, this makes σ_t be a constant A to get that $\sigma_t' = 0$.

$$C'(t) = \left(a - \frac{b}{A} \mu_t - \frac{1}{2} b^2\right) C(t)$$

Here we can simply take $\mu_t = a - \frac{1}{2} b^2$ and $\sigma_t = A = b, C(t) = B$. Then

$$\begin{aligned} Y_t &= Bf(t, X_t) \\ &= B \exp(X_t) \\ &= B \exp\left(\int_0^t (\mu_t) dt + \epsilon \leftarrow \mathcal{N}(0, \int_0^t \sigma_t^2 ds)\right) \\ &= B \exp\left(\int_0^t (a - \frac{1}{2} b^2) dt + \epsilon \leftarrow \mathcal{N}(0, \int_0^t b^2 ds)\right) \\ &= B \exp\left((a - \frac{1}{2} b^2) t + \epsilon \leftarrow \mathcal{N}(0, b^2 t)\right) \\ &= B \exp\left((a - \frac{1}{2} b^2) t + bB_t\right) \end{aligned}$$

But if we want go a little further (for more solution of $C(t), \mu_t, \sigma_t$) :

$$C(t) = B \exp\left(D(t)\right) = B \exp\left(\int_0^t (a - \frac{b}{A}\mu_s - \frac{1}{2}b^2)ds\right)$$

then

$$Y_t = f(t, X_t)$$
$$= B \exp\left(D(t)\right) \exp\left(\frac{b}{A}X_t\right)$$
$$= B \exp\left(D(t) + \frac{b}{A}X_t\right)$$
$$= B \exp\left(\int_0^t (a - \frac{b}{A}\mu_s - \frac{1}{2}b^2)ds + \frac{b}{A}\left(\int_0^t (\mu_s)dt + \mathcal{N}(0, \int_0^t A^2 ds)\right)\right)$$
$$= B \exp\left(\int_0^t (a - \frac{1}{2}b^2)dt + \frac{b}{A}\mathcal{N}(0, A^2 t)\right)$$
$$= B \exp\left(\int_0^t (a - \frac{1}{2}b^2)dt + \mathcal{N}(0, b^2 t)\right)$$
$$= B \exp\left((a - \frac{1}{2}b^2)t + bB_t)\right)$$

We found the same solution as before.

Finally, the SP Y_t is well determined (defined) if we have the initial condition, which means $Y_0 = \beta$. Then in this case, we have $B = \beta$.

3   Stochastic Differentiation

It means an operator for stochastic differentiation, also called infinitesimal generator.

- The input of this operator is a Feller process, which means a particular kind of Markov process that satisfies certain regularity conditions.

- This operator is a Fourier multiplier operator, which is thus the multiplication between the Fourier transform $\mathcal{F}$ of a function g and a specified function known as the multiplier $h : \mathcal{F}_g(f) \times h(f)$. This operator encodes a great deal of information about the process.

Chapter 6

Maths quantitative

1 Sharpe ratio

$$\frac{R_{gain} - R_{loss}}{\sigma_{gain}}$$

Note that in sharpe ratio we use rate of return (R) and covariance and correlation are applied on this factor. On the other hand, in pairs trading, corr and cov may applied on price.

Portfolio $(\sum \gamma_i = 1)$:

$$Y = \gamma_1 X_1 + \gamma_2 X_2 + ... + \gamma_n X_n$$

Diversifiable risk:

$$\sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} 2\gamma_i \gamma_j Cov(X_i, X_j)$$

Undiversifiable risk:

$$\gamma_1^2 Var(X_1) + \gamma_2^2 Var(X_2) + ... + \gamma_n^2 Var(X_n)$$

2 Co-integration

TODO

3 Security return

Two types of rate of return :

- Simple rate of return (arithmetic return) : When dealing with multiple assets over the same timeframe.

- Logarithmic rate of return (log return) : When you make calculations about a single asset over multip time periodes. log return is better in approximately normal distribution and when working with distribution in general.

4 Markowitz's theory or Modern portfolio theory

4.1 Mathematical model

- R_i is the return on asset i.

- w_i is the weighting of component asset i, $\sum_i w_i = 1$.

- σ_i is the (sample) standard deviation of the periodic returns on an asset i.

- σ_{ij} is the covariance of the periodic returns on the two assets i and j. Alternatively, $\sigma_{ij} = \sigma_i \sigma_j \rho_{ij}$, where ρ_{ij} is the correlation.

Then the expected return and std of the portfolio :

- $E(R_p) = \sum_i w_i E(i)$

- $\sigma_p = \sum_i w_i^2 \sigma_i^2 + \sum_i \sum_{j \neq i} w_i w_j \sigma_{ij}$

For a given "risk tolerance" $q \in [0, \infty)$ as parameter, the optimal portfolio is found by minimizing the following expression:

$$\min_w w^T \Sigma w - q \times E[R]^T w$$

where Σ is the covariance matrix for the returns on the assets in the portfolio.

4.2 Optimal portfolio

Set-up

- N is the number of risky assets

- $E[r_i] = \mu_i$ is the expect (mean) return rate of asset i

- σ_{ij} is the covariance of asset i and j.

- W is the amount of wealth that is to be allocated in the portfolio.

- Z is the expected (required) amount of wealth.

- $X = [x_1, ... x_N]$ is the amount invested in each asset. Hence $x_1 + ... x_N = W$

Let C be the amount after the investement. Then $C = x_1(1 + r_1) + \ldots + x_N(1 + r_N) = W + x_1 r_1 + \ldots + x_N r_N$ and its mean of C is $E[C] = W + x_1\mu_1 + \ldots + x_N\mu_N$.

Minimize

$$Var(C) = E[C - E[C]]^2$$
$$= E[(r_1 - \mu_1)x_1 + \ldots + (r_N - \mu_N)x_N]^2$$
$$= \sum_i x_i^2 \sigma_{ii} + 2\sum_{i>j} x_i x_j \sigma_{ij}$$

subject to:

$$\begin{cases} W + x_1\mu_1 + \ldots + x_N\mu_N = Z \\ x_1 + \ldots + x_N = W \end{cases}$$

The Lagrangian for this constrained optimization problem ;

$$\sum_i x_i^2 \sigma_{ii} + 2\sum_{i>j} x_i x_j \sigma_{ij} + \lambda(Z - W - x_1\mu_1 - \ldots - x_N\mu_N) + \eta(W - x_1 - \ldots - x_N)$$

Then we need to solve the first-order conditions, for obtaining the optimal X.

$$\begin{cases} 2x_1\sigma_{11} + 2x_2\sigma_{12} + \ldots + 2x_N\sigma_{1N} = \lambda\mu_1 + \eta \\ \vdots \\ 2x_1\sigma_{N1} + 2x_2\sigma_{N2} + \ldots + 2x_N\sigma_{NN} = \lambda\mu_N + \eta \\ Z - W = x_1\mu_1 + \ldots + x_N\mu_N \\ W = x_1 + \ldots + x_N \end{cases}$$

In the sample case that $N = 2$:

$$\begin{cases} 2x_1\sigma_{11} + 2x_2\sigma_{12} = \lambda\mu_1 + \eta \\ 2x_1\sigma_{21} + 2x_2\sigma_{22} = \lambda\mu_2 + \eta \\ Z - W = x_1\mu_1 + x_2\mu_2 \\ W = x_1 + x_2 \end{cases}$$

$$\begin{cases} x_1 = \frac{(\sigma_{22}\mu_1 - \sigma_{12}\mu_2)\lambda + \eta(\sigma_{22} - \sigma_{12})}{2(\sigma_{11}\sigma_{22} - \sigma_{12}^2)} \\ x_2 = \frac{(\sigma_{11}\mu_2 - \sigma_{12}\mu_1)\lambda + \eta(\sigma_{11} - \sigma_{12})}{2(\sigma_{11}\sigma_{22} - \sigma_{12}^2)} \\ Z - W = x_1\mu_1 + x_2\mu_2 \\ W = x_1 + x_2 \end{cases}$$

4.3 Mutual fund separation theorem

Hypothesis : The market is the optimal portfolio.

Investor will choose different combinations of the safe asset and the mutual fund; but they will not choose different mixes of risky assets (since cost transaction).
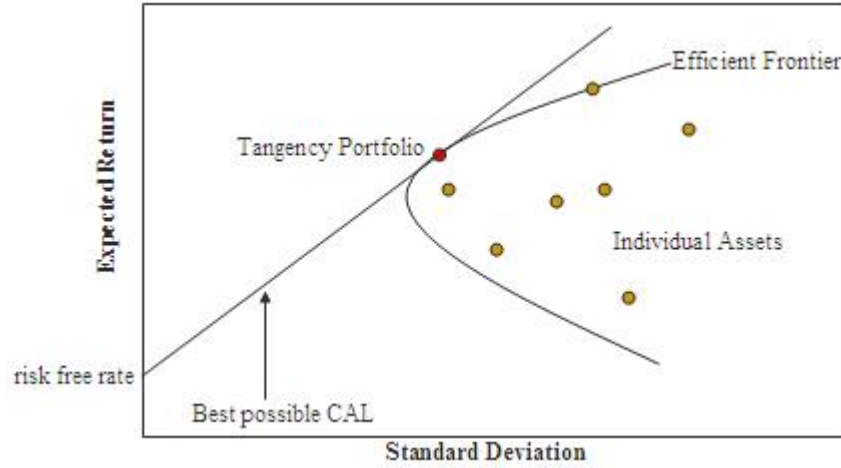
Figure 6.1: This figure shows expected return on the vertical axis, and the standard deviation on the horizontal axis (volatility or risk). In this risk (x) and expected return (y) space, we can plot every possible combination of risky assets. The horizontal hyperbolic means that, given the risk tolerance q, the relation between the risk and expected return (optimal) is $\sigma_p = aE(R_p)^2 + bE(R_p) + c$. The upper part of the hyperbolic boundary is the **efficient frontier**. We can say, for a given risk that we can support, by seeing the efficient frontier, we know the optimal expected return. The tangent to the upper part of the hyperbolic boundary is the capital allocation line (CAL).

Mutual fund separation theorem states that any investor's optimal portfolio can be constructed by holding each of certain mutual funds in appropriate ratios. In this case, the number of mutual funds is smaller than the number of individual assets in the portfolio. Here a mutual fund refers to any specified benchmark portfolio of the available assets.

4.4 Efficient frontier

See figure 6.1

4.5 Capital allocation line

If the Efficient frontier help us to observe the relation between (optimal) expected return and risk about a **combination of risky assets**, then the capital allocation line help us to observe the same thing, but about a **combination of a risky asset and a risk-free asset**.

Suppose that the relation between risk premium and the volatility (risk) is a linear relation, with slope b_P, characterize for risky asset P :

$$E(r_P) - r_F = b_P \sigma_P$$

If a portfolio C composed of the risky asset P and the risk-free asset F, which means:

$$E(r_C) = w_1 r_F + w_2 E(r_P)$$

and

$$\sigma_C = w_2 \sigma_P$$

where $w_1 + w_2 = 1$. We have :

$$\begin{aligned}
E(r_C) &= w_1 r_F + w_2 E(r_P) \\
&= w_1 r_F + w_2 (r_F + b_P \sigma_P) \\
&= r_F + w_2 b_P \sigma_P \\
&= r_F + + b_P \sigma_C
\end{aligned}$$

From

$$\begin{cases} E(r_P) - r_F = b_P \sigma_P \\ E(r_C) - r_F = b_P \sigma_C \end{cases}$$

we infer that

$$E(r_C) - r_F = \frac{E(r_P) - r_F}{\sigma_P} \sigma_C$$

The slope $\frac{E(r_P) - r_F}{\sigma_P}$ is known as the "reward-to-variability ratio". In the figure 6.1, we can observe it (half-line) between the risk free rate point $((0, r_F)$ and the optimal expected return and std point, which is thus (σ_P, r_P). This line can be considered as a new efficient frontier while taking into account free-rate asset.

Note that, in the prolongation line, we have $w_1 < 0$ and $w_2 > 1$, Then $w_1 < 0$ means that we borrow money with a risk-free rate and $w_2 > 2$ means that we invest more than what we have to asset P.

4.6 Capital market line

Given p is a combination of the market portfolio and the risk-free asset, as the CAL, in the Capital market line (CML) we have the return rate R_p is a function of σ_p :

$$\text{CML} : \sigma_p \mapsto R_f + \sigma_p \frac{E(R_M) - R_f}{\sigma_M}$$

All points along the CML have superior risk-return profiles to any portfolio on the efficient frontier (see figure 6.2). It is because that the tangency point M represents the market portfolio, so all rational investors (minimum variance criterion) should hold their risky assets in the same proportions as their weights in the market portfolio.

Points to the left of and above the CML are infeasible, whereas points to the right/below are attainable but inefficient.
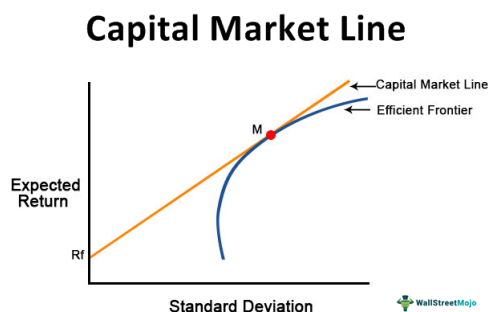
293

Figure 6.2: Capital Market Line

A stock picking rule of thumb is to buy assets whose Sharpe ratio will be above the CML ($\frac{E(R_M)-R_f}{\sigma_M}$) and sell those whose Sharpe ratio will be below. Indeed, from the efficient market hypothesis it follows that it's impossible to beat the market. Therefore, all portfolios should have a Sharpe ratio less than or equal to the market's.

In consequence, if there is a portfolio (or asset) whose Sharpe ratio will be bigger than the market's then it means an abnormal extra return over the market's return at a given level of risk. This is called the alpha 13.

5   Exploratory data analysis (EDA)

TODO Find correlation between signals.

- Spearman

- Pearson

- Kendall

- Phik

6   Greeks (finance)

In finance, "Greeks" refer to a set of mathematical measures used to quantify the risk and sensitivity of options and other derivative securities to various factors. These factors include changes in the price of the underlying asset, time decay, changes in volatility, and changes in interest rates.

Sensitivity here means the variation of $Y = f(X)$ in term of X. In short, we want to know if X is changed a quantity Δ_X, what is Δ_Y ?

6.1 Delta

Δ measures the rate of change of the theoretical option value with respect to changes in the underlying asset's price. Delta is the first derivative of the value V of the option with respect to the underlying instrument's price S:

$$\Delta = \frac{\partial V}{\partial S}$$

or we can write :

$$dV = \Delta dS$$

Since the option cost of a share is always much smaller than the uDderlying instrument's price ($V \ll S$) then we have $|\Delta| \leq 1$. Conventionally, we express Δ as the variation of option dV, given that $dS = \$1$.

For a vanilla option, Δ can be positive or negative, to facilitate understanding the profit or loss

- $0 \leq \Delta \leq 1$ for a long call (or a short put). Since in this context, we expect that the gain of profit is in the same direction with the increase of underlying instrument's price.

- $-1 \leq \Delta \leq 0$ for a long put (or a short call). Since in this context, we expect that the gain of profit is in the opposite direction with the increase of underlying instrument's price.

Depending on underlying instrument's price, a call option behaves as:

- owning 1 share of the underlying stock (if deep in the money)

- owns nothing (if far out of the money)

- something in between

6.1.1 Relationship between call and put delta

The difference between Δ of a call and Δ of a put at the same strike is equal to one. This is by the put–call parity in section 9.7, we have

$$\frac{\partial C}{\partial S} - \frac{\partial P}{\partial S} = 1$$

The total delta of a complex portfolio of positions **on the same underlying asset** can be calculated by simply taking the sum of the deltas for each individual position, which means Δ of a portfolio is linear in the constituents.

6.1.2 Share equivalent understanding

The Δ and the number of acquired options are commonly presented as the number of shares. For example, if a portfolio of 100 American call options on XYZ each have a delta of 0.25 (=25%), it will gain or lose value just like 2,500 shares of XYZ as the price changes for *small price movements* (1 option contracts covers 100 shares).

The sign of Δ and percentage notation are often dropped, .e.g, the above example, we say 25 delta call.

6.1.3 Delta hedging or Delta neutral

By buying or selling (adjustment) the underlying asset, we aim to keep the overall delta of our portfolio close to zero.

For example, if the delta of a portfolio of options in XYZ (expressed as shares of the underlying) is +2.75, the trader would be able to delta-hedge the portfolio by selling short 2.75 shares of the underlying. This portfolio will then retain its total value regardless of which direction the price of XYZ moves.

This helps to minimize the impact of price changes in the underlying asset. Albeit (mac du) for only small movements of the underlying, a short amount of time and not-withstanding (bat chap, khong chiu anh huong boi) changes in other market conditions such as volatility and the rate of return for a risk-free investment.

Note that :

- Δ can vary in term of time. Therefore, **the frequency of adjustments** depends on various factors, including the volatility of the underlying asset, time remaining until expiration, and the desired level of risk management.

- Delta hedging is not a perfect strategy, and it cannot completely eliminate all risks. However, it can help traders and investors manage and reduce their exposure to price fluctuations in the underlying asset while holding options positions.

- In practice, maintaining a zero delta is very complex because there are risks associated with re-hedging on large movements in the underlying stock's price, and research indicates portfolios tend to have lower cash flows if re-hedged too frequently

6.1.4 As a proxy for probability

The (absolute value of) Δ is close to, but not identical with, the percent moneyness of an option, i.e., *the probability that the option will expire in-the-money.*

For example, if a call option has a delta of 0.15, the trader might estimate that the option has approximately a 15% chance of expiring in-the-money. Similarly, if a put contract has a delta of -0.25, the trader might expect the option

to have a 25% probability of expiring in-the-money. At-the-money calls and puts have a delta of approximately 0.5 and -0.5.

The actual probability of an option finishing in the money is its dual delta, which is the first derivative of option price with respect to strike

$$\frac{\partial V}{\partial K}$$

6.2 Vega

Vega measures sensitivity to volatility σ. Vega is the derivative of the option value with respect to the volatility of the underlying asset.

$$\mathcal{V} = \frac{\partial V}{\partial \sigma}$$

Vega is not the name of any Greek letter. The glyph used is a non-standard majuscule version of the Greek letter nu ν, written as $\mathcal{V}$. Vega is typically expressed as the amount of money per underlying share that the option's value will gain or lose as volatility rises or falls by 1 percentage point.

- All options (both calls and puts) will gain value with rising volatility.

- Vega can be an important Greek to monitor for an option trader, especially in volatile markets, since the value of some option strategies can be particularly sensitive to changes in volatility.

- The value of an at-the-money option straddle, for example, is extremely dependent on changes to volatility.

6.3 Theta

Theta Θ, measures the sensitivity of the value of the derivative to the passage of time (rate of decay value) :

$$\Theta = -\frac{\partial V}{\partial \tau}$$

Sometime Θ is expressed in value per year. By convention, it is usual to divide the result by the number of days in a year, to arrive at the amount an option's price will move, after one day.

- Theta is almost always negative for long calls and puts (since time value descrease with passage of time), and positive for short (or written) calls and puts. An exception is a deep in-the-money European put.

- The total theta for a portfolio of options can be determined by summing the thetas for each individual position.

6.4 Rho

Rho, ρ, measures sensitivity to the interest rate: it is the derivative of the option value with respect to the risk-free interest rate (for the relevant outstanding term).

$$\rho = \frac{\partial V}{\partial r}$$

Except under extreme circumstances, the value of an option is less sensitive to changes in the risk-free interest rate than to changes in other parameters. For this reason, rho is the least used of the first-order Greeks.

Rho is typically expressed as the amount of money, per share of the underlying, that the value of the option will gain or lose as the risk-free interest rate rises or falls by 1.0% per annum (100 basis points).

6.5 Lambda

Lambda, λ or omega Ω, or elasticity is the percentage change in option value per percentage change in the underlying price :

$$\lambda = \Omega = \frac{\frac{\partial V}{V}}{\frac{\partial S}{S}} = \frac{\partial V}{\partial S} \times \frac{S}{V} = \Delta \times \frac{S}{V}$$

6.6 Epsilon

Epsilon, ϵ (also known as psi, ψ), is the sensibility (first derivative) of option price w.r.t the dividend yield rate of underlying :

$$\epsilon = \psi = \frac{\partial V}{\partial q}$$

7 Second-order Greeks

7.1 Gamma

Gamma (Γ) measures the rate of change in the Δ with respect to changes in the underlying price. Gamma is also the second derivative of the value function with respect to the underlying price.

$$\Gamma = \frac{\partial \Delta}{\partial S} = \frac{\partial^2 V}{\partial S^2}$$

Most long options have positive gamma and most short options have negative gamma. This is because if the price increases, Gamma increases as well.

Gamma is greatest when the price is approximately at-the-money (ATM) and diminishes the further out when the price go either in-the-money (ITM) or out-of-the-money (OTM).

Gamma is important because it corrects for the convexity of value.

When a trader seeks to establish an effective delta-hedge for a portfolio, the trader may also seek to neutralize the portfolio's gamma, as this will ensure that the hedge will be effective over a wider range of underlying price movements.

...

8 Capital asset pricing model

In finance, the capital asset pricing model (CAPM) is a model used to determine a theoretically **appropriate required rate of return** of an asset, to make decisions about adding assets to a well-diversified portfolio.

The model estimates the asset's sensitivity to non-diversifiable risk (also known as systematic risk or market risk). The non-diversifiable risk that affects the overall market or a particular segment of the market. It is inherent to the entire market or a specific industry and cannot be eliminated, regardless of the level of diversification within a portfolio.

This model is often represented by the quantity of sensitivity (β), as well as the expected return of the market and the expected return of a theoretical risk-free asset.

CAPM assumes a particular form of functions, in which only first and second moments matter (the relations between variable can be only expected value or variance), and zero transaction costs. Under these conditions, CAPM shows that the cost of equity capital (von chu so huu) is determined only by beta.

Despite its failing numerous empirical tests, and the existence of more modern approaches to asset pricing and portfolio selection (such as arbitrage pricing theory and Merton's portfolio problem), the CAPM still remains popular due to its simplicity and utility in a variety of situations.

In CAPM model, we assume that there is no intercept parameter :

$$E(r_i) - r_f = \beta_i \left(E(r_m) - r_f \right)$$

- $r_i, E(r_i)$ are return rate and the expected return on the capital asset.

- r_f is the risk-free rate of interest such as interest arising from government bonds.

- $r_m, E(r_m)$ are return rate and the expected return of the market.

- $E(r_m) - r_f$ is sometimes known as the market premium.

- $E(r_i) - r_f$ is also known as the risk premium.

- β_i is the sensitivity of the expected excess asset returns to the expected excess market returns, calculated by (as in subsection 2.11.1):

$$\beta_i = \frac{Cov(r_i, r_m)}{Var(r_m)} = \rho_{i,m} \frac{\sigma_i}{\sigma_m}$$

where r_i and r_m are respectively return of a security and of the market. ρ, σ means the correlation and the std.

- $|\beta_i| > 1$: agressive

- $|\beta_i| < 1$: defensive

- Some investments — such as put options — have negative betas, meaning that they would be expected to move in the opposite direction of the index.

- The expected market rate of return (r_m) is usually estimated by measuring the arithmetic average of the historical returns on a market portfolio (e.g. S&P 500).

- The risk free rate of return (r_f) is usually the arithmetic average of historical risk free rates of return and not the current risk free rate of return.

- β_i is normally calculated with 5 past years, and used to infer for future value.

8.1 Beta calculation and alpha

Remind that the return rate for a periode can be calulated by:
(Ending Price - Beginning Price + Dividends) / Beginning Price

In practive, we can the closed price of asset call p_n in day n, then $r_i^{n-k} = \frac{p_n - p_{n-k}}{p_{n-k}}$. For big number of n, we have vector $R_i = [r_i^0, r_i^1, \ldots, r_i^{n-k}]$ For the market return rate, we perform in the corresponding way and get $R_m = [r_m^0, r_m^1, \ldots, r_m^{n-k}]$

Fitting : Finding $\beta_i = \frac{Cov(r_i, r_m)}{Var(r_m)}$ or by uisng implemented solution for $y = \beta x + \alpha$ (degree 1). After fitting, usually, in inference, we use return rate which is annually expressed.

For example can see : https://www.mlq.ai/capital-asset-pricing-model-python/

Or, even easier, we could just look up a stock's beta online (Yahoo Finance and FinViz).

8.2 Security market line

This can be described by the figure 6.3.

8.3 Inference

The objective is to infer expected or required rate of return, or equity discount rate $E(r_i)$. Let take an example :

The current risk-free rate is 2% on a short-term U.S. Treasury. The long-term average rate of return for the market is 10%. Let's say Company A has a beta of 1.50, meaning that it is riskier than the overall market (which has a beta of 1).To invest in Company A :

$$RRR = 2\% + 1.50X(10\% - 2\%) = 14\%$$

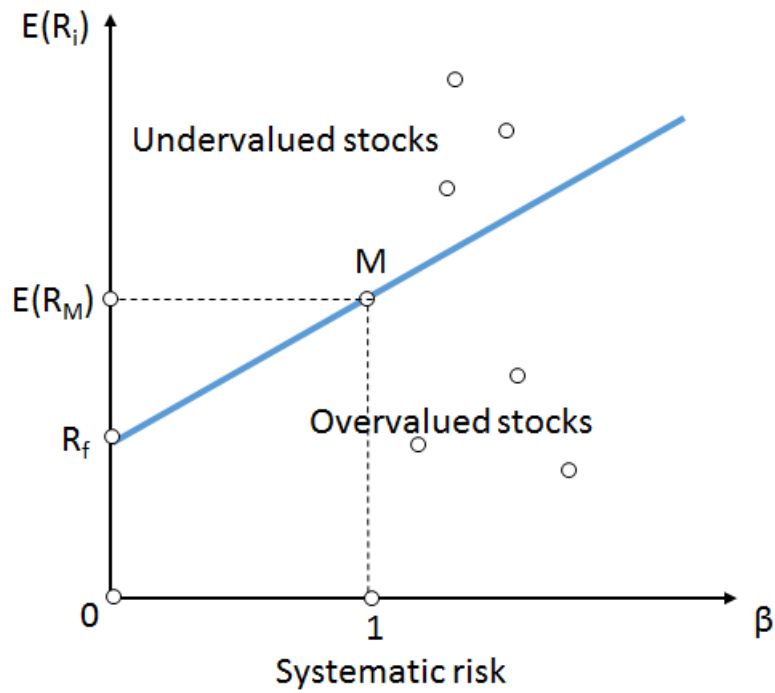Figure 6.3: The SML graphs the results from the capital asset pricing model (CAPM) formula. The x-axis represents the risk (β), and the y-axis represents the expected return. After estimating β_i for a given stock i from a long recent period, along with $E(r_m) - r_f$ (fixed), we can check the actual stock i is over-valued or undervalued. The market risk premium is determined from the slope of the SML.

Company B has a beta of 0.50, which implies that it is less risky than the overall market. To invest in Company B:

$$RRR = 2\% + 0.50X(10\% - 2\%) = 6\%$$

8.4   Beta Analysis

If you feel that your portfolio is too conservative or stagnant — that you're missing out on gains because your investments don't move very much — then it's likely that your portfolio as a whole has a beta of less than 1. Adding stocks with a beta greater than 1 would add more volatility relative to the market (and the potential for higher returns).

Conversely, if you feel that your portfolio is too risky — that you can't stomach its big upward and downward swings — then there's a good chance that your portfolio has a beta greater than 1. Adding stocks with betas of less than 1 would help decrease volatility relative to the market.

In short, **adding high-beta stocks tend to increase a portfolio's overall volatility and low-beta stocks tend to decrease it**.

8.5   Limitations of beta analysis

Beta analysis can be a useful way to manage the level of risk in your portfolio, but like any financial technique, it's not perfect.

One major drawback of beta is that it's a backward-looking metric. It's calculated based on past returns, which may not be consistent with future returns.

Beta's backward-facing nature also means it's not very useful for evaluating younger publicly traded companies without long-term track records.

It's also worth noting that beta only measures one kind of risk: systematic risk, or how a stock will respond to upturns or downturns in the overall market. It doesn't include any information on how a company is actually run. Earnings, cash flow, debt and dividends are all important fundamental measures of a company, but none directly affects a stock's beta (or relfected by stock's beta).

8.6   Asset pricing

Given P_i^T the future price of security or portfolio (can be obtained by simulation or by the future market ...). Let consider 2 cases:

- If we know r_i, which is the return rate associated with this security, usually annual. Then the present value P_i^0 can be obtained by using r_i as discounted rate:
$$P_i^0 = \frac{P_i^T}{1 + r_i}$$

- If we do not know r_i, we apply the CAPM, where the price P_0 is calculated by (the lines after is the demonstration) :

$$P_i^0 = \frac{1}{1+r_f}\left(E(P_i^T) - \frac{Cov(P_i^T, r_m)}{Var(r_m)}(E(r_m) - r_f)\right)$$

$$\Leftrightarrow 1 = \frac{1}{1+r_f}\left(\frac{E(P_i^T)}{P_i^0} - \frac{Cov(\frac{P_i^T}{P_i^0}, r_m)}{Var(r_m)}(E(r_m) - r_f)\right)$$

$$\Leftrightarrow 1 = \frac{1}{1+r_f}\left(E(1+r_i) - \frac{Cov(\frac{P_i^T}{P_i^0} - 1, r_m)}{Var(r_m)}(E(r_m) - r_f)\right)$$

$$\Leftrightarrow 1 = \frac{1}{1+r_f}\left(E(1+r_i) - \frac{Cov(r_i, r_m)}{Var(r_m)}(E(r_m) - r_f)\right)$$

$$\Leftrightarrow 1 = \frac{1}{1+r_f}\left(E(1+r_i) - (E(r_i) - r_f)\right)$$

$$\Leftrightarrow 1 = \frac{1}{1+r_f}(1+r_f)$$

9 Dividend discount model

In finance and investing, the dividend discount model (DDM) is a method of valuing the price of a company's stock based on *the fact that its stock is worth the sum of all of its future dividend payments, discounted back to their present value.* This model DDM is sometimes referred to as the Gordon growth model (GGM).

Let's note:

- P is the current stock price.

- g is the constant growth rate in perpetuity expected for the dividends.

- r is the constant cost of equity capital for that company

- D_t is the value of dividends at the end of the t first period.

Assumption: the price of stock, which is the money that investor must pay (never reselling after), is equal to the sum infinity of discounted dividends (back to present value). Hence:

$$P = \sum_{t=1}^{+\infty} \frac{D_t}{(1+r)^t} = \sum_{t=1}^{+\infty} \frac{D_0(1+g)^t}{(1+r)^t}$$

As usually (hypothesis, but not always true) that $g < r$ means that growth rate of dividend is less than one of stock's price. Using geometric suite, we have :

$$P = \frac{D_1}{r - g}$$

or the cost of equity (return rate) can be infered by :

$$r = \frac{D_1}{P} + g$$

9.1 Example

A company is expected to pay an annual dividend of \$3 next year, and its stock is currently trading at \$100 a share. The company has been steadily raising its dividend each year at a 4% growth rate.

$$RRR = (3/100) + 4 = 7\%$$

10 Fama-French three-factor model

This is an extention of the traditional CAPM described in section 8. In CAPM, we use only one variable β to compare the returns of a portfolio or stock with the returns of the market as a whole. In contrast, the Fama–French model uses three variables. Fama and French started with the observation that two classes of stocks have tended to do better than the market as a whole:

- Stock with small caps (market capitalization)

- Stock with high book-to-market ratio (B/P)

They then added these two factors to CAPM to reflect a portfolio's exposure to these two classes :

$$E(r_i) - r_f = \beta_i(E(r_m) - r_f) + s_i \cdot SMB + h_i \cdot HML + \alpha_i$$

Here r is the portfolio's expected rate of return, r_f is the risk-free return rate, and r_m is the return of the market portfolio. The parameter β is analogous to the classical β but not equal to it, since there are now two additional factors to do some of the work. SMB stands for Small (market capitalization) Minus Big and HML for High (book-to-market ratio) Minus Low.

In general, the Fama-French three-factor model, stock returns are explained by three factors: the market factor, the size factor, and the value factor. The market factor represents the return of the overall stock market, the size factor SMB represents the performance of small-cap stocks relative to large-cap stocks, and the value factor HML represents the performance of value stocks (those with low price-to-book ratios) relative to growth stocks (those with high price-to-book ratios).

The SMB factor is calculated as the difference between the returns of a portfolio of small-cap stocks and a portfolio of large-cap stocks. Specifically,

the SMB factor is the average monthly return of the small-cap portfolio minus the average monthly return of the large-cap portfolio. A positive SMB value indicates that small-cap stocks outperformed large-cap stocks during that period, while a negative SMB value indicates that large-cap stocks outperformed small-cap stocks.

The HML factor is calculated as the average monthly return of a portfolio of high B/M stocks minus the average monthly return of a portfolio of low B/M stocks. A positive HML value indicates that high B/M stocks outperformed low B/M stocks during that period, while a negative HML value indicates that low B/M stocks outperformed high B/M stocks.

10.1 Estimating parameters

Given $E(r_i), r_f, E(r_m), SMB, HML$, the parameters need to be estimated is $\theta = \{\beta, s, h, \alpha\}$. One of methods is using linear regression. We can also employ the Generalized method of moments method as described in section **??**.

11 Arbitrage pricing theory

Arbitrage pricing theory (APT) is a multi-factor model for asset pricing which relates various macro-economic (systematic) risk variables to the pricing of financial assets. Proposed by economist Stephen Ross in 1976, it is widely believed to be an improved alternative to its predecessor, the Capital Asset Pricing Model (CAPM) in section 8.

The law of one price (LOOP) states that in the absence of trade frictions (such as transport costs and tariffs), and under conditions of free competition and price flexibility (where no individual sellers or buyers have power to manipulate prices and prices can freely adjust), identical goods sold in different locations must sell for the same price when prices are expressed in a common currency

11.1 Model's formula

$$r_j = a_j + \beta_{j_1} f_1 + \beta_{j_2} f_2 + \ldots + \beta_{j_n} f_n + \epsilon_j$$

where:

- a_j is a constant for asset

- f_n is a systematic factor. A systematic factor, also known as a systematic risk factor or a market factor, refers to a common source of risk that affects a broad set of assets or securities within a particular market or industry.

- β_{j_n} is the sensitivity of the asset to factor n, also called factor loading

- ϵ_j is the risky asset's idiosyncratic random shock with mean zero.

The APT model states that if asset returns follow a **factor structure** then the following relation exists between expected returns and the factor sensitivities:

$$E(r_j) = r_f + \beta_{j_1} RP_1 + \beta_{j2} RP_2 + \ldots + \beta_{jn} RP_n$$

where:

- RP_n is the risk premium of the factor

- r_f is the risk-free rate

11.2 Factor structure

In finance, factor structure refers to the *underlying relationships* or *patterns* among a set of variables, typically used to explain the returns of assets or securities. It involves identifying the common factors that drive the variation in asset returns and understanding how these factors influence the pricing and risk characteristics of investments.

- Factor Analysis: Factor analysis is a statistical technique used to identify and analyze the underlying factors that explain the **covariance structure** of a set of variables. It helps reduce the dimensionality of the data by identifying a smaller number of latent factors that account for the majority of the variation in the observed variables. In finance, factor analysis is often applied to stock returns to identify the common factors driving their variation.

- Factor Models: Factor models are asset pricing models that incorporate **systematic factors** to explain the returns of securities or portfolios. These models assume that asset returns can be explained by a **linear combination** of one or more factors, along with a random error term.

- Risk Factors: Risk factors are key components of the factor structure in finance. They represent the underlying **sources of risk** that drive the variation in asset returns. Examples of risk factors include market risk, interest rate risk, inflation risk, currency risk, industry-specific risk, and various macroeconomic factors.

11.3 Model by matrix

For a set of m assets with returns $r \in \mathbb{R}^m$, factor loadings (parameters) $B \in \mathbb{R}^{m \times n}$, and factors $F \in \mathbb{R}^n$, error (noise) $\varepsilon \in \mathbb{R}^m$, a general (matrix) factor model that is used in APT is:

$$r = r_f + BF + \varepsilon, \quad \varepsilon \sim N(0, \Psi)$$

and $F \sim N(\mu, \Omega)$, where μ is the expected risk premium vector and Ω is the factor covariance matrix. Assuming that the noise terms for the returns and factors

are uncorrelated, the mean and covariance for the returns are respectively:

$$E(r) = r_f + B\mu, \quad Cov(r) = B\Omega B^T + \Psi$$

Solving for matrix B : It is generally assumed that we know the factors F in a model, which allows least squares to be utilized. However, an alternative to this is to assume that the factors are latent variables (in this case factors F do not need to be known) and employ factor analysis (6).

12 Weighted average cost of capital

The weighted average cost of capital (WACC) is a financial metric that represents the average rate of return a company needs to generate to cover the costs of its various sources of financing, including **equity** and **debt**. It provides a measure of the minimum acceptable return a company should achieve on its investments.

equity o day duoc hieu la phat hanh co phieu. Both equity and debt are used to raise capital.

12.1 General definition

$$WACC = \frac{\sum_{i=1}^{N} r_i \cdot MVO_i}{\sum_{i=1}^{N} MVO_i} = w_i r_i$$

where N is the number of sources of capital, which include equity (popular or preferred security,...) and (bond, debt,...), r_i is the **required** rate of return (or cost of equity) security i, MVO means the market value of all outstanding securities.

The market value of all outstanding securities refers to the total value of all publicly traded securities (contrary to no-publicly traded securities) issued by a company or available in the market.

Example For example, let's consider a company with the following outstanding securities:

Common Stock:

- Market price per share: $50

- Number of shares outstanding: 10 million

Preferred Stock:

- Market price per share: $100

- Number of shares outstanding: 2 million

Bonds:

- Market price per bond: $1,000

- Number of bonds outstanding: 5,000

Then

- MVO of common stock = $50 × 10,000,000 = $500,000,000

- MVO of preferred stock = $100 × 2,000,000 = $200,000,000

- MVO of bonds = $1,000 × 5,000 = $5,000,000

Finally

$$WACC = \frac{500}{705}r_{CS} + \frac{200}{705}r_{PS} + \frac{5}{705}r_B$$

where r_{CS}, r_{PS}, r_B are respectively the rate of return of common stock, preferred stock and the interest rate of bond.

12.2 Short definition

$$WACC = \frac{E}{D+E}R_e + \frac{D}{D+E}R_d$$

where E is the total shareholder's equity, D is the total debt, R_e is the cost of equity, R_d is the cost of debt.

Sometime, there is an intervertion of the *corporate tax rate t*. The corporate tax rate refers to the percentage of profits that corporations are required to pay in taxes on their taxable income. The corporate tax rate can vary between countries and jurisdictions and may be subject to changes over time.

The main point is that the profit is taxable and the debt is not. Hence, if we want a return after tax of $a\%$, then the return before tax is $\frac{a}{1-t}$. Then

$$WACC\frac{1}{(1-t)} = \frac{E}{D+E}R_e\frac{1}{(1-t)} + \frac{D}{D+E}R_d$$

$$\Leftrightarrow WACC = \frac{E}{D+E}R_e + \frac{D}{D+E}R_d(1-t)$$

Example A company that have the following state

- Equity: $4,000,000

- Debt: $6,000,000

- Cost of equity (R_e): 10%

- Cost of debt (R_d): 5%

- Corporate tax rate: 30%

Next, calculate the WACC using the formula:

$$WACC = (0.40.10) + (0.60.05)(1-0.30)$$

$$WACC = 0.061 = 6.1\%$$

This means that the company needs to generate a return of at least 6.1% on its investments to cover its financing costs and provide value to its shareholders.

13 Alpha

We saw in CAPM 8, APT 11 the beta, which the sensitivity to non-diversifiable or systematic factor. In this section, we try to understand alpha. Before going into definition of alpha, we discover some helpful terminologies :

13.1 Helpful terminologies

Passive Management: Also known as passive investing or index investing, passive management aims to replicate the performance of a specific market index or benchmark. The goal is to achieve a return that closely matches the performance of the chosen index, rather than outperforming it. Passive managers typically employ a buy-and-hold strategy and minimize trading activity. Examples of passive management strategies include:

- Index Funds: These funds replicate the performance of a specific market index, such as the S&P 500. They hold a diversified portfolio of securities that closely mimics the index composition.

- Exchange-Traded Funds (ETFs): ETFs are similar to index funds but trade on stock exchanges (e.g. S&P 500) like individual stocks. They provide exposure to various asset classes or market sectors and track specific indices.

- Target-Date Funds: These funds are commonly used for retirement savings and automatically adjust the asset allocation over time based on a predetermined target retirement date.

Active Management: Active management involves making investment decisions with the goal of outperforming the market or a specific benchmark. Active managers employ research, analysis, and trading strategies to select securities they believe will generate superior returns. They aim to beat the market by identifying mispriced assets or taking advantage of market trends. Active management typically involves higher costs, including management fees and transaction costs, as compared to passive management. Additionally, the performance of active managers can vary, with some outperforming the market and others underperforming. Examples of active management strategies include:

- Mutual Funds: Mutual funds are professionally managed investment vehicles that actively select and manage a diversified portfolio of securities. Fund managers make ongoing investment decisions to generate returns higher than the benchmark.

- Hedge Funds: Hedge funds are investment partnerships that employ various strategies to generate absolute returns. Hedge fund managers actively seek opportunities in multiple markets, often using complex investment techniques.

- Quantitative Strategies: These strategies rely on computer-based algorithms and quantitative models to make investment decisions. They analyze large amounts of data and historical patterns to identify trading opportunities.

Active return, also known as active performance or excess return, refers to the difference in investment performance between an actively managed portfolio and a benchmark or passive investment strategy. It measures the extent to which an investment portfolio has outperformed or underperformed its benchmark. Active return is calculated by subtracting the return of the benchmark from the return of the portfolio. The formula for active return is as follows:

$$ActiveReturn = PortfolioReturn - BenchmarkReturn$$

A positive active return indicates that the portfolio has outperformed the benchmark, while a negative active return signifies underperformance.Active return is a key metric used to evaluate the success of active management strategies. It helps investors assess whether the additional risks and costs associated with active management have led to superior investment performance compared to a passive approach.

It's important to note that active return is just one aspect of evaluating an actively managed portfolio. Other factors, such as risk-adjusted return, consistency of performance, and investment style, should also be considered when assessing the effectiveness of an active management strategy.

13.2   Definition of alpha

Alpha is a **measure** of the active return on an investment, the performance of that investment compared with a suitable market index. An alpha of 1% means the investment's return on investment over a selected period of time was 1% better than the market during that same period; a negative alpha means the investment underperformed the market. Alpha, along with beta, is one of two key coefficients in the capital asset pricing model used in modern portfolio theory and is closely related to other important quantities such as *standard deviation, R-squared and the Sharpe ratio.*

As funds (using Active Management) include various fees normally expressed in percent terms, theese funds has to maintain an alpha greater than its fees in order to provide positive gains, compared with an index fund. Historically, the vast majority of traditional funds have had negative alphas, which has led to a flight of capital to index funds and non-traditional hedge funds.

Example: At the date x, Faye invested in a stock. At the date Faye's invesment has a return of 20%. In the same periode from x to y. The average return or the benchmark return is 9%. Then alpha is simply 11%.

It is also possible to analyze a portfolio of investments and calculate a theoretical performance, most commonly using the capital asset pricing model (CAPM).
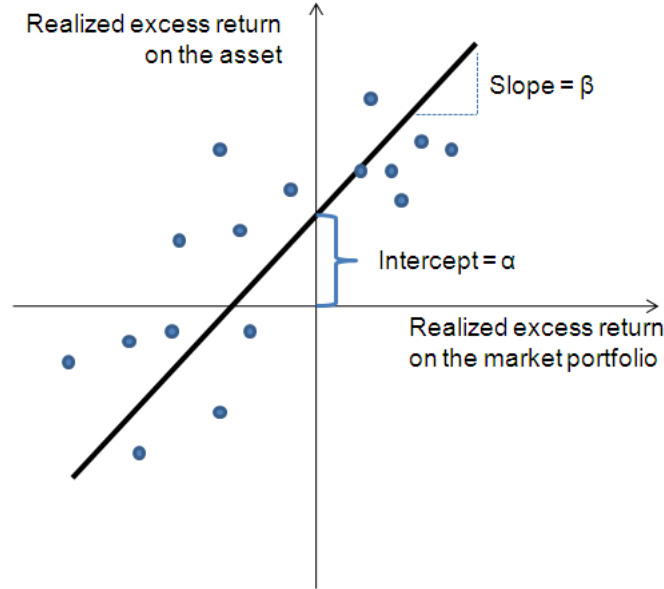
Figure 6.4: Security characteristic line

13.3 Definition of alpha in single index model

The alpha coefficient (α_i) is a parameter in the **single index model** (SIM) :

$$R_{i,t} - R_f = \alpha_i + \beta_i(R_{M,t} - R_f) + \varepsilon_{i,t}$$

where the following inputs are:

- R_i: the portfolio return

- R_M: the market return

- R_f: the risk-free rate of return

- β_i: the beta of the portfolio

In this definition, alpha and beta are both parameter to be estimated by fitting the SIM with data of portfolio return and market return.

Alpha is also the *intercept* of the **security characteristic line** (SCL), as in figure 6.4 (do not confuse with SML in figure 6.3)

It can be shown that in an efficient market, the expected value of the alpha coefficient is zero (the assumtion in CAPM). Therefore, the alpha coefficient indicates how an investment has performed after accounting for the risk it involved:

- $\alpha_i < 0$: the investment has earned too little for its risk (or, was too risky for the return)

- $\alpha_i = 0$: the investment has earned a return adequate for the risk taken

- $\alpha_i > 0$: the investment has a return in excess of the reward for the assumed risk

For instance, although a return R_i of 20% may appear good, the investment can still have a negative alpha if it's involved in an excessively risky position.

13.4 Definition by Jensen or Jensen's alpha

In finance, Jensen's alpha is used to determine the abnormal return of a security or portfolio of securities over the theoretical expected return. Alpha is calculated based on a theoretical performance instead of a market index as before. This is based on the concept that riskier assets should have higher expected returns than less risky assets.

$$\alpha_i = R_{i,t} - R_f - \beta_i(R_{M,t} - R_f)$$

In this definition, beta had been estimated from CAPM, and alpha is just inferred with the formula.

Example Now, let us assume Faye's portfolio returned 8%, a risk-free ROR of 11%, and a beta of 1.8% against a benchmark of 9%. Then alpha is 8-11-1.8(9-11) = 0.6

14 Black–Scholes model

14.1 Option reminding

By Fischer Black and Myron Scholes; Robert C. Merton,
Pricing of European call or put option (calculate value of an option).
Hypothesis for Black–Scholes model (but can be relaxed):

- Efficient market

- Absence of transisition cost

- No dividend payment

- Know Volatility and risk-free

E.g, an invester want to realize an option :

- Share price > Strike price : Exercise

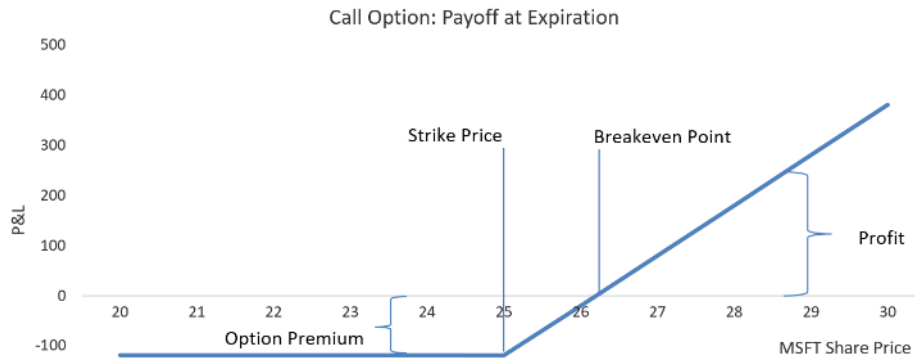- Strike price > Share price : Don't Exercise

Figure 6.5: Call option payoff. 100 shares, cost for call option is $120. Strike price is $25. "In-the-money" is the zone after strike price and "out-of-the-money" is the zone before strike price point. "At the price" is on strike price point

"In-the-money" is a term used in options trading to describe a situation where an option has intrinsic value. An option is considered "in-the-money" if the current market price of the underlying asset, such as a stock, is favorable in comparison to the option's strike price.

For example, if an investor holds a *call option* with a strike price of $50 and the stock price is currently trading at $60, the call option is "in-the-money" because the holder has the right to buy the stock at $50, which is lower than its current market price. The same, if an investor holds a *put option* with a strike price of $50 and the stock price is currently trading at $40, the put option is "in-the-money" because the holder has the right to sell the stock at $50, which is higher than its current market price.

Being "in-the-money" means the option has intrinsic value and, if exercised, would result in a profit for the option holder. Conversely, if an option is "out-of-the-money", it means it has no intrinsic value and would result in a loss if exercised.

14.2   Annualizing Volatility

Annualizing Volatility equal to *std(daily log return) multiplied by sqrt(number of trade days)*

14.3   Black–Scholes reasoning

The Black-Scholes formula is used to calculate the theoretical value of a European option from the following five data:

- S_0 the current value of the stock.

- T the time remaining for the option before expiration.

- K the strike price set by the option.

- r is risk-free rate.

- σ or s the volatility of the stock price.

- C Option premium or cost for option buy.

If the first four data are obvious, the volatility σ of the asset is difficult to assess. Two analysts may have a different opinion on the value of σ to choose.

$$C \exp(rT) = \mathbb{E}[(S_T - K)^+]$$

LHS is the cost multiplied by interest rate after the time T and RHS is the mean of payoff, hence these two quantities need to be equal.

$$
\begin{aligned}
\mathbb{E}[(S_T - K)^+] &= \mathbb{E}[(S_T - K)\mathbb{1}_{S_T > K}] \\
&= \mathbb{E}(S_T \mathbb{1}_{S_T > K}) - \mathbb{E}(K \mathbb{1}_{S_T > K}) \\
&= \mathbb{E}(S_T \mathbb{1}_{S_T > K}) - P(S_T > K)K
\end{aligned}
$$

Applying stock simulation (in sec 16.2) :

$$S_T = S_0 \exp\left((r - \frac{1}{2}s^2)T + sW_T \right)$$

For the second term $P(S_T > K)K$:

$$
\begin{aligned}
P(S_T > K) &= P\left(S_0 \exp\left((r - \frac{1}{2}s^2)T + sW_T \right) > K \right) \\
&= P\left((r - \frac{1}{2}s^2)T + sW_T > \ln(\frac{K}{S_0}) \right) \\
&= P\left(\frac{1}{\sqrt{T}}W_T > \frac{\ln(\frac{K}{S_0}) - (r - \frac{1}{2}s^2)T}{s\sqrt{T}} \right) \\
&= P\left(\frac{1}{\sqrt{T}}W_T < \frac{(r - \frac{1}{2}s^2)T - \ln(\frac{K}{S_0})}{s\sqrt{T}} \right)
\end{aligned}
$$

Let $d_2 = \frac{(r - \frac{1}{2}s^2)T - \ln(\frac{K}{S_0})}{s\sqrt{T}}$ then $P(S_T > K) = N(d_2)$ where N is cumulative distribution function of normal distribution.

For the first term $\mathbb{E}(S_T \mathbb{1}_{S_T > K})$, let f_{S_T} be the density function of S_T which is log-normal distribution:

$$f_{S_T}(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2} \right)$$

where $\mu = ln(S_0) + (r - \frac{1}{2}s^2)T$ and $\sigma = s\sqrt{T}$

$$
\begin{aligned}
\mathbb{E}(S_T \mathbb{1}_{S_T > K}) &= \int_K^{+\infty} x f_{S_T}(x) dx \\
&= \int_K^{+\infty} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right) dx \\
&= \int_K^{+\infty} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right) dx \\
&= \int_{\frac{\ln K - \mu}{\sigma}}^{+\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right) \exp(\sigma y + \mu) dy \\
&= \int_{-d_2}^{+\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2} + \sigma y + \mu\right) dy \\
&= \int_{-d_2}^{+\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(y-\sigma)^2 + \mu - \frac{\sigma^2}{2}\right) dy \\
&= \int_{-d_2-\sigma}^{+\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right) \exp\left(\mu + \frac{\sigma^2}{2}\right) dz \\
&= N(d_1) \exp\left(\mu + \frac{\sigma^2}{2}\right) = N(d_1) \exp(\ln S_0 + rT) = N(d_1) S_0 \exp(rT)
\end{aligned}
$$

where $d_1 = d_2 + \sigma = \frac{(r + \frac{1}{2}s^2)T - \ln(\frac{K}{S_0})}{s\sqrt{T}}$. Finally,

$$
C = N(d_1)S_0 - N(d_2)K \exp(rT)
$$

Note that, for calculating expected value, !!! avoid to write direct own variable S_T such as :

$$
\int_K^{+\infty} S_T P(S_T) dS_T
$$

Second, avoid to using probability $P(x = a)$

$$
\int_K^{+\infty} x P(S_T = x) dx
$$

Must using density function f with an own varible associated to, such as :
f_{S_T}

14.4 Why not using the own drift of given stock ?

In the Black-Scholes model, the drift term is assumed to be equal to the risk-free rate of return, which is the return an investor would receive on a risk-free investment such as a U.S. Treasury bond. The reason for this assumption is

that the risk-free rate represents the opportunity cost of investing in the asset, and is thus a key factor in determining the asset's fair value.

On the other hand, the own drift of a given stock represents the expected rate of return on that stock based on its past performance and future growth prospects. While the own drift of a stock is an important factor in determining its fair value, it may not be directly relevant to the option pricing problem that the Black-Scholes model is designed to solve. Instead, the model focuses on the risk-neutral probability of the underlying asset's future price movements, which can be calculated using the risk-free rate and the asset's volatility.

By using the risk-free rate in the Black-Scholes model, we can calculate the fair value of an option based on the assumption that the underlying asset follows a risk-neutral process. This means that the option price reflects the market's expectation of the future price movements of the underlying asset, without being influenced by the asset's own drift or other factors that may affect its expected return. This approach is widely used in option pricing and other financial modeling applications, and has been shown to produce accurate results in many cases.

15 Forward price

Forward price is the price at which two parties agree to buy or sell an asset at a future date. This price is agreed upon at the time of the contract, but the actual exchange of the asset and payment occurs on the future date specified in the contract.

The forward price is typically based on the current market price of the asset, adjusted for any costs or interest rates that may be applicable. It is used by investors and traders to manage their exposure to future price movements of an asset and to lock in prices for future transactions.

$$F = S_0 e^{(r-q)T} - \sum_{i=1}^{N} D_i e^{(r-q)(T-t_i)}$$

where :

- F is the forward price to be paid at time T.

- r is the risk-free interest rate.

- q is the convenience yield. Its represents the benefits that accrue to the holder of an asset due to the ability to use or consume the asset, or to gain some other benefit from owning it.

- S_0 is the spot price of the asset (i.e. what it would sell for at time 0 or simply current price).

- D_i is a dividend that is guaranteed to be paid at time t_i where $0 < t_i < T$.

16 Binomial options pricing model

Invented frist by Cox, Ross et Rubinstein (1979). There are also other versions.

- In finance, the binomial options pricing model (BOPM) provides a generalizable numerical method for the valuation of options.

- The BOPM approach has been widely used since it is able to handle a variety of conditions for which other models cannot easily be applied. This is largely because the BOPM is based on the description of an underlying instrument over a period of time (with discretization) rather than a single point.

- As Black-Scholes model can only applicable to EU options, then for US options or Bermudan options, which holder exercise the option at any time (before the expiration date), we must have a tool to value options. Hence BOPM.

- Although computationally slower than the Black–Scholes formula, it is more accurate, particularly for longer-dated options on securities with dividend payments.

- For options with several sources of uncertainty (e.g., real options) and for options with complicated features (e.g., Asian options), binomial methods are less practical due to several difficulties, and Monte Carlo option models are commonly used instead

16.1 Method description

Three main steps:

- Step 1: Create the binomial price tree. From S_0 price of underlying instrument at the valuation date.

$$S_{i+1} = \begin{cases} S_i * u \text{ if price raises} \\ S_i * d \text{ if price decreases} \end{cases}$$

where u, d are chosen by different versions of BOPM.

- Step 2: Find option value (payoff) at each final node $\begin{cases} \max(S_n - K, 0) \text{ if call option} \\ \max(K - S_n, 0) \text{ if put option} \end{cases}$

- Step 3: Find option value at earlier nodes. Once the step 2 is complete, the option value is then found for each node, starting at the penultimate time step, and working back to the first node of the tree (the valuation date) where the calculated result is the value of the option.
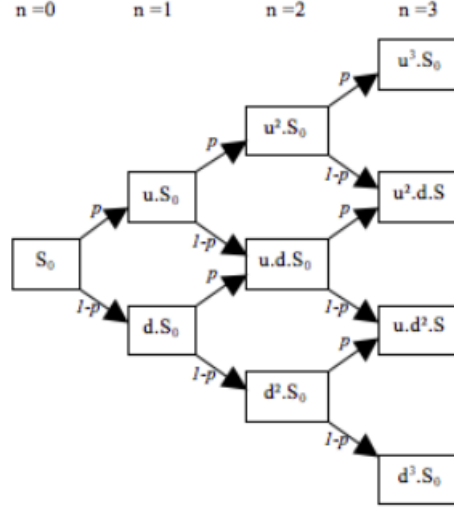
Figure 6.6: Binary tree with Bernoulli(p,u,d).

$OptionValue(i,k)\exp{(r\Delta t)} = p{\times}OptionValue(i{+}1,k){+}(1{-}p){\times}OptionValue(i{+}1,k{+}1)$

where i index means time instant (or n in horizontal axe), and k index means for vertical axe according to figure 6.6.

Then for, e.g., in call option :

if $OptionValue(i,k) < \max(S_{i,k}{-}K,0)$ then $OptionValue(i,k) = \max(S_{i,k}{-}K,0)$.

In this option, we make a simple hypothesis that the investors are smart, they know what is a greater payoff between $OptionValue(i,k)$ and $\max(S_{i,k}{-}K,0)$. This point is a advantage of time discritizating, which does not exist in BS model.

16.2 General conditions for all versions

In the step 3, p is chosen such that the related binomial distribution simulates the geometric Brownian motion of the underlying stock with parameters r and σ, which means the *expected value of price* and *variance of price*, simulated by the binomial distribution matches ones simulated by the geometric Brownian motion, at all time instant t. Let $S_A(n)$ be the price simulated by BOPM and $S_B(t) := S_2(n\Delta_t)$ be the price simulated by the geometric Brownian motion :

- $S_A(n) = S_0 u^k d^{n-k}$

- $S_B(t) = S_0 \exp\left((a - \frac{b^2}{2})t + b\sqrt{t}\frac{B_t}{\sqrt{t}}\right)$

318

Conditions need to be satisfied :

$$\begin{cases} \mathbb{E}[S_A(n)] = \mathbb{E}[S_B(n\Delta_t)] \\ Var[S_A(n)] = Var[S_B(n\Delta_t)] \end{cases}$$

Note that using these conditions may hard for calculating, we can think to using *log* function, sush as :

$$\begin{cases} \mathbb{E}\left[log(\frac{S_A(n)}{S_0})\right] = \mathbb{E}\left[log(\frac{S_B(n\Delta_t)}{S_0})\right] \\ Var\left[log(\frac{S_A(n)}{S_0})\right] = Var\left[log(\frac{S_B(n\Delta_t)}{S_0})\right] \end{cases}$$

16.3 First attempt

Expected value

(*) The expected value of price simulated by the geometric Brownian motion :

- Forward price simulation :

$$S_B(t) = S_0 \exp\left((a - \frac{b^2}{2})t + b\sqrt{t}\frac{B_t}{\sqrt{t}}\right)$$

- As the above $S(t)$ follows the log normal distribution, then :

$$E[S_B(t)] = S_0 \exp\left((a - \frac{b^2}{2})t + \frac{b^2 t}{2}\right) = S_0 \exp(at) = S_0 \exp(an\Delta_t)$$

(*) The expected value of price simulated by the binomial distribution :

- In this way, $S_A(i)$ is calculated from $S_A(i-1)$ in combining with $X(i)$, sample from Bernoulli distribution (p, u, d)

$$S_A(i) = S_A(i-1)X(i)$$

- Then

$$S_A(n) = S_0\Pi_{i=1}^n X(i)$$

- The expected value :

$$E[S_A(n)] = S_0 E[X]^n = S_0(pu + (1-p)d)^n$$

Hence by identifing S by 2 methods :

$$p = \frac{e^{a\Delta_t} - d}{u - d}$$

Note that p must be in $(0, 1)$ then $\frac{e^{a\Delta_t} - d}{u-d}$ has to be between 0 and 1.

Variance

(**) The variance of price simulated by the geometric Brownian motion :

$$Var[S_B(n\Delta_t)] = S_0^2(e^{b^2 t} - 1)e^{2(a - \frac{b^2}{2})t + b^2 t} = S_0^2(e^{b^2 n\Delta_t} - 1)e^{2an\Delta_t}$$

(**) The variance value of price simulated by the binomial distribution :

$$Var[S_A(n)] = S_0^2 Var[X...X] = S_0^2[(pu^2 + (1-p)d^2)^n - (pu + (1-p)d)^{2n}]$$

16.4 Second attempt

In the second attempt, we use the second conditions for log return.

Expected value

(*) By the geometric Brownian motion:

$$\mathbb{E}\left[log(\frac{S_B(n\Delta_t)}{S_0})\right] = (a - \frac{b^2}{2})t = (a - \frac{b^2}{2})n\Delta_t$$

(*) By the binomial distribution :

$$\mathbb{E}\left[log(\frac{S_A(n)}{S_0})\right] = \mathbb{E}[k\log(u) + (n-k)\log(d)]$$

$$= \mathbb{E}[k\log(\frac{u}{d}) + n\log(d)]$$

$$= \mathbb{E}[k]\log(\frac{u}{d}) + n\log(d)$$

$$= np\log(\frac{u}{d}) + n\log(d)$$

$$= n\left(p\log(\frac{u}{d}) + \log(d)\right)$$

Hence :

$$p\log(\frac{u}{d}) + \log(d) = (a - \frac{b^2}{2})\Delta_t$$

Variance

(**) By the geometric Brownian motion:

$$Var\left[log(\frac{S_B(n\Delta_t)}{S_0})\right] = Var[b\sqrt{t}\frac{B_t}{\sqrt{t}}] = b^2 t = b^2 n\Delta_t$$

(**) By the binomial distribution :

$$Var\left[log(\frac{S_A(n)}{S_0})\right] = Var[k\log(u) + (n-k)\log(d)]$$

$$= Var[k\log(\frac{u}{d}) + n\log(d)]$$

$$= np(1-p)\log(\frac{u}{d})^2$$

Hence :

$$p(1-p)\log(\frac{u}{d})^2 = b^2\Delta_t$$

16.5 Cox-Ross-Rubinstein

By imposing the third equation : $u = \frac{1}{d}$ and $\begin{cases} u = e^{\sigma\sqrt{\Delta_t}} \\ d = e^{-\sigma\sqrt{\Delta_t}} \end{cases}$

17 Monte Carlo methods for option pricing

Monte Carlo methods are a popular class of computational algorithms used in option pricing to estimate the value of an option by simulating multiple possible scenarios of the underlying asset's price movement over time.

The Monte Carlo method for option pricing involves the following steps:

1. Model the underlying asset's price movement using a stochastic process, such as the geometric Brownian motion model.

2. Generate a large number of simulated asset price paths using the stochastic process and random number generators.

3. Calculate the payoff of the option at the expiration date for each simulated price path.

4. Discount the payoffs back to the present using a risk-free interest rate.

5. Average the discounted payoffs to obtain the estimated option value.

The Monte Carlo method is particularly useful for pricing options with complex payoffs or when closed-form solutions are not available, such as American options. It also allows for the incorporation of various market factors and assumptions, such as volatility and interest rates, into the pricing model.

However, Monte Carlo simulations can be computationally intensive and may require significant computing power and time to produce accurate results. They also rely on the accuracy of the underlying stochastic model and the quality of the random number generators used in the simulation.

18 Kelly criterion

18.1 Gambling formula

$$f^* = p - \frac{q}{b} = p - \frac{1-p}{b}$$

where:

- f^* is the fraction of the current bankroll (tai khoan, so tien) to wager (danh cuoc).

- p is the probability of a win.

- q is the probability of a loss ($q = 1 - p$)

- b is the proportion of the bet gained with a win. E.g., if betting \$10 on a 2-to-1 odds bet (upon win you are returned \$30, means you win \$20), then $b = \$20/\$10 = 2.0$. Note that, if losing, we lose all what we bet, so b is also called b against 1.

We consider the three following cases :

- If a bet offers a 60% chance of winning the bet ($p = 0.60, q = 0.40, b = 1$), then the player should wager 20% of his budget at each opportunity ($f^* = 0.20$), in order to maximize long-term rates of capital growth.

- If the player has no advantage (his average statistical gain equals his average statistical loss), i.e. if $b = \frac{q}{p}$, then the rule is to bet nothing.

- If the player is at a disadvantage (his average statistical gain is less than his average statistical loss) ($b < \frac{q}{p}$) the formula yields a negative result, indicating that the player should take the other side of the bet.

18.2   With losing rate c

Always with the case of only two results, for the investissement, there is a small difference from the above formula. We introduce the parameter c, which is the ratio part of amout if we lose. E.g, if we invest 2\$, in case of losing, we lose 0.4\$, then $c = 0.2$.

$$f^* = \frac{p}{c} - \frac{q}{b}$$

Proof

- f, p, q, b, c are defined as above.

- We start with wealth of A_0, then at the game n, we have the wealth A_n.

- Suppose that we bet $n \to \infty$ times, in that we have pn times we win and qn we lose.

Then
$$A_n = A_0(1 + fb)^{pn}(1 - fc)^{qn}$$

And the average growth rate r is :

$$r = \sqrt[n]{\frac{A_n}{A_0}} = (1 + fb)^p(1 - fc)^q$$

r is also called expected geometric growth rate. We want to find the maximum r of this curve (as a function of f), which involves finding the derivative of the equation. This is more easily accomplished by taking the logarithm of each side first. The resulting equation is:

$$E = \log(r) = p\log(1 + fb) + q\log(1 - fc)$$

Solving for $\frac{dE}{df} = 0$, we get

$$f^* = \frac{p}{c} - \frac{q}{b}$$

In case that we do not know f before, this is because we do not know p. Suppose that we play N games and win K games, we know b and c, let Δ be a different amount from optimal f (which is fixed) :

$$A_0(1 + (f + \Delta)b)^K(1 - (f + \Delta)c)^{N-K}$$

We need to find optimal Δ :

$$\frac{Kb}{1 + (f + \Delta)b} - \frac{(N - K)c}{1 - (f + \Delta)c} = 0$$

$$\Leftrightarrow Kb - fKbc - \Delta Kbc - (N - K)c - (N - K)cfb - (N - K)c\Delta b = 0$$

$$\Leftrightarrow Kb - (N - K)c - Ncfb = Nc\Delta b$$

$$\Leftrightarrow \Delta = \frac{Kb - (N - K)c - Ncfb}{Ncb} = \frac{Kb - (N - K)c}{Ncb} - f$$

Then :

$$f + \Delta = \frac{Kb - (N - K)c}{Ncb}$$

This means that after each game, we update N and K and then use the value $\frac{Kb-(N-K)c}{Ncb}$.

18.3   Betting on multiple possible outcomes

Example: pari-mutuel betting on horse racing. Each horse i has a probability p_i to win, and proportion b_i in case of winning. Let f_i is the bet for each horse i. Then if the horse i win, we lose a fraction $1 - \sum_{j=1, j\neq i}^{N} f_j$ and win a fraction $f_i b_i$. The expected geometric growth rate r :

$$r = (1 - \sum_{j=2}^{N} f_j + f_1 b_1)^{p_1} \ldots (1 - \sum_{j=1}^{N-1} f_j + f_N b_N)^{p_N}$$

$$= \prod_{i=1}^{N}(1 - \sum_{j=1}^{N} f_j + f_i + f_i b_i)^{p_i}$$

The optimal bets is :

$$\begin{cases} f_i^* = \max(p_i - R/b_i, 0) \\ R = 1 - \sum_{j=1}^{N} f_j^* \end{cases}$$

323

18.4 Application to the stock market

In mathematical finance, if **security weights** maximize the **expected geometric growth rate**, then a portfolio is growth optimal.

Computations of growth optimal portfolios can suffer tremendous garbage in, garbage out problems. The "Garbage In, Garbage Out" (GIGO) problem refers to the concept that the quality of the output or results produced by a computer system is determined by the quality of the input data it receives. In simpler terms, if you provide a system with inaccurate, incomplete, or faulty data, the output or predictions generated by the system will also be inaccurate, unreliable, or nonsensical.

18.4.1 One risky asset and one risk free asset

Apply the stock simulation for a given risky asset (in 16.2), with estimated mean return rate μ and estimated votality σ (from one periode) :

$$S_t = S_0 \exp\left(\left(\mu - \frac{\sigma^2}{2}\right)t + \sigma W_t\right)$$

Now suppose that we use a fraction f on this risky asset and the rest $1 - f$ for the risk free asset, with the return rate r. Let R_c is the expected return of the combinaison of this risky asset and the free risk asset, then :

$$\begin{cases} E[R_c] = f\mu + (1 - f)r \\ \sigma_{R_c} = f\sigma \end{cases}$$

Then the stock simulation for combination :

$$S_t^C = S_0^C \exp\left(\left(f\mu + (1 - f)r - \frac{(f\sigma)^2}{2}\right)t + f\sigma W_t\right)$$

Then the expected log return R_1^C for an unit time length 1 :

$$f\mu + (1 - f)r - \frac{(f\sigma)^2}{2}$$

Then the optimal f^* is :

$$f^* = \frac{\mu - r}{\sigma^2}, \quad 0 \le f \le 1$$

The keypoint for the difference between optimal expected return found by this method and the moderne portfolio theory (MPT in sec 4) is :

- In this method, all investors do not care about risk.

- In this method, the expected return is calculated based on Geometric Brownian motion, which use parameters estimated from logarithmic returns.

- The logarithmic returns must follow the normal distribution.

18.4.2   Multi risky asset and one risk free asset

Let R_c is the expected return of the combinaison of multi risky assets and the free risk asset, then :

$$\begin{cases} E[R_c] = f_1\mu_1 + \ldots f_N\mu_N + (1 - \sum_{i=1}^{N} f_i)r \\ \sigma_{R_c} = \sqrt{f_1^2\sigma_1^2 + \ldots + f_N^2\sigma_N^2 + 2\sum_i^N \sum_{j>i}^N f_i f_j \sigma_{ij}} \end{cases}$$

Then the expected log return R_1^C for an unit time length 1 :

$$f_1\mu_1 + \ldots f_N\mu_N + (1 - \sum_{i=1}^{N} f_i)r - \frac{f_1^2\sigma_1^2 + \ldots + f_N^2\sigma_N^2 + 2\sum_i^N \sum_{j>i}^N f_i f_j \sigma_{ij}}{2}$$

Then the optimal f^* is :

$$\begin{cases} u_1 - r - f_1\sigma_1^2 - \sum_{j=1,j\neq 1}^N f_j \sigma_{1j} = 0 \\ u_2 - r - f_2\sigma_2^2 - \sum_{j=1,j\neq 2}^N f_j \sigma_{2j} = 0 \\ \vdots \\ u_N - r - f_N\sigma_N^2 - \sum_{j=1,j\neq N}^N f_j \sigma_{Nj} = 0 \\ 0 \leq \sum_i^N f_i \leq 1 \end{cases}$$

or let Σ is the covariance matrix among risky assets then :

$$\Sigma f^* = \mu - r, \quad 0 \leq \sum_i^N f_i \leq 1$$

19   Utility

In economics, utility refers to the pleasure, satisfaction or happiness that individuals derive from consuming goods and services. Utility is a subjective measure, meaning it varies from person to person and cannot be directly observed or measured. Instead, economists use utility as a theoretical concept to understand and analyze individuals' preferences and decision-making.

19.1   Utility function

A utility function is for the goal to quantify an individual's preferences and measures the level of utility (satisfaction) they derive from consuming different combinations (also called *alternative or bundle*) of goods and services. The utility function assigns a numerical value to each possible alternative, reflecting the individual's subjective preference for that alternative.

19.1.1 Example

Suppose James has utility function $U = \sqrt{xy}$ such that x is the number of apples and y is the number of chocolates. Alternative A has 9 apples and 16 chocolates; alternative B has 13 apples and 13 chocolates. Putting the values x, y into the utility function yields $\sqrt{9 \times 16} = 12$ for alternative A and $\sqrt{13 \times 13} = 13$ for B, so James prefers alternative B.

In general economic terms, a utility function ranks preferences concerning a set of goods and services.

19.1.2 Conditions required for utility function

For a finite set of alternatives, these require only that the preference ordering is complete (so the individual is able to determine which of any two alternatives is preferred or that they are indifferent).

If the set of alternatives is not finite (for example because even if the number of goods is finite), we may have a **continuous** utility function that maps from alternative space to the real space.

19.2 Cardinal or ordinal

Cardinal utility and ordinal utility are two different approaches to measuring and comparing the level of satisfaction or preferences of individuals. We use either cardinal or ordinal method, depending on whether they are or are not interpreted as providing more information than simply the rank ordering of preferences among bundles of goods, such as information concerning the strength of preferences.

Cardinal utility states that the utilities obtained from consumption **can be measured and are representable by numbers**. When cardinal utility is assumed, the magnitude of utility differences is an important quantity. For example, suppose a cup of orange juice has utility of 120 *utils*, a cup of tea has a utility of 80 *utils*, and a cup of water has a utility of 40 *utils*. With cardinal utility, it can be concluded that the cup of orange juice is better than the cup of tea by exactly the same amount by which the cup of tea is better than the cup of water. However, we cannot conclude, that the cup of tea is two thirds of the goodness of the cup of juice, because this conclusion would depend not only on magnitudes of utility differences, but also on the 'zero' of utility. For example, if the zero of utility was located at -40, then a cup of orange juice would be 160 utils more than zero, a cup of tea 120 utils more than zero. A cardinal utility function can be transformed to another utility function by a positive linear transformation (multiplying by a positive number, and adding some other number)

Ordinal. Instead of giving actual numbers over different bundles, ordinal utilities are only **the rankings** of utilities received from different bundles of goods or services. For example, ordinal utility could tell that having two ice creams provide a greater utility to individuals in comparison to one ice cream but could not tell exactly how much extra utility received by the individual. Ordinal
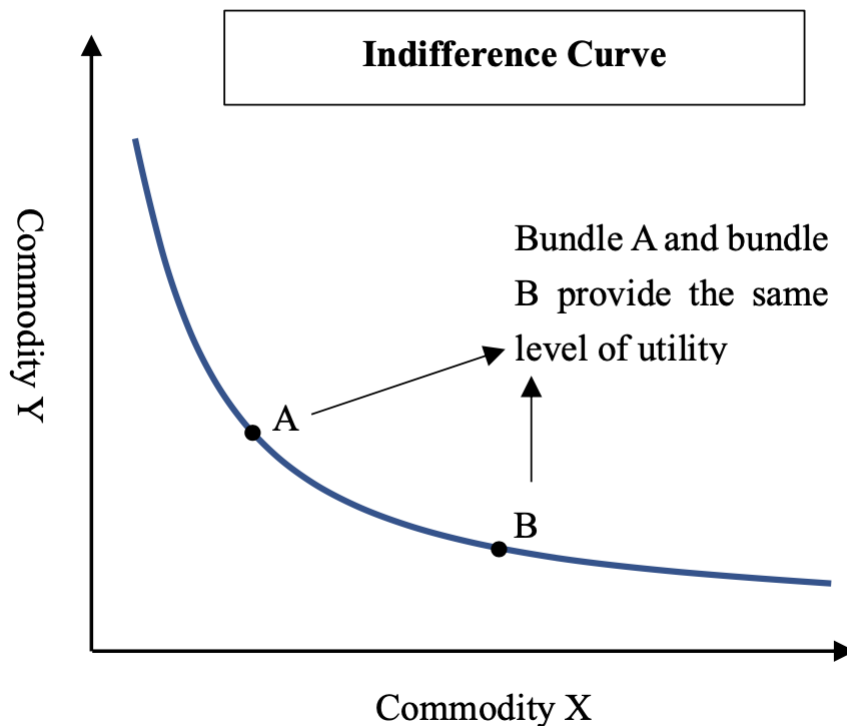
Figure 6.7: The blue line represents an indifference curve, on which bundles (composed by a quantity of commodity X and a quantity of commodity Y) have the same utility.

utility, it does not require individuals to specify how much extra utility he or she received from the preferred bundle of goods or services in comparison to other bundles. They are only required to tell which bundles they prefer.

If a function $u(x) > 0$ is ordinal, it is equivalent to the function $u(x)^2$. In contrast, if $u(x)$ is cardinal, it is not equivalent to $u(x)^2$, since $|u(x_1) - u(x_2)| \neq |u(x_1)^2 - u(x_2)^2|$.

19.3   Indifference curves

Indifference curves are graphical representations that depict **different combinations** of goods or services that yield the **same level of utility** for an individual. See figure 6.7.

19.4   Marginal Utility

Marginal utility refers to the additional utility gained from consuming an additional unit of a good or service. As individuals consume more of a particular good, the marginal utility tends to decrease. For instance, the first slice of pizza may provide high utility, but as more slices are consumed, the satisfaction from each additional slice diminishes.

This is also a reason for that utility function must be concave. By the marginal utility, we have if $w_1 < w_2$:

$$\begin{cases} U(w_1 + \Delta) - U(w_1) > U(w_2 + \Delta) - U(w_2) & \text{if } \Delta > 0 \\ U(w_1 + \Delta) - U(w_1) < U(w_2 + \Delta) - U(w_2) & \text{if } \Delta < 0 \end{cases}$$

In the both case :

$$\frac{U(w_1 + \Delta) - U(w_1)}{\Delta} > \frac{U(w_2 + \Delta) - U(w_2)}{\Delta}$$

or

$$U'(w_1) > U'(w_2)$$

This means the derivative function U' is strictly monotonically decreasing, then U must be concave (this is a property of concave function). If U is a concave function, then the Jensen's inequality can be applied :

$$U(E[X]) \geq E[U(X)]$$

where X is a random variable.

19.5   Utility Maximization

In economic theory, individuals are assumed to be *rational decision-makers* who aim to maximize their utility based on their preferences and budget constraints. Utility maximization involves allocating resources (income or wealth) to different goods and services in a way that maximizes overall satisfaction.

19.6   Expected utility

Given the probability distribution of possible outcomes $P(outcome)$ and the associated utility function of outcome $P(outcome)$. The expected utility is calculated by :

$$\sum_{outcome} P(outcome) \times U(outcome)$$

19.7   St. Petersburg paradox

The St. Petersburg paradox is a paradox related to probability and decision theory in economics. It consists of a lottery game modeled by a random variable whose mathematical expectation is infinite, but for which participants would only agree to pay a small amount of money to play. Different approaches have been proposed to resolve this paradox.

19.7.1 Announcement

The game : The player pays an initial value c to the bank (dealer). A coin is tossed heads or tails as long as it comes up *tails*. The game **ends** when first *head* appear and then the bank gives the payoff to the player. This gain is initially $2, doubled for each appearance of tail. Thus, the payout is $2 if *head* appear on the first toss, $4 if *head* appears on the second, $8 on the third, $16 on the fourth, etc. So if *head* first appears on the nth toss, the bank pays 2^n. A sample calculation for expected value for gain :

$$E = \frac{1}{2} \cdot 2 + \frac{1}{4} \cdot 4 + \frac{1}{8} \cdot 8 + \frac{1}{16} \cdot 16 + \ldots = 1 + 1 + 1 + 1 + \ldots = \infty$$

The paradox comes here, the player just put a finite amount of money but he can get infinity payoff.

The question : What is the player's initial bet so that the game is fair, i.e. so that the player's initial bet is equal to his expectation of winning, and that neither the bank nor the player are advantaged by this game ?

19.7.2 Solution by using utility function

The classical resolution of the paradox involved the explicit introduction of a utility function, an expected utility hypothesis, and the presumption of diminishing marginal utility of money. Let's suppose that we use the utility function $U(w) = \ln(w)$ (also known as *log utility*), where w is palyer's (gambler's) total wealth. Using log utility function means that if w_1, w_2 are big, hence the distance between $\ln(w_1) - \ln(w_2)$ is much less then $w_1 - w_2$. To find the bet that the player must put c, we need to resolve (for c):

$$\Delta E(U) = \sum_{k=1}^{+\infty} \frac{1}{2^k} \left[\ln \left(w + 2^k - c \right) - \ln(w) \right] = 0$$

19.7.3 Solution by finite amount of bank

If we assume that the bank only has a finite amount, the calculation of expected gain are the same as in announcement part, except that the series is no longer infinite. For example, if we assume that the bank has "only" 2^N euros, then it cannot pay more if heads appear after $N + 1$ rolls.

$$E = \sum_{k=1}^{N} \frac{1}{2^k} \cdot 2^k = N$$

So if the bank has $1050000, then the bet that the player must put is $c = \log_2^{1050000} \approx \$20$.

19.8 Hyperbolic absolute risk aversion (HARA)

$$U(W) = \frac{1-\gamma}{\gamma} \left(\frac{aW}{1-\gamma} + b \right)^{\gamma}$$

- Utility is linear (the risk neutral case) if $\gamma \to 1$

- Utility is quadratic (an implausible though very mathematically tractable case, with increasing absolute risk aversion) if $\gamma = 2$

- The exponential utility function, which has constant absolute risk aversion, occurs if $b = 1$ and $\gamma \to -\infty$.

- The power utility function occurs if $\gamma < 1$ and $a = 1 - \gamma$.

20 Risk aversion

Risk aversion refers to the tendency of individuals or investors to prefer lower-risk choices, even if the higher-risk choices offer potentially higher returns.

An investor with high risk aversion is someone who prioritizes capital preservation and is willing to accept potentially lower returns in exchange for a higher level of certainty and security.

20.1 Example

A person is given the choice between two scenarios: one with a guaranteed payoff, and one with a risky payoff with same average value. In the former scenario, the person receives \$50. In the uncertain scenario, a fair coin is flipped to decide whether the person receives \$100 or nothing. The expected payoff for both scenarios is \$50, meaning that an individual who was insensitive to risk would not care whether they took the guaranteed payment or the gamble. However, individuals may have different risk attitudes

- risk averse (or risk avoiding) : they would even accept a certain payoff of less than \$50 (for example, \$45) for a stable situation.

- risk neutral : if they are indifferent between the bet and a certain \$50 payment.

- risk loving (or risk seeking) : if they would accept the bet even when the guaranteed payment is more than \$50 (for example, \$55).

20.2 Certainty equivalent

The certainty equivalent (CE) is amount that would make indifferent between taking the guaranteed amount and the uncertain one. The CE is different for each person. Using the above example, then the certainty equivalent in three cases are :

- risk averse (or risk avoiding) : $45

- risk neutral : $50

- risk loving (or risk seeking) : $55.

The risk premium (as described in 27.3) in this case, is the difference between the expected value and the certainty equivalent :

$$ExpectedValue - CertaintyEquivalent$$

With the above example:

- For risk-averse individuals, risk premium is positive

- For risk-neutral persons it is zero

- For risk-loving individuals their risk premium is negative

Given a utility of money function U of an individual, then the CE is defined by :

$$U(CE) = \mathbb{E}[U(X)]$$

Then we can see three risk premium (positive, negative, zero) for these three type of peopel in the figure 6.8.

20.3   Utility of money

Respectively to three cases : risk-averse, risk-neutral, risk-loving, we can have three models about utility function for money. These utility functions have two key properties: **slope** and **concavity**.

20.3.1   Concavity

- For risk-averse individuals : the utility function is a concave function (to have the diminishing marginal utility, which mean if $x_3 - x_2 = x_2 - x_1 > 0$, then $U(x_3) - U(x_2) < U(x_2) - U(x_1)$.

- For risk-neutral persons : the utility function is a linear function (to have if $x_3 - x_2 = x_2 - x_1 > 0$, then $U(x_3) - U(x_2) = U(x_2) - U(x_1)$

- For risk-loving individuals : the utility function is a convex function (to have if $x_3 - x_2 = x_2 - x_1 > 0$, then $U(x_3) - U(x_2) > U(x_2) - U(x_1)$)

20.3.2   Upward slope

The upward slope control the intensity of risk-aversion or risk loving :

- Risk-averse individual : low slope

- Very risk loving individual : high slope

For risk-loving individuals, they would prefer a bet which is first-order stochastically dominant over an alternative bet, to have better reward.
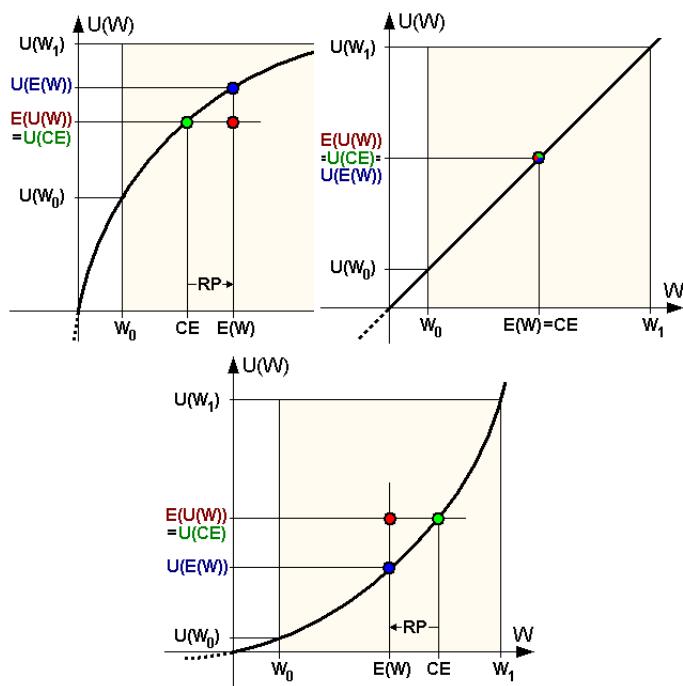
Figure 6.8: Risk premium for three types. Suppose that the wealth W is equally distributed from W_0 to W_1.

20.4   Measures of risk aversion under expected utility theory

Given an utility function, there are various measures of the risk aversion.

20.4.1   Absolute risk aversion

The higher the curvature of $u(c)$, the higher the risk aversion. However, since expected utility functions are not uniquely defined (can be equivalently obtained by a affine transformations), a measure that stays constant with respect to these transformations is needed rather than just the second derivative of $u(c)$. One such measure is the *absolute risk aversion* (ARA), after the economists Kenneth Arrow and John W. Pratt, also known as the *coefficient of absolute risk aversion*, defined as :

$$A(c) = -\frac{u''(c)}{u'(c)}$$

For example, if $u(c) = \alpha + \beta ln(c)$, so $A(c) = 1/c$.

A proof. Let π the risk premium then $\pi = E[X] - CE = \bar{X} - CE$, we have the definition of CE:

$$U(CE) = E[U(X)] \quad \text{or} \quad U(\bar{X} - \pi) = E[U(X)]$$

One one hand, using the first order approximation :

$$U(\bar{X} - \pi) \approx U(\bar{X}) - \pi U'(\bar{X})$$

On other hand, using the Taylor expansion for $U(X)$:

$$U(X) \approx U(\bar{X}) + (X - \bar{X})U'(\bar{X}) + \frac{1}{2}(X - \bar{X})^2 U''(\bar{X})$$

then we take the exptected value of two sides:

$$E[U(X)] \approx U(\bar{X}) + \frac{1}{2}Var[X]U''(\bar{X})$$

From both hands :

$$U(\bar{X}) - \pi U'(\bar{X}) = U(\bar{X}) + \frac{1}{2}Var[X]U''(\bar{X})$$

or

$$\pi = -\frac{1}{2}\frac{U''(\bar{X})}{U'(\bar{X})}Var[X]$$

Here we observe that the ralation between the risk premium π and the risk $Var[X]$. Suppose that 2 investors invest in the asset X, each investor have his own utility function U_1 and U_2. If the first investor is more risk-averse than the second, we must have $CE_1 < CE_2$ or $\pi_1 > \pi_2$, this implies that $-\frac{U_1''(\bar{X})}{U_1'(\bar{X})} > -\frac{U_2''(\bar{X})}{U_2'(\bar{X})}$. Then $-\frac{u''(c)}{u'(c)}$ reflect the risk-aversion.

20.4.2 Relative risk aversion

The relative risk aversion (RRA) or coefficient of relative risk aversion is defined as :

$$R(c) = cA(c) = -c\frac{u''(c)}{u'(c)}$$

20.5 Risk tolerance

The risk tolerance, noted $T(c)$ is the reciprocal of absolute risk aversion $A(c)$:

$$T(c) = \frac{1}{A(c)}$$

Unlike ARA whose units are in $\$^{-1}$, RRA is a dimensionless quantity, which allows it to be applied universally.

21 Intertemporal portfolio choice

Intertemporal portfolio choice refers to the process of determining how to allocate investment resources over time in order to optimize an investor's wealth or utility. It involves making decisions about the composition and allocation of assets across different periods, taking into account factors such as risk, return, and individual preferences for consumption or wealth accumulation.

Typically the criterion is the expected value of some concave function of the value of the portfolio after a certain number of time periods—that is, the expected utility of final wealth. Alternatively, it may be a function of the various levels of goods and services consumption that are attained by withdrawing some funds from the portfolio after each time period.

We discover in the section for the time discret case, for the continous time case, we will discover the Merton's portfolio problem in chapter 10. In time discret we have Time-independent and Time-dependent cases.

21.1 Time-independent decisions

We suppose that a periode of time can be divided into T time intervals, where the stochastic properties (return rates) in each interval is independent with others. Let W_0 be the initial wealth and let the *stochastic portfolio rate of growth* in interval t be R_t. Then the wealth after T intervals is:

$$W_T = W_0 R_1 R_2 \dots R_T$$

where R_t can be decomposed into the weighted sum of underlysing asset i:

$$R_t = w_{1t}r_{1t} + w_{2t}r_{2t} + \dots + w_{it}r_{it} + \dots + w_{nt}r_{nt}$$

where r_{it} refers to the **stochastic rate of growth** of underlysing asset i at interval t. The weights must satisfy $\sum_{i=1}^{n} w_{it} = 1$. Now associating the wealth with an

utility function, which means $U(W)$. Then the objective is to find w_{it} for each interval t to maximize the expected utility $\mathbb{E}[U(W_T)]$ or equivalently :

$$\max_{w_{it}} \mathbb{E}[U(W_0 R_1 R_2 \ldots R_T)]$$

21.1.1 Log utility

If we use the log utility, which means $U(W) = \ln(W)$:

$$\max_{w_{it}} \mathbb{E}[U(W_0 R_1 R_2 \ldots R_T)]$$
$$= \max_{w_{it}} \mathbb{E}[\ln(W_0 R_1 R_2 \ldots R_T)]$$
$$= \max_{w_{it}} \mathbb{E}[\ln(W_0) + \ln(R_1) + \ldots \ln(R_T)]$$
$$= \max_{w_{it}} \mathbb{E}[\ln(W_0)] + \mathbb{E}[\ln(R_1)] + \ldots \mathbb{E}[\ln(R_T)] \quad \text{Time (interval) independent}$$
$$= \max_{w_{it}} \ln(W_0) + \sum_{t=1}^{T} \mathbb{E}[\ln(w_{1t} r_{1t} + w_{2t} r_{2t} + \ldots + w_{it} r_{it} + \ldots + w_{nt} r_{nt})]$$

21.1.2 Power utility

If we use the log utility, which means $U(W) = aW^a$, where $0 < a < 1$:

$$\max_{w_{it}} \mathbb{E}[U(W_0 R_1 R_2 \ldots R_T)]$$
$$= \max_{w_{it}} aW_0^a \prod_{t=1}^{T} \mathbb{E}[(w_{1t} r_{1t} + w_{2t} r_{2t} + \ldots + w_{it} r_{it} + \ldots + w_{nt} r_{nt})^a]$$

21.2 Time-dependent decisions

As per the above, the expected utility of final wealth with a power utility function is

$$\max_{w_{it}} \mathbb{E}[U(W_0 R_1 R_2 \ldots R_T)]$$

The solution can be obtained by :

- **Dynamic programming**. We can view the above problem in more simply case : $\max_{w_{it}} \mathbb{E}[U(W_0 R_1 R_2)]$. Given a configuration of w_{i1} in R_1, then the optimal configuration w_{i2}^* that maximizes only R_2 is depend on w_{i1}. Here the dynamic programming is used to find the optimal configuration for $(\sum_{i=1}^{n} w_{i1} r_{i1})(\sum_{j=1}^{n} w_{j2} r_{j2})$. This method becomes complex very quickly if there are more than a few time periods T or more than a few assets n.

- **Dollar cost averaging**. Take $w_{i1} = w_{i2} = \ldots = w_{iT}, \forall i$. Indeed, Dollar cost averaging method means that the investor regularly invests a fixed amount of money into a particular asset or a fixed configuration of portfolio at regular intervals. Example, each month, the investor by a fixed \$500

amount of bitcoin. The basic idea behind dollar cost averaging is to reduce the impact of market volatility.

If there is not serial independence of rate of growth through time, then the expectations operator cannot be applied separately to the various multiplicative terms. Thus, the probability distribution of rate of growth for the various assets depends on their previous-period realizations, and so cannot be determined in advance. Not only depending on previous-period realizations, the optimal portfolio in a particular interval depends also the knowledge of how decisions will be made in future intervals.

Chapter 7

Analysis methods

Fundamental analysis and technical analysis are two different approaches used to analyze financial assets such as stocks, bonds, and currencies.

Fundamental analysis focuses on evaluating the intrinsic value of an asset by analyzing its underlying economic, financial, and qualitative factors such as the company's financial statements, industry trends, management team, and macroeconomic factors like interest rates, inflation, and GDP growth. The goal is to determine the asset's true worth, and whether it is undervalued or overvalued in the market. Fundamental analysts use this information to make long-term investment decisions based on the asset's fundamentals, looking to buy assets that are undervalued and sell them when they become overvalued.

On the other hand, *technical analysis* focuses on studying past market data, primarily charts, to identify patterns and trends that could indicate future price movements. Technical analysts use various tools and techniques such as moving averages, trend lines, and technical indicators to identify support and resistance levels, and potential entry and exit points. The goal is to identify short-term price movements and profit from them. Technical analysts assume that market movements are not random, and patterns in price and volume data can reveal valuable insights.

- Focus: Fundamental analysis focuses on the intrinsic value of a security by examining the underlying economic, financial, and industry factors that affect its value. Technical analysis, on the other hand, focuses on past market data, primarily price and volume, to identify patterns and trends that can be used to predict future price movements.

- Data Used: Fundamental analysis uses data such as financial statements, economic indicators, and industry trends to determine the value of a security. Technical analysis uses charts, graphs, and other visual representations of market data to identify patterns and trends.

- Timeframe: Fundamental analysis is typically a long-term approach that considers the overall health and prospects of a company or industry. Tech-

nical analysis is generally a short-term approach that focuses on identifying short-term trading opportunities.

- Approach to risk: Fundamental analysis is considered a lower-risk approach to investing because it involves a deeper understanding of the underlying factors that drive market trends. Technical analysis is considered a higher-risk approach because it relies on predicting future price movements based on past data.

- Application: Fundamental analysis is more commonly used by long-term investors who are interested in buying and holding securities for an extended period of time. Technical analysis is more commonly used by short-term traders who are looking to profit from short-term price movements.

In summary, fundamental analysis looks at the underlying factors that affect an asset's value, while technical analysis focuses on studying past market data to identify trends and patterns to predict future price movements. Both approaches have their advantages and limitations and are often used in combination to make investment decisions.

1   Technical analysis

Principles

- Supply & demand for a security determine its price.

- Change in supply & demand for a security causes a change in its price.

- Price can be projected by chart or other technical tools

Assumptions:

- There are rational and irrational market participants. Irrational means that it is led by emotion or animal spirit (tam ly dam dong).

- Technical analysis represents the study of collective investor' psychology or sentiment.

- Projecting the future price of a security by studying past investor' behavior

Technical analysis (TA) is widely used in the analysis of commodities, currencies and futures. This is because these assets don't have income flow (e.g dividend).

A problem may arise if the considered assets are low volume trading (illiquid). TA can be used on a stand-alone way or in combination with fundamental analysis. In both cases, TA is an indicator for optimal price-entry and exit points of a security.

Relative strength analysis compares a security's performance vs other securities in its own industry, or the market in general as a benchmark.

- Trend : Upward, downward, stepway. Retracement : ending of a trend, which can be trend reversal or sideway movement.

- Support and resistance. Support is a price level, which is expected to protect a security's price from going under (local minimum ?). On the other hand, resistance is a cap or ceiling in the price level of a security (local maximum).

- Change in Polarity. This is the idea that once a support level is breached, its becomes resistance level and vice versa.

2   Chart partterns

The art of trade is in reading charts correctly by spotting partterns that aren't obvious for all traders.

Two types of partterns considered:

- Reversal patterns :

    - Head and shoulder, neckline. Inverse head and shoulder.
    - Double/ Triple tops (and inverse)

- Continuations partterns : Ascending and descending triangle.

- Bullish rectangle, bearish rectangle

3   Indicators

Based on :

- Price and volume

- Momentum

- Market sentiment

- Non-priced-based mesures

3.1   Rate of change oscillator

$$ROC = \left(\frac{ClosePrice(p)}{ClosePrice(p-n)} - 1 \right) \times 100$$

Relation between ROC and price : When ROC switches from negative to positive, price moves upward and vice versa.

3.2 RSI Relative strength index

C : Price and i : time instant

- If $C_i > C_{i-1}$: $U_i = C_i - C_{i-1}$ and $D_i = 0$.

- If $C_i < C_{i-1}$: $U_i = 0$ and $D_i = C_{i-1} - C_i$.

$$RSI = 100 \frac{F(U)}{F(U) + F(D)}$$

where F is low-pass filter, usually smoothed or modified moving average (SMMA or MMA), or exponential moving average (EMA).

One reminds the property of RSI :

- If the market is steadily rising or sharply rising, the RSI tends towards 100.

- If the market is steadily falling or sharply falling, the RSI tends towards 0.

Using the following classical strategy to trade:

- IF CURRENT RSI ≤ 30 : BUY SIGNAL

- IF CURRENT RSI ≥ 70 : SELL SIGNAL

3.3 Larry Connors

- Sell if the price is above $SMA(200)$ and $RSI(2)$ is greater than 90.

- Buy if the price is below $SMA(200)$ and $RSI(2)$ is less than 10.

3.4 Bollinger Bands

:

$$SMA(20) + 2 \times \sigma$$

where σ is the std of last 20 signals.

3.5 Moving Average Convergence Divergence (MACD)

:

- $MACD = EMA(12) - EMA(26)$ (of prices)

- Signal $= EMA(9)$ (of $MACD$)

- Histogram $= MACD$ - Signal

Analysis : MACD using the second order (Fase Stochastic - Slow Stochastic) Relation with price :

- MACD crosses signal line from above : Sell signal

- MACD crosses signal line from below : Buy signal

3.6 Stochastic oscillator

$$\%K = 100 \times \frac{ClosePrice - L}{H - L}$$

where :

- L is lowest price of last N days

- H is highest price of last N days

The stochastic oscillator is very reactive and volatile. Unlike the RSI, it does not integrate an average, but directly price variations between the highest and lowest of the period considered. The indicator can go from 0 to 100 very quickly. Then we need a low-passed signal such as $\%D$ which is, let say SMA_N of $\%K$.

Relation with price:

- When $\%K$ crosses upward $\%D$ (vuot len tren) : Buy signal

- When $\%K$ crosses downward $\%D$: Sell signal

- When $\%K$ crosses upward 20 :Buy signal

- When $\%K$ crosses downward 80 :Sell signal

4 Non-price-based indicators

Such as unemployement, consumer confidence ...

4.1 Put Call ratio

The put-call ratio is calculated by dividing the number of traded put options (sell) by the number of traded call options (buy).

- $P = 1$ then bull = bear .

- $P > 0.7$ then bearish environment.

- $P < 0.5$ then bullish environment.

4.2 VIX

Measure volatility of options.
Formula from Wiki:

$$VIX = \sqrt{\frac{2e^{r\tau}}{\tau} \left(\int_0^F \frac{P(K)}{K^2} dK + \int_F^\infty \frac{C(K)}{K^2} dK \right)}$$

where:

- r is the risk-free rate.

- τ is the number of average days in a month (30 days).

- F is the 30-day forward price on the S&P 500.

- $P(K)$ and $C(K)$ are prices for puts and calls with strike K and 30 days to maturity.

Other formula:

$$\sigma^2 = \frac{2}{T} \sum_i \frac{\Delta K_i}{K_i^2} e^{RT} Q(K_i) - \frac{1}{T} \left[\frac{F}{K_0} - 1 \right]^2$$

where

σ	VIX index $= \sigma \times 100$	ΔK_i	Interval between strike prices – half the difference between the strike on either side of K_i:
T	Time to expiration (in years)		
F	Option-implied forward price		$\Delta K_i = \dfrac{K_{i+1} - K_{i-1}}{2}$
K_0 First strike equal to or otherwise immediately below the forward index level, F		R	Risk-free interest rate to expiration
K_i Strike price of the ith out-of-the-money option; a call if $K_i > K_0$ and a put if $K_i < K_0$; both put and call if $K_i = K_0$.		$Q(K_i)$	The midpoint of the bid-ask spread for each option with strike K_i.

Figure 7.1: VIX formula.

Realtion with price :

- High VIX is a signal for downward market

- Low VIX is a signal for upward market

4.3 Margin debt level

- Margin high $\rightarrow$ more lending : Bullish

- Margin low $\rightarrow$ les lending : Bearish

4.4 Mutual fund cash position

Observing the behavior of funds (mutual, pension) as a way to measure investor confidence.

- When cash is low : Bullish

- When cash is high : Bearish

4.5 New equity issuances

(Phat hanh co phieu moi)

- IPO : Initial public offering

- SEO : seasoned/secondary equity offering, also known as a follow-on offering, is a financial term used when a company offers additional shares of its stock to the public after its initial public offering (IPO). This allows the company to raise more capital from investors and increase its ownership base.

Inventment banks are interested in carrying out these equity offerings in periodes when the economy is stable => a sign to indicate that market is stable ?

Other points

- Price action: When a company issues new equity, the price of its existing shares may be affected. Technical analysts can use price charts to track how the market reacts to the news of the new equity issuance. For example, if the price of the stock falls after the issuance, it may indicate that the market perceives the new shares to dilute the value of the existing shares.

- Volume analysis: The volume of trading in a stock can be an important indicator of market sentiment. Technical analysts can track the volume of trading in a stock after a new equity issuance to determine whether there is increased demand for the shares or if investors are selling off existing shares.

- Support and resistance levels: Technical analysts can also use support and resistance levels to analyze the impact of new equity issuances on a stock's price. If a stock's price falls below a support level after a new equity issuance, it may indicate that investors are less confident in the company's future prospects. Conversely, if the price rises above a resistance level, it may indicate that the new equity issuance has boosted investor confidence.

4.6 TRIM or arms index

$$TRIN = \frac{AdvancingStocks/DecliningStocks}{AdvancingVolume/DecliningVolume}$$

where :

- Advancing Stocks : Number of stocks that are higher on the day

- Declining Stocks : Number of stocks that are lower on the day

- Advancing Volume : Total volume of all advancing stocks

- Declining Volume : Total volume of all declining stocks

Relation with market:

- $TRIN = 1$: Same advancing and declining

- $TRIN > 1$: More advancing $\rightarrow$ bullish

- $TRIN < 1$: More declining $\rightarrow$ bearish

5   Cycles

Expansion, contraction, peak, trough, trend.
Cycle periode or business cycle (by Nikolay Dmitrievich Kondratiev) :

- Some analysts consider the four year US presidential election cycle to be relevant.

- Other consider the 10, 18 or 54 year periode (called contra wave) to be of more relevant.
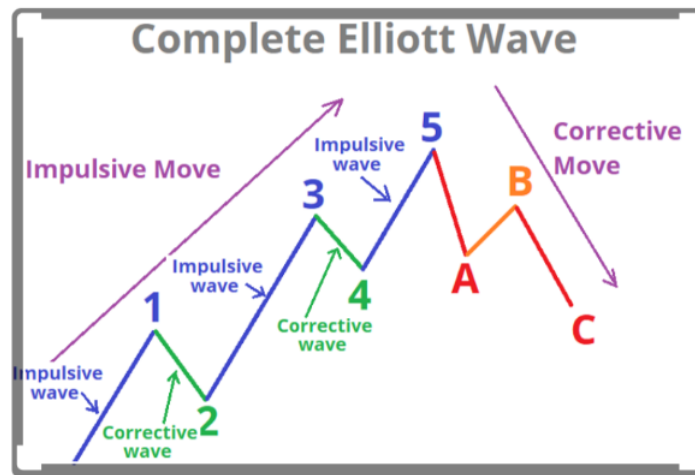
Elliott wave:

Figure 7.2: Elliott waves : In a bull market we can observe 5 waves pattern characterized by 3 distinct strong upward (impulse waves) with 2 retracement waves (correction waves) between 3 impulse waves. The 5 waves pattern is then followed by 3 waves : 2 corrective and one upward

- Grand super-cycle : Long periode of time (year or even centery).

- Sub-minuette cycle : Small cycles (minutes, hours)
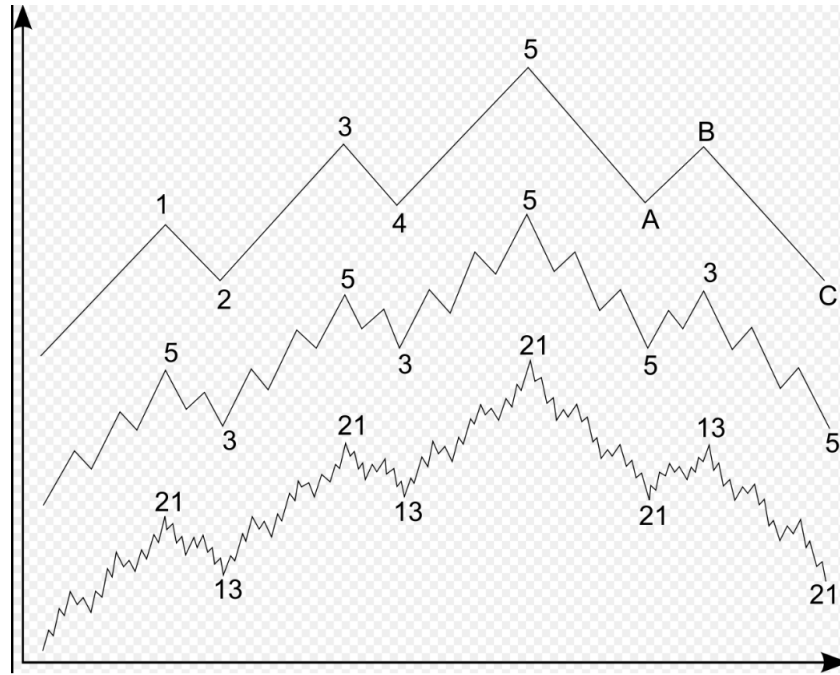
Other scenario of Elliott wave :

Figure 7.3: Other senario of Elliott wave.

Relation with Fibonaci sequence (0,1,1,2,3,5,8, ...) : The relationship between the Elliott Wave Principle and the Fibonacci sequence is that the waves identified in the Elliott Wave Principle tend to follow Fibonacci ratios. Specifically, the **lengths of the waves** tend to correspond to specific ratios of the preceding wave. The most commonly used Fibonacci ratios in Elliott Wave analysis are 0.618, 1.618 (golden rate), 2.618, 0.382, and 0.236. These ratios are derived from dividing one number in the Fibonacci sequence by another number in the sequence, and they are believed to represent key levels of support and resistance in financial markets.

Remind of Golden rate : $1 + \frac{1}{x} = x$ or $x = \frac{a+b}{a} = \frac{a}{b}$.

- $\frac{3}{2} = 1.5$; $\frac{5}{3} = 1.666$; $\frac{8}{5} = 1.6$; $\frac{13}{8} = 1.625 \rightarrow 1.618$.

- $\frac{2}{3} = 0.666$; $\frac{3}{5} = 0.6$; $\frac{5}{8} = 0.625 \rightarrow 0.618$

Chapter 8

Deterministic optimal control

1   Brachistochrone curve

This is a classical example for the optimal control.

1.1   Statement

The word *brachistochrone* designates a curve in a vertical plane on which a heavy material point placed in a uniform gravitational field, sliding **without friction and without initial speed**, presents a **minimum travel time** among all the curves joining two fixed points: we speak of problem of the brachistochrone curve (see figure 8.1).

Our challenge is to find out the form or equation that describe the curve corresponding to optimal time. The input is two points : initial point $(0,0)$ and terminal point (a,b) where $b < 0$.

1.2   Solution

Since the initial velocity is 0, let $h > 0$ be the altitude loss (ordinate) from point $(0,0)$ then the actual velocity is:
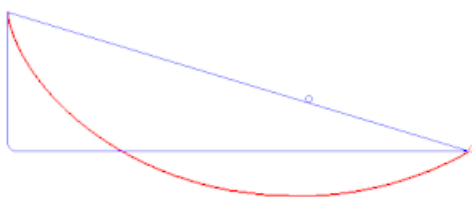
Figure 8.1: Brachistochrone illustration. The blue one is the minimum distance and the red one is minimum time.

$$v = \sqrt{2gh}$$

The law of refraction, according to Fermat's principle, indicates that, throughout its trajectory, a light ray obeys the rule :

$$\frac{\sin \theta}{v} = Constant$$

where θ represents the angle between the vertical axe and the direction of velocity.

An important property that if this condition is satisfied, then we have the optimal time. In figure 8.1, the blue trajectory have the constant θ and v is accelerated (hence not constant), then $\frac{\sin \theta}{v}$ is not constant, so this trajectory do not correspond to the optimal time.

Let $y < 0$ is the ordinate, in term of time, the optimal condition infers that :

$$\frac{\sin \theta}{\sqrt{-2gy}} = C$$

Let D be the altitude loss that the particle passes through the horizontal, then $\theta = \frac{\pi}{2}$ and $y = -D$

$$\frac{1}{\sqrt{2gD}} = C$$

From two above equations:

$$\frac{\sin \theta}{\sqrt{-2gy}} = \frac{1}{\sqrt{2gD}}$$

Since $\sin \theta = \frac{dx}{\sqrt{dx^2 + dy^2}}$ and $y' = \frac{dy}{dx}$:

$$(1 + y'^2)y = -D$$

1.3   Solving the differential equation

Suppose the D is known, then the solution for this well-known problem, by posing $y' = cotan(\frac{\alpha}{2})$:

$$\begin{cases} x(\alpha) = \frac{D}{2}(\alpha - \sin(\alpha)) + c \\ y(\alpha) = -\frac{D}{2}(1 - \cos(\alpha)) \end{cases}$$

At the initial point $y = 0$, then $\alpha = 0$, then $x(0) = 0$ infers that $c = 0$. Then we need to find D by using the terminal point (a, b) :

$$\begin{cases} a = \frac{D}{2}(\alpha - \sin(\alpha)) \\ b = -\frac{D}{2}(1 - \cos(\alpha)) \end{cases}$$

We can find first α and then D.

More solutions can be found in document "Curves & Cash: Finding the Optimal Path Using Control Theory".

2 Euler–Lagrange equation

Given the functional J (which take input a function and outputs a function) :

$$J(x) = \int_{t_0}^{t_1} \mathcal{L}\left(t, x(t), \dot{x}(t)\right) \, \mathrm{d}t$$

We wish to find a function $x(t)$ that **extremize** J, where :

- $x(t_0) = x_0$ and $x(t_1) = x_1$ fixed.

- $x : [t_0, t_1] \to E$ is of class $\mathcal{C}^1$, where E is a normed vector space.

- $\mathcal{L} : [t_0, t_1] \times E^2 \to \mathbb{R}$ is of class $\mathcal{C}^1$

$\mathcal{L}$ is also call Lagrange function.
Then a necessary condition (stationary point or Euler equation) for $x(t)$ is :

$$\frac{\partial \mathcal{L}}{\partial x} - \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial \mathcal{L}}{\partial \dot{x}}\right) = 0$$

2.1 Example 1

A standard example is finding the real-valued function $y(x)$ on the interval $[a, b]$, such that $y(a) = c$ and $y(b) = d$, for which the *path length along the curve traced by y is as short as possible* :

$$= \min_{y} \int_{a}^{b} \sqrt{\mathrm{d}x^2 + \mathrm{d}y^2}$$

$$= \min_{y} \int_{a}^{b} \sqrt{1 + y'^2}\,\mathrm{d}x$$

$$= \min_{y} \int_{a}^{b} \mathcal{L}(x, y, y')$$

Apply the Euler–Lagrange equation :

$$\frac{\partial \mathcal{L}}{\partial y} = \frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\partial \mathcal{L}}{\partial y'}\right)$$

$$\Leftrightarrow 0 = \frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{y'}{\sqrt{1 + y'^2}}\right)$$

$$\Leftrightarrow C = \frac{y'}{\sqrt{1 + y'^2}}$$

$$\Rightarrow y' = \frac{C}{1 - C^2} := A$$

$$\Leftrightarrow y(x) = Ax + B$$

It means is a straight line from $(a, y(a))$ to $(b, y(b))$.

2.2 Example 2

Euler–Lagrange equation is used in Horn–Schunck method for calculating the optical flow.

2.3 Example 3

We discover an other solution for Brachistochrone problem. Let $x(t)$ and $y(t)$ are abscissa and ordinate of the heavy material point. From the kinetic energy and the potential energy, we must have :

$$\frac{\dot{x}(t)^2 + \dot{y}(t)^2}{2} = gy(t)$$

$$\Leftrightarrow \frac{(dx)^2 + (dy)^2}{2(dt)^2} = gy$$

$$\Leftrightarrow dt = \sqrt{\frac{(dx)^2 + (dy)^2}{2gy}} = \frac{\sqrt{1 + \frac{(dy)^2}{(dx)^2}}}{\sqrt{2gy}} dx$$

We need to optimize the total time from point A to point B :

$$\int_A^B dt = \int_A^B \frac{\sqrt{1 + y'(x)}}{\sqrt{2gy}} dx = \int_A^B \mathcal{L}(x, y, y')$$

Euler-Lagrange equation gives that :

$$\frac{\partial \mathcal{L}}{\partial y} = \frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\partial \mathcal{L}}{\partial y'}\right)$$

After, we find the same equation as in section 1.

3 Hamiltonian

Consider a *dynamical system* or *dynamical constraint* :

$$\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}(t), \mathbf{u}(t), t)$$

where $\mathbf{x}$ denotes **state vector** (vertical) and $\mathbf{u}$ denotes **control vector**. The goal is to optimize a *performance index* $I(\mathbf{x}(t), \mathbf{u}(t), t)$ at each point time between t_0 and t_1, which means:

$$\max_{\mathbf{u}(t)} \mathbf{J} = \max_{\mathbf{u}(t)} \int_{t_0}^{t_1} I(\mathbf{x}(t), \mathbf{u}(t), t)\mathrm{d}t$$

We define the **Hamiltonian** :

$$H(\mathbf{x}(t), \mathbf{u}(t), \lambda(t), t) \equiv I(\mathbf{x}(t), \mathbf{u}(t), t) + \lambda^{\mathsf{T}}(t)\mathbf{f}(\mathbf{x}(t), \mathbf{u}(t), t)$$

where $\lambda(t)$, refers to **costate variables**, are functions of time rather than constants.

Then using *Pontryagin's maximum principle* says if we want to maximize **J**, the three functions $\mathbf{u}^*(t)$, $\mathbf{x}^*(t)$ and $\lambda^*(t)$ must satify :

$$H(\mathbf{x}^*(t), \mathbf{u}^*(t), \lambda^*(t), t) \geq H(\mathbf{x}(t), \mathbf{u}(t), \lambda(t), t), \quad \forall \mathbf{u}(t)$$

There are three equations that we can use to solve for $\mathbf{u}(t)$:

1. The first-order *necessary* condition for $\mathbf{u}(t)$:

$$\frac{\partial H(\mathbf{x}(t), \mathbf{u}(t), \lambda(t), t)}{\partial \mathbf{u}} = 0$$

2. The costate equation :

$$\frac{\partial H(\mathbf{x}(t), \mathbf{u}(t), \lambda(t), t)}{\partial \mathbf{x}} = -\dot{\lambda}^{\mathsf{T}}(t)$$

3. The dynamical constraint :

$$\frac{\partial H(\mathbf{x}(t), \mathbf{u}(t), \lambda(t), t)}{\partial \lambda} = \mathbf{f}(\mathbf{x}(t), \mathbf{u}(t), t) = \dot{\mathbf{x}}$$

All three equations form a system of equations, which is only necessary conditions for solving $\mathbf{u}(t)$. It depends on the system ($\mathbf{I}$ and f) that we can have solutions, it means that we can have cases that the system of equations have no root.

A **sufficient condition** for the maximum of **J** is the concavity of the Hamiltonian evaluated at the solution, i.e. :

$$H_{\mathbf{uu}}(\mathbf{x}^*(t), \mathbf{u}^*(t), \lambda^*(t), t) \leq 0$$

where $H_{\mathbf{uu}} = \frac{\partial^2 H}{\partial \mathbf{u}^2}$.

3.1 Derivation from the Lagrangian

Let's break the integral into a partition :

$$\int_{t_0}^{t_1} I(\mathbf{x}(t), \mathbf{u}(t), t) \mathrm{d}t = \sum_{n \to \infty} I(\mathbf{x}(t_{n,n+1}), \mathbf{u}(t_{n,n+1}), t_{n,n+1})(t_{n+1} - t_n)$$

where $t_n \leq t_{n,n+1} \leq t_{n+1}$.

Then we can maximize $I(\mathbf{x}(t_{n,n+1}), \mathbf{u}(t_{n,n+1}), t_{n,n+1})$ for **each time** $t_{n,n+1}$. Let $t_{n+1} - t_n$ is fixed, using $\lambda(t_{n,n+1})$ as Lagrangian multiplier, we need to maximize :

$$I(\mathbf{x}(t_{n,n+1}), \mathbf{u}(t_{n,n+1}), t_{n,n+1}) + \lambda(t_{n,n+1}) \left(\mathbf{f}(\mathbf{x}(t_{n,n+1}), \mathbf{u}(t_{n,n+1}), t_{n,n+1}) - \dot{\mathbf{x}}(t_{n,n+1})\right), \quad \forall t_{n,n+1}$$

Back to integral notation, it means we need to maximize :

$$\sum_{n\to\infty} \left[I(\mathbf{x}(t_{n,n+1}), \mathbf{u}(t_{n,n+1}), t_{n,n+1}) + \lambda(t_{n,n+1})\left(\mathbf{f}(\mathbf{x}(t_{n,n+1}), \mathbf{u}(t_{n,n+1}), t_{n,n+1}) - \dot{\mathbf{x}}(t_{n,n+1})\right)\right](t_{n+1} - t_n)$$

$$= \int_{t_0}^{t_1} \left(I(\mathbf{x}(t), \mathbf{u}(t), t) + \lambda(t)\left(\mathbf{f}(\mathbf{x}(t), \mathbf{u}(t), t) - \dot{\mathbf{x}}(t)\right)\right)\mathrm{d}t$$

$$= \int_{t_0}^{t_1} \left(I(\mathbf{x}(t), \mathbf{u}(t), t) + \lambda(t)\mathbf{f}(\mathbf{x}(t), \mathbf{u}(t), t)\right)\mathrm{d}t - \int_{t_0}^{t_1} \lambda(t)\dot{\mathbf{x}}(t)\mathrm{d}t$$

$$= \int_{t_0}^{t_1} \left(I(\mathbf{x}(t), \mathbf{u}(t), t) + \lambda(t)\mathbf{f}(\mathbf{x}(t), \mathbf{u}(t), t)\right)\mathrm{d}t + \int_{t_0}^{t_1} \dot{\lambda}(t)\mathbf{x}(t)\mathrm{d}t - \lambda(t)\mathbf{x}(t)\mid_{t_0}^{t_1}$$

$$= \int_{t_0}^{t_1} \left(I(\mathbf{x}(t), \mathbf{u}(t), t) + \lambda(t)\mathbf{f}(\mathbf{x}(t), \mathbf{u}(t), t) + \dot{\lambda}(t)\mathbf{x}(t)\right)\mathrm{d}t - \lambda(t_1)\mathbf{x}(t_1) + \lambda(t_0)\mathbf{x}(t_0)$$

$$= \sum_{n\to\infty}\left[I(\mathbf{x}(t_{n,n+1}), \mathbf{u}(t_{n,n+1}), t_{n,n+1}) + \lambda(t_{n,n+1})(\mathbf{f}(\mathbf{x}(t_{n,n+1}), \mathbf{u}(t_{n,n+1}), t_{n,n+1})) + \right.$$

$$\left. \dot{\lambda}(t_{n,n+1})\mathbf{x}(t_{n,n+1})\right] - \lambda(t_1)\mathbf{x}(t_1) + \lambda(t_0)\mathbf{x}(t_0)$$

where we did use the integrating by parts (to remove $\dot{\mathbf{x}}(t)$) and rebreak the integral into a partition. We define the extended Hamiltonian (which is inside of the above integral) :

$$\tilde{H}(\mathbf{x}(t), \mathbf{u}(t), \lambda(t), t) \equiv I(\mathbf{x}(t), \mathbf{u}(t), t) + \lambda(t)\mathbf{f}(\mathbf{x}(t), \mathbf{u}(t), t) + \dot{\lambda}(t)\mathbf{x}(t)$$

$$= H(\mathbf{x}(t), \mathbf{u}(t), \lambda(t), t) + \dot{\lambda}(t)\mathbf{x}(t)$$

To maximize each $\tilde{H}(\mathbf{x}(t_{n,n+1}), \mathbf{u}(t_{n,n+1}), \lambda(t_{n,n+1}), t_{n,n+1})$ at each instant $t_{n,n+1}$, or in general t, we must have the following necessary condtions:

- $\frac{\partial \tilde{H}(\mathbf{x}(t), \mathbf{u}(t), \lambda(t), t)}{\partial \mathbf{u}} = 0 \Rightarrow \frac{\partial H(\mathbf{x}(t), \mathbf{u}(t), \lambda(t), t)}{\partial \mathbf{u}} = 0$

- $\frac{\partial \tilde{H}(\mathbf{x}(t), \mathbf{u}(t), \lambda(t), t)}{\partial \mathbf{x}} = 0 \Rightarrow \frac{\partial H(\mathbf{x}(t), \mathbf{u}(t), \lambda(t), t)}{\partial \mathbf{x}} = -\dot{\lambda}(t)$

Then we find the same conditions as above section. We do not write the $\frac{\partial \tilde{H}}{\partial \lambda}$ since it is complicated as the present of $\dot{\lambda}$, but it will lead to the dynamical constraint.

Transversality condition : This is a boundary condition for the terminal values of the costate variables (here λ). It is one of the necessary conditions for optimality infinite-horizon optimal control problems without an endpoint constraint on the state variables.

If the terminal value of state variable (here $\mathbf{x}$) is free, as is often the case, the additional condition $\lambda(t_1) = 0$ or $\lambda(t_1)x(t_1) = 0$ is necessary for optimality.

We will find the application of Hamiltonian in the Ramsey–Cass–Koopmans model.

3.2 Relation with Euler-Lagrange equation

In case that if we can express $\mathbf{u}(t)$ in term of $\dot{\mathbf{x}}(t)$, .e.g.:

$$\mathbf{u}(t) = \mathbf{g}(\dot{\mathbf{x}}(t))$$

Then there is an equivalence between Euler-Lagrange equation and the necessary conditions related to Hamiltonian. Let's take $\mathbf{u}(t) = \dot{\mathbf{x}}(t)$ ($f(\mathbf{u}(t)) = \mathbf{u}(t)$ is then the indentical function), using the two first necessary conditions:

- $\frac{\partial H}{\partial \mathbf{u}} = 0 \Leftrightarrow \frac{\partial I}{\partial \mathbf{u}} + \lambda = 0 \Leftrightarrow \frac{\partial I}{\partial \mathbf{u}} = -\lambda$ or $\frac{\partial I}{\partial \dot{\mathbf{x}}} = -\lambda$

- $\frac{\partial H}{\partial \mathbf{x}} = -\dot{\lambda} \Leftrightarrow \frac{\partial I}{\partial \mathbf{x}} + \lambda \frac{\partial f}{\partial \mathbf{x}} = -\dot{\lambda} \Leftrightarrow \frac{\partial I}{\partial \mathbf{x}} = -\dot{\lambda}$

From the above two equation, we have Euler-Lagrange equation :

$$\frac{\partial I}{\partial \mathbf{x}} = \frac{\mathrm{d}}{\mathrm{d}t} \frac{\partial I}{\partial \dot{\mathbf{x}}}$$

3.3 Example of Hamiltonian: Ramsey–Cass–Koopmans model

This is a neoclassical model of economic growth.

3.3.1 Statement

By assumption, the only productive factors are capital :

- The capital $K(t)$

- The labour force $L(t)$, can be seen as the population. L is assumed to grow at a constant rate n, i.e. $\dot{L} = \frac{dL}{dt} = nL$, implying that $L(t) = L_0 e^{nt}$.

Next, we suppose that :

- $C(t)$ denote aggregate (total) consumption at instant t. Then the average consumption (per capita or per labour) is $c(t) = \frac{C(t)}{L(t)}$.

- $Y(t)$ denote aggregate (total) production at instant t. Then the average production is $y(t) = \frac{Y(t)}{L(t)}$.

- $k(t) = \frac{K(t)}{L(t)}$ is the capital per person (also called capital intensity).

- Depreciation rate δ (constant) for the decay of capital.

Then the captital at time instant $t + \Delta_t$ (in small Δ_t, we suppose $Y(t), C(t)$ constant) is calculated from time instant t by :

$$K(t + \Delta_t) = K(t) + (Y(t) - \delta K(t) - C(t)) \Delta_t$$

Or

$$\lim_{\Delta_t \to 0} \frac{K(t + \Delta_t) - K(t)}{\Delta_t} = \dot{K}(t) = Y(t) - \delta K(t) - C(t)$$

Or in *per capita* :
$$\frac{\dot{K}(t)}{L(t)} = \frac{Y(t)}{L(t)} - \delta\frac{K(t)}{L(t)} - \frac{C(t)}{L(t)}$$

Or
$$\frac{\dot{K}(t)L(t) - \dot{L}(t)K(t)}{L^2(t)} + \frac{\dot{L}(t)K(t)}{L^2(t)} = \frac{Y(t)}{L(t)} - \delta k(t) - c(t)$$

Or
$$\dot{k}(t) = y(t) - (\delta + n)k(t) - c(t)$$

Thus, the average production $y(t)$ can be expressed as a function of capital intensity $k(t)$ such as $y(t) = f(k(t))$, e.g., $y(t) = k(t)^{0.5}$. In particular, a common choice for f is Cobb–Douglas production function. However, we can use any production function $f : \mathbb{R}^n_+ \to \mathbb{R}_+$ satisfying the *Inada conditions* :

- $f(0_n) = 0$

- The function f is concave, i.e. the Hessian matrix of f is negative-semi definite.

- $\lim_{x_i \to 0} \frac{\partial f(\mathbf{x})}{\partial x_i} = +\infty$

- $\lim_{x_i \to +\infty} \frac{\partial f(\mathbf{x})}{\partial x_i} = 0$

Finally, we have :
$$\dot{k}(t) = f(k(t)) - (\delta + n)k(t) - c(t)$$

Next, we discover the objective function, which tries to maximize the utility U of the consumption $C(t)$
$$\int_0^\infty e^{-\rho t} U(C(t))dt$$

which is also called social welfare function. Note that $U(C(t)) = U(L(t)c(t)) = L(t)u(c(t))$, where u is a utility function that takes $c(t)$ as input. Then maximizing the above objective function is equivalent to :
$$\int_0^\infty e^{-(\rho - n)t} u(c(t))dt$$

3.3.2 Solution

In resume, we want to
$$\max_c \int_0^\infty e^{-(\rho - n)t} u(c)dt$$

where $\rho > n$ to ensure the limite of integral. Subject to:
$$c(t) = f(k(t)) - (\delta + n)k(t) - \dot{k}(t)$$

or for short :

$$c = f(k) - (\delta + n)k - \dot{k}$$

Then the objective of Ramsey–Cass–Koopmans model is **to find the optimal comsumption** $c(t)$.

Hamiltonian application. Let's define the Hamiltonian function :

$$H(k(t), c(t), \lambda(t)) = e^{-(\rho-n)t}u(c) + \lambda\left(f(k) - (\delta + n)k - c\right)$$

Then we need :

- $\frac{\partial H}{\partial c} = 0$:

$$e^{-(\rho-n)t}u'(c) - \lambda = 0$$

- $\frac{\partial H}{\partial \lambda} = \dot{k}$:

$$f(k) - (\delta + n)k - c = \dot{k}$$

- $\frac{\partial H}{\partial k} = -\dot{\lambda}(t)$:

$$\lambda\left(f'(k) - \delta - n\right) = -\lambda'$$

Note that $f'(k)$ is not constant. If we take the derivative (in term of t) of the first condition :

$$-(\rho - n)e^{-(\rho-n)t}u'(c) + e^{-(\rho-n)t}u''(c)c'(t) - \lambda'(t) = 0$$

Then using the third equation:

$$e^{-(\rho-n)t}u'(c)\left(f'(k) - \delta - n\right) = (\rho - n)e^{-(\rho-n)t}u'(c) - e^{-(\rho-n)t}u''(c)c'(t)$$
$$\Leftrightarrow u'(c)\left(f'(k) - \delta - n\right) = (\rho - n)u'(c) - u''(c)c'(t)$$
$$\Leftrightarrow u''(c)c'(t) = \left(-f'(k) + \delta + \rho\right)u'(c)$$
$$\Leftrightarrow c'(t) = \left(f'(k) - \delta - \rho\right)c\left(-\frac{u'(c)}{cu''(c)}\right)$$
$$\Leftrightarrow c'(t) = \sigma(c)\left(f'(k) - \delta - \rho\right)c$$

where $\sigma(c) = -\frac{u'(c)}{cu''(c)} = -\frac{d\ln c}{d\ln(u'(c))}$ is called *elasticity of intertemporal substitution* (which is the inverse of relative risk aversion in 20.4.2). In some litterature, $f'(k)$ can be noted as $f_k(k)$.

The above equation and the second equation form a sytem of differential equation:

$$\begin{cases} \dot{k} = f(k) - (n + \delta)k - c = g_1(k, c) \\ \dot{c} = \sigma(c)\left(f'(k) - \delta - \rho\right)c = g_2(k, c) \end{cases}$$

is called *Ramsey–Cass–Koopmans dynamical system*. If using matrix notation :

$$\begin{bmatrix} \dot{k} \\ \dot{c} \end{bmatrix} = \begin{bmatrix} -(n + \delta) & -1 \\ 0 & \sigma(c)\left(f'(k) - \delta - \rho\right) \end{bmatrix} \begin{bmatrix} k \\ c \end{bmatrix} + \begin{bmatrix} f(k) \\ 0 \end{bmatrix}$$

Unfortunately, this system is **not always linear** since there are $\sigma(c)$, $f'(k)$ and $f(k)$. If the utility function $u(c) = \ln(c)$, then $\sigma(c) = 1$.

Study of steady state : where $\dot{k} = 0$ and $\dot{c} = 0$, we have three solutions (k_0, c_0):

1. $\begin{cases} f'(k_0) = \delta + \rho \\ c_0 = f(k_0) - (n + \delta)k_0 \end{cases}$

2. $\begin{cases} c_0 = 0 \\ k_0 = 0 \end{cases}$

3. $\begin{cases} c_0 = 0 \\ f(k_0) = (n + \delta)k_0, k_0 \neq 0 \end{cases}$

The second point is trivial because $f(k) = 0$ (by Inada condition) but the third point is more complicated. We need to show that there exists $k_0 \neq 0$ such that $f(k_0) = (n + \delta)k_0$.
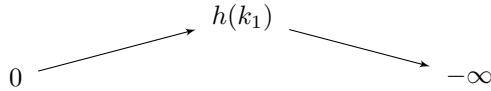
From $\lim_{k \to 0^+} f'(k) = +\infty$ and $\lim_{k \to +\infty} f'(k) = 0$ (Inada conditions), then by the *intermediate value theorem*, there exists k_1 such that $f'(k_1) = n + \delta$ or k_1 is a critical point of function $h(k) = f(k) - (n + \delta)k$. We have :

- $f(k)$ is concave and $(n + \delta)k$ is linear, then $h(k)$ is concave (by using the definition of concave), then $h''(k) < 0$. This infers that h has only critical point k_1.

- $h'(k) > 0, \forall k \in [0, k_1)$ and $h'(k) < 0, \forall k \in (k_1, +\infty)$ since $\lim_{k \to 0^+} f'(k) = +\infty$ and $\lim_{k \to +\infty} f'(k) = 0$.

- $\lim_{k \to +\infty} h(k) = -\infty$. Proof : Since $\lim_{k \to +\infty} h'(k) = -(n + \delta) < 0$, then there exists m and $r < 0$ such that $\forall k > m$, $h'(k) < r$. Using the *mean value theorem*, there exists $c \in (m, k)$ such that

$$\frac{h(k) - h(m)}{k - m} = h'(c) < r$$

or $h(k) < h(m) + r(k - m)$. As m is fixed, then if $k \to +\infty$ then $h(k) \to -\infty$

From all above statements, we have the variation table :

k	0		k_1		$+\infty$
$h''(k)$			$-$		
$h'(k)$		$+$	0	$-$	
$h(k)$					

This shows that there exists uniquely $k_0 > k_1 > 0$ such that $h(k_0) = 0$ or $f(k_0) = (n + \delta)k_0$.

Due to the non-linearity, we study the property of steady state by appoximating (or linearizing) $g_1(k, c)$ and $g_2(k, c)$. Using Taylor series (up to 1 order) and by using Jacobian matrix (for first derivative) :

$$\begin{bmatrix} g_1(k, c) \\ g_2(k, c) \end{bmatrix} = \mathbf{J}(k_0, x_0) \begin{bmatrix} k - k_0 \\ c - c_0 \end{bmatrix}$$

$$\begin{bmatrix} g_1(k, c) \\ g_2(k, c) \end{bmatrix} = \begin{bmatrix} \frac{\partial g_1(k_0, c_0)}{\partial k} & \frac{\partial g_1(k_0, c_0)}{\partial c} \\ \frac{\partial g_2(k_0, c_0)}{\partial k} & \frac{\partial g_2(k_0, c_0)}{\partial k} \end{bmatrix} \begin{bmatrix} k - k_0 \\ c - c_0 \end{bmatrix}$$

$$\begin{bmatrix} g_1(k, c) \\ g_2(k, c) \end{bmatrix} = \begin{bmatrix} f'(k_0) - (n + \delta) & -1 \\ \sigma(c_0)f''(k_0)c_0 & (f'(k_0) - \delta - \rho)(\sigma'(c_0)c_0 + \sigma(c_0)) \end{bmatrix} \begin{bmatrix} k - k_0 \\ c - c_0 \end{bmatrix}$$

Also, we check the concavity of the Hamiltonian with respect to control vector c (the sufficient condition) :

$$H_{cc}(k(t), c(t), \lambda(t)) = e^{-(\rho - n)t} u''(c(t)),$$

which is always negative since $u''(c) < 0, \forall c$. *This shows that each path $c(t)$ is optimal path for the maximization problem.*

Phase portrait study

We first draw isolines. The isoline here is the line that $k' = 0$ or $c' = 0$, in figure 8.2.

For the first case, $\begin{cases} f'(k_0) = \delta + \rho \\ c_0 = f(k_0) - (n + \delta)k_0 \end{cases}$

$$\mathbf{J} = \begin{bmatrix} \rho - n & -1 \\ \sigma(c_0)f''(k_0)c_0 & 0 \end{bmatrix}$$

$\det(\mathbf{J}) = -\sigma(c_0)f''(k_0)c_0 < 0$ since :

- $c_0 > 0$

- $\sigma(c_0) = -\frac{u'(c_0)}{cu''(c_0)} > 0$ since $u'(c_0) > 0$ and $u''(c_0) < 0$.

- $f''(k_0) < 0$ since f is concave (such as $-x^2$) by Inada condition.
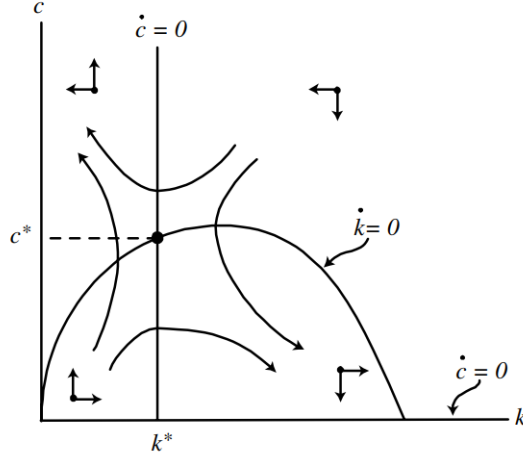
356

Figure 8.2: Isolines and some representative trajectories. We draw orthogonal arrows when passing isoline. The orthogonal arrows show the evoluation of k and c when time inscreases.

Since $\det(\mathbf{J}) < 0$, the first steady state is a **saddle point and instable** (Liapounov, since there is at least one eigenvalue is positive). See more the phase portrait of a linear differential system in stability theory https://en.wikipedia.org/wiki/Stability_theory. This is the only interior steady state, and for this point, we have two arms (figure 8.3), one is stable (eigenvalue is negative), one is unstable (eigenvalue is positive).

For the second case, $\begin{cases} c_0 = 0 \\ k_0 = 0 \end{cases}$

$$\mathbf{J} = \begin{bmatrix} f'(k_0) - (n + \delta) & -1 \\ 0 & (f'(k_0) - \delta - \rho)\,\sigma(c_0) \end{bmatrix}$$

Since $f'(k_0) \to +\infty$, then $\det(\mathbf{J}) > 0, \operatorname{tr}(\mathbf{J}) > 0$. We have also $\Delta = \operatorname{tr}(\mathbf{J})^2 - 4\det(\mathbf{J}) > 0$ because $(a + b)^2 \geq 4ab$, the equality happens when $a = b$, which means $f'(k_0) - (n+\delta) = (f'(k_0) - \delta - \rho)\,\sigma(c_0)$ or $f'(k_0)(1-\sigma(c_0)) = n-\rho$, but this can not be possible since $f'(k_0) \to +\infty$ and $\rho > n$. Then this steady state is a **source** that pushes.

For the third case, $\begin{cases} c_0 = 0 \\ f(k_0) = (n + \delta)k_0, k_0 > 0 \end{cases}$

$$\mathbf{J} = \begin{bmatrix} f'(k_0) - (n + \delta) & -1 \\ 0 & (f'(k_0) - \delta - \rho)\,\sigma(c_0) \end{bmatrix}$$

Using the above variation table, we see that $h'(k_0) = f'(k_0) - (n + \delta) < 0$. This infers $f'(k_0) - \delta - \rho < 0$ since $\rho > n$. Therefore, $\det(\mathbf{J}) > 0$ and $\operatorname{tr}(\mathbf{J}) < 0$. Then we can have a **sink** or a **degenerate sink** if $f'(k_0)(1 - \sigma(c_0)) = n - \rho$.
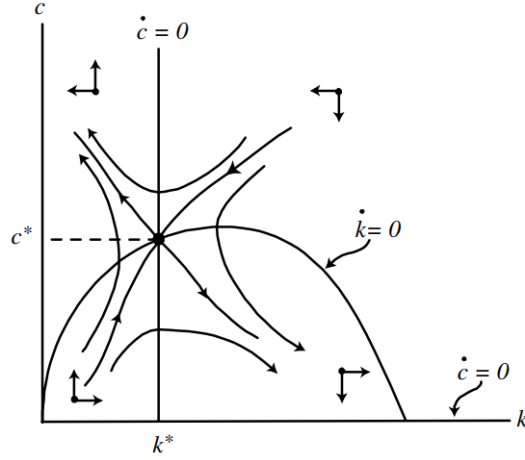
Figure 8.3: Arms (trajectories connected to an interior steady state), normally, there are two arms for an interior steady state.

We see these three steady state in the phase portrait (figure 8.4). The first point is in the top, the second point is the origin and the third point is on the abscissa. In case of we can not find the explicit solution, the phase portrait can show the evolution of a solution.

Consider some following representative trajectories :

- *Profligacy* : $c(t)$ increases and $k(t)$ decreases, on the left of isoline $c' = 0$, except the upward stable arm.

- *Wasteful over-accumulation or all saving* : converging to the third steady state $(k_0, 0)$ where $f(k_0) = (n + \delta)k_0$.

- Stable arms to interior steady state.

For the profligate trajectories, from the first differential equation $\dot{k} = f(k) - (n + \delta)k - c$. For each k, we define c_k such that $f(k) - (n + \delta)k - c_k = 0$, then $\dot{k} = f(k) - (n+\delta)k - c_k - (c - c_k)$. As $c(t)$ increases, then there exists a moment from that $k(t)' < 0$ (upper the isoline $k' = 0$). Then $k(t)$ go to negative, which is not feasible.

For the all saving trajectories, we remind the **Transversality condition** :

$$\lim_{t \to +\infty} k(t)\lambda(t) = 0$$

and combining with the first-order condition of Hamiltonian, we must have :

$$\lim_{t \to +\infty} k(t)e^{-(\rho - n)t}u_c(c(t)) = 0$$

358

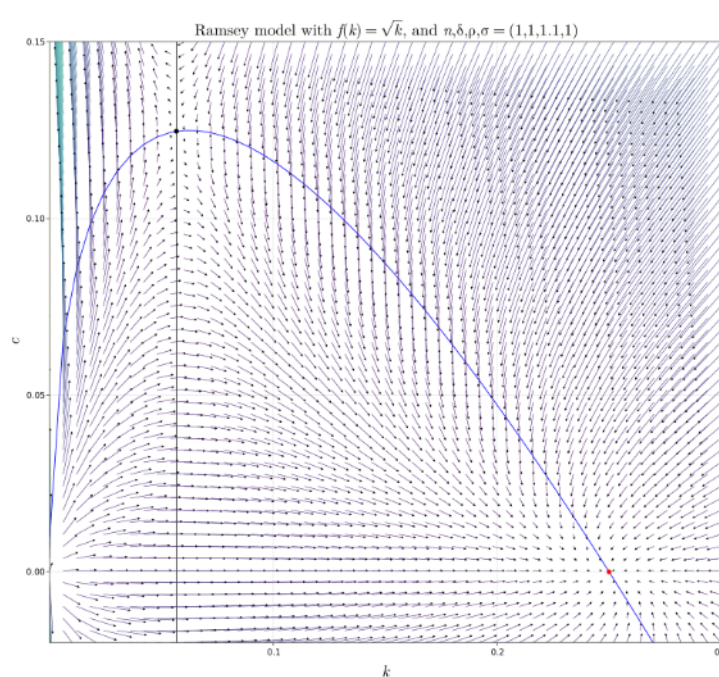Figure 8.4: Phase portrait of Ramsey Cass Koopmans model. There are three steady states (points). Abscissa is k and ordinate is c. The arrows show the optimal trajectorys of $k(t), c(t)$ when $t \to +\infty$. Each optimal trajectory is defini by initial value $k(t_0), c(t_0)$. Note that if $k(t_0) = k$ is given, then we need to choose the control variable $c(t_0)$ to have the optimal trajectory.

Remind an equation that we have from above :

$$u''(c)c'(t) = (-f'(k) + \delta + \rho)\,u'(c)$$

$$\Leftrightarrow \frac{u''(c)c'(t)}{u'(c)} = (-f'(k) + \delta + \rho)$$

$$\Leftrightarrow \frac{\partial \ln(u'(c))}{\partial t} = (-f'(k) + \delta + \rho)$$

or

$$u'(c) = \exp\left(\int_0^t (-f'(k) + \delta + \rho)ds + u'(c(0))\right)$$

Since $k(t) \to k_0$ **of third steady state**, then we have :

$$\lim_{t \to +\infty} k(t)e^{-(\rho-n)t}u_c(c(t))$$

$$= \lim_{t \to +\infty} k_0 e^{-(\rho-n)t}e^{\left((-f'(k_0)+\delta+\rho)t+u'(c(0))\right)}$$

$$= \lim_{t \to +\infty} bk_0 e^{(-f'(k_0)+\delta+n)t}$$

As k_0 in this case is on the right of k_1 $(k_0 > k_1)$ from the above demonstration and f' decreases beyond of k_1, then we must have $-f'(k_0) + \delta + n > 0$. This shows that $\lim\limits_{t \to +\infty} k(t)\lambda(t) \to +\infty$, equivalently, all saving trajectories violate the transversality condition and they are not optimal trajectories.

Finally, there is only trajectories **stable arms to interior steady state** are optimal.

Note that, steady states do not mean equilibrium. For more detailed, read *Lecture Notes on The Optimal Growth Problem* of Todd Keister at https://mpra.ub.uni-muenchen.de/1461/1/MPRA_paper_1461.pdf.

4 Linear–quadratic regulator

The case where :

- System dynamics are described by a set of **linear** differential equations.

- The cost is described by a **quadratic** function.

For horizon, we have finite and infinite case. For time, we have continuous and discrete time. So we have 4 combinations, they are lightly differents about the cost function, the boundary condition, the paramter matrix Q, R.... We discover here the continous time, combined with finite and infinite horizon. However, let discover a indispensable tool for Linear–quadratic regulator (LQR), the Riccati differential equation.

4.1 Riccati equation

In the narrowest sense, it means that any *first-order* (y') *ordinary differential equation* (y' in term of only x) is *quadratic* in the unknown function :

$$y'(x) = q_0(x) + q_1(x)\, y(x) + q_2(x)\, y^2(x)$$

To solve this problem, we can first convert it to a *second-order ordinary differential equation* :

$$u'' - Ru' + Su = 0.$$

where $y = -u'/(q_2 u)$, $R = q_1 + \dfrac{q_2'}{q_2}$ and $S = q_2 q_0$. Then we solve it as a standard second-order differential equation.

4.2 Finite-horizon, continuous-time

4.2.1 Announcement

Cost function to minimize or we need to find $u(t)$ that minimize :

$$J = x^T(t_1)F(t_1)x(t_1) + \int_{t_0}^{t_1} \left(x^T Q x + u^T R u + 2x^T N u \right) dt$$

Dynamic constraint :

$$\dot{x} = Ax + Bu$$

4.2.2 Solution

$$u = -Kx,$$

where K is given by:

$$K = R^{-1}(B^T P(t) + N^T),$$

and P is found by solving the continuous time Riccati differential equation:

$$A^T P(t) + P(t)A - (P(t)B + N)R^{-1}(B^T P(t) + N^T) + Q = -\dot{P}(t),$$

with the boundary condition:

$$P(t_1) = F(t_1).$$

4.3 Finite-horizon, continuous-time

4.3.1 Announcement

Cost function to minimize is lightly different :

$$J = \int_0^{\infty} \left(x^T Q x + u^T R u + 2x^T N u \right) dt$$

But the dynamic constraint is the same :

$$\dot{x} = Ax + Bu$$

4.3.2 Solution

The solution is the same above, except that

$$A^T P + PA - (PB + N)R^{-1}(B^T P + N^T) + Q = 0,$$

in this case, we have **no longer** the Riccati differential equation but the algebraic Riccati equation.

4.4 Note

In linear–quadratic regulator, we can have **explicite form** of $u(t)$ and $x(t)$ since the cost function and the dynamic constraint is more simple, for example, compared to Ramsey–Cass–Koopmans model (or optimal growth model). Remind that in the latter, we need the phase portrait (evaluated at each point) to know approximately the evolution of $u(t)$ and $x(t)$.

Futhermore, LQR is a particular case. If we apply the Hamiltonian, we refind the same way for the solution.

5 Hamilton–Jacobi–Bellman equation

The Hamilton-Jacobi-Bellman (HJB) equation is a *differential equation* that provides **necessary and sufficient** conditions for (deterministic) optimality of a control with respect to a loss function. Furthermore, it can be generalized to stochastic systems.

5.1 Annoucement

Consider the following problem in deterministic optimal control over the time period $[t_0, t_1]$:

$$V(x(t_0), t_0) = \min_u \int_{t_0}^{t_1} I(x(t), u(t), t)\, \mathrm{d}t + \phi(x(t_1))$$

subject to (dynamic system):

$$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = f(x(t), u(t), t)$$

5.2 Value function

To applying HJB equation, it is necessary to define the **value function**.

In general, the value function is a function that associates to an optimization problem. This function gives the value of objective function, evaluated at the solution. Hence, the value function depends only on the parameters of the problem.

In particular, for the problem of optimal control, the value function represents the optimal cost over the interval $[t, t_1]$ when started at the time t. For example, the value function of the above problem (in annoucement) is :

$$V(x(t), t) = \min_u \int_t^{t_1} I(x(s), u(s), s) \, \mathrm{d}s + \phi(x(t_1))$$

The value function can be interpreted as the cost to finish the process, and is thus also referred as "cost-to-go function".

5.3 HJB differential equation

For each instant t, we have to solve the following PDE :

$$\min_{u(t)} \left\{ I(x(t), u(t), t + \frac{\partial V(x(t), t)}{\partial x} f(x(t), u(t), t)) \right\} + \frac{\partial V(x(t), t)}{\partial t} = 0$$

subject to the terminal condition :

$$V(x(t_1), t_1) = \phi(x(t_1))$$

5.4 Deriving the equation

By Bellman's principle of optimality (that we may apply a lot in dynamic programming but do not know the name), given a small $\mathrm{d}t$ we must have :

$$V(x(t), t) = \min_{u(s), s \in [t, t+\mathrm{d}t]} \int_t^{t+\mathrm{d}t} I(x(s), u(s), s) \, \mathrm{d}s + V(x(t+\mathrm{d}t), t+\mathrm{d}t)$$
$$= \min_{u(t)} I(x(t), u(t), t)\mathrm{d}t + V(x(t+\mathrm{d}t), t+\mathrm{d}t)$$

Then by Taylor expansion :

$$V(x(t+\mathrm{d}t), t+\mathrm{d}t) = V(x(t), t) + \frac{\partial V(x(t), t)}{\partial x} \dot{x}(t)\mathrm{d}t + \frac{\partial V(x(t), t)}{\partial t}\mathrm{d}t + o(\mathrm{d}t)$$
$$= V(x(t), t) + \frac{\partial V(x(t), t)}{\partial x} \min_{u(t)} f(x(t), u(t), t)\mathrm{d}t + \frac{\partial V(x(t), t)}{\partial t}\mathrm{d}t + o(\mathrm{d}t)$$

By replacing, we refind HJB differential equation :

$$\min_{u(t)} \left\{ I(x(t), u(t), t) + \frac{\partial V(x(t), t)}{\partial x} f(x(t), u(t), t) \right\} + \frac{\partial V(x(t), t)}{\partial t} = 0$$

5.5 Relation to Hamiltonian

If using Hamiltonian for the value function, we have :

$$H(x(t), u(t), \lambda(t), t) = I(x(t), u(t), t) + \lambda(t) f(x(t), u(t), t)$$

Then one of necessary condition is :

$$I_x(x(t), u(t), t) + \lambda(t) f_x(x(t), u(t), t) + \dot{\lambda}(t) = 0$$

From HJB equation, suppose that $u(t)$ is optimal, then we differentiate both sides with respect to x and note that $\dot{x}(t) = f(x(t), u(t), t)$:

$$I_x(x(t), u(t), t) + \frac{\partial^2 V(x(t), t)}{\partial x^2} f(x(t), u(t), t) + \frac{\partial V(x(t), t)}{\partial x} f_x(x(t), u(t), t) + \frac{\partial^2 V(x(t), t)}{\partial t \partial x} = 0$$

$$I_x(x(t), u(t), t) + \frac{\partial^2 V(x(t), t)}{\partial x^2} \dot{x}(t) + \frac{\partial V(x(t), t)}{\partial x} f_x(x(t), u(t), t) + \frac{\partial^2 V(x(t), t)}{\partial t \partial x} = 0$$

$$I_x(x(t), u(t), t) + \frac{\partial V(x(t), t)}{\partial x} f_x(x(t), u(t), t) + \frac{\partial^2 V(x(t), t)}{\partial x^2} \dot{x}(t) + \frac{\partial^2 V(x(t), t)}{\partial t \partial x} = 0$$

$$I_x(x(t), u(t), t) + \frac{\partial V(x(t), t)}{\partial x} f_x(x(t), u(t), t) + \frac{\partial}{\partial t} \frac{\partial V(x(t), t)}{\partial x} = 0$$

From above two equation, we have :

$$\lambda(t) = \frac{\partial V(x(t), t)}{\partial x}$$

Chapter 9

Stochastic optimal control

This is also optimal control but there is additional stochastic factor such as noise, either in observations in evolution of the system.

1   Kalman filter

In statistics and control theory, Kalman filtering, also known as *linear quadratic estimation* (LQE), is an algorithm that uses a series of measurements observed over time, including statistical noise and other inaccuracies. This algorithm produces estimates of unknown variables that tend to be *more accurate than those based on a single measurement alone*, by estimating a joint probability distribution over the variables for each timeframe.

The algorithm works by a **two-phase process** :

1. The Kalman filter produces estimates of the current state variables, along with their uncertainties.

2. Once the outcome of the next measurement (necessarily corrupted with some error, including random noise) is observed, these estimates are updated using a weighted average. The bigger weights are given to estimates with greater certainty.

The algorithm is recursive. It can operate in real time, using three main factors : the measurements or observations $\mathbf{z}$, the state $\mathbf{x}$ and the certainty matrix (or covariance matrix of errors) $\mathbf{P}$.

1.1   Announcement

We consider the following announcement, with two phases characterized by *evolution equation* and *observation equation*:

Evolution equation :

$$\mathbf{x}_k = \mathbf{F}_k \mathbf{x}_{k-1} + \mathbf{B}_k \mathbf{u}_k + \mathbf{w}_k$$

where:

- k is time instant

- $\mathbf{x}_k$ is the state vector

- $\mathbf{F}_k$ is the state-transition matrix

- $\mathbf{B}_k$ is the controller (model) matrix

- $\mathbf{u}_k$ is the input vector for the controller

- $\mathbf{w}_k$ is supposed to be a centered normal noise

Observation equation

$$\mathbf{z}_k = \mathbf{H}_k \mathbf{x}_k + \mathbf{v}_k$$

where:

- z_k is the observation vertor (for what are observed)

- $\mathbf{H}_k$ is the observation matrix

- $\mathbf{v}_k$ is supposed to be a centered normal noise

$\mathbf{x}_0, \mathbf{w}_1, \ldots, \mathbf{w}_k, \mathbf{v}_1, \ldots, \mathbf{v}_k$ are all assumed to be mutually independent.

1.2   Objective

The objective is to have an estimation $\hat{\mathbf{x}}_k$ of $\mathbf{x}_k$, for all k, given (what we know) all $\mathbf{F}, \mathbf{B}, \ldots$ except $\mathbf{x}_k, \mathbf{w}_k, \mathbf{v}_k$. For the two latter, we only know their distribution (centered normal).

1.3   Algorithm

The notation $_{n|m}$ represents : at instant n given observations from 0 up to and including instant $m \leq n$. Now, we understand that $\hat{\mathbf{x}}_k = \hat{\mathbf{x}}_{k|k}$.

Phase 1, Prediction. Given the estimate $\hat{\mathbf{x}}_{k-1|k-1}$, the objective of this phase is just to estimate the state vector and the certainty matrix (or covariance matrix of errors), *before the observation k*, then :

$$\hat{\mathbf{x}}_{k|k-1} = \mathbf{F}_k \hat{\mathbf{x}}_{k-1|k-1} + \mathbf{B}_k \mathbf{u}_k$$

and the covariance matrix of errors :

$$
\begin{aligned}
\mathbf{P}_{k|k-1} &= Cov(\mathbf{x}_k - \hat{\mathbf{x}}_{k|k-1}) \\
&= Cov(\mathbf{F}_k \mathbf{x}_{k-1} + \mathbf{B}_k \mathbf{u}_k + \mathbf{w}_k - \mathbf{F}_k \hat{\mathbf{x}}_{k-1|k-1} - \mathbf{B}_k \mathbf{u}_k) \\
&= Cov\left(\mathbf{F}_k(\mathbf{x}_{k-1} - \hat{\mathbf{x}}_{k-1|k-1}) + \mathbf{w}_k\right) \\
&= Cov\left(\mathbf{F}_k(\mathbf{x}_{k-1} - \hat{\mathbf{x}}_{k-1|k-1})\right) + Cov(\mathbf{w}_k) \\
&= \mathbf{F}_k Cov(\mathbf{x}_{k-1} - \hat{\mathbf{x}}_{k-1|k-1})\mathbf{F}_k^T + \mathbf{Q}_k \\
&= \mathbf{F}_k \mathbf{P}_{k-1|k-1} \mathbf{F}_k^T + \mathbf{Q}_k
\end{aligned}
$$

where $\mathbf{Q}_k = Cov(\mathbf{w}_k)$

Phase 2, Update or adjustment with new observation

The objective of this phase is to update the state vector and the covariance matrix of errors, *after the observation* k. Let's define a new notation called innovation, which is thus the difference between the observation and the prediction :

$$\tilde{\mathbf{y}}_k = \mathbf{z}_k - \mathbf{H}_k \hat{\mathbf{x}}_{k|k-1}$$

We expect to use an unknown matrix $\mathbf{K}_k$ to update $\hat{\mathbf{x}}_{k|k-1}$ into $\hat{\mathbf{x}}_{k|k}$, by :

$$\hat{\mathbf{x}}_{k|k} = \hat{\mathbf{x}}_{k|k-1} + \mathbf{K}_k \tilde{\mathbf{y}}_k$$

Now we need a criterion for determining $\mathbf{K}_k$. Let's take the one that minimize :

$$\mathbb{E}\left[\left\|\mathbf{x}_k - \hat{\mathbf{x}}_{k|k}\right\|^2\right]$$

For the utility, we first calculate the covariance matrix become:

$$\begin{aligned}
\mathbf{P}_{k|k} &= Cov(\mathbf{x}_k - \hat{\mathbf{x}}_{k|k}) \\
&= Cov(\mathbf{x}_k - \hat{\mathbf{x}}_{k|k-1} - \mathbf{K}_k \tilde{\mathbf{y}}_k) \\
&= Cov\left(\mathbf{x}_k - \hat{\mathbf{x}}_{k|k-1} - \mathbf{K}_k(\mathbf{z}_k - \mathbf{H}_k \hat{\mathbf{x}}_{k|k-1})\right) \\
&= Cov\left(\mathbf{x}_k - \hat{\mathbf{x}}_{k|k-1} - \mathbf{K}_k(\mathbf{H}_k \mathbf{x}_k + \mathbf{v}_k) + \mathbf{K}_k \mathbf{H}_k \hat{\mathbf{x}}_{k|k-1}\right) \\
&= Cov\left((\mathbf{I} - \mathbf{K}_k \mathbf{H}_k)(\mathbf{x}_k - \hat{\mathbf{x}}_{k|k-1}) - \mathbf{K}_k \mathbf{v}_k\right) \\
&= Cov\left((\mathbf{I} - \mathbf{K}_k \mathbf{H}_k)(\mathbf{x}_k - \hat{\mathbf{x}}_{k|k-1})\right) + Cov(\mathbf{K}_k \mathbf{v}_k) \\
&= (\mathbf{I} - \mathbf{K}_k \mathbf{H}_k)Cov(\mathbf{x}_k - \hat{\mathbf{x}}_{k|k-1})(\mathbf{I} - \mathbf{K}_k \mathbf{H}_k)^T + \mathbf{K}_k Cov(\mathbf{v}_k)\mathbf{K}_k^T \\
&= (\mathbf{I} - \mathbf{K}_k \mathbf{H}_k)\mathbf{P}_{k|k-1}(\mathbf{I} - \mathbf{K}_k \mathbf{H}_k)^T + \mathbf{K}_k \mathbf{R}_k \mathbf{K}_k^T
\end{aligned}$$

where $\mathbf{R}_k = Cov(\mathbf{v}_k)$.

Note that $\mathbb{E}\left[\left\|\mathbf{x}_k - \hat{\mathbf{x}}_{k|k}\right\|^2\right] = \mathrm{tr}(\mathbf{P}_{k|k})$. Then we perform the first-order optimal condition, let's remind first some properties of trace and its derivative :

- $\mathrm{tr}\,(ABCD) = \mathrm{tr}\,(BCDA) = \mathrm{tr}\,(CDAB) = \mathrm{tr}\,(DABC)$

- $\displaystyle\nabla_A \,\mathrm{tr}\,(AB) = \nabla_A \,\mathrm{tr}\,(B^T A^T) = B^\top$

- $\displaystyle\nabla_A \,\mathrm{tr}\,(ABA^\top C) = C^\top AB^\top + CAB$

Then (with the note that covariance matrix P and R are sysmetric) :

$$\begin{aligned}
\frac{\partial \,\mathrm{tr}(\mathbf{P}_{k|k})}{\partial \mathbf{K}_k} &= \frac{\partial \,\mathrm{tr}\left(\mathbf{P}_{k|k-1} - \mathbf{K}_k \mathbf{H}_k \mathbf{P}_{k|k-1} - \mathbf{P}_{k|k-1}(\mathbf{K}_k \mathbf{H}_k)^T + \mathbf{K}_k \mathbf{H}_k \mathbf{P}_{k|k-1}(\mathbf{K}_k \mathbf{H}_k)^T + \mathbf{K}_k \mathbf{R}_k \mathbf{K}_k^T\right)}{\partial \mathbf{K}_k} \\
&= -2\mathbf{P}_{k|k-1}\mathbf{H}_k^T + 2\mathbf{K}_k \mathbf{H}_k \mathbf{P}_{k|k-1}\mathbf{H}_k^T + 2\mathbf{K}_k \mathbf{R}_k
\end{aligned}$$

This given :

$$\mathbf{K}_k = \mathbf{P}_{k|k-1}\mathbf{H}_k^T(\mathbf{H}_k\mathbf{P}_{k|k-1}\mathbf{H}_k^T + \mathbf{R}_k)^{-1} = \mathbf{P}_{k|k-1}\mathbf{H}_k^T(\mathbf{S}_k)^{-1}$$

where

$$\mathbf{S}_k = \mathbf{H}_k\mathbf{P}_{k|k-1}\mathbf{H}_k^T + \mathbf{R}_k$$

Now with the optimal $\mathbf{K}_k$, which implies $\mathbf{K}_k\mathbf{S}_k = \mathbf{P}_{k|k-1}\mathbf{H}_k^T$, we replace it into the above formula of $\mathbf{P}_{k|k}$, this leads to :

$$\mathbf{P}_{k|k} = \mathbf{P}_{k|k-1} - \mathbf{K}_k\mathbf{H}_k\mathbf{P}_{k|k-1} - \mathbf{P}_{k|k-1}(\mathbf{K}_k\mathbf{H}_k)^T + \mathbf{K}_k\mathbf{S}_k\mathbf{K}_k^T$$
$$= (\mathbf{I} - \mathbf{K}_k\mathbf{H}_k)\mathbf{P}_{k|k-1} \quad \text{(the last two terms cancel out)}$$

1.4   Resumed algorithm

Phase 1:

- $\mathbf{P}_{k|k-1} = \mathbf{F}_k\mathbf{P}_{k-1|k-1}\mathbf{F}_k^T + \mathbf{Q}_k$

Phase 2 :

- $\tilde{\mathbf{y}}_k = \mathbf{z}_k - \mathbf{H}_k\hat{\mathbf{x}}_{k|k-1}$ (innovation)

- $\mathbf{S}_k = \mathbf{H}_k\mathbf{P}_{k|k-1}\mathbf{H}_k^T + \mathbf{R}_k$ (also be the covariance of innovation)

- $\mathbf{K}_k = \mathbf{P}_{k|k-1}\mathbf{H}_k^T\mathbf{S}_k^{-1}$ (Kalman gain)

- $\hat{\mathbf{x}}_{k|k} = \hat{\mathbf{x}}_{k|k-1} + \mathbf{K}_k\tilde{\mathbf{y}}_k$ (update state)

- $\mathbf{P}_{k|k} = (I - \mathbf{K}_k\mathbf{H}_k)\mathbf{P}_{k|k-1}$ (update covariance)

1.5   Discretization for non-linear case

In the above problem, we have $\mathbf{x}_k = \mathbf{F}_k\mathbf{x}_{k-1}$. $\mathbf{F}_k$ is constant from instant k to $k+1$ and it means a *linear dynamical system*. For the non-linear case, we can take a discretization such that the duration between instant k to $k+1$ is short and then we can consider $\mathbf{F}_k$ is constant. If F is something differentiable then we can take:

$$\mathbf{F}_k = \frac{d\mathbf{x}(t)}{dt}$$

where t is anything between instant k to $k+1$.

1.6 Continous case

The continuous time version of Kalman filtering is also called Kalman–Bucy filtering.

Dynamic system

$$\mathbf{x}'(t) = \mathbf{F}(t)\mathbf{x}(t) + \mathbf{B}(t)\mathbf{u}(t) + \mathbf{w}(t)$$
$$\mathbf{z}(t) = \mathbf{H}(t)\mathbf{x}(t) + \mathbf{v}(t)$$

Solution

1. We find the covariance matrix $\mathbf{P}(t)$ by solving the following differential equation (Riccati equation):

$$\mathbf{P}'(t) = \mathbf{F}(t)\mathbf{P}(t) + \mathbf{P}(t)\mathbf{F}^{\mathsf{T}}(t) + \mathbf{Q}(t) - \mathbf{P}(t)\mathbf{H}^{\mathsf{T}}(t)\mathbf{R}^{-1}(t)\mathbf{H}(t)\mathbf{P}^{\mathsf{T}}(t)$$

2. Then we get the Kalman gain :

$$\mathbf{K}(t) = \mathbf{P}(t)\mathbf{H}^{\mathsf{T}}(t)\mathbf{R}^{-1}(t)$$

3. Finally, find the estimate of state by solving :

$$\hat{\mathbf{x}}'(t) = \mathbf{F}(t)\hat{\mathbf{x}}(t) + \mathbf{B}(t)\mathbf{u}(t) + \mathbf{K}(t)\left(\mathbf{z}(t) - \mathbf{H}(t)\hat{\mathbf{x}}(t)\right)$$

2 Linear–quadratic–Gaussian control

The linear–quadratic–Gaussian (LQG) control problem is one of the most fundamental optimal control problems.

2.1 Announcement

Consider the continuous-time linear dynamic system :

$$\begin{cases} \dot{\mathbf{x}}(t) = A(t)\mathbf{x}(t) + B(t)\mathbf{u}(t) + \mathbf{v}(t) \\ \mathbf{y}(t) = C(t)\mathbf{x}(t) + \mathbf{w}(t) \end{cases},$$

as in Kalman filter (sec 1), $\mathbf{x}$ is state vector, $\mathbf{u}$ is control vector, $\mathbf{v}$ and $\mathbf{w}$ are white Gaussian noise. $A(t), B(t), C(t)$ are parameters.

Given this dynamic system, then the objective is to find the control $\mathbf{u}(t)$ such that the following cost function is minimized:

$$J = \mathbb{E}\left[\mathbf{x}^{\mathsf{T}}(T)F\mathbf{x}(T) + \int_0^T \mathbf{x}^{\mathsf{T}}(t)Q(t)\mathbf{x}(t) + \mathbf{u}^{\mathsf{T}}(t)R(t)\mathbf{u}(t)\, dt\right],$$

$$F \geq 0, \quad Q(t) \geq 0, \quad R(t) > 0,$$

The final time or T may be either finite or infinite. If the horizon tends to infinity the first term $\mathbf{x}^{\mathsf{T}}(T)F\mathbf{x}(T)$ of the cost function becomes negligible and irrelevant to the problem. Indeed, LQG problem is LQR problem (sec 4) with additional noise in state vector and observation vector.
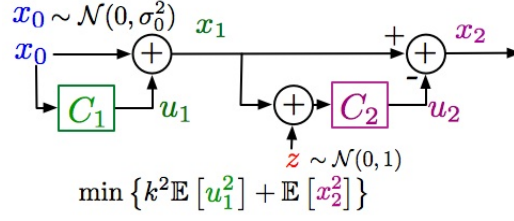
Figure 9.1: Diagram for Witsenhausen's counterexample and its cost function. C_1 and C_2 are two controllers.

2.2 Certainty equivalence

This is a property show that, **if** :

- Dynamic system is linear.

- Noise is additive $(x + \varepsilon)$, in observation or in evoluation.

- Objective function is the expected value of a quadratic form.

- Control system is centralized. It means a system that has a single central authority that makes decisions and controls the entire system. In contrast, a decentralized control system distributes decision-making authority across multiple entities or nodes. For example, see section 3.

Then the optimal control solution in this case is the same as would be obtained in the absence of the additive noise.

More over, the quadratic cost allows for the optimal control vector $\mathbf{u}(t)$ to be linear functions of the observations $\mathbf{y}(t)$:

$$\mathbf{u}(t) = M(t)\mathbf{y}(t) + N(t)$$

2.3 Solution

The solution for LQG controller, by the certainty equivalence, can be seen a combination of two steps:

1. Using Kalman filter to find the estimate $\hat{\mathbf{x}}$ of state vector $\mathbf{x}$.

2. Then applying LQR for $\hat{\mathbf{x}}$ instead of $\mathbf{x}$.

3 Witsenhausen's counterexample

This example is shown in figure 9.1.
Input for controller :

- The first controler u_1 take the observation x_0 as input : $u_1 = f(x_0)$

- The second controler u_2 take the observation $y_1 = x_1 + z$ as input : $u_2 = g(y_1)$

Since the controller (C_1) and (C_2) can not communicate with each other (information of C_1 can not be seen for C_2), then the control system is **decentralized**.

3.1 Annoucement

Dynamic system :

$$\begin{cases} x_1 = x_0 + u_1, \quad x_0 \sim \mathcal{N}(0, \sigma_0^2) \\ x_2 = x_1 - u_2 \end{cases}$$

Observation :

$$\begin{cases} x_0 \\ y_1 = x_1 + z, \quad z \sim \mathcal{N}(0, 1) \end{cases}$$

z and x_0 are independent. Cost function to minimize:

$$J = k^2 E[u_1^2] + E[x_2^2]$$

3.2 Notes

We are curious to find the certainty equivalence, when there is no noise z. Then

$$\begin{aligned} x_2 &= x_1 - g(x_1 + z) \\ &= x_0 + f(x_0) - g(x_0 + f(x_0) + z) \\ &= x_0 + f(x_0) - g(x_0 + f(x_0)) \end{aligned}$$

As $J \geq 0$, then if we take $f(a) = 0$ (zero mapping) and $g(a) = a$ (identical function), then $x_2 = 0$ and $J = k^2 E[u_1^2] + E[x_2^2] = k^2 0 + 0 = 0$.

Let's use this solution with uncertain factor (noise). This implies that $x_2 = -z$ and $J = E[x_2^2] = E[(-z)^2] = 1$.

However, let's try an other simple solution, f is always zero mapping and now $g(a) = 0.5a$, then $x_2 = 0.5x_0 - 0.5z$ and $J = E[(0.5x_0 - 0.5z)^2] = E[(0.5x_0)^2] + E[(0.5z)^2] = \frac{1}{4}(\sigma_0^2 + 1)$. If $\sigma_0 = 1$, then $J = \frac{1}{2}$ and it is a better solution then the one of certainty equivalence.

4 Hamilton–Jacobi–Bellman equation

This is the stochastic version of HJB equation. Consider the following problem, which is the stochastic version in section 5 :

$$V(X_{t_0}, t_0) = \min_u \int_{t_0}^{t_1} I(t, X_t, u_t) \, \mathrm{d}t + \phi(X_{t_1})$$

where :

- $X_t, t \in [0, T]$ the stochastic process

- u_t the steering or control function.

By first using Bellman and then expanding $V(X_t, t)$ with Itô's rule, one finds the stochastic HJB equation

5 Stochastic programming

Stochastic programming is a mathematical programming framework that is used to solve optimization problems involving uncertainty. A stochastic program is an optimization problem in which some or all problem parameters are uncertain, but follow known probability distributions. This framework contrasts with deterministic optimization, in which all problem parameters are assumed to be known exactly. Because many real-world decisions involve uncertainty, stochastic programming has found applications in a broad range of areas ranging from finance to transportation to energy optimization.

Example :

http://cgm.cs.mcgill.ca/~avis/courses/567/roussos/intro.pdf

Chapter 10

Merton's portfolio problem

This chapter is reported from "On Merton's Portfolio Problem" at https://umu.diva-portal.org/smash/get/diva2:1667781/FULLTEXT01.pdf

We classify the Merton's portfolio problem as a chapter since it requires a lot of knowledges in stochastic calculus and optimal control.

There are two main approaches to this method: the traditional approach, and the modern approach.

- The traditional approach uses dynamic programming, pioneered by Bellman, to solve the stochastic problems utlizing the Hamilton-Jacobi-Bellman equation.

- The modern approach on the other hand uses martingales or the duality approach. By using the martingale representation of wealth, the more direct martingale approach can be used to solve the problem. This approach is however more complicated for incomplete markets as it does not conclude a unique solution whereas the traditional method can be used for both complete and incomplete markets.

As most economic market are considered as incomplete markets, the traditional method is most likely to be favoured.

1 Reminding stochastic calculus

We remind the Brownian motion (section 8), denoted by $W_s(\omega)$, where s is the time. If necessary, the Stochastic Integral (sec 2) and an example Geometric Brownian Motion (sec 2.4) are also recommanded to revise.

In resume, give the SDE (or dynamic equation) with μ, σ are known:

$$dX_t = \mu(t, X_t)\mathrm{d}t + \sigma(t, X_t)\mathrm{d}W_t$$

with the initial condition $X_0 = x$ (need to be well determined).

We need to know how to solve it by Itô lemma to get SP X_t.

2 Stochastic Optimal Control

We start by giving some definitions :

- t_0 : starting time
- t : starting time in dynamic programming fashion
- T: ending time
- s : instantaneous time $(t \leq s \leq T)$
- α_s : is the control process.
- f : utility function
- g : terminal utility function
- W_s : standardised Brownian motion
- $X_s^{t,x}$: Stochastic Process X_s with the initial condition $X_t = x$.

2.1 Objective function

We define first the objective function :

$$J(t_0, x_0, \alpha) = E\left[\int_{t_0}^{T} f(s, X_s^{t_0, x_0}, \alpha_s)\mathrm{d}s + g(X_T^{t_0, x_0})\right],$$

where $X_s^{t_0, x_0}$ is the SP, which now taking into account the control factor α_s in the **dynamic equation**:

$$dX_s = \mu(s, X_s, \alpha_s)\mathrm{d}s + \sigma(s, X_s, \alpha_s)\mathrm{d}W_s,$$

with the initial condition $X_{t_0} = x_0$.

2.2 Value function

We define also **value function**, in a dynamic programming fashion :

$$J(t, x, \alpha) = E\left[\int_{t}^{T} f(s, X_s^{t,x}, \alpha_s)\mathrm{d}s + g(X_T^{t,x})\right]$$

2.3 Requirements on control process

As now there is a control process α_s in both the objective function and the dynamic equation, we need some conditions on this control process α_s to make sure that the optimization is solvable. Thus, we need that α_s is admissible

2.3.1 Admissible

Then the control process α_s is said **admissible** if **for any** given initial condition (t, x) there exists **a unique solution** to the SDE :

$$dX_s = \mu(s, X_s, \alpha_s)\mathrm{d}s + \sigma(s, X_s, \alpha_s)\mathrm{d}W_s,$$

with $X_t = x$.

We note $\alpha \in \mathcal{A}$ if α is admissible.

This admissible condition implies the following condition.

2.3.2 Adapted to $\mathcal{F}_s^W$-filtration

Note that the stochastic part of the state process X_s (without taking into account α_s) :

$$dX_s = \mu(s, X_s)\mathrm{d}s + \sigma(s, X_s)\mathrm{d}W_s,$$

is supposedly Brownian motion W_s by definition. So X_s is continuous and adapted to the $\mathcal{F}_s^W$ -filtration.

Then to make sure that

$$dX_s = \mu(s, X_s, \alpha_s)\mathrm{d}s + \sigma(s, X_s, \alpha_s)\mathrm{d}W_s,$$

has solution or well determined, we need that α_s is also adapted to $\mathcal{F}_s^W$-filtration. For this, we can suppose that α_s is a function of s and X_s :

$$\alpha_s = g(s, X_s)$$

2.4 Optimal value function

The optimal value function $v(t, x)$ is defined as

$$v(t, x) = J(t, x, \hat{\alpha}) = \sup_{\alpha \in \mathcal{A}} E\left[\int_t^T f(s, X_s^{t,x}, \alpha_s)\mathrm{d}s + g(X_T^{t,x})\right]$$

Then the optimal objective function is $v(t_0, x_0)$

2.5 Pipeline

The pipeline to solve this solution can be divided into four steps :

1. We need to know how the value function is defined (the explicite form of utility function f and of terminal utility function g, ...) and to find an SDE that represents the state process (the explicite form of μ and σ).

2. Use dynamic programming to divide the problem into sub-problems from which we derive a PDE called a Hamilton-Jacobi-Bellman equation which should hold

3. The third step starts with choosing an "ansatz", i.e. making a qualified guess of what a solution to the Hamilton-Jacobi-Bellman equation should look like. Differentiate and plug the ansatz into the Hamilton-Jacobi-Bellman equation. Now try to solve the equation.

4. The fourth step is to take the solution of the Hamilton-JacobiBellman equation and verify that this solution is indeed the optimal value function and that it is utilizing the optimal control.

2.6 Dynamic Programming Principle

We can divide the optimization problem of value function into the sub-intervals $[t, \theta], [\theta, T]$ such as :

$$v(t, x) = \sup_{\alpha \in \mathcal{A}} E \left[\int_t^T f(s, X_s^{t,x}, \alpha_s) \mathrm{d}s + g(X_T^{t,x}) \right]$$
$$= \sup_{\alpha \in \mathcal{A}} E \left[\int_t^\theta f(s, X_s^{t,x}, \alpha_s) \mathrm{d}s + v(\theta, X_\theta^{t,x}) \right]$$

2.7 Using Hamilton Jacobi Bellman Equation

Chapter 11

Maths analytics

1   Differential equation

Ordinary differential equation (ODE) is a differential equation whose unknown(s) consists of one (or more) function(s) of **one variable** (here x) and involves the derivatives of those functions :

$$a_0(x)u + a_1(x)u' + a_2(x)u'' + \ldots + a_n(x)u^{(n)} + b(x) = 0$$

Partial differential equation (PDE) Have **more than one independent variable** (here x, y, z) and usually in constrast with ODE. For example :

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} = 0.$$

2   Finite difference method

(Sai phan huu han) Finite-difference methods (FDM) are a class of numerical techniques for solving differential equations by approximating derivatives with finite differences. Both the spatial domain and time interval (if applicable) are discretized, or broken into a finite number of steps, and the value of the solution at these discrete points is approximated by solving algebraic equations containing finite differences and values from nearby points.

2.1   Finite differences

3 types of finite differences :

- Forward difference, denoted $\Delta_h[f]$:

$$\Delta_h[f](x) = f(x + h) - f(x)$$

- Backward difference, denoted $\nabla_h[f]$:

$$\nabla_h[f](x) = f(x) - f(x - h)$$

- Central difference, denoted $\delta_h[f]$:

$$\delta_h[f](x) = f(x + \frac{h}{2}) - f(x - \frac{h}{2})$$

2.2 An example of finite difference method

Considering the following problem :

$$u'(x) - \tau u(x) = 0, \forall x \in [0, 1], \ u(0) = u_0$$

By using finite difference with step h :

$$\frac{u_{n+1} - u_n}{h} - \tau u_n = 0, \forall 0 \leq n \leq M.$$

where $M = \frac{1}{h} - 1$ and $u_n = u(nh)$.

Then:

$$u_{n+1} = (1 + h\tau)u_n = (1 + h\tau)^{n+1}u_0$$

Or:

$$u_n = (1 + h\tau)^n u_0 \underset{h \to 0}{=} e^{h\tau n}u_0$$

On other hand, we know that the solution exact is :

$$u(x) = u_0 e^{\tau x}$$

$$u_n = u(nh) = u_0 e^{\tau nh}$$

3 Weak formulation

Linear system of equations

Let $V = \mathbb{R}^n$ and $A : V \to V$ be a linear mapping. Then, the weak formulation of the equation :

$$Au = f$$

involves finding $u \in V$ such that for all $v \in V$ the following equation holds:

$$\langle Au, v \rangle = \langle f, v \rangle$$

where $\langle \cdot, \cdot \rangle$ denotes an inner product.

Simple idea :With $a, b \in \mathbb{R}$ if $ax = bx, \forall x \in R$, then we must have $a = b$.

Let $v_1, ..., v_n$ be n linear independent vectors, then if we can find u that satifies:

$$\langle Au, v_i \rangle = \langle f, v_i \rangle, \forall i = 1, \ldots, n$$

Now suppose that u is decomposed into this new basis vector $v_1, \ldots, v_n$:

$$u = \sum_j^n u_j^v v_j$$

where u_i^v are scalar. Then

$$\langle Au, v_i \rangle = \langle A \sum_j^n u_j^v v_j, v_i \rangle$$
$$= \sum_j^n \langle Av_j, v_i \rangle u_j^v$$
$$= \langle f, v_i \rangle$$

Now we need to solve $\mathbf{u}^v \in \mathbb{R}^n$ with :

$$\mathbf{A}^v \mathbf{u}^v = \mathbf{f}^v$$

where $\mathbf{A}_{ij}^v = \langle Av_j, v_i \rangle, \mathbf{f}_i^v = \langle f, v_i \rangle$. Note that once we find out $\mathbf{u}^v$, we need to change back to u by :

$$u = \sum_j^n u_j^v v_j$$

4   Finite element method

The following two problems (boundary value problem BVP) demonstrate the finite element method.

4.1   Example problems

$$\text{P1} : \begin{cases} u''(x) = f(x), x \in (0, 1) \\ u(0) = u(1) = 0 \end{cases}$$

where f is given, u is an unknown function of x.

$$\text{P2} : \begin{cases} u_{xx}(x, y) + u_{yy}(x, y) = f(x, y), (x, y) \in \Omega \\ u = 0 \text{ on } \partial\Omega \end{cases}$$

where Ω is a connected open region in the (x, y) and $\partial\Omega$ is its boundary.

4.2 General method

Two steps :

- In the first step, one rephrases the original BVP in its weak form. Little to no computation is usually required for this step. The transformation is done by hand on paper.

- The second step is the discretization, where the weak form is discretized in a finite-dimensional space.

After this second step, we have concrete formula for a large but finite-dimensional linear problem whose solution will approximately solve the original BVP. This finite-dimensional problem is then implemented on a computer.

4.3 Step 1 : Weak formulation

For P1, u is solution if for every smooth functions v, we have :

$$\langle f, v \rangle = \langle u'', v \rangle$$
$$\int_0^1 fv dx = \int_0^1 u'' v dx$$

As we need to find $u(x)$ where $0 < x < 1$, then we can take $v(0) = v(1) = 0$. Applying the integral by parts :

$$\int_0^1 fv dx = \int_0^1 u'' v dx$$
$$= u'v|_0^1 - \int_0^1 u'v' dx$$
$$= -\langle u', v' \rangle$$
$$= -\phi(u, v)$$

For P2, u is solution if for every smooth functions v, we have :

$$\langle f, v \rangle = \langle u'', v \rangle$$
$$\int_\Omega fv ds = \int_\Omega u'' v ds$$

s here can be considered as small area (since we are in 2 dimensinal).As we need to find $u(x, y)$ where x, y are not on the boundary $\partial\Omega$, then we can take $v(x, y) = 0$ if $(x, y) \in \partial\Omega$. Applying the integral by parts :
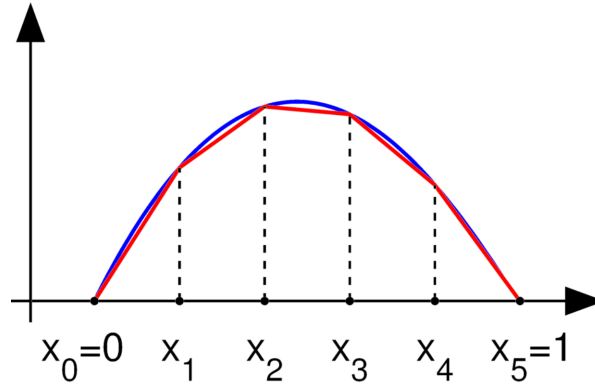
Figure 11.1: Linear approximation for curved function.

$$\int_\Omega fv ds = \int_\Omega \Delta u v ds$$
$$= \oint_{\partial\Omega} \nabla u v dl - \int_\Omega \nabla u \nabla v ds$$
$$= 0 - \langle \nabla u, \nabla v \rangle$$
$$= -\phi(u, v)$$

Here, we apply the first identity of Green, and l can be considered as small path 1d.

4.4 Step 2 : Discretization

This is the main idea of FEM.

For P1, one discretizes the segment $[0, 1]$ into n segments $[x_i, xi + 1]$, where $x_0 < x_1 < \ldots < x_{n+1}$ ($x_{i+1} - x_i$ are not necessary equidistant). Given function f, if n is big enough, one can consider function f, for each x_i to x_{x+1} is approximately a linear function (figure 11.1) :

$$f_{i,i+1}^a(x) = \frac{f(x_{i+1})(x - x_i) - f(x_i)(x - x_{i+1})}{x_{i+1} - x_i}$$

This approximation is something like theory of integration.

The advantage of piecewise linear approximation is that we can construct (find) a basis that can generate any piecewise linear function. A common basis is composed of tent functions.

Tent function basis

$$v_k(x) = \begin{cases} \frac{x - x_{k-1}}{x_k - x_{k-1}} & \text{if } x \in [x_{k-1}, x_k] \\ \frac{x_{k+1} - x}{x_{k+1} - x_k} & \text{if } x \in [x_k, x_{k+1}] \\ 0 \end{cases}$$
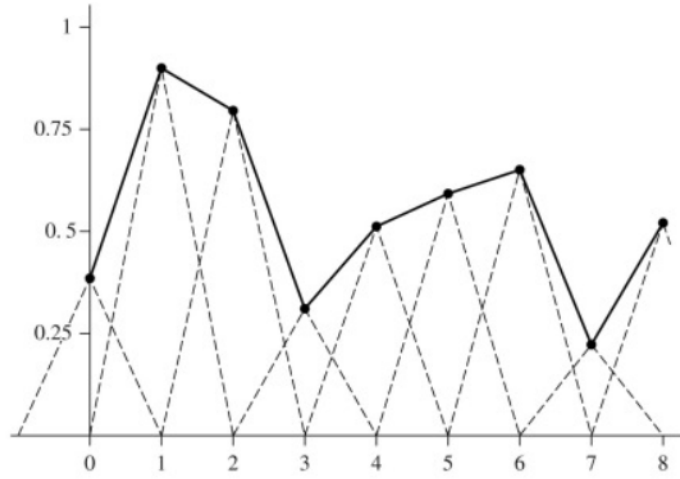
Figure 11.2: An example that show all piecewise linear function can be decomposed into tent functions basis.

where $k = 0, \ldots, n+1$. Note that if $k = 0$ or $k = n+1$, we use only the half right or half left of tent function. The number of discritized points is equal to the number of tent functions.

Here is an example show that, for any piecewise linear function, we can reconstruct it by tent functions (figure 11.2). Indeed, for any linear function, we can reconstruct it by the sum of two "diagonal" functions. Also, the value at discretized point of function is the peak of tent.

For P2 : The discrization is shown as in 11.3. The points are set on the boundery and in the definition domain of function. A question raised here is how we can connect these points to define a pieceweise linear function ? (a surface with 3 vertexes in this case). A intuitive is to use triangulation. This means for any polygon, we can devide it into many triangles. Therefore, given a definition domain and its discretized points, we create polygons from these points (without intersection and need to cover the whole domain possible). Then using triangulation for each polygon.

Tent function basis

As in case 1d, we adapt for tent function for the case 2d. Remind that pieceweise linear approximation is a surface with 3 vertexes. Then we need 3 "diagonal" functions for reconstruct this surface. A "diagonal" function is made by 3 points : 2 points are discretized points and the resting point is the value of considering function evaluated at the resting discretized point.

Tent function here are all pyramids, where top vertex lies on discretized point s, and the resting points are discretized points that has connection to this discretized point s.
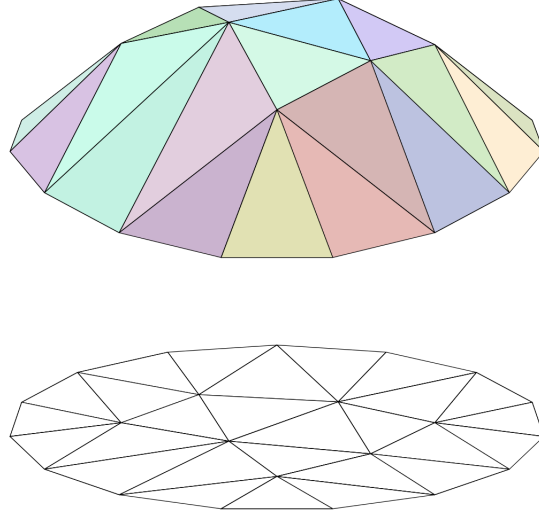
Figure 11.3: An example that show all piecewise linear function can be decomposed into tent functions basis.

4.5 Characteristic of basis

P1:

$$\langle v_j, v_k \rangle = \int_0^1 v_j v_k dx$$

and

$$\phi(v_j, v_k) = \int_0^1 v'_j v'_k dx$$

will be zero for almost all j, k if $|j - k| > 1$.

P2:

$$\langle v_j, v_k \rangle = \int_\Omega v_j v_k dx$$

and

$$\phi(v_j, v_k) = \int_\Omega \nabla v'_j \nabla v'_k dx$$

will be zero if x_j and x_k do not share an edge of the triangulation, then the integrals

4.6 Matrix form of the problem

$$u(x) = \sum_{k=0}^{n+1} u_k v_k$$

383

u_k is scalar.

$$f(x) = \sum_{k=0}^{n+1} f_k v_k$$

f_k is scalar.

Then for each v_j, we must have

$$-\phi(u, v_j) = \langle f, v_j \rangle$$

$$-\int_0^1 u' v_j' dx = \int_0^1 f v_j dx$$

$$-u_k \sum_{k=0}^{n+1} v_k' v_j' dx = f_k \sum_{k=0}^{n+1} v_k v_j dx$$

$$-\mathbf{L}[u_0, \ldots, u_{n+1}]^T = \mathbf{M}[f_0, \ldots, f_{n+1}]^T$$

$$-\mathbf{Lu} = \mathbf{Mf}$$

where $\mathbf{L}[i,j] = \phi(v_i, v_j)$ and $\mathbf{M}[i,j] = \langle v_i, v_j \rangle$. Finally, we can solve $-\mathbf{Lu} = \mathbf{Mf}$ for variable $\mathbf{u}$.

In an practice way, if n is suficiently big, we do not need to decompose f into basis function, we can solve for the second equation above :

$$-\mathbf{Lu} = \mathbf{b}$$

where $\mathbf{b}[i] = \int_0^1 f v_i dx$.

5   FEM and FDF comparison

Finite element method (FEM) and finite difference method (FDM) are both numerical methods used for solving partial differential equations (PDEs) and other mathematical problems.

The main difference between the two methods lies in how they approximate the solutions to the differential equations.

In FEM, the domain of the problem is divided into a set of smaller elements, which are connected at their boundaries to form a mesh. The solution to the PDE is then approximated by representing it as a linear combination of basis functions over each element of the mesh. The coefficients of the basis functions are determined by solving a set of linear equations that arise from applying the boundary conditions.

On the other hand, FDM involves discretizing the domain into a grid of points and approximating the derivatives in the PDEs using finite differences between the grid points. The PDEs are then replaced by a set of algebraic equations, which can be solved numerically using matrix algebra.

One advantage of FEM is that it can handle irregular geometries and boundaries more easily than FDM. FEM also tends to produce more accurate solutions, especially for problems with complex geometry or irregular boundaries. However, FEM can be more computationally expensive than FDM, and the formulation of the basis functions can be more complex.

FDM, on the other hand, is simpler to implement and can be more computationally efficient, especially for problems with regular geometries or boundaries. However, FDM may require a finer grid to achieve accurate results, which can increase the computational cost. FDM may also be less accurate than FEM for problems with complex geometry or irregular boundaries.

In summary, the choice between FEM and FDM depends on the specific problem being solved and the available computational resources.

6 Factor analysis

Factor analysis is a statistical method used to describe variability among observed, correlated variables in terms of a potentially lower number of unobserved variables called **factors** (or **latent variable** or embbedding representation). The observed variables are modelled as linear combinations of the potential factors plus "error" terms, hence factor analysis can be thought of as a special case of errors-in-variables models. The **factor loading** (or coef associated) of a variable quantifies the extent (muc do, level) to which the variable is related to a given factor.

A common rationale behind factor analytic methods is that the information gained about the interdependencies between observed variables can be used later to reduce the set of variables in a dataset. It may help to deal with data sets where there are large numbers of observed variables that are thought to reflect a smaller number of underlying/latent variables. It is one of the most commonly used inter-dependency techniques and is used when the relevant set of variables shows a systematic inter-dependence and the objective is to find out the latent factors that create a commonality.

6.1 Statistical model

The model attempts to explain a set of p observations in each of n individuals (n : number of original dimensions) with a set of k common factors (atoms, $f_{i,j}$) where there are fewer factors per unit than observations per unit ($k < p$). Each individual has k of their own common factors, and these are related to the observations via the factor loading (coef) matrix ($L \in \mathbb{R}^{p \times k}$, for a single observation, according to

$$x_{i,m} - \mu_i = l_{i,1} f_{1,m} + \ldots + l_{i,k} f_{k,m} + \varepsilon_{i,m}$$

where

- $x_{i,m}$ is the value of the ith observation of the mth individual (component)

- μ_i is the observation mean for the ith observation

- $l_{i,j}$ is the loading (coef) for the ith observation of the jth factor

- $f_{j,m}$ is the value of the jth factor of the mth individual

- $\varepsilon_{i,m}$ is the $(i,m)^{th}$ unobserved stochastic error term with mean zero and finite variance.

In matrix notation
$$X - M = LF + \varepsilon$$
where observation matrix $X \in \mathbb{R}^{p \times n}$, mean matrix $M \in \mathbb{R}^{p \times n}$ where $M_{i,m} = \mu_i$, loading matrix $L \in \mathbb{R}^{p \times k}$, factor matrix $F \in \mathbb{R}^{k \times n}$, error term matrix $\varepsilon \in \mathbb{R}^{p \times n}$. Note that in some litteratures, these matrix can be using in the transposed form (e.g., dictionary learning) :

$$X^T - M^T = F^T L^T + \varepsilon^T$$

Also we will impose the following assumptions on F:

- F and ε are independent.

- $E(F_{j,:})$=0, means the mean of a factor must be 0.

- $Var(F) = I$, means factors are uncorrelated or linearly independent, one and the other.

Properties :

$$
\begin{aligned}
Var(X - M) &= Var(LF + \varepsilon) \\
&= Var(LF + \varepsilon) \\
&= Var(LF) + 2Cov(LF, \varepsilon) + Var(\varepsilon) \\
&= Var(LF) + Var(\varepsilon) \\
&= E[(LF)^*2] - E[LF]^*2 + \Psi \\
&= E[(LF)(LF)^T] - (LE[F])^*2 + \Psi \\
&= E[LFF^T L^T] - 0 + \Psi \\
&= LE[FF^T]L^T + \Psi \\
&= L(E[F^*2] - 0)L^T + \Psi \\
&= L(E[F^*2] - E[F]^*2)L^T + \Psi \\
&= LVar(F)L^T + \Psi \\
&= LL^T + \Psi
\end{aligned}
$$

6.2   Popular Methods

In case of F is unknown, we try to find out the factor loading or representation that respect as much as possible the proximities and distances between points. Several popular methods can be listed

- Principal component analysis (PCA)

- Correspondence analysis (CA)

- Multiple correspondence analysis (MCA)

- Factor analysis of mixed data (FAMD)

In case of F is known, we can factor an example is the model APT (see 11). The F represents systematic factor such as return of market, inflation, ... in term of time and X represents return rate of independent stocks.