Stochastic Modelling and Random Processes

Yiming MA

Contents

Chapter 1

Introduction to Probability Theory

1.1 Introduction

Any realistic model of a real-world phenomenon must take into account the possibility of randomness. That is, more often than not, the quantities we are interested in will not be predictable in advance but, rather, will exhibit an inherent variation that should be taken into account by the model. This is usually accomplished by allowing the model to be probabilistic in nature. Such a model is, naturally enough, referred to as a probability model.

The majority of the chapters of this book will be concerend with different probability models of natural phenomena. Clearly, in order to master both the "model building" and the subsequent analysis of these models, we must have a certain knowledge of basic probability theory. The remainder of this chapter, as well as the next two chapters, will be concerned with a study of this subject.

1.2 Sample Space and Events

uppose that we are about to perform an experiment whose outcome is not predictable in advance. However, while the outcome of the experiment will not be known in advance, let us suppose that the set of all possible outcomes is known.

Definition 1.2.1 (Sample Spaces)**.** This set of all possible outcomes of an experiment is konwn as the **sample space** of the experiment and is denoted by $\mathcal{S}$.

Some examples are the following.

1. If the experiment consits of the flipping of a coin, then

$$\mathcal{S} = \{H, T\}$$

where H means that the outcome of the toss is head and T that it is a tail.

2. If the experiment consists of rolling a die, then the sample space is

$$\mathcal{S} = \{1, 2, 3, 4, 5, 6\}$$

where the outcome i means that i appeared on the die, $i = 1, 2, 3, 4, 5, 6$.

3. If the experiment consists of flipping two coins, then the sample sapce consists of the following four points:

$$\mathcal{S} = \{(H, H), (H, T), (T, H), (T, T)\}.$$

The outcome will be (H, H) if both coins come up heads; it will be (H, T) if the first coin comes up heads and the second comes up tails; it will be (T, H) if the first comes up tails and the second heads; and it will be (T, T) if both coins come up tails.

4. If the experiment consists of rolling two dice, then the sample space consists of the following 36 points:

$$\mathcal{S} = \begin{bmatrix} (1,1) & (1,2) & (1,3) & (1,4) & (1,5) & (1,6) \\ (2,1) & (2,2) & (2,3) & (2,4) & (2,5) & (2,6) \\ (3,1) & (3,2) & (3,3) & (3,4) & (3,5) & (3,6) \\ (4,1) & (4,2) & (4,3) & (4,4) & (4,5) & (4,6) \\ (5,1) & (5,2) & (5,3) & (5,4) & (5,5) & (5,6) \\ (6,1) & (6,2) & (6,3) & (6,4) & (6,5) & (6,6) \end{bmatrix}$$

where the outcome (i, j) is said to occur if i appears on the first die and j on the second die.

5. If the experiment consists of measuring the lifetime of a car, then the sample space consists of all nonnegative real numbers. That is,

$$\mathcal{S} = [0, \infty).$$

Definition 1.2.2 (Events)**.** Any subset E of the sample space $\mathcal{S}$ is konwn as an event.

Some examples of events are the following.

1′. In Example (1), if $E = \{H\}$, then E is the event that a head appears on the flip of the coin. Similarly, if $E = \{T\}$, then E would be the event that a tail appears.

2′. In Example (2), if $E = \{1\}$, then E is the event thhat one apperas on the roll of the die. If $E = \{2, 4, 6\}$, then E would be the event that an even number appears on the roll.

3′. In Example (3), if $E = \{(H, H), (H, T)\}$, then E is the event that a head appears on the first coin.

4′. In Example (4), if $E = \{(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)\}$, then E is the event that the sum of the dice equals seven.

5′. In Example (5), if $E = (2, 6)$, then E is the event that the car lasts between two and six years.

We say that event E occurs when the outcome of the experiment lies in E. For any two events E and F of a sample space $\mathcal{S}$ we define the new event $E \cup F$ to consist of all outcomes that are either in E or in F or in both E and F. That is, the event $E \cup F$ will occur if either E or F occurs. For example, in (1) if $E = \{H\}$ and $F = \{T\}$, then

$$E \cup F = \{H, T\}.$$

That is, $E \cup F$ would be the whole sample space $\mathcal{S}$. In (2) if $E = \{1, 3, 5\}$ and $F = \{1, 2, 3\}$, then

$$E \cup F = \{1, 2, 3, 5\},$$

and thus $E \cup F$ would occur if the outcome of the die is 1 or 2 or 3 or 5.

Definition 1.2.3 (Unions)**.** The event $E \cup F$ is often referred to as the **union** of the event E and the event F.

Definition 1.2.4 (Intersections)**.** For any two events E and F, we may also define the new event EF, sometimes written $E \cap F$, and reffered to as the **intersection** of E and F.

As follows, EF consists of all outcomes which are both in E and F. That is, the event EF will occur only if both E and F occur. In Example (2), if $E = \{1, 3, 5\}$ and $F = \{1, 2, 3\}$, then

$$EF = \{1, 3\},$$

and thus EF occur if the outcome of the die is either 1 or 3. In Example (1), if $E = \{H\}$ and $F = \{T\}$, then the event EF would not consist of any outcomes and hence could not occur.

Definition 1.2.5 (The Null Event)**.** To give such an event a name, we shall refer to it as the **null event** and denote it by $\emptyset$. (That is, $\emptyset$ refers to the event consisting of no outcomes.)

Definition 1.2.6 (Mutual Exclusivity)**.** If $EF = \emptyset$, then E and F are said to be **mutually exclusive**.

We also define unions and intersections of more than two events in a similar manner. If $E_1, E_2, \cdots$ are events, then the union of these events, denoted by $\bigcup_{n=1}^{\infty} E_n$, is defined to be the event that consists of all outcomes that are in E_n for at least one value of $n = 1, 2, \cdots$. Similarly, the intersection of the events E_n, denoted by $\bigcap_{n=1}^{\infty} E_n$, is defined to be the event consisting of thoese outcomes that are in all of the events E_n, $n = 1, 2, \cdots$.

Definition 1.2.7 (The Complement)**.** Finally, for any event E we define the new event E^c, referred to as the **complement** of E, to consist of all outcomes in the sample space $\mathcal{S}$ that are not in E. That is, E^c will occur if and only if E does not occur.

In Example (4) if $E = \{(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)\}$, the E^c will occur if the sum of the dice does not equal seven.

Remark. Also note that since the experiment must result in some outcome, it follows that $\mathcal{S}^c = \emptyset$.

1.3 probabilities Defined on Events

Definition 1.3.1 (The Probability of an Event)**.** Consider an experiment whose sample sapce is $\mathcal{S}$. For each event E of the sample space $\mathcal{S}$, we assume that a number $\mathbb{P}(E)$ is defined and satisfies the following three conditions:

(i) $0 \leq \mathbb{P}(E) \leq 1$.

(ii) $\mathbb{P}(\mathcal{S}) = 1$.

(iii) For any sequence of events $E_1, E_2, \cdots$ that are mutually exclusive, that is, events for which $E_n E_m = \emptyset$ when $n \neq m$, then

$$\mathbb{P}\left(\cup_{n=1}^{\infty} E_n\right) = \sum_{n=1}^{\infty} \mathbb{P}(E_n)$$

We refer to $\mathbb{P}(E)$ as the **probability** of the event E.

Example 1.3.1. In the coin tossing example, if we assume that a head is equally likely to appear as a tail, then we would have

$$\mathbb{P}(\{H\}) = \mathbb{P}(\{T\}) = \frac{1}{2}.$$

On the other hand, if we had a biased coin and felt that a head was twice as likely to appear as a tail, then we would have

$$\mathbb{P}(\{H\}) = \frac{2}{3}, \quad \mathbb{P}(\{T\}) = \frac{1}{3}.$$

Example 1.3.2. In the die tossing example, if we supposed that all six numbers were equally likely to appear, then we would have

$$\mathbb{P}(\{1\}) = \mathbb{P}(\{2\}) = \mathbb{P}(\{3\}) = \mathbb{P}(\{4\}) = \mathbb{P}(\{5\}) = \mathbb{P}(\{6\}) = \frac{1}{6}.$$

From (iii) it would follow that the probability of getting an even number would equal

$$\begin{aligned}\mathbb{P}(\{2,4,6\}) =&\mathbb{P}(\{2\}) + \mathbb{P}(\{4\}) + \mathbb{P}(\{6\})\\ =&\frac{1}{2}.\end{aligned}$$

Remark. We have chosen to give a rather formal definition of probabilities as being functions defined on the events of a sample space. However, it turns out that these probabilities have a nice intuitive property. Namely, if our experiment is repeated over and over again then (with probability 1) the proportion of time that evetn E occurs will just be $\mathbb{P}(E)$.

Proposition 1.3.1. *For any event E, we have*

$$\mathbb{P}(E^c) = 1 - \mathbb{P}(E). \tag{1.1}$$

Proof. Since the events E and E^c are always mutually exclusive and since $E \cup E^c = S$, we have by (ii) and (iii) that

$$1 = \mathbb{P}(S) = \mathbb{P}(E \cup E^c) = \mathbb{P}(E) + \mathbb{P}(E^c).$$

□

Remark. In words, Eq.(1.1) states that the probability that an event does not occur is one minux the probability that it does occur.

We shall now derive a formula for $\mathbb{P}(E \cup F)$, the probability of all outcomess either in E or F. To do so, consider $\mathbb{P}(E) + \mathbb{P}(F)$, which is the probability of all outcomes in E plus the probability of all points in F. Since any outcome that is in both E and F will be counted twice in $\mathbb{P}(E) + \mathbb{P}(F)$ and only once in $\mathbb{P}(E \cap F)$, we must have

$$\mathbb{P}(E) + \mathbb{P}(F) = \mathbb{P}(E \cup F) + \mathbb{P}(EF),$$

or equiavlently,

$$\mathbb{P}(E \cup F) = \mathbb{P}(E) + \mathbb{P}(F) - \mathbb{P}(EF).$$

Proposition 1.3.2. *For any two events E and F, we have*

$$\mathbb{P}(E \cup F) = \mathbb{P}(E) + \mathbb{P}(F) - \mathbb{P}(EF). \tag{1.2}$$

Proof. By $E \cup F = E \cup [F/(E \cap F)]$, we have

$$\begin{aligned}\mathbb{P}(E \cup F) =&\mathbb{P}(E \cup /[F/(E \cap F)])\\ =&\mathbb{P}(E) + \mathbb{P}\left(F/(E \cap F)\right)\\ =&\mathbb{P}(E) + \mathbb{P}(F) - \mathbb{P}(E \cap F).\end{aligned}$$

□

Remark. Note that when E and F are mutually exclusive (that is, when $EF = \emptyset$), then Eq.(1.2) states that

$$\begin{aligned}\mathbb{P}(E \cup F) =&\mathbb{P}(E) + \mathbb{P}(F) - \mathbb{P}(\emptyset)\\ =&\mathbb{P}(E) + \mathbb{P}(F).\end{aligned}$$

Example 1.3.3. Suppose that we toss two coins, and suppose that we assume that each of the four outcomes in the sample space

$$\mathcal{S} = \{(H,H),(H,T),(T,H),(T,T)\}$$

is equally likely and hence has probability $\frac{1}{4}$. Let

$$E = \{(H,H)\} \quad \text{and} \quad F = \{(H,H),(T,H)\}.$$

That is, E is the event that the first coin falls heads, and F is the event that the second coin falls heads.

By Eq.(1.2) we have that $\mathbb{P}(E \cup F)$, the probability that either the first or the second coin falls heads, is given by

$$\begin{aligned}\mathbb{P}(E \cup F) =&\mathbb{P}(E) + \mathbb{P}(F) - \mathbb{P}(EF)\\ =&\frac{1}{2} + \frac{1}{2} = \mathbb{P}(\{(H,H)\})\\ =&1 - \frac{1}{4}\\ =&\frac{3}{4}.\end{aligned}$$

This probability cound, of course, have been computed directly since

$$\mathbb{P}(E \cup F) = \mathbb{P}(\{(H,H),(H,T),(T,H)\}) = \frac{3}{4}.$$

We may also calculate the probability that any one of the three events E or F or G occurs. This is done as follows:

$$\mathbb{P}(E \cup F \cup G) = \mathbb{P}\left((E \cup F) \cup G\right)$$

which by Eq.(1.2) equals

$$\mathbb{P}(E \cup F) + \mathbb{P}(G) - \mathbb{P}\left((E \cup F)G\right).$$

It is easy to check that $(E \cup F)G = (EG) \cup (FG)$. (Remember for two sets A and B, $A = B \iff A \subset B \;\&\; B \subset A$.) Hence the preceding equals

$$\begin{aligned}&\mathbb{P}(E \cup F \cup G)\\ &\quad= \mathbb{P}(E) + \mathbb{P}(F) - \mathbb{P}(EF) + \mathbb{P}(G) - \mathbb{P}(EG \cup FG)\\ &\quad= \mathbb{P}(E) + \mathbb{P}(F) - \mathbb{P}(EF) + \mathbb{P}(G) - \mathbb{P}(EG) - \mathbb{P}(FG) + \mathbb{P}(EG \cap FG)\\ &\quad= \mathbb{P}(E) + \mathbb{P}(F) - \mathbb{P}(EF) + \mathbb{P}(G) - \mathbb{P}(EG) - \mathbb{P}(FG) + \mathbb{P}(EFG) \qquad (1.3)\end{aligned}$$

Proposition 1.3.3. *In fact, it can be shown by induction that, for any n events $E_1, E_2, E_3, \cdots, E_n$,*

$$\begin{aligned}\mathbb{P}(E_1 \cup E_2 \cup \cdots \cup E_n) =& \sum_i \mathbb{P}(E_i) - \sum_{i<j} \mathbb{P}(E_iE_j) + \sum_{i<j<k} \mathbb{P}(E_iE_jE_k)\\ &- \sum_{i<j<k<l} \mathbb{P}(E_iE_jE_kE_l)\\ &+ \cdots + (-1)^{n+1}\mathbb{P}(E_1E_2\cdots E_n). \qquad (1.4)\end{aligned}$$

This identity is called the ***inclusion-exclusion identity****.*

Remark. In words, the inclusion-exclusion identity states that the probability of the union of n events equals the sum of the probability of these events taken one at a time minus the sum of the probabilities of these events taken two at a time plus the sum of the probabilities of these events taken three at a time, and so on.

1.4 Conditional probabilities

Suppose that we toss two dice and that each of the 36 possible outcomes is equally likely to occur and hence has probability $\frac{1}{36}$. Suppose that we observe that the first die is a four. Then, given this information, what is the probability that the sum of the two dice equals six? To calculate this probability we reason as follows: Given that the initial die is a four, it follows that there can be at most six possible outcomes of our experiment, namely, $(4,1),(4,2),(4,3),(4,4),(4,5)$, and $(4,6)$. Since each of these outcomes originally had the same probability of occurring, they should still have equal probabilities. That is, given that the first die is a four, then the (conditional) probability of each of the outcomes $(4,1),(4,2),(4,3),(4,4),(4,5),(4,6)$ is $\frac{1}{6}$ while the (conditional) probability of the other 30 points in the sample space is 0. Hence, the desired probability will be $\frac{1}{6}$.

If we let E and F denote, respectively, the event that the sum of the dice is six and the event that the first die is a four, then the probability just obtained is called the conditional probability that E occurs given that F has occurred and is denoted by

$$\mathbb{P}(E|F)$$

A general formula for $\mathbb{P}(E|F)$ that is valid for all events E and F is derived in the same manner as the preceding. Namely, if the event F occurs, then in order for E to occur it is necessary for the actual occurrence to be a point in both E and in F , that is, it must be in EF. Now, because we know that F has occurred, it follows that F becomes our new sample space and hence the probability that the event EF occurs will equal the probability of EF relative to the probability of F. That is,

$$\mathbb{P}(E|F) = \frac{\mathbb{P}(EF)}{\mathbb{P}(F)}. \tag{1.5}$$

Note that Eq.(1.5) is only well defined when $\mathbb{P}(F) > 0$ and hence $\mathbb{P}(E|F)$ is only defined when $\mathbb{P}(F) > 0$.

Example 1.4.1. Suppose cards numbered one through ten are placed in a hat, mixed up,and then one of the cards is drawn. If we are told that the number on the drawn card is at least five, then what is the conditional probability that it is ten?

Sol. Let E denote the event that the number of the drawn card is ten, and let F be the event that it is at least five. The desired probability is $\mathbb{P}(E|F)$. Now, from Eq.(1.5)

$$\mathbb{P}(E|F) = \frac{\mathbb{P}(EF)}{\mathbb{P}(F)}.$$

However, $EF = E$ since the number of the card will be both ten and at least five if and only if it is number ten. Hence,

$$\mathbb{P}(E|F) = \frac{\frac{1}{10}}{\frac{6}{10}} = \frac{1}{6}.$$

Example 1.4.2. A family has two children. What is the conditional probability that both are boys given that at least one of them is a boy? Assume that the sample space $\mathcal{S}$ is given by $\mathcal{S} = \{(b,b),(b,g),(g,b),(g,g)\}$, and all outcomes are equally likely. ((b,g) means, for instance, that the older child is a boy and the younger child a girl.)

Sol. Letting B denote the event that both children are boys, and A the event that at least one of them is a boy, then the desired probability is given by

$$\begin{aligned}\mathbb{P}(B|A) =& \frac{\mathbb{P}(BA)}{\mathbb{P}(A)}\\ =& \frac{\mathbb{P}(\{(b,b)\})}{\mathbb{P}(\{(b,b),(b,g),(g,b)\})} = \frac{\frac{1}{4}}{\frac{3}{4}} = \frac{1}{3}.\end{aligned}$$

Example 1.4.3. Bev can either take a course in computers or in chemistry. If Bev takes the computer course, then she will receive an A grade with probability $\frac{1}{2}$; if she takes the chemistry course then she will receive an A grade with probability $\frac{1}{3}$. Bev decides to base her decision on the flip of a fair coin. What is the probability that Bev will get an A in chemistry?

Sol. If we let C be the event that Bev takes chemistry and A denote the event that she receives an A in whatever course she takes, then the desired probability is $\mathbb{P}(AC)$. This is calculated by using Eq.(1.5) as follows:

$$\begin{aligned}\mathbb{P}(AC) =&\mathbb{P}(C)\mathbb{P}(A|C)\\ =&\frac{1}{2}=\frac{1}{6}.\end{aligned}$$

Example 1.4.4. Suppose an urn contains seven black balls and five white balls. We draw two balls from the urn without replacement. Assuming that each ball in the urn is equally likely to be drawn, what is the probability that both drawn balls are black?

Sol. Let F and E denote, respectively, the events that the first and second balls drawn are black. Now, given that the first ball selected is black, there are six remaining black balls and five white balls, and so $\mathbb{P}(E|F)=\frac{6}{11}$. As $\mathbb{P}(F)$ is clearly $\frac{7}{12}$, our desired probability is

$$\begin{aligned}\mathbb{P}(EF) =&\mathbb{P}(E)\mathbb{P}(F|E)\\ =&\frac{7}{12}\cdot\frac{6}{11}=\frac{42}{132}.\end{aligned}$$

Example 1.4.5. Suppose that each of three men at a party throws his hat into the center of the room. The hats are first mixed up and then each man randomly selects a hat. What is the probability that none of the three men selects his own hat?

Sol. We shall solve this by first calculating the complementary probability that at least one man selects his own hat. Let us denote by $E_i, i=1,2,3$, the event that the ith man selects his own hat. To calculate the probability $\mathbb{P}(E_1\cup E_2\cup E_3)$, we first note that

$$\begin{aligned}\mathbb{P}(E_i) =&\frac{1}{3}, \quad i=1,2,3\\ \mathbb{P}(E_iE_j) =&\frac{1}{6}, \quad i\neq j\\ \mathbb{P}(E_1E_2E_3) =&\frac{1}{6}.\end{aligned} \tag{1.6}$$

To see why Eq.(1.6) is correct, consider first

$$\mathbb{P}(E_iE_j)=\mathbb{P}(E_i)\mathbb{P}(E_j|E_i)$$

Now $\mathbb{P}(E_i)$, the probability that the ith man selects his own hat, is clearly $\frac{1}{3}$ since he is equally likely to select any of the three hats. On the other hand, give that the ith man selected his own hat, then there remain two hats that the jth man may select, and as one of these two is his own hat, it follows that with probability $\frac{1}{2}$ he will select it. That is, $\mathbb{P}(E_j|E_i)=\frac{1}{2}$ and so

$$\mathbb{P}(E_i,E_j)=\mathbb{P}(E_i)\mathbb{P}(E_j|E_i)=\frac{1}{3}\cdot\frac{1}{2}=\frac{1}{6}.$$

To calculate $\mathbb{P}(E_1E_2E_3)$ we write

$$\begin{aligned}\mathbb{P}(E_1E_2E_3) =&\mathbb{P}(E_1E_2)\mathbb{P}(E_3|E_1E_2)\\ =&\frac{1}{6}\mathbb{P}(E_3|E_1E_2).\end{aligned}$$

However, given that the first tow men get their own hats it follows that the third man must also get his own hat (since there are no other hats left). That is $\mathbb{P}(E_3|E_1E_2) = 1$ and so

$$\mathbb{P}(E_1E_2E_3) = \frac{1}{6}.$$

Now from Eq.(1.4) we have

$$\begin{aligned}\mathbb{P}(E_1 \cup E_2 \cup E_3) =& \mathbb{P}(E_1) + \mathbb{P}(E_2) + \mathbb{P}(E_3) - \mathbb{P}(E_1E_2) \\ & - \mathbb{P}(E_1E_3) - \mathbb{P}(E_2E_3) + \mathbb{P}(E_1E_2E_3) \\ =& 1 - \frac{1}{2} + \frac{1}{6} \\ =& \frac{2}{3}.\end{aligned}$$

Hence, the probability that none of the men selects his own hat is $1 - \frac{2}{3} = \frac{1}{3}$.

1.5 Independent Events

Definition 1.5.1 (Independence of 2 Events)**.** Two events E and F are said to be **independent** of

$$\mathbb{P}(EF) = \mathbb{P}(E)\mathbb{P}(F).$$

Remark. By Eq.(1.5) this implies that E and F are independent if $\mathbb{P}(E|F) = \mathbb{P}(E)$ (which also implies that $\mathbb{P}(F|E) = \mathbb{P}(F)$). That is, E and F are independent if knowledge that F has occurred does not affect the probability that E occurs. That is, the occurrence of E is independent of whether or not F occurs.

Definition 1.5.2 (Dependence of 2 Events)**.** Two events E and F that are not independent are said to be **dependent**.

Example 1.5.1. Suppose we toss two fair dice. Let E_1 denote the event that the sum of the dice is six and F denote the event that the first die equals four. Then

$$\mathbb{P}(E_1F) = \mathbb{P}(\{4, 2\}) = \frac{1}{36}$$

while

$$\mathbb{P}(E_1)\mathbb{P}(F) = \frac{5}{36} \cdot \frac{1}{6} = \frac{5}{216}$$

and hence E_1 and F are not independent. Intuitively, the reason for this is clear for if we are interested in the possibility of throwing a six (with two dice), then we will be quite happy if the first die lands four (or any of the numbers 1, 2, 3, 4, 5) because then we still have a possibility of getting a total of six. On the other hand, if the first die landed six, then we would be unhappy as we would no longer have a chance of getting a total of six. In other words, our chance of getting a total of six depends on the outcome of the first die and hence E_1 and F cannot be independent.

Let E_2 be the event that the sum of the dice equals seven. Is E_2 independent of F? The answer is yes since

$$\mathbb{P}(E_2F) = \mathbb{P}(\{(4, 3)\}) = \frac{1}{36}$$

while

$$\mathbb{P}(E_2)\mathbb{P}(F) = \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36}.$$

The definition of independence can be extended to more than two events.

Definition 1.5.3 (Independence of n Events)**.** The events E_1, E_2, $\cdots$, E_n are said to be **independent** if for every subset $E_{1'}, E_{2'}, \cdots, E_{r'}$, $r \leq n$, of these events

$$\mathbb{P}(E_{1'}E_{2'}\cdots E_{r'}) = \mathbb{P}(E_{1'})\mathbb{P}(E_{2'})\cdots\mathbb{P}(E_{r'}).$$

Remark. Intuitively, the events $E_1, E_2, \cdots, E_n$ are independent if knowledge of the occurrence of any of these events has no effect on the probability of any other event. But one should notice that *pairwise independence generally does not indicate joint independence*.

Example 1.5.2 (Pairwise Independent Events That Are Not Independent)**.** Let a ball be drawn from an urn containing four balls, numbered 1, 2, 3, 4. Let $E = \{1, 2\}$, $F = \{1, 3\}$, $G = \{1, 4\}$. If all four outcomes are assumed equally likely, then

$$\begin{aligned}\mathbb{P}(EF) &= \frac{1}{4} = \mathbb{P}(E)\mathbb{P}(F)\\ \mathbb{P}(EG) &= \frac{1}{4} = \mathbb{P}(E)\mathbb{P}(G)\\ \mathbb{P}(FG) &= \frac{1}{4} = \mathbb{P}(F)\mathbb{P}(G).\end{aligned}$$

However,

$$\frac{1}{4} = \mathbb{P}(EFG) \neq \mathbb{P}(E)\mathbb{P}(F)\mathbb{P}(G).$$

Hence, even though the events E, F, G are pairwise independent, they are not jointly independent.

Example 1.5.3. There are r players, with player i initially having n_i units, $n_i > 0$, $i = 1, \cdots, r$. At each stage, two of the players are chosen to play a game, with the winner of the game receiving 1 unit from the loser. Any player whose fortune drops to 0 is eliminated, and this continues until a single player has all $n \equiv \sum_{i=1}^{r} n_i$ units, with that player designated as the victor. Assuming that the results of successive games are independent, and that each game is equally likely to be won by either of its two players, find the probability that player i is the victor.

Sol.

Chapter 2

Discrete-Time Markov Chains

2.1 Stochastic Processes

Definition 2.1.1 (Stochastic Processes)**.** A **stochastic process** is a collection of random variables. That is, for each $t \in T$, $X(t)$ is a random variable.

Remark. The set T is called the **index set** of the process.

- When T is a countable set, the stochastic process is said to be a **discrete-time** process.

- If T is an uncountable set, such as an interval of the real line, the stochastic process is said to be a **continuous-time** process.

Remark. The index t is often interpreted as time, and as a result, we refer $X(t)$ as the **state** of the process at time t. And, the **state space** of a stochastic process is defined as the set of all possible values that the random variables $X(t)$ can assume.

2.2 Discrete-Time Markov Chains

Consider a process that has a value in each time period. Let X_n denote its value in time period n, and suppose we want to make a probability model for the sequence of successive values $X_0, X_1, X_2, \cdots$. The simplest model would probably be to assume that the X_n are independent random variables, but often such an assumption is clearly unjustified. For instance, starting at some time suppose that X_n represents the price of one share of some security, such as Google, at the end ot n additional trading days. Then it certainly seems unreasonable to suppose that the price at the end of day $n+1$ is independent of the prices on days n, $n-1$, $n-2$ and so down to day 0. However, it might be reasonable to suppose that the price at the end of trading day $n+1$ depends on the previous end-of-day prices only throught the price at the end of day n. That is, it might be reasonable to assume that the conditional distribution of X_{n+1} given all the past end-of-day prices $X_n, X_{n-1}, \cdots, X_0$ depends on these past prices only through the price at the end of day n. Such an assumption defines a Markov chain, a type of stochastic process that will be studied in this chapter, and which we now formally define.

Definition 2.2.1 (Markov Chains)**.** Let $\{X_n, n = 0, 1, 2, \cdots\}$ be a stochastic process that takes on a finite or countable number of possible values. Denote the set of all these values by S. The process is called a **Markov chain** if for all $A \subset S$, $n \in \mathbb{N}$ and $s_0, \cdots, s_n \in S$,

$$\mathbb{P}[X_{n+1} \in A | X_n = s_n, \cdots, X_0 = s_0] = \mathbb{P}[X_{n+1} \in A | X_n = s_n].$$

- S is called the **state space** of the process.

- If $X_n = i$, for some $n \in \mathbb{N}$ and $i \in S$, then the process is said to be in state i at time n.
- The property that the the conditional probability distribution of future states of the process (conditional on both past and present states) depends only upon the present state, not on the sequence of events that preceded it, is called the **Markov property**.

Definition 2.2.2 (Homogeneity)**.** A Markov process is called **homogeneous** if

$$\mathbb{P}[X_{n+1} = a | X_n = b],$$

$\forall a, b \in S$, is independent of n.

Theorem 2.2.1. *Markov property implies the past variable X_{n-1} and the future variable X_{n+1} are independent conditional on the present X_n.*

Proof.

$$\begin{aligned}
&\mathbb{P}[X_{n-1} = s_{n-1}, X_{n+1} = s_{n+1} | X_n = s_n] \\
=&\frac{\mathbb{P}[X_{n-1} = s_{n-1}, X_n = s_n, X_{n+1} = s_{n+1}]}{\mathbb{P}[X_n = s_n]} \\
=&\frac{\mathbb{P}[X_{n-1} = s_{n-1}]}{\mathbb{P}[X_n = s_n]}\mathbb{P}[X_n = s_n | X_{n-1} = s_{n-1}]\mathbb{P}[X_{n+1} = s_{n+1} | X_n = s_n, X_{n-1} = s_{n-1}] \\
=&\frac{\mathbb{P}[X_{n-1} = s_{n-1}]}{\mathbb{P}[X_n = s_n]}\mathbb{P}[X_n = s_n | X_{n-1} = s_{n-1}]\mathbb{P}[X_{n+1} = s_{n+1} | X_n = s_n] \\
=&\frac{\mathbb{P}[X_{n-1} = s_{n-1}]}{\mathbb{P}[X_n = s_n]}\frac{\mathbb{P}[X_{n-1} = s_{n-1} | X_n = s_n]\mathbb{P}[X_n = s_n]}{\mathbb{P}[X_{n-1} = s_{n-1}]}\mathbb{P}[X_{n+1} = s_{n+1} | X_n = s_n] \\
=&\mathbb{P}[X_{n-1} = s_{n-1} | X_n = s_n]\mathbb{P}[X_{n+1} = s_{n+1} | X_n = s_n].
\end{aligned}$$

□

Definition 2.2.3 (Transition Matrices)**.** If a discrete-time Markov chain is homogeneous, then we can define its **transition matrix** P by setting

$$P_{i,j} = \mathbb{P}[X_{n+1} = j | X_n = i].$$

- The transition matrix of a homogeneous discrete-time Markov chain is made of constants due to homogeneity.
- The dimension of the transition matrix is $\#S \times \#S$, where $\#S$ denotes the size of the state space.
- The sum of each row of P is 1:

  $$\sum_{j \in S} P_{i,j} = 1.$$

  So $\vec{1}$ is an eigenvector of P with eigenvalue 1.

Example 2.2.1 (Forecasting the Weather)**.** Suppose that the chance of rain tomorrow depends on previous weather conditions only through whether or not it is raining today and not on past weather conditions. Suppose also that if it rains today, then it will rain tomorrow with probability α; and if it does not rain today, then it will rain tomorrow with probability β. If we say that the process is in state 0 when it rains and state 1 when it does not rain, then the preceding is a two-state Markov chain whose transition probabilities are given by

$$\begin{bmatrix} \alpha & 1-\alpha \\ \beta & 1-\beta \end{bmatrix}.$$

Example 2.2.2 (Random Walk with Boundaries)**.** Let $\{X_n : n \in \mathbb{N}\}$ be a simple random walk on $S = \{1, \cdots, L\}$ with

$$\mathbb{P}[X_{n+1} = j | X_n = i] = \begin{cases} p & j = i+1, \, 1 \leq i \leq L-1 \\ q & j = i-1, \, 2 \leq i \leq L \\ 0 & \text{Otherwise} \end{cases}.$$

The boundary conditions are

- **periodic** if $P_{L,1} = p$, $P_{1,L} = q$;
- **absorbing** if $P_{L,L} = 1$, $P_{1,1} = 1$;
- **closed** if $P_{L,L} = p$, $P_{1,1} = q$;
- **reflecting** if $P_{L,L-1} = 1$, $P_{1,2} = 1$.

Definition 2.2.4 (Absorbing Stetes)**.** A state $s \in S$ is called **absorbing** for a discrete-time Markov chain with transition matrix P, if

$$P_{s,y} = \delta_{s,y}, \text{ for all } y \in S.$$

Example 2.2.3 (Random Walk with Absorbing Boundary Conditions)**.** Consider the simple random walk model $\{X_n : n \in \mathbb{N}\}$ on $S = \{1, \cdots L\}$, with probability p of moving right and probability q of moving left, and also assume the boundary conditions are absorbing. Let h_k be the absorption probability for $X_0 = k \in S$, i.e.

$$h_k = \mathbb{P}[X_n \in \{1, L\} \text{ for some } n \geq 0 | X_0 = k].$$

Conditioning on the first jump, we have

$$\begin{aligned} h_k =& p \times \mathbb{P}[X_n \in \{1, L\} \text{ for some } n \geq 0 | X_0 = k, X_1 = k+1] \\ & + q \times \mathbb{P}[X_n \in \{1, L\} \text{ for some } n \geq 0 | X_0 = k, X_1 = k-1] \\ =& p \times \mathbb{P}[X_n \in \{1, L\} \text{ for some } n \geq 0 | X_1 = k+1] \\ & + q \times \mathbb{P}[X_n \in \{1, L\} \text{ for some } n \geq 0 | X_1 = k-1] \\ =& p h_{k+1} + q h_{k-1}, \end{aligned}$$

for $k = 2, \cdots, L-1$. And obviously, we have $h_1 = h_L = 1$. The characteristic equation for the above linear recurrence relation is

$$\lambda = p\lambda^2 + q,$$

whose solutions are $\lambda_1 = 1$, $\lambda_2 = q/p$. So the general solution to the recursion is

$$h_k = c_1 + c_2 \left(\frac{q}{p}\right)^k,$$

where $c_1, c_2 \in \mathbb{R}$ are constants. By the initial conditions $h_1 = h_k = 1$, we konw the coefficients $c_1 = 1$ and $c_2 = 0$. Thus, we have $h_k \equiv 1$

2.3 Chapman-Kolmogorov Equations

We have already defined the one-step transition probabilities $P_{i,j}$. We now define the n-step transition probabitlies $P_{i,j}^n$.

Definition 2.3.1. $P_{i,j}^n$ is defined to be the probability that a homogeneous process in state i will be in state j after n additional transitions. That is,

$$P_{i,j}^n = \mathbb{P}[X_{n+k} = j | X_k = i], \qquad n \geq 0,\, i, j \in S.$$

Remark. Two simple observations are

- $P_{i,j}^1 = P_{i,j}$,
- $P_{i,j}^0 = \delta(i, j)$.

Theorem 2.3.1 (Chapman-Kolmogorov Equations)**.** *The **Chapman-Kolmogorov equations** provide a method for computing n-step transition probabilities. These equations are*

$$P_{i,j}^{n+m} = \sum_{k \in S} P_{i,k}^n P_{k,j}^m \qquad \textit{for all } n, m \geq 0, \textit{ and all } i, j \in S. \tag{2.1}$$

Proof. Formally, we have

$$\begin{aligned} P_{i,j}^{n+m} =& \mathbb{P}[X_{n+m} = j | X_0 = i] \\ =& \sum_{k \in S} \mathbb{P}[X_{n+m} = j, X_n = k | X_0 = i] \\ =& \sum_{k \in S} \mathbb{P}[X_{n+m} = j | X_n = k, X_0 = i] \mathbb{P}[X_n = k | X_0 = i] \\ =& \sum_{k \in S} \mathbb{P}[X_{n+m} = j | X_n = k] \mathbb{P}[X_n = k | X_0 = i] \\ =& \sum_{k \in S} P_{k,j}^m P_{i,k}^n. \end{aligned}$$

□

Corollary 2.3.1. *If we let $P^{(n)}$ denote the matrix of n-step transition probabilities $P_{i,j}^n$, then (2.1) asserts that*

$$P^{(n+m)} = P^{(n)} P^{(m)}.$$

Hence, in particular,

$$P^{(2)} = PP = P^2$$

and by induction

$$P^{(n)} = P^n.$$

Example 2.3.1. Consider Example 2.2.1 in which the weather is considered as a two-state Markov chain. If $\alpha = 0.7$ aans $\beta = 0.4$, then calculate the probability that it will train four days from today given that it is raining today. *Sol.* The one-step transition probability matrix is given by

$$P = \begin{bmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{bmatrix}$$

Hence,

$$P^{(2)} = P^2 = \begin{bmatrix} 0.61 & 0.39 \\ 0.52 & 0.48 \end{bmatrix},$$

$$P^{(4)} = (P^{(2)})^2 = \begin{bmatrix} 0.5749 & 0.4251 \\ 0.5668 & 0.4332 \end{bmatrix}$$

and the desired probability $P^4_{0,0}$ equals 0.5749.

Remark. Suppose the initial distribution (i.e. the distribution of X_0) is given by $\pi_0(x) = \mathbb{P}[X_0 = x]$, then the distribution of X_n is then

$$\pi_n(x) = \sum_{y \in S} \sum_{s_1 \in S} \cdots \sum_{s_{n-1} \in S} \pi_0(y) P_{y,s_1} \cdots P_{s_{n-1},x}.$$

If we represent the distribution of X_n as a row vector $\boldsymbol{\pi}_n$, then

$$\boldsymbol{\pi}_n = \boldsymbol{\pi}_0 P^n.$$

Now consider a discrete-time Markov chain on a finite state space with $|S| = L$, let $\lambda_1, \cdots, \lambda_L \in \mathbb{C}$ be the eigenvalues of the transition matrix P with corresponding left (row) eigenvectors $\langle u_i|$ and right (column) eigenvector $|v_i\rangle$, $i = 1, \cdots L$, in bracket notation. Assuming that all eigenvalues are distinct, we have

$$P = \sum_{i=1}^{L} \lambda_i |v_i\rangle\langle u_i| \quad \text{and} \quad P^n = \sum_{i=1}^{L} \lambda_i^n |v_i\rangle\langle u_i,$$

since eigenvectors can be chosen orthonomal $\langle u_i|v_j\rangle = \delta_{i,j}$.

Since $\langle \boldsymbol{\pi}_n| = \langle \boldsymbol{\pi}_0|P^n$, we get

$$\langle \boldsymbol{\pi}_n| = \langle \boldsymbol{\pi}_n|v_1\rangle \lambda_1^n \langle u_1| + \cdots + \langle \boldsymbol{\pi}_0|v_L\rangle \lambda_L^n \langle u_L|.$$

- By the Gershgorin theorem, we know there exists at least one eigenvalue λ_i such that $|\lambda_i| \leq 1$. Contributions with such λ_i decay exponentially.
- $\lambda_1 = 1$ corresponds to the stationary distribution $\langle \pi| = \langle u_1|$ and $|v_1\rangle = \vec{1}$.
- Other $\lambda_i \neq 1$ with $|\lambda_1| = 1$ corresponds to persistent oscillations.

Definition 2.3.2 (Lazy Versions)**.** Let $\{X_n : n \in \mathbb{N}\}$ be a discrete-time Markov chain with transition matrix P. The DTMC with transition matrix

$$P^\epsilon_{i,j} = \epsilon \delta_{i,j} + (1-\epsilon) P_{i,j},\ \epsilon \in (0,1)$$

is called a **lazy version** of the original chain.

- P^ϵ has the same eigenvectors as P with eigenvalues $\lambda_i^\epsilon = \lambda_i(1-\epsilon) + \epsilon$ since

$$\langle u_i|P^\epsilon = \epsilon\langle u_i| + \lambda(1-\epsilon)\langle u_i| \qquad \text{analogously for } |v_i\rangle$$

- This implies $|\lambda_i^\epsilon| < |\lambda_i| \leq 1$ unless $\lambda_i = 1$. Such a matrix P^ϵ is called **aperiodic**, and there are no persistent oscillations. (See Section 2.5.1.)
- The stationary distribution is unique if and only if the eigenvalue $\lambda = 1$ is unique (has multiplicity 1), which is independent of lazyness (discussed later).

2.4 Classification of States

Definition 2.4.1 (Accessibility)**.** State $j4$ is said to be **accessible** from state i if $P_{i,j}^n > 0$ for some $n \geq 0$.

Theorem 2.4.1. *State j is accessible from state i if and only if, starting in i, it is possible that the state will ever enter state j.*

Proof. The necessity is implied by the definition of accessibility. Now, suppose the process will ever enter state j from i with a positive probability, but j is not accessible from i, then

$$\begin{aligned}\mathbb{P}[\text{ever be in } j|\text{start in } i] =&\mathbb{P}\left\{\bigcup_{n=0}^{\infty}\{X_n = j\}|X_0 = i\right\}\\ \leq&\sum_{n=0}^{\infty}\mathbb{P}[X_n = j|X_0 = i]\\ =&\sum_{n=0}^{\infty}P_{i,j}^n\\ =&0,\end{aligned}$$

which is a contradiction. □

Definition 2.4.2 (Communication)**.** Two states i and j that are accessible to each other are said to **communicate**, and we write $i \leftrightarrow j$.

Remark. Note that any state communicates with itself, by definition,

$$P_{i,i}^0 = \mathbb{P}[X_n = i|X_n = i] = 1.$$

Theorem 2.4.2 (Properties of Communication)**.** *The raltion of communication satisfies the following three properties:*

1. *State i communicates with state i, $\forall i \in S$.*
2. *If state i communicates with state j, then state j communicates with state i.*
3. *If state i communicates with state j, and state j communicates with state k, then state i communicates with state k.*

Proof. Properties 1 and 2 follow immediately fromthe definition of communication. To Prove 3 suppose that i communicates with j, and j communicates with k. Thus, there exsit interger n and m such that $P_{i,j}^n > 0$, $P_{j,k}^m > 0$. Now by the Chapman-Kolmogorov equations, we have

$$P_{i,k}^{n+m} = \sum_{r\in S} P_{i,r}^n P_{i,r}^m \geq P_{i,j}^n P_{j,k}^m > 0.$$

Hence, state k is accessible from state i. Similarly. we can show that state i is accessible from state k. Hence, state i and k communicate. □

Definition 2.4.3 (Classes)**.** Two state that communicate are said to be in the same **class**.

Remark. It is an easy consequence of properties 1, 2, 3 that any two classes of states are either identical or disjoint. In other words, the concept of communication divides the state space up into a number of separate classes.

Definition 2.4.4 (Irreducibility)**.** The Markov chain is said to be **irreducible** if there is only one class, that is, if all states communicates with each other.

For any state i, we let f_i denote the probability that, starting in state i, the process will ever reenter state i, i.e.

$$f_i = \mathbb{P}[X_n = i \text{ for some } n > 0 | X_0 = i].$$

Definition 2.4.5 (Recurrence & Transience)**.** State i is said to be **recurent** if $f_i = 1$ and **transient** if $f_i < 1$.

In Stephan's notes, the difinitions of recurrence and transience are the same as above, but new notations are introduced.

Definition 2.4.6 (Recurrence & Transience)**.** Let $T_x = \inf\{n \geq 1 : X_n = x\}$, then state x is called

- **transient**, if $\mathbb{P}[T_x = \infty | X_0 = x] > 0$;
- **recurrent**, if $\mathbb{P}[T_x < \infty | X_0 = x] = 1$.

Suppose that the process starts in state i and i is recurent. Hence, with probability 1, the process will eventually reenter state i. However, by the definition of a homogeneous Markov chain, it follows that the process will be starting over again when it reenters state i and, therefore, state i will eventually be visited again. Continual repetition of this argument leads to the following conclusion.

Remark. If state i is recurrent then, starting in state i, the process will reenter state i again and again and again — in face, infinitely often.

On the other hand, suppose that state i is transient. Hence, each time the process enters state i there will be a positive probability, namely, $1 - f_i$, that it will never again enter the state. Therefore, starting in state i, the probability that the process will be in state i for exactly n time periods equals $f_i^{n-1}(1 - f_i)$, $n \geq 1$.

Remark. If state i is transient then, starting in state i, the number of time periods that the process will be in state i has a *geometric* distribution with finite mean $1/(1 - f_i)$.

From the preceding two paragraphs, it follows that state i is recurrent if and only if, starting in state i, the expected number of time periods that the process is in state i is infinite. That is, Let

$$I_n = \begin{cases} 1, & \text{if } X_n = i \\ 0, & \text{if } X_n \neq i \end{cases}$$

then state i is recurrent if and only if

$$E\left[\sum_{n=0}^{\infty} I_n | X_0 = i\right] = \infty,$$

and it is transient if and only if

$$E\left[\sum_{n=0}^{\infty} I_n | X_0 = i\right] < \infty.$$

But one question is how to calculate the expectation of the infinite sum of the indicator variables I_n. To get this term, notice that our indicator variables are nonnegative, and using Fubini's theorem for nonnegative functions, we have

$$\begin{aligned} E\left[\sum_{n=0}^{\infty} I_n|X_0=i\right] &= \sum_{n=0}^{\infty} E[I_n|X_0=i] \\ &= \sum_{n=0}^{\infty} \mathbb{P}[X_n=i|X_0=i] \\ &= \sum_{n=0}^{\infty} P_{i,i}^n. \end{aligned}$$

Proposition 2.4.1. *For a homogeneous discrete-time Markov chain with its state space S and transition matrix P, for any state $i \in S$, i is recurrent if and only if*

$$\sum_{n=0}^{\infty} P_{i,i}^n = \infty,$$

and transient if and only if

$$\sum_{n=0}^{\infty} P_{i,i}^n < \infty.$$

The argument leading to the proceding proposition is doubly important because it also shows that a transient state will only be visited a finite number of times (hence the name transient). This leads to the conclusion that *in a finite-state Markov chain not, not all states can be transient.*

To see this, suppose the states are $0, 1, \cdots, M$ and suppose that they are all transient. Then after a finite amount of time (say, after time T_0) state 0 will never be visited, and after a time (say T_1) state 1 will never be visited, and after a time (say T_2) state 2 will never be visited, and so on. Thus, after a finite time $T = \max\{T_0, T_1, \cdots, T_M\}$ no state will be visited. But as the process must be in some states after time T we arrive at a contradiction, which shows that at leasat one of the states must be recurrent.

Another use of Proposition 2.4.1 is that it enables us to show that recurrence is a class property.

Corollary 2.4.1. *If state i is recurrent, and state i communicates with state j, then state j is recurrent.*

Proof. To prove this we first note that, since state i communicates with state j, there exist intergers k and m such that $P_{i,j}^k > 0$, $P_{j,i}^m > 0$. Now for any integer n

$$P_{j,j}^{m+n+k} \geq P_{j,i}^m P_{i,i}^n P_{i,j}^k.$$

This follows since the left side of the preceding is the probability of going from j to j in $m+n+k$ steps, while the right side is the probability of going from j to j in $m_n + k$ steps via a path that goes from j to i in m steps, then from i to i in an additional steps, then from i to j in an additional k steps.

From the preceding we obtain, by summing over n, that

$$\sum_{n=1}^{\infty} P_{j,j}^n \geq \sum_{n=1}^{\infty} P_{j,j}^{m+n+k} \geq P_{j,i}^m P_{i,j}^k \sum_{n=1}^{\infty} P_{j,j}^n = \infty$$

since $P_{j,i}^m P_{i,j}^k > 0$ and $\sum_{n=1}^{\infty} P_{j,j}^n$ is infinite since state i is recurrent. Thus, by Proposition 2.4.1 it follows that state j is also recurrent. □

Remark. 1. Corollary 2.4.1 also implies that *transience is a class propertya.* For if state i is transient and communicates with state j , then state j must also be transient. For if j were recurrent then, by Corollary 2.4.1, i would also be recurrent and hence could not be transient.

2. Corollary 2.4.1 along with our previous result that not all states in a finite Markov chain can be transient leads to the conclusion that *all states of a finite irreducible Markov chain are recurrent.*

2.5 Long-Run Proportions and Limiting Probabilities

For pairs of states $i \neq j$, let $f_{i,j}$ denote the probability that the Markov chain, starting in state i, will ever make a transition into state j. That is,

$$f_{i,j} = \mathbb{P}[X_n = j \text{ for some } n > 0 | X_0 = i].$$

We then have the following result.

Proposition 2.5.1. *If i is recurrent and i communicates with j, then $f_{i,j} = 1$.*

Proof. Because i and j communicate there is a value n such that $P_{i,j}^n > 0$. Let $X_0 = i$ and say that the first opportunity is a success if $X_n = j$, and note that the first opportunity is a success with probability $P_{i,j}^n > 0$. If the first opportunity is not a success, then consider the next time (after time n) that the chain enters state i. (Because state i is recurrent we can be certain that it will eventually reenter state i after time n.) Say that the second opportunity is a success if n time periods later the Markov chain is in state j . If the second opportunity is not a success then wait until the next time the chain enters state i and say that the third opportunity is a success if n time periods later the Markov chain is in state j.

Continuing in this manner, we can define an unlimited number of opportunities, each of which is a success with the same positive probability $P_{i,j}^n$. Because the number of opportunities until the first success occurs is geometric with parameter $P_{i,j}^n$, it follows that with probability 1 a success will eventually occur and so, with probability 1, state j will eventually be entered. □

If state j is recurrent, let m_j denote the expected number of transitions that it takes the Markov chain when starting in state j to return to that state. That is, with

$$N_j = \min\{n > 0 : X_n = j\}$$

equal to the number of transitions until the Markov chain makes a transition into state j ,

$$m_j = \mathbb{E}[N_j | X_0 = j].$$

Remark. In Stephan's notes, there is a similar notation to N_j, which is called T_x and defined by

$$T_x = \inf\{n > 0 : T_n = x\}.$$

T_x is called the **first return time** to state x.

If state j is recurrent, then these two notations are equivalent. Otherwise, N_j may not exist and T_j in this case will be ∞.

Definition 2.5.1 (Positive Recurrence & Null Recurrence)**.** Say that the recurrent state j is **positive recurent** if $m_j < \infty$ and say that it is **null recurrent** if $m_j = \infty$.

Stephan's definitions are similar.

Definition 2.5.2 (Transient & Positive Recurrence & Null Recurrence)**.** Let T_x be the **first return time** to state x (i.e. $T_x = \inf\{n > 0 : X_n = x\}$). A state $x \in S$ is called

- **transient**, if $\mathbb{P}[T_x = \infty] > 0$;
- **positive recurent**, if $\mathbb{P}[T_x < \infty] = 1$ and $\mathbb{E}[T_x|X_0 = x] < \infty$;
- **null recurent**, if $\mathbb{P}[T_x < \infty] = 1$ and $\mathbb{E}[T_x|X_0 = x] = \infty$.

Now suppose that the Markov chain is irreducible and recurrent. In this case we now show that *the long-run proportion of time that the chain spends in state j is equal to $1/m_j$.* That is, letting π_j denote the **long-run proportion of time** that the Markov chain is in state j , we have the following proposition.

Proposition 2.5.2. *If the Markov chain is irreducible and recurrent, then for any initial state*

$$\pi_j = 1/m_j.$$

Proof. Suppose that the Markov chain starts in state i, and let T_1 denote the number of transitions until the chain enters state j; then let T_2 denote the additional number of transitions from time T_1 until the Markov chain next enters state j; then let T_3 denote the additional number of transitions from time T_1+T_2 until the Markov chain next enters state j, and so on. Note that T_1 is finite because Proposition 2.5.1 tells us that with probability 1 a transition into j will eventually occur. Also, for $n \geq 2$, because T_n is the number of transitions between the $(n1)$th and the nth transition into state j , it follows from the Markovian property that T_2 , T_3 , $\cdots$ are independent and identically distributed with mean m_j. Because the nth transition into state j occurs at time $T_1 + \cdots + T_n$ we obtain that π_j , the long-run proportion of time that the chain is in state j, is

$$\begin{aligned}\pi_j &= \lim_{n\to\infty} \frac{n}{\sum_{i=1}^n T_i} \\ &= \lim_{n\to\infty} \frac{1}{\frac{1}{n}\sum_{i=1}^n T_i} \\ &= \lim_{n\to\infty} \frac{1}{\frac{T_1}{n} + \frac{T_2+\cdots+T_n}{n}} \\ &=\frac{1}{m_j},\end{aligned}$$

where the last inequality follows because $\lim_{n\to\infty} T_1/n = 0$ and, from the strong law of large numbers, $\lim_{n\to\infty} \frac{T_2+\cdots+T_n}{n} = \lim_{n\to\infty} \frac{T_2+\cdots+T_n}{n-1}\frac{n-1}{n} = m_j$. □

Remark. Because $m_j < \infty$ is equivalent to $1/m_j > 0$, it follows from the preceding that state j is positive recurrent if and only if $\pi_j > 0$.

We now exploit the above remark to show that positive recurrence is a class property.

Proposition 2.5.3. *If i is positive recurrent and $i \leftrightarrow j$ then j is positive recurrent.*

Proof. Suppose that i is positive recurrent and that $i \leftrightarrow j$. Now, let n be such that $P^n_{i,j} > 0$. Because π_i is the long-run proportion of time that the chain is in state i, and $P^n_{i,j}$ is the probability that it will be in state j after n transitions from i.

$$\begin{aligned}\pi_i P^n_{i,j} =&\text{long-run proportion of time the chain is in } i \\ &\text{and will be in } j \text{ after } n \text{ transitions} \\ =&\text{long-run proportion of time the chain is in } j \\ &\text{and was in } i\ n \text{ transitions ago} \\ \leq&\text{long-run proportion of time the chain is in } j.\end{aligned}$$

Hence, $\pi_j \geq \pi_i P^n_{i,j} > 0$, showing that j is positive recurrent. □

Remark.

1. It follows from the preceding result that *null recurrence is also a class property.* For suppose that i is null recurrent and $i \leftrightarrow j$. Because i is recurrent and $i \leftrightarrow j$ we can conclude that j is recurrent. But if j were positive recurrent then by the preceding proposition i would also be positive recurrent. Because i is not positive recurrent, neither is j.

2. *An irreducible finite state Markov chain must be positive recurrent.* For we know that such a chain must be recurrent; hence, all its states are either positive recurrent or null recurrent. If they were null recurrent then all the long run proportions would equal 0, which is impossible when there are only a finite number of states. Consequently, we can conclude that the chain is positive recurrent.

3. The classical example of a null recurrent Markov chain is the one dimensional symmetric random walk.

To determine the long-run proportions $\{\pi_j, j \geq 1\}$, note, because π_i is the long-run proportion of transitions that come from state i, that

$$\pi_i P_{i,j} = \text{long-run proportion of transitions that go from state } i \text{ to state } j,$$

summing the preceding over all i now yields that

$$\pi_j = \sum_{i \in S} \pi_i P_{i,j}.$$

Indeed, the following important theorem can be proven.

Theorem 2.5.1. *Consider an irreducible Markov chain. If the chain is positive recurrent, then the long-run proportions are the unique solution of the equations*

$$\begin{aligned} \pi_j &= \sum_{i \in S} \pi_i P_{i,j} \\ \sum_{j \in S} \pi_j &= 1 \end{aligned}$$

Moreover, if there is no solution of the preceding linear equations, then the Markov chain is either transient or null recurrent and all $\pi_j = 0$.

Example 2.5.1. Consider Example 2.2.1, in which we assume that if it rains today, then it will rain tomorrow with probability ; and if it does not rain today, then it will rain tomorrow with probability β. If we say that the state is 0 when it rains and 1 when it does not rain, then by Theorem 2.5.1 the long-run proportions π_0 and π_1 are given by

$$\begin{aligned} \pi_0 &= \alpha\pi_0 + \beta\pi_1, \\ \pi_1 &= (1-\alpha)\pi_0 + (1-\beta)\pi_1, \\ \pi_0 + \pi_1 &= 1, \end{aligned}$$

which yields that

$$\pi_0 = \frac{\beta}{1+\beta-\alpha}, \ \pi_1 = \frac{1-\alpha}{1+\beta-\alpha}.$$

For example, if $\alpha = 0.7$ and $\beta = 0.4$, then the long-run proportion of rain is $\pi_0 = \frac{4}{7} = 0.571$.

Proposition 2.5.4. *If the initial state is chosen according to the probabilities π_j, then the probability of being in state j at any time n is also equal to π_j. That is, if*

$$\mathbb{P}[X_0 = j] = \pi_j, \forall j \in S,$$

then

$$\mathbb{P}[X_n = j] = \pi_j, \forall n \geq 0.$$

Proof. This can be easily proven by induction, for it is true when $n = 0$, and if we suppose it is true for $n - 1$, then writing

$$\begin{aligned}\mathbb{P}[X_n = j] =& \sum_{i\in S} \mathbb{P}[X_n = j|X_{n-1} = i]\mathbb{P}[X_{n-1} = i] \\ =& \sum_{i\in S} P_{i,j}\pi_i \qquad \text{by induction hypothesis} \\ =& \pi_j \qquad \text{by Theorem 2.5.1}\end{aligned}$$

□

Definition 2.5.3 (Stationary Probabilities)**.** The long-run propoertions π_j, are aften called **stationary probabilities**.

Definition 2.5.4 (Stationary MC)**.** A MC is **stationary** if it has stationary probabilitiy $\boldsymbol{\pi}$.

Theorem 2.5.2 (Existence of Stationary Probabilities)**.** *Every homogeneous discrete-time MC with a finite state space has a stationary probability distribution.*

Proof. Let $\Delta = \{\text{probabilities } \boldsymbol{\pi} \text{ on the MC } |\boldsymbol{\pi}\vec{1} = 1, \boldsymbol{\pi}_i \geq 0\}$. Since $P_{i,j} \geq 0$ and $P\vec{1} = \vec{1}$, we have $\boldsymbol{\pi} \in \Delta \implies \boldsymbol{\pi} P \in \Delta$.

Notice that Δ is compact (which is equivalently closed and bounded, since it is in a finite linear space) and convex, and P is continuous (actually it is linear). By Brouwer fixed-point theorem, P has a fixed point $\boldsymbol{\pi}^*$ in Δ, i.e. $\boldsymbol{\pi}^* P = \boldsymbol{\pi}^*$. □

Remark. The uniqueness of stationary distributions is not guaranteed. And note that any convex combination of two stationary distributions is also a stationary distributions.

The proof of Theorem 2.5.2 is not constructive, and an even better way is to prove that the mean fraction of time spent in each state from a given initial state is a stationary probability. But we need first a decomposition theorem for homogeneous finite discrete-time Markov chain.

Definition 2.5.5 (Cycles)**.** A **cycle** is a closed path (or walk) in S along the graph of allowed transitions (defined by P) of length greater than 0.

Robert used the following definitions for transience and recurrence. which are equivalent to those previously defined but more practical.

Definition 2.5.6 (Transience & Recurrence)**.** Say $i \in X$ is **transient** if $\nexists$ cycles throught i. Otherwise, i is **recurrent**.

Definition 2.5.7 (Communication)**.** Say i, j **communicates**, and denote this as $i \leftrightarrow j$, if there exists a cycle through i and j.

Proposition 2.5.5. *Communication is an equivalent relation on the set of recurrent states:*

- $i \leftrightarrow i$;
- $i \leftrightarrow j \iff j \leftrightarrow i$;
- $i \leftrightarrow j,\ j \leftrightarrow k \implies i \leftrightarrow k$.

Definition 2.5.8 (Communicating Components)**.** Equivalent classes are called **communicating components**.

Definition 2.5.9 (Absorbing Communicating Components)**.** A communicating componen is **absorbing** if it is impossible to leave it.

The transition graph for a MC can be quotiented by communication relations, and the resulting graph is acyclic (DAG). The basal communicating components are absorbing ones.

Finiteness of a MC indicates that there exists an absorbing communicating component. We can reduce a MC to an absorbing component, and the result is called an **irreducible** MC.

Let $x \in A$, where A is an absorbing component, and let m_x be the mean time to return to x conditional on starting at x. Finiteness implies $m_x < \infty$.

For an arbitrary $y \in S$, let $\gamma_x(y)$ be the mean time in y before return to x given that the process starts at x. Notice that $\gamma_x(y) \geq 0$ and $\gamma_x(y) = 0$ if $y \notin A$. Then $\gamma_x P = \gamma_x$. (Use $\mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[X]$.) Let $\boldsymbol{\pi}_x(y) = \gamma_x(y)/m_x$, then $\boldsymbol{\pi}_x \vec{1} = 1$. So $\boldsymbol{\pi}_x$ is a stationary probability.

Remark. Next, $\boldsymbol{\pi}_x = \boldsymbol{\pi}_y$ if and only if x, y are in the same absorbing component A, so denote it by $\boldsymbol{\pi}_A$.

Remark. Also,

$$\boldsymbol{\pi}_A = \begin{cases} \frac{1}{m_x} & \text{for } x \in A \\ 0 & \text{for } x \notin A \end{cases}$$

Furthermore, $\boldsymbol{\pi}_A$ is DS-ergodic.

Definition 2.5.10 (Dynamical-System Ergodicty)**.** A stationary probability is **dynamical-system ergodic** if it is not a convex combination of other stationary probabilities.

Theorem 2.5.3 (DS-Ergodic Theorem)**.** *If $\boldsymbol{\pi}$ is DS-ergodic, then $\forall x \in S$ with $\boldsymbol{\pi}(x) > 0$, the fraction of times $0, \cdots, T-1$ spent in any $y \in S$ given that the process starts in x converges almost surely to $\boldsymbol{\pi}(y)$ as $T \to \infty$.*

Remark. If a *finite discrete-time* MC has a *unique absorbing component* A, then the fraction of time spent in y given that the process starts anywhere converges *almost surely* to $\boldsymbol{\pi}_A(y)$ and $\boldsymbol{\pi}_A$ is the *only* stationary probability. DS theorists would say the MC is **uniquely ergodic**.

Remark. If the MC has more than one absorbing components A_j, then there are ”commutor probabilities” C_{z,A_j} for which absorbing component A_j you eventually land in given that starts in $z \in S$. Then the fraction of time spent in y converges to $\boldsymbol{\pi}_{A_j}$ with probability C_{z,A_j}, as $T \to \infty$.

2.5.1 Limiting Probabilities

In Example 2.3.1 we considered a two-state Markov chain with transition probability matrix

$$P = \begin{bmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{bmatrix}$$

and showed that

$$P^{(4)} = \begin{bmatrix} 0.5749 & 0.4251 \\ 0.5668 & 0.4332 \end{bmatrix}.$$

From this it follows that $P^{(8)} = P^{(4)} P^{(4)}$ is given by

$$P^{(8)} = \begin{bmatrix} 0.571 & 0.429 \\ 0.571 & 0.429 \end{bmatrix}.$$

Note that the matrix $P^{(8)}$ is almost identical to the matrix $P^{(4)}$, and that each of the rows of $P^{(8)}$ has almost identical values. Indeed, it seems that $P^n_{i,j}$ is converging to some value as $n \to \infty$, with this value not depending on i. Moreover, in Example 2.5.1 we showed that the long-run proportions for this chain are $\pi_0 = 4/70.571$, $\pi_1 = 3/70.429$, thus making it appear that these long-run proportions may also be limiting probabilities. Although this is indeed the case for the preceding chain, it is not always true that the long-run proportions are also limiting probabilities. To see why not, consider a two-state Markov chain having

$$P_{0,1} = P_{1,0} = 1.$$

Because this Markov chain continually alternates between states 0 and 1, the long-run proportions of time it spends in these states are

$$\pi_0 = \pi_1 = 1/2.$$

However,

$$P_{0,0}^n = \begin{cases} 1, & \text{if } n \text{ is even} \\ 0, & \text{if } n \text{ is odd} \end{cases},$$

and so $P_{0,0}^n$ does not have a limiting value as n goes to infinity.

Definition 2.5.11 (Periodicity)**.** In general, a chain that can only return to a state in a multiple of $d > 1$ steps (where $d = 2$ in the preceding example) is said to be **periodic** and does not have limiting probabilities.

Proposition 2.5.6. *For an irreducible chain that is not periodic, and such chains are called **aperiodic**, the limiting probabilities will always exist and will not depend on the initial state. Moreover, the limiting probability that the chain will be in state j will equal π_j, the long-run proportion of time the chain is in state j.*

Proof. That the limiting probabilities, when they exist, will equal the long-run proportions can be seen by letting

$$\alpha_j = \lim_{n\to\infty} \mathbb{P}[X_n = j],$$

and using that

$$\mathbb{P}[X_{n+1} = j] = \sum_{i\in S} \mathbb{P}[X_{n+1} = j|X_n = i]\mathbb{P}[X_n = i] = \sum_{i\in S} P_{i,j}\mathbb{P})[X_n = i]$$

and

$$1 = \sum_{i\in S} \mathbb{P}[X_n = i].$$

Letting $n \to \infty$ in the preceding two equations yields, upon assuming that we can bring the limit inside the summation, that

$$\alpha_j = \sum_{i\in S} \alpha_i P_{i,j}$$
$$1 = \sum_{i\in S} \alpha_i$$

Hence, $\{\alpha_j, j \geq 0\}$ satisfies the equations for which $\{\pi_j, j \geq 0\}$ is the unique solution, showing that $\alpha_j = \pi_j, j \geq 0$. □

Definition 2.5.12 (Ergodicity)**.** An irreducible, positive recurrent, aperiodic Markov chain is said to be **ergodic**.

Robert talked about the stochastic-process ergodicity in the case that the MC has a finite state space.

Definition 2.5.13 (Stochastic-Process Ergodicity)**.** Let $\{X_n : n \in \mathbb{N}\}$ be a Markov chain with transition matrix P. If $\exists \boldsymbol{\pi} \in \Delta$, where $\Delta = \{\boldsymbol{x} \in \mathbb{R}^{1\times|S|} | \boldsymbol{x}\vec{1} = 1, \boldsymbol{x}_i \geq 0, i = 1, \cdots |S|\}$, such that $\forall \boldsymbol{\pi}_0 \in \Delta, \boldsymbol{\pi}_0 P^n \to \boldsymbol{\pi}$, as $n \to \infty$, then the MC is **stochastic-process ergodic**.

Remark. All senses of convergence are equivalent in finite state space, but let's say in l_1.

Remark. This tells us not only that the fraction of time spent in state y converges almost surely to $\boldsymbol{\pi}_y$ for any initial condition, but also that the probability of $X_n = y$ converges to $\boldsymbol{\pi}_y$ as $n \to \infty$.

Remark. There are finite discrete-time MCs with a unique absorbing component (so dynamical-system ergodic) which are not stochastic-process ergodic. The obstruction is periodicity. The two-state MC discussed at the beginnig of this section is an example.

Theorem 2.5.4. *A homogeneous finite discrete-time Markov chain is stochastic-process ergodic if it has a unique absorbing component and it is aperiodic.*

Proof. The proof is by the Perro-Frobenius theorem. For an irreducible, aperiodic, non-negative matrix, there is a unique and simple eigenvalue of maximum modulus, and it has a positive eigenvector.

In our case, the largest modulus of eigenvalue P can have is 1 by conservation of probability (or the Gershgorin theorem). And we konw it has eigenvalue 1, since $P\vec{1} = \vec{1}$.

So we can restrict the MC to the unique absorbing component, and thus, 1 is simple and has a positive left eigenvector $\boldsymbol{\pi}$ which we normalize to $\boldsymbol{\pi}\vec{1} = 1$. Then $\boldsymbol{\pi}_0 P^n = a\boldsymbol{\pi} +$ terms going to 0 as $n \to \infty$, and $\boldsymbol{\pi}_0\vec{1} = 1$, $P\vec{1} = \vec{1}$, so $a = 1$. Hence the MC is SP-ergodic. □

2.6 Mean Time Spent in Transient States

Consider now a finite-state Markov chain and suppose that the states are numbered so that $T = \{1, 2, \cdots, t\}$ denotes the set of transient states. Let

$$P_T = \begin{bmatrix} P_{1,1} & P_{1,2} & \cdots & P_{1,t} \\ P_{2,1} & P_{2,2} & \cdots & P_{2,t} \\ \vdots & \vdots & \ddots & \vdots \\ P_{n,1} & P_{n,2} & \cdots & P_{t,t} \end{bmatrix}$$

and note that since P_T specifies only the transition probabilities from transient states into transient states, some of its row sums are less than 1 (otherwise, T would be a closed class of states). For transient states i and j , let $s_{i,j}$ denote the expected number of time periods that the Markov chain is in state j , given that it starts in state i. Let $\delta_{i,j} = 1$ when $i = j$, and let it be 0 otherwise. Condition on the initial transition to obtain

$$\begin{aligned} s_{i,j} =& \delta_{i,j} + \sum_{k \in S} P_{i,k} s_{k,j} \\ =& \delta_{i,j} + \sum_{k=1}^{t} P_{i,k} s_{k,j} \end{aligned} \tag{2.2}$$

where the final equality follows since it is impossible to go from a recurrent to a transient state, implying that $s_{k,j} = 0$ when k is a recurrent state.

Let S denote the matrix of values $s_{i,j}, i, j = 1, \cdots, t$. That is,

$$S = \begin{bmatrix} s_{1,1} & s_{1,2} & \cdots & s_{1,t} \\ s_{2,1} & s_{2,2} & \cdots & s_{2,t} \\ \vdots & \vdots & \ddots & \vdots \\ s_{n,1} & s_{n,2} & \cdots & s_{t,t} \end{bmatrix}$$

In matrix notation, Eq. (2.2) can be written as

$$S = I + P_T S.$$

where I is the identity matrix of size t. Because the preceding equation is equivalent to

$$(I - P_T)S = I$$

we obtain, upon multiplying both sides by $(IP_T)^1$,

$$S = (I - P_T)^{-1}.$$

That is, the quantities $s_{i,j}$, $i \in T$, $j \in T$, can be obtained by inverting the matrix $I - P_T$. (The existence of the inverse is established on the fact that $I - P_T$ is diagonally dominant.)

For $i \in T$, $j \in T$, the quantity $f_{i,j}$, equal to the probability that the Markov chain ever makes a transition into state j given that it starts in state i, is easily determined from P_T. To determine the relationship, let us start by deriving an expression for $s_{i,j}$ by conditioning on whether state j is ever entered. This yields

$$\begin{aligned} s_{i,j} =& \mathbb{E}[\text{time in } j | \text{start in } i, \text{ ever transit to } j] f_{i,j} \\ &+ \mathbb{E}[\text{time in } j | \text{start in } i, \text{ never transit to } j](1 - f_{i,j}) \\ =&(\delta_{i,j} + s_{j,j}) f_{i,j} + \delta_{i,j}(1 - f_{i,j}) \\ =&\delta_{i,j} + s_{j,j} f_{i,j}. \end{aligned}$$

Solving the preceding equation yields

$$f_{i,j} = \frac{s_{i,j} - \delta_{i,j}}{s_{j,j}}.$$

Suppose we are interested in the expected time until the Markov chain enters some sets of states A, which need not be the set of recurrent states. We can reduce this back to the previous situation by making all states in A absorbing states. That is, reset the transition probabilities of states in A to satisfy

$$P_{i,i} = 1, \, i \in A.$$

This transforms the states of A into recurrent states, and transforms any state outside of A from which an eventual transition into A is possible into a transient state. Thus, our previous approach can be used.

2.7 Time Reversible Markov Chains

Consider a stationary ergodic Markov chain (that is, an ergodic Markov chain that has been in operation for a long time) having transition probabilities $P_{i,j}$ and stationary probabilities π_i , and suppose that starting at some time we trace the sequence of states going backward in time. That is, starting at time n, consider the sequence of states $X_n, X_{n-1}, X_{n-2}, \cdots$. It turns out that this sequence of states is itself a Markov chain with transition probabilities $Q_{i,j}$ defined by

$$\begin{aligned} Q_{i,j} =& \mathbb{P}[X_m = j | X_{m+1} = i] \\ =& \frac{\mathbb{P}[X_m = j, X_{m+1} = i]}{\mathbb{P}[X_{m+1} = i]} \\ =& \frac{\mathbb{P}[X_m = j]\mathbb{P}[X_{m+1} = i | X_m = j]}{\mathbb{P}[X_{m+1} = i]} \\ =& \frac{\pi_j P_{j,i}}{\pi_i} \end{aligned}$$

To prove that the reversed process is indeed a Markov chain, we must verify that

$$\mathbb{P}[X_m = j | X_{m+1} = i, X_{m+2} = s_{m+2}, X_{m+3} = s_{m+3}, \cdots] = \mathbb{P}[X_m = j | X_{m+1} = i].$$

To see that this is so, suppose that the present time is $m + 1$. Now, since $X_0, X_1, X_2, \cdots$ is a Markov chain, it follows that the conditional distribution of the future $X_{m+2}, X_{m+3}, \cdots$ given the present state X_{m+1} is independent of the past state X_m . However, independence is a symmetric relationship (that is, if A is independent of B, then B is independent of A), and so this means that given X_{m+1}, X_m is independent of $X_{m+2}, X_{m+3}, \cdots$. But this is exactly what we had to verify.

Thus, the reversed process is also a Markov chain with transition probabilities given

$$Q_{i,j} = \frac{\pi_j P_{j,i}}{\pi_i}.$$

Remark. In general, the reverse chain is not homogeneous even if the original MC is, but if the MC is stationary with stationary probability $\boldsymbol{\pi}$, then

$$Q_{i,j} = \frac{\pi_j}{\pi_i} P_{i,j}.$$

So the reverse MC is also homogeneous and stationary with the same probability $\boldsymbol{\pi}$.

Definition 2.7.1 (Time Reversibility)**.** If $Q_{i,j} = P_{i,j}$ for all i, j, then the Markov chain is said to be **time reversible**.

Definition 2.7.2 (Detailed Balance Conditions)**.** The condition for time reversibility, namely, $Q_{i,j} = P_{i,j}$ can also be expressed as

$$\pi_i P_{i,j} = \pi_j P_{j,i} \qquad \text{for all } i, j \tag{2.3}$$

Thesse conditions are called **detailed balance conditions**.

The condition in Eq. (2.3) can be stated that, for all states i and j , the rate at which the process goes from i to j (namely, $\pi_i P_{i,j}$) is equal to the rate at which it goes from j to i (namely, $\pi_j P_{j,i}$). It is worth noting that this is an obvious necessary condition for time reversibility since a transition from i to j going backward in time is equivalent to a transition from j to i going forward in time; that is, if $X_m = i$ and $X_{m1} = j$, then a transition from i to j is observed if we are looking backward, and one from j to i if we are looking forward in time. Thus, the rate at which the forward process makes a transition from j to i is always equal to the rate at which the reverse process makes a transition from i to j; if time reversible, this must equal the rate at which the forward process makes a transition from i to j.

Proposition 2.7.1. *If we can find nonnegative numbers, summing to one, that satisfy Eq.* (2.3)*, then it follows that the Markov chain is time reversible and the numbers represent the limiting probabilities.*

Proof. This is so since if

$$x_i P_{i,j} = x_j P_{j,i} \qquad \text{for all } i,\ j, \sum_{i \in S} x_i = 1, \tag{2.4}$$

then summing over i yields

$$\sum_{i \in S} x_i P_{i,j} = x_j \sum_{i \in S} P_{j,i} = x_j, \sum_{i \in S} x_i = 1,$$

and, because the limiting probabilities π_i are the unique solution of the preceding, it follows that $x_i = \pi_i$ for all i. □

If we try to solve Eq. (2.4) for an arbitrary Markov chain with states $0, 1, \cdots, M$, it will usually turn out that no solution exists. For example, from Eq. (2.4), implying (if $P_{i,j}P_{j,k} > 0$) that

$$\frac{x_i}{x_k} = \frac{P_{j,i}P_{k,j}}{P_{i,j}P_{j,k}}$$

which in general need not equal $P_{k,i}/P_{i,k}$. Thus, we see that a necessary condition for time reversibility is that

$$P_{i,k}P_{k,j}P_{j,i} = P_{i,j}P_{j,k}P_{k,i} \quad \text{for all } i, j, k \tag{2.5}$$

which is equivalent to the statement that, starting in state i, the path $i \to k \to j \to i$ has the same probability as the reversed path $i \to j \to k \to i$. To understand the necessity of this, note that time reversibility implies that the rate at which a sequence of transitions from i to k to j to i occurs must equal the rate of ones from i to j to k to i, and so we must have

$$\pi_i P_{i,k}P_{k,j}P_{j,i} = \pi_i P_{i,j}P_{j,k}P_{k,i}$$

implying Eq. (2.5) when $\pi_i > 0$.

In fact, we have the following

Theorem 2.7.1. *A stationary Markov chain for which $P_{i,j} = 0$ whenever $P_{j,i} = 0$ is time reversible if and only if starting in state i, any path back to i has the same probability as the reversed path. That is, if*

$$P_{i,i_1}P_{i_1,i_2}\cdots P_{i_k,i} = P_{i,i_k}P_{i_k,i_k-1}\cdots P_{i_1,i} \tag{2.6}$$

for all states $i, i_1, \cdots, i_k$.

Proof. We have already proven necessity. To prove sufficiency, fix states i and j and rewrite Eq. (2.6) as

$$P_{i,i_1}P_{i_1,i_2}\cdots P_{i_k,j}P_{j,i} = P_{i,j}P_{j,i_k}\cdots P_{i_1,i}$$

Summing the preceding over all states $i_1, \cdots, i_k$ yields

$$P^{k+1}_{i,j}P_{j,i} = P_{i,j}P^{k+1}_{j,i}$$

Consequently,

$$\frac{P_{j,i}\sum_{k=1}^{m}P^{k+1}_{i,j}}{m} = \frac{P_{i,j}\sum_{k=1}^{m}P^{k+1}_{j,i}}{m}$$

Letting $m \to \infty$ yields

$$P_{j,i}\pi_j = P_{i,j}\pi_i$$

which proves the theorem. □

The concept of the reversed chain is useful even when the process is not time reversible. To illustrate this, we start with the following proposition.

Proposition 2.7.2. *Consider an irreducible Markov chain with transition probabilities $P_{i,j}$. If we can find positive numbers π_i , $i \in S$, summing to one, and a transition probability matrix $Q = [Q_{i,j}]$ such that*

$$\pi_i P_{i,j} = \pi_j Q_{j,i} \tag{2.7}$$

then the $Q_{i,j}$ are the transition probabilities of the reversed chain and the π_i are the stationary probabilities both for the original and reversed chain.

2.8 Monte Carlo Markov Chains

Given probability $\boldsymbol{\pi}$ on S and a random variable $f : S \to \mathbb{R}$, we may want to compute the mean $\mathbb{E}[f] =: \boldsymbol{\pi}(f)$. Or might be given an unnormalized probability $\tilde{\pi}$ and want $\boldsymbol{\pi}(f)$ with $\boldsymbol{\pi} = \frac{\tilde{\boldsymbol{\pi}}}{Z}$ and Z is the normalizing coefficient $Z = \tilde{\boldsymbol{\pi}}\vec{1}$, or might want to compute Z. Such problems arise in statistical mechanics where probability of state x is proportional to $e^{-\beta H(x)}$ with H interpreted as energy and $\beta = \frac{1}{R_B T}$ interpreted as coolness. Then the normalizing coefficient is $Z = \sum_{x \in S} e^{-\beta H(x)}$ and mean energy is

$$\frac{1}{Z} \sum_{x \in S} H e^{-\beta H}.$$

They also arise in statistical inference for Bayesian model comparison. For example, suppose model M has parameterd μ and consists in a specification of

$$\mathbb{P}[\text{data}|M, \mu].$$

Bayesian inference gives

$$\mathbb{P}[\mu|\text{data}, M] = \frac{\mathbb{P}[\text{data}|M, \mu]\mathbb{P}[\mu]}{Z(M)}$$

To compare 2 models M_1 and M_2, we need to look at

$$\frac{\mathbb{P}[M_1|\text{data}]}{\mathbb{P}[M_2|\text{data}]} = \frac{Z(M_1)\mathbb{P}[M_1]}{Z(M_2)\mathbb{P}[M_2]}.$$

Design a MC with unique absorbing component on which $m\tilde{b}f\pi$ is stationary. Then the fraction of time spent in state x by a typical realization converges to $\boldsymbol{\pi}(x)$ as $T \to \infty$, and the time-average of $\tilde{\boldsymbol{\pi}}(X_n) \to Z$.

The easiest way to achieve $\tilde{\pi}$ stationary is to choose P so that $\tilde{\pi}_i P_{i,j} = \tilde{\boldsymbol{\pi}}_j P_{j,i}$ $\forall i, j \in S$ and $P\vec{1} = \vec{1}$ with $P_{i,j} \geq 0$. We can do this by taking any proposal transition probabilities $Q_{i,j}$ and use acceptance probabilities $A_{i,j}$ like

$$\text{Metropolis–Hastings } A_{i,j} = \begin{cases} 1 & \text{if } \tilde{\boldsymbol{\pi}}_j Q_{j,i} \geq \tilde{\boldsymbol{\pi}}_i Q_{i,j} \\ \frac{\tilde{\boldsymbol{\pi}}_j Q_{j,i}}{\tilde{\boldsymbol{\pi}}_i Q_{i,j}} & \text{otherwise} \end{cases}.$$

$$\text{Heatbath } A_{i,j} = \frac{\tilde{\boldsymbol{\pi}}_j Q_{j,i}}{\tilde{\boldsymbol{\pi}}_i Q_{i,j} + \tilde{\boldsymbol{\pi}}_j Q_{j,i}},$$

and set $P_{i,j} = Q_{i,j} A_{i,j}$, $j \neq i$, and $P_{i,i}$ be the reject.

We may as well take $Q_{i,i} = 0$, $\forall i \in S$, and require that Q to have unique absorbing component (in general the whole of S). Then with the above choices, P has detailed balance for $\tilde{\boldsymbol{\pi}}$:

$$\frac{\tilde{\boldsymbol{\pi}}_i Q_{i,j} A_{i,j}}{\tilde{\boldsymbol{\pi}}_j Q_{j,i} A_{j,i}} = 1.$$

This can be done without rejection: given i, let $w_{i,j} = Q_{i,j} A_{i,j}$, and $W_i = \sum_{j \in S} w_{i,j}$. Let

$$P_{i,j} = \frac{w_{i,j}}{W_i},$$

and weight time spent in i by W_i.

Question of time needed to explore the state space, or mixing time (really only need DS ergodization time, but P^T is SP ergodic if P is DS ergoditc and T is the period, for a cyclic component).

Mixing time

$$T(\epsilon) = \min\{T \in \mathbb{N} | d(\sigma P^n, \boldsymbol{\pi}) \leq \epsilon, \forall n \geq T \text{ and } \sigma \in \Delta\},$$

using total variation distance (which is defined as one half of l_1 distance).

Define

$$\|P\|_Z = \sup_{\boldsymbol{v} \neq 0, \boldsymbol{v}\vec{1}=0} \frac{\|vP\|_1}{\|v\|_1},$$

where Z is interpreted as "zero charge".

If $\|P\|_Z < 1$, then P is a contraction on Δ, so get SP-ergodicity.

If $\|P^n\|_Z \leq Cr^n$ with $0 < r < 1$ then $T(\epsilon) \leq \frac{\log(\epsilon/c)}{\log r}$.

One can get such bounds on $\|P^n\|_Z$ from Dobrushin's ergodicity coefficients:

$$\|P\|_Z = 1 - \min_{i,j} \sum_k \min\{P_{i,k}, P_{j,k}\} = \frac{1}{2} \max_{i,j} \sum_k |P_{i,k} - P_{j,l}|$$

But what is really wanted is to know $d(\mu_T, \boldsymbol{\pi})$, where $\mu_T = \frac{1}{T} \sum_{n=0}^{T-1} \delta_{X_n}$. Typically, this is $O(T^{-1/2})$ (CLT generalization).

Can reformulate: find $\min T$ such that $\mathbb{P}[d(\mu_T, \boldsymbol{\pi}) > \epsilon] > \eta$. Large Deviation Theorem implies

$$\mathbb{P}[d(\mu_T, \boldsymbol{\pi}) > \epsilon] \leq C \exp(-\frac{T\epsilon^2}{K + \frac{1}{2}}),$$

where $K = \|(I - P)^{-1}\|_Z$. So $T \sim (K + \frac{1}{2})\epsilon^{-2} \log \frac{C}{\eta}$.

Can adapt the MCMC to the function f whose mean we want: importance sampling/variance reduction.

2.9 Countable Discrete-Time Markov Chains

One can extend much of what we've done for finite DTMC to the countably infite case, for example, the simple random walk (SRW) on $\mathbb{Z}$. But some results become more subtile, for example, SRW is not SP-ergodic, despite being irreducible (actually it also fails to have a stationary probability; also it is not aperiodic — it has a period 2.)

One have to refine various concepts. Let

$$T_x = \inf\{n > 0 : X_n = x\}.$$

The following definitions are the same as those defined before.

Definition 2.9.1 (Transience)**.** Say $x \in S$ is **transient** if

$$\mathbb{P}[T_x = \infty | X_0 = x] > 0,$$

Remark. x is transient, then with probability 1 X_n comes back to x only finitely many times.

Definition 2.9.2 (Null Recurrence)**.** Say $x \in S$ is **null recurent** if

$$\mathbb{P}[T_x < \infty | X_0 = x] = 1,$$

and

$$\mathbb{E}[T_x | X_0 = x] = \infty.$$

Definition 2.9.3 (Positive Recurrence)**.** Say $x \in S$ is **positive recurrent** if

$$\mathbb{P}[T_x < \infty | X_0 = x] = 1,$$

and

$$\mathbb{E}[T_x | X_0 = x] < \infty.$$

Definition 2.9.4 (Communicating Classes)**.** A **communicating class** is either null recurrent or positive recurrent.

Theorem 2.9.1. *An absorbing class has a unique stationary probability if and only if it is positive recurrent. And* $\boldsymbol{\pi}_x = 1/T_x$*.*

Remark. Positive recurrence guarantees the existence and uniqueness of the stationary distribution.

2.10 Countable Discrete-Time Markov Chains

One can extend much of what we have done for finite discrete-time Markov chains to the countably infinite case, e.g. the **simple random walk** on $\mathbb{Z}$, but some results become more subtle. For example, the simple random walk is *not SP-ergodic*, despite being *irreducible*. Actually, it even *fails to have a stationary probability*; also it is *not aperiodic*, and it has a *period* 2.

Example 2.10.1. Using definition of the simple random walk:

$$Y_n = \sum_{i=0}^{n-1} X_i,$$

where X_i's are independent and identically distributed, with

$$X_i = \begin{cases} +1 & \text{with probability}\, p \\ -1 & \text{with probability}\, 1-p \end{cases},$$

Compute the $\mathbb{E}[Y_n]$ and $\text{Var}[Y_n]$.

One has to refine various concepts.

Definition 2.10.1 (The First Return Time)**.** The **first return time** to state x is defined as

$$T_x = \inf\{n \geq 1 : X_n = x | X_0 = x\}.$$

Remark. Notice that when the state space is finite and x is recurrent, T_x is finite. Since the state space here is countably infinite, T_x is allowed to be infinte.

Definition 2.10.2 (Transience)**.** Say $x \in S$ is **transient** if

$$\mathbb{P}[T_x = \infty] > 0.$$

Remark. If $x \in S$ is transient, then with probability 1 X_n comes back to x only finitely many times.

Definition 2.10.3 (Null Recurrence)**.** Say $x \in S$ is **null recurrent** if

$$\mathbb{P}[T_x < \infty] = 1 \quad \text{and} \quad \mathbb{E}[T_x] = \infty.$$

Definition 2.10.4 (Positive Recurrence)**.** Say $x \in S$ is **positive recurrent** if

$$\mathbb{P}[T_x < \infty] = 1 \quad \text{and} \quad \mathbb{E}[T_x] < \infty.$$

Remark. A communicating class is either **null recurrent**, which means every member is null recurrent, or **positive recurrent** which means every member is positive recurrent.

Theorem 2.10.1 (Stationarity $\iff$ Positive Recurrence)**.** *An absorbing class has a stationary probability if and only if it is positive recurrent. Furthermore, if the class has one stationary probability, then it is uniquely determined by*

$$\boldsymbol{\pi}_x = \frac{1}{\mathbb{E}[T_x]}.$$

Chapter 3

Continuous-Time Markov Chains

3.1 Continuous-Time Markov Chains

We are now considering a continuous-time markov chain with a countable state space S and the domain $T \in \mathbb{R}$ (or $T \in \mathbb{R}_+$), and we restrict $X : \mathbb{R} \mapsto S$ to those which are *piecewise constant* and *right-continuous*, meaning

$$X(t) = \begin{cases} \vdots & \vdots \\ s & t \in [J_s, J_{s'}) \\ s' & t \in [J_{s'}, J_{s''}) \\ \vdots & \vdots \end{cases}$$

Definition 3.1.1 (Continuous-Time Markov Chains)**.** $X(t) : \mathbb{R} \mapsto S$ is a **continuous-time Markov chain**, if it satisfies the **Markov property**

$$\mathbb{P}[X(t_{n+1}) \in A | X(t_n) = s_n, \cdots, X(t_1) = s_1] = \mathbb{P}[X(t_{n+1}) \in A | X(t_n) = s_n],$$

where $A \subset S$ and $t_1 < \cdots t_n < t_{n+1}$.

Definition 3.1.2 (Homogeneity)**.** A continuous-time Markov chain is **homogeneous** if

$$\mathbb{P}[X(t+u) \in A | X(u) = s] = \mathbb{P}[X(t) \in A | X(0) = s].$$

Remark. Homogeneity means time translation invariance.

Definition 3.1.3 (Transition Matrices)**.** Let $(P_t)_{i,j} := \mathbb{P}[X(t) = j | X(0) = i]$, then P_t is the transition matrix with time step t.

Remark. The (i, j) element of the transition matrix P_t can also be expressed as $P_t(i, j)$.

Theorem 3.1.1 (Chapman-Kolmogorov Equation)**.** *The transition matrix P of a homogeneous Markov chain satisfies*

$$P_{t+u} = P_t P_u, \; P_0 = I.$$

Proof. Notice that

$$\begin{aligned}(P_{t+u})_{i,j} =&\mathbb{P}[X(t+u)=j|X(0)=i]\\ =&\sum_{k\in S}\mathbb{P}[X(t+u)=j|X(t)=k,\ X(0)=i]\mathbb{P}[X(t)=k|X(0)=i]\\ =&\sum_{k\in S}\mathbb{P}[X(t+u)=j|X(t)=k]\mathbb{P}[X(t)=k|X(0)=i]\\ =&\sum_{k\in S}\mathbb{P}[X(u)=j|X(0)=k]\mathbb{P}[X(t)=k|X(0)=i]\\ =&\sum_{k\in S}(P_u)_{k,j}(P_t)_{i,k}\\ =&(P_t)_{i,:}\ (P_u)_{:,j},\end{aligned}$$

where $(P_t)_{i,:}$ is the i-th row of P_t and $(P_u)_{:,j}$ is the j-th column of P_u. Thus, $P_{t+u}=P_tP_u$. And by definition, $(P_0)_{i,j}=\mathbb{P}[X_0=j|X_0=i]=\delta_{i,j}$, so $P_0=I$. □

3.1.1 The Rate Matrix

Definition 3.1.4 (Rate Matrix)**.** Suppose P_t is differentiable with respect to t at $t=0$, then

$$G:=\left.\frac{\mathrm{d}P_t}{\mathrm{d}t}\right|_{t=0}$$

is called the **generator** or the **rate matrix** of the process.

Proposition 3.1.1. *$P_t=\exp(tG)$ in the sense of power series.*

Proof. By the Chapman-Kolmogorov equation, we have

$$\begin{aligned}P_{t+u} =&P_tP_u\\ P_{t+u}-P_t =&P_t(P_u-I)\\ \frac{P_{t+u}-P_t}{u} =&P_t\cdot\frac{P_u-I}{u}\\ \lim_{u\to0}\frac{P_{t+u}-P_t}{u} =&\lim_{u\to0}P_t\cdot\frac{P_u-I}{u}\\ \lim_{u\to0}\frac{P_{t+u}-P_t}{u} =&P_t\cdot\lim_{u\to0}\frac{P_u-I}{u}\\ \frac{\mathrm{d}P_t}{\mathrm{d}t} =&P_tG,\end{aligned}$$

So $P_t=C\cdot\exp(tG)$, where C is a constant diagonal matrix with diagonal elements being equal. By $P_0=I$, we konw $C=I$. □

Proposition 3.1.2. *The generator G also satisfies*

$$G\vec{1}=\vec{0}.$$

Proof. For any probability distribution $\boldsymbol{\pi}_t=\boldsymbol{\pi}_0P_t$ with initial distribution $\boldsymbol{\pi}_0$, evolves by

$$\frac{\mathrm{d}\boldsymbol{\pi}_t}{\mathrm{d}t}=\boldsymbol{\pi}_0\frac{\mathrm{d}P_t}{\mathrm{d}t}=\boldsymbol{\pi}_0P_tG=\boldsymbol{\pi}_tG.$$

And by conservation of probability, we have $\boldsymbol{\pi}_t\vec{1}=\vec{1}$, which implies $\boldsymbol{\pi}_tG\vec{1}=\frac{\mathrm{d}\boldsymbol{\pi}_t\vec{1}}{\mathrm{d}t}=0$. Since $\boldsymbol{\pi}_t$ is arbitrary, we have $G\vec{1}=0$. □

Theorem 3.1.2 (The Master Equation). *The equation*

$$\frac{\mathrm{d}\boldsymbol{\pi}_t}{\mathrm{d}t} = \boldsymbol{\pi}_t G$$

can be written into

$$\frac{\mathrm{d}(\boldsymbol{\pi}_t)_i}{\mathrm{d}t} = \underbrace{\sum_{j\neq i}(\boldsymbol{\pi}_t)_j G_{j,i}}_{\text{“gain”}} - \underbrace{\sum_{j\neq i}(\boldsymbol{\pi}_t)_i G_{i,j}}_{\text{“loss”}},$$

which is called the ***master equation****.*

Proof. For $i \neq j$, since $G_{i,j}$ is the rate at which the process goes from state i to j, we have $G_{i,j} \geq 0$. By $G\vec{1} = \vec{0}$, we have

$$G_{i,i} = -\sum_{j\neq i} G_{i,j}.$$

So

$$\begin{aligned}\frac{\mathrm{d}(\boldsymbol{\pi}_t)_i}{\mathrm{d}t} =& \boldsymbol{\pi}_t G_{:,i} \\ =& \sum_{j\in S}(\boldsymbol{\pi}_t)_j G_{j,i} \\ =& \sum_{j\neq i}(\boldsymbol{\pi}_t)_j G_{j,i} - \sum_{j\neq i}(\boldsymbol{\pi}_t)_i G_{i,j}.\end{aligned}$$

□

Remark. The name “master equation” is exaggerated; it does not tell everything about the process, such as the correlations between states at different times.

Example 3.1.1 (Poisson Processes). The **Poisson process** with rate $\lambda > 0$ has the state space $S = \mathbb{N}$, $X(0) = 0$, and the transition matirx G such that

$$G_{i,j} = \begin{cases} \lambda & j = i + 1 \\ -\lambda & j = 1 \end{cases}.$$

It has $\mathbb{P}[X(t+u) = n + k | X(u) = n] = e^{-\lambda t}\frac{(\lambda t)^k}{k!}$, $\forall n, k \in \mathbb{N}$, $\forall t, u \in \mathbb{R}_+$.

Example 3.1.2 (Birth and Death Processes). Suppose we have the birth rates α_i and the death rates β_i ($\beta_0 = 0$), for $i \in S = \mathbb{N}$. The rate matrix G is defined by

$$G_{i,j} = \begin{cases} \alpha_i & j = i + 1 \\ \beta_i & j = i - 1 \\ -(\alpha_i + \beta_i) & j = i \end{cases}.$$

Then the process is called the **Birth and Death Process**.

Example 3.1.3 ($M/M/1$ queue). The birth and death process has a special case - the ***M/M/*1 queue**, in which $\alpha_i = \alpha$, $\beta_i = \beta$ for $i \neq 0$ and $\beta_0 = 0$. M means “memoryless”, and 1 means there is only one cashier to serve customers.

Example 3.1.4 ($M/M/\infty$ queue)**.** Another example is the ***M/M/∞* queue**, in which there are infinitely many servers so that customers do not have to wait for people in front of them. In this model $\alpha_i = \alpha$ and $\beta = i\beta$.

Example 3.1.5 (Population Growth)**.** Population growth can be modelled by the birth and death process with $\alpha_i = i\alpha$ and $\beta_i = i\beta$, where i is the size of population.

3.1.2 Stationarity and Reversibility

Definition 3.1.5 (Stationarity)**.** Say $\boldsymbol{\pi} \in \Delta$ is **stationary** if $\boldsymbol{\pi} G = 0$.

Definition 3.1.6 (Reversibility)**.** Say $\boldsymbol{\pi} \in \Delta$ is **reversible** if

$$\boldsymbol{\pi}_i G_{i,j} = \boldsymbol{\pi}_j G_{j,i},\ \forall i, j \in S.$$

Proposition 3.1.3 (Reversibility $\implies$ Stationarity)**.** *If $\boldsymbol{\pi} \in \Delta$ is reversibile, then it is also stationary.*

Proposition 3.1.4. *S is fintie $\implies$ $\exists$ stationary $\boldsymbol{\pi}$.*

There is an analogous decomposition of the state space S into transient and recurrent states, and of the set of recurrent states into communicating components. And we have the same definition of an absorbing component.

Proposition 3.1.5. *If S is finite, then each absorbing component has a unique stationary probability $\boldsymbol{\pi}$, and the space of starionary $\boldsymbol{\pi}$ for the whole continuous-time Markov chain (up to normalisation) is the span of those for its absorbing components. Furthermore, 0 is a semisimple eigenvalue of G.*

Theorem 3.1.3. *Suppose S is finite and G has a unique absorbing component, then the process is SP-ergodic, which means*

$$\lim_{t\to\infty} \boldsymbol{\pi}_t = \boldsymbol{\pi}_A,$$

where $\boldsymbol{\pi}_A$ is the stationary distribution of the absorbing component.

Remark. Aperiodicity is automatic in continuous time.

3.1.3 The Jump Chain

Definition 3.1.7 (Waiting Times)**.** The **waiting time** or the **holding time** W_x is defined as

$$W_x = \inf\{t > 0 : X(t) \neq x | X(0) = x\}.$$

Proposition 3.1.6. *The waiting time W_x is exponentially distributed with mean $\frac{1}{|G_{x,x}|}$.*

Proof.

$$\begin{aligned}\mathbb{P}[W_x > t + u | W_x > t] =& \mathbb{P}[W_x > t + u | X(s) = x,\ \forall s \le t] \\ =& \mathbb{P}[W_x > t + u | X(t) = x] \\ =& \mathbb{P}[W_x > u | X(0) = x] \\ =& \mathbb{P}[W_x > u].\end{aligned}$$

So $\mathbb{P}[W_x > t + u] = \mathbb{P}[W_x > u]\mathbb{P}[W_x > t]$. So $\exists \gamma \in \mathbb{R}$, such that

$$\mathbb{P}[W_x > t] = e^{-\gamma t}.$$

$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{P}[W_x > t]\Big|_{t=0} = G_{x,x}$ shows $\gamma = -G_{x,x}$. □

Definition 3.1.8 (Jump Times)**.** Define **jump times** $J_{n+1} = \inf\{t > J_n : X(t) \neq X(J_n)\}$, with $J_0 = 0$.

Remark. The jump times are an example of "stopping times", i.e. random variables such that $\{J_n \leq t\}$ is independent of $\{X(s) : s > t\}$ given $\{X(s) : s \leq t\}$.

Theorem 3.1.4. *Markov chains satisfiy the **strong Markov property**: let T be a stopping time conditional on $X_T = i$, then X_{T+t} $(t \geq 0)$ is Markov and independent of $\{X(s) : s \leq T\}$.*

Definition 3.1.9 (The Jump Chain)**.** Let $Y_n = X(J_n)$, then $\{Y_n : n \in \mathbb{N}\}$ is called the **jump chain** of $\{X_t : t \in \mathbb{R}\}$.

Remark. The jump chain $\{Y_n : n \in \mathbb{N}\}$ is a discrete-time Markov chain.

Proposition 3.1.7. *The one-step transition matrix of the jump chain $\{Y_n : n \in \mathbb{N}\}$ is*

$$P_{i,j} = \begin{cases} 0 & j = i \\ \frac{G_{i,j}}{|G_{i,i}|} & j \neq i \,\&\, G_{i,i} = 0 \\ \delta_{i,} & G_{i,i} = 0 \end{cases}.$$

Remark. We can make sample paths for the continuous-time Markov chain by making paths for the associated jump chain and choosing independent waiting times W_{Y_n} with mean $1/|G_{Y_n,Y_n}|$, and let

$$J_n = \sum_{0 \leq k < n} W_{Y_k}.$$

3.2 Countable Continuous-Time Markov Chains

Now suppose the state space S of a continuous-time Markov chains is countable. We can define the null and positive recurrence as in the discrete-time case, but we have to find the return time differently.

Definition 3.2.1 (First Return Time)**.** The **first return time** to state $x \in S$ is defined as

$$\inf\{t > J_1 : X(t) = x\},$$

for $X(0) = x$.

Proposition 3.2.1. *Each positive recurrent absorbing component has a unique stationary probability distribution $\boldsymbol{\pi}$, and*

$$\boldsymbol{\pi} = \frac{\mathbb{E}[W_x]}{\mathbb{E}[T_x]}.$$

In continuous time, the process can get "explosion".

Definition 3.2.2 (Explosion)**.** Let $J_\infty = \lim_{n\to\infty} J_n$. If $\mathbb{P}[J_\infty = \infty] < 1$, then the continuous-time Markov chain is called **explosive**, which means there is a positive probability for infinitely many events in a bounded time.

Proposition 3.2.2. *If $\sup_{i\in S} |G_{i,i}| < \infty$, then the continuous-time Markov chain is not eplosive.*

Example 3.2.1 (Explosion)**.** Consider a birth and death process with $X(0) = 1$, $\alpha_i = i^2$ and $\beta_i = 0$. Then

$$\mathbb{E}[J_\infty] = \sum_{i=2}^{\infty} \mathbb{E}[W_i] = \sum_{i=2}^{\infty} \frac{1}{\alpha_i} = \sum_{i=2}^{\infty} \frac{1}{i^2} < \infty,$$

which means with probability 1 J_∞ is finite.

3.3 Semi-Markov Chains

Definition 3.3.1 (Semi-Markov Chains)**.** Take a discrete-time Markov chain and make a continuous-time process by waiting a time W_x in each state $x \in S$ independently of previous and future states but not necessarily exponentially distributed.

Remark. Semi-Markov chains allow for latent periods and variations of infectivity with time from infection.

3.4 Gaussian Processes

Definition 3.4.1 (Gaussian Processes)**.** Let $X : T \mapsto \mathbb{R}$ be a stochasti process. $X(t)$ is called a **Gaussian process** if $\forall t_1, \cdots, t_n \in T$, $(X(t_1), \cdots, X(t_2))$ is a multivariate Gaussian random vector, i.e. it has the probability density function

$$f(x_1, \cdots x_n) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \exp\left(-\frac{1}{2}(\vec{x} - \vec{\mu}^T)\Sigma^{-1}(\vec{x} - \vec{\mu})\right),$$

for some $\vec{\mu} = [\mu_1, \cdots, \mu_n]^T$ and some positive definite symmetric $n \times n$ matrix Σ.

Remark. A Guassian process is not necessarily Markov.

Proposition 3.4.1. *There exist functions $m : T \mapsto \mathbb{R}$ and $c : T \times T \mapsto \mathbb{R}$ such that $\mu_i = m(t_i)$ and $\Sigma_{i,j} = c(t_i, t_j)$ with c being "positive definite" i.e. such that Σ is positive definite $\forall t_1, \cdots, t_n \in T$.*

Example 3.4.1 (Stationary Ornstein-Uhlenbeck Processes)**.** Let $T = \mathbb{R}$, $m(t) = 0$ and $c(t, t') = e^{-|t'-t|}$, then the process is called a **stationary Ornstein-Uhlenbeck process**.

One can allow degenerate Gaussians, e.g. Ornstein-Uhlenbeck with specified initial condition $X(0) = 0$, then $f(x_0) = \delta_0(x_0)$, which is not a Gaussian probability density function but can be viewed as the limit of a Gaussian density.

The best way to generate a Gaussian distribution is to use its characteristic function instead of its PDF.

Definition 3.4.2 (Characteristic Functions)**.** Let $\vec{X}$ be a random vector, then its characteristic function is

$$\phi(\vec{\theta}) := \mathbb{E}[e^{i\vec{\theta}^T \vec{X}}].$$

Remark. For a multivariate Gaussian distribution with the mean vector $\vec{\mu}$ and covariance matrix $\mathbf{\Sigma}$ (which is allowed to be positive semi-definite), its characteristic function is

$$\phi(\vec{\theta}) = e^{i\vec{\theta}^T \vec{\mu} - \frac{1}{2}\vec{\theta}^T \mathbf{\Sigma} \vec{\theta}}.$$

We can include vector-valued Gaussian processes.

Definition 3.4.3 (Multivariate Gaussian Processes)**.** $\vec{X} : T \mapsto \mathbb{R}^n$ is a **multivariate Gaussian process** if $X : T \times \{1, \cdots, n\} \mapsto \mathbb{R}$ is a Gaussian process.

Definition 3.4.4 (Stationary Gaussian Processes)**.** Suppose $T = \mathbb{R} \times \mathbb{K}$. The Gaussian process is **stationary**, if its mean function $m(t, k)$ is independent of t, and its covariance function $c(t, k; t', k')$ is dependent only on $t - t'$ and $k - k'$.

Remark. Gaussian processes are great for inference, because $\mathbb{P}$[parameters|data] reduces to linear algebra.

3.5 Markov Processes with $S = \mathbb{R}$

Suppose $X : T \mapsto \mathbb{R}$, where T can be $\mathbb{Z}$ or $\mathbb{R}$.

Definition 3.5.1 (Markov Processes)**.** $\{X(t) : t \in T\}$ is a **Markov Processes** if it satisfies the Markov property

$$\mathbb{P}[X(t_{n+1}) \in A | X(t_n) = s_n, \cdots, X(t_1) = s_1] = \mathbb{P}[X(t_{n+1}) \in A | X(t_n) = s_n],$$

where $A \subset \mathbb{R}$ and $t_{n+1} > t_n > \cdots > t_1$.

Remark. There is a technical problem in the definition. The conditional probability is not well defined, since random variables $X_{t_n}, \cdots, X(t_1)$ now take values in $\mathbb{R}$, and the probability that they take particular values is 0. This will not be a problem if we restrict to any choice of interpretation of conditional probability such that

$$\mathbb{P}[X(t) \in A] = \int \mathbb{P}[X(t) \in A | X(0) = x] \, \mathrm{d}\mathbb{P}[X(0) \le x] \qquad \text{(a Stieltjes integral)}.$$

Definition 3.5.2 (Homogeneity)**.** A Markov process is **homoegeneous** if

$$\mathbb{P}[X(t) \in A | X(t') = s] = \mathbb{P}[X(t - t') \in A | X(0) = s].$$

It is unlikely that $\mathbb{P}[X(t) = y | X(0) = x] > 0$, so instead we specify $\mathbb{P}[X(t) \in A | X(0) = x]$ for any measurable set $A \subset \mathbb{R}$ as

$$\int_A p_t(x, y) \, \mathrm{d}y$$

for a transition density $p_t(\cdot, \cdot)$.

Definition 3.5.3 (Transition Densities)**.** A **transition probability** is a function $p_t(\cdot, \cdot) : \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}$ such that

$$\mathbb{P}[X(t) \in A | X(0) = x] = \int_A p_t(x, y) \, \mathrm{d}y$$

Theorem 3.5.1 (The Chapman-Kolmogorov Equation)**.** *The Markov property and homogeneity implies the* ***Chapman-Kolmogorov equation***

$$p_{t+u}(x, y) = \int_{\mathbb{R}} p_t(x, z) p_u(z, y) \, \mathrm{d}z.$$

3.5.1 Jump Processes

Definition 3.5.4 (Jump Processes)**.** $\{X(t) : t \in\}$ is a **jump process** if

- there is a **jump rate density** $r(x, y)$ with the **exit rate**

$$R(x) = \int_{\mathbb{R}} r(x, y) \, \mathrm{d}y \le M < \infty, \ \forall x \in \mathbb{R},$$

  where $M \in \mathbb{R}$ is a constant;

- its transition density satisfies

$$p_{\Delta t}(x, y) = r(x, y)\Delta t + \big(1 - R(x)\Delta t\big)\, \delta(y - x) + o(\Delta t), \ \text{as}\, \Delta t \to 0.$$

Theorem 3.5.2 (The Kolmogorov-Feller Equation)**.** *The Chapman-Kolmogorov equation of a jump process turns into the* ***Kolmogorov-Feller equation*** *for initial condition $x \in \mathbb{R}$*

$$\frac{\partial}{\partial t} p_t(x, y) = \int_{\mathbb{R}} p_t(x, z) r(z, y) - p_t(x, y) r(y, z) \, \mathrm{d}z$$

3.5.2 Diffusion Processes

Definition 3.5.5 (The Brownian Motion)**.** The **Brownian motion** is a Gaussian process $B : \mathbb{R}_+ \mapsto \mathbb{R}$ with $m(t) = 0$ and $c(t, t') = \min(t, t')$ and almost surely continuous paths.

Proposition 3.5.1 (Brownian Motions are Markov)**.** *A Brownian motion is Markov, and it has independent increments: $\forall t_1 < \cdots < t_n$, $(X(t_{k+1}) - X_{t_k})_{k=1,\cdots,n-1}$ are independent variables.*

Proposition 3.5.2 (Brownian Motions are Homeogeneous)**.** *Furthermore, the increments are stationary: $X(t) - X(s)$ and $X(t - s) - X(0) = X(t - s)$ have the same distribution, for t s. So $B(t)$ is homoegeneous.*

Remark. $B(t)$ is not stationary.

Proposition 3.5.3. *The transition density $p_t(x, y)$ of a Brownian motion is a Gaussian PDF with mean $y - x$ and variance t, which satisfies the heat equation (or diffusion equation):*

$$\frac{\partial p_t}{\partial t} = \frac{1}{2} \frac{\partial^2 p_t}{\partial y^2}$$

with the initial condition $p_0(x, y) = \delta(y - x)$.

Proposition 3.5.4. *Brownian motions are normally distributed: $B(t) \sim \mathcal{N}(0, t)$.*

Proposition 3.5.5. *$B(t)$ is scale-invariant: $B(\lambda t)$ and $\sqrt{\lambda} B(t)$ have the same distribution.*

Proposition 3.5.6. *$B(t)$ is almost surely continuous, but it is also almost surely nowhere differentiable. Actually,*

$$\xi_{t,h} := \frac{B(t + h) - B(t)}{h} \sim \mathcal{N}\left(0, \frac{1}{h}\right).$$

Although Brownian motions are almost surely nowhere differentiable, we can still informally talk about the limit proecss $\xi_t := \lim_{h \to 0} \xi_{t,h}$.

Definition 3.5.6 (Gaussian White Noises)**.** $\xi_t := \lim_{h \to 0} \xi_{t,h}$ is called the **Gaussian white noise**.

Remark. The Gaussian white noise can be considered as a limiting case of a Gaussian process with mean $m(t) = 0$ and $c(t, t') = \delta(t - t')$.

Proposition 3.5.7. *$B(t) = \int_0^t \xi_{t'} \, \mathrm{d}t'$, or we can write it as a stochastic differential equation*

$$\frac{\mathrm{d}B}{\mathrm{d}t} = \xi,$$

with $B(0) = 0$.

3.6 Generators as Operators

3.6.1 Generators of Discrete Continuous-Time Markov Chains

For a continuous-time Markov chain with a countable state space S, for any function $f : S \mapsto \mathbb{R}$, we have

$$\mathbb{E}[f(X(t))] = \sum_{x \in S} \boldsymbol{\pi}_t(x) f(x) = \boldsymbol{\pi}_t \vec{f},$$

where $\vec{f}$ is a column vector of values of f at all the state $x \in S$.

We may be interested in how fast $\mathbb{E}[f(X(t))]$ varies with time t, so

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}[f(X(t))] = \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{\pi}_t \vec{f} = \boldsymbol{\pi}_t G \vec{f}.$$

Thus, we can think of the generator G as acting on the function f by

$$(Gf)(x) = \sum_{y \in S} G_{x,y} f(y) = \sum_{\substack{y \neq x \\ y \in S}} G_{x,y}(f(y) - f(x)).$$

3.6.2 Generators of Continuous Continuous-Time Markov Chains

The idea of generators as operators can be extended to $S = \mathbb{R}$ by replacing matrices and vectors with operators and functions.

Generators of Brownian Motions

For a Brownian motion,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}[f(X(t))] =& \frac{\partial}{\partial t}\int_{\mathbb{R}} p_t(x,y) f(y)\,\mathrm{d}y \\ =& \int_{\mathbb{R}} \frac{\partial}{\partial t} p_t(x,y) f(y)\,\mathrm{d}y \\ =& \frac{1}{2}\int_{\mathbb{R}} \frac{\partial^2}{\partial y^2} p_t(x,y) f(y)\,\mathrm{d}y \\ =& \mathbb{E}[(\mathcal{L}f)(X(t))]\end{aligned}$$

with $(\mathcal{L}f)(x) := \frac{1}{2}f''(x)$, assuming f is twice differentiable and $f(x)\,\&f'(x) \to 0$ as $x \to \pm\infty$ (integration by parts). $\mathcal{L}$ is the generator but now a linear operator on functions.

Generators of Jump Processes

For a jump process on $\mathbb{R}$,

$$(\mathcal{L}f)(x) = \int_{\mathbb{R}} r(x,y)[f(y) - f(x)]\,\mathrm{d}y.$$

We can obtain the Brownian motion as a scaling limit of a jump process. Take a jump process $X(t)$ with $r(x,y) = q(y-x)$ such that $\int_{\mathbb{R}} zq(z)\,\mathrm{d}z = 0$ and $\int_{\mathbb{R}} z^2 q(z)\,\mathrm{d}z = \sigma^2 \in (0,\infty)$. Then $\forall T > 0$, with $X(0) = 0$,

$$\frac{\epsilon}{\sigma}X(\frac{t}{\epsilon^2})\bigg|_{t\in[0,T]} \xrightarrow{\mathrm{d}} B(t)\big|_{t\in[0,T]}\,, \text{ as } \epsilon \to 0.$$

We can prove this by Taylor expansion of the generator

$$f(y) = f(x) + (y-x)f'(x) + \frac{1}{2}(y-x)^2 f''(x) + \cdots,$$

and tightness of the set S of probability distributions for the scaled jump process: $\forall \eta > 0$, $\exists K \in \mathbb{R}$ such that for all $\mu \in \bar{S}$, $\mu(K^c) < \eta$.

3.7 General Diffusion Processes

Definition 3.7.1 (General Diffusion Processes)**.** A **general diffusion process** is a Markov process on $\mathbb{R}$ with the generator of the form

$$(\mathcal{L}f)(x) = a(x,t)f'(x) + \frac{1}{2}\sigma^2(x,t)f''(x),$$

for some functions a (which is called the **drift**) and σ (which is called the **noise**).

Example 3.7.1 (Ornstein-Uhlenbeck Processes)**.** An **Ornstein-Uhlenbeck process** has the generator

$$(\mathcal{L}f)(x) = -\alpha x f'(x) + \frac{1}{2}\sigma^2 f''(x),$$

for some $\alpha > 0$ and $\sigma > 0$. The drift is $-\alpha x$, which is **mean reverting**.

Remark. We have already seen a definition of the Ornstein-Uhlenbeck process as a Gaussian process. We will also formulate it as a stochastic differential equation

$$\frac{\mathrm{d}X}{\mathrm{d}t} = -\alpha X + \sigma\xi,$$

where ξ is the Gaussian white noise, or

$$\mathrm{d}X = -\alpha X\,\mathrm{d}t + \sigma\,\mathrm{d}B \quad \text{(to be explained)}.$$

Example 3.7.2 (Brownian Bridges)**.** A **Brownian bridge** has the generator

$$(\mathcal{L}f)(x) = -\frac{x}{1-t}f'(x) + \frac{1}{2}f''(x),$$

which is only defined on $t \in [0,1)$.

Equivalently, it is a Brownian motion conditioned on $B(1) = 0$.

Example 3.7.3 (Branching Processes)**.** A **branching process** is a diffusion process with $a(x,t) = \alpha x$ for some constant $\alpha > 0$ and $\sigma^2(x) = \beta x$ for some $\beta > 0$, defined on $x \geq 0$.

3.8 More on Generators

The generator $\mathcal{L}$ are defined on functions on the state space but also tell you how probability distributions evolve, using the adjoint $\mathcal{L}^*$.

Probability distributions are linear functionals on a set of continuous functions $S \mapsto \mathbb{R}$, in comparison with row vectors in the case of discrete state space in which a row vector is a linear functional on a set of possible column vectors. Represent a linear functional when $S = \mathbb{R}$ by an integral with respect to a probability density p:

$$f \mapsto \int_{\mathbb{R}} p(x)f(x)\,\mathrm{d}x \in \mathbb{R}.$$

We start from

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}[f(X(t))] = \mathbb{E}[f(X(t))].$$

Notice that

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathbb{R}} p_t(x,y)f(y)\,\mathrm{d}y = \int_{\mathbb{R}} p_t(x,y)\mathcal{L}f(y)\,\mathrm{d}y.$$

Suppose we are considering the diffusion process with $\mathcal{L}(f) = af' + \frac{1}{2}\sigma^2 f''$, and assume $p\,\&\frac{\partial p}{\partial y} \to 0$ as $y \to \infty$. Integrate by parts (twice) to get

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathbb{R}} p_t(x,y)f(y)\,\mathrm{d}y &= \int_{\mathbb{R}} p_t(x,y)\mathcal{L}f(y)\,\mathrm{d}y\\ &= \int_{\mathbb{R}}\left[-\frac{\partial}{\partial y}\left(a(y,t)p_t(x,y)\right) + \frac{1}{2}\frac{\partial^2}{\partial y^2}\left(\sigma^2(y,t)p_t(x,y)\right)\right] f(y)\,\mathrm{d}y,\end{aligned}$$

which is true for all $f \in C^2(\mathbb{R})$. Thus

$$\frac{\partial p_t}{\partial t} = -\frac{\partial}{\partial y}\left(ap_t\right) + \frac{1}{2}\frac{\partial^2}{\partial y^2}\left(\sigma^2 p_t\right),$$

which is called the **Fokker-Planck equation**. Regard $-\frac{\partial}{\partial y}\left(ap_t\right) + \frac{1}{2}\frac{\partial^2}{\partial y^2}\left(\sigma^2 p_t\right)$ as a function of y, and denote it as $\mathcal{L}^* p_t$.

Definition 3.8.1 (The Fokker-Planck equation)**.** The **Fokker-Planck equation** for a diffusion process is

$$\frac{\partial p_t}{\partial t} = -\frac{\partial}{\partial y}\left(ap_t\right) + \frac{1}{2}\frac{\partial^2}{\partial y^2}\left(\sigma^2 p_t\right).$$

Suppose the a, σ are t-independent, then we get the stationary density

$$p^*(x) = \frac{1}{Z}\exp\left(\int_0^x \frac{2a(y) - (\sigma^2)'(y)}{\sigma^2(y)}\right)\mathrm{d}y,$$

where Z is the normalisation constant.

Example 3.8.1. For an Orstein-Uhlenbeck process,

$$p*(x) = \frac{1}{Z}\exp\left(\int_0^x -\frac{2\alpha y}{\sigma^2}\,\mathrm{d}y\right) = \frac{1}{Z}\exp\left(-f\frac{\alpha x^2}{\sigma^2}\right),$$

which is the density function of $\mathcal{N}(0, \frac{\sigma^2}{2\alpha})$.

Proposition 3.8.1. *The Fokker-Planck equation is an advection-diffusion equation with diffusion* $D = \frac{\sigma^2}{2}$ *and advection velocity* $v = a - \sigma\sigma'$*.*

Definition 3.8.2 (The Advection-Diffusion Equation)**.** A general **advection-diffusion equation** for the density of a conserved quantity ρ is

$$\frac{\partial \rho}{\partial t} + \mathrm{div}(\rho v - D\nabla\rho) = 0.$$

Remark. For an advection-diffusion equation, the stationary density ρ corresponds to $\mathrm{div}(\rho v - D\nabla\rho) = 0$, so in 1-D

$$\begin{aligned}\rho v =& D\nabla\rho\\ \rho =& \frac{1}{Z}\exp\left(\int_0^x \frac{v(y)}{D(y)}\,\mathrm{d}y\right).\end{aligned}$$

Definition 3.8.3 (Real Brownian Motions)**.** A **real Brownian motion** is better modelled by a Langevin equation:

$$m\ddot{X} + \gamma\dot{X} = \sigma\xi.$$

Remark. Note that this is an Ornstein-Uhlenbeck process for the velocity $\dot{X}$, so real Brownian motion is an integrated Ornstein-Uhlenbeck process. It is almost surely differentiabl in contrast to Brownian motion. But as Largvin noted the timescale for the mean reversion of γ is about 10^{-8} seconds. As a result, if you look on timescales greater than 10^{-8} seconds, it looks like the Brownian motion

$$\gamma\dot{X} = \sigma\xi.$$

3.9 Stochastic Differential Equations

Example 3.9.1 (Diffusion Processes)**.** For a diffusion process on $\mathbb{R}$, it satisfies

$$\mathrm{d}X = a(X,t)\,\mathrm{d}t + \sigma(X,t)\,\mathrm{d}B.$$

We interpret this as the limit of timestep for a computational method with $a, \sigma \in C^1(\mathbb{R})$, but there are many different interpretations if σ depends on X.

3.9.1 Ito's Interpretation

The Euler-Maruyama Step

Evaluate σ at the beginning of the step:

$$X(t+h) - X(t) = a(X(t),t)h + \sigma(X(t),t)[B(t+h) - B(t)] + o(h) \quad \text{as} \quad h \to 0.$$

Remark. We can use $B(t+h) - B(t) = h\xi_h(t) \sim \mathcal{N}(0,h)$ to avoid the implicit term $B(t+h)$.

Remark. This corresponds to $(\mathcal{L}f)(x) = af' + \frac{1}{2}\sigma^2 f''$ for $f \in C^2(\mathbb{R})$, because

$$f(X(t+h)) - f(X(t)) = f'(X(t))[ah + \sigma h\xi_h + o(h)] + \frac{1}{2}f''(X(t))[ah + \sigma h\xi_h + o(h)]^2 + o(h).$$

Thus

$$\begin{aligned}\mathbb{E}[f(X(t+h)) - f(X(t))|X(t)] =& f'(X(t))[ah + o(h)] + \frac{1}{2}f''(X(t))\sigma^2 h + o(h)\\ =& (\mathcal{L}f)(X(t)) + o(h).\end{aligned}$$

3.9.2 The Stratonovich Rule

We can also interpret a stochastic differential equation using the midpoint rule.

For $\mathrm{d}X = b\,\mathrm{d}t + \sigma\,\mathrm{d}B$ or $dotX = b + \sigma\xi$ where ξ is the Gaussian white noise, it means

$$\begin{aligned}X(t+h) - X(t) =& \frac{1}{2}\left[b(X(t+h)) - b(X(t))\right] + \frac{1}{2}\left[\sigma(X(t+h)) + \sigma(X(t))\right]h\xi_h + o(h)\\ =& b(X(t))h + o(h) + \sigma(X(t)) + \frac{1}{2}\sigma'\sigma h\xi_h + o(\sqrt{h}).\end{aligned}$$

This is an implicit method, but it becomes explicit if we move $\frac{1}{2}\sigma'\sigma$ into b, i.e. $a = b + \frac{1}{2}\sigma'\sigma$. So this converts between the Stratonovich's and Ito's interpretations.