Contents

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** over 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 $F_k \coloneqq E_k \backslash \bigcup_{i=1}^{k-1} E_i$ 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 such that for every outcome measured by a random variable X in a sample space S, there exists a function $f : \Omega \to S$ such that $X = f(Z)$. 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**.

Proposition 1.3.2 ▸ Countable Additivity in Probability Spaces

Let (Ω, P) be a probability space. If $\{A_i\}_{i \in \mathbb{N}^+}$ is a family of subsets of Ω such that $A_i \subseteq A_{i+1}$ for all $i \in \mathbb{N}^+$, then

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

If $\{B_i\}_{i \in \mathbb{N}^+}$ is a family of subsets of Ω such that $B_{i+1} \subseteq B_i$ for all $i \in \mathbb{N}^+$, then

$$P\left(\bigcap_{i \in \mathbb{N}^+} B_i\right) = \lim_{n \to \infty} P(B_n).$$

Proof. Define $A_0 := \varnothing$, then

$$\{A_{i+1} \setminus A_i : i \in \mathbb{N}\}$$

is a countable collection of pairwise disjoint subsets of Ω. Therefore,

$$\begin{aligned}
P\left(\bigcup_{i\in\mathbb{N}^+} A_i\right) &= P\left(\bigcup_{i\in\mathbb{N}} (A_{i+1} \setminus A_i)\right) \\
&= \sum_{i=0}^{\infty} P(A_{i+1} \setminus A_i) \\
&= \lim_{n\to\infty} \sum_{i=0}^{n} P(A_{i+1} \setminus A_i) \\
&= \lim_{n\to\infty} P\left(\bigcup_{i=0}^{n} (A_{i+1} \setminus A_i)\right) \\
&= \lim_{n\to\infty} P(A_n).
\end{aligned}$$

Define $C_i := \Omega \setminus B_i$. Since $B_{i+1} \subseteq B_i$ for all $i \in \mathbb{N}^+$, we have $C_i \subseteq C_{i+1}$ for all $i \in \mathbb{N}^+$. Therefore,

$$\begin{aligned}
P\left(\bigcup_{i\in\mathbb{N}^+} B_i\right) &= P\left(\bigcup_{i\in\mathbb{N}^+} (\Omega \setminus C_i)\right) \\
&= \sum_{i=1}^{\infty} P(\Omega \setminus C_i) \\
&= 1 - \sum_{i=1}^{\infty} P(C_i) \\
&= 1 - P\left(\bigcup_{i\in\mathbb{N}^+} C_i\right) \\
&= 1 - \lim_{n\to\infty} P(C_n) \\
&= 1 - \lim_{n\to\infty} P(\Omega \setminus B_n) \\
&= 1 - \lim_{n\to\infty} (1 - P(B_n)) \\
&= \lim_{n\to\infty} P(B_n).
\end{aligned}$$

□

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 1.3.3 ▸ 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 1.3.4 ▸ 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 $X : \Omega \to \mathcal{X}$ 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 $(\Omega, \mathcal{F})$ to $(\mathcal{X}, \mathcal{B})$.

We shall verify that P_X as defined above is indeed a probability measure. Notice that we have $P_X(A) \in [0,1]$ for all $A \subseteq \mathcal{X}$ and that $P_X(\mathcal{X}) = 1$ and $P_X(\varnothing) = 0$. Let $\{A_i\}_{i \in \mathbb{N}}$ be a family of pairwise disjoint events. Consider

$$\begin{aligned} X^{-1}\left(\bigcup_{i \in \mathbb{N}} A_i\right) &= \left\{\omega \in \Omega : X(\omega) \in \bigcup_{i \in \mathbb{N}} A_i\right\} \\ &= \bigcup_{i \in \mathbb{N}} \{\omega \in \Omega : X(\omega) \in A_i\} \\ &= \bigcup_{i \in \mathbb{N}} X^{-1}(A_i). \end{aligned}$$

Therefore,

$$\begin{aligned} P_X\left(\bigcup_{i\in\mathbb{N}} A_i\right) &= P\left(\bigcup_{i\in\mathbb{N}} X^{-1}(A_i)\right) \\ &= \sum_{i\in\mathbb{N}} P\left(X^{-1}(A_i)\right) \\ &= \sum_{i\in\mathbb{N}} P_X(A). \end{aligned}$$

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

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

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 1.3.5 ▸ 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 1.3.6 ▸ 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). By using Proposition 1.3.2, we can prove the following result with some analysis tools:

Proposition 1.3.7 ▸ Properties of Distribution Functions

If F_X is a distribution function of a real-valued random variable X, then

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 1.3.8 ▸ 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.

Proof. Notice that for all $x \in \mathbb{R}$, we have $P(X \leq x) = P\left(F^{-1}(Z) \leq x\right)$. One may check that $F^{-1}(z) \leq x$ if and only if $z \leq F(x)$, so

$$P(X \leq x) = P\big(Z \leq F(x)\big) = F(x).$$

□

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

Definition 1.3.9 ▶ 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:

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

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

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

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

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

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

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

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

Correspondingly, we give the definition for continuous random variables.

Definition 1.3.10 ▶ 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 $a < b$,

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

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 $\lambda > 0$:
$$f_X(x) = \lambda \mathrm{e}^{-\lambda x} \mathbf{1}_{[0,\infty)}(x).$$
- $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$ 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

$$X \coloneqq \mathbf{1}_{\{W=1\}} Y + \mathbf{1}_{\{W=0\}} Z,$$

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

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

In the real-valued continuous case, we have

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

Furthermore, we have the following properties:

Proposition 1.3.11 ▸ Basic Properties of Expectation

Let X be a real-valued random variable, then

1. *if $P\left(X \geq 0\right) = 1$, then $\mathbb{E}\left[X\right] \geq 0$;*
2. *if $P\left(X = 1\right) = 1$, then $\mathbb{E}\left[X\right] = 1$;*
3. *$\mathbb{E}\left[aX + b\right] = a\mathbb{E}\left[X\right] + b$ for any $a, b \in \mathbb{R}$;*
4. *if Y is another real-valued random variable such that X and Y are independent, then $\mathbb{E}\left[XY\right] = \mathbb{E}\left[X\right]\mathbb{E}\left[Y\right]$.*

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

Theorem 1.3.12 ▸ 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

$$\begin{aligned}\mathbb{E}\left[X\right] &= \int_0^{\infty} x f_X\left(x\right)\,\mathrm{d}x \\ &\geq \int_a^{\infty} x f_X\left(x\right)\,\mathrm{d}x \\ &\geq a \int_0^{\infty} f_X\left(x\right)\,\mathrm{d}x \\ &= P\left(X \geq a\right).\end{aligned}$$

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

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

Theorem 1.3.13 ▸ Law of Total Expectations

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

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 1.3.14 ▶ Variance

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

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

$\sqrt{\operatorname{Var}(X)}$ 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 $f(x) = \frac{c}{1+x^3}\mathbf{1}_{\{x\in[0,\infty)\}}$ where $c > 0$ is appropriately chosen. One can check that $\mathbb{E}[X]$ is finite but $\mathbb{E}\left[X^2\right]$ is unbounded.

Similar to our discussion of expectation, we propose the following basic property:

Proposition 1.3.15 ▶ Basic Property of Variance

Let X be a real-valued random variable, then

$$\operatorname{Var}(aX + b) = a^2\operatorname{Var}(X)$$

for all $a, b \in \mathbb{R}$.

The following result is important:

Theorem 1.3.16 ▶ 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 1.3.12, we have

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

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

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

□

Remark. In general,

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

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 1.3.17 ▶ Covaraince

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

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

Remark. Note that $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$ and $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$. If X and Y are independent, then $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X) = 0$.

Covariance can be seen as a measure for the **joint** variation of two random variables, which offers a way to assess the dependency between two random variables to some extent.

Definition 1.3.18 ▶ Correlation

Let X and Y be random variables. The **correlation** between X and Y is defined as

$$\mathrm{Corr}(X, Y) := \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}}.$$

If $\mathrm{Corr}(X, Y) = 0$, we say that X and Y are **unrelated**.

Correlation is obviously symmetric. Furthermore, we have the following result:

Proposition 1.3.19 ▶ Range of Correlation

Let X and Y be random variables such that their expectation and variance are both finite, then $\operatorname{Corr}(X, Y) \in [-1, 1]$.

Remark. In particular, $\operatorname{Corr}(X, X) = 1$ and $\operatorname{Corr}(X, -X) = -1$.

Lastly, we propose the following result on the variance of the sum of random variables:

Proposition 1.3.20 ▶ Variance of a Sum

Let $\{X_i\}_{i=1}^n$ be a family of random variables such that $\mathbb{E}\left[X_i^2\right]$ is finite for all $i = 1, 2, \cdots, n$. then

$$\operatorname{Var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \operatorname{Var}(X_i) + 2 \sum_{1 \leq i < j \leq n} \operatorname{Cov}\left(X_i, X_j\right).$$

1.4 Generating Functions

Definition 1.4.1 ▶ Probability Generating Function

Let X be an $\mathbb{N}$-valued discrete random variable, then its **probability generating function** is a map $\phi_X : T \to \mathbb{R}$ defined as

$$\phi_X(t) := \mathbb{E}\left[t^X\right] = \sum_{k \in \mathbb{N}} P(X = k) t^k,$$

where $T \subseteq \mathbb{R}$ is the set of all values of t such that the sum converges.

Notice that

$$\phi_X^{(n)}(t) = \sum_{k=n}^{\infty} k! P(X = k) t^{k-n},$$

so it is clear that

$$P(X = x) = \frac{\phi_X^{(x)}(0)}{x!},$$

hence the name "probability generating function". Moreover, we have $\phi_X(0) = P(X = 0)$ and $\phi(1) = 1$. Using the properties of expectation, it is easy to prove the following result:

Proposition 1.4.2 ▶ Probability Generating Function of Sum of Random Variables

If X and Y are independent $\mathbb{N}$-valued discrete random variables, then $\phi_{X+Y} = \phi_X \phi_Y$.

Analogously, another kind of generating functions concern the *moments* of a random variable.

Definition 1.4.3 ▶ Moment Generating Function

Let X be an $\mathbb{N}$-valued random variable, then its **moment generating function** is a map $M_X : T \to \mathbb{R}$ defined as

$$M_X(t) = \mathbb{E}\left[\mathrm{e}^{tX}\right] = \sum_{k \in \mathbb{N}} P(X = k)\,\mathrm{e}^{tk},$$

where $T \subseteq \mathbb{R}$ is the set of all values of t such that the sum converges.

By observation, we see that

$$M_X^{(n)}(0) = \sum_{k \in \mathbb{N}} k^n P(X = k)$$

yields the n-th moment of X. However, notice that $M_X(t) = \phi_X\left(e^t\right)$, so actually

$$\mathbb{E}\left[X^n\right] = \left.\frac{\mathrm{d}^n \phi\left(\mathrm{e}^t\right)}{\mathrm{d}t^n}\right|_{t=0}.$$

Next, we state a result without proof on the use of generating functions to study the convergence of distributions.

Proposition 1.4.4 ▶ Convergence of Generating Functions

Let $\{X_i\}_{i \in \mathbb{N}}$ be a family of $\mathbb{N}$-valued random variables such that ϕ_i is the probability generating function for X_i and X be an $\mathbb{N}$-valued random variable with probability generating function ϕ. If there exists some $a \in \mathbb{R}$ such that $\{\phi_i\}_{i \in \mathbb{N}}$ converges to ϕ point-wisely, then

$$\lim_{n \to \infty} P(X_n = k) = P(X = k)$$

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

Note that in reality, many random variables are not integer-valued. Therefore, we should extend the notion of generating functions to the general case.

Definition 1.4.5 ▶ Laplace Transform of Random Variables

Let X be a non-negative random variable. The **Laplace transform** of X is defined as a map $\Lambda : \mathbb{R}_0^+ \to \mathbb{R}$ such that $\Lambda(\lambda) := \mathbb{E}\left[\mathrm{e}^{-\lambda X}\right]$.

One may check that the properties of a Laplace transform is analogous to those of generating functions:

1. $\Lambda_X(\lambda) \in [0,1]$ for all $\lambda \geq 0$;
2. $\Lambda_X(0) = 1$ and Λ_X is decreasing in λ;
3. $\Lambda_X^{(n)}(0) = (-1)^n \mathbb{E}[X^n]$, so Λ_X determines the moments;
4. if $\{\Lambda_n\}_{n\in\mathbb{N}}$ is a family of Laplace transforms for random variables $\{X_n\}_{n\in\mathbb{N}}$ respectively, then $\{\Lambda_n\}_{n\in\mathbb{N}}$ converging to Λ, the Laplace transform of a random variable X, point-wisely on $[0,a]$ for some $a \in \mathbb{R}$ implies that $\left\{p_{X_n}\right\}_{n\in\mathbb{N}}$ converges point-wisely to p_X on $\mathbb{R}_0^+$.

Note that we choose $\lambda \geq 0$ which will ensure that $\mathbb{E}\left[\mathrm{e}^{-\lambda X}\right]$ is finite if X is non-negative. However, for general random variables which can take negative values, we need to use a different construction.

Definition 1.4.6 ▶ Fourier Transform of Random Variables

Let X be a real-valued random variable. The **Fourier transform** of X is defined as a map $\varphi_X : \mathbb{R} \to \mathbb{R}$ such that $\varphi_X(t) \coloneqq \mathbb{E}\left[\mathrm{e}^{\mathrm{i}tX}\right]$.

One may check that the properties of a Fourier transform is analogous to those of Laplace transforms:

1. $|\varphi_X(t)| \in [0,1]$ for all $t \in \mathbb{R}$;
2. $\varphi_X(0) = 1$;
3. $\varphi_X^{(n)}(0) = \mathrm{i}^n \mathbb{E}[X^n]$, so φ_X determines the moments;
4. if $\{\varphi_n\}_{n\in\mathbb{N}}$ is a family of Laplace transforms for random variables $\{X_n\}_{n\in\mathbb{N}}$ respectively, then $\{\varphi_n\}_{n\in\mathbb{N}}$ converging to φ, the Fourier transform of a random variable X, point-wisely on $[-a,a]$ for some $a > 0$ implies that $\left\{p_{X_n}\right\}_{n\in\mathbb{N}}$ converges point-wisely to p_X on $\mathbb{R}$.

1.5 Limit Theorems

Chebyshev's Inequality can be used to prove a very important result regarding large numbers.

Theorem 1.5.1 ▸ Weak Law of Large Numbers

Let $\{X_i\}_{i\in\mathbb{N}}$ be a family of independent and identically distributed random variables with mean μ and finite variance, then for any $\epsilon > 0$,

$$\lim_{n\to\infty} P\left(\left|\frac{\sum_{i=1}^{n} X_i}{n} - \mu\right| > \epsilon\right) = 0.$$

2 Markov Chains

2.1 Stochastic Processes

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

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

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.

In a Markov chain, we have the following representation of Bayes's Rule:

Proposition 2.1.4 ▸ Bayes' Rule for Markov Chains

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

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

Proof. If $n = 2$, by Theorem 1.2.4, we know that

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

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

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

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

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

By using Theorem 1.2.4, 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 **??** 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 2.1.5 ▸ 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 $P\left(X_{n+k} = j \mid X_n = i\right)$ with respect to X_{n+1} to obtain

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

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

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

Therefore,

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

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 2.1.6 ▸ Transition Probability Matrix

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

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

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

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 2.1.7 ▸ Stationary Markov Chain

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

For a stationary Markov chain X, suppose $\boldsymbol{P}^{(m)}$ is the m-step transition probability matrix, then we can write its distribution as

$$\pi_{t+m} = \pi_t \boldsymbol{P}^{(m)} = \pi_t \boldsymbol{P}^m.$$

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

Definition 2.1.8 ▸ 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 2.1.9 ▸ Chapman-Kolmogorov Equations

Let X be a Markov chain with state space S, then

$$\begin{aligned}\boldsymbol{P}^{(m)} &= \boldsymbol{P}\boldsymbol{P}^{(m-1)} \\ &= \boldsymbol{P}^{(m-1)}\boldsymbol{P}.\end{aligned}$$

In particular, if X is a stationary Markov chain with transition probability matrix $\boldsymbol{P}$, *then 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 $S := \{x_1, x_2, \cdots, x_s\}$. Notice that if we have a column vector

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

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

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

2.2 Stationary Distribution

One classic problem related to Markov chains is **Gambler's Ruin**, which states the following scenario:

> Two gamblers A and B bet against each other by flipping a fair coin. Upon each flip, if a head shows up, B gives 1 dollar to A and if a tail shows up, A gives 1

> dollar to B. Let X_i and Y_i be the amount of money A and B have after the i-th coin flip respectively. If $X_0 = m$ and $Y_0 = n$, what is the probability that A is ruined first?

We can model this problem with a discrete-time Markov chain X with finite state space S. Fix some $a, b \in S$ with $a \neq b$, and define

$$H := \begin{cases} 1 & \text{if there exists } i < j \text{ such that } X_i = a \text{ and } X_j = b \\ 0 & \text{otherwise} \end{cases}.$$

Define $f(x) = P(H = 1 \mid X_0 = x)$. Notice that

$$\begin{aligned} \mathbb{E}\left[f(X_1) \mid X_0 = x\right] &= \sum_{x \in S} P(H = 1 \mid X_0 = X_1, X_0 = x) P(X_1 = x \mid X_0 = x) \\ &= \sum_{x \in S} P(H = 1 \mid X_1 = x, X_0 = x) P(X_1 = x \mid X_0 = x) \\ &= P(H = 1 \mid X_0 = x). \end{aligned}$$

Therefore, $f(x) = (\boldsymbol{P}\mu)(x)$. In other words, $\boldsymbol{P}\mu$ gives the conditional distribution for the event that a is reached before b given the starting state X_0.

Furthermore, suppose $\mu = \mathbf{1}$ is a constant vector, then clearly $\boldsymbol{P}^n\mu = \mu = \mathbf{1}$ for all $n \in \mathbb{N}$. Therefore, $\mathbf{1}$ is an eigenvector for $\boldsymbol{P}$ with eigenvalue 1. Note that $\boldsymbol{P}^{\mathrm{T}}$ and $\boldsymbol{P}$ have the same eigenvalues, so there exists some row vector ν such that

$$\nu\boldsymbol{P} = \nu.$$

If ν is a probability vector, then if $X_0 \sim \nu$, we have $X_n \sim \nu$ for all $n \in \mathbb{N}$. Such ν is known as the *stationary distribution*.

Definition 2.2.1 ▸ Stationary Distribution

A distribution π for a Markov chain X with transition probability matrix $\boldsymbol{P}$ is a **stationary distribution** if $\pi\boldsymbol{P} = \pi$.

Intuitively,

$$\lim_{n \to \infty} \boldsymbol{P}^n = \begin{bmatrix} \pi \\ \pi \\ \vdots \\ \pi \end{bmatrix}$$

if π is a stationary distribution. To see this, we consider the following proposition:

Proposition 2.2.2 ▶ Eigenvalues of Transition Probability Matrix

Let $\boldsymbol{P}$ be the transition probability matrix for a finite-state discrete-time Markov chain X. If λ is an eigenvalue of $\boldsymbol{P}$, then $|\lambda| \leq 1$.

Proof. Let μ be the eigenvector associated to λ. Note that there exists some state x_0 such that $|\mu(x_0)|$ attains the maximum. Let S be the state space. Consider

$$\begin{aligned}
|\lambda\mu(x_0)| &= |(\boldsymbol{P}\mu)(x_0)| \\
&= \left|\sum_{y\in S} \boldsymbol{P}(y, x_0)\,\mu(y)\right| \\
&\leq |\mu(x_0)|\left|\sum_{y\in S} \boldsymbol{P}(y, x_0)\right| \\
&= |\mu(x_0)|.
\end{aligned}$$

Therefore, $|\lambda| \leq 1$. □

One important question is whether a stationary distribution exists for a Markov chain, and if it exists, whether it is unique. To investigate such problems, we first need to impose some regularity conditions to Markov chains.

To visualise such regularity, it is helpful to model a Markov chain as a digraph using its state space as the vertex set. The following definition helps us to define the edges.

Definition 2.2.3 ▶ Accessibility

Let X be a Markov chain with probability transition matrix $\boldsymbol{P}$. A state j is said to be **accessible** from another state i if $P_{ij}^{(m)} > 0$ for some $m \in \mathbb{N}$.

In other words, j is accessible from i of the Markov chain can reach j with **non-zero probability** when starting from i using finitely many steps. We sometimes use $i \rightarrow j$ to denote such accessibility.

It is easy to check that accessibility is **transitive**.

Markov chains might exhibit some pathological behaviours.

Definition 2.2.4 ▶ Disconnected States

Let S be the state space of a discrete-time Markov chain with probability transition matrix $\boldsymbol{P}$. If there exist $X, Y \subseteq S$ such that $P_{ij} = 0$ for all $(i, j) \in X \times Y$ and for all $(i, j) \in Y \times X$, then X and Y are said to be **disconnected**.

Note the disconnected states are mutually unreachable. There is another situation where reachability is one-way.

Definition 2.2.5 ▸ Source and Sink

Let S be the state space of a discrete-time Markov chain with probability transition matrix $\boldsymbol{P}$. A state $i \in S$ is a **source** if for all $j \neq i$, we have $P_{ji} = 0$, and a **sink**, or **absorbing state**, if for all $j \neq i$, we have $P_{ij} = 0$.

Remark. Colloquially, a source is a state which cannot be reached again once left and a sink is a state which the process cannot leave once reached.

Therefore, the ideal case is that every pair of states are mutually reachable.

Definition 2.2.6 ▸ Intercommunication

For a Markov chain X with state space S, two states x and y are said to **intercommunicate**, denoted as $x \leftrightarrow y$, if

$$P(X_m = y \mid X_0 = x) > 0 \quad \text{and} \quad P(X_n = x \mid X_0 = y) > 0$$

for some $m, n \in \mathbb{N}$.

It is clear that intercommunication is an **equivalence relation** on the state space, the proof of which is easy enough to be left as an exercise to the reader.

Now, the state space will be partitioned by the quotient with $\leftrightarrow$. Each equivalence class, denoted by $\mathcal{C}$, is a closed walk with states as the vertices. More precisely, if we model the Markov chain as a digraph D such that $V(D)$ is the state space and $E(D) := \{(x, y) : x \to y\}$, then each $V \in V(D)\big/\leftrightarrow$ induces a **strongly connected component** of D.

Clearly it suffices to study the Markov chain on a sink class to investigate the long-term behaviour. In particular, we may consider a Markov chain such that the entire state space intercommunicates.

Definition 2.2.7 ▸ Irreducibility

A Markov chain is **irreducible** if its state space is an equivalence class under $\leftrightarrow$.

Note that a Markov chain is irreducible if and only if it has only one intercommunicating class. Furthermore, an irreducible chain can have finitely or infinitely many states.

Irreducible Markov chains can be further classified to *transient* ones and *recurrent* ones.

Definition 2.2.8 ▶ Return Probability

Let i be a state of a Markov chain X. The **return probability** to i is defined as

$$P_{ii}^{(n)} := P\left(X_n = i \mid X_0 = i\right).$$

By this definition, we have $P_{ii}^{(n)} = (\boldsymbol{P}^n)_{ii}$ for stationary chains. Note that 3 cases could be possible when a state i is re-visited in n steps:

1. starting from i, the chain self-loops at i for n times;
2. starting from i, the chain visits some different states and re-visits i for the first time at the n-th step;
3. starting from i, the chain re-visits i for several times and happens to reach i again at the n-th step.

However, we should observe that the key to computing the probability for either case is to find the time taken for the chain to re-visit the starting state **for the first time**.

Definition 2.2.9 ▶ First Return Probability

Let X be a Markov chain and i be a state. The **first return probability** to i at the n-th transition is defined as

$$f_{ii}^{(n)} := P\left(X_1 \neq i, X_2 \neq i, \cdots, X_{n-1} \neq i, X_n = i \mid X_0 = i\right).$$

Remark. We define $f_{ii}^{(0)} = 0$.

Clearly, $f_{ii}^{(n)} \leq P_{ii}^{(n)}$ for all $n \in \mathbb{N}$. In general, the following formula is a way to compute the return probability using first return probability:

Theorem 2.2.10 ▶ Formula for Return Probability

Let X be a Markov chain and i be a state, then

$$P_{ii}^{(n)} = \sum_{k=0}^{n} f_{ii}^{(k)} P_{ii}^{(n-k)}.$$

Proof. Let

$$T_i := \min\left\{n \in \mathbb{N}^+ : X_n = i\right\}$$

be the first timestamp at which X returns to i starting from i. Notice that

$$\begin{aligned} P_{ii}^{(n)} &= \sum_{k=0}^{n} P\left(T_i = k \mid X_0 = i\right) P\left(X_n = i \mid X_k = i\right) \\ &= \sum_{k=0}^{n} f_{ii}^{(k)} P_{ii}^{(n-k)}. \end{aligned}$$

□

Notice that since $f_{ii}^{(0)} = 0$, we only need $P_{ii}^{(1)}, \cdots, P_{ii}^{(n-1)}$ to compute $P_{ii}^{(n)}$ iteratively.

It is natural to classify the states based on whether we can ascertain that the Markov chain can return to the state in finite time.

Definition 2.2.11 ▶ Transient and Recurrent States

Let

$$f_{xx} \coloneqq \sum_{n=0}^{\infty} f_{xx}^{(n)}$$

be the probability that X returns to x in finite time. The state x is said to be **recurrent** if $f_{xx} = 1$ and **transient** if $f_{xx} < 1$.

In short, a recurrent state is one which the process will definitely return to in finite time, and a transient state always has a non-zero probability that the process never returns to it in finite time. Therefore, an alternative definition considers

$$T_x \coloneqq \min\left\{n \in \mathbb{N}^+ : X_n = x\right\}$$

and define

$$f_{xx} \coloneqq \sum_{n=0}^{\infty} P\left(T_x = n \mid X_0 = x\right).$$

It is easy to see that the two definitions are equivalent, which also follows from Definition 2.2.9.

Notice that the probability that the process cannot return to x is given by $1 - f_{xx}$.

Note that if x is a recurrent state, then the probability of X returning to x in finite time after starting at x is 1. By the Markovian property, we know that X will also return to x in finite time for the second time. This implies that for any recurrent state x, the process will visit it for infinitely many times in the long-run with a probability of 1. Consider $\mathbf{1}_{\{X_n = x\}}$ to be a

Bernoulli random variable with parameter f_{xx}, then if x is recurrent, we have

$$P\left(\sum_{n=1}^{\infty} \mathbf{1}_{\{X_n=x\}} = \infty\right) = 1.$$

This means that the expected number of returns to x in the long-run is given by

$$\mathbb{E}\left[\sum_{n=0}^{\infty} \mathbf{1}_{\{X_n=x\}} \mid X_0 = x\right] = \sum_{n=0}^{\infty} \boldsymbol{P}^n (x, x) = \infty.$$

A small note is that this does not mean that $\boldsymbol{P}^n (x, x)$ must not converge to 0.

If x is a transient state, then there is a non-zero probability that X never returns to x in finite time. Let N_x be the number of times X returns to x before it never returns to x in finite time. It is clear that $N_x \sim \text{Geo}(1 - f_{xx})$. Therefore, the expected number of returns to x is

$$\mathbb{E}[N_x] = \frac{f_{xx}}{1 - f_{xx}}.$$

We formalise this observation into the following proposition:

Proposition 2.2.12 ▸ Expected Number of Returns of Markov Chains

Let X be a discrete-time finite-state irreducible Markov chain and x be some state. Let N_x be the number of returns to x given $X_0 = x$ before X never returns to x in finite time, then

$$\mathbb{E}[N_x \mid X_0 = x] = \frac{f_{xx}}{1 - f_{xx}}.$$

Proof. Let T_n be the return time for the n-th return to x and define $T_0 = 0$. Consider

$$\begin{aligned}
P(N_x = k \mid X_0 = x) &= P(T_{k+1} = \infty, T_k < \infty) \\
&= \sum_{n=0}^{\infty} P(T_{k+1} = \infty, T_k = n) \\
&= \sum_{n=0}^{\infty} P(T_{k+1} = \infty \mid T_k = n) P(T_k = n) \\
&= \sum_{n=0}^{\infty} P(T_1 = \infty \mid X_0 = x) P(T_k = n) \\
&= P(T_1 = \infty \mid X_0 = x) \sum_{n=0}^{\infty} P(T_k = n) \\
&= (1 - f_{xx}) P(T_k < \infty).
\end{aligned}$$

Note that $P(T_n < \infty) = P(T_{n-1} < \infty) f_{xx}$, so we have

$$P(T_n < \infty) = f_{xx}^n$$

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

$$P(N_x = k \mid X_0 = x) = (1 - f_{xx}) f_{xx}^k$$

and so

$$\mathbb{E}[N_x \mid X_0 = x] = \frac{f_{xx}}{1 - f_{xx}}.$$

□

Remark. A direct consequence of this result is that the expected number of **visits** including the initial one to x is $\frac{1}{1-f_{xx}}$.

We can related this result to Definition 2.2.8 to derive a characterisation for recurrent and transient states.

Proposition 2.2.13 ▸ Characterisation of Recurrent and Transient States

Let X be a Markov chain and i be a state, then i is transient if and only if

$$\sum_{n=1}^{\infty} P_{ii}^{(n)}$$

is finite.

Proof. Let N_i be the number of returns to i starting from i, then

$$N_i = \sum_{n=1}^{\infty} \mathbf{1}_{\{X_n = i\}}.$$

Therefore,

$$\begin{aligned}\mathbb{E}\left[N_i \mid X_0 = i\right] &= \mathbb{E}\left[\sum_{n=1}^{\infty} \mathbf{1}_{\{X_n=i\}} \middle| X_0 = i\right] \\ &= \sum_{n=1}^{\infty} \mathbb{E}\left[\mathbf{1}_{\{X_n=i\}} \middle| X_0 = i\right] \\ &= \sum_{n=1}^{\infty} P\left(X_n = i \mid X_0 = i\right) \\ &= \sum_{n=1}^{\infty} P_{ii}^{(n)}.\end{aligned}$$

By Proposition 2.2.12, i is transient if and only if $\sum_{n=1}^{\infty} P_{ii}^{(n)}$ is finite. □

An important implication of this result is that, if i is a transient state, then $\sum_{n=m}^{\infty} P_{ii}^{(n)}$ is non-increasing as m increases. Therefore, by monotone convergence theorem,

$$\lim_{m\to\infty} \sum_{n=m}^{\infty} P_{ii}^{(n)} = 0,$$

which means that the probability of returning to i is **vanishing** in the long-run.

Intuitively, if a transient state x intercommunicates with some other state y, then y will also be transient because we can return to y as long as we return to x. In general, we propose the following result:

Proposition 2.2.14 ▸ Recurrence and Transience as a Class Property

Let x and y be intercommunicating states of a Markov chain, then they are either both transient or both recurrent.

Proof. It suffices to show that x is recurrent if and only if y is recurrent. Let $\boldsymbol{P}$ be the transition probability matrix of the Markov chain. Since $x \leftrightarrow y$, there exists $k, \ell \in \mathbb{N}$ such that

$$\boldsymbol{P}^k(x,y) > 0, \qquad \boldsymbol{P}^{\ell}(x,y) > 0.$$

Notice that

$$\boldsymbol{P}^{k+\ell+n}(x,x) \geq \boldsymbol{P}^k(x,y)\,\boldsymbol{P}^n(y,y)\,\boldsymbol{P}^{\ell}(y,x)$$

for all $n \in \mathbb{N}$. Notice that the expected number of returns to x is given by

$$\begin{aligned}\sum_{i=0}^{\infty} \boldsymbol{P}^{i}(x,x) &\geq \sum_{n=0}^{\infty} \boldsymbol{P}^{k+\ell+n}(x,x) \\ &\geq \sum_{n=0}^{\infty} \boldsymbol{P}^{k}(x,y)\,\boldsymbol{P}^{n}(y,y)\,\boldsymbol{P}^{\ell}(y,x) \\ &= \boldsymbol{P}^{k}(x,y)\,\boldsymbol{P}^{\ell}(y,x)\sum_{n=0}^{\infty} \boldsymbol{P}^{n}(y,y).\end{aligned}$$

Therefore, x is recurrent if and only if y is recurrent. □

Therefore, for each intercommunicating class in a Markov chain, we can classify it as either a “recurrent” class or a “transient” class. This motivates the following definition:

Definition 2.2.15 ▶ Recurrent and Transient Markov Chains

An irreducible Markov chain with a countable state space is called **recurrent** if one of its states is recurrent, and **transient** if one of its states is transient.

Intuitively, in a recurrent class, the chain will return to each state indefinitely and in a transient class, the chain will eventually “escape” from the class.

Theorem 2.2.16 ▶ Path Properties

Let X be an irreducible Markov chain with a countable state space S, initial distribution μ and probability transition matrix $\boldsymbol{P}$. If X is recurrent, then

$$P\left(\sum_{n=0}^{\infty} \mathbf{1}_{\{X_n=i\}} > M\right) = 1$$

for all $M \in \mathbb{R}$ and all $i \in S$. If X is transient, then there exists some $M_i \in \mathbb{R}$ for each $i \in S$ such that

$$P\left(\sum_{n=0}^{\infty} \mathbf{1}_{\{X_n=i\}} \leq M_i\right) = 1$$

Proof. Suppose that X is recurrent and fix some $x \in S$. For every $i \in S$ with $i \neq x$, since X is irreducible, we have $i \leftrightarrow x$ and so there exists some $n_i > 0$ such that $P\left(X_{n_i} = i \mid X_0 = x\right) > 0$. Without loss of generality, we can take n_i to be the least integer with such property. By Proposition 2.2.12, we have

$$\mathbb{E}\left[\sum_{n=n_i}^{\infty} \mathbf{1}_{\{X_n=i\}} \middle| X_{n_i} = i\right] > M$$

for all $M \in \mathbb{R}$. This means that

$$\begin{aligned}\mathbb{E}\left[\sum_{n=0}^{\infty} \mathbf{1}_{\{X_n=i\}}\right] &= \mathbb{E}\left[\mathbb{E}\left[\sum_{n=0}^{\infty} \mathbf{1}_{\{X_n=i\}} \middle| X_0\right]\right] \\ &= \sum_{x\in S} P\left(X_{n_i} = i \mid X_0 = x\right) \mathbb{E}\left[\sum_{n=0}^{\infty} \mathbf{1}_{\{X_n=i\}} \middle| X_0 = i\right] \\ &> M\end{aligned}$$

for all $M \in \mathbb{R}$. Therefore, for all $M \in \mathbb{R}$,

$$P\left(\sum_{n=0}^{\infty} \mathbf{1}_{\{X_n=i\}} > M\right) = 1.$$

If X is transient, by Proposition 2.2.13, there exists some $M_i \in \mathbb{R}$ for each $i \in S$ such that

$$\sum_{n=0}^{\infty} \mathbf{1}_{\{X_n=i|X_0=i\}} \le M_i.$$

Therefore,

$$\sum_{n=0}^{\infty} \mathbf{1}_{\{X_n=i\}} \le \sum_{n=0}^{\infty} \mathbf{1}_{\{X_n=i|X_0=i\}} \le M_i.$$

□

Now we states a result about finite-state irreducible chains:

Corollary 2.2.17 ▸ Recurrence of Finite-State Irreducible Markov Chains

All finite-state irreducible Markov chains are recurrent.

Proof. Suppose on contrary that X is a finite-state irreducible transient Markov chain with state space S, then by Theorem 2.2.16, there exists some $M \in \mathbb{R}$ such that

$$\sum_{i\in S} \sum_{n=0}^{\infty} \mathbf{1}_{\{X_n=i\}} \le M,$$

which is not possible because this implies that $X_{M+1} \neq S$. □

We now briefly discuss the **long-run behaviour** of reducible Markov chains. Note that if X is a reducible Markov chain, then we can decompose the state space of X into intercommunicating classes $\mathcal{C}_1, \mathcal{C}_2, \cdots$. It is clear that

$$\lim_{n\to\infty} P\left(X_n \in \mathcal{C}_k\right) = 0$$

if C_k is transient.

Proposition 2.2.18 ▶ Limiting Distribution of Transient Irreducible Markov Chains

Let X be an irreducible transient Markov chain with state space S and transition probability matrix $\boldsymbol{P}$, then for any $x, y \in S$, we have

$$\lim_{n\to\infty} P_{xy}^{(n)} = 0 \quad \textit{and} \quad \lim_{n\to\infty} P(X_n = y) = 0.$$

Proof. By Proposition 2.2.12, there exists some $M_{xy} \in \mathbb{R}^+$ such that

$$\mathbb{E}\left[\sum_{n=0}^{\infty} \mathbf{1}_{X_n=y} \middle| X_0 = x\right] = \sum_{n=0}^{\infty} P_{xy}^{(n)} < M_{xy}.$$

Therefore, $\sum_{n=0}^{\infty} P_{xy}^{(n)}$ converges for all $x, y \in S$, which means that

$$\lim_{n\to\infty} P_{xy}^{(n)} = 0.$$

Therefore,

$$\begin{aligned}\lim_{n\to\infty} P(X_n = y) &= \lim_{n\to\infty} \sum_{x\in S} P(X_0 = x) P_{xy}^{(n)} \\ &= \sum_{x\in S} P(X_0 = x) \lim_{n\to\infty} P_{xy}^{(n)} \\ &= 0.\end{aligned}$$

□

Now consider the general case. Suppose that π is a limiting distribution, i.e.,

$$\pi := \lim_{n\to\infty} \mu \boldsymbol{P}^n$$

for some initial distribution μ, then it is clear that $\pi\boldsymbol{P} = \pi$. Informally, this means that π will not change with time.

Definition 2.2.19 ▶ Stationary Measure and Stationary Distribution

Let X be a Markov chain with a countable state space S and transition probability matrix $\boldsymbol{P}$, a measure $\mu : S \to [0, \infty)$ is called a **stationary measure** if $\mu\boldsymbol{P} = \mu$ and

$$\sum_{x\in S} \mu(x) \in (0, \infty].$$

If μ is a probability measure, then it is a **stationary distribution** of X.

Remark. Note that if μ is a stationary measure such that $\sum_{x \in S} \mu(x) = C$ is finite, then we can normalise it by $\widehat{\mu} \coloneqq \frac{\mu}{C}$ to obtain a stationary distribution on X.

Note that a stationary distribution might not be a limiting distribution. Consider a two-state Markov chain with

$$\boldsymbol{P} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

It is easy to see that $\pi \coloneqq \left(\frac{1}{2}, \frac{1}{2}\right)$ is the only stationary distribution, but if we have initial distribution $\mu \coloneqq (0, 1)$, the distribution will not converge to π.

Clearly, if a Markov chain has a finite state space S, we can find the stationary distribution by taking

$$\boldsymbol{Q} \coloneqq \begin{bmatrix} \boldsymbol{P} & \boldsymbol{I}_{n \times 1} \end{bmatrix}$$

and solving the linear system $\boldsymbol{x}\boldsymbol{Q} = [\, x \;\, 1 \,]$, so every finite-state Markov chain has a stationary distribution, but it may not be unique.

If $\mathcal{C}_k$ is a recurrent class, we say that $P(X_n \in \mathcal{C}_k)$ is the *entering probability*. Note that for a recurrent $\mathcal{C}_k$, we can restrict X to $\mathcal{C}_k$ as the new state space to construct an irreducible recurrent Markov chain. If π is a stationary distribution of the restricted chain and p is the entering probability to $\mathcal{C}_k$, then for the original chain we have

$$P(X_n = i) = \pi_i p.$$

Therefore, to investigate the long-run behaviour of a reducible Markov chain, it suffices to discuss the long-run behaviour of the restricted chains on all of its recurrent classes.

2.3 Regularity

Definition 2.3.1 ▶ Period

Let X be a Markov chain and i be a state. The **period** of i is defined as

$$d(i) \coloneqq \gcd\left\{n \in \mathbb{N}^+ : P_{ii}^{(n)} > 0\right\}.$$

If $P_{ii}^{(n)} = 0$ for all $n \in \mathbb{N}^+$, we define $d(i) = 0$. If $d(i) = 1$, then i is said to be an **aperiodic state**.

If we simulate the chain with a graph, then $\left\{n \in \mathbb{N}^+ : P_{ii}^{(n)} > 0\right\}$ is the set of lengths of all i-i walks and the period is just the greatest common divisor of all these walk lengths. Intuitively, all states in the same intercommunicating class should have the same period.

Proposition 2.3.2 ▶ Periodicity as a Class Property

If $i \leftrightarrow j$ in a Markov chain, then $d(i) = d(j)$.

Proof. Since $i \leftrightarrow j$, there exists some $k, \ell \in \mathbb{N}^+$ such that $P_{ji}^{(k)} > 0$ and $P_{ij}^{(\ell)} > 0$. Therefore,

$$P_{ii}^{(\ell+k)} \geq P_{ij}^{(\ell)} P_{ji}^{(k)} > 0$$

and so $d(i) \mid (\ell + k)$. Now, for all $n \in \left\{n \in \mathbb{N}^+ : P_{jj}^{(n)} > 0\right\}$, we have

$$P_{ii}^{(\ell+n+k)} \geq P_{ij}^{(\ell)} P_{jj}^{(n)} P_{ji}^{(k)} > 0.$$

Therefore, $d(i) \mid (\ell + n + k)$ and so $d(i) \mid n$. This means that $d(i) \geq d(j)$. Similarly, we can show that $d(j) \geq d(i)$, and so $d(i) = d(j)$. □

It is clear that for any state i, we can find some positive integer N such that $P_{ii}^{(Nd(i))} > 0$. In general, we have the following result:

Proposition 2.3.3 ▶ Long-Run Behaviour of Periodicity

For any state i in a Markov chain, there is some $N \in \mathbb{N}$ such that for any integer $n \geq N$, we have

$$P_{ii}^{(nd(i))} > 0$$

and

$$P_{ji}^{(m+nd(i))} > 0$$

whenever $P_{ji}^{(m)} > 0$ for some $m \in \mathbb{N}^+$.

Consider an irreducible Markov chain X. Note that X only has one intercommunicating class, so every state of X shares the same period. Therefore, it makes sense to talk about the period of the Markov chain, denoted by d. If $d = 1$, then we say that X is an *aperiodic Markov chain*.

Proposition 2.3.4 ▶ Long-Run Behaviour of Irreducible Aperiodic Chains

Let X be a finite-state irreducible aperiodic Markov chain with transition probability matrix $\boldsymbol{P}$, then there exists some $N \in \mathbb{N}^+$ such that $\boldsymbol{P}^{(N)}$ has all positive entries.

Proof. Let S be the finite state space. For all $i \in S$, by Proposition 2.3.3 there exists some $N_i \in \mathbb{N}^+$ such that for any integer $n \geq N_i$ we have $P_{ii}^{(n)} > 0$. Take

$$N \coloneqq \max_{i \in S} N_i,$$

then we have $P_{ii}^{(N)} > 0$ for all $i \in S$, i.e., $\boldsymbol{P}^{(N)}$ has positive entries along its diagonal. Take any $j \in S$, then by Proposition 2.3.3 we can find some $M_{ij} \in \mathbb{N}^+$ such that

$$P_{ij}^{(M_{ij})} > 0.$$

Define $N_{ij} \coloneqq N + M_{ij}$, then for any integer $n \geq N_{ij}$, we can write $n = n_0 + m$ for some $m \geq M_{ij}$ and some $n_0 \geq N$. Therefore,

$$\begin{aligned} P_{ij}^{(n)} &= P_{ij}^{(n_0+m)} \\ &\geq P_{ii}^{(n_0)} P_{ij}^{(m)} \\ &> 0. \end{aligned}$$

Take $M \coloneqq \max_{i,j \in S} N_{ij}$, then $P_{ij}^{M} > 0$ for all $i, j \in S$, i.e., $\boldsymbol{P}^{(M)}$ has all positive entries. □

The above result implies that when N is large enough, a finite-state irreducible aperiodic Markov chain can reach any state after exactly N steps!

Definition 2.3.5 ▶ Regular Markov Chain

A Markov chain is said to be **regular** if there exists some $k \in \mathbb{N}^+$ such that the k-step transition probability matrix $\boldsymbol{P}^{(k)}$ has all positive entries.

Such a $\boldsymbol{P}^{(k)}$ is called a *regular transition probability matrix* and it implies that

$$P\left(X_k = j \mid X_0 = i\right) > 0$$

for any states i and j. Using Proposition 2.3.4, we have already characterised one type of regular chains.

Proposition 2.3.6 ▶ Irreducible Aperiodic Finite-State Markov Chains Are Regular

All irreducible aperiodic finite-state Markov chains are regular.

It is easy to see that all regular Markov chains must be irreducible. We claim that if $\boldsymbol{P}^k$ has all positive entries, then so does $\boldsymbol{P}^n$ for all $n \geq k$. The intuition is as follows: $\boldsymbol{P}^k$ having all positive entries means that we can reach any state after k steps. Note that irreducibility means that for any state i, there exists at least some other state $j \neq i$ such that $j \to i$. Since it is possible to reach j after k steps, then it must be possible to reach i after $(k+1)$ steps.

Proposition 2.3.7 ▸ Regular n-step Transition Probability Matrix

In a regular Markov chain, if $\boldsymbol{P}^{(k)}$ is regular, then $\boldsymbol{P}^{(n)}$ is regular for all $n \geq k$.

We now state the main theorem for regular Markov chains:

Theorem 2.3.8 ▸ Main Theorem for Regular Markov Chains

Let $\boldsymbol{P}$ be a regular transition probability matrix for some regular Markov chain with state space $S := \{1, 2, \cdots, N\}$, then there exists a unique probability distribution

$$\pi := \left(\pi_1, \pi_2, \cdots, \pi_N\right)$$

such that $\pi \boldsymbol{P} = \pi$ and

$$\pi_j = \lim_{n \to \infty} P_{ij}^{(n)} = \sum_{k \in S} \pi_k P_{kj}$$

for all $i \in S$.

Remark. In particular, $\pi \boldsymbol{P} = \pi$ is known as the *global balance equations*.

In practice, it is not so easy to solve the system $\pi \boldsymbol{P} = \pi$, so we often consider the following system instead:

Definition 2.3.9 ▸ Local Balance Equations

Let π be a distribution on a Markov chain X with probability transition matrix $\boldsymbol{P}$. The **local balance equations** are given by

$$\pi_i P_{ij} = \pi_j P_{ji}$$

for all i, j in the state space.

It can be proven that local balance equations are stronger than global balance equations.

The distribution π obtained from the above theorem is known as the *limiting distribution*.

Clearly, a limiting distribution is a stationary distribution. Notice that

$$\begin{aligned}\lim_{n\to\infty} P(X_n = j) &= \sum_{i=1}^{N} \lim_{n\to\infty} P(X_n = j \mid X_0 = i) P(X_0 = i) \\ &= \sum_{i=1}^{N} \pi_j P(X_0 = i) \\ &= \pi_j \sum_{i=1}^{N} P(X_0 = i) \\ &= \pi_j.\end{aligned}$$

Therefore, π_j gives the marginal probability of $X_n = j$ in the long-run. Furthermore, π_j is intuitively the proportion of time where $X_n = j$. Formally, we have

$$\begin{aligned}\mathbb{E}\left[\frac{1}{n}\sum_{k=0}^{n-1} \mathbf{1}_{\{X_k=j\}} \,\middle|\, X_0 = i\right] &= \frac{1}{n}\sum_{k=0}^{n-1} \mathbb{E}\left[\mathbf{1}_{\{X_k=j\}} \,\middle|\, X_0 = i\right] \\ &= \frac{1}{n}\sum_{k=0}^{n-1} P(X_k = j \mid X_0 = i).\end{aligned}$$

Next, let us consider irreducible Markov chains with infinitely many states.

Definition 2.3.10 ▸ Null Recurrent and Positive Recurrent Markov Chains

Let X be an irreducible Markov chain with a countable state space S. Let

$$R_i \coloneqq \min\{n \in \mathbb{N}^+ : X_n = i\}$$

be the first return time of $i \in S$, then i is called **positive recurrent** if $\mathbb{E}[R_i \mid X_0 = i]$ is finite, and **null recurrent** otherwise. A Markov chain is **positive recurrent** if all of its states are positive recurrent, and **null recurrent** if all of its states are null recurrent.

Similar to the recurrence and transience of states, we can prove that both null and positive recurrence are a class property.

Proposition 2.3.11 ▸ Null and Positive Recurrence as a Class Property

If x and y are two recurrent states such that $x \leftrightarrow y$, then either both are null recurrent or both are positive recurrent.

Clearly, any irreducible recurrent Markov chain is either null recurrent or positive recurrent. We shall first study irreducible positive recurrent Markov chains.

Definition 2.3.12 ▶ Ergodic Markov Chain

An **ergodic** Markov chain is an irreducible aperiodic positive recurrent Markov chain.

We can prove that an ergodic Markov chain exhibits the following limiting behaviour:

Theorem 2.3.13 ▶ Basic Limit Theorem

Let X be an ergodic Markov chain with state space S and transition probability matrix $\mathbf{P}$. Define

$$m_i \coloneqq \mathbb{E}\left[R_i \mid X_0 = i\right] = \sum_{n=1}^{\infty} n f_{ii}^{(n)},$$

where R_i is the first return time to $i \in S$, then for any $i, j \in S$, the limit $\lim_{n\to\infty} P_{ij}^{(n)}$ exists and

$$\lim_{n\to\infty} P_{ij}^{(n)} = \lim_{n\to\infty} P_{jj}^{(n)} = \frac{1}{m_j} = \pi_j,$$

where π is a probability distribution with $\pi\mathbf{P} = \pi$.

Usually, it is easier to compute π first, from which we can then derive m_j for each state j. Additionally, for null recurrent Markov chains, we can prove the following limit theorem:

Proposition 2.3.14 ▶ Long-Run Limits of Null Recurrent Markov Chains

Let X be a null recurrent Markov chain with state space S and transition probability matrix $\mathbf{P}$, then for all $i, j \in S$, we have $\lim_{n\to\infty} P_{ij}^{(n)} = 0$.

We can summarise the above results into the following theorem:

Theorem 2.3.15 ▶ Limit Theorem for Markov Chains

Let X be an irreducible recurrent Markov chain with state space S and define

$$m_i \coloneqq \mathbb{E}\left[R_i \mid X_0 = i\right] = \sum_{n=1}^{\infty} n f_{ii}^{(n)},$$

where R_i is the first return time to i, then for any $i, j \in S$,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} P_{ij}^{(k)} = \frac{1}{m_j}.$$

If X is aperiodic, then

$$\lim_{n\to\infty} P_{ij}^{(n)} = \frac{1}{m_j}$$

and if X is periodic with period d, then

$$\lim_{n\to\infty} P_{jj}^{(nd)} = \frac{d}{m_j}.$$

2.4 Markov Chain Monte Carlo

Monte Carlo algorithms are computer algorithms which utilise repeated random sampling. By repeating the simulation to obtain identical and independent samples, we can obtain a *sample mean* for the random variable of interest, which is an unbiased estimator for the true mean of the random variable.

Suppose the random variable of interest has distribution p and we wish to obtain many independent samples distributed according to p, then we can simply set up a Markov chain for which p is a stationary distribution. Note that this means that we need a transition probability matrix $\boldsymbol{P}$ such that $p\boldsymbol{P} = p$. However, this equation is too difficult to solve.

To deal with this, we take inspiration from **hill-climbing algorithm**. Suppose we construct any initial transition probability matrix $\boldsymbol{Q}$, such that the corresponding Markov chain is aperiodic and irreducible, then we can aim at a function g such that $pg(\boldsymbol{Q}) = p$. Recall that if p and $g(\boldsymbol{Q})$ satisfy the local balance equations, they will definitely satisfy the global balance equations. Therefore, consider any i, j from the state space such that

$$p_i Q_{ij} \neq p_j Q_{ji}.$$

Take $p_{\min} \coloneqq \min\{p_i Q_{ij}, p_j Q_{ji}\}$ and fix some function α such that

$$p_i Q_{ij} \alpha(i, j) = p_{\min} = p_j Q_{ji} \alpha(j, i).$$

By observation, we have

$$\alpha(i, j) = \frac{p_{\min}}{p_i Q_{ij}} = \min\left\{\frac{p_j Q_{ji}}{p_i Q_{ij}}, 1\right\}.$$

Next, to make sure that the matrix still have row sum as 1 after each iteration, we adjust its diagonal where the local balance equations always hold. This means that

$$P_{ii} = Q_{ii} + \sum_{k\neq i} Q_{ik}\big(1 - \alpha(i, k)\big) = 1 - \sum_{k\neq i} P_{ik}.$$

Theoretically, we could apply this procedure to all pairs of states, but that is computationally unrealistic. Instead, suppose we have a sequence of random variables $\{X_n\}_{n\in\mathbb{N}}$ where

$$P(Y_{n+1} = X_{n+1} \mid Y_n = X_n) = P_{X_n X_{n+1}},$$

then it can be proven that the distribution of Y converges to p. Therefore, for some sufficiently large $M \in \mathbb{N}^+$, we can be confident that $X_M \sim p$ approximately.

Technique 2.4.1 ▸ Hastings-Metropolis Algorithm

- Suppose p is our target distribution with sample space S.
- Let M be the maximum depth of iterations.
- Choose an irreducible Markov chain with state space S and transition probability matrix $\boldsymbol{Q}$ which is symmetric.
- Select some $X_0 \in S$.
- At the n-th iteration:
    - Generate some $Y \sim q$ where $q_j = Q_{X_n j}$ for all $j \in S$.
    - Generate some $U \sim \mathrm{U}(0, 1)$.
    - Compute $\alpha(X_n, Y) := \min\left\{\frac{p_Y Q_{Y X_n}}{p_{X_n} Q_{X_n Y}}, 1\right\}$.
    - If $U < \alpha(X_n, Y)$, set $X_{n+1} = Y$; else set $X_{n+1} = X_n$.
- Return $Z := X_M \sim p$ approximately.

Notice that

$$\begin{aligned} P(Z_{n+1} = X_{n+1} \mid Z_n = X_n) &= Q_{X_n X_{n+1}} \alpha(X_n, X_{n+1}) = P_{X_n X_{n+1}}, \\ P(Z_{n+1} = X_n \mid Z_n = X_n) &= Q_{X_n X_n} + \sum_{i \neq X_n} Q_{X_n i}\big(1 - \alpha(X_n, i)\big). \end{aligned}$$

Another interesting thing to notice here is that what we need to run the algorithm is the ratio $\frac{p_i}{p_j}$ over all pairs of states, and not the distribution p itself. This means that if $p = c\boldsymbol{b}$ for some vector $\boldsymbol{b}$ (known as a *kernel function*) and a *normalising constant* $c \in \mathbb{R}$, we can compute $\frac{p_i}{p_j} = \frac{b_i}{b_j}$ without knowing c.

> *Remark.* We can modify Algorithm 2.4.1 to apply to continuous random variables by changing the one-step transition probabilities with the one-step transition density functions.

2.5 Branching Processes

One special type of Markov chains is the *branching processes*, which can be used to simulate population dynamics. Consider a family of random variables $\{X_n\}_{n\in\mathbb{N}}$, where at the end of each time period (known as a *generation*), every individual generates $\xi \sim F$ offsprings independently, then

$$X_{n+1} = \sum_{i=1}^{X_n} \xi_i,$$

which is a random sum.

Definition 2.5.1 ▸ Branching Process

Let $\left\{\xi_i^{(n)}\right\}_{i,n\in\mathbb{N}}$ be a family of independent and identically distributed random variables. A stochastic process where

$$X_{n+1} = \sum_{i=1}^{X_n} \xi_i^{(n)}$$

for all $n \in \mathbb{N}^+$ is called a **branching process**.

Given the initial value X_0, it suffices to specify ξ in order to specify a branching process. It is clear that a branching process is a Markov chain because the $\xi_i^{(n)}$'s are independent of X_n. Notice that

$$P\left(X_{n+1} = j \mid X_n = i\right) = P\left(\sum_{k=1}^{i} \xi_k = j \mid X_n = i\right),$$

so the Markov chain is time-homogeneous. Furthermore, this means that if $\boldsymbol{P}$ is the transition probability matrix, then

$$P_{ij} = P\left(\sum_{k=1}^{i} \xi_k = j\right).$$

Consider a branching process with respect to ξ whose distribution is known, then we can write out ϕ_ξ explicitly. We will use this to specify a branching process.

Proposition 2.5.2 ▸ Specification of Branching Processes

Let X be a branching process with respect to ξ with probability generating function ϕ_ξ, then

$$\phi_{X_{n+1}|X_n}(t) = \phi_\xi(t)^{X_n}$$

for all $t \in \mathbb{R}$.

Proof. Let $X_n = x$, then for all $t \in \mathbb{R}$,

$$\begin{aligned}\phi_{X_{n+1}|X_n=x}(t) &= \mathbb{E}\left[t^{X_{n+1}} \,\middle|\, X_n = x\right] \\ &= \mathbb{E}\left[t^{\sum_{i=1}^{x} \xi_i}\right] \\ &= \mathbb{E}\left[\prod_{i=1}^{x} t^{\xi_i}\right] \\ &= \prod_{i=1}^{x} \mathbb{E}\left[t^{\xi_i}\right] \\ &= \prod_{i=1}^{x} \phi_{\xi_i}(t).\end{aligned}$$

Since the ξ_i's are independent, we have

$$\phi_{X_{n+1}|X_n}(t) = \phi_{\xi}(t)^{X_n}$$

for all $t \in \mathbb{R}$. □

Furthermore, we can remove the condition altogether by conditioning on the initial distribution of X_0 only instead.

Corollary 2.5.3 ▸ Specification of Branching Processes in Initial Distribution

Let X be a branching process with respect to ξ with probability generating function ϕ_ξ, then

$$\phi_{X_n|X_0}(t) = \phi_\xi^n(t)^{X_0}$$

for all $t \in \mathbb{R}$.

Proof. Let $X_0 = k$. Notice that for all $t \in \mathbb{R}$,

$$\begin{aligned}\phi_{X_n|X_0=k}(t) &= \mathbb{E}\left[t^{X_n} \,\middle|\, X_0 = k\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\prod_{i=1}^{X_{n-1}} t^{\xi_i} \,\middle|\, X_{n-1}\right] \,\middle|\, X_0 = k\right] \\ &= \mathbb{E}\left[\prod_{i=1}^{X_{n-1}} \mathbb{E}\left[t^{\xi_i}\right] \,\middle|\, X_0 = k\right] \\ &= \mathbb{E}\left[\phi_\xi(t)^{X_{n-1}} \,\middle|\, X_0 = k\right] \\ &= \phi_{X_{n-1}|X_0=k}\big(\phi_\xi(t)\big).\end{aligned}$$

Therefore,

$$\phi_{X_n|X_0=k}(t) = \phi_{X_1|X_0=k}\big(\phi_\xi^{n-1}(t)\big).$$

By Proposition 2.5.2,

$$\phi_{X_1|X_0=k}\big(\phi_\xi^{n-1}(t)\big) = \prod_{i=1}^{k} \phi_\xi\big(\phi_\xi^{n-1}(t)\big),$$

which implies that

$$\phi_{X_n|X_0}(t) = \phi_\xi^n(t)^{X_0}$$

for all $t \in \mathbb{R}$. □

Using the same idea of recursion, we can find the expectation and variance of a branching process.

Proposition 2.5.4 ▸ Expectation and Variance of Branching Processes

Let X be a branching process with respect to ξ with $\mathbb{E}[\xi] = \mu$ and $\operatorname{Var}(\xi) = \sigma^2$, then

$$\mathbb{E}[X_n \mid X_0] = \mu^n X_0 \quad \textit{and} \quad \operatorname{Var}(X_n \mid X_0) = \begin{cases} \frac{1-\mu^n}{1-\mu}\left(\mu^{n-1}\sigma^2\right)X_0 & \textit{if } \mu \neq 1 \\ n\left(\mu^{n-1}\sigma^2\right)X_0 & \textit{if } \mu = 1 \end{cases}.$$

Proof. Notice that

$$\begin{aligned} \mathbb{E}[X_n \mid X_0] &= \mathbb{E}\left[\sum_{i=1}^{X_{n-1}} \xi_i^{(n-1)} \,\middle|\, X_0\right] \\ &= \mathbb{E}[X_{n-1}\xi \mid X_0] \\ &= \mu\mathbb{E}[X_{n-1} \mid X_0], \end{aligned}$$

so $\mathbb{E}[X_n \mid X_0] = \mu^n\mathbb{E}[X_0 \mid X_0] = \mu^n X_0$. Consider

$$\begin{aligned} \operatorname{Var}(X_n) &= \mathbb{E}\left[(X_n - \mathbb{E}[X_n])^2\right] \\ &= \mathbb{E}\left[(X_n - \mu\mathbb{E}[X_{n-1}])^2\right] \\ &= \mathbb{E}\left[\left[(X_n - \mu X_{n-1}) + (\mu X_{n-1} - \mu\mathbb{E}[X_{n-1}])\right]^2\right]. \end{aligned}$$

Let $\mathbb{E}\left[X_{n-1}\right] = \mu'$, then

$$\begin{aligned}
\mathbb{E}\left[\left(X_n - \mu X_{n-1}\right)^2\right] &= \sum_{x=0}^{\infty} \mathbb{E}\left[\left(X_n - \mu X_{n-1}\right)^2 \middle| X_{n-1} = x\right] P\left(X_{n-1} = x\right) \\
&= \sum_{x=0}^{\infty} \mathbb{E}\left[\left(\sum_{i=1}^{x} \xi_i - \mu x\right)^2 \middle| X_{n-1} = x\right] P\left(X_{n-1} = x\right) \\
&= \sum_{x=0}^{\infty} x\sigma^2 P\left(X_{n-1} = x\right) \\
&= \sigma^2 \mu', \\
\mathbb{E}\left[\left(\mu X_{n-1} - \mu\mu'\right)^2\right] &= \mu^2 \mathbb{E}\left[\left(X_{n-1} - \mu'\right)^2\right] \\
&= \mu^2 \operatorname{Var}\left(X_{n-1}\right), \\
\mathbb{E}\left[\left(X_n - \mu X_{n-1}\right)\left(\mu X_{n-1} - \mu\mu'\right)\right] &= \mu \sum_{x=0}^{\infty} \left(x - \mu'\right) \mathbb{E}\left[X_n - x\mu \mid X_{n-1} = x\right] P\left(X_{n-1} = x\right) \\
&= 0.
\end{aligned}$$

Therefore, $\operatorname{Var}\left(X_n \mid X_0\right) = \mu^2 \operatorname{Var}\left(X_{n-1} \mid X_0\right) + \sigma^2 \mathbb{E}\left[X_{n-1} \mid X_0\right]$, and so

$$\begin{aligned}
\operatorname{Var}\left(X_n \mid X_0\right) &= \mu^{2n} \operatorname{Var}\left(X_0 \mid X_0\right) + \sigma^2 \sum_{i=n-1}^{2n-2} \mu^i X_0 \\
&= \sigma^2 X_0 \sum_{i=n-1}^{2n-2} \mu^i.
\end{aligned}$$

If $\mu = 1$, then $\operatorname{Var}\left(X_n \mid X_0\right) = n\sigma^2 X_0$. Otherwise,

$$\operatorname{Var}\left(X_n \mid X_0\right) = \frac{1 - \mu^n}{1 - \mu}\left(\mu^{n-1}\sigma^2\right) X_0.$$

□

Note that for any branching process, 0 is always an absorbing state, so the probability that the population becomes extinct by the n-th generation is $P\left(X_n = 0\right)$.

Definition 2.5.5 ▸ Extinction Probability

Let X be a branching process, then **extinction probability** is the probability that there is no offspring at the n-th generation, given by $u_n^{(k)} \coloneqq P\left(X_n = 0 \mid X_0 = k\right)$.

Alternatively, let $T \coloneqq \min\left\{t \in \mathbb{N} : X_t = 0\right\}$ be the time of extinction, the extinction probability is also given by $P(T \leq n \mid X_0)$. Applying first-step analysis yields

$$\begin{aligned} u_n^{(k)} &= \sum_{i \in S} P(T \leq n \mid X_1 = i) P(X_1 = i \mid X_0 = k) \\ &= \sum_{i \in S} P(T \leq n-1 \mid X_0 = i) P_{ki} \\ &= \sum_{i \in S} P_{ki} u_{n-1}^{(i)}. \end{aligned}$$

Note that this is a huge system. To simplify the calculations, consider the branching process with $X_0 = k$ as k independent branching processes with $X_0 = 1$ and let T_i be the extinction time for the i-th process. Clearly, $\{T \leq n\} = \bigcap_{i=1}^{k} \{T_i \leq n\}$. Therefore,

$$u_n^{(k)} = \prod_{i=1}^{k} P(T_i \leq n \mid X_0 = 1) = \left(u_n^{(1)}\right)^k.$$

For the sake of simplicity, we denote $u_n := u_n^{(1)}$, then $u_n^{(k)} = u_n^k$. Clearly,

$$u_n = \sum_{i \in S} P(X_1 = i \mid X_0 = 1) u_{n-1}^{(i)} = \sum_{i \in S} P(\xi = i) u_{n-1}^{i}.$$

Recall that $\phi_\xi(t) = \sum_{i \in S} P(\xi = i) t^i$, so we actually have $u_n = \phi_\xi(u_{n-1})$. This obviously means that

$$\begin{aligned} u_n &= \phi_\xi^{n-1}(u_1) \\ &= \phi_\xi^{n-1}\big(P(T \leq 1 \mid X_0 = 1)\big) \\ &= \phi_\xi^{n-1}\big(P(X_1 = 0 \mid X_0 = 1)\big) \\ &= \phi_\xi^{n-1}\big(P(\xi = 0)\big). \end{aligned}$$

So we obtain a general formula.

Proposition 2.5.6 ▸ Formula for Extinction Probability

Let $u_n^{(k)}$ be the extinction probability by the n-th generation of a branching process X with respect to ξ with $X_0 = k$, then

$$u_n^{(k)} = \phi_\xi^{n-1}\big(P(\xi = 0)\big)^k,$$

where ϕ_ξ is the probability generating function of ξ.

Furthermore, we can define

$$u_\infty^{(k)} := \lim_{n\to\infty} u_n^{(k)} = \lim_{n\to\infty} u_n^k = u_\infty^k$$

as the probability that the population will eventually extinct in finite time. Notice that

$$1 \geq P(T \leq n+1 \mid X_0 = 1) \geq P(T \leq n \mid X_0 = 1) \geq 0,$$

so by the monotone convergence theorem, $u_\infty := \lim_{n\to\infty} u_n$ always exists. By our first-step analysis, $u_n = \phi_\xi(u_{n-1})$ so if we take the limit on both sides, we have

$$u_\infty = \lim_{n\to\infty} \phi_\xi(u_{n-1}).$$

Since ϕ_ξ is continuous,

$$u_\infty = \phi_\xi\left(\lim_{n\to\infty} u_{n-1}\right) = \phi_\xi(u_\infty).$$

With some differentiation, it is easy to see that ϕ_ξ is a strictly convex function on $(0,1]$ and so the equation has at most 2 solutions.

In particular, if $P(\xi = 0) = 0$, then there is nothing to be calculated in the first place because the population will never become extinct as long as $X_0 > 0$. By definition, it is easy to see that $\phi_\xi(1) = 1$, so $u_\infty = 1$ is always a solution. However, it can be proven that if $\phi'(1) > 1$, then the other solution in $[0,1)$ is the actual probability of extinction in finite time.

Proposition 2.5.7 ▸ Probability of Extinction in Finite Time

Let X be a branching process with respect to ξ and define $u_\infty^{(k)} := \lim_{n\to\infty} u_n^{(k)}$ as the probability that the population eventually becomes extinct in finite time, then

- *if $P(\xi = 0) = 0$, then $u_\infty^{(k)} = 1$;*
- *if $P(\xi = 0) > 0$, then*
  - *if $\mathbb{E}[\xi] \leq 1$, then $u_\infty^{(k)} = 1$;*
  - *otherwise, $u_\infty \in [0,1)$ satisfies*

    $$u_\infty = \phi_\xi(u_\infty)$$

    where ϕ_ξ is the probability generating function of ξ.

> *Remark.* When $P(\xi = 0) > 0$, the branching process is called *subcritical*, *critical* and *supercritical* if $\mathbb{E}[\xi] < 1$, $\mathbb{E}[\xi] = 1$ and $\mathbb{E}[\xi] > 1$, respectively.

2.6 Poisson Process

Another important type of stochastic processes is the *Poisson process*, which is used to count the number of discrete events in a given interval and is closely related to both the Poisson distribution and the exponential distribution.

Definition 2.6.1 ▶ Poisson Process

Poisson process with **rate** or **intensity** $\lambda > 0$ is a discrete-time integer-valued stochastic process $\{X(t) : t \in \mathbb{N}\}$ such that $X(0) = 0$ and

- for any $t_0, t_1, \cdots, t_n \in \mathbb{N}$ with $t_0 = 0$, $\{X(t_i) - X(t_{i-1}) : i = 1, 2, \cdots, n\}$ are mutually independent;
- for any $s \in \mathbb{N}$ and $t \in \mathbb{N}^+$, we have $X(s+t) - X(s) \sim \operatorname{Pois}(\lambda t)$.

An immediate result from the definition is that $X(t) \sim \operatorname{Pois}(\lambda t)$ for all $t \in \mathbb{N}^+$. Recall that for a Poisson random variable, we have $\mathbb{E}[X(t)] = \operatorname{Var}(X(t)) = \lambda t$. Intuitively, the Poisson process counts the number of events in unit intervals such that $X(t)$ is the number of events up to time t. Now, each unit interval can be divided infinitesimally such that we can approximate the distribution using binomial random variables.

Theorem 2.6.2 ▶ Law of Rare Events

Let $X_n \sim \operatorname{Bin}(n, p_n)$ be a binomial random variable where $\lim_{n\to\infty} np_n = \lambda$, then as $n \to \infty$, $X_n \sim \operatorname{Pois}(\lambda)$ approximately.

Proof. Consider

$$\begin{aligned}
\lim_{n\to\infty} P(X_n = x) &= \lim_{n\to\infty} \binom{n}{x} p_n^x (1-p_n)^{n-x} \\
&\approx \lim_{n\to\infty} \frac{\prod_{i=1}^{x}(n-i+1)}{x!} \left(\frac{\lambda}{n}\right)^x \left(1-\frac{\lambda}{n}\right)^n \lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^{-x} \\
&= \lim_{n\to\infty} \frac{\prod_{i=1}^{x}(n-i+1)}{x!} \left(\frac{\lambda}{n}\right)^x \left(1-\frac{\lambda}{n}\right)^n \\
&\approx \lim_{n\to\infty} \frac{n^x \lambda^x}{x! n^x} \left(1-\frac{\lambda}{n}\right)^n \\
&= \frac{\lambda^x}{x!} \mathrm{e}^{-\lambda}.
\end{aligned}$$

Therefore, $\lim_{n\to\infty} X_n \sim \operatorname{Pois}(\lambda)$ approximately. □

The term *rare events* refers to the situation where the probability of two events occurring at the same time for some small interval h is $o(h)$, i.e., $\lim_{h\to 0} \frac{p}{h} = 0$.

Definition 2.6.3 ▶ Poisson Point Process

Let $N\big((s,t]\big)$ be the number of events in the interval $(s,t]$. It is called a **Poisson point process** of **intensity** λ if

- for any $t_0, t_1, \cdots, t_n \in \mathbb{R}$ with $t_0 = 0$ and $t_i < t_{i+1}$, $\big\{N\big((t_i, t_{i+1}]\big) : i = 1, 2, \cdots, n\big\}$ are mutually independent;
- $P\Big(N\big((t, t+h]\big) \geq 1\Big) = \lambda h + o(h)$ as $h \to 0$.
- $P\Big(N\big((t, t+h]\big) \geq 2\Big) = o(h)$ as $h \to 0$.

Remark. If the intensity λ is dependent on t, then the Poisson point process is non-homogeneous and non-stationary.

We can use a Poisson process to derive the exponential distribution by considering the waiting time of events.

Definition 2.6.4 ▶ Waiting Time

Let $X(t)$ be a Poisson process of rate λ, then the **waiting time** for the n-th event is defined as W_n such that $X(W_n - 1) = n - 1$ and $X(W_n) = n$. The **sojourn time** between two consecutive events is defined as $S_n : W_{n+1} - W_n$.

Let us consider the event $\{W_1 \leq t\}$, i.e., the event where the first event occurs by time t, then

$$P(W_1 \leq t) = P\big(X(t) \geq 1\big) = 1 - P\big(X(t) = 0\big) = 1 - \mathrm{e}^{-\lambda t}.$$

It is clear that $W_1 \sim \operatorname{Exp}(\lambda)$. This is generalised as the following:

Theorem 2.6.5 ▶ Gamma Distribution of Waiting Times

Let W_n be the waiting time for the n-th event in a Poisson process $X(t)$ where $n \in \mathbb{N}^+$, then we have $W_n \sim \operatorname{Gamma}(n, \lambda)$ where

$$f_{W_n}(t) = \frac{\lambda^n t^{n-1}}{(n-1)!}\mathrm{e}^{-\lambda t}.$$

Recall that the exponential distribution is memoryless, so it is obvious that the sojourn time $S_n = W_1$ for each $n \in \mathbb{N}$.

Theorem 2.6.6 ▶ Exponential Distribution of Sojourn Times

Let S_n be the sojourn time after the n-th event in a Poisson process of rate λ, then we have $S_n \sim \operatorname{Exp}(\lambda)$ for all $n \in \mathbb{N}$ where $f_{S_n}(t) = \lambda\mathrm{e}^{-\lambda t}$.

It can be proven that given $X(t) = 1$, the time at which this first event occurs is uniformly distributed, i.e., $W_1 \mid X(t) = 1 \sim \mathrm{U}(0, t)$. In the generalised case where $X(t) = n$, we can apply Bayes's rule to obtain

$$\begin{aligned} f\big(w_1, w_2, \cdots, w_n \mid X(t) = n\big) &= \frac{P(S_n > t - w_n) \prod_{i=1}^{n} f_{S_{i-1}}(w_i - w_{i-1})}{\frac{(\lambda t)^n \mathrm{e}^{-\lambda t}}{n!}} \\ &= \frac{n! \mathrm{e}^{-\lambda(t-w_n)} \prod_{i=1}^{n} \lambda \mathrm{e}^{-\lambda(w_i - w_{i-1})}}{(\lambda t)^n \mathrm{e}^{-\lambda t}} \\ &= \frac{n!}{t^n}. \end{aligned}$$

It can be shown that this is the joint distribution of n independent uniform random variables. If we focus on the k-th event, we can prove the following result:

Theorem 2.6.7 ▸ Arriving Time

Let $X(t)$ be a Poisson process. Given $X(t) = n$, the waiting time to the k-th event, defined by $W_k \mid X(t) = n$, has the probability density function

$$f_k(x) = \frac{n! x^{k-1} (t-x)^{n-k}}{(n-k)!(k-1)! t^n}.$$