Probability Theory and Distribution Theory

沈威宇

June 29, 2025

Contents

Chapter 1 Probability Theory (機率論) and Distribution Theory (分布論)

1 Probability (機率)

I Experiment or Trial (試驗)

- (Random) Experiment or Trial (（隨機）試驗): A process that can be repeated and the results may be different each time. It has repeatability, that is, the test can be repeated under the same conditions, and randomness, that is, the results of each test may be different, with randomness and uncertainty.
- Sample space (樣本空間): The set of all possible outcomes of an experiment. For example, the sample space for flipping a coin is $\{\text{head}, \text{tail}\}$.
- Event (事件): A subset of the sample space. For example, the event of rolling a die and getting an even number is $\{2, 4, 6\}$.
- Impossible event (空事件): An event with zero probability.
- Sure event (全事件): An event with one probability.
- Sum event (和事件): The sum event of event A and event B is $A \cup B$.
- Product event (積事件): The product event of event A and event B is $A \cap B$.
- Complement event (餘事件): The complement event of event A in sample space S is $A' = S \setminus A$.
- Partitions (分割或劃分): If $\{A_i\}_{i \in I}$ is a set of partitions of the sample space Ω then the following conditions are met:

$$\begin{aligned} &\forall i \in I : A_i \subseteq \Omega, \\ &\forall i, j \in I \wedge i \neq j : A_i \cap A_j = \varnothing, \\ &\bigcup_{i \in I} A_i = \Omega. \end{aligned}$$

- Repeated trials: A trial that consists of repeated independent trials. The probability of an event A where events A_i occured in each trial i is $\prod_i P(A_i)$.
- Objective probability (客觀機率) or Frequency probability (頻率機率): An objective probability value obtained based on past experience or statistical data, usually the frequency of past events or repeated experiments to obtain the probability of an event occurring.

- Subjective probability (主觀機率): A probability value that is not supported by statistical data.

II Probability Rules

- Probability (機率): The likelihood of an event occurring, a number between 0 and 1. The closer the probability is to 1, the more likely the event is to occur.
- Classical probability (古典機率): If the number of all possible outcomes of an event is finite or countable infinite, and the chance of each outcome occurring in the sample space is equal, then the probability of the event occurring can be calculated by:

$$P(A) = \frac{\text{The number of outcomes of event A}}{\text{The number of all possible outcomes}}$$

- Conditional probability (條件機率): The probability of another event occurring given that a certain event has occurred. Usually expressed as $P(A|B)$, which is the probability of event A occurring given that event B has occurred.$P(A|B) = \frac{P(A \cap B)}{P(B)}$.
- Independent events (獨立事件): Two or more events have no effect on each other, that is,

$$\left(\forall J \neq \varnothing \subseteq \{a \mid a \in \mathbb{N} \wedge 1 \leq a \leq n\} : P\left(\bigcap_{j \in J} A_j\right) = \prod_{j \in J} P\left(A_j\right) \iff \left(\left(A_1, A_2, \ldots, A_n\right) \text{ are independent e}\right.\right.$$

- Mutually exclusive events (互斥事件): Two or more events cannot occur at the same time, that is,

$$\left(\forall J \neq \varnothing \subseteq \{a \mid a \in \mathbb{N} \wedge 1 \leq a \leq n\} : P\left(\bigcap_{j \in J} A_j\right) = 0 \text{ are mutually exclusive events.}\right)$$

- Bayes' Theorem (貝葉斯定理或貝氏定理): If $\{A_i\}_{i \in I}$ is a set of partitions of a sample space Ω, then,

$$\forall 1 \leq j \leq |I| : P(A_j|B) = \frac{P\left(A_j\right) \times P\left(B \middle| A_j\right)}{\sum_{k=1}^{|I|} P\left(A_k\right) \times P\left(B \middle| A_k\right)}$$

III Random Variable (隨機變數)

The random variable is a measurable function of which the domain is the sample space of a probability space and the range is typically a subset of real numbers. If the range of it is finite or countable infinite, it is called a discrete random variable; if the range of it is continuous, it is called a continuous random variable; if the range of it is continuous in some parts and discrete in others, it is called a mixed random variable. Given ramdom variable X and sample space Ω, $X = x$ means that event $\{\omega \in \Omega \mid X(\omega) = x\}$ occured. If a random variable X follows a probability distribution P, we write $X \sim P$.

IV Probability Function (機率函數)

i Probability Mass Function (PMF) (機率質量函數) or Probability Function of Discrete Random Variable

The probability mass function, denoted as $P(x)$, is a function of which the domain is the range of a discrete random variable and the codomain is $[0, 1]$, indicating the probability of events, such that the probability sum of the probability mass function of all elements in its domain is 1.

ii Probability Density Function (PDF) (機率分布函數) or Probability Function of Continuous Random Variable

The probability density function, denoted as $f(x)$, is a function of which the domain is the range of a continuous random variable and the codomain is $[0, 1]$, indicating the probability of events, such that the probability sum of the probability density function of all elements in its domain is 1.

V Cumulative Distribution Function (CDF) (累積分布函數)

Cumulative distribution function is a function $F : \mathbb{R} \to [0, 1]$ satisfying

$$\lim_{x\to-\infty} F(x) = 0, \quad \lim_{x\to\infty} F(x) = 1.$$

i For discrete random variables

For a discrete random variable X with probability mass function $P(x)$, the cumulative distribution function F is given by:

$$F(x) = P(X \leq x) = \sum_{k\leq x} P(X = k).$$

ii For continuous random variables

For a continuous random variable X with probability density function $f(x)$, the cumulative distribution function F is given by:

$$F(x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t.$$

VI Probability space (機率空間)

A probability space (Ω, Σ, μ) is a measure space of which Ω is the sample space of an experiment, μ is called the probability measure, indicating the probability of events, such that $\mu : \Sigma \to [0, 1]$, $\mu(\Omega) = 1$. The composition function of the random variable and the probability function of an experiment is the probability measure of that experiment.

VII (Mathematical) Expected Value/Expectation/Expectancy (期望值) or Mean (平均值)

i For discrete random variables

For a discrete random variable X with probability mass function $P(x)$, the expected value, $E[X]$ or $E(X)$, or mean μ_X, is given by:

$$E[X] = \mu_X = \sum_{x \in \text{range}(X)} x \cdot P(x).$$

ii For continuous random variables

For a continuous random variable X with probability density function $f(x)$, the expected value, $E[X]$ or $E(X)$, or mean μ_X, is given by:

$$E[X] = \mu_X = \int_{-\infty}^{\infty} x \cdot f(x)\,\mathrm{d}x.$$

VIII Variance (變異數)

i For discrete random variables

For a discrete random variable X with probability mass function $P(x)$, the variance, $\text{Var}(X)$, is given by:

$$\text{Var}(X) = {\sigma_X}^2 = E[(X - E[X])^2] = \sum_{x \in \text{range}(X)} (x - E[X])^2 \cdot P(x).$$

ii For continuous random variables

For a continuous random variable X with probability density function $f(x)$, the variance, $\text{Var}(X)$, is given by:

$$\text{Var}(X) = {\sigma_X}^2 = E[(X - E[X])^2] = \int_{-\infty}^{\infty} (x - E[X])^2 \cdot f(x)\,\mathrm{d}x.$$

IX Standard Deviation (標準差)

The standard deviation, $\text{SD}(X)$ or σ_X, is the positive square root of the variance, that is,

$$\text{SD}(X) = \sqrt{\text{Var}(X)}.$$

X Mode (眾數)

For a discrete or continuous random variable X with probability function $p(x)$, the mode, $\text{Mode}(X)$, is given by:

$$\text{Mode}(X) = \arg\max_x(p(x)).$$

XI Median (中位數)

i For all random variables

For a random variable X, the median, $\text{Median}(X)$, is an element in the range of X that satisfies:

$$P(X \leq \text{Median}(X)) \geq \frac{1}{2} \wedge P(X \geq \text{Median}(X)) \geq \frac{1}{2}.$$

ii For continuous random variables

For a continuous random variable X with probability density function $f(x)$, the median, $\text{Median}(X)$, is an element in the range of X such that the cumulative distribution function $F(x)$ of it is $\frac{1}{2}$, that is,

$$F(\text{Median}(X)) = \int_{-\infty}^{\text{Median}(X)} f(x)\,\mathrm{d}x = \frac{1}{2}.$$

XII Entropy (熵)

i For discrete random variables

For a discrete random variable X with probability mass function $P(x)$, the entropy, $H(X)$, is given by:

$$H(X) = -\sum_{i \in D_P} P(i) \log P(i).$$

ii For continuous random variables

For a continuous random variable X with probability density function $f(x)$, the entropy, $H(X)$, is given by:

$$H(X) = -\int_{-\infty}^{\infty} f(x) \log f(x)\,\mathrm{d}x.$$

XIII Affine Transformation

There are two random variable X and $Y = aX + b$ where $a, b \in \mathbb{R}$, then,

$$E[X] = aE[X] + b.$$

$$\text{Var}(Y) = a^2\text{Var}(X).$$

$$\text{SD}(Y) = |a|\text{SD}(X).$$

$$\text{Mode}(Y) = a\text{Mode}(X) + b.$$

$$\text{Median}(Y) = a\text{Median}(X) + b.$$

If X is discrete,

$$H(Y) = \begin{cases} H(X), & a \neq 0 \\ 0, & a = 0 \end{cases};$$

if X is continuous,

$$H(Y) = \begin{cases} H(X) + \log|a|, & a \neq 0 \\ 0, & a = 0 \end{cases}.$$

2 Common Discrete Distribution

I Bernoulli Trial or Binomial Trial

i Bernoulli Trial (伯努力試驗) or Binomial Trial (二項試驗)

A Bernoulli trial is a random experiment with exactly two possible outcomes, ”success” and ”failure”, in which the probability of success is the same every time the experiment is conducted.

ii Binomial Distribution

If the random variable X follows the binomial distribution with number of Bernoulli trials $n \in \mathbb{N}$ and probability of success p, we write $X \sim B(n, p)$. The probability of getting exactly k successes in n independent Bernoulli trials (with the same success probability p) is given by the probability mass function:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

$P(X = k)$ when $X \sim B(n, p)$ has the following property:

- If $(n+1)p \in \mathbb{N} \wedge p \neq 1$, the probability mass function has two modes at $k = (n+1)p$ and $k = (n+1)p - 1$; if $p = 1$, it has one mode at $k = (n+1)p - 1$; otherwise, it has one mode at $\lfloor (n+1)p \rfloor$. The probability mass function is strictly increasing for $k < (n+1)p$ and strictly decreasing for $k > (n+1)p$

  Proof.

  $$K := \arg \max_{k \in \mathbb{Z}, 0 \leq k \leq n} \left(\binom{n}{k} p^k (1-p)^{n-k} \right), \quad n \in \mathbb{N}, p > 0.$$

  If $p = 1$, it has one mode at $k = (n+1)p - 1$; otherwise, let:

  $$g(k) := \frac{P(X = k+1)}{P(X = k)} = \frac{(n-k)p}{(k+1)(1-p)}.$$

  $$g'(k) = \frac{-(n+1)p}{(k+1)^2(1-p)} \leq 0.$$

  $$K - 1 = \{\min \left(k \text{ s.t. } \frac{n-k}{k+1} > \frac{1-p}{p} \right), \min \left(k \text{ s.t. } \frac{n-k}{k+1} \geq \frac{1-p}{p} \right)\}.$$

  Solve $g(k*) = 0$ for $k*$:

  $$\frac{n - (k^* + 1)}{(k^* + 1) + 1} = \frac{1-p}{p}$$

  $$np - (k^* + 1)p = (k^* + 1) + 1 - (k^* + 1) - p.$$

  $$k^* = (n+1)p.$$

  □

- The expected value of it is np.

Proof.

$$\begin{aligned}
E[X] &= \sum_{k=1}^{n} k\binom{n}{k}p^k(1-p)^{n-k} \\
&= \sum_{k=1}^{n} n\binom{n-1}{k-1}p^k(1-p)^{n-k} \\
&= np\sum_{k=1}^{n}\binom{n-1}{k-1}p^{k-1}(1-p)^{n-k} \\
&= np(p+1-p)^{n-1} \\
&= np.
\end{aligned}$$

□

- The variance of it is $np(1-p)$.

  Proof.

$$\begin{aligned}
E[X^2] &= \sum_{k=0}^{n} k^2\binom{n}{k}p^k(1-p)^{n-k} \\
&= np\sum_{k=1}^{n} k\binom{n-1}{k-1}p^{k-1}(1-p)^{n-k} \\
&= np\sum_{k=1}^{n-1} k\binom{n-1}{k}p^{k}(1-p)^{n-1-k} + np \\
&= np(n-1)p + np \\
&= n^2p^2 - np^2 + np
\end{aligned}$$

$$\mathrm{Var}(X) = E[X^2] - (E[X])^2 = n^2p^2 - np^2 + np - n^2p^2 = np(1-p).$$

□

iii Geometric Distribution

If the random variable X follows the geometric distribution with probability of success p, we write $X \sim G(p)$ or $X \sim Geo(p)$. The number of Bernoulli trials (with the same success probability p) needed to get one success is given by the probability mass function:

$$P(X = k) = p(1-p)^{k-1}.$$

$P(X = k)$ when $X \sim G(p)$ has the following property:

- The probability mass function is strictly decreasing.

  Proof.

$$g(k) := p(1-p)^{k-1}.$$

$$g'(k) = p\ln(1-p)(1-p)^{k-1} < 0.$$

□

- $k > j \in \mathbb{N}$,

$$P(X = k|X > j) = P(X = k - j).$$

Proof.

$$\frac{p(1-p)^k}{(1-p)^j} = p(1-p)^{k-j}.$$

□

- The expected value of it is $\frac{1}{p}$.

  Proof.

$$E[X] = \sum_{k=1}^{\infty} kp(1-p)^{k-1} = \sum_{k=1}^{\infty}(k-1)p(1-p)^{k-1} + \sum_{k=1}^{\infty} p(1-p)^{k-1}$$

$$(1-p)E[X] = \sum_{k=1}^{\infty} kp(1-p)^k$$

$$E[X] - (1-p)E[X] = pE[X] = \sum_{k=1}^{\infty} p(1-p)^{k-1} = p\frac{1}{1-(1-p)} = 1.$$

□

- The variance of it is $\frac{1-p}{p^2}$.

  Proof.

$$E[X^2] = \sum_{k=1}^{\infty} k^2 p(1-p)^{k-1} = \sum_{k=1}^{\infty}(k-1)^2 p(1-p)^{k-1} + 2\sum_{k=1}^{\infty} kp(1-p)^{k-1} - \sum_{k=1}^{\infty} p(1-p)^{k-1}.$$

$$(1-p)E[X^2] = \sum_{k=1}^{\infty} k^2 p(1-p)^k.$$

$$E[X^2] - 2E[X] + 1 - (1-p)E[X^2] = pE[X^2] - 2E[X] + 1 = 0.$$

$$E[X^2] = \frac{2E[X]-1}{p} = \frac{2-p}{p^2}.$$

$$\text{Var}(X) = E[X^2] - (E[X])^2 = \frac{1-p}{p^2}.$$

□

3 Common Continuous Distribution

I Normal distribution (常態分布/正態分布)/Gaussian distribution (高斯分布)

i Probability density function

The general normal distribution or Gaussian distribution given by mean μ and standard deviation σ is a continuous probability distribution for a real-valued random variable defined by the probability density function:

$$f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}.$$

ii Cumulative distribution function

$$F(x|\mu,\sigma^2) = \frac{1}{2}\left(1 + \operatorname{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right),$$

where erf is the error function, which is a $\mathbf{C} \to \mathbf{C}$ function defined by:

$$\operatorname{erf}(z) = \frac{2}{\sqrt{\pi}}\int_0^z e^{-t^2}\,\mathrm{d}t.$$

iii Mode

$$\operatorname{Mode}(x) = \mu$$

iv Median

$$\operatorname{Median}(x) = \mu$$

v Entropy

$$H(x) = \frac{1}{2}\log\left(2\pi e\sigma^2\right).$$

vi Notation

If the random variable X follows the general normal distribution given by mean μ and standard deviation σ, we write $X \sim \mathcal{N}(\mu,\sigma^2)$.

vii Standard (標準)/Unit (単位) Normal Distribution

The standard normal distribution or unit normal distribution is the normal distribution given by mean $\mu = 0$ and standard deviation $\sigma = 1$.

- The probability density function of standard normal distribution, $\varphi(z)$, is

$$\varphi(z) = \frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}.$$

- The cumulative distribution function of standard normal distribution, $\Phi(z)$, is

$$\Phi(z) = \frac{1}{2}\left(1 + \operatorname{erf}\left(\frac{z}{\sqrt{2}}\right)\right).$$

- PDF of general normal distribution:

$$f(x|\mu,\sigma^2) = \frac{1}{\sigma}\varphi\left(\frac{x-\mu}{\sigma}\right).$$

- CDF of general normal distribution:

$$F(x|\mu,\sigma^2) = \Phi\left(\frac{x-\mu}{\sigma}\right).$$

4 Discrete-Time Markov Chain (DTMC) (離散時間馬可夫鏈) and Discrete-Time Markov Decision Process (DTMDP) (離散時間馬可夫決定過程)

I Stochastic matrix (隨機矩陣), transition matrix (轉移矩陣), probability matrix (機率矩陣), transition probability matrix (轉移機率矩陣), substitution matrix (替代矩陣), or Markov matrix (馬可夫矩陣)

- (Left) stochastic matrix, (left) transition matrix, (left) probability matrix, (left) transition probability matrix, (left) substitution matrix, or (left) Markov matrix: A square matrix in which the sum of each row is 1 and each element is greater than or equal to zero.
- Right stochastic matrix, right transition matrix, right probability matrix, right transition probability matrix, right substitution matrix, or right Markov matrix: A square matrix in which the sum of each column is 1 and each element is greater than or equal to zero.
- Doubly stochastic matrix, doubly transition matrix, doubly probability matrix, doubly transition probability matrix, doubly substitution matrix, or doubly Markov matrix: A matrix that is both a left stochastic matrix and a right stochastic matrix.
- Closed property (封閉性) of left stochastic matrix: The matrix product of several left stochastic matrices, the arithmetic mean of several left stochastic matrices, and the positive integer power of a left stochastic matrix are all stochastic matrices.
- Closed property of right stochastic matrix: The matrix product of several right stochastic matrices, the arithmetic mean of several right stochastic matrices, and the positive integer power of a right stochastic matrix are all stochastic matrices.
- Regularity: A left or right stochastic matrix $\mathbf{A}$ is regular if there exists at least one $n \in \mathbb{N}_0$ such that every entries in $\mathbf{A}^n$ is positive.
- Singularity: A left or right stochastic matrix is singular if it is not regular.

II Discrete-time Markov Chain (DTMC)

- State s: The state of the agent with respect to the environment.
- State space S: The set of all possible states.
- State transition: The transition from a state to next state.
- State transition probability $p(s'|s)$: A probability mass function that defines the likelihood of an agent transitioning to s' from s.
- (State) transition (probability) matrix P: Suppose the states could be indexed as $s_i \left(_{i=1}^{n}\right)$. The state transition matrix is defined to be:

$$P \in [0,1]^{n \times n} \wedge P_{ij} := p(s_j|s_i).$$

It satisfies:

$$\forall i \in \mathbb{N} \leq n : \sum_{j=1}^{n} P_{ij} = 1$$

- Trajectory: A finite or infinite state chain that an agent can take. In a trajectory, the ith state is called s_{i-1}, making the trajectory $s_0 s_1 \dots s_{i-1} s_i \dots$.
- System model, Transition model, or Model: the state transition probability of each state of a DTMC.
- Discrete-time Markov chain (DTMC) $(S, p(s'|s))$: A stochastic process describing a sequence of possible events in which the probability of each event depends only on the current state in the previous event. A DTMC is given by a two-tuple of the state space S and state transition probability $p(s'|s)$, that the following property, called Markov property (馬可夫性) or memoryless property (無記憶性或無後效性), holds:

$$p(s_{t+1}|s_t, s_{t-1}, \dots, s_0) = p(s_{t+1}|s_t).$$

- Irreducibility (不可約性): a DTMC is irreducible if for any two states s_i and s_j, there exists at least one positive integer k such that $(P^k)_{ij} > 0$. A DTMC is irreducible if and only if its state transition matrix P is a regular stochastic matrix.
- Period (週期): the period of a state s_i in S is the greatest common divisor of all positive integer k such that $(P^k)_{ii} > 0$.
- Aperiodicity (非週期性): a state is aperiodic if its period is 1; a DTMC is aperiodic if all states in its state space is aperiodic.
- Steady state (穩定狀態或穩態): A probability distribution π over S such that $\pi P = \pi$. If a DTMC is irreducible and aperiodic, then it has a unique steady-state distribution π, and for any probability distribution π' over S:

$$\lim_{k \to \infty} \pi' P^k = \pi$$

III Discrete-time Markov decision process (DTMDP)

- State s: The state of the agent with respect to the environment.
- State space S: The set of all possible states.
- State transition: The transition from a state to next state.
- Action: A choice the agent can make to interact with the environment, changing its state.
- Action space of a state A_s: The set of all possible actions of a state s.
- (State) transition (probability) matrix (of an action) $p(s'|s, a)$: A probability mass function that defines the likelihood of an agent transitioning to s' given that the agent takes an action a in a state s.

- Policy $\pi(a|s)$: A policy is a probability function from the state space to the action spaces an agent follows to select actions based on its current state. It defines the conditional probability of the agent taking action a when in state s.
- Deterministic policy: A policy is deterministic if, for each state s, there exists exactly one action a such that $\pi(a|s) = 1$ and $\pi(a'|s) = 0$ for all other actions $a' \neq a$.
- Stochastic policy: A policy that is not deterministic.
- State transition probability (given a policy) $p_\pi(s'|s)$:

$$p_\pi(s'|s) := \sum_a \pi(a|s)p(s'|s,a).$$

- State transition matrix (given a policy) P_π: Suppose the states could be indexed as $s_i \left(\begin{smallmatrix} n \\ i=1 \end{smallmatrix}\right)$. The state transition matrix of a policy π is defined to be:

$$P_\pi \in [0,1]^{n\times n} \wedge (P_\pi)_{ij} := p_\pi(s_j|s_i).$$

  It satisfies:

$$\forall i \in \mathbb{N} \leq n : \sum_{j=1}^{n} P_{ij} = 1$$

- Reward r: A real number the agent gets after taking a action.
- Reward transition probability $p(r|s,a)$: A probability mass function that defines the likelihood of an agent receiving reward r given that the agent takes an action a in a state s.
- Reward (given a policy) $r_\pi(s)$:

$$r_\pi(s) := E[r|S_t = s, A_t = a] = \sum_a \pi(a|s) \sum_r p(r|s,a)r.$$

- Reward vector (given a policy) r_π: Suppose the states could be indexed as $s_i \left(\begin{smallmatrix} n \\ i=1 \end{smallmatrix}\right)$. The reward vector of a policy π is defined to be:

$$r_\pi := [r_\pi(s_i) \left(\begin{smallmatrix} n \\ i=1 \end{smallmatrix}\right)]^\top \in \mathbb{R}^n.$$

- Trajectory: A finite or infinite state-action-reward chain that an agent can take by taking a chain of actions in the action space of the state it's in, moving along a chain of states in the state space, and receiving rewards along the way. In a trajectory, the ith state is called s_{i-1}, the ith action taken is called a_{i-1}, the ith reward received is called r_i, making the trajectory $s_0 \xrightarrow[r_1]{a_0} s_1 \dots s_{i-1} \xrightarrow[r_i]{a_{i-1}} s_i \dots$.
- Return: The sum of all rewards the agent receives along a trajectory.
- Discounted return G_t: The discounted reward G_t at step t given that the discount rate is $\gamma \in [0,1)$, the reward at step i is r_i, and the final step in the trajectory is the step $t+n$ ($n = \infty$ for infinite trajectory), is defined to be:

$$G_t := \sum_{i=0}^{n} \gamma^i r_{t+i}.$$

- Terminal state: The state that the agent is in after its last action in a finite trajectory.
- Episode or Trial: A trajectory with a terminal state.
- Episodic task: A task with a terminal state.
- Continuing task: A task without a terminal state.
- Target state: The terminal state in a finite trajectory, or the state that the agent stays in since a specific action is taken and that the agent takes a same action that doesn't change its state afterwards in an infinite trajectory. Not all infinite trajectories have a target state.
- Absorbing state: A target state in an infinite trajectory that any action of the agent after it yields zero reward. Not all infinite trajectories have an absorbing state.
- System model, Transition model, or Model: the state transition probability of each action in the action space of each state, and the reward transition probability of each action in the action space of each state of a DTMDP.
- Discrete-time Markov decision process (DTMDP) $(S, A_s, p(s'|s,a), p(r|s,a))$: A DTMDP is given by a four-tuple of the state space S, the action spaces A_s of each state s, the state transition probability $p(s'|s,a)$ of each action a in the action space of each state s, and the reward transition probability $p(r|s,a)$ of each action a in the action space of each state s, that the following property, called Markov property or memoryless property, holds:

  $$p(S_{t+1}|A_t, S_t, A_{t-1}, S_{t-1}, \dots, A_0, S_0) = p(S_{t+1}|A_t, S_t),$$
  $$p(R_{t+1}|A_t, S_t, A_{t-1}, S_{t-1}, \dots, A_0, S_0) = p(R_{t+1}|A_t, S_t).$$

  If a policy is given, a DTMDP becomes a DTMC.
- State value $v_\pi(s)$: The state value, a function of state s given the policy π, is the expected value of the discounted return G_t given $S_t = s$, that is,

  $$v_\pi(s) := E[G_t|S_t = s].$$

- State value vector v_π: Suppose the states could be indexed as $s_i \left(_{i=1}^{n}\right)$. The state value vector of a policy π is defined to be:

  $$v_\pi := [v_\pi(s_i) \left(_{i=1}^{n}\right)]^\top \in \mathbb{R}^n.$$

- Policy evaluation: Given a policy, finding out the corresponding state values of all states is called policy evaluation.
- Action value or Q-value $q_\pi(s,a)$ or $Q^\pi(s,a)$: The action value, a function of state-action pair (s,a) given the policy π, is the expected value of the discounted return G_t given $S_t = s$ and $A_t = a$, that is,

  $$q_\pi(s,a) := E[G_t|S_t = s, A_t = a] = E[R_{t+1} + \gamma v_\pi(S_{t+1})|S_t = s, A_t = a].$$

- Optimal Policy π^*: Given two policies π_1 and π_2, if

  $$v_{\pi_1}(s) \geq v_{\pi_2}(s), \quad \forall s \in S,$$

then we say π_1 is "better" than π_2.

A policy π^* is optimal if for any other policy π

$$v_{\pi^*}(s) \geq v_\pi(s), \quad \forall s \in S.$$

Given a DTMDP, there must exist an optimal policy, but it is not necessarily unique.

IV Bellman Equation

i Bellman Equation Elementwise Form

The elementwise form of the Bellman equation of a given policy π is

$$\begin{aligned}
v_\pi(s) =& E[G_t|S_t = s] \\
=& E[R_{t+1}|S_t = s] + \gamma E[G_{t+1}|S_t = s] \\
=& \sum_a \pi(a|s) \sum_r p(r|s,a)r + \gamma \sum_a \pi(a|s) \sum_{s'} p(s'|s,a)v_\pi(s') \\
=& \sum_a \pi(a|s) \left(\sum_r p(r|s,a)r + \gamma \sum_{s'} p(s'|s,a)v_\pi(s') \right) \\
=& r_\pi(s) + \gamma \sum_{s'} p_\pi(s'|s)v_\pi(s'), \quad \forall s \in S.
\end{aligned}$$

ii Bellman Equation Matrix-Vector Form

The matrix-vector form of the Bellman equation of a given policy π is

$$v_\pi = r_\pi + \gamma P_\pi v_\pi.$$

iii Bellman Equation Closed Form Solution

$$v_\pi = (I - \gamma P_\pi)^{-1} r_\pi.$$

iv Bellman Equation Iterative Solution Matrix-Vector Form

Consider a sequence $\{v_k\}$ where v_0 is any arbitrary vector $\in \mathbb{R}^{|S|}$, and

$$v_k = r_\pi + \gamma P_\pi v_{k-1}, \quad k \in \mathbb{N},$$

then

$$v_\pi = \lim_{k \to \infty} v_k.$$

In practice, we usually stop when $\|v_k - v_{k-1}\|$ is sufficiently small or when k is sufficiently large.

v Bellman Equation Iterative Solution Elementwise Form

Consider a sequence $\{v_k(s)\}$ where $v_0(s)$ is any arbitrary value, and

$$v_k = \sum_a \pi(a|s) \left(\sum_r p(r|s,a)r + \gamma \sum_{s'} p(s'|s,a)v_{k-1}(s) \right), \quad k \in \mathbb{N},$$

then

$$v_\pi(s) = \lim_{k \to \infty} v_k.$$

In practice, we usually stop when $|v_k(s) - v_{k-1}(s)|$ is sufficiently small or when k is sufficiently large.

vi Value Improvement Theorem

Consider a sequence $\{v_k\}$ where v_0 is any arbitrary vector $\in \mathbb{R}^{|S|}$, and

$$v_k = r_\pi + \gamma P_\pi v_{k-1}, \quad k \in \mathbb{N},$$

then

$$v_{k+1} \geq v_k.$$

V State-Action Value Function or Value Function (VF)

Compare

$$v_\pi(s) = \sum_a \pi(a|s) q_\pi(s,a)$$

and the Bellman equation, we have the action-value function:

$$q_\pi(s,a) = \sum_r p(r|s,a) r + \gamma \sum_{s'} p(s'|s,a) v_\pi(s').$$

VI Contraction Mapping Theorem

i Fixed Point

Given $f: X \to X$, $x \in X$ is a fixed point if

$$f(x) = x.$$

ii Contraction Mapping or Contractive Function

$f: X \to X$ is a contraction mapping if

$$\exists \gamma \in [0,1): \ \|f(x_1) - f(x_2)\| \leq \gamma \|x_1 - x_2\|, \quad \forall x_1, x_2 \in X,$$

where $\|\cdot\|$ can be any vector norm.

iii Contraction Mapping Theorem

For any contraction mapping,

- Existence: there exists a fixed point x^* satisfying $f(x^*) = x^*$.
- Uniqueness: the fixed point x^* is unique.
- Algorithm: Consider a sequence $\{x_k\}$ where x_0 is any arbitrary value and $x_{k+1} = f(x_k), \quad k \in \mathbb{N}$, then

  $$\lim_{k \to \infty} x_k = x^*.$$

  Moreover, the convergence rate is exponential and determined by γ.

VII Bellman Optimality Equation (BOE)

i Bellman Optimality Equation Elementwise Form

$$v(s) = \max_{\pi} \left(\sum_{a} \pi(a|s) q(s,a) \right), \quad s \in S.$$

ii Bellman Optimality Equation Matrix-Vector Form

$$v = \max_{\pi} \left(r_\pi + \gamma P_\pi v_\pi \right).$$

iii Value Iteration (VI) Matrix-Vector Form

Let

$$f(v) = \max_{\pi} \left(r_\pi + \gamma P_\pi v \right),$$

where v is a state value given policy π.
Because $f(v)$ is a contraction mapping, it satisfies the contraction mapping theorem, that is,

- Existence and uniqueness:

  $$\exists! v^* \text{ such that } v^* = f(v^*),$$

- Iterative algorithm: Consider a sequence $\{v_k\}$ where v_0 is any arbitrary value, and $v_k = f(v_{k-1}), \quad k \in \mathbb{N}$. It converges to v^* in an exponential rate determined by γ as k approaching ∞.

One iteration in the value iteration algorithm,

$$v_{k+1} = f(v_k) = \max_{\pi} \left(r_\pi + \gamma P_\pi v_k \right), \quad k+1 \in \mathbb{N},$$

can be decomposed into two steps,

1. **Policy update (PU)**: Solve

   $$\pi_{k+1} = \arg\max_{\pi} (r_\pi + \gamma P_\pi v_k)$$

   for π_{k+1} given v_k.

2. **Value update (VU)**: Solve

   $$v_{k+1} = r_{\pi_{k+1}} + \gamma P_{\pi_{k+1}} v_k),$$

   for v_{k+1} given π_{k+1} and v_k.

In practice, we usually stop when $\|v_k - v_{k-1}\|$ is sufficiently small or when k is sufficiently large.

iv Value Iteration (VI) Elementwise Form

In elementwise form, the two steps of value iteration can be written as,

1. **Policy update (PU)**: Solve

   $$\pi_{k+1}(s) = \arg\max_{\pi} \sum_{a} \pi(a|s) q_{\pi_k}(s,a), \quad s \in S$$

for $\pi_{k+1}(s)$ given $v_k(s')$ for all $s' \in S$.
Let $a_k^*(s) = \arg\max_a q_{\pi_k}(s,a)$. We select

$$\pi_{k+1}(a|s) = \begin{cases} 1, & a = a_k^*(s) \\ 0, & a \neq a_k^*(s) \end{cases},$$

called "greedy policy" because it simply selects the greatest policy value.

2. **Value update (VU)**: Solve

$$v_{k+1}(s) = \sum_a \pi_{k+1}(a|s) q_{\pi_k}(s,a), \quad s \in S$$

for $v_{k+1}(s)$ given $\pi_{k+1}(a|s)$ for all $a \in A_s$ for all $s \in S$, and $v_k(s')$ for all $s' \in S$.
Since $\pi_{k+1}(a|s)$ is greedy,

$$v_{k+1}(s) = \max_a q_{\pi_k}(s,a).$$

In practice, we usually stop when $|v_k(s) - v_{k-1}(s)|$ is sufficiently small or when k is sufficiently large.

v Optimality Theorem

Suppose v^* is the solution to a Bellman optimality equation, that is,

$$v^* = \max_\pi \left(r_\pi + \gamma P_\pi v^* \right).$$

Suppose

$$\pi^* = \arg\max_\pi \left(r_\pi + \gamma P_\pi v^* \right).$$

Then

$$v^* = r_{\pi^*} + \gamma P_{\pi^*} v^*.$$

v^* is the optimal state value, and π^* is the optimal policy.

vi Optimal Policy Invariance Theorem

Consider a Markov decision process with $v^* \in \mathbb{R}^{|S|}$ as the optimal state value satisfying $v^* = \max_\pi(r_\pi + \gamma P_\pi v^*$. If every reward r is changed by an affine transformation to $ar + b$, where $a, b \in \mathbb{R}$ and $a \neq 0$, then the corresponding optimal state value v' is also an affine transformation of v^*:

$$v' = av^* + \frac{b}{1-\gamma}\mathbf{1},$$

where $\gamma \in [0,1)$ is the discount rate and $\mathbf{1} = [1, \ldots, 1]^\top \in \mathbb{R}^{|S|}$.
Consequently, the optimal policies are invariant to any affine transformation of the reward signals.

vii Policy Improvement Theorem

If

$$\pi_{k+1} = \arg\max_\pi(r_\pi + \gamma P_\pi v_{\pi_k}),$$

then $\|v_{\pi_{k+1}}\| \geq \|v_{\pi_k}\|$ for any k.

viii Convergence of Policy Iteration Theorem

The state value sequence $\{v_{\pi_k}\}_{k=0}^{\infty}$ generated by the policy iteration algorithm converges to the optimal state value v^*. Consequently, the policy sequc $\{\pi_k\}_{k=0}^{\infty}$ coverages to an optimal policy.

ix Policy Iteration (PI) Matrix-Vector Form

An arbitrary initial policy π_0 is given. One iteration in the policy iteration algorithm can be decomposed into two steps,

1. **Policy evaluation (PE)**: Solve the Bellman equation

$$v_{\pi_k} = r_{\pi_k} + \gamma P_{\pi_k} v_{\pi_k}$$

   for v_{π_k} given π_k.

2. **Policy improvement (PI)**: Solve

$$\pi_{k+1} = \arg\max_{\pi}(r_{\pi} + \gamma P_{\pi} v_{\pi_k})$$

   for π_{k+1} given v_{π_k}.

In practice, we usually stop when $\|v_{\pi_k} - v_{\pi_{k-1}}\|$ is sufficiently small or when k is sufficiently large.

x Policy Iteration (PI) Elementwise Form

In elementwise form, the two steps of policy iteration can be written as,

1. **Policy evaluation (PE)**: Solve the Bellman equation

$$v_{\pi_k}(s) = \text{☼}_a \pi(a|s) q_{\pi_k}(s,a), \quad s \in S$$

   for v_{π_k} given π_k.

2. **Policy improvement (PI)**:
   Solve

$$\pi_{k+1}(s) = \arg\max_{\pi} \sum_a \pi(a|s) q_{\pi_k}(s,a), \quad s \in S$$

   for $\pi_{k+1}(s)$ given $v_{\pi_k}(s)$.
   Let $a_k^*(s) = \arg\max_a q_{\pi_k}(s,a)$. We select

$$\pi_{k+1}(a|s) = \begin{cases} 1, & a = a_k^*(s) \\ 0, & a \neq a_k^*(s) \end{cases},$$

   called ”greedy policy” because it simply selects the greatest policy value.

In practice, we usually stop when $|v_{\pi_k}(s) - v_{\pi_{k-1}}(s)|$ is sufficiently small or when k is sufficiently large.

xi Truncated Policy Iteration

The truncated policy iteration is the same as policy iteration with the policy evaluation step using the iterative solution but stopped when $\|v_k - v_{k-1}\|$ is sufficiently small or when k is sufficiently large. If stopping when $k = 1$, the truncated policy iteration becomes value iteration except that the first iteration lacks value update and is initialized with an arbitrary policy; if stopping when $k = \infty$, the truncated policy iteration becomes policy iteration.