Contents

1

Probability

1.1 Probability Spaces

In an elementary level, we have been viewing probability as the quotient between the number of desired outcomes and the number of all possible outcomes. This definition, though intuitive, is not very solid when it comes to an infinite sample space. In this introductory chapter, we would establish the theories of probability using a more modern and rigorous structure.

Suppose we perform an experiment. This experiment might have many possible outcomes, but we are interested in only one or some of them. This leads to the following notions:

Definition 1.1.1 ▸ Sample Space and Events

A **sample space** of some experiment is the set of all possible outcomes of the experiment. An **event** is a subset of the sample space.

In a naïve attempt to devise a probability model, if the sample space S is countable, then it suffices to define a *probability mass function* $P : S \to \mathbb{R}$ such that $\sum_{\omega \in S} P(\omega) = 1$. Naturally, the probability for an event $E \subseteq S$ is defined as $P(E) = \sum_{\omega \in E} P(\omega)$. This summation is compatible with the infinite case because if we have countably many pairwise disjoint events $E_1, E_2, \cdots$, we can compute

$$P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i) = \lim_{n\to\infty} \sum_{i=1}^{n} E_i,$$

which is clearly convergent by monotone-convergent theorem.

However, when S is uncountable, this construction leads to weird behaviours. For example, suppose S is the sample space for the experiment of tossing a fair coin for uncountably many times. It is clear that for any $\omega \in S$, we have

$$P(\omega) = \lim_{n\to\infty} \frac{1}{2^n} = 0,$$

but at the same time we must have

$$1 = P(S) = \sum_{\omega \in S} P(\omega) = 0,$$

which is ridiculous. Therefore, we need to find a better way to construct the probability model. Notice that here the incompatibility arises because we build our model by considering the probabilities of individual outcomes. Our next attempt try to bypass this issue by considering the probabilities of events only.

Since the set of all events in a sample space S is simply $\mathcal{P}(S)$, let us instead consider a more generalisable algebraic structure for this collection of subsets.

Definition 1.1.2 ▸ Set Algebra

Let X be a set. A **set algebra** over X is a family $\mathcal{F} \subseteq \mathcal{P}(X)$ such that

- $X \backslash F \in \mathcal{F}$ for all $F \in \mathcal{F}$ (closed under complementation);
- $X \in \mathcal{F}$;
- $X_1 \cup X_2 \in \mathcal{F}$ for any $X_1, X_2 \in \mathcal{F}$ (closed under binary union).

There are several immediate implications from the above definition.

First, by closure under complementation, we know that an algebra over any set X must contain the empty set.

Second, by De Morgan's Law, one can easily check that if the first 2 axioms hold, the closure under binary union is equivalent to

- $X_1 \cap X_2 \in \mathcal{F}$ for any $X_1, X_2 \in \mathcal{F}$;
- $\bigcup_{i=1}^{n} X_i \in \mathcal{F}$ for any $X_1, X_2, \cdots, X_n \in \mathcal{F}$ for all $n \in \mathbb{N}$;
- $\bigcap_{i=1}^{n} X_i \in \mathcal{F}$ for any $X_1, X_2, \cdots, X_n \in \mathcal{F}$ for all $n \in \mathbb{N}$.

$(X, \mathcal{F})$ is known as a *field of sets*, where the elements of X are called *points* and those of $\mathcal{F}$, *complexes* or *admissible sets* of X.

In probability theory, what we are interested in is a special type of set algebras known as σ-*algebras*.

Definition 1.1.3 ▸ σ-Algebra

A **σ-Algebra** over a set A is a non-empty set algebra over A that is closed under countable union.

Of course, by the same argument as above, we known that any σ-algebra is closed under countable intersection as well.

Roughly speaking, we could now define the probability of an event $E \subseteq S$ as the ratio of the size of E to that of S. The remaining question now is: how do we define the size of a set (and in particular, an infinite set) properly?

Definition 1.1.4 ▸ Measure

Let X be a set and Σ be a σ-algebra over X. A **measure** over Σ is a function

$$\mu : \Sigma \to \mathbb{R} \cup \{-\infty, +\infty\}$$

such that

- $\mu(E) \geq 0$ for all $E \in \Sigma$ (non-negativity);
- $\mu(\varnothing) = 0$;
- $\mu\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} \mu(E_i)$ for any countable collection of pairwise disjoint elements of Σ (countable additivity or σ-additivity).

The triple (X, Σ, μ) is known as a **measure space** and the pair (X, Σ), a **measurable space**.

One thing to note here is that if at least one $E \in \Sigma$ has a finite measure, then $\mu(\varnothing) = 0$ is automatically guaranteed for obvious reasons.

Definition 1.1.5 ▸ Probability Measure

Let $\mathcal{F}$ be a σ-algebra over a sample space S. A **probability measure** on S is a measure $P : \mathcal{F} \to [0, 1]$ such that $P(S) = 1$.

Obviously, the above definition immediately guarantees that

1. $P(A^c) = 1 - P(A)$;
2. $P(A) \leq P(B)$ if $P(A) \subseteq P(A)$;
3. $P(A \cup B) \leq P(A) + P(B)$.

The third result follows from a direct application of the principle of inclusion and exclusion. By induction, one can easily check that

$$P\left(\bigcup_{i=1}^{n} E_i\right) \leq \sum_{i=1}^{n} P(E_i)$$

for any finitely many events. The following proposition extends this result to countable collections of events:

Proposition 1.1.6 ▸ Union Bound of Countable Collections of Events

Let $(S, \mathcal{F}, P)$ be a probability space and $E_1, E_2, \cdots, E_n, \cdots \in \mathcal{F}$ is any countable sequence of events, then

$$P\left(\bigcup_{i=1}^{\infty} E_i\right) \leq \sum_{i=1}^{\infty} P(E_i).$$

Proof. Define $F_1 \coloneqq E_1$ and $F_k \coloneqq E_k \setminus \bigcup_{i=1}^{k-1} E_i$ for $k \geq 2$. Clearly, the F_i's are pairwise disjoint. By Definition 1.1.3, the F_i's are elements of $\mathcal{F}$. Note that $P(F_i) \leq \mathbb{E}_{\mathbb{i}}$ for all $i \in \mathbb{N}^+$, so

$$\begin{aligned} P\left(\bigcup_{i=1}^{\infty} E_i\right) &= P\left(\bigcup_{i=1}^{\infty} F_i\right) \\ &= \sum_{i=1}^{\infty} P(F_i) \\ &\leq \sum_{i=1}^{\infty} P(E_i). \end{aligned}$$

□

Intuitively, the equality is attained if and only if the events are pairwise disjoint.

1.2 Conditional Probability

Suppose E and F are events in the same sample space. We should reassess the probability of E given that F has occurred because now we have gained some new information which could alter our prediction for future events.

Notice that by given the condition on the occurrence of F, we have effectively reduced the sample space to F and the event to $E \cap F$.

Definition 1.2.1 ▸ Conditional Probability

Let P be a probability measure on a sample space S. For any events $E, F \subseteq S$, the **conditional probability** of E given F is defined as

$$P(E \mid F) = \frac{P(E \cap F)}{P(F)}.$$

Clearly, the above definition is equivalent to $P(E \cap F) = P(F)P(E \mid F)$, which is natural in a sense that if we wish both E and F to happen, we just need F to happen first and E to happen given the occurrence of F. This can be generalised into the following result:

Theorem 1.2.2 ▶ Law of Total Probabilities

Let $F_1, F_2, \cdots, F_n$ be a partition of a sample space S with probability measure P. For any event $A \subseteq S$,

$$P(A) = \sum_{i=1}^{n} P(A \mid F_i) P(F_i).$$

We can generalise the formula in Definition 1.2.1 into any finite number of events.

Proposition 1.2.3 ▶ Generalised Formula for Conditional Probability

Let $E_1, E_2, \cdots, E_n$ be events in a sample space S with probability measure P, then

$$P\left(\bigcap_{i=1}^{n} E_i\right) = P(E_1) \prod_{i=1}^{n-1} P(E_{i+1} \mid E_1, \cdots, E_i).$$

Additionally, recall that the Bayes' theorem states the following:

Theorem 1.2.4 ▶ Bayes' Theorem

Let A and B be events in a sample space with probability measure P, then

$$P(A \mid B) = \frac{P(B \mid A) P(A)}{P(B)}.$$

Note that it is not necessary that $P(E \mid F) < P(E)$. In some cases, the occurrence of F does not affect the occurrence of E.

Definition 1.2.5 ▶ Independent Events

Let S be a sample space with probability measure P. Two events $E, F \subseteq S$ are **independent** if $P(E \mid F) = P(E)$, or equivalently, $P(E \cap F) = P(E) P(F)$. A collection of events $E_1, E_2, \cdots, E_n$ are said to be **jointly independent** if for any $I \subseteq \{1, 2, \cdots, n\}$,

$$P\left(\bigcap_{i \in I} E_i\right) = \prod_{i \in I} P(E_i),$$

or equivalently, $P(E_1 \mid E_2, \cdots, E_n) = P(E_1)$.

Note that E and F are independent if and only if E, F, E^c, F^c are pairwise independent. Moreover, for any jointly independent collection of events $E_1, E_2, \cdots, E_n$ and any disjoint

index sets $I, J \subseteq \{1, 2, \cdots, n\}$,

$$P\left(\bigcap_{i\in I} E_i \cap \bigcap_{j\in J} E_j^c\right) = \left(\prod_{i\in I} P(E_i)\right)\left(\prod_{j\in J} P\left(E_j^c\right)\right).$$

Remark. Joint independence is a strictly stronger result than pairwise independence, i.e., there exists pairwise independent events E_1, E_2, E_3 such that

$$P(E_1 \cap E_2 \cap E_3) \neq P(E_1)P(E_2)P(E_3).$$

1.3 Random Variables

A random variable can be viewed as a function that maps the outcomes in a sample space to some measurable co-domain. We first introduce a few preliminary definitions.

Definition 1.3.1 ▶ Probability Space

A **probability space** is a tuple $(S, \mathcal{F}, P)$ where S is a sample space, $\mathcal{F}$ is a σ-algebra on S and P is a probability measure on S.

It can be troublesome to consider different sample spaces for different experiments. Therefore, we define the *abstract probability space* as $(\Omega, \mathcal{F}, P)$ with a uniform random variable $Z : \Omega \to \mathcal{Z}$ such that for every random variable $X : S \to \mathcal{X}$, there exists a function $f : \mathcal{Z} \to \mathcal{X}$ such that X is simply $f(Z)$ restricted to S. For convenience, we often choose $\Omega = [0, 1]$ and P to be the uniform measure on $[0, 1]$.

One important property of a probability space is **countable additivity**. Specifically, if P is a probability measure on a sample space Ω and $A_1 \subseteq A_2 \subseteq \cdots \subseteq \Omega$, then

$$P\left(\bigcup_{i\in\mathbb{N}} A_i\right) = \lim_{n\to\infty} A_n.$$

On the other hand, if $\Omega \supseteq A_1 \supseteq A_2 \supseteq \cdots$, then

$$P\left(\bigcap_{i\in\mathbb{N}} A_i\right) = \lim_{n\to\infty} A_n.$$

It is important that the co-domain of a random variable is measurable. For this purpose, we construct some structure to generalise open intervals in $\mathbb{R}$.

Definition 1.3.2 ▸ Borel Algebra

Let X be a topological space. A **Borel set** on X is a set which can be formed via countable union, countable intersection and relative complementation of open sets in X. The smallest σ-algebra over X containing all Borel sets on X is known as the **Borel algebra** over X.

Note that the Borel set over $\mathbb{R}$ is just the family of all open intervals.

Clearly, the Borel algebra over X contains all open sets in X according to the above axioms from Definition 1.1.3. This helps us define the following:

Definition 1.3.3 ▸ Random Variable

Let $(\Omega, \mathcal{F}, P)$ be the abstract probability space and $(\mathcal{X}, \mathcal{B})$ be a measurable space where $\mathcal{B}$ is the Borel algebra over $\mathcal{X}$. A **random variable** is a function $X : S \to \mathcal{X}$ such that

$$\{\omega \in S : X(\omega) \in B\} \in \mathcal{F}$$

for all $B \in \mathcal{B}$. The probability measure P_X on $\mathcal{X}$ induced by P with

$$P_X(A) \coloneqq P(X \in A) \coloneqq P\big(\{\omega \in \Omega : X(\omega) \in A\}\big)$$

is known as the **distribution** of X.

Remark. Rigorously, such a random variable X is a *measurable function* or *measurable mapping* from $(S, \mathcal{F})$ to $(\mathcal{X}, \mathcal{B})$.

A random variable describes the random outcome of an "experiment" or "phenomenon". When we perform a series of experiments (for example, model the weather for n consecutive days), it is reasonable to use one random variable to capture the outcome for each experiment. In this way, we obtain a collection of random variables denoted as

$$X_i^n \coloneqq (X_1, X_2, \cdots, X_n).$$

Clearly, if X is a real-valued random variable, we have $\{\omega \in S : X(\omega) > x\} \in \mathcal{F}$. Moreover, we claim that

$$\{\omega \in S : X(\omega) < x\} = \bigcup_{y<x} \{\omega \in S : X(\omega) \leq y\}.$$

The proof is quite straightforward and is left to the reader as an exercise. By Definition 1.1.3, this means that

$$\{\omega \in S : X(\omega) < x\} \cup \{\omega \in S : X(\omega) > x\} \in \mathcal{F}.$$

Therefore, $\{\omega \in S : X(\omega) = x\} \in \mathcal{F}$. This argument justifies the probabilities $P(X < x)$ and $P(X = x)$.

When a collection of many random variables is concerned, we may consider their *joint distribution*.

Definition 1.3.4 ▸ Joint Distribution

Let $(\Omega, \mathcal{F}, P)$ be the abstract probability space and $X_i : \Omega \to \mathcal{X}_i$ be random variables for $i = 1, 2, \cdots, n$. The **joint distribution** of $\boldsymbol{X} := (X_1, X_2, \cdots, X_n)$ is the probability measure $P_{\boldsymbol{X}}$ with domain $\mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_n$ such that

$$\begin{aligned} P_{\boldsymbol{X}}(A_1 \times A_2 \times \cdots \times A_n) &:= P\big(\{\omega \in \Omega : X_1(\omega) \in A_1, X_2(\omega) \in A_2, \cdots, X_n(\omega) \in A_n\}\big) \\ &= P(X_1 \in A_1, X_2 \in A_2, \cdots, X_n \in A_n). \end{aligned}$$

Now we can define independence between random variables in a manner similar to Definition 1.2.5.

In this course, we focus on real-valued random variables, which can be fully determined by their *distribution functions*.

Definition 1.3.5 ▸ Distribution Function

Let X be a real-valued random variable over the abstract probability space, the **distribution function** of X is a function $F_X : \Omega \to [0, 1]$ such that

$$F_X(x) = P(X \leq x).$$

Note that for all $a < b \in \mathbb{R}$,

$$P\big(X \in (a, b]\big) = F_X(b) - F_X(a).$$

It can be shown from here that F_X fully determines the distribution of X (which is non-trivial).

A distribution function F_X has the following properties:

1. F_X is **non-decreasing**.
2. $\lim_{x \to -\infty} F_X(x) = 0$ and $\lim_{x \to +\infty} F_X(x) = 1$.
3. For all $x \in \mathbb{R}$, $F_X(x) = \lim_{y \to x^+} F_X(y)$ and $F_X(x^-) := \lim_{y \to x^-} F_X(y)$ exists. In particular, $P(X = x) = F_X(x) - F_X(x^-)$.

Remark. Conversely, every function $F : \mathbb{R} \to [0,1]$ satisfying the above properties induces a probability measure P on $\mathbb{R}$ with $P\big((-\infty, x]\big) = F_X(x)$.

In computer programs, a random number is often generated via the uniform random variable Z over $[0,1]$. Z can be used to generate a random variable X associated to any given distribution function F.

Theorem 1.3.6 ▸ Distribution Simulation

Let $F : \mathbb{R} \to [0,1]$ be any distribution function. Define $F' : [0,1] \to \mathbb{R}$ by

$$F'(y) = \inf\{x \in \mathbb{R} : F(x) \geq y\}.$$

Let Z be the uniform random variable on $[0,1]$, then $X := F'(Z)$ is a random variable with distribution function F.

A random variable can be discrete or continuous. We first define the discrete case.

Definition 1.3.7 ▸ Discrete Random Variable

A random variable is **discrete** if its range is countable.

Here we list down some commonly used discrete random variables and their distributions:

- $X \sim \text{Bernoulli}(p)$ where $0 < p < 1$:

$$P(X = i) = \begin{cases} 1-p & \text{if } i = 0 \\ p & \text{if } i = 1 \end{cases}.$$

- $X \sim \text{B}(n, p)$ where $0 < p < 1$ and $n \in \mathbb{N}^+$:

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} = \frac{n!}{x!\,(n-x)!} p^x (1-p)^{n-x}.$$

- $X \sim \text{Geo}(p)$ where $0 < p < 1$:

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

- $X \sim \text{Pois}(\lambda)$ where $\lambda > 0$:

$$P(X = x) = \mathrm{e}^{-\lambda} \frac{\lambda^x}{x!}.$$

Correspondingly, we give the definition for continuous random variables.

Definition 1.3.8 ▸ Continuous Random Variables

A random variable X is **continuous** if there exists a function $f_X : \mathbb{R} \to [0, \infty)$ called the **probability density function** such that for all $a < b$,

$$P(X \in (a, b]) = \int_a^b f_X(x)\, \mathrm{d}x.$$

The commonly used continuous random variables are as follows:

- $X \sim \mathrm{U}(a, b)$ where $b \geq a$:
$$f_X(x) = \frac{x - a}{b - a}.$$
- $X \sim \mathrm{Exp}(\lambda)$ where $\lambda > 0$:
$$f_X(x) = \lambda \mathrm{e}^{-\lambda x} \mathbf{1}_{[0,\infty)}(x).$$
- $X \sim \mathcal{N}(\mu, \sigma^2)$ where $\sigma^2 > 0$:
$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

A random variable can be neither discrete nor continuous. For example, let Y be the result of rolling a fair die, $Z \sim \mathrm{U}(0, 1)$ and $W \sim \mathrm{Bernoulli}(p)$. Define

$$X := \mathbf{1}_{W=1} Y + \mathbf{1}_{W=2} Z,$$

then $P(X \in A) = pP(Y \in A) + (1 - p) P(Z \in A)$.

1.3.1 Expectation

Recall that we have defined expectations for discrete and continuous random variables in elementary probability theory. In terms of measure theory, the two formulae can be unified as the Lebesgue integral

$$\mathbb{E}[X] = \int_S X(\omega)\, \mathrm{d}P(\omega).$$

In the discrete case, we have

$$\mathbb{E}[g(x)] = \sum_{x \in \mathcal{X}} g(x) P(X = x).$$

If X is non-negative integer-valued, this is equivalent to

$$\mathbb{E}\left[g\left(x\right)\right] = \sum_{n=0}^{\infty} P\left(X \geq n\right).$$

In the real-valued continuous case, we have

$$\mathbb{E}\left[g\left(x\right)\right] = \int_{-\infty}^{\infty} g\left(x\right) f_X\left(x\right) \, \mathrm{d}x.$$

The following result gives a way to approximate probabilities by the expectation of a random variable:

Theorem 1.3.9 ▸ Markov's Inequality

If X is a non-negative random variable, then $P\left(X \geq a\right) \leq \frac{\mathbb{E}[X]}{a}$ for all $a > 0$.

Proof. It suffices to prove for the continuous case. Notice that

$$\begin{aligned}
\mathbb{E}[X] &= \int_{0}^{\infty} x f_X\left(x\right) \, \mathrm{d}x \\
&\geq \int_{a}^{\infty} x f_X\left(x\right) \, \mathrm{d}x \\
&\geq a \int_{0}^{\infty} f_X\left(x\right) \, \mathrm{d}x \\
&= P\left(X \geq a\right).
\end{aligned}$$

Therefore, $P\left(X \geq a\right) \leq \frac{\mathbb{E}[X]}{a}$. □

Note that $\mathbb{E}[X]$ is a real number while $\mathbb{E}[X \mid Y]$ is a **random variable** formed as a function of Y. In a way, Y partitions the sample space into regions where $\mathbb{E}[X \mid Y = y_i]$ gives the expectation of X in the region induced by $Y = y_i$ for each $y_i \in \mathcal{Y}$. In general, the following result holds:

Theorem 1.3.10 ▸ Law of Iterated Expectations

Let X and Y be random variables, then $\mathbb{E}\big[\mathbb{E}[X \mid Y]\big] = \mathbb{E}[X]$.

The above formula can be interpreted as the fact that $\mathbb{E}[X \mid Y]$ is a best estimator for X.

1.3.2 Variance

Note that the expectation is insufficient in describing a random variable because probability mass on exceptionally large values can influence the expectation significantly. For example, let X_n be a random variable with $\Pr(X_n = n) = \frac{1}{n}$ and $\Pr(X_n = 0) = 1 - \frac{1}{n}$. Notice that by taking the limit, $\lim_{n\to\infty} \Pr(X_n = 0) = 1$ but $\mathbb{E}[X_n] = 1$ for all $n \in \mathbb{N}^+$. Therefore, we define the *variance* as another parameter to specify a distribution.

Definition 1.3.11 ▸ Variance

Let X be a random variable. The **variance** of X is defined as

$$\operatorname{Var}(X) \coloneqq \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2.$$

$\sqrt{\operatorname{Var}(X)}$ is called the **standard deviation** of X.

Note that the variance might not be finite. Consider a continuous random variable X with probability density function $f(x) = \frac{c}{1+x^3}\mathbf{1}_{[0,\infty)}(x)$ where $c > 0$ is appropriately chosen. One can check that $\mathbb{E}[X]$ is finite but $\mathbb{E}\left[X^2\right]$ is unbounded.

The following result is important:

Theorem 1.3.12 ▸ Chebyshev's Inequality

For any real-valued random variable X with finite variance,

$$P\left(|X - \mathbb{E}[X]| > a\sqrt{\operatorname{Var}(X)}\right) \leq \frac{1}{a^2}$$

for all $a > 0$.

Proof. Define $g(X) : (X - \mathbb{E}[X])^2$, which is clearly non-negative. By Theorem 1.3.9, we have

$$P\big(g(X) > a^2\operatorname{Var}(X)\big) \leq \frac{\mathbb{E}[g(X)]}{a^2\operatorname{Var}(X)}.$$

Note that $\mathbb{E}[g(X)] = \operatorname{Var}(X)$, so

$$P\left(|X - \mathbb{E}[X]| > a\sqrt{\operatorname{Var}(X)}\right) = P\big(g(X) > a^2\operatorname{Var}(X)\big) \leq \frac{1}{a^2}.$$

□

Remark. In general,

$$P(|X - \mathbb{E}[X]| > a) \leq \frac{\operatorname{Var}(X)}{a^2}.$$

1.3.3 Correlation

Given any random variables X and Y, they are not necessarily independent in general. We wish to investigate how correlated they are to each other. For this purpose, we introduce the following notion:

Definition 1.3.13 ▸ Covaraince

If X and Y are real-valued random variables, the **covariance** between X and Y is defined as

$$\operatorname{Cov}(X,Y) \coloneqq \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y] = \mathbb{E}\big[(X - \mathbb{E}[X])\,(Y - \mathbb{E}[Y])\big].$$

Remark. Note that $\operatorname{Cov}(X,Y) = \operatorname{Cov}(Y,X)$ and $\operatorname{Cov}(X,X) = \operatorname{Var}(X)$.

2

Stochastic Process

2.1 Markov Chains

Definition 2.1.1 ▸ Stochastic Process

A **stochastic process** is a collection of random variables $\{X(t) : t \in T\}$ where T is an **index set** and $X(t)$ is known as the **current state**. The set of all possible states is known as the **state space**.

Let S be a sample space, a stochastic process defined over the space can be thought of a sequence of random variables where $X(t)$ describes the distribution of an outcome $\omega \in S$ at timestamp t. The state space S is simply the co-domain of the $X(t)$'s.

A stochastic process is said to be

- *discrete-time* if the index set is countable;
- *continuous-time* if the index set is a continuum;
- *discrete-state* if the state space is countable;
- *finite-state* if the state space is finite;
- *continuous-state* if the state space is a continuum.

The term "continuum" refers to a **non-empty compact connected metric space**.

In this course, we focus on discrete-time discrete-state stochastic processes. One important property we will discuss now is the *Markovian property*.

Definition 2.1.2 ▸ Markovian Property

Let $\{X_n : n \in T\}$ be a discrete-time stochastic process over some probability space $(S, \mathcal{F}, P)$. The stochastic process is called **Markovian** if

$$P\big(X_{n+1} = x_{n+1} \mid X_0^n = (x_0, x_1, \cdots, x_n)\big) = P\left(X_{n+1} = x_{n+1} \mid X_n = x_n\right)$$

for all $n \in \mathbb{N}$.

The Markovian property essentially says that, given X_n, what has happened before, i.e., X_k for all $k < n$, is independent of what happens afterwards, i.e., X_{n+m} for $m \in \mathbb{N}^+$.

As the name suggests, the Markovian property is closely related to the Markov chains, which can be defined rigorously as follows:

Definition 2.1.3 ▶ Discrete-Time Markov Chain

A **Markov chain** is a discrete-time discrete-state stochastic process satisfying the Markovian property.

Recall that the *Bayes's Theorem* states the following:

Theorem 2.1.4 ▶ Bayes's Theorem

For any random variables X and Y,

$$p_{X|Y}(x \mid y) = \frac{p_{Y|X}(y \mid x)\, p_X(x)}{\sum_{x' \in \mathcal{X}} p_{Y|X}(y \mid x')\, p_X(x')}.$$

Theorem 2.1.4 leads to the following important result:

Corollary 2.1.5 ▶ Bayes' Rule for Markov Chains

Let $X_1, X_2, \cdots, X_n$ be any n random variables forming a Markov chain, then

$$p_{X_1^n}(x_1, x_2, \cdots, x_n) = p_{X_1}(x_1) \prod_{i=1}^{n-1} p_{X_{i+1}|X_i}(x_{i+1} \mid x_i).$$

Proof. If $n = 2$, by Theorem 2.1.4, we know that

$$\begin{aligned} p_{X_1,X_2}(x_1, x_2) &= p_{X_1|X_2}(x_1 \mid x_2)\, p_{X_2}(x_2) \\ &= p_{X_1}(x_1)\, p_{X_2|X_1}(x_2 \mid x_1). \end{aligned}$$

Suppose that there exists some integer $k \geq 2$ such that

$$p_{X_1^k}(x_1, x_2, \cdots, x_k) = p_{X_1}(x_1) \prod_{i=1}^{k-1} p_{X_{i+1}|X_i}(x_{i+1} \mid x_i)$$

For any k random variables X_1^k forming a Markov chain. Let X_{k+1} be any random variable such that X_1^{k+1} forms a Markov chain, then

$$p_{X_{k+1}|X_1^k}(x_{k+1} \mid x_1, x_2, \cdots, x_k) = p_{X_{k+1}|X_k}(x_{k+1} \mid x_k).$$

By using Theorem 2.1.4, we have

$$\begin{aligned} p_{X_1^{k+1}}(x_1, x_2, \cdots, x_{k+1}) &= p_{X_{k+1}|X_1^k}(x_{k+1} \mid x_1, x_2, \cdots, x_k)\, p_{X_1^k}(x_1, x_2, \cdots, x_k) \\ &= p_{X_{k+1}|X_k}(x_{k+1} \mid x_k)\, p_{X_1}(x_1) \prod_{i=1}^{k-1} p_{X_{i+1}|X_i}(x_{i+1} \mid x_i) \\ &= p_{X_1}(x_1) \prod_{i=1}^{k} p_{X_{i+1}|X_i}(x_{i+1} \mid x_i). \end{aligned}$$

□

Consider a discrete-time discrete-state stochastic process $\{X_n : n \in T\}$ with state space S over some probability space $(S, \mathcal{F}, P)$. Here, the σ-algebra $\mathcal{F}$ can be generated using simple events $\{\omega \in S : X_n(\omega) = s\}$ for all $n \in T$ and $s \in S$. Notice that this means that we need to find the joint distribution

$$p_{X_{n_1}^{n_k}}(s_1, s_2, \cdots, s_k)$$

for any tuple of random variables $X_{n_1}^{n_k}$ in the stochastic process, where $k \in \mathbb{N}$ and $k \leq |T|$ if T is finite, and any $(s_1, s_2, \cdots, s_k) \in S^k$. In general, this joint distribution might be hard to find, but things become easier if the stochastic process is a Markov chain because by Corollary 2.1.5 we have

$$p_{X_{n_1}^{n_k}}(s_1, s_2, \cdots, s_k) = p_{X_{n_1}}(s_1) \prod_{i=1}^{k-1} p_{X_{n_{i+1}}|X_{n_i}}(s_{i+1} \mid s_i).$$

If we can find $p_{X_m|X_n}(s_m \mid s_n)$ for any $m > n$, we could simplify this expression further!

Definition 2.1.6 ▸ Transition Probability

The **transition probability** is defined as

$$p_{ij}^{n,m} := P(X_m = j \mid X_n = i).$$

In particular, $p_{ij}^{n,n+1}$ is known as the **one-step transition probability** or **jump probability**.

Take some $k \in \mathbb{N}^+$ and consider

$$p_{ij}^{n,n+k} = P(X_{n+k} = j \mid X_n = i).$$

We first marginalise $P(X_{n+k} = j \mid X_n = i)$ with respect to X_{n+1} to obtain

$$P(X_{n+k} = j \mid X_n = i) = \sum_{s \in S} P(X_{n+k} = j \mid X_n = i, X_{n+1} = s) P(X_{n+1} = s \mid X_n = i).$$

Since X_n, X_{n+1} and X_{n+k} form a Markov chain, we have

$$P(X_{n+k} = j \mid X_n = i, X_{n+1} = s) = P(X_{n+k} = j \mid X_{n+1} = s).$$

Therefore,

$$\begin{aligned} p_{ij}^{n,n+k} &= P(X_{n+k} = j \mid X_n = i) \\ &= \sum_{s \in S} P(X_{n+k} = j \mid X_{n+1} = s) P(X_{n+1} = s \mid X_n = i) \\ &= \sum_{s \in S} p_{sj}^{n+1,n+k} p_{is}^{n,n+1}. \end{aligned}$$

Notice that now we have reduced the gap by 1. By repeatedly applying this process to $p_{sj}^{n+1,n+k}$, we eventually arrive at

$$p_{ij}^{n,n+k} = \sum_{s_1, s_2, \cdots, s_{k-1} \in S} p_{is_1}^{n,n+1} \left(\prod_{r=1}^{m-n-2} p_{s_r s_{r+1}}^{n+r,n+r+1} \right) p_{s_{m-1} j}^{n+k-1,n+k}$$

It is useful to see the one-step transition probability $p_{ij}^{n,n+1}$ as a function

$$f : T \times S \times S \to \mathbb{R}.$$

Thus far, we have basically shown that to specify a Markov chain fully, we will need to define the **index set** T, the **state space** S and the **one-step transition probabilities** $p_{ij}^{n,n+1}$ for all $i, j \in S$.

We can write the transition probabilities as a matrix.

Definition 2.1.7 ▸ Transition Probability Matrix

For any Markov chain, the **transition probability matrix** is a matrix $\boldsymbol{P}^{n,n+1}$ where $P_{ij}^{n,n+1} = p_{ij}^{n,n+1}$.

Let π_t be the distribution at time t for a Markov chain $\{X_i\}_{i \in \mathbb{N}^+}$, then we can write

$$\pi_t := \left[P(X_t = x_1), P(X_t = x_2), \cdots, P\left(X_t = x_{|\mathcal{X}|}\right). \right]$$

Therefore, the distribution at time $t+1$ is given by

$$\pi_{t+1} = \pi_t \boldsymbol{P}^{n,n+1}.$$

Iterate this process and we have

$$\pi_t = \pi_0 \prod_{i=0}^{t-1} \boldsymbol{P}^{i,i+1}.$$

Generally, the (i,j) entry of $\prod_{i=n}^{m-1} \boldsymbol{P}^{i,i+1}$ is exactly $p_{ij}^{n,m}$.

Notice that in general, $p_{ij}^{n,n+1}$ is dependent on n, but when the transition probability is independent of n, the computation will become much easier.

Definition 2.1.8 ▸ Time-Homogeneous Markov Chain

A Markov Chain is **time-homogeneous** if $p_{ij}^{n,n+1} = p_{ij}^{1,2}$ for all $n \in \mathbb{N}^+$.

For a time-homogeneous Markov chain X, we can write its distribution as

$$\pi_t = \pi_0 \boldsymbol{P}^t.$$

This means that any time-homogeneous Markov chain is fully determined by its state space, transition probability matrix $\boldsymbol{P}$ and the starting distribution π_0.

Definition 2.1.9 ▸ Stochastic Matrix

Let S be a countable index set. A matrix $\boldsymbol{P}$ is called a **stochastic matrix** if $P_{ij} \geq 0$ for all $i, j \in S$ and $\sum_{j \in S} P_{ij} = 1$ for all $i \in S$.

The above computation can be summarised into the following result:

Theorem 2.1.10 ▸ Chapman-Kolmogorov Equations

Let X be a time-homogeneous Markov chain with state space S and transition probability matrix $\boldsymbol{P}$. For any $m, n \in \mathbb{N}$,

$$P(X_{m+n} = j \mid X_0 = i) = \sum_{k \in S} P(X_m = k \mid X_0 = i) P(X_n = j \mid X_0 = k).$$

Furthermore, for all $n \in \mathbb{N}$,

$$P(X_n = j \mid X_0 = i) = \boldsymbol{P}_{ij}^n.$$

Take any state space $S \coloneqq \{x_1, x_2, \cdots, x_s\}$. Notice that if we have a column vector

$$\mu \coloneqq \begin{bmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ f(x_s) \end{bmatrix}$$

for some function f, then

$$\begin{aligned} (\boldsymbol{P}^n \mu)_i &= \sum_{j=1}^{s} \boldsymbol{P}_{ij}^n \mu_j \\ &= \sum_{j=1}^{s} P\left(X_n = x_j \mid X_0 = x_i\right) f\left(x_j\right) \\ &= \mathbb{E}\left[f(X_n) \mid X_0 = x_i\right]. \end{aligned}$$

Suppose $X_0 \sim \lambda$, then clearly

$$\begin{aligned} \mathbb{E}\left[f(X_n)\right] &= \sum_{i=1}^{s} \mathbb{E}\left[f(X_n) \mid X_0 = x_i\right] \lambda_i \\ &= \sum_{i=1}^{s} \lambda_i \left(\boldsymbol{P}^n \mu\right)_i \\ &= \lambda \boldsymbol{P}^n \mu. \end{aligned}$$