Random Processes: 2023/24

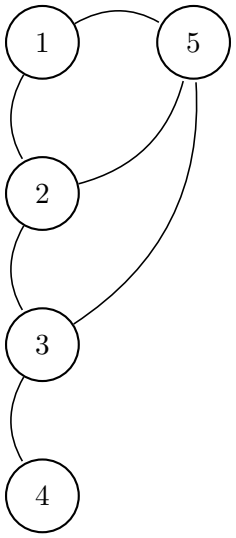
Dr J Robert Johnson

Contents

1 Introduction

1.1 Motivating Example

Imagine a maze of rooms numbered $1, 2, \ldots, n$ with passages joining some of them as in the figure below. We start in room 1 and moves around the maze randomly. Every minute we choose a passage leading from our current location at random (with each such choice equally likely) and move along it. What happens?



Of course, on one level this is a silly question; since the moves are at random anything could happen! However, there are sensible questions we can pose (and maybe answer):

- What is the probability that we reach room 2 before room 3?
- What is the expectation of the number of steps before our first visit to room 4?
- In the long run what proportion of time do we expect to spend in each room?

By the end of the first half of this module you will have seen some techniques to answer questions like these.

1.2 Stochastic Processes

Our route through the maze is an example of a stochastic process (that is a process that evolves randomly over time). Let's formalise that idea a bit.

Definition 1. A *stochastic process* is a collection of random variables $\{X_t : t \in T\}$, indexed by a set T, where each random variable X_t takes values in a set S. The index-set T is usually thought of as 'time'. The set S is called the *state space* of the stochastic process[1].

In this module we will often consider time in discrete steps starting from time 0 so it will be convenient to take $T = \mathbb{N} = \{0, 1, 2, 3, \dots\}$.

For our maze example, let's take $T = \mathbb{N}$ and let X_t be the number of the room we are in after we have been in the maze for t minutes. Our starting location is room 1 so $X_0 = 1$. At time 1 minute we move to either room 2 (with probability 1/2) or room 5 (with probability 1/2). So X_1 is a random variable with

$$\mathbb{P}(X_1 = 2) = 1/2, \quad \mathbb{P}(X_1 = 5) = 1/2.$$

With more work we could calculate more probabilities but things quickly get quite involved. For instance

$$\begin{aligned}\mathbb{P}(X_3 = 3) &= \mathbb{P}(X_1 = 2, X_2 = 5, X_3 = 3) + \mathbb{P}(X_1 = 5, X_2 = 2, X_3 = 3) \\ &= \left(\frac{1}{2} \times \frac{1}{3} \times \frac{1}{3}\right) + \left(\frac{1}{2} \times \frac{1}{3} \times \frac{1}{3}\right) \\ &= \frac{1}{9}\end{aligned}$$

Make sure you can see how this calculation works; we have figured all possible trajectories which take us to room 3 at time 3 and then found the probability of each one.

To work out something more complicated like the probability that $X_8 = 3$ or the probability of reaching room 2 before room 3 we will need some different ideas.

Before we leave this example let's notice one more feature. Suppose that I know I am in room k at time t. Conditioned on this the random variable X_{t+1} (i.e. the location after the next move) does not depends on my location at times $t-1, t-2, \dots$. For instance:

$$\mathbb{P}(X_{t+1} = 2 \mid X_t = 1, X_{t-1} = 2) = \mathbb{P}(X_{t+1} = 2 \mid X_t = 1, X_{t-1} = 5) = \frac{1}{2}.$$

This is an extremely important property which is the basis of the definition of a *Markov chain*.

In the first half of the course, we will take $T = \mathbb{N}$, so the stochastic process can be written as $(X_0, X_1, X_2, \dots)$. This is called a *discrete-time stochastic process*. It is a good model to use when the system changes only at fixed times (for example in the maze example we moved at fixed 1 minute intervals). It can also be used if we are only interested in the state of a system at fixed times. For example, we might be observing the number of yeast cells in a culture, at the end of each hour after the start of an experiment, or the

[1] In this module, the state space S will always be either finite or countably infinite.

number of people in a supermarket queue n minutes after the shop opens.

In the second half of the course, we will consider *continuous-time stochastic processes*, where $T = \mathbb{R}_{\geqslant 0} = \{x \in \mathbb{R} : x \geqslant 0\}$. This is a good model to use for systems which can change state at any time. For example, we might want to keep track of the number of people in a queue at *all* times after it opened (meaning, we would want to know exactly when people joined or left the queue). This is most naturally described in continuous time since people may join the queue, or be served and leave, at any time.

1.3 Summary of the Module

Our motivating example of moving randomly through a maze is a special type of process called a Markov chain (roughly this is because our choice at each step depends only on the room we are in and not how we got there). This module is about Markov chains and develops techniques to answer questions like the ones posed for the maze example above. The first half of the module covers discrete-time Markov chains (where movement is at fixed time intervals) in some detail. We will:

- Define and introduce the basics of Markov chains, how to describe them, and some important terminology. (Week 1)
- Define absorbing states and describe a method for answering questions like what is the probability I reach room 2 before room 3. (Week 2)
- Study the long-term behaviour of Markov chains. This includes questions such as what proportion of time do we expect to spend in each room over a long period of time. Tools from linear algebra appear in the theory here. (Weeks 3, 4)
- Define and see how to use the concepts of recurrence and transience. These express a dichotomy involving the chance of returning to our starting state. (Week 5)
- Study a range of examples to illustrate the theory of the previous weeks. (Week 6)

The second half looks at some special cases of continuous time Markov chains leading up to continuous-time Markov chains. Being in continuous-time, techniques from continuous mathematics (calculus and differential equations) will play a role.

- Define the Poisson process which models events which happen randomly in continuous time at a regular rate (such as customers arriving in queue). The Poisson process is the most fundamental continuous-time stochastic process. We will see how to calculate simple probabilities and study various random variables associated with it. (Week 8,9)
- Study the birth process which generalises the Poisson process in allowing arrival rates that depend on the current state of the system. (Week 10)

- Define continuous-time Markov chains in generality. We will develop some of the theory here but in less detail than the discrete-time case, without too many technicalities, and with lots of examples such as birth-death processes and queues. (Week 11, 12)

2 Markov Chain Fundamentals

In this section we will define Markov chains and see how to describe them. Towards the end of the section we will see a variety of examples illustrating all the concepts. For now, a good example to have in mind is the random walk through a maze from the introduction.

2.1 The Markov property

Recall that if A and B are events with $\mathbb{P}(A) \neq 0$, then we define the *conditional probability of B given A* as

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

This can be thought of as the probability that the event B occurs, given that we know already that the event A occurs.

Definition 2. Let S be a finite or countable set. A discrete-time stochastic process $(X_0, X_1, X_2, \ldots)$ with state space S is said to be a *Markov chain* if it satisfies, for all $t \in \mathbb{N}$,

$$\mathbb{P}(X_{t+1} = i_{t+1} \mid X_t = i_t, X_{t-1} = i_{t-1}, \ldots, X_1 = i_1, X_0 = i_0) = \mathbb{P}(X_{t+1} = i_{t+1} \mid X_t = i_t),$$

for all $i_{t+1}, i_t, i_{t-1}, \ldots, i_1, i_0 \in S$ for which $\mathbb{P}(X_t = i_t, X_{t-1} = i_{t-1}, \ldots, X_1 = i_1, X_0 = i_0) \neq 0$. (Note that if $\mathbb{P}(X_t = i_t, X_{t-1} = i_{t-1}, \ldots, X_1 = i_1, X_0 = i_0) = 0$, then the conditional probability on the left-hand side is not defined.)

This is called the *Markov property*. It says that if we know the state of the chain at a certain time, then we know the probability distribution of the state of the chain at any future time. This is possibly the most important definition of the module. Make sure you understand what it saying (the formal description makes it look rather more intimidating than it really is!). One way to remember it is with the slogan:

The future, conditioned on the present, does not depend on the past.

2.2 Transition Probabilities

Definition 3. If $\mathbb{P}(X_{t+1} = j \mid X_t = i)$ is independent of t, then we say that the Markov chain is *homogeneous*. In this case, we write $\mathbb{P}(X_{t+1} = j \mid X_t = i) = p_{ij}$, and we call the p_{ij}'s the *transition probabilities*[2].

For the rest of the module, unless stated otherwise, 'Markov chain' will always mean 'homogeneous Markov chain'.

[2]For notational clarity, we will sometimes write $p_{i,j}$ rather than p_{ij} for the transition probabilities.

The idea is that we have a transition probability $\mathbb{P}(X_{t+1} = j \mid X_t = i)$ of going from state i to state j and this does not depend on t (this is the homogeneous condition) or on any information I may give you about $X_{t-1}, X_{t-2}, \ldots, X_0$ (this is the Markov property).

To show that a process is *not* a Markov chain we need to find some specific instance where the Markov property fails. For example if I can find states a, b, c, d such that

$$\mathbb{P}(X_{t+1} = a \mid X_t = b, X_{t-1} = c) \neq \mathbb{P}(X_{t+1} = a \mid X_t = b, X_{t-1} = d)$$

then I have shown that the process $(X_0, X_1, X_2, \ldots)$ is not a Markov chain.

Suppose $(X_0, X_1, X_2, \ldots)$ is a (homogeneous, discrete-time) Markov chain, and we know the probability distribution of X_0, and all the transition probabilities. Then the following easy lemma tells us the joint probability distribution of the random variables $(X_0, X_1, \ldots, X_t)$, for any $t \in \mathbb{N}$. In other words, the initial distribution and the transition probabilities contain all the information we need to work out the probability of following any trajectory through the state space.

Lemma 2.1. *If $(X_0, X_1, X_2, \ldots)$ is a (homogeneous) Markov chain with state space S, with $\mathbb{P}(X_0 = i) = \mu_i^{(0)}$ for all $i \in S$, and with transition probabilities p_{ij} $(i, j \in S)$, then*

$$\mathbb{P}(X_0 = i_0, X_1 = i_1, X_2 = i_2, \ldots, X_t = i_t) = \mu_{i_0}^{(0)} p_{i_0,i_1} p_{i_1,i_2} \cdots p_{i_{t-2},i_{t-1}} p_{i_{t-1},i_t},$$

for all $i_0, i_1, \ldots, i_t \in S$.

Proof.

$$\begin{aligned}
&\mathbb{P}(X_0 = i_0, X_1 = i_1, \ldots, X_t = i_t) \\
&\quad = \mathbb{P}(X_0 = i_0) \times \mathbb{P}(X_1 = i_1 \mid X_0 = i_0) \times \mathbb{P}(X_2 = i_2 \mid \{X_1 = i_1, X_0 = i_0\}) \\
&\qquad \times \cdots \times \mathbb{P}(X_t = i_t \mid \{X_{t-1} = i_{t-1}, \ldots, X_0 = i_0\}) \\
&\qquad\qquad \text{(by the definition of conditional probability)} \\
&\quad = \mathbb{P}(X_0 = i_0)\mathbb{P}(X_1 = i_1 \mid X_0 = i_0)\mathbb{P}(X_2 = i_2 \mid X_1 = i_1) \cdots \mathbb{P}(X_t = i_t \mid X_{t-1} = i_{t-1}) \\
&\qquad\qquad \text{(by the Markov property)} \\
&\quad = \mu_{i_0}^{(0)} p_{i_0,i_1} p_{i_1,i_2} \cdots p_{i_{t-1},i_t}. \qquad \text{(by definition of } p_{ij} \text{ and } \mu_i^{(0)})
\end{aligned}$$

□

We look at two ways of writing the transition probabilities of a Markov chain. The first (transition graphs) is easier for visualising the process,the second (transition matrices) is good for calculations.

2.2.1 Transition graphs

A convenient way of visualising a Markov chain is by drawing its *transition graph*. The set of 'vertices' (or 'nodes') of the transition graph of a Markov chain is the set of states S. For each pair or states i, j for which the transition probability p_{ij} is non-zero, there is a directed 'edge' (or 'arc') from vertex i to vertex j, labelled with the transition probability p_{ij}.

2.2.2 Transition matrices

Definition 4. Let $(X_0, X_1, X_2, \ldots)$ be a (homogeneous) Markov chain with state space S and transition probabilities $(p_{ij} : \ i, j \in S)$. Its *transition matrix* P is the matrix with rows and columns indexed by S, with (i, j)-th entry $P_{i,j} = p_{ij}$. In other words, it is just the matrix of transition probabilities.

Since we usually insist that matrices have finitely many rows and columns, we will only use the transition matrix when S is finite.

Remark. Note that for any transition matrix P, all the row-sums are 1:

$$\sum_{j \in S} p_{ij} = 1 \quad \forall i \in S.$$

This is clear, as if $X_0 = i$, then we must have $X_1 = j$ for exactly one state $j \in S$, so

$$\sum_{j \in S} p_{ij} = \sum_{j \in S} \mathbb{P}(X_1 = j \mid X_0 = i) = 1.$$

A square matrix of non-negative numbers with all its row-sums equal to 1 is often called a *stochastic* matrix. Notice that any stochastic matrix is the transition matrix of some Markov chain. (Can you see why?)

The transition probabilities are the probabilities of moving between states. In Lemma 2.1 the other quantity we needed to know was the distribution of the X_0 (in other words where the chain starts).

Definition 5. The *initial distribution* of a Markov chain $(X_0, X_1, X_2, \ldots)$ is the probability distribution of X_0, i.e. the row-vector $\boldsymbol{\mu}^{(0)}$ with entries indexed by S, where

$$\mu_i^{(0)} = \mathbb{P}(X_0 = i)$$

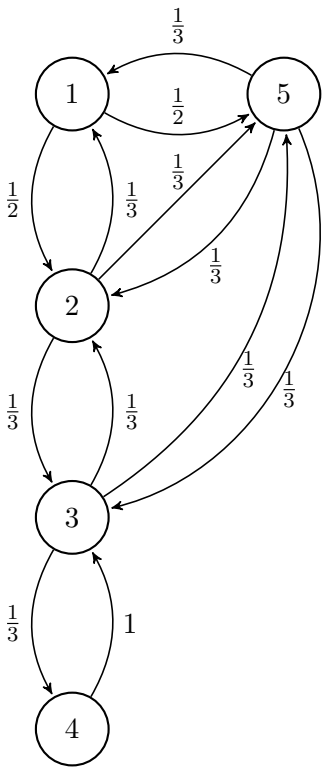
for all $i \in S$.

We will see later that using a matrix and a row vector to specify the transition probabilities and initial distribution is not just a convenient way of writing them down. Ideas from linear algebra will make an appearance.

2.3 Examples

In this section I'll describe a variety of examples of Markov chains. We will see how to represent them as transition graphs and transition matrices, and how to go from a description in words to a mathematical model. There are more examples on the first problem sheet.

Example 1 (Rooms and Passages)**.** Recall the maze example in the introduction, with X_t being the number of the room occupied at time t minutes. We have $(X_0, X_1, X_2, \ldots)$ is a Markov chain with the following transition graph.



The transition probabilities on this picture come from the formula:

$$p_{ij} = \begin{cases} \frac{1}{\text{number of passages out of room } i} & \text{if there is a passage from } i \text{ to } j; \\ 0 & \text{otherwise.} \end{cases}$$

We could also express the transition probabilities as a transition matrix:

$$\begin{pmatrix} 0 & 1/2 & 0 & 0 & 1/2 \\ 1/3 & 0 & 1/3 & 0 & 1/3 \\ 0 & 1/3 & 0 & 1/3 & 1/3 \\ 0 & 0 & 1 & 0 & 0 \\ 1/3 & 1/3 & 1/3 & 0 & 0 \end{pmatrix}.$$

I haven't mentioned the initial distribution yet. In the description from the introduction we started in room 1 so $X_0 = 1$. This could also be expressed by saying that the initial distribution is the row vector:

$$\boldsymbol{\mu}^{(0)} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

We could change the chain by making the starting state random as well. Suppose we start by picking a room uniformly at random. Then X_0 is a random variable with $\mathbb{P}(X_0 = i) = 1/5$ for $1 \leqslant i \leqslant 5$. Equivalently the initial distribution is the row vector:

$$\boldsymbol{\mu}^{(0)} = \begin{pmatrix} 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \end{pmatrix}.$$

You could pick some more complicated probability distribution as the initial distribution if you wanted to.

Non-Example 2. Suppose we modify the previous rooms and passages example to forbid backtracking. In other words, after each step, if the room I am in has d passages, I ignore the one I came in through and select randomly from the remaining $d-1$ passages. This is not a Markov chain. This is intuitively clear: at any point my next step depends not just on my location but where I was previously. To make it formal let's use that idea to show that the definition of the Markov property is violated.

Suppose that $X_t = 2$ then the distribution of X_{t+1} depends on whether X_{t-1} is 1, 3 or 5. In particular,

$$\begin{aligned} &\mathbb{P}(X_{t+1} = 1 \mid X_t = 2, X_{t-1} = 1) = 0 \qquad \text{(because of the no backtracking rule)} \\ &\mathbb{P}(X_{t+1} = 1 \mid X_t = 2, X_{t-1} = 3) = 1/2 \end{aligned}$$

So with this modification $(X_0, X_1, X_2, \ldots)$ is not a Markov chain.

Non-Example 3. Again we will modify the previous rooms and passages example. This time suppose that the passage joining rooms 1 and 2 becomes blocked at time 1.5 minutes. The effect of this is to make the transition probabilities depend on time. For example:

$$\begin{aligned} &\mathbb{P}(X_1 = 2 \mid X_0 = 1) = 1/2 \\ &\mathbb{P}(X_2 = 2 \mid X_1 = 1) = 0 \end{aligned}$$

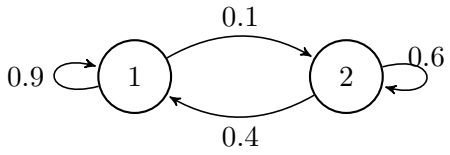
The Markov property still holds because the probability of moving from room i to room j only depends on my current location X_t and time t.

So with this modification $(X_0, X_1, X_2, \ldots)$ is still a Markov chain but it is not homogeneous.

Example 4 (Simple weather model)**.** Suppose that each day is either dry (state 1) or wet (state 2) and that the state of the weather tomorrow is a random variable depending only on the state of the weather today[3] with

$$\mathbb{P}(\text{a dry day is followed by a wet day}) = 0.1$$
$$\mathbb{P}(\text{a wet day is followed by a dry day}) = 0.4.$$

If we let X_t be the state of the weather on day t of observations then $(X_0, X_1, X_2, \ldots)$ is a Markov chain with transition graph



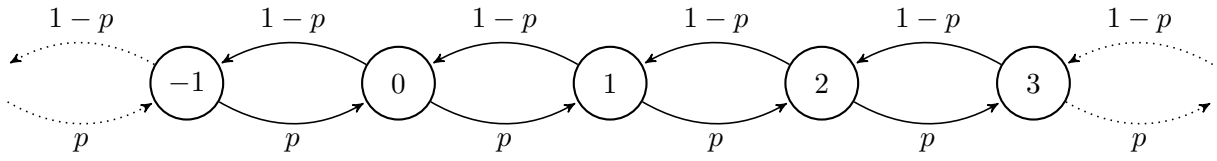
and transition matrix

$$\begin{pmatrix} 0.9 & 0.1 \\ 0.4 & 0.6 \end{pmatrix}$$

Example 5 (Random walk on $\mathbb{Z}$)**.** Suppose I have a coin which has probability p of showing Heads. I start at 0 and perform a random walk on $S = \mathbb{Z}$ by tossing the coin and taking one step in the positive direction if it shows Heads and one step in the negative direction if it shows Tails. My position X_t after t coin tosses is a Markov chain. It has $X_0 = 0$ and transition probabilities:

$$p_{ij} = \begin{cases} p & \text{if } j = i + 1; \\ 1 - p & \text{if } j = i - 1; \\ 0 & \text{otherwise.} \end{cases}$$

The transition graph is:



[3]In reality this assumption is probably not very realistic. However, sometimes this kind of simple model can be a useful approximation.

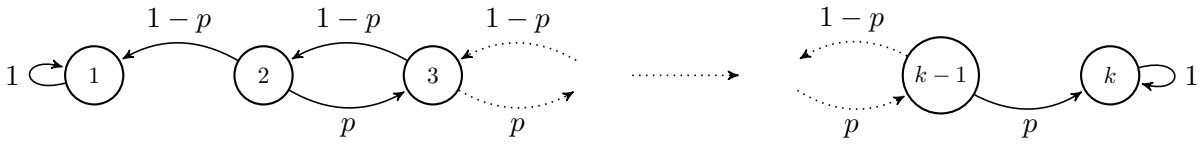
Unlike our earlier examples, this one has an infinite state space (the set of integers). Some parts of the theory of Markov chains become more complicated in this case. Most of the time we will deal with finite state spaces but we will see something of how things extend to the infinite (countable) case.

I won't attempt to write down a transition matrix for this example because we usually don't consider matrices with an infinite number of rows and columns.

Example 6 (1-dimensional random walk with absorbing boundaries)**.** It is possible to modify the previous example by insisting that if we ever reach 1 or k (some fixed large integer) we never move again (we will also need to pick a value or probability distribution for X_0). The state space will be $\{1, 2, 3, \ldots, k\}$ with transition probabilities:

$$p_{ij} = \begin{cases} 1 & \text{if } i = j = 1; \\ 1 & \text{if } i = j = k; \\ p & \text{if } 2 \leqslant i \leqslant k-1 \text{ and } j = i+1; \\ 1-p & \text{if } 2 \leqslant i \leqslant k-1 \text{ and } j = i-1; \\ 0 & \text{otherwise.} \end{cases}$$

and transition diagram:



The transition matrix is:

$$\begin{pmatrix} 1 & 0 & 0 & 0 & \cdots & 0 \\ 1-p & 0 & p & 0 & \cdots & 0 \\ 0 & 1-p & 0 & p & \cdots & 0 \\ \vdots & \ddots & & \ddots & & \vdots \\ 0 & \cdots & 0 & 1-p & 0 & p \\ 0 & \cdots & 0 & 0 & 0 & 1 \end{pmatrix}$$

This process could model a gambler who plays a sequence of games. In each game he has a probability p of winning £1 and probability $1-p$ of losing £1. If he ever loses all his money or accumulates £k then he stops (this corresponds to the special states 0 and k). If his initial fortune is £m and we set $X_0 = m$ then X_t represents the amount of money he has after playing t games. We will see later how to figure out the probability that our gambler ends up with £k rather than £0.

Example 7 (Number of Successes)**.** I repeatedly toss a coin which has probability p of showing Heads. Let X_t be the number of heads seen in the first t tosses. This is a Markov chain. The state space is $\mathbb{N}$. Suppose that $X_t = k$ (so after t tosses, I have seen k Heads). There are two possibilities for the next toss:

- It is a Head in which case I have now seen $k+1$ Heads;
- It is a Tail in which case I have still only seen k Heads.

These happen with probabilities p and $1-p$ respectively. So

$$\mathbb{P}(X_{t+1} = k+1 \mid X_t = k) = p, \qquad \mathbb{P}(X_{t+1} = k \mid X_t = k) = 1-p$$

and these probabilities would stay the same under any extra information on $X_{t-1}, X_{t-2} \ldots$. So $(X_0, X_1, X_2, \ldots)$ is a Markov chain with $X_0 = 0$ (after 0 tosses I have seen no Heads) and transition probabilities:

$$p_{ij} = \begin{cases} p & \text{if } j = i+1; \\ 1-p & \text{if } j = i; \\ 0 & \text{otherwise.} \end{cases}$$

A process with $S = \mathbb{N}$ with the property that X_t is non-decreasing in t (that is $X_s \leqslant X_t$ for all $s \leqslant t$) is called a *counting process*. We will see many examples with this property.

3 First-step Analysis

3.1 Conditioning

In this section we will use the Theorem of Total Probability (and variants of it) to calculate probabilities and expectations. This approach is usually called *conditioning*.

Reminder. Supose that $E_1, \ldots, E_m$ is a partition of the sample space, A is an event and T is a random variable.

- $\mathbb{P}(A) = \sum_{i=1}^{m} \mathbb{P}(A \mid E_i)\mathbb{P}(E_i)$ (usually called the Theorem of Total Probability)
- $\mathbb{E}(T) = \sum_{i=1}^{m} \mathbb{E}(T \mid E_i)\mathbb{P}(E_i)$

Often in our examples there will be another event which everything is conditioned on. This gives the following versions in which B is another event.

- $\mathbb{P}(A \mid B) = \sum_{i=1}^{m} \mathbb{P}(A \mid E_i \cap B)\mathbb{P}(E_i \mid B)$
- $\mathbb{E}(T) = \sum_{i=1}^{m} \mathbb{E}(T \mid E_i \cap B)\mathbb{P}(E_i \mid B)$

Finally, the partition will often be given by another random variable. Suppose that X is a random variable taking values in $\{1, \ldots, m\}$. Then setting $E_i = \mathbb{P}(X = i)$ gives a partition of the sample space and the results above can be used. The first one for instance becomes:

- $\mathbb{P}(A) = \sum_{i=1}^{m} \mathbb{P}(A \mid X = i)\mathbb{P}(X = i)$

Strictly speaking, if we were stating these formally we would need to insist that all the conditional probabilities are defined. For instance, in the ordinary Theorem of Total Probability we need that $\mathbb{P}(E_i) > 0$ for all i. However, in practice this should never cause confusion.

We will use these implicitly throughout this section so you need to be comfortable with what they say and fluent in applying them. If you need to brush up on any of this, you can find these Theorems and examples of their use in your Introduction to Probability notes.

3.2 Absorbing states

Look back at example 6. The states 0 and k in this example are special; once the chain has reached one of them it can never leave. This idea motivates the next definition.

Definition 6. Let $(X_0, X_1, X_2, \ldots)$ be a Markov chain with state space S and transition probabilities $(p_{ij} : i, j \in S)$. A state $i \in S$ is said to be *absorbing* if $p_{ii} = 1$.

Absorption is said to occur when a Markov chain first reaches an absorbing state.

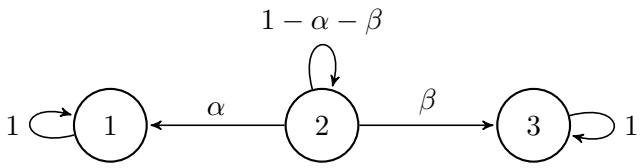
When a Markov chain reaches an absorbing state, it cannot leave: it remains in this state forever.

When a Markov chain has one or more absorbing states, there are several questions we may wish to answer. For example, what is the probability of reaching a certain absorbing state? What is the expected time until absorption occurs? These questions can be answered by 'first-step analysis', also known as 'conditioning on the first step'. The computation reduces to solving simultaneous linear equations.

3.3 First-step analysis for absorption probabilities

We will look at a couple of simple examples before describing the general approach.

Example 8. Let $(X_0, X_1, X_2, \ldots)$ be the Markov chain with $X_0 = 2$ and transition graph



where $0 < \alpha < 1$, $0 < \beta < 1$, $\alpha + \beta < 1$.

Define $T = \min\{n : X_n \neq 2\}$ (the time of absorption). We would like to find the probability that absorption occurs at state 1 rather than state 3.

Define $a = \mathbb{P}(X_T = 1 \mid X_0 = 2)$. Conditioning on X_1 we have

$$\begin{aligned}
a &= \mathbb{P}(X_T = 1 \mid X_1 = 1, X_0 = 2)\mathbb{P}(X_1 = 1 \mid X_0 = 2) + \mathbb{P}(X_T = 1 \mid X_1 = 2, X_0 = 2)\mathbb{P}(X_1 = 2 \mid X_0 = 2) \\
&\quad + \mathbb{P}(X_T = 1 \mid X_1 = 3, X_0 = 2)\mathbb{P}(X_1 = 3 \mid X_0 = 2) \\
&= \mathbb{P}(X_T = 1 \mid X_1 = 1)\mathbb{P}(X_1 = 1 \mid X_0 = 2) + \mathbb{P}(X_T = 1 \mid X_1 = 2)\mathbb{P}(X_1 = 2 \mid X_0 = 2) \\
&\quad + \mathbb{P}(X_T = 1 \mid X_1 = 3)\mathbb{P}(X_1 = 3 \mid X_0 = 2) \qquad \text{(By the Markov property)} \\
&= 1 \times \alpha + a \times (1 - \alpha - \beta) + 0 \times \beta
\end{aligned}$$

In the last line we are using the fact that $\mathbb{P}(X_T = 1 \mid X_1 = 2) = \mathbb{P}(X_T = 1 \mid X_0 = 2)$. This is because each of these gives the probability that the process is absorbed at 1 given that it starts in state 2.

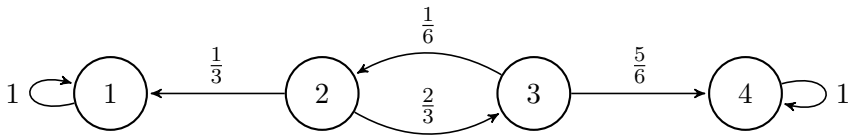
Solving this we get that $a = \mathbb{P}(X_T = 1 \mid X_0 = 2) = \frac{\alpha}{\alpha+\beta}$.

By a similar method we get that $\mathbb{P}(X_T = 3 \mid X_0 = 2) = \frac{\beta}{\alpha+\beta}$.

Since $\frac{\alpha}{\alpha+\beta} + \frac{\beta}{\alpha+\beta} = 1$, one or other of these events happens with probability 1 and the probability that the chain is never absorbed is 0.

That example was not too difficult because after the first step we had either been absorbed or were back where we started. This was because there was only one non-absorbing state. Let's look at a slightly more complicated example where this doesn't hold.

Example 9. Let $(X_0, X_1, X_2, \ldots)$ be the Markov chain with $X_0 = 2$ and transition graph



Define $T = \min\{n : X_n = 1, 4\}$ (the time of absorption).

As before, we would like to work out $a = \mathbb{P}(X_T = 1 \mid X_0 = 2)$ by conditioning on X_1. Unfortunately, the first step could take us to state 3 which is a different starting place. The trick is to introduce a new variable to deal with this. Let

$$\begin{aligned} a_2 &= \mathbb{P}(X_T = 1 \mid X_0 = 2), \\ a_3 &= \mathbb{P}(X_T = 1 \mid X_0 = 3). \end{aligned}$$

Now, conditioning on X_1 in the same way as the previous example we get an equation for a_2 involving a_3:

$$\begin{aligned} a_2 &= \mathbb{P}(X_T = 1 \mid X_1 = 1)p_{21} + \mathbb{P}(X_T = 1 \mid X_1 = 3)p_{23} \\ &= 1 \times \tfrac{1}{3} + \tfrac{2}{3}a_3 \end{aligned}$$

and an equation for a_3 involving a_2:

$$\begin{aligned} a_3 &= \mathbb{P}(X_T = 1 \mid X_1 = 2)p_{32} + \mathbb{P}(X_T = 1 \mid X_1 = 4)p_{34} \\ &= \tfrac{1}{6}a_2 + 0 \times \tfrac{5}{6}. \end{aligned}$$

Solving the simultaneous equations:

$$\begin{aligned} a_2 &= \tfrac{1}{3} + \tfrac{2}{3}a_3, \\ a_3 &= \tfrac{1}{6}a_2, \end{aligned}$$

we get $\mathbb{P}(X_T = 1 \mid X_0 = 2) = a_2 = \frac{3}{8}$.

We can use the same approach more generally.

Suppose $(X_0, X_1, X_2, \ldots)$ is a Markov chain with state space S, transition probabilities $(p_{ij} : \ i, j \in S)$, and absorbing states $A \subseteq S$ (in other words $p_{ii} = 1$ if and only if $i \in A$). Let $T = \min\{n : X_n \in A\}$ be the time of absorption.

Fix $k \in A$ and suppose that we want to find $\mathbb{P}(X_T = k)$. The set up for the method of first-step analysis is as follows:

- Let $a_i = \mathbb{P}(X_T = k \mid X_0 = i)$. We will find equations for a_i and solve them.

We now establish equations for the a_i:

- It is immediate that $a_k = 1$ and $a_l = 0$ for all $l \in A \setminus \{k\}$.
- If $i \in S \setminus A$ then conditioning on X_1 we get

$$a_i = \sum_{s \in S} \mathbb{P}(X_1 = s \mid X_0 = i)\mathbb{P}(X_T = k \mid X_1 = s) = \sum_{s \in S} p_{is} a_s$$

This is because $\mathbb{P}(X_T = k \mid X_1 = s)$ is just the probability that the chain is absorbed in k given that we are in state s. So $\mathbb{P}(X_T = k | X_1 = s) = a_s$.

Substituting the value of a_i for $i \in A$ we get

$$a_i = p_{ik} + \sum_{s \in S \setminus A} p_{is} a_s$$

The result of all this is $|S|$ equations for the $|S|$ variables a_i. We now solve them to find the a_i that we want.

We have proved the first part of the following theorem.

Theorem 3.1. *Suppose $(X_0, X_1, X_2, \ldots)$ is a Markov chain with state space S, transition probabilities $(p_{ij} : i, j \in S)$, and absorbing states $A \subseteq S$. Let $T = \min\{n : X_n \in A\}$ denote the time of first absorption.*

(i) Fix $k \in A$. For each $i \in S$, let $a_i = \mathbb{P}(X_T = k \mid X_0 = i)$. Then the $(a_i : i \in S)$ satisfy the following system of simultaneous linear equations:

$$a_i = \begin{cases} 1, & \text{if } i = k; \\ 0, & \text{if } i \in A \setminus \{k\}; \\ p_{ik} + \sum_{s \in S \setminus A} p_{is} a_s & \text{if } i \in S \setminus A. \end{cases}$$

(ii) If S is finite and from every state it is possible to reach an absorbing state in some number of steps, then absorption occurs with probability 1, and the system of equations in (i) has a unique solution.

The proof of part (ii) is omitted and non-examinable. Later we will see an example of how the conclusion of part (ii) may fail when S is infinite.

3.4 First-step analysis for absorption time and related quantities

We can also use first-step analysis to determine the expectation of the time of absorption. Again, let's do an example first.

Example 10. Let $(X_0, X_1, X_2, \ldots)$ be the Markov chain on state space $S = \{1, 2, 3, 4\}$ with $X_0 = 1$ and transition matrix:

$$\begin{pmatrix} 0 & 1/3 & 1/3 & 1/3 \\ 2/3 & 0 & 1/3 & 0 \\ 1/3 & 2/3 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

(Why not draw the transition graph to visualise this better?)

State 4 is the only absorbing state so we define $T = \min\{n : X_n = 4\}$ (the time of absorption). We want to find $\mathbb{E}(T \mid X_0 = 1)$. As before the trick is to make the starting state a variable. So let $u_i = \mathbb{E}(T \mid X_0 = i)$.

Clearly $u_4 = 0$. For $i \neq 4$ we have

$$\begin{aligned} u_i &= 1 + \mathbb{E}(\text{the number of further steps before absorption after the first one}) \\ &= 1 + \sum_{s \in S} p_{is} \mathbb{E}(\text{number of further steps given first step is to } s) \\ &= 1 + \sum_{s \in S} p_{is} \mathbb{E}(T \mid X_0 = s) \\ &= 1 + \sum_{s \in S} p_{is} u_s \end{aligned}$$

For our example we get the equations:

$$\begin{aligned} u_1 &= 1 + \tfrac{1}{3}u_2 + \tfrac{1}{3}u_3, \\ u_2 &= 1 + \tfrac{2}{3}u_1 + \tfrac{1}{3}u_3, \\ u_3 &= 1 + \tfrac{1}{3}u_1 + \tfrac{2}{3}u_2. \end{aligned}$$

Solving these we get $\mathbb{E}(T \mid X_0 = 1) = u_1 = \frac{48}{7}$.

Example 11. Let $(X_0, X_1, X_2, \ldots)$ be the same Markov chain as in the previous example. What is the expected number of visits to state 2 the chain makes before it is absorbed? Let V be the random variable which counts these visits to state 2. So

$$V = |\{t : 0 \leqslant t \leqslant T, X_t = 2\}|$$

As before it helps to make the starting state a variable and define $v_i = \mathbb{E}(V \mid X_0 = i)$.

Clearly $v_4 = 0$. For $i \neq 4$ we have

$$v_i = \mathbb{E}(\text{the number of further visits to state 2 after the next step}) + \begin{cases} 1 & \text{if } i = 2 \\ 0 & \text{if } i \neq 2 \end{cases}$$

If we write $w(i) = \begin{cases} 1 & \text{if } i = 2 \\ 0 & \text{if } i \neq 2 \end{cases}$ then

$$\begin{aligned} v_i &= w(i) + \sum_{s \in S} p_{is} \mathbb{E}(\text{number of further visits to 2 given first step is to } s) \\ &= w(i) + \sum_{s \in S} p_{is} \mathbb{E}(V \mid X_0 = s) \\ &= w(i) + \sum_{s \in S} p_{is} v_s \end{aligned}$$

For our example we get the equations:

$$\begin{aligned} v_1 &= \tfrac{1}{3} v_2 + \tfrac{1}{3} v_3, \\ v_2 &= 1 + \tfrac{2}{3} v_1 + \tfrac{1}{3} v_3, \\ v_3 &= \tfrac{1}{3} v_1 + \tfrac{2}{3} v_2. \end{aligned}$$

Solving these we get $\mathbb{E}(V \mid X_0 = 1) = v_1 = \frac{15}{7}$.

Both of these examples are special cases of the following more general result.

Theorem 3.2. *Suppose $(X_0, X_1, X_2, \ldots)$ is a Markov chain with state space S, transition probabilities $(p_{ij} : i, j \in S)$, and absorbing states $A \subseteq S$, and suppose $w : S \setminus A \to \mathbb{R}$ is a weight function on the non-absorbing states. Let $T = \min\{n : X_n \in A\}$ denote the time of first absorption, and let $W = \sum_{t=0}^{T-1} w(X_t)$. Then the following facts hold.*

(i) For each $i \in S$, let $b_i = \mathbb{E}(W \mid X_0 = i)$. Then the $(b_i : i \in S)$ satisfy the following system of simultaneous linear equations:

$$b_i = \begin{cases} 0, & \text{if } i \in A; \\ w(i) + \sum_{j \in S \setminus A} p_{ij} b_j & \text{if } i \notin A. \end{cases}$$

(ii) If S is finite and from every state it is possible to reach an absorbing state in some number of steps, then absorption occurs with probability 1, and the system of equations in (i) has a unique solution.

We have seen all the ideas needed to prove part (i) of this in our examples.

Proof. (i) If $i \in A$ then $T = 0$ and $b_i = 0$. If $i \in S \setminus A$, then conditioning on the first step gives:

$$\begin{aligned} b_i &= w(i) + \sum_{j \in S} \mathbb{P}(X_1 = j \mid X_0 = i) \mathbb{E}\Big(\sum_{t=1}^{T-1} w(X_t) \mid X_1 = j \Big) \\ &\qquad \text{(using the Law of Total Expectation)} \\ &= w(i) + \sum_{j \in S} p_{ij} b_j \\ &= w(i) + \sum_{j \in S \setminus A} p_{ij} b_j. \end{aligned}$$

(ii) Omitted and non-examinable.

□

If we take $w(i) = 1$ for all $i \in S \setminus A$ in the above theorem, then $W = T$, the time of first absorption, so we get the following corollary.

Corollary 3.3. *Suppose $(X_0, X_1, X_2, \ldots)$ is a Markov chain with state space S, transition probabilities $(p_{ij} : i, j \in S)$, and absorbing states $A \subseteq S$. Let $T = \min\{n : X_n \in A\}$ denote the time of first absorption. Then the following facts hold.*

(i) Let $v_i = \mathbb{E}(T \mid X_0 = i)$. Then the v_i satisfy the following system of simultaneous linear equations:

$$v_i = \begin{cases} 0, & \text{if } i \in A; \\ 1 + \sum_{j \in S \setminus A} p_{ij} v_j & \text{if } i \notin A. \end{cases}$$

(ii) If S is finite and from every state it is possible to reach some absorbing state in some number of steps, then absorption occurs with probability 1, and the above system of equations has a unique solution.

As we saw, another natural application of Theorem 3.2 is finding the expected number of visits to state k. To do this we would apply the Theorem with weight function

$$w(i) = \begin{cases} 1 & \text{if } i = k \\ 0 & \text{if } i \neq k \end{cases}.$$

3.5 More Examples

We've done lots of examples already in this section and there are more on Problem Sheet 2. We'll look at a couple more examples where a bit of extra ingenuity is needed.

Example 12. Let $(X_0, X_1, X_2, \ldots)$ be the Markov chain on state space $S = \{1, 2, 3, 4\}$ with $X_0 = 1$ and transition matrix:

$$\begin{pmatrix} 1/4 & 1/4 & 1/4 & 1/4 \\ 1/3 & 1/3 & 1/6 & 1/6 \\ 1/6 & 1/2 & 1/6 & 1/6 \\ 1/3 & 1/6 & 1/3 & 1/6 \end{pmatrix}$$

What is the probability that the chain visits state 2 before state 3?

This Markov chain has no absorbing states so it looks as if the methods we have seen do not apply. However, there is a trick we can use.

Let $(Y_0, Y_1, Y_2, \ldots)$ be the Markov chain on state space $S = \{1, 2, 3, 4\}$ with $Y_0 = 1$ and transition matrix:

$$\begin{pmatrix} 1/4 & 1/4 & 1/4 & 1/4 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1/3 & 1/6 & 1/3 & 1/6 \end{pmatrix}$$

The only difference between these two Makov chains is that I have made states 2 and 3 into absorbing states (in the matrix this corresponds to replacing the second and third rows as above).

If we run both these processes they will follow exactly the same rules until they reach state 2 or state 3. From this you can see that the probability we reach state 2 before state 3 in the chain (X_i) is exactly the same as the probabilty we are absorbed at state 2 in the chain (Y_i). This later probability is something we can calculate using the usual first-step analysis method.

Definining $T = \min\{n : Y_n = 2, 3\}$ and setting $a_i = \mathbb{P}(Y_T = 2 \mid Y_0 = i)$ we get

$$\begin{aligned} a_1 &= \tfrac{1}{4}a_1 + \tfrac{1}{4} + \tfrac{1}{4}a_4, \\ a_4 &= \tfrac{1}{3}a_1 + \tfrac{1}{6} + \tfrac{1}{6}a_4. \end{aligned}$$

(of course $a_2 = 1$ and $a_3 = 0$).

Solving these (I leave the details to you) we get $a_1 = \frac{6}{13}$ so this is the probability that the chain (X_i) visits state 2 before state 3.

Example 13. A fair coin is repeatedly tossed until the sequence HTH is seen. By considering an appropriate Markov chain, calculate the expectation of the total number of tosses.

The most direct approach would be to define a Markov chain with $2^3 = 8$ states, one state for each possible sequence of three tosses $(HHH, HHT, \ldots, TTT)$.

However, there is a much more economical way of modelling it, as a Markov chain $(X_0, X_1, X_2, \ldots)$ with only 4 states:

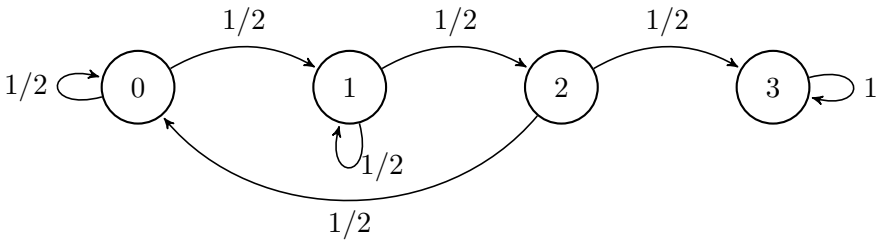
State 3: The last three tosses were HTH;

State 2: The last two tosses were HT;

State 1: The last toss was an H, and the last three tosses were not HTH;

State 0: None of the above (i.e., the last two tosses were TT, or there has only been one toss so far and it was a T, or there have been no tosses so far).

Notice that state i corresponds to the situation where 'we have just seen the first i tosses we need'. This Markov chain has transition graph as follows.



There is just one absorbing state, namely the state 3. Let T denote the time of first absorption. We can now use the usual first-step analysis method to calculate the expectation of the time of absorption starting from state 0. For each $i \in \{0, 1, 2, 3\}$, let $u_i = \mathbb{E}(T \mid X_0 = i)$. We want to calculate u_0.

The usual approach (Corollary 3.3) gives the following system of simultaneous linear equations:

$$\begin{aligned} u_0 &= 1 + \tfrac{1}{2}u_0 + \tfrac{1}{2}u_1, \\ u_1 &= 1 + \tfrac{1}{2}u_1 + \tfrac{1}{2}u_2, \\ u_2 &= 1 + \tfrac{1}{2}u_0, \\ u_3 &= 0. \end{aligned}$$

Solving these we get $u_0 = 10$.

4 Long-term behaviour

The previous section was concerned with Markov chains with absorbing states. The eventual behaviour of such a chain is to get stuck in an absorbing state[4].

If a Markov chain does not have absorbing states[5] then it will continually evolve and the long-term behaviour is more interesting. In this section we look at questions about this long-term behaviour. Linear algebra will turn out to be a powerful tool for this.

4.1 Multi-step transition probabilities

Lemma 2.1 gives us a crude way of calculating $\mathbb{P}(X_t = i)$, for any $t \in \mathbb{N}$ and any $i \in S$: we can simply sum over all the possible ways of getting to the state i at time t, giving

$$\begin{aligned}\mathbb{P}(X_t = i) &= \sum_{i_0, i_1, \ldots, i_{t-1} \in S} \mathbb{P}(X_0 = i_0, X_1 = i_1, \ldots, X_{t-1} = i_{t-1}, X_t = i) \\ &= \sum_{i_0, i_1, \ldots, i_{t-1} \in S} \mu_{i_0}^{(0)} p_{i_0, i_1} p_{i_1, i_2} \cdots p_{i_{t-2}, i_{t-1}} p_{i_{t-1}, i}.\end{aligned}$$

Soon we will see a much cleaner way of calculating $\mathbb{P}(X_t = i)$, using transition matrices.

To help our analysis of Markov chains using transition matrices, we need to introduce the following definition.

Definition 7. Let $(X_0, X_1, X_2, \ldots)$ be a (homogeneous) Markov chain with state space S. Then for any $r \in \mathbb{N}$, the *r-step transition probabilities* are defined by

$$p_{ij}^{(r)} = \mathbb{P}(X_r = j \mid X_0 = i) \quad \forall i, j \in S.$$

In other words, $p_{ij}^{(r)}$ is the probability that the chain transitions from state i to state j in exactly r steps. Note that, since the chain is homogeneous, we have

$$\mathbb{P}(X_{t+r} = j \mid X_t = i) = \mathbb{P}(X_r = j \mid X_0 = i) = p_{ij}^{(r)} \quad \forall t \in \mathbb{N},\ \forall i, j \in S.$$

Similarly to Lemma 2.1, we could calculate these by enumerating all possible r-step paths from state i to state j and summing the corresponding probabilities. This gives that

$$p_{ij}^{(r)} = \sum_{s_1, \ldots, s_{r-1} \in S} p_{i s_1} p_{s_1 s_2} \cdots p_{s_{r-2} s_{r-1}} p_{s_{r-1} j}.$$

Fortunately there is a better way.

[4] provided that S is finite and every state leads to some absorbing state

[5] or if S is infinite

Lemma 4.1 (Chapman-Kolmogorov relations)**.** *For all states $i, j \in S$ and all $r, s \in \mathbb{N}$*

$$p_{ij}^{(r+s)} = \sum_{k \in S} p_{ik}^{(r)} p_{kj}^{(s)}.$$

Proof.

$$\begin{aligned}
p_{ij}^{(r+s)} &= \mathbb{P}(X_{r+s} = j \mid X_0 = i) \\
&= \sum_{k \in S} \mathbb{P}(X_{r+s} = j, X_r = k \mid X_0 = i) \\
&\qquad\qquad \text{(since the events } X_r = k \text{ partition the probability space)} \\
&= \sum_{k \in S} \mathbb{P}(X_r = k \mid X_0 = i)\, \mathbb{P}(X_{r+s} = j \mid X_r = k, X_0 = i) \\
&\qquad\qquad \text{(by the definition of conditional probability)} \\
&= \sum_{k \in S} \mathbb{P}(X_r = k \mid X_0 = i)\, \mathbb{P}(X_{r+s} = j \mid X_r = k) \qquad \text{(by the Markov property)} \\
&= \sum_{k \in S} p_{ik}^{(r)} p_{kj}^{(s)}.
\end{aligned}$$

□

The expression on the right-hand side of Lemma 4.1 should remind you of the formula for matrix multiplication. Indeed, this lemma enables us to relate the r-step transition probabilities to the rth power of the transition matrix, in the following corollary.

Corollary 4.2. *Let $(X_0, X_1, X_2, \ldots)$ be a (homogeneous) Markov chain with finite state space S and transition probabilities $(p_{ij} : i, j \in S)$. Let P be the transition matrix of this Markov chain. Then $p_{ij}^{(r)} = (P^r)_{i,j}$ for all $i, j \in S$. In other words, the matrix P^r is the matrix of r-step transition probabilities. So if $S = \{1, 2, \ldots, n\}$, then*

$$P^r = \begin{pmatrix} p_{11}^{(r)} & p_{12}^{(r)} & \cdots & p_{1n}^{(r)} \\ p_{21}^{(r)} & p_{22}^{(r)} & \cdots & p_{2n}^{(r)} \\ \vdots & \vdots & & \vdots \\ p_{n1}^{(r)} & p_{n2}^{(r)} & \cdots & p_{nn}^{(r)} \end{pmatrix}.$$

Proof. The proof is by induction on r, with the base case $r = 1$ being immediate. For the induction step, let $r \geqslant 1$ and assume the inductive hypothesis that $p_{ij}^{(r)} = (P^r)_{i,j}$ for all $i, j \in S$. We need to prove that $p_{ij}^{(r+1)} = (P^{r+1})_{i,j}$ for all $i, j \in S$. Now for any $i, j \in S$, we

have

$$\begin{aligned} p_{ij}^{(r+1)} &= \sum_{k\in S} p_{ik}^{(r)} p_{kj} && \text{(by Lemma 4.1 with } s=1) \\ &= \sum_{k\in S} (P^r)_{i,k} P_{k,j} && \text{(by inductive hypothesis)} \\ &= (P^r P)_{i,j} && \text{(by the formula for matrix multiplication))} \\ &= (P^{r+1})_{i,j} \end{aligned}$$

This completes the inductive step, proving the corollary. □

This corollary gives us a (relatively) clean way of calculating the probability distribution of X_r, for all r.

Definition 8. Let $(X_0, X_1, \ldots)$ be a Markov chain with state space S. Its *distribution at time r* is the probability distribution of X_r, i.e. it is the row-vector $\boldsymbol{\mu}^{(r)}$ with entries indexed by S, and with $\mu_i^{(r)} = \mathbb{P}(X_r = i)$ for all $i \in S$. (In particular, as before, $\boldsymbol{\mu}^{(0)}$ is the initial distribution of the Markov chain.)

Corollary 4.3. *Let $(X_0, X_1, \ldots)$ be a (homogeneous) Markov chain with finite state space S, initial distribution $\boldsymbol{\mu}^{(0)}$ and transition matrix P. Then*

$$\boldsymbol{\mu}^{(r)} = \boldsymbol{\mu}^{(0)} P^r.$$

Proof. For any $i \in S$, we have

$$\begin{aligned} \mu_i^{(r)} &= \mathbb{P}(X_r = i) \\ &= \sum_{k\in S} \mathbb{P}(X_r = i, X_0 = k) \quad \text{(since the events } X_0 = k \text{ partition the probability space)} \\ &= \sum_{k\in S} \mathbb{P}(X_r = i \mid X_0 = k)\mathbb{P}(X_0 = k) \quad \text{(by the definition of conditional probabilities)} \\ &= \sum_{k\in S} p_{ki}^{(r)}\, \mu_k^{(0)} && \text{(by the definition of } r\text{-step transition probabilities)} \\ &= \sum_{k\in S} (P^r)_{k,i}\, \mu_k^{(0)} && \text{(by Corollary 4.2)} \\ &= (\boldsymbol{\mu}^{(0)} P^r)_i. \end{aligned}$$

□

Hence, if we can calculate P^r, and we know the initial distribution (the probability distribution of X_0), then we can calculate the probability distribution of X_t, so we know everything there is to know about the Markov chain at time r. Sometimes it is possible to calculate P^r fairly easily, but sometimes it is not, so we may need to fall back on other methods to analyse the Markov chain, as we will see later.

4.2 Computing P^r

One situation which is easy to deal with is the case that P is diagonalisable. In other words there is some invertible matrix M with $M^{-1}PM = D$ being a diagonal matrix. In this case we have

$$P^r = (MDM^{-1})(MDM^{-1})\dots(MDM^{-1}) = MD^rM^{-1}$$

which is easy to calculate because

$$\begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ 0 & 0 & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{pmatrix}^r = \begin{pmatrix} \lambda_1^r & 0 & \cdots & 0 \\ 0 & \lambda_2^r & \cdots & 0 \\ 0 & 0 & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n^r \end{pmatrix}$$

Here is a quick example.

Example 14 (Simple weather model)**.** We saw an example earlier of a two state chain with transition matrix:

$$P = \begin{pmatrix} 0.9 & 0.1 \\ 0.4 & 0.6 \end{pmatrix}$$

You can check that

$$P = \begin{pmatrix} 1 & 0.1 \\ 1 & -0.4 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 0.5 \end{pmatrix} \begin{pmatrix} 0.8 & 0.2 \\ 2 & -2 \end{pmatrix}$$

For details of where this comes from see your Linear Algebra I notes or later in these notes.

It follows that

$$P^r = \begin{pmatrix} 1 & 0.1 \\ 1 & -0.4 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & (0.5)^r \end{pmatrix} \begin{pmatrix} 0.8 & 0.2 \\ 2 & -2 \end{pmatrix}$$

So we have a relatively simple expression for the r-step transition probabilities. For instance, given that today is dry, the probability that it will be dry in 7 days time is given by the top left corner of the matrix P^7 which the above formula shows is $0.8 + 0.2 \times (0.5)^7 \approx 0.8016$

$$\mathbb{P}(X_7 = 1 \mid X_0 = 1) \approx 0.8016$$

This is much quicker than writing down all possible 7 step trajectories from state 1 to state 1 as in Lemma 2.1.

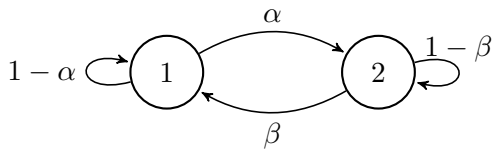
Question: Work out the bottom left corner entry of the matrix P^7. What does this number represent and what do you notice?

4.3 Long-term Behaviour Example: The General 2-state chain

The structure of a Markov chain with only two states is rather simple. Take $S = \{1, 2\}$. If $p_{12} = \alpha$ and $p_{21} = \beta$, for some $0 \leqslant \alpha, \beta \leqslant 1$, then the other transition probabilities are determined and the transition matrix must be

$$P = \begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix}$$

and the transition graph is



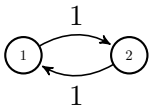
As usual, let $\boldsymbol{\mu}^{(0)}$ be the initial distribution and $\boldsymbol{\mu}^{(n)}$ be the distribution at time n. There are two boring special cases:

- If $\alpha = \beta = 0$ then $P = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ so $P^n = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ and $\lim_{n\to\infty} P^n = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$.

- $\alpha = \beta = 1$, $P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ so

$$P^n = \begin{cases} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, & \text{if } n \text{ is even;} \\ \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, & \text{if } n \text{ is odd.} \end{cases}$$

Clearly $\lim_{n\to\infty} P^n$ does not exist.



In both of these cases the behaviour is completely deterministic. We either stay where we start (first case) or alternate between the two states (second case) for ever.

The more important and interesting case is when neither of these occur so $0 < \alpha+\beta < 2$. The matrix $P = \begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix}$ has eigenvalues:

- 1 with eigenvector $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$
- $1-\alpha-\beta$ with eigenvector $\begin{pmatrix} \alpha \\ -\beta \end{pmatrix}$

So if we let $M = \begin{pmatrix} 1 & \alpha \\ 1 & -\beta \end{pmatrix}$ and D be the diagonal matrix with eigenvalues down the diagonal then $P = MDM^{-1}$. More explicitly, writing $\lambda = 1-\alpha-\beta$,

$$P = \begin{pmatrix} 1 & \alpha \\ 1 & -\beta \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & \lambda \end{pmatrix} \begin{pmatrix} \frac{\beta}{\alpha+\beta} & \frac{\alpha}{\alpha+\beta} \\ \frac{1}{\alpha+\beta} & \frac{-1}{\alpha+\beta} \end{pmatrix}.$$

Having written P like this it is easy to find P^n.

$$\begin{aligned} P^n &= \begin{pmatrix} 1 & \alpha \\ 1 & -\beta \end{pmatrix} \begin{pmatrix} 1^n & 0 \\ 0 & \lambda^n \end{pmatrix} \begin{pmatrix} \frac{\beta}{\alpha+\beta} & \frac{\alpha}{\alpha+\beta} \\ \frac{1}{\alpha+\beta} & \frac{-1}{\alpha+\beta} \end{pmatrix} \\ &= \frac{1}{\alpha+\beta} \begin{pmatrix} 1 & \alpha \\ 1 & -\beta \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & \lambda^n \end{pmatrix} \begin{pmatrix} \beta & \alpha \\ 1 & -1 \end{pmatrix} \\ &= \frac{1}{\alpha+\beta} \begin{pmatrix} 1 & \alpha \\ 1 & -\beta \end{pmatrix} \begin{pmatrix} \beta & \alpha \\ \lambda^n & -\lambda^n \end{pmatrix} \\ &= \frac{1}{\alpha+\beta} \begin{pmatrix} \beta & \alpha \\ \beta & \alpha \end{pmatrix} + \frac{\lambda^n}{\alpha+\beta} \begin{pmatrix} \alpha & -\alpha \\ -\beta & \beta \end{pmatrix} \end{aligned}$$

The second of these term is negligible for large n because $-1 < \lambda < 1$ (this is the only point in the argument where we use that we are not in the two boring cases we looked at first). So

$$\lim_{n\to\infty} P^n = \frac{1}{\alpha+\beta} \begin{pmatrix} \beta & \alpha \\ \beta & \alpha \end{pmatrix} = W$$

You will notice that the two rows of the matrix W are identical; each is equal to $\mathbf{w} = \begin{pmatrix} \frac{\beta}{\alpha+\beta} & \frac{\alpha}{\alpha+\beta} \end{pmatrix}$. This means that for any distribution $\boldsymbol{\mu}$ we have $\boldsymbol{\mu} W = \mathbf{w}$. This allows us to determine the behaviour of the chain for large n.

$$\lim_{n\to\infty} \boldsymbol{\mu}^{(n)} = \lim_{n\to\infty} \boldsymbol{\mu}^{(0)} P^n = \boldsymbol{\mu}^{(0)} W = \mathbf{w}$$

Whatever the initial distribution is, the chain tends towards the limiting distribution $\mathbf{w}$. Another way of saying this (thinking about which multi-step transition probabilities the entries of P^n represent) is:

$$\begin{aligned}
\mathbb{P}(X_n = 1 \mid X_0 = 1) &\to \frac{\beta}{\alpha+\beta} \quad \text{(top left corner of } P^n\text{)} \\
\mathbb{P}(X_n = 1 \mid X_0 = 2) &\to \frac{\beta}{\alpha+\beta} \quad \text{(bottom left corner of } P^n\text{)} \\
\mathbb{P}(X_n = 2 \mid X_0 = 1) &\to \frac{\alpha}{\alpha+\beta} \quad \text{(top right corner of } P^n\text{)} \\
\mathbb{P}(X_n = 2 \mid X_0 = 2) &\to \frac{\alpha}{\alpha+\beta} \quad \text{(bottom right corner of } P^n\text{)}
\end{aligned}$$

The vector $\mathbf{w}$ has another nice property; it is the unique probability vector which satisfies $\mathbf{w}P = \mathbf{w}$. We call such a vector a *equilibrium distribution*. If $\mathbf{w}$ is an equilibrium distribution and $\boldsymbol{\mu} = \mathbf{w}$ then $\boldsymbol{\mu}^{(n)} = \mathbf{w}$ for all n.

You can check that $\begin{pmatrix} \frac{\beta}{\alpha+\beta} & \frac{\alpha}{\alpha+\beta} \end{pmatrix}$ is also the unique equilibrium distribution in the special case $\alpha = \beta = 1$ (because $\begin{pmatrix} \frac{1}{2} & \frac{1}{2} \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \end{pmatrix}$). However when $\alpha = \beta = 0$ there are many solutions to $\mathbf{w}P = \mathbf{w}$.

That completes the story for long-term behaviour of the two-state chain. What happens if there are more states? When do limiting and equilibrium distributions exist and how do we find them?

In the next sections we will develop some of this theory.

4.4 Definition of limiting and equilibrium distributions

We have been writing distributions on S as row vectors. A convenient bit of terminology for this is:

Definition 9. A *probability vector* is a row-vector whose entries are non-negative real numbers summing to 1.

Remark. Note that a probability vector with n entries is equivalent to a probability distribution on a set of size n.

Let's define the two concepts we saw in the two-state chain example more formally.

Definition 10. Let $(X_0, X_1, X_2, \ldots)$ be a Markov chain with finite state space S. We say a probability vector $\mathbf{w}$ (with entries indexed by S) is a *limiting distribution* for the chain if for any $i, j \in S$, we have

$$\mathbb{P}(X_t = j \mid X_0 = i) \to w_j \quad \text{as } t \to \infty.$$

A key point is that the limiting value w_j does not depend on i. If a Markov chain has a limiting distribution then after it has been running for a long time, the probability that it is in particular state follows the distribution $\mathbf{w}$. This does not depend on X_0 so once the chain has been running for a long time it forgets its starting state.

Definition 11. Let $(X_0, X_1, X_2, \ldots)$ be a Markov chain with finite state space S and transition matrix P. We say a probability vector $\mathbf{w}$ (with entries indexed by S) is an *equilibrium distribution*[6]. for the chain if $\mathbf{w}P = \mathbf{w}$.

If a Markov chain starts in at equilibrium distribution, then it has that distribution forever.

We will see that these two concepts are related. We will also develop some theory which gives conditions for the existence of limiting and equilibrium distributions.

Example 15 (Two-state chain)**.** Let's summarise what we found out for the two-state chain above with transition matrix

$$\begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix}$$

We saw that:

- If $0 < \alpha + \beta < 2$ then P^n tends to a limit W. The rows of the limit are both equal to the limiting distribution $\begin{pmatrix} \frac{\beta}{\alpha+\beta} & \frac{\alpha}{\alpha+\beta} \end{pmatrix}$. This is also the unique equilibrium distribution.
- If $\alpha = \beta = 1$ then P^n does not tend to a limit. There is no limiting distribtion but $\begin{pmatrix} \frac{1}{2} & \frac{1}{2} \end{pmatrix}$ is the unique equilibrium distribution.
- If $\alpha = \beta = 0$ then $P^n = P$ for all n. So P^n does tend to a limit but it does not have identical rows. There is no limiting distribution. Any probability vector $\begin{pmatrix} x & 1-x \end{pmatrix}$ is an equilibrium distribution so there are many equilibrium distributions.

The limiting distribution can be tricky to calculate directly (for the two-state chain we needed to mess around with eigenvalues). However, the equilibrium distribution can be found just by solving a matrix equation.

Example 16. Let $(X_0, X_1, X_2, \ldots)$ be the Markov chain with state space $S = \{1, 2, 3, 4\}$ and transition matrix:

$$P = \begin{pmatrix} 1/8 & 0 & 1/2 & 3/8 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1/2 & 1/2 \end{pmatrix}$$

[6] Also called an *invariant distribution* or a *stationary distribution*

Does this Markov chain have an equilibrium distribution?

We need to solve $\mathbf{w}P = \mathbf{w}$ which is equivalent to

$$\begin{aligned} w_1 &= \tfrac{1}{8}w_1 + w_2 \\ w_2 &= w_3 \\ w_3 &= \tfrac{1}{2}w_1 + \tfrac{1}{2}w_4 \\ w_4 &= \tfrac{3}{8}w_1 + \tfrac{1}{2}w_4 \end{aligned}$$

If you solve these you get:

$$\mathbf{w} = \begin{pmatrix} w_1 & \frac{7}{8}w_1 & \frac{7}{8}w_1 & \frac{3}{4}w_1 \end{pmatrix}$$

We know that $\mathbf{w}$ must be a probability vector, so we have one more equation

$$w_1 + w_2 + w_3 + w_4 = 1.$$

This gives $w_1 = \frac{2}{7}$. So this chain has a unique equilibrium distribution:

$$\mathbf{w} = \begin{pmatrix} \frac{2}{7} & \frac{1}{4} & \frac{1}{4} & \frac{3}{14} \end{pmatrix}$$

4.5 More about limiting distributions

Our definition of limiting distribution is equivalent to two other statements. We have seen both of these in the examples above. Here is a more formal statement and proof.

Lemma 4.4. *Let $(X_0, X_1, X_2, \ldots)$ be a Markov chain with state space $S = \{1, 2, \ldots, n\}$ and transition matrix P. Let*

$$\mathbf{w} = (w_1, w_2, \ldots, w_n)$$

be a probability vector.

The following are equivalent:

(i) $\mathbf{w}$ *is a limiting distribution for the Markov chain.*

(ii)

$$P^r \to \begin{pmatrix} w_1 & w_2 & \ldots & w_n \\ w_1 & w_2 & \ldots & w_n \\ \vdots & \vdots & & \vdots \\ w_1 & w_2 & \ldots & w_n \end{pmatrix} \quad \text{as } r \to \infty.$$

(iii) For any initial distribution $\boldsymbol{\mu}^{(0)}$, the distributions $\boldsymbol{\mu}^{(r)}$ satisfy

$$\boldsymbol{\mu}^{(r)} \to \mathbf{w} \text{ as } r \to \infty.$$

Proof. Suppose that $\mathbf{w} = \begin{pmatrix} w_1 & \dots & w_n \end{pmatrix}$ is a probability vector and write

$$W = \begin{pmatrix} w_1 & w_2 & \dots & w_n \\ w_1 & w_2 & \dots & w_n \\ \vdots & \vdots & & \vdots \\ w_1 & w_2 & \dots & w_n \end{pmatrix}.$$

$(i) \implies (ii)$ Suppose that $\mathbf{w}$ is a limiting distribution. By definition

$$p_{ij}^{(r)} \to w_j \quad \text{as } r \to \infty.$$

But the left hand side of this is the (i, j) entry of the matrix P^r. So

$$P^r \to \begin{pmatrix} w_1 & w_2 & \dots & w_n \\ w_1 & w_2 & \dots & w_n \\ \vdots & \vdots & & \vdots \\ w_1 & w_2 & \dots & w_n \end{pmatrix} \quad \text{as } r \to \infty.$$

as required.

$(ii) \implies (iii)$ Suppose that $P^r \to W$ as above and $\boldsymbol{\mu}^{(0)}$ is any probability vector. Then

$$\boldsymbol{\mu}^{(r)} = \boldsymbol{\mu}^{(0)} P^r \to \boldsymbol{\mu}^{(0)} \begin{pmatrix} w_1 & w_2 & \dots & w_n \\ w_1 & w_2 & \dots & w_n \\ \vdots & \vdots & & \vdots \\ w_1 & w_2 & \dots & w_n \end{pmatrix} = \mathbf{w} \quad \text{as } r \to \infty.$$

(The limit follows from the fact that $P^r \to W$ and continuity of matrix multiplication[7].)

$(iii) \implies (i)$ Suppose that $\boldsymbol{\mu}^{(r)} \to \mathbf{w}$ for any $\boldsymbol{\mu}^{(0)}$. Then, in particular, this holds when $\boldsymbol{\mu}^{(0)} = \mathbf{e}_i$ the ith standard basis (row) vector[8]. So

$$\mathbf{e}_i P^r \to \mathbf{w}$$

But the left hand side of this is just $\begin{pmatrix} p_{i1}^{(r)} & p_{i2}^{(r)} & \dots & p_{in}^{(r)} \end{pmatrix}$ so

$$\begin{pmatrix} p_{i1}^{(r)} & p_{i2}^{(r)} & \dots & p_{in}^{(r)} \end{pmatrix} \to \mathbf{w}$$

So $p_{ij}^{(r)} \to w_j$ for all i, j. That is $\mathbf{w}$ is a limiting distribution.

□

[7]This means that if we have a sequence of matrices $A_1, A_2, \cdots$ with $A_n \to M$ as $n \to \infty$ then $\mathbf{x}A_n \to \mathbf{x}M$ as $n \to \infty$.

[8]That is $\mathbf{e}_i$ is the row vector has a 1 in position i and 0 is every other position.

4.6 More about equilibrium distributions

We noted earlier that if we start in an equilibrium distribution then we stay following that distribution for ever. Here is a lemma that formalises this.

Lemma 4.5. *Let $(X_0, X_1, X_2, \ldots)$ be a Markov chain with finite state space S and transition matrix P. Suppose that $\mathbf{w}$ is an equilibrium distribution for the Markov chain. If $\boldsymbol{\mu}^{(0)} = \mathbf{w}$, then $\boldsymbol{\mu}^{(r)} = \mathbf{w}$ for all $r \in \mathbb{N}$.*

Proof. We use induction on r. For $r = 0$, the statement holds by assumption. For the inductive step, let $r \geqslant 0$ and suppose that $\boldsymbol{\mu}^{(r)} = \mathbf{w}$. Then

$$\begin{aligned} \boldsymbol{\mu}^{(r+1)} &= \boldsymbol{\mu}^{(r)} P \\ &= \mathbf{w}P && \text{(by the induction hypothesis)} \\ &= \mathbf{w}. && \text{(since } \mathbf{w} \text{ is an equilibrium distribution)} \end{aligned}$$

□

You will also have probably noticed that in all the examples so far which have a limiting distribution, that limiting distribution is also the unique equilibrium ditribution. This is true in general.

Theorem 4.6. *Let $(X_0, X_1, X_2, \ldots)$ be a Markov chain with finite state space S and transition matrix P. Suppose $\mathbf{w}$ is a limiting distribution for the Markov chain. Then $\mathbf{w}$ is the unique equilibrium distribution.*

Proof. Suppose that $\mathbf{w}$ is a limiting distribution for the Markov chain. Then (by part (iii) of Lemma 4.4) for any initial distribution $\boldsymbol{\mu}^{(0)}$ we have $\mathbf{w} = \lim_{r\to\infty} \boldsymbol{\mu}^{(0)} P^r$.

So

$$\begin{aligned} \mathbf{w}P &= \left(\lim_{r\to\infty} \boldsymbol{\mu}^{(0)} P^r \right) P \\ &= \lim_{r\to\infty} \boldsymbol{\mu}^{(0)} P^{r+1} && \text{(by continuity of matrix multiplication)} \\ &= \mathbf{w} && \text{(by Lemma 4.4 again)} \end{aligned}$$

So $\mathbf{w}$ is an equilibrium distribution as required.

For the uniqueness, suppose that $\mathbf{u}$ is another equilibrium distribution. Then $\mathbf{u}P = \mathbf{u}$ so $\mathbf{u}P^r = \mathbf{u}$ for all r. This means that

$$\mathbf{u}P^r \to \mathbf{u} \quad \text{as } r \to \infty.$$

But by our assumption that the chain has a limiting distribution $\mathbf{w}$ we know that $\boldsymbol{\mu}^{(0)} P^r \to \mathbf{w}$ for all $\boldsymbol{\mu}^{(0)}$. In particular,

$$\mathbf{u}P^r \to \mathbf{w} \quad \text{as } r \to \infty.$$

So $\mathbf{u} = \mathbf{w}$. □

This is an extremely useful result. It means that once we know that a limiting distribiution exists, we can find it by solving $\mathbf{w}P = \mathbf{w}$. This is generally a much easier task than working out the limiting distribution directly.

4.7 Existence of limiting and equilibrium distributions

Definition 12. Let $(X_0, X_1, \ldots)$ be a Markov chain with state space S (finite or countable) and transition probabilities $(p_{i,j} : \ i, j \in S)$. The chain is:

- *irreducible* if for any $i, j \in S$, there exists $r \in \mathbb{N}$ such that $p_{i,j}^{(r)} > 0$.
- *regular* if there exists $r \in \mathbb{N}$ such that $p_{i,j}^{(r)} > 0$ for all $i, j \in S$.

A Markov chain being irreducible means that for any two states i and j, we can get from i to j in some number r of steps (typically r will depend on the pair (i, j)).

A Markov chain being regular means that it is possible to go between any two states in exactly r steps. Equivalently, the matrix P^r has no 0 entries. There is no single best recipe to follow to check whether a chain is regular. We will see a few different ways of doing this soon.

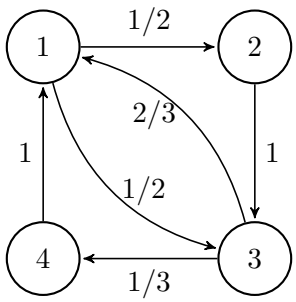
Clearly if a chain is regular then it is irreducible, but a chain may be irreducible but not regular.

Example 17 (Some Two-state examples)**.**

- $P = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$. Not irreducible (and hence not regular). There is no way to get from state 1 to state 2 so $p_{12}^{(r)} = 0$ for all r.
- $P = \begin{pmatrix} 1 & 0 \\ 1/2 & 1/2 \end{pmatrix}$. Not irreducible (and hence not regular). Again there is no way to get from state 1 to state 2 so $p_{12}^{(r)} = 0$ for all r.
- $P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$. Irreducible but not regular. There is a way to get between any two states ($p_{12}^{(1)} = p_{21}^{(1)} \neq 0$, $p_{11}^{(2)} = p_{22}^{(2)} \neq 0$). But there is no single r for which P^r has no 0 entries.
- $P = \begin{pmatrix} 1/3 & 2/3 \\ 3/4 & 1/4 \end{pmatrix}$. Irreducible and regular. The matrix $P^1 = P$ has no 0 entries. It is possible to get between any pair of states in exactly 1 step.

One way of deciding whether a chain is regular is to look at powers of P. We only need to keep track of which entries are non-zero.

Example 18. Let $(X_0, X_1, \ldots)$ be the Markov chain with transition graph as follows.



This chain is certainly irreducible; you can find a path from state i to state j for any $i, j \in \{1, 2, 3, 4\}$.

Is it regular? If you play with this example for a while you will see that you cannot get between any two states in 3 steps (for instance $p_{24}^{(3)} = 0$). However you can get between any two states in 4 steps. So the chain is regular.

Here is a more systematic way of seeing this. We need to show that P^4 has no 0 entries. To do this we only need keep track of which entries are positive and which are 0. So let's write P as

$$\begin{pmatrix} 0 & + & + & 0 \\ 0 & 0 & + & 0 \\ + & 0 & 0 & + \\ + & 0 & 0 & 0 \end{pmatrix}$$

Now P^2 has the form

$$\begin{pmatrix} 0 & + & + & 0 \\ 0 & 0 & + & 0 \\ + & 0 & 0 & + \\ + & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 0 & + & + & 0 \\ 0 & 0 & + & 0 \\ + & 0 & 0 & + \\ + & 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} + & 0 & + & + \\ + & 0 & 0 & + \\ + & + & + & 0 \\ 0 & + & + & 0 \end{pmatrix}$$

Continuing $P^4 = (P^2)^2$ has the form

$$\begin{pmatrix} + & 0 & + & + \\ + & 0 & 0 & + \\ + & + & + & 0 \\ 0 & + & + & 0 \end{pmatrix} \begin{pmatrix} + & 0 & + & + \\ + & 0 & 0 & + \\ + & + & + & 0 \\ 0 & + & + & 0 \end{pmatrix} = \begin{pmatrix} + & + & + & + \\ + & + & + & + \\ + & + & + & + \\ + & + & + & + \end{pmatrix}$$

as required. Of course you could calculate P^4 exactly by squaring P twice but if all we want to know is whether the chain is regular, this method is enough.

One of the key features of regularity is that it gives a sufficient condition for the existence of a limiting distribution.

Theorem 4.7. *Let $(X_0, X_1, \ldots)$ be a regular Markov chain with finite state space S and transition probabilities $(p_{i,j} : i, j \in S)$. Then the Markov chain has a limiting distribution.*

Non-examinable. Sketch of idea of proof. Since P is a stochastic matrix, it satisfies:

$$P \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$$

So 1 is an eigenvalue with eigenvector $\begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$.

It can be shown (proof omitted) that the other eigenvalues $\lambda_2, \lambda_3, \ldots, \lambda_n$ (not necessarily distinct) of P all satisfy $-1 < \lambda_i < 1$.

Suppose that P is diagonalisable then $P = MDM^{-1}$ where the diagonal entries of D are the eigenvalues and the columns of M are the corresponding eigenvectors.

$$P = \underbrace{\begin{pmatrix} 1 & * & \cdots & * \\ 1 & \vdots & & \vdots \\ 1 & * & \cdots & * \end{pmatrix}}_{M} \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ 0 & 0 & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{pmatrix} M^{-1}$$

This allows us to write P^r as

$$P^r = \begin{pmatrix} 1 & * & \cdots & * \\ 1 & \vdots & & \vdots \\ 1 & * & \cdots & * \end{pmatrix} \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & \lambda_2^r & \cdots & 0 \\ 0 & 0 & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n^r \end{pmatrix} M^{-1}$$

Now since each of $\lambda_2, \cdots, \lambda_n$ satisfies $|\lambda_i| < 1$, these tend to 0 in the limit:

$$\lim_{r\to\infty} P^r = \begin{pmatrix} 1 & * & \cdots & * \\ 1 & \vdots & & \vdots \\ 1 & * & \cdots & * \end{pmatrix} \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ 0 & 0 & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix} M^{-1}$$

$$= \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 1 & \vdots & & \vdots \\ 1 & 0 & \cdots & 0 \end{pmatrix} \begin{pmatrix} w_1 & w_2 & \cdots & w_n \\ * & * & \cdots & * \\ \vdots & & & \vdots \\ * & * & \cdots & * \end{pmatrix} \quad \text{(where } \mathbf{w} \text{ is the first row of } M^{-1}\text{)}$$

$$= \begin{pmatrix} w_1 & w_2 & \cdots & w_n \\ w_1 & w_2 & \cdots & w_n \\ \vdots & & & \vdots \\ w_1 & w_2 & \cdots & w_n \end{pmatrix}$$

So there is a limiting distribution $\mathbf{w}$.

If the matrix is not diagonalisable then a similar argument (omitted) works using an upper triangular matrix of a suitable form. □

Remark. This Theorem gives a procedure for finding the limiting distribution for a Markov chain.

- First check that the chain is regular
- If it is, calculate the equilibrium distribution by solving $\mathbf{w}P = \mathbf{w}$.
- By Theorem 4.6, this equilibrium distribution is also a limiting distribution.

Unfortunately, if at the first step you discover the chain is not regular then you can't deduce anything. The chain may have a limiting distribution or it may not. We will see examples of both of these cases later.

The condition of being irreducibile also has a nice consequence.

Theorem 4.8. *Let $(X_0, X_1, \ldots)$ be a Markov chain with finite state space S and transition probabilities $(p_{i,j} : i, j \in S)$. If it is irreducible, then it has a unique equilibrium distribution.*

Proof. We will deduce this from Theorem 4.7.

Let P be the transition matrix of $(X_0, X_1, \ldots)$.

Let $(Y_0, Y_1, \ldots)$ be the Markov chain with state space S and transition matrix

$$Q = \tfrac{1}{2}(P + I),$$

where I is the identity matrix. Then the Markov chain $(Y_0, Y_1, \ldots)$ is irreducible (since $(X_0, X_1, \ldots)$ is, and $p_{i,j} > 0 \Rightarrow q_{i,j} > 0$), and moreover it has a loop at every state. By Question 3 on Problem Sheet 4, an irreducible Markov chain with finite state space and a loop at some state, is regular. It follows that $(Y_0, Y_1, \ldots)$ is regular. Hence, by Theorem 4.7, it has a limiting distribution $\mathbf{w}$. By 4.6 this $\mathbf{w}$ is also its unique equilibrium distribution.

Now observe that a probability vector $\mathbf{u}$ is an equilibrium distribution for $(X_0, X_1, \ldots)$ if and only if it is an equilbrium distribution for $(Y_0, Y_1, \ldots)$. Indeed, suppose $\mathbf{u}$ is an equilibrium distribution for $(X_0, X_1, \ldots)$ so $\mathbf{u}P = \mathbf{u}$. Now,

$$\mathbf{u}Q = \mathbf{u}(\tfrac{1}{2}(P + I)) = \tfrac{1}{2}\mathbf{u}P + \tfrac{1}{2}\mathbf{u}I = \tfrac{1}{2}\mathbf{u} + \tfrac{1}{2}\mathbf{u} = \mathbf{u},$$

so $\mathbf{u}$ is an equilibrium distribution for $(Y_0, Y_1, \ldots)$.

Conversely, suppose $\mathbf{u}$ is an equilibrium distribution for $(Y_0, Y_1, \ldots)$; then $\mathbf{u}Q = \mathbf{u}$, so $\mathbf{u}(\frac{1}{2}(P + I)) = \mathbf{u}$, so $\frac{1}{2}\mathbf{u}P + \frac{1}{2}\mathbf{u}I = \mathbf{u}$, so $\mathbf{u}P = \mathbf{u}$, so $\mathbf{u}$ is an equilibrium distribution for $(X_0, X_1, \ldots)$.

It follows that $\mathbf{w}$ is the unique equilibrium distribution for $(X_0, X_1, \ldots)$. □

4.8 Time spent in each state

We have seen that the limiting distribution (if it exists) tells us how the chain behaves in the long run in the sense that the probability it is in any particular state (that is $\mathbb{P}(X_r = i)$) tends to a limit. Here is a different way of describing the long run behaviour of a Markov chain which tells us something whenever the chain has a unique equilibrium distribution even if does not have a limiting distribution.

Suppose we are interested in the amount of time that a chain spends in a particular state. If we fix $k \in S$ and write $V_k(t)$ for the number of visits to state k in the first t time steps. That is

$$V_k(t) = |\{i \in \{0, 1, \ldots, t-1\} : X_i = k\}|$$

(this is t time steps because there are t elements in $\{0, 1, \ldots, t-1\}$). The proportion of those first t steps that we spend in state k is

$$\frac{V_k(t)}{t}$$

This quantity is a random variable so we can consider its expectation. If the expectation of this quantity tends to a limit as $t \to \infty$ then we can say after the chain has been running for a long time, the expectation of the proportion of time spent in each state tends to a limit. It turns out that this happens if the chain has a unique equilibirum distribution and in this case the limit is given by the equilibrium distribution. In particular, the theorem below applies to any irreducible chain (with finite S) and any chain with a limiting distribution.

Theorem 4.9. *Let $(X_0, X_1, \ldots)$ be a Markov chain with finite state space S and transition matrix P. Supose that* $\mathbf{w}$ *is a unique equilibrium distribution. Then for any $k \in S$,*

$$\mathbb{E}\left(\frac{V_k(t)}{t}\right) \to w_k \quad \text{as } t \to \infty.$$

Proof. Omitted and non-examinable. □

4.9 Summary and examples

Let $(X_0, X_1, X_2, \ldots)$ be a Markov chain on finite state space S with transition matrix P.

- If the chain is irreducible and regular there is a unique equilibrium distribution which is also the limiting distribution. That is
  - There is a unique solution to $\mathbf{w}P = \mathbf{w}$
  - $P^n \to \begin{pmatrix} \mathbf{w} \\ \mathbf{w} \\ \vdots \\ \mathbf{w} \end{pmatrix}$ (the $n \times n$ matrix in which each row is $\mathbf{w}$).
  - The expected proportion of time the chain spends in state k tends to w_k.
- If the chain is irreducible but not regular then it has a unique equilibrium distribution but may not have a limiting distribution.
  - There is a unique solution to $\mathbf{w}P = \mathbf{w}$
  - The expected proportion of time the chain spends in state k tends to w_k.
  - The limit $\lim_{n\to\infty} P^n$ may not exist.
- If the chain is not irreducible then we can't deduce anything.

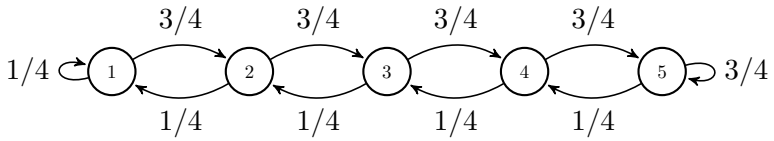
If I have a regular Markov chain, then it is usually not too difficult to:

- check that a chain is regular (proving that the limiting distribution exists),
- solve $\mathbf{w}P = \mathbf{w}$ (to find the limiting distribution).

We will see something of what happens with infinite S and the non-irreducible case later.

Let's see a range of examples illustrating this.

Example 19. Let $(X_0, X_1, X_2, \ldots)$ be the Markov chain with state space $S = \{1, 2, 3, 4, 5\}$ and transition graph as follows.



This Markov chain is clearly irreducible.

It is also regular. There are several ways to see this. Here is one (why not see if you can find any other proofs of regularity).

- For any i there is a path from i to 1 in at most 4 steps.
- For any j there is a path from 1 to j in at most 4 steps.
- Putting these together we get a path from i to j via 1 in at most 8 steps.
- But there is a loop at 1 so we can add steps to turn this into a path from i to j in exactly 8 steps.
- We conclude that the chain is regular (in the definition $r = 8$ will do).

Now,

- Since it is regular it has a limiting distribution (by Theorem 4.7).
- The limiting distribution is the unique equilibrium distributiom (by Theorem 4.6).

So to find the limiting distribution we find a probability vector which satisfies:

$$\mathbf{w} \begin{pmatrix} 1/4 & 3/4 & 0 & 0 & 0 \\ 1/4 & 0 & 3/4 & 0 & 0 \\ 0 & 1/4 & 0 & 3/4 & 0 \\ 0 & 0 & 1/4 & 0 & 3/4 \\ 0 & 0 & 0 & 1/4 & 3/4 \end{pmatrix} = \mathbf{w}$$

Which is equivalent to

$$\begin{aligned} w_1 &= \tfrac{1}{4}w_1 + \tfrac{1}{4}w_2 \\ w_2 &= \tfrac{3}{4}w_1 + \tfrac{1}{4}w_3 \\ w_3 &= \tfrac{3}{4}w_2 + \tfrac{1}{4}w_4 \\ w_4 &= \tfrac{3}{4}w_3 + \tfrac{1}{4}w_5 \\ w_5 &= \tfrac{3}{4}w_4 + \tfrac{3}{4}w_5 \end{aligned}$$

We get

$$\mathbf{w} = \begin{pmatrix} w_1 & 3w_1 & 9w_1 & 27w_1 & 81w_1 \end{pmatrix}.$$

But $\mathbf{w}$ is a probability vector so $1 = w_1 + w_2 + w_3 + w_4 + w_5 = 121w_1$.

We finally get that the limiting distribution is:

$$\mathbf{w} = \begin{pmatrix} \frac{1}{121} & \frac{3}{121} & \frac{9}{121} & \frac{27}{121} & \frac{81}{121} \end{pmatrix}.$$

Example 20. Let $(X_0, X_1, X_2, \ldots)$ be the Markov chain with state space $\{1, 2, 3, 4\}$ and transition matrix

$$P = \begin{pmatrix} 0 & 1/3 & 0 & 2/3 \\ 1/2 & 0 & 1/2 & 0 \\ 0 & 1/3 & 0 & 2/3 \\ 1/2 & 0 & 1/2 & 0 \end{pmatrix}$$

The chain is plainly irreducible. However, at every step the parity of the state changes (from odd to even or even to odd). This means that

- If r is odd then (for instance) $p_{11}^{(r)} = 0$
- If r is even then (for instance) $p_{12}^{(r)} = 0$

So the chain is not regular (there is no r for which P^r has no zero entries).

For a similar reason the chain has no limiting distribution.

Since it is irreducible it must have a unique equilibrium distribution. You can find this by solving:

$$\mathbf{w} \begin{pmatrix} 0 & 1/3 & 0 & 2/3 \\ 1/2 & 0 & 1/2 & 0 \\ 0 & 1/3 & 0 & 2/3 \\ 1/2 & 0 & 1/2 & 0 \end{pmatrix} = \mathbf{w}$$

You get

$$\mathbf{w} = \begin{pmatrix} \frac{1}{4} & \frac{1}{6} & \frac{1}{4} & \frac{1}{3} \end{pmatrix}$$

Since there is a unique equilibrium distribution, we know that the expected proportion of time in state k tends to w_k (by Theorem 4.9). So after the chain has been running for a long time the expectation of the proportion of time spent in state 1 will be close to $\frac{1}{4}$.

This example shows that a chain can be irreducible but not regular. We have seen an example of this already (the two-state chain with $\alpha = \beta = 1$). However, if you found that example to be a bit degenerate (it contains no genuine probability) then this one may be more convincing.

Example 21. Let $(X_0, X_1, X_2, \ldots)$ be the Markov chain with state space $\{1, 2, 3, 4\}$ and transition matrix

$$P = \begin{pmatrix} 2/3 & 1/3 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 \\ 0 & 0 & 2/3 & 1/3 \\ 0 & 0 & 1/2 & 1/2 \end{pmatrix}.$$

This is not irreducible (there is no path from states 1 or 2 to states 3 or 4). This means that:

$$p_{1,3}^{(r)} = p_{1,4}^{(r)} = p_{2,3}^{(r)} = p_{2,4}^{(r)} = 0$$

for any $r \in \mathbb{N}$. This observation (or thinking about matrix multiplication) shows that P^n must have the form:

$$P^n = \begin{pmatrix} ? & ? & 0 & 0 \\ ? & ? & 0 & 0 \\ 0 & 0 & ? & ? \\ 0 & 0 & ? & ? \end{pmatrix}$$

So there is no way that P^n can tend to a limit with all rows being equal. So the chain has no limiting distribution.

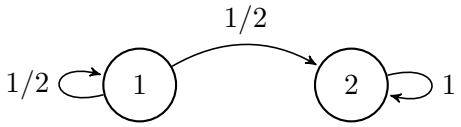
The vectors

$$\mathbf{w} = \begin{pmatrix} \frac{3}{5} & \frac{2}{5} & 0 & 0 \end{pmatrix} \quad \text{and} \quad \mathbf{w}' = \begin{pmatrix} 0 & 0 & \frac{3}{5} & \frac{2}{5} \end{pmatrix}$$

are both equilibrium distributions so there is no unique equilibrium distribution. Can you see where these equiibrium distributions came from and can you find some more?

Rather like the previous example, this behaves a bit like one of the degenerate cases of the two-state chain (in this case the $\alpha = \beta = 0$ one). If you objected to that example as not having any genuine probability in it then you will like this one more.

Example 22. Let $(X_0, X_1, \ldots)$ be the Markov chain with state space $S = \{1, 2\}$ and transition graph



The transition matrix is

$$P = \begin{pmatrix} 1/2 & 1/2 \\ 0 & 1 \end{pmatrix}.$$

This chain is clearly not irreducible, as 2 an absorbing state (so there is no way of getting from state 2 to state 1). So Theorem 4.7 and Theorem 4.6 can not be applied to it.

However, it does have a unique equilibrium distribution, namely $\begin{pmatrix} 0 & 1 \end{pmatrix}$. To see this, check that it is the unique solution to $\mathbf{w}P = \mathbf{w}$.

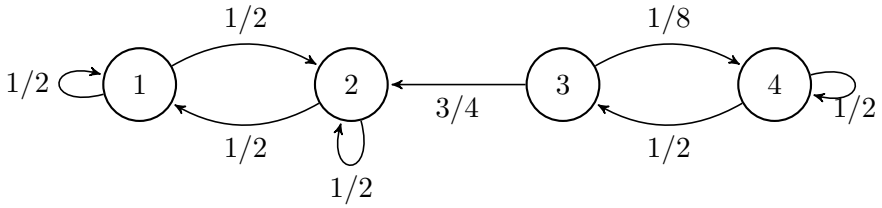
Moreover, if I start the chain from state 1, the probability it has not been absorbed after r steps is $\left(\frac{1}{2}\right)^r$). It follows that

$$\mathbb{P}(X_r = 2 \mid X_0 = i) = 1 - \left(\tfrac{1}{2}\right)^r \to 1 \quad \text{as } r \to \infty$$

So $\begin{pmatrix} 0 & 1 \end{pmatrix}$ is also a limiting distribution.

This example shows that a Markov chain may have a unique equilibrium distribution and a limiting distribution even if it is not irreducible. Theorem 4.7 and Theorem 4.6 give sufficient but not necessary conditions. In Example 22 this is because there is an absorbing state but that is not the only reason for this behaviour. Here is an example of a Markov chain which is not irreducible and has no absorbing states but still has a unique equilibrium distribution and a limiting distribution.

Example 23. Let $(X_0, X_1, \ldots)$ be the Markov chain with state space $S = \{1, 2, 3, 4\}$ and transition graph

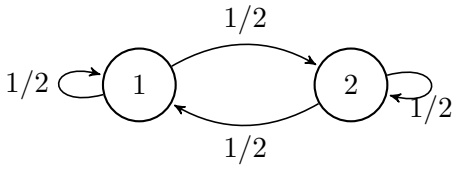


$$\begin{pmatrix} 1/2 & 1/2 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 \\ 0 & 3/4 & 1/8 & 1/8 \\ 0 & 0 & 1/2 & 1/2 \end{pmatrix}$$

This chain is clearly not irreducible as there is no way of getting from state 1 to state 4 (for instance). So Theorem 4.7 and Theorem 4.6 can not be applied to it.

However, if you solve $\mathbf{w}P = \mathbf{w}$, you will find that $\mathbf{w} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \end{pmatrix}$ is the unique solution so we have a unique equilibrium distribution.

Think about how this chain behaves starting from state 4. Eventually, we will follow the transition from state 3 to state 2 (there is probability 0 that we avoid this step for ever). From that point onwards we can never return to states 3 and 4 and the chain behaves exactly like the chain with transition graph:



This chain has a limiting distribution $\begin{pmatrix} \frac{1}{2} & \frac{1}{2} \end{pmatrix}$ and so $\mathbf{w} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \end{pmatrix}$ is a limiting distribution for the original chain.

4.10 Equilibrium and limiting distributions for infinite S

When the state space is infinite things get a bit more complicated. The definitions of irreducible, regular, and limiting distribution all work for infinite S. The definition of equilibrium distribution needs to be re-written but contains the same idea.

Definition 13. Let $(X_0, X_1, \ldots)$ be a Markov chain on state space $\mathbb{N}$. We say that $(w_i)_{i\geqslant 0}$ is an *equilibrium distribution* if

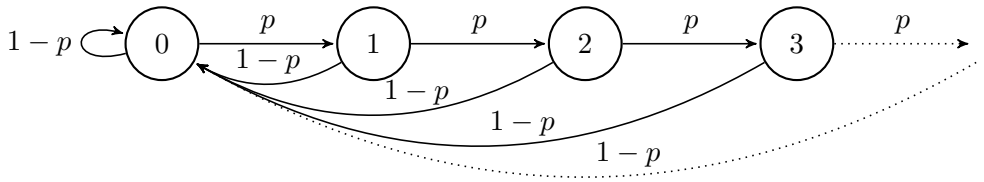
- $w_i \geqslant 0$ for all $i \in \mathbb{N}$ and $\sum_{i=0}^{\infty} w_i = 1$
- $w_k = \sum_{i=0}^{\infty} w_i p_{ik}$ for all $k \in \mathbb{N}$

The first condition is analogous to $\mathbf{w}$ being a probability vector, the second is analogous to the matrix equation $\mathbf{w}P = \mathbf{w}$.

Let's look at two examples showing different behaviour.

Example 24 (The success runs chain)**.** Suppose that we toss a coin which has probability p of showing Heads and let X_i be the length of the longest current run of Heads (that is the number of consecutive Heads we have seen up to and including the ith coin toss). In other words, $X_i = k$ means that after i tosses the last $k+1$ tosses were $T\underbrace{H\cdots H}_{k}$.

The transition graph is:



This chain is irreducible (we can get between any two states) but not regular (for any r we have $p_{0,r+1}^{(r)} = 0$).

For an equilibrium distribution we would need to satisfy

$$w_0 = (1-p)w_0 + (1-p)w_1 + (1-p)w_2 + \cdots = (1-p)\sum_{i=0}^{\infty} = 1-p$$

and

$$\begin{aligned} w_1 &= pw_0 \\ w_2 &= pw_1 \\ &\vdots \\ w_k &= pw_{k-1} \text{ for all } k \geqslant 1 \end{aligned}$$

But this means that $w_1 = p(1-p)$, $w_2 = p^2(1-p)$ and in general $w_k = p^k(1-p)$.

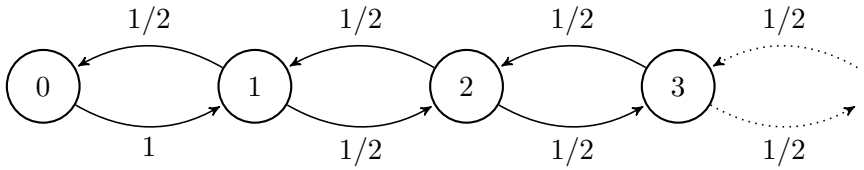
This should look familiar, if $Z \sim \text{Geom}(p)$ then the equilibrium distribution is the distribution of $Z - 1$.

In fact, if r is large ($r \geqslant k+1$ will do) then

$$\mathbb{P}(X_r = k) = \mathbb{P}(\text{last } k+1 \text{ tosses were } T\underbrace{H \cdots H}_{k}) = (1-p)p^k$$

So this is also a limiting distribution.

Example 25. The random walk on $\mathbb{N}$ with reflecting boundary is the Markov chain with state space $\mathbb{N}$ and transition graph:



This is irreducible (we can get between any two states). It it not regular (the chain alternates between even and odd states, also given any r we have $p^{(r)}_{0,r+1} = 0$).

For an equilibrium distribution we would need to satisfy

$$\begin{aligned} w_0 &= \tfrac{1}{2}w_1 \\ w_1 &= w_0 + \tfrac{1}{2}w_2 \\ w_2 &= \tfrac{1}{2}w_1 + \tfrac{1}{2}w_3 \\ w_k &= \tfrac{1}{2}w_{k-1} + \tfrac{1}{2}w_{k+1} \text{ for all } k \geqslant 2 \end{aligned}$$

The first two equations give that $w_1 = 2w_0$ and $w_2 = 2w_1 - 2w_0 = 2w_0$. The remaining equations give that $w_k = 2w_0$ for all k (you could prove this by induction on k for instance).

In order to have a distribution we would need $\sum_{i=0}^{\infty} w_i = 1$. But, we have

$$\sum_{i=0}^{\infty} w_i = w_0 + 2w_0 + 2w_0 + 2w_0 + \cdots .$$

There is no way to choose w_0 to make this sum equal to 1.

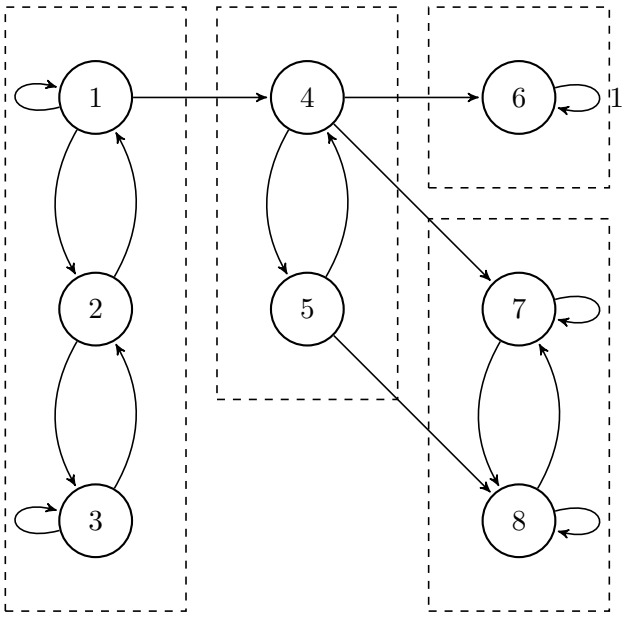
So there is no equilibrium distribution.

We won't say much about the theory of chains with infinite state spaces in this module. However, we will see some more examples and a reason for the difference between the two behaviours in Examples 24 and 25.

5 Recurrence and Transience

5.1 Communicating Classes

Consider the Markov chain with following transition graph (ignore the dashed boxes for now):



(The arrows can have any positive probabilities on provided that the numbers on the arrows out of each state sum to 1)

If we start at state 1 the chain will behave as follows:

- Move among $\{1, 2, 3\}$ then escape to $\{4, 5\}$
- Move among $\{4, 5\}$ (we can never return to $\{1, 2, 3\}$) then either leave to
  - state 6 where we are absorbed or
  - states $\{7, 8\}$ and continue to move among $\{7, 8\}$ never leaving.

The sets $\{1, 2, 3\}, \{4, 5\}, \{6\}, \{7, 8\}$ partition S and have the property that once we leave we can never return. They are the minimal sets with this property in the sense that no proper subset of any of them has this property.

Let's make some definitons to formalise this.

Definition 14. Let $a, b \in S$ be states

- *a communicates with b* (written $a \to b$) if $p_{ab}^{(r)} > 0$ for some $r \in \mathbb{N}$ (that is we can move from a to b in some number of steps).

- a and b *intercommunicate* (written $a \leftrightarrow b$) if $a \to b$ and $b \to a$.
- The *communicating class* for state a is $\{s \in S : a \leftrightarrow s\}$.

Remark. Intercommunicates is a relation on S. If you took the module Introduction to Algebra you will have seen the concept of an equivalence relation. It turns out that $\leftrightarrow$ is an equivalence relation and the communicating classes are its equivalence classes. A consequence of this is that the communicating classes form a partition of S.

Example 26. For the Markov chain at the start of this section, the sets $\{1,2,3\}, \{4,5\}, \{6\}, \{7,8\}$ that we identified are the communicating classes.

Example 27. These definitions fit with some other concepts we have seen:

- $a \leftrightarrow a$ for any chain and any $a \in S$ (since $p_{aa}^{(0)} = 1$). If a is an absorbing state then the singleton $\{a\}$ is a communicating class.
- A chain is irreducible if and only if S has exactly one communicating class (namely the whole of S).

5.2 First-return probabilities

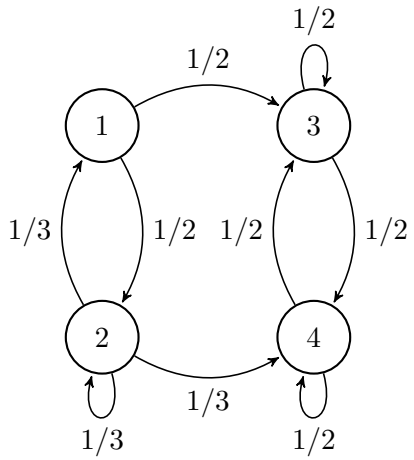
Definition 15. Let $(X_0, X_1, X_2, \ldots)$ be a Markov chain with state space S. For $i \in S$ and $t \geqslant 1$ let

- $f_i^{(t)} = \mathbb{P}(X_t = i, X_{t-1} \neq i, X_{t-2} \neq i, \cdots, X_1 \neq i \mid X_0 = i)$ (the probability that the *first* return to state i is at time t).
- $f_i = \sum_{t=1}^{\infty} f_i^{(t)}$ (the probability of return to state i).

Definition 16. Let $(X_0, X_1, X_2, \ldots)$ be a Markov chain with state space S. A state $i \in S$ is

- *recurrent* if $f_i = 1$ (we are certain to return to i);
- *transient* if $f_i < 1$ (there is some positive probability that we never return to i).

Example 28. Let $(X_0, X_1, X_2, \ldots)$ be the Markov chain with state space $\{1, 2, 3, 4\}$ and transition graph



Consider each of the states:

- State 1: $f_1^{(1)} = 0$, $f_1^{(2)} = \frac{1}{2} \times \frac{1}{3} = \frac{1}{6}$. If $t \geqslant 2$, $f_1^{(t)} = \frac{1}{2} \times \left(\frac{1}{3}\right)^{t-2} \frac{1}{3} = \frac{1}{6}\left(\frac{1}{3}\right)^{t-2}$. Summing these

$$f_1 = \sum_{t=2}^{\infty} \tfrac{1}{6}\left(\tfrac{1}{3}\right)^{t-2} = \tfrac{1}{4} \qquad \text{(sum of a geometric progression)}$$

  So state 1 is transient; there is a probabilty of 3/4 that we never return.

- State 3: Similarly $f_3^{(t)} = \frac{1}{2}\left(\frac{1}{2}\right)^{t-2}\frac{1}{2} = \left(\frac{1}{2}\right)^t$ for $t \geqslant 1$. Summing these

$$f_3 = \sum_{t=1}^{\infty} \left(\tfrac{1}{2}\right)^t = 1$$

  So state 3 is recurrent; the probability that we never return is 0.

- State 4 behaves just like state 3. So it is also recurrent.

- State 2: $f_2^{(1)} = \frac{1}{3}$, $f_2^{(2)} = \frac{1}{6}$, $f_2^{(t)} = 0$ for $t > 2$. So $f_2 = \frac{1}{3} + \frac{1}{6} = \frac{1}{2} < 1$. State 2 is transient.

Notice that states in the same communicating class behave in the same way.

- The set $\{1, 2\}$ is a communicating class and 1 and 2 are both transient states.

- The set $\{3, 4\}$ is a communicating class and 3 and 4 are both recurrent states.

Example 29. In that example we could calculate f_i exactly. If we only care about classifying states as recurrent or transient we can be a bit less accurate. Look back at the 8 state example at the start of this section.

- State 1: There is a chance that we leave to state 4 immediately and then we can never return. So $f_1 \leqslant 1 - p_{1,4} < 1$. State 1 is transient.
- State 2: There is a path 2 to 1 to 4 which if we follow we can never return. So $f_2 \leqslant 1 - p_{2,1}p_{1,4} < 1$. State 2 is transient.
- Similarly states $3, 4, 5$ are all transient.
- State 6 is certainly recurrent as it is absorbing.
- State 7: The only way we can never return is if we go to 8 then loop for ever. This has probability 0. So state 7 is recurrent.
- State 8 is recurrent for the same reason.

Example 30. If a Markov chain has a finite communicating class C which can never be left then every state in it is recurrent. In particular, every state in a finite irreducible chain is recurrent. To see this. suppose we start at $s \in C$. Because C is a communicating class, from every state $i \in C$, there is a path back to s with probability p_i of being followed. Let $p = \min\{i \in C : p_i\}$. (we can take this minimum since C is finite). Suppose we leave s and go to state i_1. In order to not return to s we need to fail to follow our path back to s; this has probability $(1 - p_{i_1})$. If we haven't returned to s we are now at some other state i_2. Again, in order to not return to s we need to fail to follow our path back to s and this has probability $(1 - p_{i_2})$. Continuing, the probability of never returning to s is an infinite product of terms each of the form $(1 - p_i)$. Since $(1 - p_i) > 1 - p$, the probability that we never return to s is at most $(1 - p)(1 - p)(1 - p) \ldots = 0$.

5.3 Properties of recurrent and transient states

There are several consequences of the dichotomy between recurrence and transience. Here are some important ones.

Theorem 5.1. *Let $(X_0, X_1, X_2, \ldots)$ be a Markov chain with state space S. Let $s \in S$ be a state and set $X_0 = s$. Let M_s be the random variable counting the number of visits to s (including the one at time 0).*

(a) If state s is recurrent then

(i) $\mathbb{P}(M_s = \infty) = 1$;

(ii) $\sum_{t=0}^{\infty} p_{ss}^{(t)} = \infty$;

(iii) All states in the same communicating class as s are recurrent.

(b) If state s is transient then

(i) $\mathbb{P}(M_s = \infty) = 0$, *moreover* $M_s \sim Geom(1 - f_s)$;

(ii) $\sum_{t=0}^{\infty} p_{ss}^{(t)} < \infty$;

(iii) All states in the same communicating class as s are transient.

Remark.

- Parts (iii) of this Theorem confirm our observation that states in the same communicating class behave in the same way. We say that reccurence and transience are *class properties* meaning that they are properties that apply to every state in a communicating class.
- Parts (ii) provide a test for checking recurrence/transience. This can be useful as the $p_{ss}^{(t)}$ are often easier to calculate than the $f_s^{(t)}$. For instance, we will use this condition for recurrence/transience in proving part (iii).
- There are a few place in this (and the theory for the next few sections) where I am being a bit informal in dealing with the infinite. Everything works but there are some technical justifications behind the scenes which we will ignore. Some of you (for instance the kind of people who were a bit suspicious of calculus before taking Differential and Integral Analysis) will need to take a few things on trust here.

Proof. (i) Suppose that s is transient. Then by definition $f_s < 1$. If we return to s then the probability we make a further return to s is also f_s. Moreover, this is independent of how many returns we have made to s already (by the Markov property). Making k visits to s means that we return to s exactly $k - 1$ times. Hence,

$$\mathbb{P}(M_s = k) = (f_s)^{k-1}(1 - f_s).$$

This is the pmf of a Geometric random variable so $M_s \sim \text{Geom}(1 - f_s)$.

It follows that $\sum_{k=0}^{\infty} \mathbb{P}(M_s = k) = 1$ so with probability 1 the chain makes only a finite number of returns to s. So $\mathbb{P}(M_s = \infty) = 0$.

On the other hand, if s is recurrent then $f_s = 1$. The probability that we make exactly k returns to s is $(f_s)^{k-1}(1 - f_s) = 0$. In other words $\mathbb{P}(M_s = k) = 0$ for any finite k. So we must have $\mathbb{P}(M_s = \infty) = 1$.

(ii) Suppose that $X_0 = s$. We will use the $p_{ss}^{(t)}$ to calculate $\mathbb{E}(M_s)$. To do this introduce the following indicator random variables. Let

$$A_t = \begin{cases} 1 & \text{if } X_t = s \\ 0 & \text{if } X_t \neq s. \end{cases}$$

Each A_t is a Bernoulli random variable and A_t takes the value 1 with probability $p_{ss}^{(t)}$. So $\mathbb{E}(A_t) = p_{ss}^{(t)}$. Since for each visit to s we have one of the A_t taking the value 1, these random variable can be used to count the number of visits to s.

$$\begin{aligned} M_s &= \sum_{t=0}^{\infty} A_t \\ \mathbb{E}(M_s) &= \mathbb{E}\left(\sum_{t=0}^{\infty} A_t\right) \\ &= \sum_{t=0}^{\infty} \mathbb{E}(A_t) && \text{(By linearity of expectation)} \\ &= \sum_{t=0}^{\infty} p_{ss}^{(t)}. \end{aligned}$$

Now, if s is transient, $M_s \sim \text{Geom}(1 - f_s)$ (by part (i))so $\mathbb{E}(M_s) = \frac{1}{1-f_s}$. That is $\sum_{t=0}^{\infty} p_{ss}^{(t)} = \frac{1}{1-f_s}$ which is finite.

If s is recurrent, then $\mathbb{E}(M_s) = \infty$. That is $\sum_{t=0}^{\infty} p_{ss}^{(t)} = \infty$.

(iii) There are two parts to this

- If $a \leftrightarrow b$ and a is transient then b is transient.
- If $a \leftrightarrow b$ and a is recurrent then b is recurrent.

These are equivalent (they both say that we can never have $a \leftrightarrow b$ with a recurrent and b transient). So it suffices to prove the first. We will use part (ii) to do this.

Suppose that $a \leftrightarrow b$ and a is transient.

- Since $a \to b$, we have $p_{ab}^{(r)} > 0$ for some $r \in \mathbb{N}$.

- Since $b \to a$, we have $p_{ba}^{(s)} > 0$ for some $s \in \mathbb{N}$.

Now if $t \geqslant r + s$ there is a route from a to a by going: a to b in r steps, b to b in $t - r - s$ steps, b to a in s steps. So

$$p_{aa}^{(t)} \geqslant p_{ab}^{(r)} p_{bb}^{(t-r-s)} p_{ba}^{(s)} = \alpha p_{bb}^{(t-r-s)},$$

where $\alpha = p_{ab}^{(r)} p_{ba}^{(s)}$.

Now,

$$\sum_{t=0}^{\infty} p_{aa}^{(t)} \geqslant \sum_{t=r+s}^{\infty} p_{aa}^{(t)} \geqslant \sum_{t=r+s}^{\infty} \alpha p_{bb}^{(t-r-s)} \geqslant \alpha \sum_{k=0}^{\infty} p_{bb}^{(k)}$$

Since a is transient the lefthand side is finite. So the righthand side is finite. So b is transient.

□

5.4 First return times

Suppose that $s \in S$ is a recurrent state and we start out chain with $X_0 = s$. The time at which we first return to s is a random variable which we can study.

Definition 17. If s is recurrent and $X_0 = s$ then the *first-return time* is the random variable

$$R_s = \min\{t \geqslant 1 : X_t = s\}.$$

We made this definition only for recurrent s because if s is transient then there is a positive probability that the set $\{t \geqslant 1 : X_t = s\}$ is empty and so has no minimum.

Th random variable R_s takes values in $\mathbb{N} \setminus \{0\}$ and its probability mass function is given by the first return probabilities.

$$\mathbb{P}(R_s = t) = f_s^{(t)}.$$

So its expectation (that is the expected first return time, or expected time between visits to s) is given by:

$$\mathbb{E}(R_s) = \sum_{t=1}^{\infty} t f_s^{(t)}.$$

For simple chains we could calculate this directly (for instance in Example 28, the formula for $f_3^{(t)}$ shows that $R_3 \sim \text{Geom}(1/2)$ so $\mathbb{E}(R_3) = 2$). If things are a bit more complicated we could use first-step analysis by setting

$$v_i = \mathbb{E}(\text{time of first visit to } s \mid X_0 = i\} \text{ for } i \neq s.$$

Conditioning on the first step as usual gives a system of equations for v_i and then $\mathbb{E}(R_s) = 1 + \sum_{i \neq s} p_{si} v_i$.

However, we will soon see a nice connection between $\mathbb{E}(R_s)$ and the equilibrium distribution which gives another method to calculate it.

5.5 Positive and null recurrence

If $\mathbb{E}(R_s)$ is large then, even though we are certain to return to s, it is likely to take a long time. This idea motivates the next definition.

Definition 18. Let $s \in S$ be a recurrent state.

- If $\mathbb{E}(R_s)$ is finite then s is said to be *positive recurrent*
- If $\mathbb{E}(R_s)$ is infinite then s is said to be *null recurrent*

Remark. You should think of null recurrent states as being recurrent but only just. We return to them with probabilty 1 but the expected time to return is infinite. This can only happen when S is infinite. For finite S, every recurrent state is positive recurrent (the argument of Example 30 can be adapted to show this).

Here is another way in which null recurrent states are 'nearly transient'. Recall that if s is transient then $\sum_{t\geqslant 1} p_{ss}^{(t)}$ is finite. This means that $p_{ss}^{(t)} \to 0$ as $t \to \infty$. However, it is also possible to have $p_{ss}^{(t)} \to 0$ as $t \to \infty$ but $\sum_{t\geqslant 1} p_{ss}^{(t)}$ being infinite. It turns out that such s are precisely the null recurrent ones.

Theorem 5.2. *Let $(X_0, X_1, X_2, \ldots)$ be a Markov chain with state space S. Let $s \in S$ be a recurrent state. Then*

- *The state s is null recurrent if and only if $p_{ss}^{(t)} \to 0$ as $t \to \infty$;*
- *The state s is positive recurrent if and only if $p_{ss}^{(t)} \not\to 0$ as $t \to \infty$.*

Proof. Omitted and non-examinable. □

This can be used (similarly to the proof of part (iii) Theorem 5.1) to show that null recurrence and positive recurrence are class properties. So if a chain is irreducible, every state behaves the same and we can speak of the chain being transient, null recurrent or positive recurrent.

We said earlier that there would be a connection between $\mathbb{E}(R_s)$ and the equilibrium distribution. This is the content of the following result.

Theorem 5.3. *An irreducible Markov chain has a unique equilibrium distribution if and only if it is positive recurrent. In this case the equilibrium distribution is given by*

$$w_s = \frac{1}{\mathbb{E}(R_s)}.$$

Proof. Omitted and non-examinable. □

Remark. This is a useful result.

- It gives a quick way of calculating $\mathbb{E}(R_s)$ via finding the equilibrium distribution which is useful even in the finite S case.
- When S is infinite, it explains the different behaviour we saw in Examples 24 and 25.

Example 31. Pick any example of a chain with an equilibrium distribution. How about Example 19 for instance. This had transition matrix

$$\begin{pmatrix} 1/4 & 3/4 & 0 & 0 & 0 \\ 1/4 & 0 & 3/4 & 0 & 0 \\ 0 & 1/4 & 0 & 3/4 & 0 \\ 0 & 0 & 1/4 & 0 & 3/4 \\ 0 & 0 & 0 & 1/4 & 3/4 \end{pmatrix}$$

and we found that

$$\mathbf{w} = \left(\tfrac{1}{121} \quad \tfrac{3}{121} \quad \tfrac{9}{121} \quad \tfrac{27}{121} \quad \tfrac{81}{121}\right).$$

was the unique equilibrium distribution.

By Theorem 5.3 we have that $\mathbb{E}(R_2) = \frac{121}{3}$. So if we start the chain with $X_0 = 2$ then the expectation of the time of the first return to state 2 is $\frac{121}{3} \approx 40.33$.

Now let's look back to Examples 24 and 25 and use positive recurrence to explain why the first of them had an equilibrium distribution but the second did not.

Example 32. Recall Example 24 (the success runs chain). We have that $\mathbb{P}(X_t = 0) = 1-p$ for all t (because being in state 0 happens if and only if the last toss was a Tail). So $p_{0,0}^{(t)} = 1 - p \not\to 0$ as $t \to \infty$. So state 0 is positive recurrent (by Theorem 5.2). So by Theorem 5.3 we must have a unique equilibrium distribution (which we calculated in our previous discussion of this example).

We can also use Theorem 5.3 to read off information about the first return times. We had that $w_k = (1-p)p^k$ so the expected first return time (or if you prefer the expected time between visits) to state k is $\frac{1}{(1-p)p^k}$.

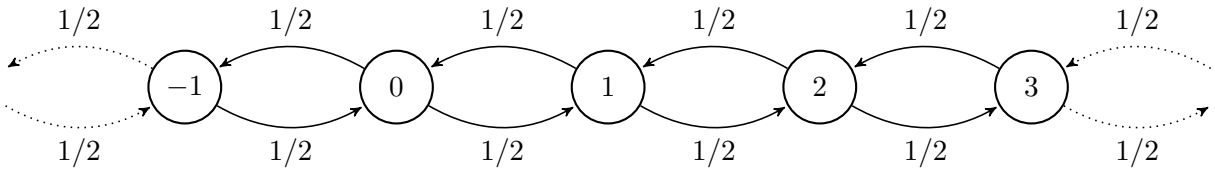
For the next two examples we will need the following approximation for the middle Binomial coefficient[9]:

$$A\frac{2^{2t}}{\sqrt{t}} < \binom{2t}{t} < B\frac{2^{2t}}{\sqrt{t}}$$

where A and B are absolute constants.

[9]This is a consequence on Stirling's approximation for $n!$ which in fact gives a more precise bound.

Example 33. Let $(X_0, X_1, X_2, \ldots)$ be the random walk on $\mathbb{N}$ with reflecting boundary (Example 25). To find the probability of returning to state 0 in t steps in that chain we will compare it with the chain $(Y_0, Y_1, Y_2, \ldots)$ with state space $\mathbb{Z}$ and transition graph:



This is the instance of Example 5 (the random walk on $\mathbb{Z}$) with $p = 1/2$.

The point is that the process $(|Y_0|, |Y_1|, |Y_2|, \ldots)$ follows exactly the same rules as $(X_0, X_1, X_2, \ldots)$. So the probability of returning to state 0 in t steps in (X_i) is exactly the same as it is in (Y_i). We will work in (Y_i) as the calculations are a little tidier.

We have $p_{0,0}^{(t)} = 0$ if t is odd, and $p_{0,0}^{(2t)} = \binom{2t}{t}\left(\frac{1}{2}\right)^{2t}$ (there are $\binom{2t}{t}$ paths of $2t$ steps from state 0 to state 0 and each one has probability $\left(\frac{1}{2}\right)^{2t}$).

Now

$$p_{0,0}^{(2t)} = \binom{2t}{t}\left(\frac{1}{2}\right)^{2t} < B\frac{2^{2t}}{\sqrt{t}}\left(\frac{1}{2}\right)^{2t} = \frac{B}{\sqrt{t}} \to 0 \text{ as } t \to \infty$$

So state 0 is either transient or null recurrent. To decide which we look at the sum

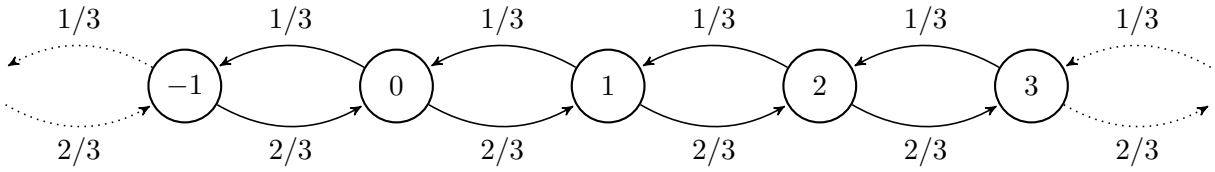
$$\sum_{t\geqslant 1} p_{0,0}^{(t)} = \sum_{t\geqslant 1} p_{0,0}^{(2t)} = \sum_{t\geqslant 1} \binom{2t}{t}\left(\frac{1}{2}\right)^{2t} > \sum_{t\geqslant 1} A\frac{2^{2t}}{\sqrt{t}}\left(\frac{1}{2}\right)^{2t} > A\sum_{t\geqslant 1}\frac{1}{\sqrt{t}}$$

But this is infinite (recall from Calculus II) so state 0 is null recurrent.

This means that every state is null recurrent (because it is a class property).

The expectation of the first return time to 0 is infinite and (as we saw in Example 25) there is no equilibrium distribution.

Example 34. For our last example of this section, let's look at the random walk on $\mathbb{Z}$ (Example 5) when $p \neq 1/2$. I will do the $p = 2/3$ case. Any $p \neq 1/2$ would show essentially the same behaviour (try it!).



As before we have $p_{0,0}^{(t)} = 0$ if t is odd. But now $p_{0,0}^{(2t)} = \binom{2t}{t}\left(\frac{1}{3}\right)^t\left(\frac{2}{3}\right)^t$ (there are $\binom{2t}{t}$ paths of $2t$ steps from state 0 to state 0 and each one has t steps to the right and t steps to the left so has probability $\left(\frac{1}{3}\right)^t\left(\frac{2}{3}\right)^t$).

Now

$$p_{0,0}^{(2t)} = \binom{2t}{t}2^{-2t} < B\frac{2^{2t}}{\sqrt{t}}\left(\frac{1}{3}\right)^t\left(\frac{2}{3}\right)^t = B\frac{x^t}{\sqrt{t}} \to 0 \text{ as } t \to \infty$$

where $x = 8/9$ and the limit is 0 because $-1 < x < 1$.

As before state 0 is either transient or null recurrent. To decide which we look at the sum

$$\sum_{t\geqslant 1} p_{0,0}^{(t)} = \sum_{t\geqslant 1} p_{0,0}^{(2t)} < B\sum_{t\geqslant 1}\frac{x^t}{\sqrt{t}} < B\sum_{t\geqslant 1} x^t = \frac{Bx}{1-x} = 8B$$

which is finite.

This means that every state is transient (because transience is a class property).

The probability that we return to state 0 is strictly less that 1. There is a positive probability that we never return.

5.6 Summary

The last four examples show all the possible behaviours for an irreducible chain. The example of an irreducible chain with all states transient (Example 34) needs an infinite state space to work. If a chain is irreducible and S is finite then we can never escape forever from a state (because the chain is irreducible) and we must return with probability 1 (see Example 30). So if S is finite then an irreducible chain must be positive recurrent.

The last four example show all the possible behaviours for an irreducible chain[10]:

- If S is finite, then the chain must be positive recurrent and have a unique equilibrium distribution (Example 19 among many others).
- If S is infinite, then we could have
  - The chain is positive recurrent and has a unique equilibrium distribution (Example 24)
  - The chain is null recurrent and has no equilibrium distribution (Example 25)
  - The chain is transient and has no equilibrium distribution (Example 34)

To decide between these for infinite S we can look at $p_{ss}^{(t)}$ (choosing any $s \in S$ since these are class propeties).

- If $p_{s,s} \not\to 0$ then the chain is positive recurrent (and $\sum_{t \geqslant 1} p_{ss}^{(t)}$ must be infinite).
- If $p_{s,s} \to 0$ but $\sum_t p_{ss}^{(t)}$ is infinite then the chain is null recurrent.
- If $p_{s,s} \to 0$ and $\sum_t p_{ss}^{(t)}$ is finite then the chain is transient.

Notice that I didn't specify exactly what range of t the sum was over. We could take $t \geqslant 0$ or $t \geqslant 1$ (or even $t \geqslant 100$). Since the only thing that matters is whether the sum is finite or infinite changing the first few terms won't make any difference.

[10] Recall that transience, null recurrence and positive recurrence are class properties, so if the chain is irreducible we can use these words to describe the whole chain.

6 Continuous-time Stochastic Processes

At the start of the module we defined a stochastic process as a collection of random variables indexed by a set T. Up to now we have mainly been working with discrete-time processes, that is $T = \mathbb{N}$ (or similar).

In the second part of the course we will study *continuous-time stochastic processes* where $T = \mathbb{R}_{\geqslant 0} = \{t \in \mathbb{R} : t \geqslant 0\}$.

Definition 19. A *continuous-time stochastic process* is a collection of random variables $(X(t) : \ t \in \mathbb{R}_{\geqslant 0})$ each taking values in some set S.

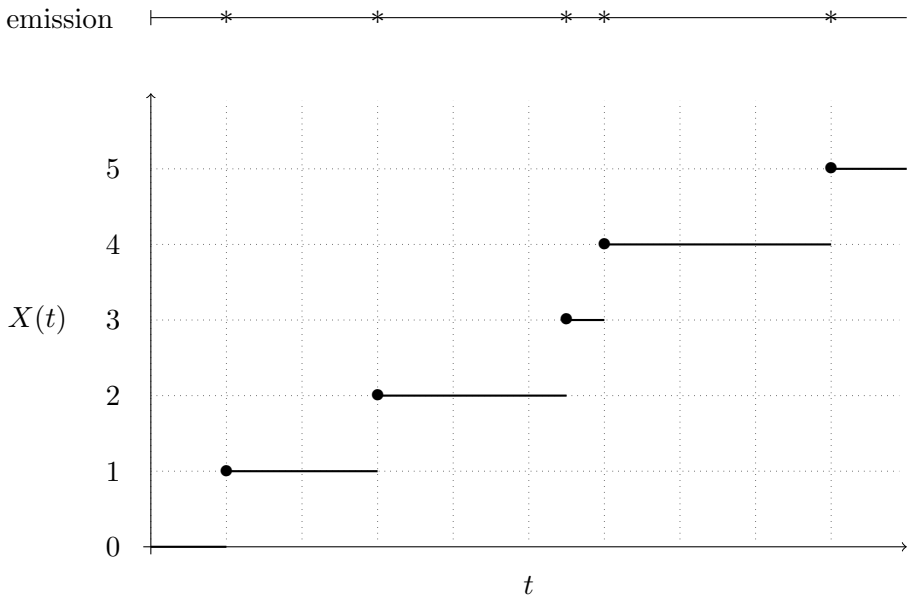
There is one random variable $X(t)$ for each non-negative real number t. In a slight change of notation from the discrete-time case we usually write this as $X(t)$ rather than X_t. Just as before we will usually think of the index set $\mathbb{R}_{\geqslant 0}$ as representing time.

If we are trying to model changes which can take place at any real-valued time (as opposed to just at integer times) then we would use a continuous-time stochastic process. Many real-world processes are modelled better as continuous-time stochastic processes, than as discrete-time stochastic processes. Here are some examples:

- $X(t) =$ the number of alpha particles emitted by a radioactive source between time 0 and time t, where times are measured in seconds from the moment a particle detector is switched on.
- $X(t) =$ the number of buses passing a bus stop outside Queen Mary between time 0 and time t, where times are measured in minutes after 08:00 this morning.
- $X(t) =$ the number of people in a supermarket queue t minutes after the supermarket opens.
- $X(t) =$ the number of holes in the first t metres of a pipeline.
- Pick any discrete-time process and modify it so that instead of making transitions at regular times $1, 2, 3, \ldots$ we make transitions according to the same probabilities but at random times. If we are in state i, we wait for an amount of time drawn from some distribution and then jump to state j with probability p_{ij}. Now our continuous-time process is given by $X(t) =$ the state occupied at time t.

In most of these we are counting discrete events (such as a particle being emitted or a person joining or leaving a queue) but they can happen at any point in continuous time. For this kind of example, the $X(t)$ take non-negative integer values so we can take $S = \mathbb{N}$. This will be typical of all our examples.

Let's draw a couple of pictures to help visualise the first example:



The first picture shows the times at which a particle is emitted (represented by $*$). The graph below shows how $X(t)$ evolves. At time 0 no particles have been emitted so $X(0) = 0$. Then $X(t)$ remains at 0 until the first particle is emitted at which point it jumps up to 1. Continuing, $X(t)$ evolves by making a jump up by 1 at each of the (random) times that a particle is emitted.

This example of counting particles emitted from a radioactive source is a classic example of the first continuous-time process we study: the Poisson process.

7 The Poisson Process

7.1 Definition and simple calculations

One of the most important continuous-time stochastic processes is the *Poisson Process*. Before defining this, recall the following.

Definition 20. An integer-valued random variable R is said to have the *Poisson distribution with parameter* λ if

$$\mathbb{P}(R = k) = e^{-\lambda}\frac{\lambda^k}{k!} \quad \text{for all } k \in \mathbb{N}$$

In this case, we write $R \sim \text{Po}(\lambda)$.

Now we are ready to define a Poisson process.

Definition 21. Let $\lambda > 0$. A continuous-time stochastic process $(X(t) : \ t \geqslant 0)$ is said to be a *Poisson process of rate* λ if the following properties hold.

(i) $X(0) = 0$.

(ii) For any $0 \leqslant t_1 < t_2 < t_3 < \ldots < t_n$, the random variables

$$X(t_2) - X(t_1),\ X(t_3) - X(t_2),\ X(t_4) - X(t_3),\ \ldots,\ X(t_n) - X(t_{n-1})$$

are mutually independent of one another[11].

(iii) For any $s, t \geqslant 0$, we have $X(s+t) - X(s) \sim \text{Po}(\lambda t)$.

Remark. This definition implies that in a Poisson process $(X(t) : \ t \geqslant 0)$, each random variable $X(t)$ takes values in $\mathbb{N}$. Moreover if $t_2 > t_1$ then $X(t_2) \geqslant X(t_1)$. Hence, a Poisson process can be thought of as 'counting' the number of events of a certain kind[12] (often called arrivals to avoid confusion with the probabilistic use of the word event), which occur between time 0 and time t. For example, the number of radioactive particles emitted, or the number of buses passing a stop between time 0 and time t.

Let's look at these properties:

(i) This is a (fairly natural) technical condition which makes calculations easier; if we liked we could allow $X(0) = s$ instead for any $s \in \mathbb{N}$ without changing too much.

[11] Recall that random variables $R_1, R_2, \ldots, R_n$ are said to be mutually independent of one another if for any $a_1, a_2, \ldots, a_n$, we have $\mathbb{P}(R_1 = a_1, R_2 = a_2, \ldots R_n = a_n) = \mathbb{P}(A_1 = a_1)\mathbb{P}(A_2 = a_2) \cdots \mathbb{P}(A_n = a_n)$.

[12] As in the discrete-time Example 7 we say that a process with $S = \mathbb{N}$ satisfying that $X(t)$ is non-decreasing in t is called a *counting process*.

(ii) This shows that the Poisson process has a rather strong 'memoryless' property. In particular, it implies that $X(t+s)-X(s)$ is independent of $X(s)$ (and of the evolution of the process before time s). This should remind you of the Markov property. A useful special case of property (ii) is that if $a < b < c < d$ (so the intervals $[a, b]$ and $[c, d]$ are disjoint) then the number of arrivals in $[a, b]$ and the number of arrivals in $[c, d]$ are independent random variables.

(iii) This property looks somewhat mysterious; why should the Poisson distribution be important here? We will see the reason for this later.

Example 35. Suppose that I receive email according to a Poisson process of rate 2 per hour. That is if I start counting messages from time 0, and let $X(t)$ be the number of messages that I have have recieved by time t hours then the $X(t)$ satisfy the definition above with $\lambda = 2$. What is:

(i) The probability that I receive 2 messages in the first hour.

(ii) The probability that I receive 2 messages in the second hour.

(iii) The probability that I receive 2 messages in the first two hours.

(iv) The probability that I receive 2 messages in the second hour given that I received none in the first hour.

(v) The probability that I receive no messages in the first hour and 2 in the second hour.

We have:

(i) $\mathbb{P}(2 \text{ messages in first hour}) = \mathbb{P}(\underbrace{X(1) - X(0)}_{\sim \mathrm{Po}(2)} = 2) = e^{-2}\frac{2^2}{2!} = \frac{2}{e^2}$

(ii) $\mathbb{P}(2 \text{ messages in second hour}) = \mathbb{P}(\underbrace{X(2) - X(1)}_{\sim \mathrm{Po}(2)} = 2) = e^{-2}\frac{2^2}{2!} = \frac{2}{e^2}$

(iii) $\mathbb{P}(2 \text{ messages in first two hours}) = \mathbb{P}(\underbrace{X(2) - X(0)}_{\sim \mathrm{Po}(4)} = 2) = e^{-4}\frac{4^2}{2!} = \frac{8}{e^4}$

(iv) We want $\mathbb{P}(2 \text{ messages in second hour} \mid \text{none in first hour})$. Using that $X(2) - X(1)$ and $X(1) - X(0)$ are independent random variables (property (ii) of Definition 21), this is

$$\begin{aligned}\mathbb{P}(X(2) - X(1) = 2 \mid X(1) - X(0) = 0) &= \mathbb{P}(X(2) - X(1) = 2) \\ &= \frac{2}{e^2}\end{aligned}$$

(v) We want $\mathbb{P}$(no messages in first hour, 2 messages in second hour). Again, using that $X(2)-X(1)$ and $X(1)-X(0)$ are independent random variables, this is

$$\begin{aligned}\mathbb{P}(X(2)-X(1)=2, X(1)-X(0)=0) &= \mathbb{P}(X(2)-X(1)=2)\cdot\mathbb{P}(X(1)-X(0)=0)\\ &= \frac{2}{e^2}\times e^{-2}\\ &= \frac{2}{e^4}\end{aligned}$$

Example 36 (Truncation)**.** Suppose that $X(t)$ is a Poisson process of rate λ. Fix some time $s>0$ and define a new process $(Y(t):\ t\geqslant 0)$ by $Y(t)=X(t+s)-X(s)$. In other words we are interested in the same events that defined the process $X(t)$ but we start counting them from time s rather than time 0. The process $Y(t)$ is also a Poisson process of rate λ. Indeed $Y(0)=X(0+s)-X(s)=0$ and the other two properties are clearly inherited from the process X. This shows that the Poisson process has a 'memoryless' property. If I start observing it from any point in time then what I see is still a Poisson process.

7.2 The infinitesimal definition of the Poisson process

It turns out that we can give an equivalent definition of the Poisson process (i.e. a definition equivalent to Definition 21) in terms of how the process behaves over small time-intervals; this is called the *infinitesimal definition* of the Poisson process. To do this we need the following notation.

Notation. We use the notation $o(h)$ to denote *any* function $f=f(h):\mathbb{R}\to\mathbb{R}$ such that $\frac{f(h)}{h}\to 0$ as $h\to 0$.

The idea is that when we are considering expressions of the form ch (where c is a constant) we will represent any neglible lower order terms as $o(h)$. For instance $h^2=o(h)$, $100h^2=o(h)$, but $\frac{1}{1000}h$ is not $o(h)$, and $\sqrt{h}$ is not $o(h)$. Also $e^{-\lambda h}=1-\lambda h+o(h)$.

We want to count events which we will refer to as arrivals. Arrivals can happen at any time, and occur randomly at a constant rate λ per unit time. We will do this using a stochastic process $(N(t):t\geqslant 0)$ with $N(t)$ being the number of arrivals occuring in the time interval $[0,t]$. From this description, we necessarily have $N(0)=0$ and $N(s)\geqslant N(t)$ for all $s\geqslant t$. It is reasonable to think that the process would also satisfy the following properties:

(i) For any $n,m\in\mathbb{N}$,

$$\mathbb{P}(N(t+h)=n\mid N(t)=m)=\begin{cases}\lambda h+o(h) & \text{if } n=m+1;\\ 1-\lambda h+o(h) & \text{if } n=m;\\ o(h) & \text{if } n\geqslant m+2;\\ 0 & \text{if } n<m.\end{cases}$$

(ii) The probability distribution of the random variable $N(s+t)-N(s)$ does not depend on s. (This property is called the *time-homogeneity condition.*)

(iii) For any $0 \leqslant t_1 < t_2 < t_3 < \ldots < t_n$, the random variables

$$N(t_2) - N(t_1),\ N(t_3) - N(t_2),\ N(t_4) - N(t_3),\ \ldots,\ N(t_n) - N(t_{n-1})$$

are mutually independent of one another.

Condition (i) above says that if, for example, $N(t)$ denotes the number of particles that arrive at a particle detector between time 0 and time t, then in a short interval $[t, t+h]$ we have:

- the probability of exactly one new particle arriving is approximately λh;
- the probability of more than one new particle arriving is negligible.

One way of seeing where this probability λh comes from is to imagine splitting the time interval $[0,1]$ into a large number $\frac{1}{h}$ intervals each of length h. Since each of them is short, we can assume that each interval contains either 1 arrival (with some probability p) or 0 arrivals (with probability $1-p$). If there are $\frac{1}{h}$ intervals and in each one there is probability p of an arrival, then the expectation of the total number of arrivals in $[0,1]$ is $\frac{p}{h}$. The rate is λ per unit time so this expected value should be λ which forces $p = \lambda h$.

This is true of many processes that occur in nature — for example, the slow radioactive decay of a large piece of radioactive material, or the arrival of photons from a distant source of light. It can also be a good approximation to many social or economic processes such as customers entering a supermarket, phone calls made to a call centre, or cars passing a certain road junction[13].

We will show that if these properties are satisfied then the process $N(t)$ is a Poisson process of rate λ. In fact these properties give an alternative way of defining the Poisson process (sometimes called the infinitesimal definition of the Poisson process).

Remark. There are several reasons why it can be good to have several equivalent definitions of the same thing. In this case:

- Condition (i) of the infinitesimal definition is a bit more intuitive, and explains the mystery of the Poisson distribution appearing in property (iii) of our original definition 21.
- We will see later that some calculations and arguments about the Poisson process are easier using the original definition and some are easier using the infinitesimal one. Knowing both of them gives us the best of both worlds.

[13] In reality time-homoegenity will not hold exactly here; there will be more cars at rush-hour than in the middle of the night. But if I watched for a few hours in the middle of the day, then a Poisson process would give a decent approximation to what I observe.

Theorem 7.1. *If the process $(N(t) : t \geqslant 0)$ satisfies conditions (i) to (iii) above then it is a Poisson process of rate λ.*

The proof introduces a method which we will use several times in dealing with continuous-time stochastic processes.

Proof. Conditions (i) and (ii) of Definition 21 are immediate. The crux of the argument is to show that $N(t+s) - N(s) \sim \text{Po}(\lambda t)$ (condition (iii) of 21).

By the time-homogeneity condition, it suffices to prove this for $s = 0$. So we must show that $N(t) \sim \text{Po}(\lambda t)$ for all $t \geqslant 0$.

For each $k \in \mathbb{N}$, define a function $p_k : \mathbb{R}_{\geqslant 0} \to [0, 1]$ by

$$p_k(t) = \mathbb{P}(N(t) = k)$$

We will derive a system of differential equations for the functions p_k and then solve them to show that for all $k \in \mathbb{N}$ and all $t \geqslant 0$:

$$p_k(t) = e^{-\lambda t} \frac{(\lambda t)^k}{k!}$$

which says precisely that $N(t) \sim \text{Po}(\lambda t)$ for all $t \geqslant 0$.

We have

$$\begin{aligned} p_0(t+h) &= \mathbb{P}(N(t+h) = 0) \\ &= \mathbb{P}(N(t+h) = 0 \mid N(t) = 0) \cdot \mathbb{P}(N(t) = 0) \\ &= \mathbb{P}(\text{no arrivals in } [t, t+h]) \cdot \mathbb{P}(N(t) = 0) \\ &= (1 - \lambda h + o(h))\mathbb{P}(N(t) = 0) \\ &= (1 - \lambda h + o(h))p_0(t) \\ &= p_0(t) - \lambda h p_0(t) + o(h). \end{aligned}$$

Hence, rearranging, we have

$$\frac{p_0(t+h) - p_0(t)}{h} = -\lambda p_0(t) + \frac{o(h)}{h}.$$

Taking the limit of both sides as $h \to 0$ gives

$$p_0'(t) = -\lambda p_0(t).$$

(Where as usual, $p_0'(t)$ denotes $\frac{d}{dt}(p_0(t))$.)

Now let $k \geqslant 1$. We have

$$\begin{aligned} p_k(t+h) &= \mathbb{P}(N(t+h) = k) \\ &= \sum_{m=0}^{k} \mathbb{P}(N(t+h) = k \mid N(t) = m) \cdot \mathbb{P}(N(t) = m) \end{aligned}$$

Now (because of property (i) of $N(t)$) every term of this sum except the $m = k-1$ and $m = k$ ones are $o(h)$ so:

$$\begin{aligned} p_k(t+h) &= \mathbb{P}(N(t+h) = k \mid N(t) = k-1) \cdot \mathbb{P}(N(t) = k-1) \\ &\quad + \mathbb{P}(N(t+h) = k \mid N(t) = k) \cdot \mathbb{P}(N(t) = k) + o(h) \\ &= (\lambda h + o(h))p_{k-1}(t) + (1 - \lambda h + o(h))p_k(t) + o(h) \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(using property (i) of } N(t)) \\ &= \lambda h p_{k-1}(t) + p_k(t) - \lambda h p_k(t) + o(h) \end{aligned}$$

Rearranging, we have

$$\frac{p_k(t+h) - p_k(t)}{h} = \lambda p_{k-1}(t) - \lambda p_k(t) + \frac{o(h)}{h}.$$

Taking the limit of both sides as $h \to 0$ gives

$$p_k'(t) = \lambda p_{k-1}(t) - \lambda p_k(t).$$

Hence, we must solve the following system of differential equations:

$$\begin{aligned} p_0'(t) &= -\lambda p_0(t), \\ p_k'(t) &= \lambda p_{k-1}(t) - \lambda p_k(t) \quad \text{for all } k \geqslant 1. \end{aligned}$$

We do this inductively. First we solve the first equation (for p_0):

$$p_0'(t) = -\lambda p_0(t).$$

This has general solution $p_0(t) = Ae^{-\lambda t}$, where A is a constant of integration. Since $N(0) = 0$, we have $p_0(0) = \mathbb{P}(N(0) = 0) = 1$, so $A = 1$, and therefore $p_0(t) = e^{-\lambda t}$, which is the formula we want.

We can now solve the second differential equation, for p_1. We have

$$p_1'(t) = \lambda p_0(t) - \lambda p_1(t) = \lambda e^{-\lambda t} - \lambda p_1(t).$$

Rearranging, we have

$$p_1'(t) + \lambda p_1(t) = \lambda e^{-\lambda t}.$$

Now we use a trick: we can rewrite the left-hand side of the above equation as

$$e^{-\lambda t} \frac{d}{dt}(p_1(t)e^{\lambda t}).$$

So the equation can be rewritten as

$$\frac{d}{dt}(p_1(t)e^{\lambda t}) = \lambda.$$

Integrating both sides gives

$$p_1(t)e^{\lambda t} = \lambda t + C,$$

where C is a constant of integration. Since $N(0) = 0$, we have $p_1(0) = \mathbb{P}(N(0) = 1) = 0$, and therefore $C = 0$, so

$$p_1(t) = \lambda t e^{-\lambda t},$$

which is the formula we want.

Continuing in this manner, we get that every $p_k(t)$ has the required form. The formal way to write this is as a proof by induction on k. The idea is exactly the same as the argument we just saw to solve for $p_1(t)$ using our solution for $p_0(t)$.

Let $k \geqslant 1$ and assume that we have proved that

$$p_{k-1}(t) = e^{-\lambda t}\frac{(\lambda t)^{k-1}}{(k-1)!}.$$

Then the differential equation for $p_k(t)$ gives

$$p_k'(t) = \lambda p_{k-1}(t) - \lambda p_k(t) = \lambda\left(e^{-\lambda t}\frac{(\lambda t)^{k-1}}{(k-1)!}\right) - \lambda p_k(t).$$

Rearranging, we get

$$p_k'(t) + \lambda p_k(t) = e^{-\lambda t}\frac{\lambda^k t^{k-1}}{(k-1)!}.$$

We can rewrite the left-hand side of the above equation as

$$e^{-\lambda t}\frac{d}{dt}(p_k(t)e^{\lambda t}),$$

so the equation can be rewritten as

$$\frac{d}{dt}(p_k(t)e^{\lambda t}) = \frac{\lambda^k t^{k-1}}{(k-1)!}.$$

Integrating both sides gives

$$p_k(t)e^{\lambda t} = \frac{\lambda^k t^k}{k!} + C',$$

where C' is a constant of integration. Since $N(0) = 0$, we have $p_k(0) = \mathbb{P}(N(0) = k) = 0$, and therefore $C' = 0$. Hence,

$$p_k(t) = e^{-\lambda t}\frac{(\lambda t)^k}{k!}.$$

This proves (by induction on k) that

$$p_k(t) = e^{-\lambda t}\frac{(\lambda t)^k}{k!} \quad \text{for all } k \in \mathbb{N},$$

and therefore $N(t) \sim \mathrm{Po}(\lambda t)$, as required. □

7.3 Superposition and Thinning

We saw in Example 36 that if we start observing a Poisson process at some fixed time s then we see another Poisson process. In this section we see two more ways in which we can build new Poisson processes from old. These are called superposition (adding together two independent Poisson processes) and thinning (randomly filtering out events from a Poisson process).

Lemma 7.2 (The Superposition Lemma)**.** *Let $\lambda, \mu > 0$. Let $(A(t) : t \geqslant 0)$ be a Poisson process of rate λ, and let $(B(t) : t \geqslant 0)$ be a Poisson process of rate μ. Suppose that $(A(t) : t \geqslant 0)$ and $(B(t) : t \geqslant 0)$ are independent of one another. Then if $C(t) = A(t) + B(t)$, the stochastic process $(C(t) : \ t \geqslant 0)$ is a Poisson process of rate $\lambda + \mu$.*

We can prove this either by using the first definition 21 or the infinitesimal definition.

First proof using Definition 21. Since $A(0) = B(0) = 0$ we have that $C(0) = 0$ which is the first condition for a Poisson process.

For any $t_1 < t_2 < \cdots < t_n$, the family of random variables $A(t_{i+1}) - A(t_i)$, $B(t_{i+1}) - B(t_i)$ with $1 \leqslant i \leqslant n$ are mutually independent. Hence, $A(t_{i+1}) - A(t_i) + B(t_{i+1}) - B(t_i) = C(t_{i+1}) - C(t_i)$ are independent random variables for $1 \leqslant i \leqslant n$. This is the third condition for a Poisson process.

The heart of the proof is showing condition (iii) that $C(s+t) - C(s) \sim \text{Po}((\lambda + \mu)t)$. This boils down to calculating the probability that the process $C(t)$ has k arrivals in the interval $[s, s+t]$. For this to happen we must have m arrivals from process A and $k - m$ arrivals from process B. So

$$\begin{aligned}
\mathbb{P}(C(s+t) - C(s) = k) &= \sum_{m=0}^{k} \mathbb{P}(A(s+t) - A(s) = m, B(s+t) - B(s) = k - m) \\
&= \sum_{m=0}^{k} \mathbb{P}(A(s+t) - A(s) = m)\mathbb{P}(B(s+t) - B(s) = k - m) \\
&\qquad \text{(as the processes } A \text{ and } B \text{ are independent)} \\
&= \sum_{m=0}^{k} e^{-\lambda t} \frac{(\lambda t)^m}{m!} e^{-\mu t} \frac{(\mu t)^{k-m}}{(k-m)!} \\
&= \frac{e^{-\lambda t - \mu t}}{k!} \sum_{m=0}^{k} (\lambda t)^m (\mu t)^{k-m} \frac{k!}{m!(k-m)!} \\
&= \frac{e^{-(\lambda+\mu)t}}{k!} (\lambda t + \mu t)^k \qquad \text{(by the Binomial Theorem)} \\
&= \frac{e^{-(\lambda+\mu)t}}{k!} ((\lambda + \mu)t)^k .
\end{aligned}$$

This is precisely the pmf of a $\text{Po}((\lambda + \mu))t$ random variable as required. □

Second proof using the infinitesimal definition. Properties (ii) and (iii) of the infinitesimal definition are inherited from the processes A and B.

We need to show that $\mathbb{P}(C(t+h) = n \mid C(t) = m)$ has the right form for small h. This boils down to showing that

$$\mathbb{P}(C(t+h)-C(t) = k) = \mathbb{P}(C \text{ has } k \text{ arrivals in } [t, t+h]) = \begin{cases} 1-(\lambda+\mu)h + o(h) & \text{if } k = 0; \\ (\lambda+\mu)h + o(h) & \text{if } k = 1; \\ o(h) & \text{if } k > 1. \end{cases}$$

Now

$$\begin{aligned} \mathbb{P}(C \text{ has no arrivals in } [t, t+h]) &= \mathbb{P}(A \text{ has no arrivals, } B \text{ has no arrivals in } [t, t+h]) \\ &= \mathbb{P}(A \text{ has no arrivals})\mathbb{P}(B \text{ has no arrivals in } [t, t+h]) \\ &\qquad\qquad \text{(as the processes } A \text{ and } B \text{ are independent)} \\ &= (1-\lambda h + o(h))(1-\mu h + o(h)) \\ &= 1 - \lambda h - \mu h + \lambda\mu h^2 + o(h) \\ &= 1 - (\lambda+\mu)h + o(h) \qquad\qquad (\text{since } h^2 = o(h)) \end{aligned}$$

Similarly,

$$\begin{aligned} \mathbb{P}(C \text{ has 1 arrival in } [t, t+h]) &= \mathbb{P}(A \text{ has 1 arrival, } B \text{ has 0 arrivals in } [t, t+h]) \\ &\quad + \mathbb{P}(A \text{ has 0 arrivals, } B \text{ has 1 arrival in } [t, t+h]) \\ &= (\lambda h + o(h))(1-\mu h + o(h)) + (1-\lambda h + o(h))(\mu h + o(h)) \\ &= \lambda h - \lambda\mu h^2 + o(h) + \mu h - \lambda\mu h^2 + o(h) \\ &= (\lambda+\mu)h + o(h) \qquad\qquad (\text{since } h^2 = o(h)) \end{aligned}$$

Finally,

$$\begin{aligned} \mathbb{P}(C \text{ has at least 2 arrivals in } [t, t+h]) &= \mathbb{P}(A \text{ has 1 arrival, } B \text{ has 1 arrival in } [t, t+h]) \\ &\quad + \mathbb{P}(\text{at least one of } A \text{ and } B \text{ has at least 2 arrivals in } [t, t+h]) \\ &= (\lambda h + o(h))(\mu h + o(h)) + o(h) \\ &= \lambda\mu h^2 + o(h) \\ &= o(h) \qquad\qquad (\text{since } h^2 = o(h)) \end{aligned}$$

□

Different people will probably prefer one or other of these proofs. The first one is a bit shorter, but the second is maybe conceptually simpler. It can be a good thing to have several proofs of the same result for several reasons. It can help us see why something is true in a more complete way. Also, it is common in maths to adapt a proof to a different situation, and this might be easier with one proof than the other. Having a variety of proofs gives me the best chance that one of them will help in my new situation.

Lemma 7.3 (The Thinning Lemma)**.** *Let $\lambda > 0$ and let $0 < p < 1$. Let $(X(t) : t \geqslant 0)$ be a Poisson process of rate λ which counts arrivals occuring in $\mathbb{R}_{\geqslant 0}$. Suppose that each arrival is deleted with probability $1 - p$ (and survives with probability p) independently of all other events. Let $Y(t)$ denote the number of events which survive between time 0 and time t. Then $(Y(t) : t \geqslant 0)$ is a Poisson process of rate $p\lambda$.*

Again this can be proved using either of the definitions. This time I will just give the proof using Definition 21. Why not have a go at giving a proof using the infinitesimal definition?

Proof. Clearly $Y(0) = 0$ so the first condition is satisfied.

For any $0 \leqslant t_1 < t_2 < t_3 < \ldots < t_n$, mutual independence of the random variables $Y(t_{i+1}) - Y(t_i)$ follows from mutual independence of the $X(t_{i+1}) - X(t_i)$ and the independence of the thinning procedure.

The heart of the proof is showing condition (iii) that $Y(s+t) - Y(s) \sim \text{Po}(p\lambda t)$. This boils down to calculating the probability that the process $Y(t)$ has k arrivals in the interval $[s, s+t]$. For this to happen we must have $m \geqslant k$ arrivals from the process X in the interval $[s, s+t]$ and for exactly k of them to survive the thinning procedure. Conditional on there being m arrivals from X, the number of arrivals surviving the thinning procedure is distributed $\text{Bin}(m, p)$. So

$$\begin{aligned}
\mathbb{P}(Y(s+t) - Y(s) = k) &= \sum_{m=k}^{\infty} \mathbb{P}(X(s+t) - X(s) = m)\mathbb{P}(Y(s+t) - Y(s) = k \mid X(s+t) - X(s) = m) \\
&= \sum_{m=k}^{\infty} e^{-\lambda t} \frac{(\lambda t)^m}{m!} \binom{m}{k} p^k (1-p)^{m-k} \\
&= e^{-\lambda t} p^k \sum_{m=k}^{\infty} \frac{(\lambda t)^m}{m!} \frac{m!}{k!(m-k)!} (1-p)^{m-k} \\
&= e^{-\lambda t} p^k \frac{(\lambda t)^k}{k!} \sum_{m=k}^{\infty} \frac{(\lambda t)^{m-k}(1-p)^{m-k}}{(m-k)!} \\
&= e^{-\lambda t} \frac{(p\lambda t)^k}{k!} \sum_{i=0}^{\infty} \frac{(\lambda t(1-p))^i}{i!} \\
&\qquad \text{(relabelling dummy variable as } i = m - k\text{)} \\
&= e^{-\lambda t} \frac{(p\lambda t)^k}{k!} e^{\lambda t(1-p)} \qquad \text{(using power series for } e^x\text{)} \\
&= \frac{(p\lambda t)^k}{k!} e^{-p\lambda t}
\end{aligned}$$

This is precisely the pmf of a $\text{Po}(p\lambda t)$ random variable as required. □

Example 37. Suppose that I receive email from students as a Poisson process of rate 1 per hour[14] and email from others as a Poisson process of rate 2 per hour. Suppose also that for each message I receive there is a 1/2 chance that it needs a reply independently of who it came from and all other messages.

In this case the total number of emails I receive forms a Poisson process of rate 3 per hour (by Lemma 7.2) and the total number of emails which need replies forms a Poisson process of rate 3/2 (using Lemma 7.3 applied to the process counting the total emails).

7.4 Arrival and Interarrival Times

Reminder: Continuous Random Variables

Recall that if R is a continuous random variable then we can't just describe it by giving its probabliity mass function (because $\mathbb{P}(R = x) = 0$ for any x). Instead we use

- The cumulative distribution function (cdf) which is defined by:

$$F_R(t) = \mathbb{P}(R \leqslant t)$$

- The probability density function (pdf) which satisfies:

$$\mathbb{P}(a \leqslant R \leqslant b) = \int_a^b f_R(t)\, dt$$

- The useful fact that the cdf and pdf are related by:

$$f_R = F_R'(t).$$

Here is one example which we will use a lot. A continuous random variable R has the Exponential distribution with parameter μ if

$$F_R(t) = 1 - e^{-\mu t} \quad \text{for } t \geqslant 0.$$

Differentiating we get the pdf of an exponentially distributed random variable to be

$$f_R(t) = \begin{cases} \mu e^{-\mu t} & \text{if } t \geqslant 0; \\ 0 & \text{if } t < 0. \end{cases}$$

We write $R \sim \text{Exp}(\mu)$. You can check (try it) that $\mathbb{E}(R) = 1/\mu$.

The exponential distribution has the following 'memoryless' property.

$$\mathbb{P}(R > s + t \mid R > s) = \mathbb{P}(R > t) \quad \text{for any } s, t \geqslant 0.$$

Exercise: prove this using the definition of conditional probability and the exponential distribution cdf.

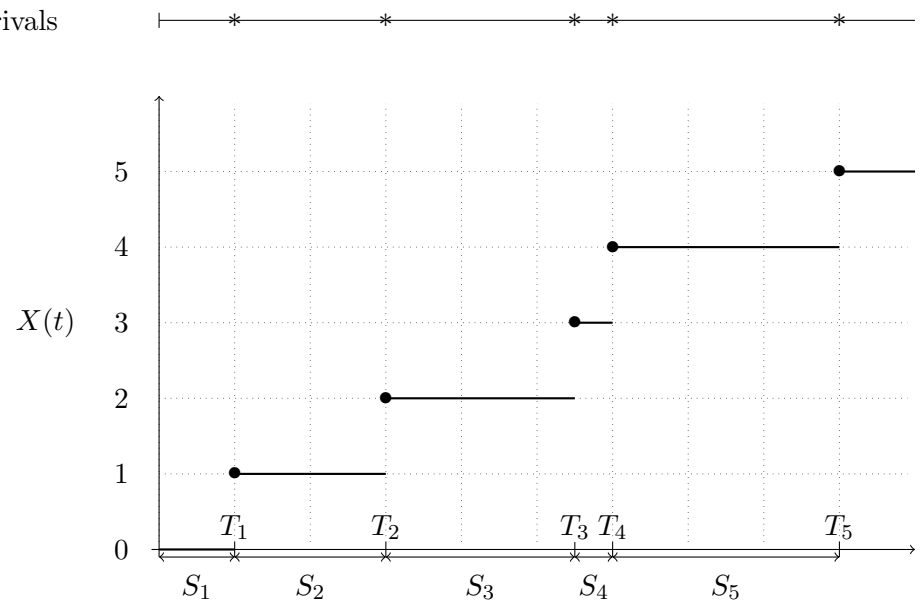
[14] Formally, the set up is as in Example 35. If we let $S(t)$ be the number of emails received from students up to time t then $S(t)$ is a Poisson process of rate 1 per hour.

Random Variables Associated with the Poisson Process

Definition 22. Suppose that $(X(t) : t \geqslant 0)$ is a Poisson process. We define

- the *arrival times*
$$T_i = \min\{t : X(t) = i\}$$
- the *interarrival times* or *sojourn times*
$$S_1 = T_1$$
$$S_i = T_i - T_{i-1} \quad \text{for } i \geqslant 2$$

The T_i give the times at which we have an arrival; the interarrival times give the time between consecutive arrivals.



Note that
$$T_n = S_1 + S_2 + \ldots + S_n.$$
So if I tell you the S_i you can find the T_i and vice versa. Also, either of the S_i and the T_i determines the evolution of the process.

Distribution of the Interarrival Times

Now, $T_1 = S_1$ and we can work out the cdf as

$$F_{S_1}(t) = \mathbb{P}(S_1 \leqslant t) = 1 - \mathbb{P}(\text{no arrival in } [0,t]) = 1 - e^{-\lambda t}.$$

So $S_1 \sim \text{Exp}(\lambda)$.

More generally,

Theorem 7.4. *If $(X(t) : t \geqslant 0)$ is a Poisson process of rate λ then the interarrival times S_i are independent random variables and each is distributed* $\text{Exp}(\lambda)$.

The memoryless property of the exponential distribution means that if we are observing a Poisson process at time s, then the time we have to wait for the next arrival is distributed $\text{Exp}(\lambda)$ regardless of how long we have been waiting since the previous arrival.

Sketch of proof. For a fixed time s we have

$$\mathbb{P}(\text{time from } s \text{ to next arrival} \leqslant t) = 1 - \mathbb{P}(\text{no arrivals in } [s, s+t]) = 1 - e^{-\lambda t}$$

Similarly,

$$\mathbb{P}(S_n \leqslant t) = 1 - \mathbb{P}(\text{no arrivals in } (T_{n-1}, T_{n-1} + t]) = 1 - e^{-\lambda t}$$

So $S_n \sim \text{Exp}(\lambda)$. Moreover, this does not depend on $S_1, \ldots, S_{n-1}$ so the S_i are independent.

I have called this a sketch because the argument is not quite rigorous. The definition of the Poisson process (property (iii)) only tells us about the distribution of the number of arrivals in a fixed interval so it is not valid to use it for the interval $(T_{n-1}, T_{n-1} + t]$ whose endpoints are random variables. This can be fixed with some technicalities but we will just go with this version which conveys the spirit of the proof. □

Three different ways to think of the Poisson process

This property of the interarrival times could be used to give a third equivalent definition of the Poisson process although I won't write that definition down in detail. The thing to remember is that any one of the following three statements captures the key idea of the Poisson process:

- $X(s+t) - X(s) \sim \text{Po}(t\lambda)$
- The probability of arrivals in a short time interval of length h satisfies:

$$\mathbb{P}(k \text{ arrivals in the interval } [t, t+h]) = \begin{cases} \lambda h + o(h) & \text{if } k = 1; \\ 1 - \lambda h + o(h) & \text{if } k = 0; \\ o(h) & \text{if } k > 1. \end{cases}$$

- The interarrival times are independent and distributed $\text{Exp}(\lambda)$

For the first two we also need to insist that the number of arrivals in disjoint intervals are mutually independent (see property ii of our original definition of the Poisson process 21).

Distribution of the Arrival Times

What about the arrival times?

Theorem 7.5. *Let $(X(t) : t \geqslant 0)$ be a Poisson process of rate λ. Then for each $n \geqslant 1$, the nth arrival time T_n has pdf*

$$f_{T_n}(t) = \begin{cases} e^{-\lambda t}\frac{\lambda^n t^{n-1}}{(n-1)!} & \text{if } t \geqslant 0; \\ 0 & \text{if } t < 0. \end{cases}$$

(This is called the Gamma distribution with parameters n and λ.)

I will give two proofs of this. In the first we figure out their distribution via the cdf as for T_1 using the original definition of the Poisson process.

Proof. Clearly $f_{T_n}(t) = 0$ for all $t < 0$. Suppose that $t \geqslant 0$. We have

$$\begin{aligned} F_{T_n} = \mathbb{P}(T_n \leqslant t) &= 1 - \sum_{k=0}^{n-1} \mathbb{P}(k \text{ arrivals in } [0,t]) \\ &= 1 - \sum_{k=0}^{n-1} e^{-\lambda t}\frac{(\lambda t)^k}{k!} \\ &= 1 - e^{-\lambda t} - \sum_{k=1}^{n-1} e^{-\lambda t}\frac{(\lambda t)^k}{k!} \end{aligned}$$

Differentiating:

$$\begin{aligned} f_{T_n}(t) = F'_{T_n} &= \lambda e^{-\lambda t} - \sum_{k=1}^{n-1}\left(-\lambda e^{-\lambda t}\frac{(\lambda t)^k}{k!} + e^{-\lambda t}\frac{\lambda^k t^{k-1}}{(k-1)!}\right) \\ &= \lambda e^{-\lambda t} + \sum_{k=1}^{n-1}\left(e^{-\lambda t}\frac{\lambda^{k+1} t^k}{k!}\right) - \sum_{k=1}^{n-1}\left(e^{-\lambda t}\frac{\lambda^k t^{k-1}}{(k-1)!}\right) \\ &= \lambda e^{-\lambda t} + \sum_{k=1}^{n-1}\left(e^{-\lambda t}\frac{\lambda^{k+1} t^k}{k!}\right) - \sum_{j=0}^{n-2}\left(e^{-\lambda t}\frac{\lambda^{j+1} t^j}{j!}\right) \end{aligned}$$

Most of the terms in these two sums cancel out and we are left with just

$$f_{T_n}(t) = \lambda e^{-\lambda t} + e^{-\lambda t}\frac{\lambda^n t^{n-1}}{(n-1)!} - e^{-\lambda t}\frac{\lambda^1 t^0}{0!}$$

(from the $k = n-1$ term of the first sum and the $j = 0$ term of the second sum)

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

as required. □

We can also prove this Theorem via the infinitesimal definition. As before different people will prefer different proofs. It is a good idea to get comfortable with using both definitions of the Poisson process and seeing both proofs should help with this.

Second proof of 7.5 using the infinitesimal definition of Poisson process. Let $n \in \mathbb{N}$. Consider the cdf $F_{T_n}(t)$ of T_n. Let $t > 0$. We have

$$\begin{aligned}
F_{T_n}(t+h) - F_{T_n}(t) &= \mathbb{P}(T_n \leqslant t+h) - \mathbb{P}(T_n \leqslant t) \\
&= \mathbb{P}(T_n \in (t, t+h]) \\
&= \sum_{k=0}^{n-1} \mathbb{P}(X(t) = k)\mathbb{P}(\text{at least } n-k \text{ arrivals in } (t, t+h]) \\
&= \mathbb{P}(X(t) = n-1)\mathbb{P}(\text{one arrival in } (t, t+h]) + o(h) \\
&\quad \text{(since the probability of more than 1 arrival in } (t, t+h] \text{ is } o(h)) \\
&= e^{-\lambda t}\frac{(\lambda t)^{n-1}}{(n-1)!}(\lambda h + o(h)) + o(h) \\
&= h e^{-\lambda t}\frac{\lambda^n t^{n-1}}{(n-1)!} + o(h).
\end{aligned}$$

Dividing both sides by h gives

$$\frac{F_{T_n}(t+h) - F_{T_n}(t)}{h} = e^{-\lambda t}\frac{\lambda^n t^{n-1}}{(n-1)!} + \frac{o(h)}{h}.$$

Taking the limit as $h \to 0$ gives

$$\begin{aligned}
f_{T_n}(t) &= \frac{d}{dt}F_{T_n}(t) \\
&= \lim_{h\to 0}\left[\frac{F_{T_n}(t+h) - F_{T_n}(t)}{h}\right] \\
&= \lim_{h\to 0}\left[e^{-\lambda t}\frac{\lambda^n t^{n-1}}{(n-1)!} + \frac{o(h)}{h}\right] \\
&= e^{-\lambda t}\frac{\lambda^n t^{n-1}}{(n-1)!}.
\end{aligned}$$

Clearly, we have $f_{T_n}(t) = 0$ for all $t < 0$. Hence, T_n has the pdf claimed. □

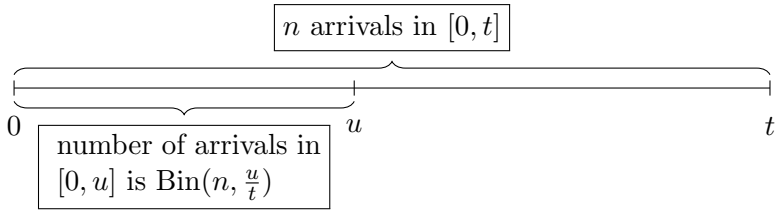
7.5 Conditioning on $X(t) = n$

Suppose that we know $X(t) = n$. This means that there are exactly n arrivals in the intreval $[0, t]$. What can we say about when those n arrivals occured?

Theorem 7.6. *Let $(X(t) : t \geqslant 0)$ be a Poisson process of rate λ. Let $t > 0$, and let $n \in \mathbb{N}$. Then for any $u \in [0, t]$, the conditional distribution of $X(u)$, conditioned on $X(t) = n$, is* $\mathrm{Bin}(n, \frac{u}{t})$. *That is*

$$\mathbb{P}(X(u) = k \mid X(t) = n) = \binom{n}{k} (\tfrac{u}{t})^k (1 - \tfrac{u}{t})^{n-k} \quad \textit{for any } k \in \{0, 1, 2, \ldots, n\}.$$

The parameter $\frac{u}{t}$ is the proportion of the big interval $[0, t]$ taken up by the small interval $[0, u]$ as in this picture:



Example 38. I receive email as a Poisson process of rate 3 per hour. I receive 5 messages between 8am and 10am. What is the probability that I received 3 of them between 8am and 9am?

I start counting emails at 8am and write $X(t)$ for the number of messages I receive after I have been counting for t hours. The question tells us that $X(2) = 5$ and asks for

$$\mathbb{P}(X(1) = 3 \mid X(2) = 5)$$

By the previous Theorem, the conditional distribution is $\mathrm{Bin}(5, 1/2)$ so

$$\mathbb{P}(X(1) = 3 \mid X(2) = 5) = \binom{5}{3} \left(\frac{1}{2}\right)^3 \left(\frac{1}{2}\right)^2 = \frac{10}{32} = \frac{5}{16}.$$

Proof of Theorem 7.6. By the definition of conditional probability we have

$$\begin{aligned}
\mathbb{P}(X(u)=k \mid X(t)=n) &= \frac{\mathbb{P}(X(u)=k \text{ and } X(t)=n)}{\mathbb{P}(X(t)=n)} \\
&= \frac{\mathbb{P}(X(u)-X(0)=k \text{ and } X(t)-X(u)=n-k)}{\mathbb{P}(X(t)=n)} \\
&= \frac{\mathbb{P}(X(u)-X(0)=k)\cdot\mathbb{P}(X(t)-X(u)=n-k)}{\mathbb{P}(X(t)-X(0)=n)} \\
&\qquad (X(u)-X(0) \text{ and } X(t)-X(u) \text{ are independent of one another}) \\
&= \frac{e^{-\lambda u}\frac{(\lambda u)^k}{k!}\cdot e^{-\lambda(t-u)}\frac{(\lambda(t-u))^{n-k}}{(n-k)!}}{e^{-\lambda t}\frac{(\lambda t)^n}{n!}} \\
&= \frac{e^{-\lambda u}e^{-\lambda(t-u)}}{e^{-\lambda t}}\frac{\lambda^k\lambda^{n-k}}{\lambda^n}\frac{n!}{k!(n-k)!}\frac{u^k(t-u)^{n-k}}{t^n} \\
&= \binom{n}{k}(\tfrac{u}{t})^k(\tfrac{t-u}{t})^{n-k} \\
&= \binom{n}{k}(\tfrac{u}{t})^k(1-\tfrac{u}{t})^{n-k}
\end{aligned}$$

as required. □

One consequence of this Theorem is that

$$\mathbb{P}(T_1 \leqslant u \mid X(t)=1) = \mathbb{P}(X(u)=1 \mid X(t)=1) = \frac{u}{t}.$$

So we have proved:

Corollary 7.7. *Let $(X(t) : t \geqslant 0)$ be a Poisson process of rate λ. Let $t > 0$. Then conditional on $X(t) = 1$, the first arrival time T_1 is uniformly distributed on $[0, t]$. That is the conditional cdf and pdf are*

$$F_{T_1|X(t)=1}(u) = \frac{u}{t}, \quad f_{T_1|X(t)=1}(u) = \begin{cases} \frac{1}{t} & \text{if } 0 \leqslant u \leqslant t; \\ 0 & \text{otherwise.} \end{cases}.$$

Similarly, suppose that $X(t) = n$, what can we say about the conditional distribution of $T_1, \ldots, T_n$?

Theorem 7.8. *Let $(X(t) : t \geqslant 0)$ be a Poisson process of rate λ. Let $t > 0$ and let $n \in \mathbb{N}$. Suppose we condition on $X(t) = n$. Then the (conditional) joint distribution of the arrival times $(T_1, T_2, \ldots, T_n)$ can be found as follows. Let $U_1, U_2, \ldots, U_n$ be mutually independent random variables which are each uniformly distributed on $[0, t]$. Take the permutation π of $\{1, 2, \ldots, n\}$ such that*

$$U_{\pi(1)} < U_{\pi(2)} < \ldots < U_{\pi(n)},$$

(so the permutation π puts the random variables $U_1, U_2, \ldots, U_n$ in increasing order of the values they take). Then the (conditional) joint distribution of the arrival times $(T_1, T_2, \ldots, T_n)$ is identical to the joint distribution of $(U_{\pi(1)}, U_{\pi(2)}, \ldots, U_{\pi(n)})$.

This result takes a bit of getting your head around, but once you have digested what is says it should look very plausible. We won't give a proof.

A very useful feature of this result is that if we are interested in some function of the arrival times which is not changed by permuting them, then we can replace the T_i with the U_i.

Corollary 7.9. *Let $T_1, T_2, \ldots$ be the arrival times of a Poisson process. Let $r(T_1, \ldots, T_n)$ be a symmetric function of n variables*[15]*. Then*

$$\mathbb{E}(r(T_1, \ldots, T_n) \mid X(t) = n) = \mathbb{E}(r(U_1, \ldots, U_n))$$

where $U_1, \ldots, U_n$ are independent random variables with $U_i \sim U[0, t]$.

Example 39. Let $T_1, T_2, \ldots$ be the arrival times of a Poisson process (of any rate λ). Let $U_1, \ldots, U_n$ be independent random variables each distributed $U[0, t]$.

- $\mathbb{E}(T_1 + T_2 + \cdots + T_n \mid X(t) = n) = \mathbb{E}(U_1 + U_2 + \cdots U_n) = \frac{nt}{2}$ (as $\mathbb{E}(U_i) = \frac{t}{2}$)
- $\mathbb{E}(T_1 T_2 \cdots T_n \mid X(t) = n) = \mathbb{E}(U_1 U_2 \cdots U_n) = \mathbb{E}(U_1)\mathbb{E}(U_2) \cdots \mathbb{E}(U_n) = \left(\frac{t}{2}\right)^n$ (as the U_i are independent).

Example 40. People arrive in a waiting room according to a Poisson process of rate λ per minute. I check in the waiting room after it has been open for 30 minutes and see that there are 8 people waiting (and nobody has left). What is the expectation of the total time spent waiting?

Let $X(t)$ be the process counting the arrivals in the waiting room. We are told that $X(t)$ is a Poisson process of rate λ and that $X(30) = 8$.

If a person arrives at time u minutes then when I check the room they have been waiting for $30 - u$ minutes. So we need to calculate

$$\mathbb{E}((30 - T_1) + (30 - T_2) + \cdots + (30 - T_8) \mid X(30) = 8)$$

[15] That is for any permutation π of $\{1, 2, \ldots, n\}$ and any $x_1, \ldots, x_n$ we have $r(x_1, x_2, \ldots, x_n) = r(x_{\pi(1)}, x_{\pi(2)}, \ldots, x_{\pi(n)})$; the value of r is unchanged if we permute the inputs.

But this function is unchanged by permuting the T_i so we can replace them by the U_i.

$$\begin{aligned}\mathbb{E}((30-T_1)+(30-T_2)+\cdots+(30-T_8)\mid X(30)=8) &= \mathbb{E}((30-U_1)+\cdots+(30-U_8))\\ &= 240-\mathbb{E}(U_1)-\cdots-\mathbb{E}(U_8)\\ &\qquad\text{(by linearity of exectation)}\\ &= 240-8\times 15 \ \ (\text{as } U_i\sim U[0,30])\\ &= 120\end{aligned}$$

Notice that this does not depend on the rate λ because conditioning on $X(30)=8$ has removed its significance.

Similarly, if I know there are m people in the room at time 30 then the total expected waiting time is

$$\begin{aligned}\mathbb{E}((30-T_1)+(30-T_2)+\cdots+(30-T_m)\mid X(30)=m) &= \mathbb{E}((30-U_1)+\cdots+(30-U_m))\\ &= 30m-15m=15m\end{aligned}$$

Suppose that I don't know how many people are in the waiting room. What can we say about the expected total waiting time of the people waiting when it has been open for 30 minutes. Conditioning on $X(30)$ we get that this is

$$\sum_{m=0}^{\infty}\mathbb{E}((30-T_1)+(30-T_2)+\cdots+(30-T_m)\mid X(30)=m)\mathbb{P}(X(30)=m)$$

Substituting in the result from above this is

$$\begin{aligned}\sum_{m=0}^{\infty}15me^{-30\lambda}\frac{(30\lambda)^m}{m!} &= 15\sum_{m=0}^{\infty}me^{-30\lambda}\frac{(30\lambda)^m}{m!}\\ &= 15\times 30\lambda\\ &\text{(since the sum is just the expectation of a Po}(30\lambda)\text{ random variable)}\\ &= 450\lambda\end{aligned}$$

8 Birth Processes

8.1 Motivating example

Suppose we have a single yeast cell in a large vat with plenty of space and food. We can model the growth of the population by assuming that each cell in the population reproduces (asexually) by generating offspring according to a Poisson process of rate λ independently of all the other cells. Let $X(t)$ be the number of cells in the population at time t. Let's assume that no cells die and ignore the fact that our population may be limited by lack of food or space (these will be reasonable assumptions at least in the early stages of growth).

The process $(X(t) : t \geqslant 0)$ is not itself a Poisson process; if there are more cells in the population at time t then we expect the arrivals (or births) to happen at a faster rate than if there were fewer cells.

Our aim is to modify the definition of the Poisson process to include the feature that the arrival rate depends on $X(t)$. This will give us the concept of a birth process.

Recall our three ways of thinking of the Poisson process:

- $X(s+t) - X(s) \sim \text{Po}(t\lambda)$ and the number of arrivals in disjoint intervals are independent.
- The probability of an arrival in a short interval of length h satisfies:

$$\mathbb{P}(k \text{ arrivals in the interval } [t, t+h]) = \begin{cases} \lambda h + o(h) & \text{if } k = 1; \\ 1 - \lambda h + o(h) & \text{if } k = 0; \\ o(h) & \text{if } k > 1. \end{cases}$$

  independently of the process up to time t.
- The interarrival times are independent and distributed $\text{Exp}(\lambda)$

It turns out that the second and third of these can be adapted to involve a parameter that depends on $X(t)$ rather than a single λ. This will give us two ways of thinking of a birth process.

This shows the power of having several equivalent definitions of a concept; if gives us more options for extending or generalising it.

8.2 Infinitesimal definition of a birth process

If we take the infinitesimal definition of the Poisson process and allow the rate of arrivals in the small interval $[t, t+h]$ to depend on $X(t)$ we get:

Definition 23. A continuous-time stochastic process $(X(t) : t \geqslant 0)$ is a *birth process* with *birth parameters* $\lambda_s, \lambda_{s+1}, \lambda_{s+2}, \ldots > 0$ if it satisfies the three properties:

(i) $X(0) = s \geqslant 0$

(ii) For any $m \geqslant s$,

$$\mathbb{P}(X(t+h) = n \mid X(t) = m) = \begin{cases} \lambda_m h + o(h) & \text{if } n = m+1; \\ 1 - \lambda_m h + o(h) & \text{if } n = m; \\ o(h) & \text{if } n \geqslant m+2; \\ 0 & \text{if } n < m. \end{cases}$$

(iii) If $0 < a < b$, then conditional on $X(a)$, the random variable $X(b) - X(a)$ is independent of the process before time a (that is, independent of $(X(u) : 0 \leqslant u < a)$).

Often $X(t)$ will represent the size of a population at time t. In this case, we are allowing the birth rate to depend on the current size of the population.

Example 41. A Poisson process of rate λ is precisely a birth process with $X(0) = 0$ and with birth parameters given by $\lambda_n = \lambda$ for all $n \in \mathbb{N}$.

Example 42 (The 'Linear' Birth Process (also called the 'Yule' Birth Process))**.** Let $\lambda > 0$. A *Linear Birth Process* with birth-rate λ is a birth process with $X(0) = 1$ and birth parameters given by $\lambda_n = n\lambda$ for all $n \in \mathbb{N}$.

Think back to our yeast cell example from the previous section. We will see that the assumption there that each cell generates births as a Poisson process leads to a population which grows as a linear birth process.

We have:

$$\begin{aligned} \mathbb{P}(X(t+h) = n \mid X(t) = n) &= \mathbb{P}(\text{no births in the interval } [t, t+h] \mid X(t) = n) \\ &= (1 - \lambda h + o(h))^n \\ &\quad \text{(since each of the } n \text{ cells reproduces according to a Poisson process of rate } \lambda) \\ &= 1 - n\lambda h + o(h) \quad \text{(all other terms are order } h^2 \text{ or smaller)} \end{aligned}$$

Also,

$$\begin{aligned}\mathbb{P}(X(t+h)=n+1 \mid X(t)=n) &= \mathbb{P}(\text{one birth in the interval } [t,t+h] \mid X(t)=n)\\ &= (\text{number of possibilities for which cell reproduces})\\ &\times \mathbb{P}(\text{that cell reproduces}) \times \mathbb{P}(\text{other } n-1 \text{ cells don't reproduce})\\ &= n(\lambda h + o(h))(1-\lambda h + o(h))^{n-1}\\ &= n\lambda h + o(h) \qquad (\text{all other terms are order } h^2 \text{ or smaller})\end{aligned}$$

Similarly,

$$\mathbb{P}(X(t+h) \geqslant n+2 \mid X(t)=n) = \mathbb{P}(\text{more than one birth in the interval } [t,t+h] \mid X(t)=n) = o(h)$$

So this is a linear birth process; it follows the definition of a birth process with $\lambda_n = n\lambda$.

8.3 Arrival and Interarrival times

We can define arrival and interarrival times almost exactly as we do for the Poisson process. The only small difference is that we allowed out birth process to start with $X(0) = s$ (rather than $X(0) = 0$ for the Poisson process) so we need to be a bit careful about how to label the arrivals.

Definition 24. Let $(X(t) : t \geqslant 0)$ be a birth process with $X(0) = s \in \mathbb{N}$. We define

- the *arrival times*
$$T_i = \min\{t : X(t) = i\} \quad \text{for } i \geqslant s+1$$
- the *interarrival times* or *sojourn times*
$$\begin{aligned} S_{s+1} &= T_{s+1}\\ S_i &= T_i - T_{i-1} \quad \text{for } i \geqslant s+2\end{aligned}$$

In other words, if $X(t)$ is the size of a certain population at time t, then the continuous random variable T_n is the time at which the size of the population first reaches n. The random variable S_n is the time the population spends at size exactly $(n-1)$. Equivalently, T_n is the time of the $(n-s)$th birth and S_n is the length of time between the $(n-s-1)$th birth and the $(n-s)$th birth.

Similarly to a Poisson process, it turns out that the interarrival times are mutually independent exponential random variables, but (unlike with the Poisson process), their parameters may differ: S_n is an exponential random variable with parameter λ_{n-1}, the birth parameter corresponding to when there are $n-1$ individuals in the population. This is the content of the following useful theorem which is the analogue of Theorem 7.4 for Birth processes.

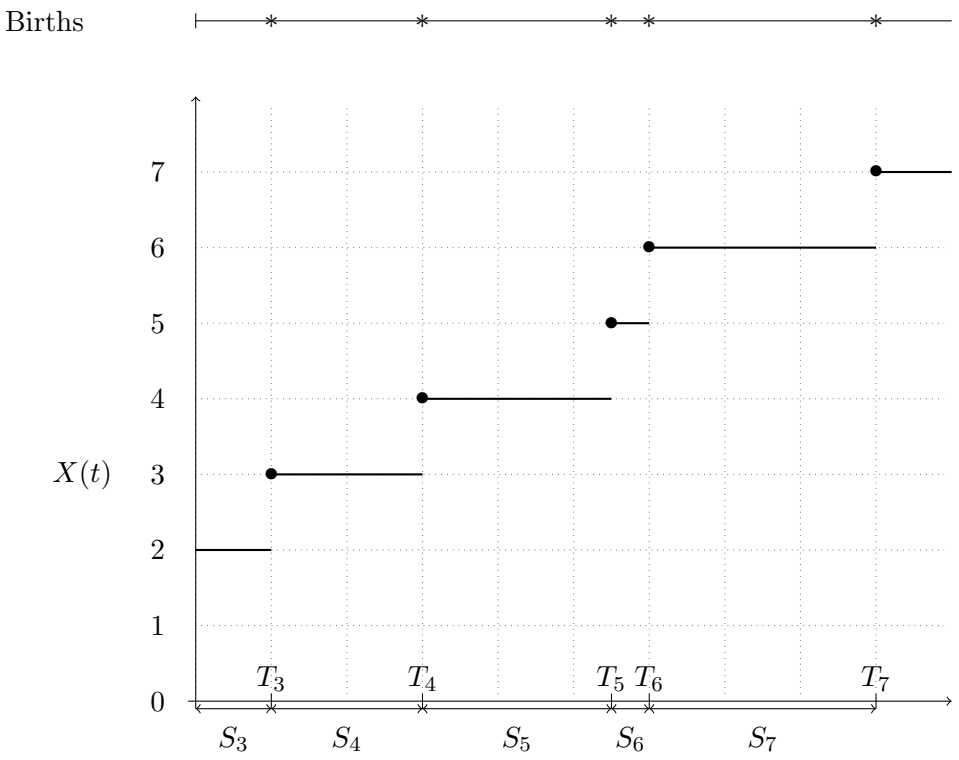
Theorem 8.1. *Let $(X(t) : t \geqslant 0)$ be a birth process with $X(0) = s \in \mathbb{N}$ and birth parameters $\lambda_0, \lambda_1, \lambda_2, \ldots$. Then the interarrival times $S_{s+1}, S_{s+2}, S_{s+3}, \ldots$ are mutually independent random variables with $S_n \sim \text{Exp}(\lambda_{n-1})$ for all $n \geqslant s+1$.*

We won't prove this but it should look very plausible. From T_{n-1} until the time of the next birth the process behaves like a Poisson process of rate λ_{n-1}, so it seems reasonable that S_n should have the same distribution as the interarrival times in such a process, namely $\text{Exp}(\lambda_{n-1})$.

There is no similarly nice expression for the distribution of the T_i. However, using linearity of expectation and the fact that $\mathbb{E}(S_n) = \frac{1}{\lambda_{n-1}}$ (by properties of the exponential distribution), we can say that if $n \geqslant s+1$

$$\mathbb{E}(T_n) = \mathbb{E}(S_{s+1} + S_{s+2} + \cdots + S_n) = \frac{1}{\lambda_s} + \frac{1}{\lambda_{s+1}} + \cdots + \frac{1}{\lambda_{n-1}}$$

Here is a figure to illustrate the T_i and S_i in the situation $X(0) = 2$ (so we start with 2 individuals in our population).



8.4 Differential Equations for Birth Process

Just as for the Poisson process we can analyse the birth process using differential equations.

Theorem 8.2. *Let $(X(t) : t \geqslant 0)$ be a birth process with $X(0) = s \in \mathbb{N}$ and birth parameters $\lambda_s, \lambda_{s+1}, \lambda_{s+2}, \ldots$. For each $k \geqslant s$, define a function $p_k : \mathbb{R}_{\geqslant 0} \to [0,1]$ by*

$$p_k(t) = \mathbb{P}(X(t) = k).$$

Then the functions $p_k(t)$ satisfy the following system of differential equations.

$$\begin{aligned} p_s'(t) &= -\lambda_s p_s(t), \\ p_k'(t) &= \lambda_{k-1} p_{k-1}(t) - \lambda_k p_k(t) \quad \text{for all } k \geqslant s+1. \end{aligned}$$

Moreover, this system of equations has a unique solution subject to the initial conditions

$$\begin{aligned} p_s(0) &= 1, \\ p_k(0) &= 0 \quad \text{for all } k \geqslant s+1. \end{aligned}$$

Notice that if we are in the Poisson process case ($\lambda_i = \lambda$ for all i) then the differential equations in this theorem are exactly the ones we derived when proving Theorem 7.1 (the equivalence of the infinitesimal and original definitions of the Poisson process). The proof will follow exactly the same approach as the proof of Theorem 7.1. We won't finish up with a nice formula for the functions $p_k(t)$ (as we did for the Poisson process), but we will have a method which in principle could be used to solve for as many k as we want.

Proof. We first derive the system of differential equations for the functions p_k.

We will look at the $p_s(t)$ first. Here we have

$$\begin{aligned} p_s(t+h) &= \mathbb{P}(X(t+h) = s) \\ &= \mathbb{P}(X(t+h) = s \mid X(t) = s) \cdot \mathbb{P}(X(t) = s) \\ &= \mathbb{P}(\text{no arrivals in } [t, t+h]) \cdot \mathbb{P}(X(t) = s) \\ &= (1 - \lambda_s h + o(h))\mathbb{P}(X(t) = s) \\ &= (1 - \lambda_s h + o(h)) p_s(t) \\ &= p_s(t) - \lambda_s h p_s(t) + o(h). \end{aligned}$$

Hence, rearranging, we have

$$\frac{p_s(t+h) - p_s(t)}{h} = -\lambda_s p_s(t) + \frac{o(h)}{h}.$$

Taking the limit of both sides as $h \to 0$ gives

$$p_s'(t) = -\lambda_s p_s(t).$$

Now let $k \geqslant s+1$. We have

$$\begin{aligned} p_k(t+h) &= \mathbb{P}(X(t+h)=k) \\ &= \sum_{m=s}^{k} \mathbb{P}(X(t+h)=k \mid X(t)=m) \cdot \mathbb{P}(X(t)=m) \end{aligned}$$

Now, because of property (ii) of the definition of a birth process every term of this sum except the $m=k-1$ and $m=k$ ones are $o(h)$ so:

$$\begin{aligned} p_k(t+h) &= \mathbb{P}(X(t+h)=k \mid X(t)=k-1) \cdot \mathbb{P}(X(t)=k-1) \\ &\quad + \mathbb{P}(X(t+h)=k \mid X(t)=k) \cdot \mathbb{P}(X(t)=k) + o(h) \\ &= (\lambda_{k-1}h + o(h))p_{k-1}(t) + (1-\lambda_k h + o(h))p_k(t) + o(h) \\ &\qquad \text{(again using property (ii) of the definition of a bith process)} \\ &= \lambda_{k-1}hp_{k-1}(t) + p_k(t) - \lambda_k h p_k(t) + o(h) \end{aligned}$$

Rearranging, we have

$$\frac{p_k(t+h)-p_k(t)}{h} = \lambda_{k-1}p_{k-1}(t) - \lambda_k p_k(t) + \frac{o(h)}{h}.$$

Taking the limit of both sides as $h \to 0$ gives

$$p_k'(t) = \lambda_{k-1}p_{k-1}(t) - \lambda_k p_k(t).$$

Hence, the $p_k(t)$ satisfy the following system of differential equations:

$$\begin{aligned} p_s'(t) &= -\lambda_s p_s(t), \\ p_k'(t) &= \lambda_{k-1}p_{k-1}(t) - \lambda_k p_k(t) \quad \text{for all } k \geqslant s+1. \end{aligned}$$

Let's show inductively that these have a unique solution. The argument will also provide a method for solving the equations one by one which in practice could be used to solve the first few. First we solve the first equation (for p_s):

$$p_s'(t) = -\lambda_s p_s(t).$$

This has general solution $p_s(t) = Ae^{-\lambda_s t}$, where A is a constant of integration. Since $X(0)=s$, we have $p_s(0) = \mathbb{P}(X(0)=s) = 1$, so $A=1$, and therefore the first equation has a unique solution $p_s(t) = e^{-\lambda_s t}$.

Now for the induction step, let's take $k \geqslant s+1$ and assume that we have a unique solution for $p_{k-1}(t)$ (unlike in the Poisson process case we don't have a nice form for this solution).

Then the differential equation for $p_k(t)$ gives

$$p_k'(t) = \lambda_{k-1}p_{k-1}(t) - \lambda_k p_k(t).$$

Rearranging, we get

$$p_k'(t) + \lambda_k p_k(t) = \lambda_{k-1} p_{k-1}(t).$$

We can rewrite the left-hand side of the above equation as

$$e^{-\lambda_k t} \frac{d}{dt}(p_k(t) e^{\lambda_k t}),$$

so the equation can be rewritten as

$$\frac{d}{dt}(p_k(t) e^{\lambda_k t}) = e^{\lambda_k t} \lambda_{k-1} p_{k-1}(t).$$

Integrating both sides gives

$$p_k(t) e^{\lambda_k t} = \int_0^t e^{\lambda_k x} \lambda_{k-1} p_{k-1}(x)\, dx.$$

where the limits of integration are chosen so make $p_k(0) = 0$ (because for $k \geqslant s+1$ we have $\mathbb{P}(X(0) = k) = 0$)). Hence,

$$p_k(t) = \lambda_{k-1} e^{-\lambda_k t} \int_0^t e^{\lambda_k x} p_{k-1}(x)\, dx.$$

So $p_k(t)$ has a unique solution for all k (by induction on k). □

The proof of this result also gives a method for finding the functions $p_s(t), p_{s+1}(t), \ldots$ (for as many as we have patience to do) as in the following example.

Example 43. Let $(X(t) : \ t \geqslant 0)$ be a birth process with $X(0) = 2$ and birth parameters $\lambda_i = 1 + \binom{i}{2}$ for $i \geqslant 2$. Let

$$p_k = \mathbb{P}(X(t) = k).$$

Find the functions $p_2(t), p_3(t), p_4(t)$.

We have $\lambda_2 = 2, \lambda_3 = 4, \lambda_4 = 7$. By the previous Theorem these satisfy the differential equations:

$$\begin{aligned} p_2'(t) &= -\lambda_2 p_2(t), \\ p_k'(t) &= \lambda_{k-1} p_{k-1}(t) - \lambda_k p_k(t) \quad \text{for all } k \geqslant 3. \end{aligned}$$

Putting in the values for λ_i we get

$$\begin{aligned} p_2'(t) &= -2p_2(t), \\ p_3'(t) &= 2p_2(t) - 4p_3(t) \\ p_4'(t) &= 4p_3(t) - 7p_4(t) \end{aligned}$$

The first equation has solution $p_2 = Ce^{-2t}$ and we know $p_2(0) = 1$ so $C = 1$. So the solution is

$$p_2(t) = e^{-2t}$$

Rearranging the second equation and substituting for $p_2(t)$ gives

$$\begin{aligned} e^{-4t}\frac{d}{dt}\left(p_3(t)e^{4t}\right) &= 2e^{-2t} \\ \frac{d}{dt}\left(p_3(t)e^{4t}\right) &= 2e^{2t}. \end{aligned}$$

And so (using the initial condition $p_3(0) = 0$),

$$p_3(t)e^{4t} = \int_0^t 2e^{2x}\,dx = e^{2t} - 1$$

So the solution is

$$p_3(t) = e^{-2t} - e^{-4t}.$$

Rearranging the third equation and substituting for $p_3(t)$ gives

$$\begin{aligned} e^{-7t}\frac{d}{dt}\left(p_4(t)e^{7t}\right) &= 4e^{-2t} - 4e^{-4t} \\ \frac{d}{dt}\left(p_4(t)e^{7t}\right) &= 4e^{5t} - 4e^{3t}. \end{aligned}$$

And so (using the initial condition $p_4(0) = 0$),

$$p_4(t)e^{7t} = \int_0^t 4e^{5x} - 4e^{3x}\,dx = \frac{4}{5}e^{5t} - \frac{4}{3}e^{3t} - \frac{4}{5} + \frac{4}{3}$$

So the solution is

$$p_4(t) = \frac{4}{5}e^{-2t} - \frac{4}{3}e^{-4t} + \frac{8}{15}e^{-7t}.$$

8.5 Explosion

Do all possible choices of birth parameter give a reasonably behaved birth process?

Suppose that we have a birth process with $X(0) = 0$ in which the interarrival times satisfy $S_i = \frac{1}{2^i}$. Using the connection between the arrival and interarrival times we have

$$T_n = \sum_{i=1}^{n} S_i = \sum_{i=1}^{n} \frac{1}{2^i}$$

Since $\sum_{i=1}^{\infty} \frac{1}{2^i} = 1$ (sum of a geometric progression) we have that $T_n < 1$ for all $n \in \mathbb{N}$. In other words by the time we reach time 1 we have had infinitely many arrivals and $X(t) = \infty$ for all $t \geqslant 1$. Of course this is not possible in the physical world, but in a mathematical model it can certainly occur. Similar behaviour happens whenever $\sum_{i=1}^{\infty} S_i$ is finite.

On the other hand if $\sum_{i=1}^{\infty} S_i$ is infinite then for every fixed time t, we have some n for which

$$t < \sum_{i=1}^{n} S_i = T_n.$$

So $X(t) < n$. This shows that $X(t)$ is finite for all t.

Definition 25. Let $(X(t) : t \geqslant 0)$ be a birth process with $X(0) = s \in \mathbb{N}$ and interarrival times $S_{s+1}, S_{s+2}, \ldots$.

- The process *explodes* if $\sum_{i=s+1}^{\infty} S_i$ is finite;
- The process *does not explode* if $\sum_{i=s+1}^{\infty} S_i$ is infinite.

If the process explodes then $\sum_{i=s+1}^{\infty} S_i$ is the *time of explosion.*

Explosion depends on the S_i which are random variables so explosion is an event and it has a probability. It turns out that for any choice of parameters explosion occurs with probability 0 or with probability 1 and there is a simple condition on the birth parameters which determines which of these cases we are in.

Theorem 8.3. *Let $(X(t) : t \geqslant 0)$ be a birth process with $X(0) = s \in \mathbb{N}$ and birth parameters $\lambda_s, \lambda_{s+1}, \lambda_{s+2}, \ldots$.*

(i) If $\sum_{i=s}^{\infty} \frac{1}{\lambda_i}$ is finite then the probability of explosion is 1;

(ii) If $\sum_{i=s}^{\infty} \frac{1}{\lambda_i}$ is infinite then the probability of explosion is 0.

Sketch of proof. The key point is that

$$\sum_{i=s}^{\infty} \frac{1}{\lambda_i} = \sum_{i=s}^{\infty} \mathbb{E}(S_{i+1}) = \mathbb{E}\left(\sum_{i=s}^{\infty} S_{i+1} \right)$$

So if $\sum_{i=s}^{\infty} \frac{1}{\lambda_i}$ is finite then so is $\mathbb{E}\left(\sum_{i=s}^{\infty} S_{i+1}\right)$. But if a random variable W has finite expectation then $\mathbb{P}(W = \infty) = 0$. Hence the probability of explosion is 0. This proves part (i).

Part (ii) is omitted but should look plausible. □

Another way of thinking of this condition is that

$$\sum_{i=s}^{n} \frac{1}{\lambda_i} = \mathbb{E}\left(\sum_{i=s}^{n} S_{i+1}\right) = \mathbb{E}(T_n).$$

So

- if $\sum_{i=s}^{\infty} \frac{1}{\lambda_i} = T$ is finite, then however large n is, we have $\mathbb{E}(T_n) < T$.
- if $\sum_{i=s}^{\infty} \frac{1}{\lambda_i}$ is infinite, then $\mathbb{E}(T_n) \to \infty$ as $n \to \infty$.

Example 44.

- If $X(t)$ is a Poisson process then $s = 0$ and $\lambda_i = \lambda$ for all i. We have $\sum_{i=0}^{\infty} \frac{1}{\lambda}$ is infinite so the probability of explosion is 0.
- If $X(t)$ is a linear birth process then $\lambda_i = i\lambda$ for all i. We have $\sum_{i=s}^{\infty} \frac{1}{i\lambda}$ is infinite so the probability of explosion is 0.
- If $X(t)$ is the birth process of Example 43 with $s = 2$ and $\lambda_i = 1 + \binom{i}{2}$ for all i. We have $\sum_{i=2}^{\infty} \frac{1}{1+\binom{i}{2}} = \sum_{i=2}^{\infty} \frac{2}{i^2-i+2} < \sum_{i=2}^{\infty} \frac{2}{(i-1)^2}$ which is finite so the probability of explosion is 1.

9 Continuous-time Markov chains

9.1 Transition Rates and the Generator Matrix

In the final part of the module we look at a more general family of continuous-time processes which include Poisson processes and birth processes. We will do this quite briefly just introducing the main ideas and seeing some examples without getting too deep into the details.[16]

Recall that for a discrete-times process we defined the t-step transition probabilities as the probability of going from state i to state j in t steps:

$$p_{i,j}^{(t)} = \mathbb{P}(X_{s+t} = j \mid X_s = i).$$

To describe a discrete-time Markov chain it was enough to give the transition probabilities

$$p_{i,j} = p_{i,j}^{(1)} = \mathbb{P}(X_{s+1} = j \mid X_s = i).$$

Note that there is no dependence on s in either of these definitions because we are dealing with homogeneous chains.

Now suppose that $(X(t) : \ t \geqslant 0)$ is a continuous-time process on state space S. How can we do something similar in this setting?

If $i, j \in S$ we can still consider the probability of going between i and j in time t. Let's define:

$$p_{i,j}(t) = \mathbb{P}(X(s+t) = j \mid X(s) = i).$$

Now, unlike in the discrete-time case, there is no smallest unit of time so we can't immediately define an analogue of the transition probabilty. The way to get round this is to mimic the infinitesimal description of the Poisson process (or birth process).

For each $j \neq i$ we have a transition rate g_{ij} (which could be 0) and:

$$\mathbb{P}(X(s+h) = j \mid X(s) = i) = g_{ij}h + o(h)$$

The way to interpret this is that in a small interval $[s, s+h]$ the probability of making a transition from i to j is approximately $g_{ij}h$. Note that the g_{ij} is a rate not a probability; it could be any non-negative real number.

We also need to think about the probability of staying in the same state. Clearly we want

$$\sum_{j \in S} \mathbb{P}(X(s+t) = j \mid X(s) = i) = 1$$

[16] I have also omitted a few technicalities in definitions and proofs. As for some of the infinite state space material earlier, if you are happiest when everything is on a very rigorous pure mathematical footing then you will need to take a few things on trust.

So

$$\begin{aligned}\mathbb{P}(X(s+h)=i \mid X(s)=i) &= 1-\sum_{j\neq i}\mathbb{P}(X(s+h)=j \mid X(s)=i)\\ &= 1-\sum_{j\neq i} g_{ij}h+o(h)\end{aligned}$$

So let's set $g_{ii}=-\sum_{j\neq i} g_{ij}$ and then

$$\mathbb{P}(X(s+h)=i \mid X(s)=i) = 1+g_{ii}h+o(h)$$

This will be fine provided that $\sum_{j\neq i} g_{ij}$ is finite.

It is convenient to represent these g_{ij} in a (possibly infinite) array with rows and columns indexed by S. This is similar to the transition matrix of a discrete-time Markov chain. This matrix is called the *generator matrix* (or just the *generator*) of the process and is often denoted by G or Q. For clarity we will sometimes write g_{ij} as $g_{i,j}$ (just as we did for transition probabilities).

If $S=\{1,2,\ldots,n\}$ we would have:

$$G=\begin{pmatrix} -\sum_{j\neq 1} g_{1,j} & g_{1,2} & g_{1,3} & \cdots & g_{1,n}\\ g_{2,1} & -\sum_{j\neq 2} g_{2,j} & g_{2,3} & \cdots & g_{2,n}\\ \vdots & & \ddots & & \vdots\\ g_{n,1} & g_{n,2} & g_{n,3} & \cdots & -\sum_{j\neq n} g_{n,j}\end{pmatrix}$$

where the sums are over all values of j in S apart from the excluded one.

From the way we constructed it, the generator satisfies:

- The off-diagonal entries are all non-negative.
- The row sums are 0.
- It is a square $|S|\times|S|$ matrix (or the infinite analogue if S is infinite)

These should be reminiscent of the properties of a transition matrix but they are not the same.

If we have a state i with $g_{ij}=0$ for all $j\neq i$ then there are no transitions out of i. Just as in the discrete-time case we say that t is an *absorbing* state. Note that $g_{ii}=-\sum_{j\neq i} g_{ij}=0$ so in the generator an absorbing state will correspond to a row in which every entry is 0.

9.2 Definition and Examples

The discussion above can be used to define a general continuous-time Markov chain. We will assume that the state space S is finite or countably infinite.

Definition 26. The process $(X(t) : \ t \geqslant 0)$ is a *continuous-time homogeneous Markov chain* with state space S if it satisfies:

- For all times $0 < t_1 < \ldots < t_m < t$ and states $i, j, s_1, \ldots, s_m \in S$

  $$\mathbb{P}(X(t+r) = j \mid X(t_1) = s_1, X(t_2) = s_2, \ldots, X(t_m) = s_m, X(t) = i) = \mathbb{P}(X(t+r) = j \mid X(t) = i)$$

  whenever the lefthand side is defined[17]. In other words, conditional on $X(t)$, we have that $X(t+r)$ is independent of the process up to time t (this is the Markov property).

- For all $i, j \in S$ and $t, r \geqslant 0$ the probability $\mathbb{P}(X(t+r) = j \mid X(t) = i)$ does not depend on t (this is the time homogeneity property).

- $$\mathbb{P}(X(t+h) = j \mid X(t) = i) = \begin{cases} g_{ij}h + o(h) & \text{if } i \neq j \\ 1 + g_{ii}h + o(h) & \text{if } i = j \end{cases}$$

  where $g_{ij} \geqslant 0$ and $g_{ii} = -\sum_{j \neq i} g_{ij}$ (exactly as above).

The matrix G of g's as above is called the *generator matrix* or the *infinitesimal generator* of the process.

To completely specify the process we would also need to say what $X(0)$ is. This could be either a value or a probability distribution on S.

Example 45. A Poisson process of rate λ satisfies this definition with

$$g_{ij} = \begin{cases} \lambda & \text{if } j = i+1 \\ 0 & \text{if } j \neq i+1 \end{cases}$$

and $g_{ii} = -\lambda$. The generator matrix is:

$$\begin{pmatrix} -\lambda & \lambda & 0 & 0 & 0 & \cdots \\ 0 & -\lambda & \lambda & 0 & 0 & \cdots \\ 0 & 0 & -\lambda & \lambda & 0 & \cdots \\ \vdots & & \ddots & & & \vdots \end{pmatrix}$$

[17] that is whenever $\mathbb{P}(X(t_1) = s_1, X(t_2) = s_2, \ldots, X(t_m) = s_m, X(t) = i) \neq 0$

Example 46. A birth process with $X(0) = 0$ and parameters $\lambda_0, \lambda_1, \ldots$ satisfies this definition with

$$g_{i,j} = \begin{cases} \lambda_i & \text{if } j = i+1 \\ 0 & \text{if } j \neq i+1 \end{cases}$$

and $g_i = -\lambda_i$. The generator matrix is:

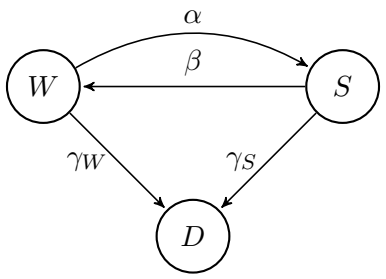
$$\begin{pmatrix} -\lambda_0 & \lambda_0 & 0 & 0 & 0 & \cdots \\ 0 & -\lambda_1 & \lambda_1 & 0 & 0 & \cdots \\ 0 & 0 & -\lambda_2 & \lambda_2 & 0 & \cdots \\ \vdots & & \ddots & & & \vdots \end{pmatrix}$$

Example 47. A person's health is represented by one of three states 'Well' (state 1), 'Sick' (state 2), 'Dead' (state 3). Let $X(t)$ for $t \geqslant 0$ be the state of someone's health at time t years after we start recording it. We are going to regard this as a continuous model so the state can change at any time and we will track it continuously. Suppose that transitions happen from state 1 to state 2 at rate α per year; and from state 2 to state 1 at rate β per year. We can also go to state 3 from either state 1 or state 2 with rates γ_W and γ_S resepectively (we would expect there to be greater chance of death when sick so γ_S will be greater than γ_W).

We have generator matrix:

$$\begin{pmatrix} -\alpha - \gamma_W & \alpha & \gamma_W \\ \beta & -\beta - \gamma_S & \gamma_S \\ 0 & 0 & 0 \end{pmatrix}$$

We could also represent this with the picture



Note that the numbers here are rates not probabilities. Also we don't show loops: the parameter g_{ii} which controls the probability of staying in a state is determined by the arrows out of it via the relation $g_{ii} = -\sum_{j \neq i} g_{ij}$.

In reality the situation we are trying to model here will not be time homogeneous; you would expect the chance of dying to be higher in old age (large t). However it could be a useful approximation in some circumstances.

9.3 Description using exponential holding times

We saw another way to think of the Poisson process using interarrival times $S_1, S_2, S_3, \ldots$. The key property was that these are independent random variables and each is distributed $\text{Exp}(\lambda)$. This viewpoint can also be applied to the general continous-time Markov chain.

Suppose that the chain is in state i and that $g_{ii} \neq 0$ (that is i is not absorbing). From the definition we know that

$$\mathbb{P}(X(t+h) = i \mid X(t) = i) = 1 + g_{ii}h + o(h) = 1 - \sum_{j \neq i} g_{ij}h + o(h)$$

this implies (similarly to deriving and solving the Poisson process differential equation for $p_0(t)$) that the time spent in state i until the next move is a random variable distributed $\text{Exp}(\sum_{j \neq i} g_{ij})$ (equivalently $\text{Exp}(-g_{ii})$). These are called the *holding times* or *waiting times*. We will then move to some other state j chosen according to the probabilities

$$\mathbb{P}(\text{jump to state } j) = \frac{g_{ij}}{-g_{ii}} = \frac{g_{ij}}{\sum_{j \neq i} g_{ij}}.$$

That is, the chain jumps according to the discrete-time chain with the same state space S and transition probabilities $p_{ij} = \frac{g_{ij}}{-g_{ii}}$ if $i \neq j$, $p_{ii} = 0$, but rather than happening at regular discrete times, these jumps happen at random times given by the holding times.

Just as for the birth process we may have problems if transitions happen at a faster and faster rate resulting in infinitely many moves from state to state before a fixed time. This is called explosion.

Definition 27. A continuous-time Markov chain is *non-explosive* if the probability of having infinitely many transitions in any finite interval of time is 0.

It is possible to give a condition on the generator which ensures that the process is non-explosive. I won't give this in detail but it implies for instance that any continuous-time Markov chain with finite S is non-explosive. For the rest of the module we will only consider non-explosive processes.

Example 48. Take $\alpha, \beta > 0$ and let $(X(t) : t \geqslant 0)$ be the two-state chain with $S = \{1, 2\}$ and generator

$$\begin{pmatrix} -\alpha & \alpha \\ \beta & -\beta \end{pmatrix}.$$

Describe how this chain evolves assuming that we start in state 1?

Starting in state 1, we wait a random amount of time distributed $\text{Exp}(\alpha)$ and then we jump to state 2. From state 2 we wait a random amount of time distributed $\text{Exp}(\beta)$ and then we jump back to state 1. This repeats with each of the waiting times being independent.

Example 49 (Birth-death process)**.** Let $(X(t) : t \geqslant 0)$ be the continuous-time Markov chain with $S = \mathbb{N}$ and generator

$$G = \begin{pmatrix} -\lambda_0 & \lambda_0 & 0 & 0 & 0 & \dots \\ \mu_1 & -\lambda_1 - \mu_1 & \lambda_1 & 0 & 0 & \dots \\ 0 & \mu_2 & -\lambda_2 - \mu_2 & \lambda_2 & 0 & \dots \\ \vdots & \ddots & \ddots & \ddots & & \vdots \end{pmatrix}$$

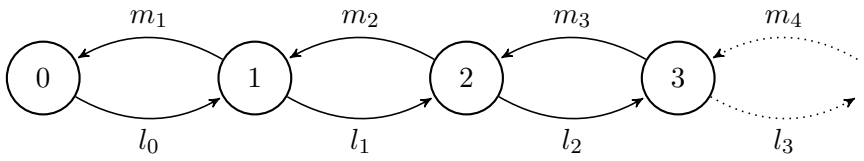
The idea is that if the population has size i, then births happen at a rate λ_i and deaths happen at a rate μ_i. When $i \geqslant 1$ we have that

$$\mathbb{P}(X(t+h) = j \mid X(t) = i) = \begin{cases} \lambda_i h + o(h) & \text{if } j = i+1 \\ \mu_i h + o(h) & \text{if } j = i-1 \\ 1 - (\lambda_i + \mu_i)h + o(h) & \text{if } j = i \\ o(h) & \text{otherwise} \end{cases}$$

The case $i = 0$ is a bit different because when the population has size 0 there can be no deaths:

$$\mathbb{P}(X(t+h) = j \mid X(t) = 0) = \begin{cases} \lambda_i h + o(h) & \text{if } j = 1 \\ 1 - \lambda_i h + o(h) & \text{if } j = 0 \\ o(h) & \text{otherwise} \end{cases}$$

Our description in terms of exponential holding times means that we can compare this with the discrete-time Markov chain with transition graph:



where

$$l_i = \frac{\lambda_i}{\lambda_i + \mu_i}, \quad m_i = \frac{\mu_i}{\lambda_i + \mu_i}.$$

If we are in state i, we wait a random amount of time distributed $\mathrm{Exp}(\lambda_i + \mu_i)$ and then perform a step of the discrete-time chain above moving either left or right with the probabilities shown.

Example 50. Suppose we have a queue of customers with a single server. Customers arrive according to a Poisson process of rate λ and wait until the server is available. Each customer takes a random amount of time to be served with each service distributed $\mathrm{Exp}(\mu)$ independently of all other customers. Suppose also that arrival times and service times are

independent. Let $Q(t)$ be the number of people in the queue (including the person being served) at time t.

It turns out that $Q(t)$ is a birth-death process (as in the previous example) with $\lambda_i = \lambda$ and $\mu_i = \mu$. So the generator is

$$G = \begin{pmatrix} -\lambda & \lambda & 0 & 0 & 0 & \dots \\ \mu & -\lambda-\mu & \lambda & 0 & 0 & \dots \\ 0 & \mu & -\lambda-\mu & \lambda & 0 & \dots \\ \vdots & \ddots & \ddots & \ddots & & \vdots \end{pmatrix}$$

To see this we can do the usual consideration of what happens in a small interval. If $m \geqslant 1$,

- $\mathbb{P}(\text{one arrival in } [t, t+h]) = \lambda h + o(h)$ as arrivals form a Poisson process of rate λ.
- $\mathbb{P}(\text{one service completed in } [t, t+h]) = 1 - e^{-\mu h} = \mu h + o(h)$ as service times are $\text{Exp}(\mu)$ and $e^{-x} = 1 - x + o(x)$ for small x
- Anything else (more than one arrival, more than one departure, an arrival and a departure) has probability $o(h)$.

9.4 Differential Equations

We analysed the Poisson process and Birth processes using differential equations. A similar approach can be used more generally. Suppose that $(X(t) : \ t \geqslant 0)$ is a continuous-time homogeneous Markov chain. For the Poisson process we defined

$$p_i(t) = \mathbb{P}(X(t) = i \mid X(0) = 0)$$

the probability of going from state 0 to state i in time t.

In the general case we need to consider more possible transitions so to extend this definition we should consider:

$$p_{i,j}(t) = \mathbb{P}(X(t) = j \mid X(0) = i)$$

for any $i, j \in S$.

Since the chain is homogenous, this is equal to the probability of going from i to j in any time interval of length t. That is:

$$p_{i,j}(t) = \mathbb{P}(X(t+s) = j \mid X(s) = i).$$

These are precisely the functions we defined at the start of this section.

Now, thinking of $p_{i,j}(t)$ as being analogous to the t-step transition probability in a discrete-time chain we can write down the same Chapman-Kolmogorov relations as in the discrete case.

Thinking of the $p_{i,j}(t)$ as generalising the $p_i(t)$ of the Poisson process (as above) we can find a system of differential equations for them. In fact, there are two ways to do this derivation leading to two different systems of differential equations, the forward and backward equations.

Theorem 9.1. *Suppose that $(X(t) : t \geqslant 0)$ is a continuous-time Markov chain with generator matrix G and $X(t)$ is non-explosive. For any state $i, j \in S$ and any times $s, t \geqslant 0$ we have:*

(i) the Chapman-Kolmogorov relations*:*

$$p_{i,j}(s+t) = \sum_{k \in S} p_{i,k}(s) p_{k,j}(t)$$

(ii) the forwards equations*:*

$$p'_{i,j}(t) = p_{i,j}(t) g_{jj} + \sum_{k \neq j} p_{i,k}(t) g_{kj} = \sum_{k} p_{i,k}(t) g_{kj}$$

(iii) the backwards equations*:*

$$p'_{i,j}(t) = g_{i,i} p_{ij}(t) + \sum_{k \neq i} g_{ik} p_{k,j}(t) = \sum_{k} g_{ik} p_{k,j}(t)$$

Remark. All three of these results can be written in matrix form (at least when S is finite). If $S = \{1, 2, \ldots, n\}$ and we write

$$P(t) = \begin{pmatrix} p_{1,1}(t) & p_{1,2}(t) & \cdots & p_{1,n}(t) \\ p_{2,1}(t) & p_{2,2}(t) & \cdots & p_{2,n}(t) \\ \vdots & \vdots & \cdots & \vdots \\ p_{n,1}(t) & p_{n,2}(t) & \cdots & p_{n,n}(t) \end{pmatrix}, \quad P'(t) = \begin{pmatrix} p'_{1,1}(t) & p'_{1,2}(t) & \cdots & p'_{1,n}(t) \\ p'_{2,1}(t) & p'_{2,2}(t) & \cdots & p'_{2,n}(t) \\ \vdots & \vdots & \cdots & \vdots \\ p'_{n,1}(t) & p'_{n,2}(t) & \cdots & p'_{n,n}(t) \end{pmatrix}$$

then the three parts of this Theorem become:

(i) $P(s+t) = P(s)P(t)$;

(ii) $P'(t) = P(t)G$ (the forwards equations);

(iii) $P'(t) = GP(t)$ (the backwards equations).

Proof. (i) Exactly same argument as proof of Lemma 4.1 (condition on $X(s)$).

(ii) Let's work out $p_{i,j}(t+h)$ using part (i).

$$\begin{aligned} p_{i,j}(t+h) &= \sum_{k\in S} p_{i,k}(t)p_{k,j}(h) \\ &= p_{i,j}(t)p_{j,j}(h) + \sum_{k\neq j} p_{i,k}(t)p_{k,j}(h) \\ &= p_{i,j}(t)(1+g_{jj}h) + \sum_{k\neq j} p_{i,k}(t)g_{kj}h + o(h) \end{aligned}$$

Rearranging we get

$$\frac{p_{i,j}(t+h)-p_{i,j}(t)}{h} = p_{i,j}(t)g_{jj} + \sum_{k\neq j} p_{i,k}(t)g_{kj} + \frac{o(h)}{h}$$

Letting $h \to 0$

$$p'_{i,j}(t) = p_{i,j}(t)g_{jj} + \sum_{k\neq j} p_{i,k}(t)g_{kj}$$

as required.

(iii) Similarly to the previous part, let's work out $p_{i,j}(h+t)$ using part (i).

$$\begin{aligned} p_{i,j}(h+t) &= \sum_{k\in S} p_{i,k}(h)p_{k,j}(t) \\ &= p_{i,i}(h)p_{j,j}(t) + \sum_{k\neq i} p_{i,k}(h)p_{k,j}(t) \\ &= (1+g_{ii}h)p_{i,j}(t) + \sum_{k\neq i} g_{ik}(t)p_{k,j}h + o(h) \end{aligned}$$

Rearranging we get

$$\frac{p_{i,j}(h+t)-p_{i,j}(t)}{h} = g_{ii}p_{i,j}(t) + \sum_{k\neq i} g_{ik}p_{k,j}(t) + \frac{o(h)}{h}$$

Letting $h \to 0$

$$p'_{i,j}(t) = g_{ii}p_{i,j}(t) + \sum_{k\neq i} g_{ik}p_{k,j}(t)$$

as required.

□

In general these equations will be hard to solve but sometimes there is a nice solution. Two examples we have seen are the Poisson process (section 8.2) and the linear birth process (Problem Sheet 9, Question 3). Sometimes, even if we can't solve the equations completely we can use them to extract some information about the process (see examples later).

Example 51. Earlier (Example 47) we saw a three state chain with $S = \{W, S, D\}$ and generator matrix:

$$\begin{pmatrix} -\alpha - \gamma_W & \alpha & \gamma_W \\ \beta & -\beta - \gamma_S & \gamma_S \\ 0 & 0 & 0 \end{pmatrix}$$

Let's work out some of the differential equations for this.

The forward equation for p_{WW} is

$$p'_{WW}(t) = p_{WW}(t)g_{WW} + p_{WS}(t)g_{SW} + p_{WD}(t)g_{DW}$$

(You can see this either by using the formula from Theorem 9.1 or looking at the top left entry of the matrix $P(t)G$).)

If we put in the values of g_{ij} from the matrix we get:

$$p'_{WW}(t) = -(\alpha + \gamma_W)p_{WW}(t) + \beta p_{WS}(t)$$

For the backward equation we get (using the formula from Theorem 9.1 or looking at the top left entry of the matrix $GP(t)$):

$$\begin{aligned} p'_{WW}(t) &= g_{WW}p_{WW}(t) + g_{WS}p_{SW}(t) + g_{WD}p_{DW}(t) \\ p'_{WW}(t) &= -(\alpha + \gamma_W)p_{WW}(t) + \alpha p_{SW}(t) \qquad \text{(using } p_{DW}(t) = 0 \text{ here)} \end{aligned}$$

Let's do one more

$$\begin{aligned} p'_{WS}(t) &= \alpha p_{WW}(t) - (\beta + \gamma_S)p_{WS}(t) \quad \text{forwards equation} \\ p'_{WS}(t) &= -(\alpha + \gamma_W)p_{WS}(t) + \alpha p_{SS}(t) \quad \text{backwards equation} \end{aligned}$$

Have a go at writing out the others. There will be 9 equations (one for each entry of thr 3×3 matrix $P'(t)$.

However three of them are very simple $p'_{DW}(t) = p'_{DS}(t) = p'_{DD}(t) = 0$. This is because $p_{DW}(t) = p_{DS}(t) = 0$ for all t and $p_{DD}(t) = 1$ for all t so these three functions are constant. You can also see this from the matrices; every entry of the last row of G is 0 and so every entry of the last rows of GP will also be 0.

9.5 Examples of Limiting Behaviour

It is possible to develop a theory of equilibrium and limiting distributions for continuous-time chains in a similar way to the discrete-time case. In this module we won't do this. However, here are a couple of examples showing how we can get information about the long-term behaviour of the chain from the relevant differential equations.

Example 52. Take $\alpha, \beta > 0$ and let $(X(t) : t \geqslant 0)$ be the two-state chain with $S = \{1, 2\}$ and generator

$$\begin{pmatrix} -\alpha & \alpha \\ \beta & -\beta \end{pmatrix}.$$

The forward equations for $p_{1,2}(t)$ is

$$p'_{1,2}(t) = \alpha p_{1,1}(t) - \beta p_{1,2}(t).$$

We know that $\mathbb{P}(X(t) = 1 \mid X(0) = 1) + \mathbb{P}(X(t) = 2 \mid X(0) = 1) = 1$, so $p_{1,1}(t) = 1 - p_{1,2}(t)$. Substituting this we get

$$p'_{1,2}(t) = \alpha - (\alpha + \beta) p_{1,2}(t).$$

This can be solved, similarly to the equations for a birth process. We have

$$\begin{aligned} p'_{1,2}(t) + (\alpha + \beta) p_{1,2}(t) &= \alpha \\ e^{-(\alpha+\beta)t} \frac{d}{dt} \left(p_{1,2}(t) e^{(\alpha+\beta)t} \right) &= \alpha \\ p_{1,2}(t) e^{(\alpha+\beta)t} &= \alpha \int_0^t e^{(\alpha+\beta)x} \, dx \\ &\qquad \text{(the limits come from the condition } p_{1,2}(0) = 0) \\ p_{1,2}(t) e^{(\alpha+\beta)t} &= \frac{\alpha}{\alpha + \beta} \left(e^{(\alpha+\beta)t} - 1 \right) \\ p_{1,2}(t) &= \frac{\alpha}{\alpha + \beta} \left(1 - \frac{1}{e^{(\alpha+\beta)t}} \right) \end{aligned}$$

A similar calculation for $p_{2,1}(t)$ (I leave the details to you) gives

$$p_{2,1}(t) = \frac{\beta}{\alpha + \beta} \left(1 - \frac{1}{e^{(\alpha+\beta)t}} \right)$$

This (and the fact that $p_{1,1}(t) + p_{1,2}(t) = p_{2,1}(t) + p_{2,2}(t) = 1$) allows us to work out the matrix

$$P(t) = \begin{pmatrix} p_{1,1}(t) & p_{1,2}(t) \\ p_{2,1}(t) & p_{2,2}(t) \end{pmatrix} = \begin{pmatrix} \left(1 + \frac{\alpha}{\beta e^{(\alpha+\beta)t}}\right) \frac{\beta}{\alpha+\beta} & \left(1 - \frac{1}{e^{(\alpha+\beta)t}}\right) \frac{\alpha}{\alpha+\beta} \\ \left(1 - \frac{1}{e^{(\alpha+\beta)t}}\right) \frac{\beta}{\alpha+\beta} & \left(1 + \frac{\beta}{\alpha e^{(\alpha+\beta)t}}\right) \frac{\alpha}{\alpha+\beta} \end{pmatrix}.$$

As $t \to \infty$ we have

$$P(t) \to \begin{pmatrix} \frac{\beta}{\alpha+\beta} & \frac{\alpha}{\alpha+\beta} \\ \frac{\beta}{\alpha+\beta} & \frac{\alpha}{\alpha+\beta} \end{pmatrix}.$$

This should remind you of the ideas on limiting distributions for discrete-time chains. The matrix tends to a limit with both rows equal. The probability that the chain is in state j given that it starts in state i tends to a limit which does not depend on i.

Example 53. In Example 49 we defined the Birth-Death process as the continuous-time stochastic process with $S = \mathbb{N}$ and generator:

$$G = \begin{pmatrix} -\lambda_0 & \lambda_0 & 0 & 0 & 0 & \dots \\ \mu_1 & -\lambda_1 - \mu_1 & \lambda_1 & 0 & 0 & \dots \\ 0 & \mu_2 & -\lambda_2 - \mu_2 & \lambda_2 & 0 & \dots \\ \vdots & \ddots & \ddots & \ddots & & \vdots \end{pmatrix}$$

For this process the forward equations are:

$$\begin{aligned} p'_{i,j}(t) &= \lambda_{j-1} p_{i,j-1}(t) - (\mu_j + \lambda_j) p_{i,j}(t) + \mu_{j+1} p_{i,j+1}(t) \quad \text{for } j \geqslant 1 \\ p'_{i,0}(t) &= -\lambda_0 p_{i,0}(t) + \mu_1 p_{i,1}(t) \end{aligned}$$

and the backwards equations are:

$$\begin{aligned} p'_{i,j}(t) &= \mu_i p_{i-1,j}(t) - (\mu_i + \lambda_i) p_{i,j}(t) + \lambda_i p_{i+1,j}(t) \quad \text{for } i \geqslant 1 \\ p'_{0,j}(t) &= -\lambda_0 p_{0,j}(t) + \lambda_0 p_{1,j}(t) \end{aligned}$$

In general there is no nice form of solution for these equations. However, it can still be possible to read off some information about the system. See the next example for a special case of this.

Example 54. In Example 50 we saw how to model a queue with a single server as a birth-death process with parameters $\lambda_i = \lambda$, $\mu_i = \mu$. This process is sometimes called the $M(\lambda)/M(\mu)/1$ queue. Look back at that example to see what assumptions on arrivals and service times led to these parameters. The generator is:

$$G = \begin{pmatrix} -\lambda & \lambda & 0 & 0 & 0 & \dots \\ \mu & -\lambda - \mu & \lambda & 0 & 0 & \dots \\ 0 & \mu & -\lambda - \mu & \lambda & 0 & \dots \\ \vdots & \ddots & \ddots & \ddots & & \vdots \end{pmatrix}$$

Using the expression in the previous example we have equations:

$$\begin{aligned} p'_{i,j}(t) &= \lambda p_{i,j-1}(t) - (\mu + \lambda) p_{i,j}(t) + \mu p_{i,j+1}(t) && \text{(forward equation for } j \geqslant 1) \\ p'_{i,0}(t) &= -\lambda p_{i,0}(t) + \mu p_{i,1}(t) && \text{(forward equation for } j = 0) \\ p'_{i,j}(t) &= \mu p_{i-1,j}(t) - (\mu + \lambda) p_{i,j}(t) + \lambda p_{i+1,j}(t) && \text{(backwards equation for } i \geqslant 1) \\ p'_{0,j}(t) &= -\lambda p_{0,j}(t) + \lambda p_{1,j}(t) && \text{(backwards equation for } i = 0) \end{aligned}$$

It can be shown (although we won't prove it) that we have a limit $p_{i,j}(t) \to w_j$ as $t \to \infty$ where w_j does not depend on i.

Now let's look at the equations above letting $t \to \infty$. The forwards equations are:

$$\lim_{t\to\infty} p'_{i,j}(t) = \lambda w_{j-1} - (\mu+\lambda)w_j + \mu w_{j+1} \quad \text{for } j \geqslant 1$$
$$\lim_{t\to\infty} p'_{i,0}(t) = -\lambda w_0 + \mu w_1$$

and the backwards equations are:

$$\lim_{t\to\infty} p'_{i,j}(t) = \mu w_j - (\mu+\lambda)w_j + \lambda w_j = 0 \quad \text{for } j \geqslant 1$$
$$\lim_{t\to\infty} p'_{0,j}(t) = -\lambda w_j + \lambda w_j = 0$$

Substituting the result of the backwards equations into the forwards equations we get:

$$\lambda w_{j-1} - (\mu+\lambda)w_j + \mu w_{j+1} = 0 \quad \text{for } j \geqslant 1$$
$$-\lambda w_0 + \mu w_1 = 0$$

Which can be re-arranged to:

$$w_1 = \tfrac{\lambda}{\mu} w_0$$
$$w_{j+1} = \tfrac{\mu+\lambda}{\mu} w_j - \tfrac{\lambda}{\mu} w_{j-1} \quad \text{for } j \geqslant 1$$

You can check (induction on j) that these have solution $w_j = \left(\frac{\lambda}{\mu}\right)^j w_0$

So

$$\sum_{j=0}^{\infty} w_j = w_0 \sum_{j=0}^{\infty} \left(\frac{\lambda}{\mu}\right)^j = \begin{cases} w_0 \frac{1}{1-\frac{\lambda}{\mu}} & \text{if } \lambda < \mu; \\ \infty & \text{if } \lambda \geqslant \mu. \end{cases}$$

So if $\lambda < \mu$ we can choose $w_0 = 1 - \frac{\lambda}{\mu}$ which makes $\sum_{j=0}^{\infty} w_j = 1$ and we have a limiting distribution given by

$$\mathbb{P}(Q(t) = j) = \left(\frac{\lambda}{\mu}\right)^j \left(1 - \frac{\lambda}{\mu}\right)$$

(that is in the limit $Q(t) + 1 \sim \text{Geom}(1 - \frac{\lambda}{\mu})$).

If $\lambda \geqslant \mu$ there is no limiting distribution and

$$\mathbb{P}(Q(t) = j) \to 0 \text{ as } t \to \infty.$$

In this case the size of the queue will grow arbitrarily large. For any $n \in \mathbb{N}$, if we take t large enough the queue has at least n people in it with probability 1.

Quite a lot of the method of the previous example works for any birth-death process. We can use the same trick of noticing that the backwards equations force $\lim_{t\to\infty} p'_{i,j}(t) = 0$ and substitute this into the forwards equations (try it!). This gives equations for the w_j

which can be solved in terms of the parameters and w_0. From that we get a condition on the parameters under which there is a limiting distribution. However, if the parameters are more complicated than for the $M(\lambda)/M(\mu)/1$ queue then the w_j won't have such a nice form. On Problem Sheet 10, Question 4 leads you through this analysis for a queue with 2 servers. Also on Problem sheet 10, Question 6 looks at how this method of using both the forwards and backwards equations without solving them can be used to say something about the long-term behaviour of a general continuous-time Markov chain.