Lecture notes for Part A Probability

Notes written by James Martin, updated by Matthias Winkel

Oxford, Michaelmas Term 2018

winkel@stats.ox.ac.uk

Version of 27 October 2018

1

Review: probability spaces, random variables, distributions, independence

1.1 Probability spaces and random variables

We start by reviewing the basic idea of a **probability space** introduced in last year's course. This framework underlies modern probability theory, even though we will seldom need to appeal to it directly in this course.

A **probability space** is a collection $(\Omega, \mathcal{F}, \mathbb{P})$ with the following structure:

(i) Ω is a set, which we call the **sample space**.

(ii) $\mathcal{F}$ is a collection of subsets of Ω. An element of $\mathcal{F}$ is called an **event**.

(iii) The **probability measure** $\mathbb{P}$ is a function from $\mathcal{F}$ to $[0,1]$. It assigns a **probability** to each event in $\mathcal{F}$.

We can think of the probability space as modelling an experiment. The sample space Ω represents the set of all possible outcomes of the experiment.

The set $\mathcal{F}$ of events should satisfy certain natural conditions:

(1) $\Omega \in \mathcal{F}$.

(2) If $\mathcal{F}$ contains a set A, then it also contains the complement A^c (i.e. $\Omega \setminus A$).

(3) If $(A_i, i \in \mathcal{I})$ is a finite or countably infinite collection of events in $\mathcal{F}$, then their union $\bigcup_{i \in \mathcal{I}} A_i$ is also in $\mathcal{F}$.

By combining (2) and (3), we can also obtain finite or countable intersections, as well as unions.

Finally, the probability measure $\mathbb{P}$ should satisfy the following conditions (the **probability axioms**):

(1) $\mathbb{P}(\Omega) = 1$.

(2) If $(A_i, i \in \mathcal{I})$ is a finite or countably infinite collection of **disjoint** events, then

$$\mathbb{P}\left(\bigcup_{i\in\mathcal{I}} A_i\right) = \sum_{i\in\mathcal{I}} \mathbb{P}(A_i). \tag{1.1}$$

A **random variable** is a function defined on Ω. We will consider real-valued random variables, i.e. functions from Ω to $\mathbb{R}$.

A random variable represents an **observable** in our experiment; something we can "measure".

Formally, for a function X from Ω to $\mathbb{R}$ to be a random variable, we require that the subset

$$\{\omega \in \Omega \colon X(\omega) \leq x\}$$

of Ω is an event in $\mathcal{F}$, for every $x \in \mathbb{R}$. (Then, by taking complements, unions and intersections, we will in fact have that the set $\{\omega \in \Omega \colon X(\omega) \in B\}$ is in $\mathcal{F}$ for a very large class of sets B).

We will usually write X rather than $X(\omega)$ for the value taken by a random variable. Thus if X is a random variable we can talk about the probability of the event

$$\{X \in B\} = \{\omega \in \Omega \colon X(\omega) \in B\},$$

which we will write as $\mathbb{P}(X \in B)$.

Within one experiment, there will be many observables! That is, on the same probability space we can consider many different random variables.

Remarks:

(a) For very simple models, there may be a natural way to set up the sample space Ω (e.g. to represent the set of possible outcomes of the throw of a die or a coin). For more complicated models, this quickly becomes less straightforward. In practice, we hardly ever want to consider Ω directly; instead we work directly with the "events" and "random variables" (the "observables") in the experiment.

(b) In contrast, there are settings in probability theory where we care a lot about the collection of events $\mathcal{F}$, and its structure. (For example, modelling a process evolving in time, we might have a family of different collections $\mathcal{F}_t$, $t \geq 0$, where $\mathcal{F}_t$ represents the set of events which can be observed by watching the evolution of the process up to time t). However, for the purposes of this course we will hardly ever worry about $\mathcal{F}$ directly; we will be safe to assume that $\mathcal{F}$ will always contain any event that we wish to consider.

1.1.1 Examples

Here are some examples of systems (or "experiments") that we might model using a probability space, and, for each one, some examples of random variables that we might want to consider within our model:

- We throw two dice, one red and one blue. Random variables: the score on the red die; the score on the blue die; the sum of the two; the maximum of the two; the indicator function of the event that the blue score exceeds the red score....

- A Geiger counter detecting particles emitted by a radioactive source. Random variables: the time of the kth particle detected, for $k = 1, 2, \dots$; the number of particles detected in the time interval $[0, t]$, for $t \in [0, \infty)$....

- A model for the evolution of a financial market. Random variables: the prices of various stocks at various times; interest rates at various times; exchange rates at various times....

- The growth of a colony of bacteria. Random variables: the number of bacteria present at a given time; the diameter of the colonised region at given times; the number of generations observed in a given time interval....

- A call-centre. The time of arrival of the kth call; the length of service required by the kth caller; the wait-time of the kth caller in the queue before receiving service....

1.2 Probability distributions

We consider the distribution of a random variable X. This can be summarised by the **distribution function** (or **cumulative distribution function**) of X, defined by

$$F(x) = \mathbb{P}(X \leq x)$$

for $x \in \mathbb{R}$. (Once we know F, we can derive the probabilities $\mathbb{P}(X \in B)$ for a very wide class of sets B by taking complements, intersections and unions. Formally, the **distribution** of a random variable X is the map $B \mapsto \mathbb{P}(X \in B)$, considered on a suitable collection of subsets $B \subseteq \mathbb{R}$. In practice, we identify distributions by identifying the cumulative distribution function or any other associated function that uniquely determines a distribution.)

Any distribution function F must obey the following properties:

(1) F is non-decreasing.

(2) F is right-continuous.

(3) $F(x) \to 0$ as $x \to -\infty$.

(4) $F(x) \to 1$ as $x \to \infty$.

Remark 1.1. Note that two different random variables can have the same distribution! For example, consider the model of two dice mentioned above. If the dice are "fair", then the distribution of the score on the blue die might be the same as the distribution of the score on the red die. However, that does not mean that the two scores are always the same! They are two different "observables" within the same experiment.

If two random variables X and Y have the same distribution, we write $X \overset{d}{=} Y$.

We single out two important classes of random variables: **discrete** and **continuous**.

1.2.1 Discrete random variables

A random variable X is **discrete** if there is a finite or countably infinite set B such that $\mathbb{P}(X \in B) = 1$.

We can represent the distribution of a discrete random variable X by its probability mass function

$$p_X(x) = \mathbb{P}(X = x)$$

for $x \in \mathbb{R}$. This function is zero except at a finite or countably infinite set of points. We have

- $\sum_x p_X(x) = 1$.
- $\mathbb{P}(X \in A) = \sum_{x \in A} p_X(x)$ for any set $A \subseteq \mathbb{R}$.

The points x where $\mathbb{P}(X = x) > 0$ are sometimes called the **atoms** of the distribution of X. In many examples these will be a set of integers such as $\{1, 2, \dots, n\}$ or $\{0, 1, 2, \dots\}$ or $\{1, 2, 3, \dots\}$.

The cumulative distribution function of X has jumps at the location of the atoms, and is constant on any interval that does not contain an atom.

1.2.2 Continuous random variables

A random variable X is called (absolutely) **continuous** if its distribution function F can be written as an integral. That is, there is a function f such that

$$\mathbb{P}(X \le x) = F(x) = \int_{-\infty}^{x} f(u)du.$$

f is called the **density function** (or **probability density function**) of X.

This certainly implies that F is a continuous function (although note that not all continuous F can be written in this way). In particular, $\mathbb{P}(X = x) = F(x) - \lim_{y \uparrow x} F(y) = 0$ for any x. The density function is not unique; for example, we can change the value of f at any single point without affecting the integral of f. At points where F is differentiable, it is natural to take $f(x) = F'(x)$. For any $a < b$, we have

$$\mathbb{P}(a \le X \le b) = \int_a^b f(u)du.$$

1.2.3 Median

The median of a distribution with cumulative distribution function F is $m \in \mathbb{R}$ such that $F(m) = 1/2$, if such m is unique, or the midpoint $m = (a + b)/2$ of the interval (a, b), where where $F(x) = 1/2$, $x \in (a, b)$.

1.3 Expectation and variance

Let X be a discrete random variable with probability mass function $p_X(x) = \mathbb{P}(X = x)$. The **expectation** (or **mean**) of X is defined by

$$\mathbb{E}\,X = \mathbb{E}\,(X) = \sum_x x p_X(x), \tag{1.2}$$

when this sum converges.

If instead X is a continuous random variable with density function f, then its expectation is given by

$$\mathbb{E}\,X = \mathbb{E}\,(X) = \int_{-\infty}^{\infty} xf(x)dx, \tag{1.3}$$

when this integral converges.

We often want to express the expectation of a function of a random variable X in terms of the density function or the mass function of X. We have

$$\mathbb{E}\,g(X) = \sum_x g(x)p_X(x)$$

in the discrete case, and

$$\mathbb{E}\,g(X) = \int_{-\infty}^{\infty} g(x)f(x)dx$$

in the continuous case, always provided that the **expectation exists**, i.e. that the sum or integral converges.

It is rather unsatisfactory that we have two different definitions of expectation for two different cases, and no definition at all for random variables which are neither continuous nor discrete. In fact it is not difficult to unify the definitions. A very natural way is to consider approximations of a general random variable by discrete random variables. This is analogous to the construction of the integral of a general function by defining the integral of a step function using sums, and then defining the integral of a general function using an approximation by step functions, which you saw in last year's analysis course.

This unifies the two definitions above, and extends the definition to all types of random variable, whether discrete, continuous or neither. We will not pursue this here – but we will collect together basic properties of expectation which we will use constantly:

(1) For any event A, write $\mathbf{1}_A$ for the indicator function of A. Then $\mathbb{E}\,\mathbf{1}_A = \mathbb{P}(A)$.

(2) If $\mathbb{P}(X \geq 0) = 1$, then $\mathbb{E}\,X \geq 0$.

(3) **(Linearity 1):** $\mathbb{E}\,(aX) = a\mathbb{E}\,X$ for any constant a.

(4) **(Linearity 2):** $\mathbb{E}\,(X + Y) = \mathbb{E}\,X + \mathbb{E}\,Y$.

Formally, $\mathbb{E}$ is a linear operator on the vector space of random variables $X\colon \Omega \to \mathbb{R}$ whose expectations exist.

1.3.1 Variance and covariance

The **variance** of a random variable X is defined by

$$\mathrm{Var}(X) = \mathbb{E}\,[(X - \mathbb{E}\,X)^2],$$

provided that the expectations exist. This can then alternatively be expressed as

$$\mathrm{Var}(X) = \mathbb{E}\,(X^2) - (\mathbb{E}\,X)^2.$$

The **covariance** of two random variables X and Y is defined by

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

if the expectations exist, which can then alternatively be expressed as

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

Note that $\mathrm{Var}(X) = \mathrm{Cov}(X,X)$. From the linearity of expectation, we get a bi-linearity property for covariance:

$$\mathrm{Cov}(aX+b, cY+d) = ac\,\mathrm{Cov}(X,Y).$$

As a special case we can obtain

$$\mathrm{Var}(aX+b) = a^2\,\mathrm{Var}(X).$$

We also have the property

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X,Y)$$

and more generally

$$\mathrm{Var}(X_1+X_2+\cdots+X_n) = \sum_{i=1}^{n}\mathrm{Var}(X_i) + 2\sum_{1\le i<j\le n}\mathrm{Cov}(X_i,X_j),$$

always provided that all expectations exist.

1.4 Independence

Events A and B are **independent** if

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

More generally, a family of events $(A_i, i\in\mathcal{I})$, possibly infinite, even uncountable, is called independent if for all finite subsets J of $\mathcal{I}$,

$$\mathbb{P}\left(\bigcap_{i\in J} A_i\right) = \prod_{i\in J}\mathbb{P}(A_i).$$

Remark 1.2. Remember that this is a stronger condition than pairwise independence! Even for three events, it is possible that A_1, A_2 are independent, A_2, A_3 are independent and A_1, A_3 are independent but that A_1, A_2, A_3 are **not** independent.

Random variables $X_1, X_2, \ldots, X_n$ are independent if for all $B_1, B_2, \ldots, B_n \subset \mathbb{R}$, the events $\{X_1\in B_1\}$, $\{X_2\in B_2\}, \ldots, \{X_n\in B_n\}$ are independent.

In fact, it turns out to be enough to check that for all $x_1, x_2, \ldots, x_n$,

$$\begin{aligned}\mathbb{P}(X_1\le x_1,\ldots,X_n\le x_n) &= \mathbb{P}(X_1\le x_1)\cdots\mathbb{P}(X_n\le x_n)\\ &= F_{X_1}(x_1)\cdots F_{X_n}(x_n).\end{aligned}$$

If the random variables are all discrete, another equivalent condition is that

$$\mathbb{P}(X_1 = x_1, \dots, X_n = x_n) = \mathbb{P}(X_1 = x_1) \cdots \mathbb{P}(X_n = x_n).$$

When X and Y are independent random variables whose expectations exist, we have $\mathbb{E}\,(XY) = \mathbb{E}\,(X)\mathbb{E}\,(Y)$, or equivalently $\mathrm{Cov}(X, Y) = 0$. That is, X and Y are uncorrelated. The converse is *not* true; uncorrelated does not imply independent!

Various of the properties above can be summarised by the phrase "**independence means multiply**".

1.5 Examples of probability distributions

We review some of the families of probability distributions which are of particular importance in applications and in theory.

1.5.1 Continuous distributions

Uniform distribution

X has the uniform distribution on an interval $[a, b]$ if its probability density function is given by

$$f(x) = \begin{cases} \frac{1}{b-a}, & \text{if } a \le x \le b, \\ 0, & \text{otherwise.} \end{cases}$$

We write $X \sim U[a, b]$.

Exponential distribution

X has exponential distribution with parameter (or *rate*) λ if its distribution function is given by

$$F(x) = \begin{cases} 0, & x < 0, \\ 1 - e^{-\lambda x}, & x \ge 0. \end{cases}$$

Its density function is

$$f(x) = \begin{cases} 0 & x < 0 \\ \lambda e^{-\lambda x} & x \ge 0 \end{cases}.$$

We write $X \sim \mathrm{Exp}(\lambda)$. We have $\mathbb{E}\,X = 1/\lambda$ and $\mathrm{Var}\,X = 1/\lambda^2$. If $X \sim \mathrm{Exp}(\lambda)$ and $a > 0$, then $aX \sim \mathrm{Exp}(\lambda/a)$. An important property of the distribution is the **memoryless property**: $\mathbb{P}(X > x + t \,|\, X > t)$ does not depend on t.

Normal distribution

X has the normal (or Gaussian) distribution with mean μ and variance σ^2 if its density function is given by

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

We write $X \sim N(\mu, \sigma^2)$. The **standard normal** distribution is $N(0, 1)$.

If $X \sim N(\mu, \sigma^2)$ then $aX + b \sim N(a\mu + b, a^2\sigma^2)$. In particular, $(X - \mu)/\sigma$ has standard normal distribution.

If $X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$ are independent, $X + Y \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$.

The normal distribution has an extremely important role in probability theory, exemplified by the fact that it appears as the limit in the Central Limit Theorem.

We often write Φ for the distribution function of the standard normal distribution:

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) dz.$$

Gamma distribution

The family of gamma distributions generalises the family of exponential distributions. The gamma distribution with *rate* λ and *shape* r has density

$$f(x) = \begin{cases} \frac{\lambda^r}{\Gamma(r)} x^{r-1} e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}.$$

Here $\Gamma(r)$ is the gamma function, defined by $\Gamma(r) = \int_0^\infty z^{r-1}e^{-z}dz$. It is the analytic continuation of the factorial function, in that $\Gamma(r) = (r-1)!$ when r is an integer.

A gamma distribution with shape $r = 1$ is an exponential distribution.

If $X \sim \text{Gamma}(r_X, \lambda)$ and $Y \sim \text{Gamma}(r_Y, \lambda)$ are independent, then we have $X + Y \sim \text{Gamma}(r_X + r_Y, \lambda)$. As a special case, if $X_1, X_2, \dots, X_n$ are i.i.d. with $\text{Exp}(\lambda)$ distribution, then $X_1 + X_2 + \cdots + X_n$ has $\text{Gamma}(n, \lambda)$ distribution.

1.5.2 Discrete distributions

Discrete uniform distribution

X has the discrete uniform distribution on a set B of size n (for example the set $\{1, 2, \dots, n\}$) if

$$p_X(x) = \begin{cases} 1/n, & x \in B \\ 0, & x \notin B \end{cases}.$$

Bernoulli distribution

X has Bernoulli distribution with parameter p if

$$p_X(1) = p, \;\; p_X(0) = 1 - p$$

(and so of course $p_X(x) = 0$ for other values of x).

We have $\mathbb{E}\, X = p$ and $\operatorname{Var} X = p(1 - p)$.

If A is an event with $\mathbb{P}(A) = p$, then its indicator function $\mathbf{1}_A$ has Bernoulli distribution with parameter p.

Binomial distribution

If $X_1, X_2, \dots, X_n$ are i.i.d. Bernoulli random variables with the same parameter p, then their sum $X_1 + \cdots + X_n$ has Binomial distribution with parameters n and p.

Equivalently, if $A_1, \dots, A_n$ are independent events, each with probability p, then the total number of those events which occur has Binomial(n, p) distribution.

If $X \sim$ Binomial(n, p) then

$$p_X(k) = \binom{n}{k} p^k (1-p)^{n-k} \quad \text{for } k \in \{0, 1, \dots, n\}.$$

$\mathbb{E}\, X = np$ and $\operatorname{Var} X = np(1-p)$.

Geometric distribution

Let $p \in (0, 1)$ and let X have mass function

$$p_X(k) = (1-p)^{k-1} p \quad \text{for } k \in \{1, 2, 3, \dots\}.$$

Let $Y = X - 1$; then Y has mass function

$$p_Y(k) = (1-p)^k p \quad \text{for } k \in \{0, 1, 2, \dots\}.$$

The terminology is not consistent; either X or Y might be said to have a geometric distribution with parameter p. (Or even sometimes with parameter $1-p$).

If we have a sequence of independent trials, with probability p of success at each trial, then X could represent the number of trials needed for the first success to occur, while Y could represent the number of failures needed before the first success occurs.

We have

$$\mathbb{P}(X > k) = \mathbb{P}(Y \geq k) = (1-p)^k \quad \text{for } k = 0, 1, 2, \dots.$$

The geometric distribution can be thought of as a discrete anaologue of the exponential distribution. It too has a memoryless property; for $k, m \in \{0, 1, 2, \dots\}$, the conditional probability $\mathbb{P}(X > k + m \,|\, X > k)$ does not depend on k.

Poisson distribution

X has Poisson distribution with mean λ if

$$\mathbb{P}(X = r) = \frac{e^{-\lambda} \lambda^r}{r!} \quad \text{for } r = 0, 1, 2, \dots.$$

$\mathbb{E}\, X = \lambda$ and $\operatorname{Var} X = \lambda$.

If $X \sim$ Poisson(λ) and $Y \sim$ Poisson(μ) are independent, then $X + Y \sim$ Poisson$(\lambda + \mu)$.

The Poisson distribution arises in many applications; it is a good model for the total number of events occurring when there are a large number of possible events which each occur with small probability. There are close connections between the Poisson distribution and the exponential distribution, which we will see in detail when we study Poisson processes at the end of the course.

2

Convergence of random variables, and limit theorems

Let X and Y be random variables. What might we mean by saying "X and Y are close"?

(1) We might be describing a particular realisation. For example, we made an observation of X and Y, and on this occasion $|X - Y| < \epsilon$.

(2) We might be making a statement about the joint distribution of X and Y, for example

$$\mathbb{P}(|X - Y| < \epsilon) > 1 - \epsilon,$$

or

$$\mathbb{E}\big(|X - Y|\big) < \epsilon.$$

(3) We might be comparing the distribution of X with the distribution of Y, for example

$$|F_X(x) - F_Y(x)| < \epsilon \text{ for all } x.$$

Correspondingly, there are several different things we might mean when we say that a sequence of random variables converges to a limit.

2.1 Modes of convergence

Let $X_1, X_2, \dots$ and X be random variables.

Note that $\{X_n \to X \text{ as } n \to \infty\}$ is an **event**. More formally we could write

$$\{\omega \in \Omega \colon X_n(\omega) \to X(\omega) \text{ as } n \to \infty\}$$

to emphasise the dependence on ω; the event might hold for some ω but not for others.

Definition. *$X_n \to X$ **almost surely** (or "with probability 1") if*

$$\boxed{\mathbb{P}\left(X_n \to X \ \textit{as}\ n \to \infty\right) = 1.} \tag{2.1}$$

We often abbreviate to "$X_n \to X$ a.s.".

Definition. $X_n \to X$ ***in probability*** *(written* $X_n \xrightarrow{P} X$*) if for every* $\epsilon > 0$,

$$\boxed{\mathbb{P}\Big(\big|X_n - X\big| < \epsilon\Big) \to 1 \text{ as } n \to \infty.} \tag{2.2}$$

Let $F_1, F_2, \dots$ and F be the distribution functions of $X_1, X_2, \dots$ and X respectively.

Definition. $X_n \to X$ ***in distribution*** *(or* ***weakly****), written* $X_n \xrightarrow{d} X$*, if, for every* x *such that* F *is continuous at* x*,*

$$\boxed{F_n(x) \to F(x) \text{ as } n \to \infty.} \tag{2.3}$$

We will see later that these formulations are in decreasing order of strength.

2.2 Convergence in distribution

Notice that in the definition of convergence in distribution in (2.3), the random variables involved appear only through their distributions. Hence we do not even need all the random variables to be defined on the same probability space. This is really a definition about convergence of *distributions*, not about convergence of random variables. The *joint* distribution of the random variables does not need to be defined. This is in contrast to the definitions of almost sure convergence in (2.1) and of convergence in probability in (2.2), where we genuinely do need all the random variables to be defined on the same space.

As a result, we might sometimes vary the notation by writing a distribution rather than a random variable on the right-hand side; e.g. "$X_n \xrightarrow{d} N(0,1)$" if the limit in distribution is the standard normal, or "$X_n \xrightarrow{d} U[0,1]$" if the limit in distribution is the uniform distribution on $[0,1]$.

In many cases the limit will be deterministic; e.g. if the limit is a distribution which puts all its mass at the value 0, then we will write $X_n \xrightarrow{d} 0$.

In (2.3), why did we ask for the limit to hold only for x which are continuity points of F, rather than at all x? The first couple of examples (which are almost trivial) make this clear.

Example 2.1. Let X_n have the uniform distribution on the interval $[-1/n, 1/n]$. Then $F_n(x) \to 0$ for all $x < 0$, and $F_n(x) \to 1$ for all $x > 0$.

So we have $X_n \xrightarrow{d} 0$, i.e. the distribution of X_n converges to that of a deterministic random variable which is equal to 0 with probability 1. Such a random variable has distribution function given by $F(x) = 0$ for $x < 0$ and $F(x) = 1$ for $x \geq 0$.

Note that $F_n(0) = 1/2$ for all n, while $F(0) = 1$. So convergence does not hold at the point 0 itself (but this is OK, since 0 is not a continuity point of F).

Example 2.2. Let X_n be a deterministic random variable taking the value $1/n$ with probability 1. Let X be a deterministic random variable taking the value 0 with probability 1 (as above). Then once again, $X_n \xrightarrow{d} X$, (even though $\mathbb{P}(X_n \leq 0) = 0$ for all n while $\mathbb{P}(X \leq 0) = 1$).

There are many situations in which a sequence of discrete random variables converges to a continuous limit. Here is one example, showing that a geometric distribution with a small parameter is well approximated by an exponential distribution:

Example 2.3. Let X_n have geometric distribution on the positive integers, with parameter p_n, i.e. $\mathbb{P}(X_n = k) = (1-p_n)^{k-1}p_n$ for $k = 1, 2, \ldots$. Show that if $p_n \to 0$ as $n \to \infty$, then p_nX_n converges in distribution to the exponential distribution with mean 1.

Solution: We have $\mathbb{P}(X_n > k) = (1-p_n)^k$ for $k = 0, 1, 2, \ldots$. For $x \geq 0$, we have

$$\begin{aligned}\mathbb{P}(p_nX_n > x) &= \mathbb{P}\left(X_n > \frac{x}{p_n}\right)\\ &= \mathbb{P}\left(X_n > \left\lfloor\frac{x}{p_n}\right\rfloor\right)\\ &= (1-p_n)^{\lfloor x/p_n \rfloor}\\ &\to e^{-x} \text{ as } n \to \infty\end{aligned}$$

because $p_n \to 0$; here we use the fact that $(1-\epsilon)^{x/\epsilon} \to e^{-x}$ as $\epsilon \to 0$, and also that $\lfloor x/p_n \rfloor - x/p_n$ is bounded.

Hence if F_n is the distribution function of p_nX_n, then $1 - F_n(x) \to e^{-x}$ as $n \to \infty$. So $F_n(x) \to 1 - e^{-x}$ for all $x > 0$, while $F_n(x) = 0$ for all $x \leq 0$ and all n.

So indeed $F_n(x) \to F(x)$ for all x, where F is the distribution function of a random variable with Exp(1) distribution.

There are several more examples on the problem sheets.

2.3 Comparison of different modes of convergence

Theorem 2.4. *The following implications hold:*

$$\boxed{X_n \to X \textit{ almost surely}} \quad \Rightarrow \quad \boxed{X_n \to X \textit{ in probability}} \quad \Rightarrow \quad \boxed{X_n \to X \textit{ in distribution}}$$

The reverse implications do not hold in general.

Before starting the proof we note a useful fact, which is a simple consequence of the countable additivity axiom for unions of disjoint sets (1.1).

Lemma 2.5. *Let $A_n, n \geq 1$, be an increasing sequence of events; that is, $A_1 \subseteq A_2 \subseteq A_3 \subseteq \ldots$. Then*

$$\lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}\left(\bigcup_{n\geq 1} A_n\right). \tag{2.4}$$

Proof. Because the sequence A_n is increasing, it is easy to rewrite the union as a disjoint union:

$$\begin{aligned}\mathbb{P}\left(\bigcup_{n\geq 1} A_n\right) &= \mathbb{P}\Big(A_1 \cup \bigcup_{n\geq 1}(A_{n+1} \setminus A_n)\Big)\\ &= \mathbb{P}(A_1) + \sum_{i=1}^{\infty} \mathbb{P}\big(A_{i+1} \setminus A_i\big) \qquad \text{(using countable additivity)}\\ &= \mathbb{P}(A_1) + \lim_{n\to\infty} \sum_{i=1}^{n-1} \mathbb{P}\left(A_{i+1} \setminus A_i\right)\end{aligned}$$

$$\begin{aligned}
&= \lim_{n\to\infty} \Big(\mathbb{P}(A_1) + \sum_{i=1}^{n-1} \mathbb{P}\left(A_{i+1} \setminus A_i\right)\Big) \\
&= \lim_{n\to\infty} \mathbb{P}\Big(A_1 \cup \bigcup_{1\le i\le n-1} (A_{i+1} \setminus A_i)\Big) \\
&= \lim_{n\to\infty} \mathbb{P}(A_n).
\end{aligned}$$

□

Proof of Theorem 2.4.
(1) First we will show that convergence in probability implies convergence in distribution. Let F_n be the distribution function of X_n, and F the distribution function of X. Fix any x such that F is continuous at x, and fix any $\epsilon > 0$.

Observe that if $X_n \le x$, then either $X \le x + \epsilon$ or $|X_n - X| > \epsilon$. Hence

$$\begin{aligned}
F_n(x) &= \mathbb{P}(X_n \le x) \\
&\le \mathbb{P}\big(X \le x + \epsilon \text{ or } |X_n - X| > \epsilon\big) \\
&\le \mathbb{P}\big(X \le x + \epsilon\big) + \mathbb{P}\big(|X_n - X| > \epsilon\big) \\
&\to F(x+\epsilon) \text{ as } n \to \infty,
\end{aligned}$$

using the convergence in probability. So $F_n(x) < F(x+\epsilon) + \epsilon$ for all large enough n.

Similarly by looking at $1 - F_n(x) = \mathbb{P}(X_n > x)$, we can obtain that $F_n(x) > F(x-\epsilon) - \epsilon$ for all large enough n.

Since $\epsilon > 0$ is arbitrary, and since F is continuous at x, this implies that $F_n(x) \to F(x)$ as $n \to \infty$.

---

(2) For convergence in distribution, we do not need the random variables to be defined on the same probability space. But even if they are, convergence in distribution does not imply convergence in probability. For example, suppose that X and Y are random variables with the same distribution but with $\mathbb{P}(X = Y) < 1$. Then the sequence $X, X, X, \dots$ converges to Y in distribution, but not in probability.

---

(3) Now we will show that almost sure convergence implies convergence in probability. Fix $\epsilon > 0$ and for $N \in \mathbb{N}$, define the event A_N by

$$A_N = \{|X_n - X| < \epsilon \text{ for all } n \ge N\}.$$

Suppose that $X_n \to X$ almost surely. If the event $\{X_n \to X\}$ occurs, then the event A_N must occur for some N, so we have $\mathbb{P}\big(\bigcup A_N\big) = 1$. A_N is an increasing sequence of events, so (2.4) then gives $\lim_{N\to\infty} \mathbb{P}(A_N) = 1$.

But A_N implies $\big|X_N - X\big| < \epsilon$, giving $\mathbb{P}\Big(\big|X_N - X\big| < \epsilon\Big) \to 1$. Since ϵ is arbitrary, this means that $X_n \to X$ in probability, as desired.

---

(4) Finally we want to show that convergence in probability does not imply almost sure convergence.

Consider a sequence of independent random variables X_n where $\mathbb{P}(X_n = 1) = 1/n$ and $\mathbb{P}(X_n = 0) = (n-1)/n$.

We have $X_n \to 0$ in probability as $n \to \infty$ because for any $\epsilon > 0$, $\mathbb{P}(|X_n - 0| < \epsilon) \geq \mathbb{P}(X = 0) \to 1$.

Since X_n only take the values 0 and 1, the event $\{X_n \to 0\}$ is the same as the event $\{X_n = 0 \text{ eventually}\}$. This is $\bigcup_{N \geq 1} B_N$ where $B_N = \{X_n = 0 \text{ for all } n \geq N\}$.

But for any N and K,

$$\begin{aligned}\mathbb{P}(B_N) \leq \mathbb{P}(X_n = 0 \text{ for all } n = N, \dots, N+K) &= \frac{N-1}{N}\frac{N}{N+1}\frac{N+1}{N+2}\cdots\frac{N+K-1}{N+K} \\ &= \frac{N-1}{N+K}.\end{aligned}$$

As $K \geq 1$ is arbitrary, we obtain that $\mathbb{P}(B_N) = 0$. Hence also by Lemma 2.5, $\mathbb{P}(\bigcup_{N\geq 1} B_N) = 0$, and so $\mathbb{P}(X_n \to 0) = 0$. Hence it is not the case that X_n converges to 0 almost surely. □

Although convergence in distribution is weaker than convergence in probability, there is a partial converse, for the case when the limit is deterministic:

Theorem 2.6. *Let $X_1, X_2, \dots$ be a sequence of random variables defined on the same probability space. If $X_n \to c$ in distribution where c is some constant, then also $X_n \to c$ in probability.*

Proof. Exercise (see problem sheet). □

2.4 Review: Weak law of large numbers

This was covered at the end of last year's course (without explicitly introducing the notion of convergence in probability).

Let $S_n = X_1 + X_2 + \cdots + X_n$, where X_i are i.i.d. with mean μ. The law of large numbers tells us that, roughly speaking, S_n behaves to first order like $n\mu$ as $n \to \infty$. The weak law phrases this in terms of convergence in probability. (Later we will see a stronger result in terms of almost sure convergence).

Theorem (Weak Law of Large Numbers)**.** *Let $X_1, X_2, \dots$ be i.i.d. random variables with finite mean μ. Let $S_n = X_1 + X_2 + \cdots + X_n$. Then*

$$\frac{S_n}{n} \xrightarrow{P} \mu \text{ as } n \to \infty.$$

That is, for all $\epsilon > 0$,

$$\mathbb{P}\left(\left|\frac{S_n}{n} - \mu\right| < \epsilon\right) \to 1 \text{ as } n \to \infty. \tag{2.5}$$

Given Theorem 2.6, we could equivalently write $\frac{S_n}{n} \xrightarrow{d} \mu$.

We will give an extremely simple proof of the weak law of large numbers, under an additional condition (that the X_i have finite variance). To do this, we need some results which give probability bounds on the tail of a distribution in terms of its mean and variance.

Theorem (Markov's inequality)**.** *Let X be random variable taking non-negative values (i.e. $\mathbb{P}(X \geq 0) = 1$). Then for any $z > 0$,*

$$\mathbb{P}(X \geq z) \leq \frac{\mathbb{E}\,X}{z}. \tag{2.6}$$

Proof. We consider a random variable $X_z = z\mathbf{1}\{X \geq z\}$. So X_z takes the value 0 whenever X is in $[0, z)$ and the value z whenever X is in $[z, \infty)$. So $X \geq X_z$ always (here we use the fact that X is non-negative).

Then $\mathbb{E}\,X \geq \mathbb{E}\,X_z = z\mathbb{E}\,\mathbf{1}\{X \geq z\} = z\mathbb{P}(X \geq z)$. Rearranging gives the result. □

Theorem (Chebyshev's inequality)**.** *Let Y be a random variable with finite mean and variance. Then for any $\epsilon > 0$,*

$$\mathbb{P}\big(|Y - \mathbb{E}\,Y| \geq \epsilon\big) \leq \frac{\operatorname{Var} Y}{\epsilon^2}.$$

Proof.

$$\begin{aligned}\mathbb{P}\big(|Y - \mathbb{E}\,Y| \geq \epsilon\big) &= \mathbb{P}\big([Y - \mathbb{E}\,Y]^2 \geq \epsilon^2\big)\\ &\leq \frac{\mathbb{E}\left([Y - \mathbb{E}\,Y]^2\right)}{\epsilon^2}\end{aligned}$$

(by applying Markov's inequality (2.6) with $X = [Y - \mathbb{E}\,Y]^2$ and $z = \epsilon^2$)

$$= \frac{\operatorname{Var} Y}{\epsilon^2}.$$

□

Proof of the weak law of large numbers in the case of random variables with finite variance. Let X_i be i.i.d. with mean μ and variance σ^2. Recall $S_n = X_1 + \cdots + X_n$. We want to show that $S_n/n \xrightarrow{P} \mu$ as $n \to \infty$.

We have $\mathbb{E}\,(S_n/n) = \mu$, and (using the independence of the X_i),

$$\begin{aligned}\operatorname{Var}\left(\frac{S_n}{n}\right) &= \frac{\operatorname{Var} S_n}{n^2}\\ &= \frac{\operatorname{Var} X_1 + \cdots + \operatorname{Var} X_n}{n^2}\\ &= \frac{n\sigma^2}{n^2}\\ &= \frac{\sigma^2}{n}.\end{aligned}$$

Fix any $\epsilon > 0$. Using Chebyshev's inequality applied to the random variable S_n/n, we have

$$\begin{aligned}\mathbb{P}\left(\left|\frac{S_n}{n} - \mu\right| \geq \epsilon\right) &\leq \frac{\operatorname{Var}\left(\frac{S_n}{n}\right)}{\epsilon^2}\\ &= \frac{\sigma^2}{n\epsilon^2}\\ &\to 0 \text{ as } n \to \infty.\end{aligned}$$

So indeed (2.5) holds, as required. □

Remark. Observe that we could relax considerably the assumptions in the weak law of large numbers, and still get the same result using almost the same proof. We never used at all the assumption that the X_i all had the same distribution. We could also relax the assumption that the X_i are independent, as long as the variance of S_n grows more slowly than n^2. For example, if we have an upper bound on the variance of each X_i, and a bound which is $o(n^2)$ on the sum $\sum_{1\leq i<j\leq n}\operatorname{Cov}(X_i, X_j)$, then exactly the same idea applies to show that $(S_n - \mathbb{E}\, S_n)/n$ converges to 0 in distribution.

2.5 Strong law of large numbers

In the weak law of large numbers, we proved convergence in distribution of the average of i.i.d. random variables to the mean. The strong law says more: the convergence occurs with probability 1.

Theorem (Strong Law of Large Numbers)**.** *Let $X_1, X_2, \dots$ be i.i.d. with mean μ. Let $S_n = X_1 + \cdots + X_n$. Then*

$$\frac{S_n}{n} \to \mu \textit{ almost surely as } n \to \infty.$$

2.5.1 Proof of the strong law of large numbers (non-examinable)

A proof of the strong law of large numbers in full generality is somewhat involved. A nice proof uses martingales and is part of next year's course on Probability, Measure and Martingales.

However, if we assume an extra condition, namely that the distribution has a finite fourth moment, then a relatively straightforward proof is possible. [NB the proof is not examinable.]

Proof of Strong Law of Large Numbers, under the additional condition $\mathbb{E}\, X_n^4 < \infty$.
Let us centre the X_n, writing $W_n = X_n - \mu$.

Then $\mathbb{E}\, W_n = 0$, and we have $\mathbb{E}\, X_n^4 < \infty \Rightarrow \mathbb{E}\, W_n^4 < \infty$ (exercise).

Note also that

$$\begin{aligned}(\mathbb{E}\, W_n^2)^2 &= \mathbb{E}\,(W_n^4) - \operatorname{Var}(W_n^2)\\ &\leq \mathbb{E}\,(W_n^4).\end{aligned}$$

We will consider $\mathbb{E}\,\left[(S_n - n\mu)^4\right]$. Expanding the fourth power and using linearity of expectation, we obtain

$$\begin{aligned}\mathbb{E}\,\left[(S_n - n\mu)^4\right] &= \mathbb{E}\,\left[(W_1 + W_2 + \cdots + W_n)^4\right]\\ &= \sum_{1\leq i\leq n}\mathbb{E}\, W_i^4 + 4\sum_{\substack{1\leq i,j\leq n\\ i\neq j}}\mathbb{E}\, W_i^3 W_j + 3\sum_{\substack{1\leq i,j\leq n\\ i\neq j}}\mathbb{E}\, W_i^2 W_j^2\\ &\qquad + 6\sum_{\substack{1\leq i,j,k\leq n\\ i,j,k \text{ distinct}}}\mathbb{E}\, W_i^2 W_j W_k + \sum_{\substack{1\leq i,j,k,l\leq n\\ i,j,k,l \text{ distinct}}}\mathbb{E}\, W_i W_j W_k W_l.\end{aligned}$$

(The exact constants in front of the sums are not too important!) Using independence and $\mathbb{E}\, W_i = 0$, most of these terms vanish. For example, $\mathbb{E}\, W_i^3 W_j = \mathbb{E}\, W_i^3 \mathbb{E}\, W_j = 0$. We are left with only

$$\begin{aligned}\mathbb{E}\left[(S_n - n\mu)^4\right] &= n\mathbb{E}\,W_1^4 + 3n(n-1)\left(\mathbb{E}\left[W_1^2\right]\right)^2 \\ &\le 3n^2\mathbb{E}\,W_1^4.\end{aligned}$$

From this we have

$$\begin{aligned}\mathbb{E}\left[\sum_{n=1}^{\infty}\left(\frac{S_n}{n} - \mu\right)^4\right] &= \sum_{n=1}^{\infty}\mathbb{E}\left[\left(\frac{S_n}{n} - \mu\right)^4\right] \\ &= \sum_{n=1}^{\infty}\frac{1}{n^4}\mathbb{E}\left[(S_n - n\mu)^4\right] \\ &\le \sum_{n=1}^{\infty}\frac{3\mathbb{E}\,W_1^4}{n^2} \\ &< \infty.\end{aligned}$$

Formally, interchanging the infinite series and expectation in the first line requires a justification refining the notion of absolute convergence to a framework involving general expectations, which is beyond the scope of this course in the present generality. This is not so hard for discrete random variables (changing the order of terms in absolutely convergent series). The theory for such interchanging of series and integrals is developed in next term's course on Integration and transferred to a general setting of expectations on probability spaces at the very beginning of next year's course on Probability, Measure and Martingales.

But if Z is a random variable with $\mathbb{E}\,Z < \infty$, then certainly $\mathbb{P}(Z < \infty) = 1$. Applying this with $Z = \left(\frac{S_n}{n} - \mu\right)^4$, we get

$$\mathbb{P}\left(\sum_{n=1}^{\infty}\left(\frac{S_n}{n} - \mu\right)^4 < \infty\right) = 1.$$

Finally, if $\sum(a_n - \mu)^4$ is finite, then certainly $a_n \to \mu$ as $n \to \infty$. So we can conclude that

$$\mathbb{P}\left(\frac{S_n}{n} \to \mu \text{ as } n \to \infty\right) = 1,$$

as required. □

2.6 Central limit theorem

The weak law of large numbers tells us that the distribution of S_n/n concentrates around μ as n becomes large. The **central limit theorem** (CLT) goes much further, telling us that (if the random variables X_i have finite variance) the "fluctuations" of S_n around $n\mu$ are of order $\sqrt{n}$. Moreover, the behaviour of these fluctuations is *universal*; whatever the distribution of the X_i, if we scale $S_n - n\mu$ by $\sqrt{n}$, we obtain a normal distribution as the limit as $n \to \infty$.

Theorem (Central Limit Theorem)**.** *Let $X_1, X_2, \ldots$ be i.i.d. random variables with mean μ and variance $\sigma^2 \in (0, \infty)$. Let $S_n = X_1 + X_2 + \cdots + X_n$. Then*

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{d} N(0,1) \text{ as } n \to \infty. \tag{2.7}$$

We will prove the CLT later using generating functions.

Remark 2.7. We can summarise the CLT in three stages:

(1) The distribution of S_n concentrates around $n\mu$;

(2) The fluctuations of S_n around $n\mu$ are of order $\sqrt{n}$;

(3) The asymptotic distribution of these fluctuations is normal.

These are somehow in increasing order of refinement. Some students take in the third of these, but not the first two; they remember that the RHS in (2.7) is a normal distribution, but are hazy about what is going on on the LHS. This is a bit perverse; without knowing the scale of the fluctuations, or what they fluctuate around, knowing their distribution is not so useful!

Example 2.8. An insurance company sells 10, 000 similar car insurance policies. They estimate that the amount paid out in claims on a typical policy has mean £240 and standard deviation £800. Estimate how much they need to put aside in reserves to be 99% sure that the reserve will exceed the total amount claimed.

Solution: Let $\mu = £240$, $\sigma = £800$, $n = 10,000$, and note $\Phi^{-1}(0.99) = 2.326$ where Φ is the distribution function of the standard normal.

Let S_n be the total amount claimed. For large n, the Central Limit Theorem tells us that

$$\mathbb{P}\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} < \Phi^{-1}(0.99)\right) \approx 0.99,$$

i.e.

$$\mathbb{P}\left(S_n < \Phi^{-1}(0.99)\sigma\sqrt{n} + n\mu\right) \approx 0.99.$$

So the amount needed in reserves is approximately $\Phi^{-1}(0.99)\sigma\sqrt{n} + n\mu$, which in this case is £2, 586, 080.

Notice that the reserve required per customer is about £258, which is £18 higher than μ. We can see from the calculation above that this surplus is proportional to $n^{-1/2}$. If we had 100 customers rather than 10,000, we would need a surplus 10 times bigger, while with 1,000,000 customers it would be 10 times smaller.

The fact that the amount per customer needed to cover the fluctuations around the mean is decreasing in the number of customers is an example of **risk pooling**.

Of course, the example of identical customers is a bit simplistic, but the effect of risk pooling that we observe is a very real one. Our analysis also assumed that the different customers are independent – is that realistic? For car insurance, it is not such a bad assumption. Similarly for life insurance. In the case of property insurance, it could be a very bad assumption (for example, floods can damage many properties simultaneously). In that situation, the effect of risk pooling is a lot smaller (which explains why obtaining insurance for a property subject to a risk of flooding can be problematic, even if the risk is not that great).

Example 2.9 (Binomial distribution: CLT and Poisson approximation)**.** Let $p \in (0, 1)$ and let Y_n have Binomial(n, p) distribution. Then we can write $Y_n = X_1 + \cdots + X_n$ where the

X_i are i.i.d. Bernoulli(p) random variables. (We can think of the X_i as indicator functions of independent events, all with the same probability, e.g. arising from random sampling).

The X_i each have mean p and variance $p(1-p)$. So we can apply the CLT to obtain

$$\frac{Y_n - np}{\sqrt{n}} \xrightarrow{d} N\big(0, p(1-p)\big) \qquad \text{as } n \to \infty.$$

Now instead of considering fixed p, consider random variables W_n with Binomial(n, p_n) distribution, where $p_n \to 0$ as $n \to \infty$. Now a very different limit applies, describing a situation in which we have a very large number of trials but each one has a very small probability of success. Let $\lambda_n = np_n$, the mean of W_n. Suppose that λ_n converges to a limit λ as $n \to \infty$, so that the expected total number of successes stays approximately constant. Then we will show that W_n converges in distribution to Poisson(λ).

It is enough to show (check!) that for each fixed $k = 0, 1, \dots,$

$$\mathbb{P}(W_n = k) \to \frac{\lambda^k}{k!} e^{-\lambda}$$

(since the RHS is the probability that a Poisson(λ) random variable takes the value k).

We have

$$\begin{aligned}
\mathbb{P}(W_n = k) &= \binom{n}{k} p_n^k (1-p_n)^{n-k} \\
&= \binom{n}{k} \left(\frac{\lambda_n}{n}\right)^k \left(1 - \frac{\lambda_n}{n}\right)^{n-k} \\
&= \frac{n(n-1)\cdots(n-k+1)}{n^k} \frac{\lambda_n^k}{k!} \left(1 - \frac{\lambda_n}{n}\right)^n \left(1 - \frac{\lambda_n}{n}\right)^{-k} \\
&\to 1 \cdot \frac{\lambda^k}{k!} e^{-\lambda} \cdot 1
\end{aligned}$$

as $n \to \infty$, as desired.

3

Generating functions

3.1 Review of probability generating functions

Let X be a random variable taking non-negative integer values, and let p_X be its probability mass function. The **probability generating function** of X is defined by

$$G(z) := \mathbb{E}\left(z^X\right) = \sum_{k=0}^{\infty} p_X(k) z^k.$$

G is a power series whose radius of convergence is at least 1. We can recover the coefficients of the power series, i.e. the values of the function p_X, from the behaviour of G and its derivatives at the point 0, and we can compute the moments of X from the behaviour of G and its derivatives at 1:

Theorem 3.1.

(a) $G^{(k)}(0) = k!\, p_X(k)$ *for* $k = 0, 1, 2, \ldots$.

(b) $G(1) = 1$ *and* $G^{(k)}(1) = \mathbb{E}\left[X(X-1)\cdots(X-k+1)\right]$ *for* $k = 1, 2, \ldots$.

Here if the radius of convergence of G is exactly 1, then $G^{(k)}(1)$ should be taken to mean $\lim_{z\uparrow 1} G^{(k)}(z)$. In this case, the limit may be finite or infinite.

From Theorem 3.1(a), we see immediately that a distribution is determined by its generating function:

Theorem 3.2 (Uniqueness theorem for probability generating functions)**.** *If X and Y have the same generating function, then they have the same distribution.*

From Theorem 3.1(b), we have, for example, $\mathbb{E}(X) = G'(1)$, $\text{Var}(X) = G''(1) + G'(1) - [G'(1)]^2$.

Generating functions are extremely useful tools for dealing with sums of independent random variables. Let X and Y be independent random variables with generating functions G_X and G_Y. Then the generating function of their sum is given by

$$\begin{aligned} G_{X+Y}(z) &= \mathbb{E}\left(z^{X+Y}\right) \\ &= \mathbb{E}\left(z^X z^Y\right) \end{aligned}$$

$$\begin{aligned} &= \mathbb{E}\left(z^X\right)\mathbb{E}\left(z^Y\right) \text{ (by independence)} \\ &= G_X(z)G_Y(z). \end{aligned}$$

(Again, "independence means multiply").

We can also treat the sum of a random number of random variables. Let $X_1, X_2, \dots$ be i.i.d. random variables (taking non-negative integer values), and let N be another random variable, also taking non-negative integer values, independent of the sequence X_i. Define $S = X_1 + \cdots + X_N$. Then we can write the generating function of S in terms of the common generating function of the X_i and the generating function of N by

$$\begin{aligned} G_S(z) &= \mathbb{E}\left(z^S\right) \\ &= \mathbb{E}\left(z^{X_1+\cdots+X_N}\right) \\ &= \mathbb{E}\left(\mathbb{E}\left(z^{X_1+\cdots+X_N}\middle|N\right)\right) \\ &= \mathbb{E}\left(\mathbb{E}\left(\left(z^{X_1}\right)\right)^N\right) \\ &= \mathbb{E}\left(\left(G_X(z)\right)^N\right) \\ &= G_N\left(G_X(z)\right). \end{aligned}$$

3.2 Moment generating functions

The probability generating function is well-adapted for handling random variables which take non-negative integer values. To treat random variables with general distribution, we now introduce two related objects, the moment generating function and the characteristic function.

The **moment generating function** of a random variable X is defined by

$$M_X(t) := \mathbb{E}\left(e^{tX}\right). \tag{3.1}$$

This expectation may be finite or infinite.

(Note that we could obtain the moment generating function by substituting $z = e^t$ in the definition of the probability generating function above. An advantage of this form is that we can conveniently consider an expansion around $t = 0$, whereas the expansion around $z = 0$, convenient when the random variables took only non-negative integer values, no longer gives a power series in the general case.)

For the same reason as for probability generating functions, the mgf of a sum of independent random variables is the product of the mgfs:

Theorem 3.3.

(a) If $Y = aX + b$, then $M_Y(t) = e^{bt}M_X(at)$.

(b) Let $X_1, \dots, X_n$ be independent random variables, with mgfs $M_{X_1}, \dots, M_{X_n}$. Then the mgf of their sum is given by

$$M_{X_1+\cdots+X_n}(t) = M_{X_1}(t)\dots M_{X_n}(t).$$

Proof. (a): easy exercise. Part (b) is also straightforward:

$$\begin{aligned} M_{X_1+\dots+X_n}(t) &= \mathbb{E}\left(e^{tX_1+\dots+tX_n}\right) \\ &= \mathbb{E}\left(e^{tX_1}\dots e^{tX_n}\right) \\ &= \mathbb{E}\left(e^{tX_1}\right)\dots\mathbb{E}\left(e^{tX_n}\right) \text{ (by independence)} \\ &= M_{X_1}(t)\dots M_{X_n}(t). \end{aligned}$$

□

An immediate disadvantage of the moment generating function is that it may not be well defined. If the positive tail of the distribution is too heavy, the expectation in the definition in (3.1) may be infinite for all $t > 0$: while if the negative tail is too heavy, the expectation may be infinite for all $t < 0$.

For the moment generating function to be useful, we will require $\mathbb{E}\, e^{t_0|X|} < \infty$ for some $t_0 > 0$. That is, X has "finite exponential moments" of some order (equivalently, the tails of the distribution function decay at least exponentially fast). Then (exercise!) the moment generating function is finite for all $t \in (-t_0, t_0)$, and also all the moments $\mathbb{E}\, X^k$ are finite.

Most of the classical distributions that we have looked at are either bounded or have tails that decay at least exponentially (for example uniform, geometric/exponential, normal, Poisson...). However, distributions with heavier tails are also of great importance, especially in many modelling contexts. For those distributions, the moment generating function is of no use; however, we can consider a variant of it, the characteristic function (see later).

The next result explains the terminology "moment generating function"; the mgf of X can be expanded as a power series around 0, in which the coefficients are the moments of X.

Theorem 3.4. *Suppose $M_X(t)$ is finite for $|t| \le t_0$, for some $t_0 > 0$. Then*

(a) $M_X(t) = \sum_{k=0}^{\infty} \frac{t^k \mathbb{E}\,(X^k)}{k!}$ for $|t| \le t_0$.

(b) $M_X^{(k)}(0) = \mathbb{E}\,(X^k)$.

Informal proof.

$$\begin{aligned} M_X(t) &= \mathbb{E}\,(e^{tX}) \\ &= \mathbb{E}\left(1 + tX + \frac{(tX)^2}{2!} + \frac{(tX)^3}{3!} + \dots\right) \\ &= 1 + tE(X) + \frac{t^2\mathbb{E}\,(X^2)}{2!} + \frac{t^3\mathbb{E}\,(X^3)}{3!} + \dots, \end{aligned}$$

using linearity of expectation. This gives (a) and taking derivatives at 0 gives (b). Exchanging expectation with an infinite sum, as we did here, really needs extra justification. In this case there is no problem (for example, it is always fine in the case where the sum of the absolute values also has finite expectation – in this case this gives $\mathbb{E}\, e^{|tX|} < \infty$ which is easily seen to be true); but we do not pursue it further here. □

The following **uniqueness** and **continuity** results will be key to our applications of the moment generating function.

Theorem 3.5. *If X and Y are random variables with the same moment generating function, which is finite on $[-t_0, t_0]$ for some $t_0 > 0$, then X and Y have the same distribution.*

Theorem 3.6. *Suppose Y and X_1, X_2, ... are random variables whose moment generating functions M_Y and $M_{X_1}, M_{X_2}, \ldots$ are all finite on $[-t_0, t_0]$ for some $t_0 > 0$. If*

$$M_{X_n}(t) \to M_Y(t) \text{ as } n \to \infty, \text{ for all } t \in [-t_0, t_0],$$

then

$$X_n \xrightarrow{d} Y \text{ as } n \to \infty.$$

The proofs of the uniqueness and continuity results for mgfs are beyond the scope of the course. They correspond to an inversion theorem from Fourier analysis, by which the distribution function of X can be written in a suitable way as a linear mixture over t of terms $\mathbb{E}\, e^{itX}$.

Example 3.7. Find the moment generating function of the exponential distribution with parameter λ.
Solution:

$$\begin{aligned} M(t) &= \mathbb{E}\,(e^{tX}) \\ &= \int_0^\infty e^{tx} f(x) dx \\ &= \int_0^\infty \lambda e^{tx} e^{-\lambda x} dx \\ &= \frac{\lambda}{\lambda - t} \int_0^\infty (\lambda - t) \exp^{-(\lambda - t)x} dx \\ &= \frac{\lambda}{\lambda - t} \text{ for } t \in (-\infty, \lambda). \end{aligned}$$

In the last step we used the fact that the integrand is the density function of a random variable, namely one with $\mathrm{Exp}(\lambda - t)$ distribution, so that the integral is 1.

Example 3.8. Find the moment generating function of a random variable with $N(\mu, \sigma^2)$ distribution. If $Y_1 \sim N(\mu_1, \sigma_1^2)$ and $Y_2 \sim N(\mu_2, \sigma_2^2)$ are independent, show that $Y_1 + Y_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.
Solution: Let $X \sim N(\mu, \sigma^2)$. Then $X = \sigma Z + \mu$, where Z is standard normal. We have

$$\begin{aligned} M_Z(t) &= \mathbb{E}\,(e^{tZ}) \\ &= \int_{-\infty}^\infty \exp(tz) \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-z^2}{2}\right) dz \\ &= \int_{-\infty}^\infty \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-(z^2 - 2tz)}{2}\right) dz \\ &= \int_{-\infty}^\infty \exp\left(\frac{t^2}{2}\right) \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-(z-t)^2}{2}\right) dz \end{aligned}$$

(this is "completing the square")

$$= \exp\left(\frac{t^2}{2}\right) \int_{-\infty}^\infty \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-(z-t)^2}{2}\right) dz$$

$$= e^{t^2/2}$$

(the same trick as before: the integrand is the density function of $N(t,1)$ so the integral is 1).

Then from the first part of Theorem 3.3, $M_X(t) = e^{\mu t} M_Z(\sigma t) = e^{\mu t + \sigma^2 t^2/2}$.

For the second part,

$$\begin{aligned} M_{Y_1+Y_2}(t) &= M_{Y_1}(t) M_{Y_2}(t) \\ &= e^{\mu_1 t + \sigma_1^2 t^2/2} e^{\mu_2 t + \sigma_2^2 t^2/2} \\ &= e^{(\mu_1+\mu_2)t + (\sigma_1^2+\sigma_2^2)t^2/2}. \end{aligned}$$

Since this is the mgf of $N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$, and it is finite on an interval $[-t_0, t_0]$ (in fact, for all t), the uniqueness theorem for mgfs tells us that indeed that must be the distribution of $Y_1 + Y_2$.

3.2.1 Proof of WLLN and CLT using moment generating functions

Let X_1, $X_2, \ldots$ be a sequence of i.i.d. random variables with finite exponential moments of some order (i.e. their moment generating function is finite on some interval containing the origin in its interior).

Let $S_n = X_1 + X_2 + \cdots + X_n$. We can use moment generating functions to give a very compact proof of the Central Limit Theorem for the sequence S_n (and, even more simply, the Weak Law of Large Numbers).

Let the X_i have mean μ and variance σ^2, and let M be their moment generating function.

Weak law of large numbers

From Taylor's Theorem and the expansion of M as a power series around 0 (Theorem 3.4) we can write, as $h \to 0$,

$$\begin{aligned} M(h) &= M(0) + hM'(0) + o(h) \\ &= 1 + h\mu + o(h). \end{aligned}$$

Here, we use notation "$f(h) = o(g(h))$ as $h \to 0$" to mean $f(h)/g(h) \to 0$ as $h \to 0$. We will similarly use notation "$a_n = o(b_n)$ as $n \to \infty$" to mean $a_n/b_n \to 0$ as $n \to \infty$. Here, specifically, we therefore have $(M(h) - M(0) - hM'(0))/h \to 0$ as $h \to 0$.

Let $\overline{M}_n$ be the mgf of S_n/n. Using the independence of the X_i, we have

$$\begin{aligned} \overline{M}_n(t) &= \mathbb{E}\left(e^{tS_n/n}\right) \\ &= \mathbb{E}\left(e^{tX_1/n} \ldots e^{tX_n/n}\right) \\ &= (M(t/n))^n \\ &= \left(1 + \frac{t}{n}\mu + o(t/n)\right)^n \quad \text{as } n \to \infty \\ &\to e^{t\mu} \text{ as } n \to \infty. \end{aligned}$$

But $e^{t\mu}$ is the mgf of a random variable which takes the constant value μ with probability 1. From the continuity theorem for mgfs, $S_n/n \xrightarrow{d} \mu$ as $n \to \infty$, and we have proved the weak law of large numbers.

Central limit theorem

Let $Y_i = X_i - \mu$, and let M_Y be the mgf of the common distribution of the Y_i. Taking one more term in the Taylor expansion, we have that as $h \to 0$,

$$\begin{aligned} M_Y(h) &= M_Y(0) + hM_Y'(0) + \frac{h^2}{2}M_Y''(0) + o(h^2) \\ &= 1 + h\mathbb{E}\,(Y) + \frac{h^2}{2}\operatorname{Var}(Y) + o(h^2) \\ &= 1 + h^2\sigma^2/2 + o(h^2). \end{aligned}$$

Let $\widetilde{M}_n$ be the mgf of $\frac{S_n - \mu n}{\sigma\sqrt{n}}$. Then we have

$$\begin{aligned} \widetilde{M}_n(t) &= \mathbb{E}\left(\exp\left(\frac{t(S_n-\mu n)}{\sigma\sqrt{n}}\right)\right) \\ &= \mathbb{E}\left(\exp\left(\frac{t(X_1-\mu)}{\sigma\sqrt{n}}\right)\dots\exp\left(\frac{t(X_n-\mu)}{\sigma\sqrt{n}}\right)\right) \\ &= M_Y\left(\frac{t}{\sigma\sqrt{n}}\right)^n \\ &= \left(1 + \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right)\right)^n \quad \text{as } n \to \infty \\ &\to \exp\left(\frac{t^2}{2}\right) \quad \text{as } n \to \infty. \end{aligned}$$

But the last line is the mgf of $N(0,1)$. Using the continuity theorem again,

$$\frac{S_n - \mu n}{\sigma\sqrt{n}} \xrightarrow{d} N(0,1),$$

and we have proved the CLT.

3.3 Using moment generating functions for tail bounds

Using a version of Markov's inequality applied to a random variable like e^{tX}, we can get bounds on the tail of a distribution in terms of the moment generating function which are much better than we would get from, for example, Chebyshev's inequality, which is the application of Markov's inequality to the random variable $(X-\mu)^2$. (Of course, this can only work if the moment generating function exists!)

For example, we can apply this to simple random walk. Let X_i be i.i.d. taking values -1 and 1 with probability $1/2$ each. Let $S_n = X_1 + \dots + X_n$, so that S_n is the position of simple random walk on $\mathbb{Z}$ after n steps.

We know from the central limit theorem that, for large n, S_n is typically on the order of $\sqrt{n}$. So an event like $\{|S_n| > na\}$, for some $a > 0$, ought to have probability which gets small as $n \to \infty$.

First we bound the probability using Chebyshev. We have $\mathbb{E}\,S_n = 0$ and $\operatorname{Var} S_n = n$. So

$$\mathbb{P}(|S_n| > na) \le \frac{\operatorname{Var} S_n}{(na)^2}$$

$$= \frac{1}{na^2}.$$

This goes to 0 as desired but not very quickly!

Let us try instead using the moment generating function. We have

$$\begin{aligned}\mathbb{E}\, e^{tX_i} &= \frac{e^t + e^{-t}}{2} \\ &= \cosh t \\ &\leq \exp\left(\frac{t^2}{2}\right) \text{ for all } t.\end{aligned}$$

(The inequality $\cosh t \leq \exp(t^2/2)$ can be checked directly by expanding the exponential functions and comparing coefficients in the power series).

For $t > 0$, we can now write

$$\begin{aligned}\mathbb{P}(S_n > na) &= \mathbb{P}\left(\exp(tS_n) > \exp(tna)\right) \\ &\leq \frac{\mathbb{E}\,\exp(tS_n)}{\exp(tna)} \text{ (this is from Markov's inequality)} \\ &= \left(\frac{\mathbb{E}\,\exp(tX_i)}{\exp(ta)}\right)^n \\ &\leq \left(\exp\left(t^2/2 - ta\right)\right)^n.\end{aligned}$$

Note that this is true for *any* positive t, so we are free to choose whichever one we like. Naturally, we want to minimise the RHS. It is easy to check (just differentiate) that this is done by choosing $t = a$, which gives

$$\mathbb{P}(S_n > na) \leq \exp\left(-na^2/2\right).$$

By symmetry the bound on $\mathbb{P}(S_n < -na)$ is exactly the same. Combining the two we get

$$\mathbb{P}(|S_n| > na) \leq 2\exp\left(-na^2/2\right).$$

This decays much quicker than the bound from Chebyshev above!

3.4 Characteristic functions

The **characteristic function** is defined by replacing t by it in the definition of the moment generating function. The characteristic function of X is given by

$$\phi_X(t) := \mathbb{E}\,(e^{itX}),$$

for $t \in \mathbb{R}$. We can write

$$\phi_X(t) = \mathbb{E}\,(\cos(tX)) + i\mathbb{E}\,(\sin(tX)).$$

As a result we can see that the characteristic function is finite for every t, whatever the distribution of X. In fact, $|\phi_X(t)| \leq 1$ for all t.

This means that many of the results for the moment generating function which depended on exponential tails of the distribution have analogues for the characteristic function which hold for any distribution. Just as before we have $\phi_{X+Y}(t) = \phi_X(t)\phi_Y(t)$. The uniqueness and continuity theorems given for mgfs hold in a similar way for characteristic functions. The Taylor expansion of the characteristic function around the origin involves the moments of the distribution in a similar way (except now with an added factor of i^k for the kth term):

$$\phi_X(t) = 1 + it\mathbb{E}\,X + i^2t^2\frac{\mathbb{E}\,X^2}{2} + \cdots + i^kt^k\frac{\mathbb{E}\,X^k}{k!} + o(t^k) \tag{3.2}$$

as $t \to 0$, whenever $\mathbb{E}\,X^k$ is finite. Hence by following exactly the same strategy, we could give a proof of the central limit theorem using characteristic functions instead of mgfs. This would now prove the CLT without any additional assumption on the distribution (only finiteness of the variance is needed). Apart from working with complex power series instead of real power series, there are no additional complications when translating the proof from mgfs to cfs.

When the mgf is finite in an interval containing the origin in its interior, the theory of analytic continuation of complex functions allows us to obtain the characteristic function easily, by replacing t with it in the mgf.

Example 3.9. (a) The mgf of $N(0,1)$ is $\exp(t^2/2)$, and the cf is $\exp((it)^2/2) = \exp(-t^2/2)$.

(b) The mgf of Exp(1) is $1/(1-t)$, and the cf is $1/(1-it)$.

(c) Suppose X has Cauchy distribution with density $f(x) = \frac{1}{\pi(1+x^2)}$. The moment generating function is infinite for all $t \neq 0$ (in fact, even the mean does not exist as $\mathbb{E}\,|X| = \infty$ – exercise). The characteristic function is given by

$$\phi_X(t) = \mathbb{E}\,e^{itX} = \int_{-\infty}^{\infty} \frac{e^{itx}}{\pi(1+x^2)}dx$$

and this can be evaluated by contour integration to give $e^{-|t|}$.

Note that ϕ_X is not differentiable at 0; from (3.2), this corresponds to the fact that the mean does not exist.

In fact, consider $X_1, X_2, \dots X_n$ i.i.d. Cauchy, and $S_n = X_1 + \cdots + X_n$. Then

$$\phi_{S_n/n}(t) = \phi\left(\frac{t}{n}\right)^n = \left(e^{-|t|/n}\right)^n = e^{-|t|} = \phi_X(t).$$

So S_n/n and X_i have the same distribution! The law of large numbers and the CLT do not apply (since the mean does not exist).

3.4.1 Comparing moment generating functions and characteristic functions

Question M4(a)(ii) on Part A paper AO2 from 2011 asks:

> *State one purpose for which you should use the characteristic function rather than the moment generating function, and one purpose for which you would want to use the moment generating function rather than the characteristic function.*

The previous section gives an obvious answer to the first part of the question: when the distribution does not have exponentially decaying tails, the moment generating function is not useful but the characteristic function certainly is (to prove the CLT, for example). In the other direction, one could refer to the use of the mgf to give bounds on the tail of a distribution. In Section 3.3 we did this using Markov's inequality applied to the random variable e^{tX}; replacing this with e^{itX} would give nothing sensible, since that function is not real-valued, let alone monotonic.

4

Joint distribution of continuous random variables

4.1 Review of jointly continuous random variables

The **joint cumulative distribution function** of two random variables X and Y is defined by

$$F_{X,Y}(x,y) = \mathbb{P}(X \leq x, Y \leq y).$$

X and Y are said to be **jointly continuous** if their joint cdf can be written as an integral:

$$F_{X,Y}(x,y) = \int_{u=-\infty}^{x} \int_{v=-\infty}^{y} f(u,v) du\, dv.$$

Then f is said to be the joint pdf of X and Y, often written as $f_{X,Y}$. As in the case of a single random variable, we might more properly say "a joint pdf" rather than "the joint pdf" because we can, for example, change the value of f at finitely many points without changing the value of any integrals of f. But it is natural to put

$$f_{X,Y}(x,y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x,y)$$

whenever $F_{X,Y}$ is differentiable at (x,y).

For general (suitably nice[1]) sets $A \subset \mathbb{R}^2$ we have

$$\mathbb{P}\left((X,Y) \in A\right) = \int \int_A f_{X,Y}(x,y) dx\, dy. \tag{4.1}$$

If $f_{X,Y}$ satisfies (4.1) for all (nice) $A \subset \mathbb{R}^2$, then, clearly, $f_{X,Y}$ is a joint pdf of (X,Y). It suffices to check (4.1) for rectangles A or just for sets of the form $A = (-\infty, u] \times (-\infty, v]$, which yield the joint cdf.

We can recover the distribution of one of the random variables X or of Y by integrating over the other one. (In this context the distribution of one of the variables is called the **marginal distribution**).

$$f_X(x) = \int_{y=-\infty}^{\infty} f_{X,Y}(x,y) dy$$

[1] The suitable definition of "nice" is "Borel measurable". See Part A Integration.

$$f_Y(y) = \int_{x=-\infty}^{\infty} f_{X,Y}(x,y)dx$$

A function of X and Y is itself a random variable. Its expectation is given by

$$\mathbb{E}\, h(X,Y) = \int_{x=-\infty}^{\infty} \int_{y=-\infty}^{\infty} h(x,y) f_{X,Y}(x,y) dx\, dy.$$

Recall that X and Y are independent if $F_{X,Y}(x,y) = F_X(x)F_Y(y)$ for all x, y. Equivalently, the joint density can be written as a product:

$$f_{X,Y}(x,y) = f_X(x) f_Y(y).$$

All the above can be naturally generalised to describe the joint distribution of more than two random variables.

4.2 Change of variables

Often there is more than one natural coordinate system in which to view a model. We have the following **change of variables** result:

Theorem 4.1. *Suppose $T : (x,y) \mapsto (u,v)$ is a one-to-one mapping from some domain $D \subseteq \mathbb{R}^2$ to some range $R \subseteq \mathbb{R}^2$.*

*Define the **Jacobian** J as a function of (u,v) by*

$$J \;=\; \det \begin{pmatrix} \dfrac{\partial x}{\partial u} & \dfrac{\partial x}{\partial v} \\[2ex] \dfrac{\partial y}{\partial u} & \dfrac{\partial y}{\partial v} \end{pmatrix} \;=\; \frac{\partial x}{\partial u}\frac{\partial y}{\partial v} - \frac{\partial x}{\partial v}\frac{\partial y}{\partial u}.$$

Assume that the partial derivatives involved exists and are continuous.

If X, Y have joint probability density function $f_{X,Y}$, then the random variables U, V defined by $(U,V) = T(X,Y)$ are jointly continuous with joint probability density function $f_{U,V}$ given by

$$f_{U,V}(u,v) = \begin{cases} f_{X,Y}\big(x(u,v), y(u,v)\big)|J(u,v)| & \text{if } (u,v) \in R \\ 0 & \text{otherwise} \end{cases}.$$

Proof. The proof is simple using the familiar formula for change of variables in an integral. Suppose that $A \subseteq D$ and $T(A) = B$. Then, since T is one-to-one,

$$\begin{aligned} \mathbb{P}\left((U,V) \in B\right) &= \mathbb{P}\left((X,Y) \in A\right) \\ &= \int\int_A f_{X,Y}(x,y) dx\, dy \\ &= \int\int_B f_{X,Y}\left(x(u,v), y(u,v)\right) |J(u,v)| du\, dv. \end{aligned}$$

Hence the final integrand is the joint pdf of (U,V). □

The formula for change of variables in the integral appeared in various contexts last year. Recall the general idea: after a suitable translation, the transformation T looks locally like a linear transformation whose matrix is the matrix of partial derivatives above. We know that the factor by which the area of a set changes under a linear transformation is given by the determinant of the matrix of the transformation. So, locally, the Jacobian $J(u,v)$ gives the ratio between the area of a rectangle $(x, x+dx) \times (y, y+dy)$ and its image under T (which is a parallelogram). Since we want the probability to stay the same, and probability is area times density, we should rescale the density by the same amount $J(u,v)$.

Example 4.2. Let X, Y be i.i.d. exponentials with rate λ. Let $U = X/(X+Y)$, $V = X+Y$. What is the joint distribution of (U, V)?
Solution:

$$\begin{aligned} f_{X,Y}(x,y) &= \lambda e^{-\lambda x} \lambda e^{-\lambda y} \\ &= \lambda^2 e^{-\lambda(x+y)} \end{aligned}$$

for $(x,y) \in (0,\infty)^2$. The transformation $(u,v) = (x/(x+y), x+y)$ takes $(0,\infty)^2$ to $(0,1) \times (0,\infty)$. It is inverted by $x = uv, y = v(1-u)$. The Jacobian is given by

$$\begin{aligned} J(u,v) \;=\; \det \begin{pmatrix} \dfrac{\partial x}{\partial u} & \dfrac{\partial x}{\partial v} \\ \\ \dfrac{\partial y}{\partial u} & \dfrac{\partial y}{\partial v} \end{pmatrix} \;&=\; \det \begin{pmatrix} v & u \\ -v & 1-u \end{pmatrix} \\ &= v(1-u) + uv \\ &= v. \end{aligned}$$

So we have

$$\begin{aligned} f_{U,V}(u,v) &= f_{X,Y}\big(x(u,v), y(u,v)\big) |J(u,v)| \\ &= \lambda^2 e^{-\lambda\left(x(u,v)+y(u,v)\right)} |J(u,v)| \\ &= v\lambda^2 e^{-\lambda v} \end{aligned}$$

for $(u,v) \in (0,1) \times (0,\infty)$.

This factorises into a product of a function of u and a function of v (the function of u is trivial). So U and V are independent, with

$$\begin{aligned} f_U(u) &= 1, \;\; u \in (0,1) \\ f_V(v) &= \lambda^2 v e^{-\lambda v}, \;\; v \in (0,\infty) \end{aligned}$$

So $U \sim U[0,1]$ and $V \sim \text{Gamma}(2,\lambda)$, independently.

Example 4.3. Let X and Y be independent $\text{Exp}(\lambda)$ as in the previous example, and now let $V = X+Y$, $W = X-Y$. This transformation takes $(0,\infty)^2$ to the set $\{(v,w) : |w| < v\}$. The inverse transformation is

$$x = \frac{v+w}{2}, \;\; y = \frac{v-w}{2}$$

with Jacobian

$$J(v,w) = \det \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{pmatrix} = -\frac{1}{2}.$$

(Notice that any linear transformation always has constant Jacobian). So we have

$$\begin{aligned} f_{V,W}(v,w) &= \begin{cases} f_{X,Y}\left(\frac{v+w}{2}, \frac{v-w}{2}\right) |J(v,w)| & \text{for } |w| < v \\ 0 & \text{otherwise} \end{cases} \\ &= \begin{cases} \frac{1}{2}\lambda^2 e^{-\lambda v} & \text{for } |w| < v \\ 0 & \text{otherwise .} \end{cases} \end{aligned}$$

It looks like the pdf factorises into a product as in the previous example. But here this is not really the case! – because of the restriction to $|w| < v$. In fact, V and W could not be independent here, otherwise we could not have $\mathbb{P}(|W| < V) = 1$.

From the previous example we already know that $V \sim \text{Gamma}(2, \lambda)$. What is the marginal distribution of W?

$$\begin{aligned} f_W(w) &= \int_{v=|w|}^{\infty} \frac{1}{2}\lambda^2 e^{-\lambda v} dv \\ &= \left[-\frac{1}{2}\lambda e^{-\lambda v} \right]_{|w|}^{\infty} \\ &= \frac{1}{2}\lambda e^{-\lambda |w|}. \end{aligned}$$

We see that the distribution of W is symmetric around 0, and by adding the density at w and $-w$, the distribution of $|W|$ has pdf $\lambda e^{-\lambda|w|}$ and so again has $\text{Exp}(\lambda)$ distribution.

Example 4.4 (General formula for the sum of continuous random variables)**.** If X and Y are jointly continuous with density function $f_{X,Y}$, what is the distribution of $X+Y$? We can change variables to $U = X+Y, V = X$. This transformation has Jacobian 1 (check!), and we obtain $f_{U,V}(u,v) = f_{X,Y}(v, u-v)$.

To obtain the marginal distribution of $X+Y$, which is U, we integrate over v:

$$f_{X+Y}(u) = \int_{-\infty}^{\infty} f_{X,Y}(v, u-v) dv.$$

An important case is when X and Y are independent. Then we obtain the **convolution** formula:

$$f_{X+Y}(u) = \int_{-\infty}^{\infty} f_X(v) f_Y(u-v) dv.$$

4.2.1 Multivariate distributions

Everything above can be generalised to the case of the joint distribution of $n > 2$ random variables. The Jacobian is now the determinant of an $n \times n$ matrix.

4.3 Multivariate normal distribution

Let $Z_1, Z_2, \ldots, Z_n$ be i.i.d. standard normal random variables. Their joint density function can be written as

$$\begin{aligned} f_{\mathbf{Z}}(\mathbf{z}) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z_i^2}{2}\right) \\ &= \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\mathbf{z}^T\mathbf{z}\right). \end{aligned}$$

Define $W_1, \ldots, W_n$ by

$$\begin{pmatrix} W_1 \\ W_2 \\ \vdots \\ W_n \end{pmatrix} = A \begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_n \end{pmatrix} + \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix}$$

where A is some $n \times n$ matrix.

Assume A is invertible. Then by change of variables (the Jacobian is constant) we get

$$f_{\mathbf{W}}(\mathbf{w}) = \frac{1}{(2\pi)^{n/2}|\det A|} \exp\left(-\frac{1}{2}(\mathbf{w}-\boldsymbol{\mu})^T \left(AA^T\right)^{-1}(\mathbf{w}-\boldsymbol{\mu})\right).$$

The matrix $\Sigma := AA^T$ is the *covariance matrix* in the sense that $\mathrm{Cov}(W_i, W_j) = (AA^T)_{ij}$ (check, e.g. for $n = 2$ if you want an easy case). $W_1, \ldots, W_n$ are said to have the **multivariate normal distribution** with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ.

For the case $n = 2$, one can manipulate to obtain (with $X = W_1$, $Y = W_2$)

$$\begin{aligned} &f_{X,Y}(x, y) \\ &= \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2\left(1-\rho^2\right)}\left[\frac{(x-\mu_X)^2}{\sigma_X^2} - \frac{2\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}\right]\right) \end{aligned}$$

where σ_X^2 and σ_Y^2 are the variances of X and Y, μ_X and μ_Y are the means, and ρ is the correlation coefficient between X and Y which is defined by

$$\rho = \frac{\mathrm{Cov}(X, Y)}{\sigma_X\sigma_Y}$$

and lies in $(-1, 1)$.

Note that

(1) The density depends only on μ_X, μ_Y, σ_X, σ_Y and ρ.

(2) X and Y are independent $\Leftrightarrow \rho = 0$. ($\Rightarrow$ is true for any joint distribution; $\Leftarrow$ is a special property of joint normal.)

A special case is the **standard bivariate normal** where $\sigma_X = \sigma_Y = 1$ and $\mu_X = \mu_Y = 0$. Then

$$f_{X,Y}(x, y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)}\right).$$

In this case $\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$.

4.4 Conditional densities

The basic definition of conditional probability: for two events A and B with $\mathbb{P}(A) > 0$, the conditional probability of B given A is

$$\mathbb{P}(B|A) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)}. \tag{4.2}$$

Applying this to the distribution of a random variable, we have for example

$$\mathbb{P}(X \leq x|A) = \frac{\mathbb{P}(\{X \leq x\} \cap A)}{\mathbb{P}(A)}.$$

The left-hand side is a cumulative distribution function. It gives the **conditional distribution** of X, given A. We might denote it by $F_{X|A}(x)$.

In the case where X is discrete, we can write the conditional probability mass function:

$$p_{X|A}(x) = \mathbb{P}(X = x|A).$$

If X is continuous, we can differentiate the conditional distribution function to get a conditional density function $f_{X|A}(x)$, and then for a set C,

$$\mathbb{P}(X \in C|A) = \int_{x \in C} f_{X|A}(x) dx.$$

The conditional expectation of X given A is the expectation of the conditional distribution, which is given by

$$\mathbb{E}\left(X|A\right) = \sum_x x p_{X|A}(x)$$

in the discrete case, and by

$$\mathbb{E}\left(X|A\right) = \int x f_{X|A}(x) dx$$

in the continuous case.

Example 4.5. Suppose X and Y are independent random variables which both have uniform distribution on $[0, 1]$. Find the conditional distribution and conditional expectation of Y given $X + Y > 1$.

Solution:

$$\mathbb{P}(Y < y|X + Y > 1) = \frac{\mathbb{P}(Y < y, X + Y > 1)}{\mathbb{P}(X + Y > 1)}.$$

Since X, Y are uniform on the square $[0, 1]^2$, the probability of a set is equal to its area.

The set $\{x + y > 1\}$ has area $1/2$, while for fixed y, the set $\{(x, v) : v < y, x + v > 1\}$ has area $y^2/2$.

So the distribution function of Y given $X + Y > 1$ is $F(y) = (y^2/2)/(1/2) = y^2$, and the conditional density is $2y$ on $[0, 1]$, and 0 elsewhere.

The conditional expectation $\mathbb{E}\left(Y|X + Y > 1\right)$ is $\int_0^1 y \times 2y \, dy = 2/3$.

A common way in which conditional distributions arise is when we have two random variables X and Y with some joint distribution; we observe the value of X and want to know what this tells us about the value of Y. That is, what is the conditional distribution of Y given $X = x$?

When X is a discrete random variable, everything works fine; since $\mathbb{P}(X = x)$ will be positive, we can use the approach above.

However, if X is continuous, then $\mathbb{P}(X = x)$ will be 0 for every x. Now we have a problem, since if the event A in (4.2) has probability 0, then the definition makes no sense.

To resolve this problem, rather than conditioning directly on $\{X = x\}$, we look at the distribution of Y conditioned on $\{x \leq X \leq x + \epsilon\}$. If the joint distribution is well-behaved (as it will be in all the cases that we wish to consider), we can obtain a limit as $\epsilon \downarrow 0$, which we define as the distribution of Y given $X = x$.

As $\epsilon \to 0$, we have

$$\begin{aligned}
\mathbb{P}\big(Y \leq y \big| x \leq X \leq x+\epsilon\big) &= \frac{\displaystyle\int_{v=-\infty}^{y} \int_{u=x}^{x+\epsilon} f_{X,Y}(u,v)du\,dv}{\displaystyle\int_{u=x}^{x+\epsilon} f_X(u)du} \\
&\sim \frac{\epsilon \displaystyle\int_{v=-\infty}^{y} f_{X,Y}(x,v)dv}{\epsilon f_X(x)} \\
&= \int_{v=-\infty}^{y} \frac{f_{X,Y}(x,v)}{f_X(x)} dv. \qquad (4.3)
\end{aligned}$$

So we define $F_{Y|X=x}(y)$, the **conditional distribution function of** Y **given** $X = x$, as the right-hand side of (4.3).

Differentiating with respect to y, we obtain the **conditional density function of** Y **given** $X = x$, written as $f_{Y|X=x}(y)$:

$$\boxed{f_{Y|X=x}(y) = \frac{f_{X,Y}(x,y)}{f_X(x)}}.$$

These definitions make sense whenever $f_X(x) > 0$. In that case, note that $f_{Y|X=x}$ is indeed a density function, because we have defined $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy$. (Notice that the denominator $f_X(x)$ does not involve y at all; it is just a normalising constant).

The idea is that the following two procedures are equivalent:

(1) generate (X, Y) according to the joint density function $f_{X,Y}$;

(2) first generate X according to the density function f_X, and then having observed $X = x$, generate Y according to the density function $f_{Y|X=x}$.

Example 4.6 (Simple example)**.** Let (X, Y) be uniform on the triangle $\{0 < y < x < 1\}$. Then

$$f_{X,Y}(x,y) = \begin{cases} 2 & 0 < y < x < 1 \\ 0 & \text{otherwise} \end{cases}.$$

For the conditional density of Y given $X = x$,

$$\begin{aligned} f_{Y|X=x}(y) &= \frac{f_{X,Y}(x,y)}{f_X(x)} \\ &= \begin{cases} 2/f_X(x) & 0 < y < x \\ 0 & \text{otherwise} \end{cases}, \end{aligned}$$

provided $x \in (0,1)$. We do not need to calculate $f_X(x)$, since it is just a normalising constant. Since the conditional density function of Y is constant in y, we see that Y is uniform on $[0,x]$, with distribution function given by

$$F_{Y|X=x}(y) = \begin{cases} 0 & y < 0 \\ \frac{y}{x} & 0 \le y \le x \\ 1 & y > x \end{cases}.$$

The conditional mean of Y given $X = x$ is $x/2$.

Example 4.7 (Bivariate normal)**.** Let X and Y be jointly normal with means μ_1 and μ_2 respectively, variances σ_1^2 and σ_2^2 respectively, and correlation coefficient ρ. What is the conditional distribution of Y given $X = x$?

Rather than working directly from the joint density function, we can proceed by writing Y as the sum of two terms, one which is a function of X and one which is independent of X.

First let us write X and Y as functions of independent standard normals Z_1 and Z_2. If we put

$$\begin{aligned} X &= \sigma_1 Z_1 + \mu_1 \\ Y &= \rho\sigma_2 Z_1 + \sqrt{1-\rho^2}\sigma_2 Z_2 + \mu_2 \end{aligned}$$

then indeed X and Y have the desired means, variances and covariance (check!).

Then we can write

$$Y = \rho\frac{\sigma_2}{\sigma_1}(X - \mu_1) + \sqrt{1-\rho^2}\sigma_2 Z_2 + \mu_2.$$

The first term is a function of X and the second term, involving only Z_2, is independent of X.

So conditional on $X = x$, the distribution of Y is the distribution of

$$\rho\frac{\sigma_2}{\sigma_1}(x - \mu_1) + \sqrt{1-\rho^2}\sigma_2 Z_2 + \mu_2,$$

which is normal with mean $\rho\frac{\sigma_2}{\sigma_1}(x - \mu_1) + \mu_2$ and variance $(1-\rho^2)\sigma_2^2$.

Note the way the variance of this conditional distribution depends on ρ. We say that ρ^2 is the "amount of the variance of Y explained by X". Consider the extreme cases. If $\rho = \pm 1$, then the conditional variance is 0. That is, Y is a function of X and once we observe X, there is no longer any uncertainty about the value of Y. If $\rho = 0$, the conditional variance and the unconditional variance are the same; observing X tells us nothing about Y.

4.5 Cautionary tale

The definition above of conditional distribution given the value of a continuous random variable makes sense in context, but keep in mind that conditioning directly on events on probability zero is not valid, and as a result the objects involved are not robust to seemingly innocent manipulation! Consider the following example:

Example 4.8 (Borel's paradox)**.** Consider the uniform distribution on the half-disc $C = \{(x, y) : y \geq 0, x^2 + y^2 \leq 1\}$. The joint density of X and Y is given by

$$f(x, y) = \begin{cases} \frac{2}{\pi} & (x, y) \in C \\ 0 & \text{otherwise} \end{cases}.$$

What is the conditional distribution of Y given $X = 0$? Its density is given by

$$f_{Y|X=0}(y) = \frac{2/\pi}{f_X(0)}$$

for $y \in [0, 1]$, and 0 elsewhere. So the distribution is uniform on $[0, 1]$ (we do not need to calculate $f_X(0)$ to see this, since it is only a normalising constant).

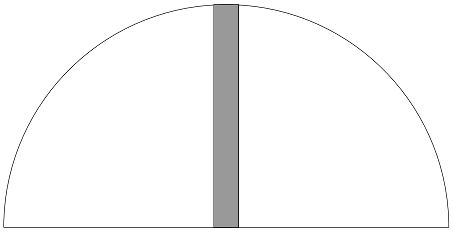
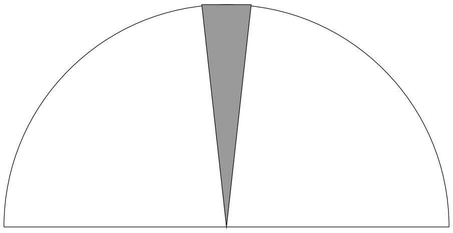
We could change variables and represent the same distribution in polar coordinates. Then R and Θ are independent; R has density $2r$ on [0,1] and Θ is uniform on $[0, \pi)$. (See first question on problem sheet 2 for the transformation to polar coordinates. But in this case where the density of X, Y is uniform on a set, one can also easily derive the joint distribution of R and Θ directly by considering areas of subsets of the set C).

Note that the events $\{X = 0\}$ and $\{\Theta = \pi/2\}$ are the same.

What is the conditional distribution of R given $\Theta = \pi/2$? Since R and Θ are independent, it still has density $2r$ on $[0, 1]$. This is *not* uniform on $[0, 1]$.

But when $X = 0$, i.e. when $\Theta = \pi/2$, R and Y are the same thing. So the distribution of R given $\Theta = \pi/2$ ought to be the same as the distribution of Y given $X = 0$, should it not?

What is happening is that, although the events $\{X = 0\}$ and $\{\Theta = \pi/2\}$ are the same, it is *not* the case that the events $\{|X| < \epsilon\}$ and $\{|\Theta - \pi/2| < \epsilon\}$ are the same. When we condition X to be within ϵ of 0, we restrict to a set which is approximately a rectangle (the left-hand picture below). However, when we condition Θ to be near $\pi/2$, we restrict to a thin sector of the circle, which is approximately a triangle (the right-hand picture below). In the second case, we bias the point chosen to lie higher up. As $\epsilon \to 0$, this bias persists; the two limits are not the same!

What this "paradox" illustrates is that conditioning for continuous random variables involves a limit, and that it can be important exactly how the limit is taken. The procedure whereby we generate X from f_X and then Y from $f_{Y|X}$ makes sense in terms of a particular set of variables; but the conditional densities involved are not robust to a change of variables.

5

Markov chains: Introduction

Let $X_n, n = 0, 1, 2, \dots$ be a "random process", taking values in some set I called the *state space*. That is, $X_0, X_1, X_2, \dots$ are random variables with $X_n \in I$ for all n.

Often X_n represents some quantity evolving in time. So far we have been working with random variables taking values which are real numbers of some kind, but there is no problem in considering a more general state space. For example, we might consider processes of the following kind:

- $I = \mathbb{Z}^2$, X_n=position at nth step of a "random walk" on the two-dimensional lattice.
- $I = \{A, B, C, \dots, a, b, c, \dots, ., ?, !, \dots\}$, X_n=nth character in a text or in an email.
- $I = \{C, G, A, T\}$ (representing cytosine, guanine, adenine, thymine, the four bases of DNA), X_n=base appearing in nth position in a DNA sequence.

We will assume that the state space I is finite or countably infinite (i.e. discrete). A **(probability) distribution** on I is a collection $\lambda = (\lambda_i, i \in I)$ with $\lambda_i \geq 0$ for all i, and $\sum \lambda_i = 1$. This is really just the same idea as the probability mass function of a discrete random variable. We will often think of λ as a **row vector**. We will say that a random variable Y taking values in I has distribution λ if $\mathbb{P}(Y = i) = \lambda_i$ for all i.

5.1 Markov chains

Let $X = (X_0, X_1, X_2, \dots)$ be a sequence of random variables taking values in I. The process X is called a **Markov chain** if for any $n \geq 0$ and any $i_0, i_1, \dots, i_{n+1} \in I$,

$$\mathbb{P}\left(X_{n+1} = i_{n+1} \middle| X_n = i_n, \dots, X_0 = i_0\right) = \mathbb{P}\left(X_{n+1} = i_{n+1} \middle| X_n = i_n\right). \tag{5.1}$$

(To be precise, we should restrict (5.1) to cases where these conditional probabilities are well-defined, i.e. where the event $\{X_n = i_n, \dots, X_0 = i_0\}$ has positive probability.)

The Markov chain is called **(time) homogeneous** if in addition $\mathbb{P}\left(X_{n+1} = j \middle| X_n = i\right)$ depends only on i and j, not on n. In that case we write

$$p_{ij} = \mathbb{P}\left(X_{n+1} = j \middle| X_n = i\right)$$

(or we will often write $p_{i,j}$ rather than p_{ij}, according to convenience). The quantities p_{ij} are known as the **transition probabilities** of the chain.

We will work almost always with homogeneous chains. To describe such a chain, it is enough to specify two things:

- the initial distribution λ of X_0. For each $i \in I$, $\lambda_i = \mathbb{P}(X_0 = i)$.
- the **transition matrix** $P = (p_{ij})_{i,j\in I}$.

P is a square (maybe infinite) matrix, whose rows and columns are indexed by I. P is a "stochastic matrix" which means that all its entries are non-negative and every row sums to 1. Equivalently, every row of P is a probability distribution. The ith row of P is the distribution of X_{n+1} given $X_n = i$.

Theorem 5.1. *For $i_0, i_1, \dots, i_n \in I$,*

$$\boxed{\mathbb{P}\left(X_0 = i_0, X_1 = i_1, \dots, X_n = i_n\right) = \lambda_{i_0} p_{i_0 i_1} p_{i_1 i_2} \dots p_{i_{n-1} i_n}.}$$

Proof. By the definition of conditional probabilities and cancellations,

$$\begin{aligned}
&\mathbb{P}\left(X_0 = i_0, X_1 = i_1, \dots, X_n = i_n\right)\\
&\quad = \mathbb{P}\left(X_0 = i_0\right) \mathbb{P}\left(X_1 = i_1 \middle| X_0 = i_0\right) \mathbb{P}\left(X_2 = i_2 \middle| X_1 = i_1, X_0 = i_0\right) \times \dots\\
&\qquad\qquad\qquad\qquad \dots \times \mathbb{P}\left(X_n = i_n \middle| X_{n-1} = i_{n-1}, \dots, X_0 = i_0\right)\\
&\quad = \mathbb{P}\left(X_0 = i_0\right) \mathbb{P}\left(X_1 = i_1 \middle| X_0 = i_0\right) \mathbb{P}\left(X_2 = i_2 \middle| X_1 = i_1\right) \dots \mathbb{P}\left(X_n = i_n \middle| X_{n-1} = i_{n-1}\right)\\
&\quad = \lambda_{i_0} p_{i_0 i_1} p_{i_1 i_2} \dots p_{i_{n-1} i_n},
\end{aligned}$$

where we used the definition of a Markov chain to get the penultimate line. □

If X is a Markov chain with initial distribution λ and transition matrix P, we will sometimes write "$X \sim \text{Markov}(\lambda, P)$".

Markov chains are "memoryless". If we know the current state, any information about previous states is irrelevant to the future evolution of the chain. We can say that "the future is independent of the past, given the present". This is known as the **Markov property**:

$$\begin{aligned}
&\mathbb{P}\left(X_{n+1} \in A_{n+1}, \dots, X_{n+m} \in A_{n+m} \middle| X_0 \in A_0, \dots, X_{n-1} \in A_{n-1}, X_n = i\right)\\
&= \mathbb{P}\left(X_{n+1} \in A_{n+1}, \dots, X_{n+m} \in A_{n+m} \middle| X_n = i\right)\\
&= \mathbb{P}\left(X_1 \in A_{n+1}, \dots, X_m \in A_{n+m} \middle| X_0 = i\right).
\end{aligned}$$

for all $A_0, \dots, A_{m+n} \subseteq I$ with $\mathbb{P}(X_0 \in A_0, \dots, X_{n-1} \in A_{n-1}, X_n = i) > 0$, or equivalently (by the definition of conditional probabilities and cancellations)

$$\begin{aligned}
&\mathbb{P}\left(X_0 \in A_0, \dots, X_{n-1} \in A_{n-1},\, X_{n+1} \in A_{n+1}, \dots, X_{n+m} \in A_{n+m} \middle| X_n = i\right)\\
&= \mathbb{P}\left(X_0 \in A_0, \dots, X_{n-1} \in A_{n-1} \middle| X_n = i\right) \mathbb{P}\left(X_{n+1} \in A_{n+1}, \dots, X_{n+m} \in A_{n+m} \middle| X_n = i\right)\\
&= \mathbb{P}\left(X_0 \in A_0, \dots, X_{n-1} \in A_{n-1} \middle| X_n = i\right) \mathbb{P}\left(X_1 \in A_{n+1}, \dots, X_m \in A_{n+m} \middle| X_0 = i\right).
\end{aligned}$$

Notation: it will be convenient to write $\mathbb{P}_i$ for the distribution conditioned on $X_0 = i$. For example $\mathbb{P}_i(X_1 = j) = p_{ij}$. Similarly $\mathbb{E}_i$ for expectation conditioned on $X_0 = i$.

5.2 n-step transition probabilities

Write $p_{ij}^{(n)} = \mathbb{P}\left(X_{k+n} \middle| X_k = i\right)$. This is an **$n$-step transition probability** of the Markov chain.

Theorem 5.2. *(Chapman-Kolmogorov equations)*

(i) $p_{ik}^{(n+m)} = \sum_{j \in I} p_{ij}^{(n)} p_{jk}^{(m)}$.

(ii) $p_{ij}^{(n)} = (P^n)_{i,j}$.

Here P^n is the nth power of the transition matrix. As ever, matrix multiplication is given by $(AB)_{i,j} = \sum_k (A)_{i,k}(B)_{k,j}$, whether the matrices are finite or infinite.

Proof. (i) We condition on X_n, i.e. we consider the partition $\{X_n = j\}$, $j \in I$, and use the Law of Total Probability:

$$\begin{aligned}\mathbb{P}\left(X_{n+m} = k \middle| X_0 = i\right) &= \sum_j \mathbb{P}\left(X_n = j \middle| X_0 = i\right) \mathbb{P}\left(X_{n+m} = k \middle| X_n = j, X_0 = i\right) \\ &= \sum_j \mathbb{P}\left(X_n = j \middle| X_0 = i\right) \mathbb{P}\left(X_{n+m} = k \middle| X_n = j\right)\end{aligned}$$

(using the Markov property)

$$= \sum_j p_{ij}^{(n)} p_{jk}^{(m)}.$$

(ii) For $n = 1$, this holds by definition of P. Inductively, if this holds for any $n \geq 1$,

$$p_{ik}^{(n+1)} = \sum_j p_{ij}^{(n)} p_{jk}^{(1)} = \sum_j (P^n)_{i,j} (P)_{j,k} = (P^n P)_{i,k} = \left(P^{n+1}\right)_{i,k}.$$

□

Theorem 5.3. *Let λ be the initial distribution (i.e. the distribution of X_0). Then the distribution of X_1 is λP, and more generally the distribution of X_n is λP^n.*

Here we are thinking of λ as a row vector, so that λP^n is also a row vector; $(\lambda A)_i = \sum_k \lambda_k A_{ki}$ as usual, whether the dimensions are finite or infinite.

Proof. Just condition on the initial state, i.e. apply the Law of Total Probability for the partition $\{X_0 = i\}$, $i \in I$:

$$\begin{aligned}\mathbb{P}(X_1 = j) &= \sum_i \mathbb{P}(X_0 = i) \mathbb{P}\left(X_1 = j \middle| X_0 = i\right) \\ &= \sum_i \lambda_i p_{ij} \\ &= (\lambda P)_j,\end{aligned}$$

and similarly for X_n with p_{ij} and P replaced by $p_{ij}^{(n)}$ and P^n. □

Using this result and the Markov property it is easy to get the following property: if $(X_0, X_1, X_2, \dots)$ is a Markov chain with initial distribution λ and transition matrix P, then $(X_0, X_k, X_{2k}, \dots)$ is a Markov chain with initial distribution λ and transition matrix P^k.

Example 5.4 (General two-state Markov chain)**.** Let $I = \{1, 2\}$ and

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

What is $p_{11}^{(n)}$? Two approaches:

(1) P has eigenvalues 1 and $1 - \alpha - \beta$ (check! Every Markov transition matrix has 1 as an eigenvalue – why?). So we can diagonalise:

$$P = U^{-1} \begin{pmatrix} 1 & 0 \\ 0 & 1-\alpha-\beta \end{pmatrix} U,$$
$$P^n = U^{-1} \begin{pmatrix} 1 & 0 \\ 0 & (1-\alpha-\beta)^n \end{pmatrix} U.$$

We get $(P^n)_{11} = A + B(1 - \alpha - \beta)^n$ for some constants A and B.

Since we know $p_{11}^{(0)} = 1$ and we have $p_{11}^{(1)} = 1 - \alpha$, we can solve for A and B to get

$$p_{11}^{(n)} = \frac{\beta}{\alpha+\beta} + \frac{\alpha}{\alpha+\beta}(1-\alpha-\beta)^n. \tag{5.2}$$

(2) Alternatively, we can condition on the state at step $n - 1$:

$$\begin{aligned} p_{11}^{(n)} &= p_{11}^{(n-1)}(1-\alpha) + p_{12}^{(n-1)}\beta \\ &= p_{11}^{(n-1)}(1-\alpha) + \left(1 - p_{11}^{(n-1)}\right)\beta \\ &= (1-\alpha-\beta)p_{11}^{(n-1)} + \beta. \end{aligned}$$

This gives a linear recurrence relation for $p_{11}^{(n)}$, which we can solve using standard methods to give (5.2) again.

5.3 A few examples

Random walk on a cycle

$I = \{0, 1, 2, \dots, M - 1\}$. At each step the walk increases by 1 (mod M) with probability p and decreases by 1 (mod M) with probability $1 - p$. That is,

$$p_{ij} = \begin{cases} p & \text{if } j \equiv i + 1 \mod M, \\ 1-p & \text{if } j \equiv i - 1 \mod M, \\ 0 & \text{otherwise,} \end{cases}$$

or

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

Simple symmetric random walk on $\mathbb{Z}^d$

$I = \mathbb{Z}^d$. At each step the walk moves from its current site to one of its $2d$ neighbours chosen uniformly at random.

$$p_{ij} = \begin{cases} \frac{1}{2d} & \text{if } |i-j| = 1, \\ 0, & \text{otherwise} \end{cases}$$

where $|i-j| = |i_1 - j_1| + \cdots + |i_d - j_d|$ for states $i = (i_1, \ldots, i_d)$ and $j = (j_1, \ldots, j_d)$.

Card-shuffling

Let I be the set of orderings of 52 cards. We can regard I as the permutation group S_{52}. There are many interesting Markov chains on permutation groups. We can think of shuffling a pack of cards. A simple and not very practical example of a shuffle: at each step, choose a and b independently and uniformly in $\{1, 2, \ldots, 52\}$ and exchange the cards in positions a and b. This gives

$$p_{\alpha\beta} = \begin{cases} \dfrac{2}{52^2} & \text{if } \alpha = \beta\tau \text{ for some transposition } \tau, \\ \dfrac{1}{52} & \text{if } \alpha = \beta, \\ 0 & \text{otherwise.} \end{cases}$$

5.4 Exploring the Markov property

Let us look at a few examples of simple processes where the Markov property holds or fails. We can do this in the context of a simple random walk on $\mathbb{Z}$.

Let X_i be i.i.d. with $\mathbb{P}(X_i = 1) = p$ and $\mathbb{P}(X_i = -1) = 1 - p$.

Let $S_0 = 0$ and $S_n = \sum_{i=1}^n X_i$.

Then:

(1) X_n is a Markov chain. In fact, X_n are i.i.d., which is a stronger property. Given any history, the next state is equal to 1 with probability p and -1 with probability $1-p$. The matrix of the chain X_n (with rows and columns indexed by $\{-1, 1\}$) is $P = \begin{pmatrix} 1-p & p \\ 1-p & p \end{pmatrix}$.

(2) The random walk S_n is also a Markov chain. Its transition probabilities are $p_{i,i+1} = p$ and $p_{i,i-1} = p$ for all $i \in \mathbb{Z}$.

(3) Consider the process $M_n = \max_{0 \leq k \leq n} S_k$. Try drawing some possible paths of the process S_n, and the corresponding paths of the "maximum process" M_n. Is this maximum process a Markov chain?

We can consider two different ways of arriving at the same state. Suppose we observe $(M_0, \ldots, M_4) = (0, 0, 0, 1, 2)$. This implies $S_4 = 2$ (the maximum process has just increased, so now the walk must be at its current maximum.) In that case, if the random walk moves up at the next step, then the maximum will also increase. So

$$\mathbb{P}(M_5 = 3 | (M_0, \ldots, M_4) = (0, 0, 0, 1, 2)) = p.$$

Suppose instead that $(M_0, \ldots, M_4) = (0, 1, 2, 2, 2)$. In that case, both $S_4 = 2$ and $S_4 = 0$ are possible (check! – find the corresponding paths). As a consequence, sometimes the maximum will stay the same at the next step, even when the random walk moves up. So we have

$$\mathbb{P}(M_5 = 3 | (M_0, \ldots, M_4) = (0, 1, 2, 2, 2)) < p.$$

We see that the *path* to $M_4 = 2$ affects the conditional probability of the next step of the process. So M_n is *not* a Markov chain.

The next result gives a criterion for the Markov property to hold.

Proposition 5.5. *Suppose that $(Y_n, n \geq 0)$ is a random process, and for some function f we can write, for each n,*

$$Y_{n+1} = f(Y_n, X_{n+1}),$$

where X_{n+1} is independent of $Y_0, Y_1, \ldots, Y_n$. Then (Y_n) is a Markov chain.

Proof. The idea is that to update the chain, we use only the current state and some "new" randomness. We have

$$\begin{aligned}
&\mathbb{P}\big(Y_{n+1} = i_{n+1} \big| Y_n = i_n, \ldots, Y_0 = i_0\big) \\
&\qquad = \mathbb{P}\left(f(i_n, X_{n+1}) = i_{n+1} \big| Y_n = i_n, \ldots, Y_0 = i_0\right) \\
&\qquad = \mathbb{P}\left(f(i_n, X_{n+1}) = i_{n+1}\right) \quad \text{(by independence of } X_{n+1} \text{ from } Y_0, \ldots, Y_n) \\
&\qquad = \mathbb{P}\left(f(i_n, X_{n+1}) = i_{n+1} \big| Y_n = i_n\right) \quad \text{(by independence of } X_{n+1} \text{ from } Y_n) \\
&\qquad = \mathbb{P}\left(Y_{n+1} = i_{n+1} \big| Y_n = i_n\right).
\end{aligned}$$

□

For example, for the simple random walk above, we can put $S_{n+1} = f(S_n, X_{n+1})$, where $f(s, x) = s + x$. For the card-shuffling example in the previous section, if $Y_n \in S_{52}$ is the permutation after step n, we can put $Y_{n+1} = f(Y_n, X_{n+1})$ where for a permutation β and a transposition τ, $f(\beta, \tau) = \beta\tau$, and where (X_n) is an i.i.d. sequence in which each member is uniform in the set of transpositions.

5.5 Class structure

Let $i, j \in I$. We say that "i leads to j" and write "$i \to j$" if $\mathbb{P}_i(X_n = j) > 0$ for some $n \geq 0$, i.e. $p_{ij}^{(n)} > 0$ for some $n \geq 0$.

If $i \to j$ and $j \to i$ then we say "i communicates with j" and write $i \leftrightarrow j$.

Then $\leftrightarrow$ is an equivalence relation (check!). It partitions the state space I into **communicating classes**.

A class C is called **closed** if $p_{ij} = 0$ whenever $i \in C, j \notin C$, or equivalently $i \not\to j$ for any $i \in C, j \notin C$. Once the chain enters a closed class, it can never escape from it. If $\{i\}$ is a closed class then $p_{ii} = 1$, and i is called an **absorbing state**. If C is not closed it is called **open**.

A chain (or more precisely a transition matrix) for which I consists of a single communicating class is called **irreducible**. Equivalently, $i \to j$ for all $i, j \in I$.

Example 5.6. Let $I = \{1, 2, 3, 4, 5, 6, 7\}$. The communicating classes for the transition matrix

$$P = \begin{pmatrix} 0 & \frac{1}{2} & 0 & 0 & 0 & 0 & \frac{1}{2} \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & \frac{1}{4} & \frac{1}{4} & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

are $\{1, 2, 3\}$, $\{4\}$, $\{5, 6\}$ and $\{7\}$. The closed classes are $\{5, 6\}$ and $\{7\}$ (so 7 is an absorbing state). Draw a diagram to visualise the chain!

5.6 Periodicity

Consider the transition matrix

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

Again, draw a diagram to visualise the chain. Note that $p_{ii}^{(n)} = 0$ whenever n is odd.

For a general chain and a state $i \in I$, the **period** of the state i is defined to be the greatest common divisor of the set $\left\{n \geq 1 \colon p_{ii}^{(n)} > 0\right\}$. (If $p_{ii}^{(n)} = 0$ for all $n > 0$, then the period is not defined). All the states in the chain above have period 2.

In Example 5.6, states 1, 2 and 3 have period 3, the period of state 4 is undefined, 5 and 6 have period 2 and the absorbing state 7 has period 1.

i is called **aperiodic** if this g.c.d. is 1 (and otherwise **periodic**). Equivalently (check!), i is aperiodic if $p_{ii}^{(n)} > 0$ for all sufficiently large n.

Fact. *All states in a communicating class have the same period.*

Proof. Suppose $i \leftrightarrow j$ and $d|n$ whenever $p_{ii}^{(n)} > 0$.

Since i and j communicate, we can find a and b with $p_{ij}^{(a)} > 0$ and $p_{ji}^{(b)} > 0$. Then also $p_{ii}^{(a+b)} > 0$.

Suppose $p_{jj}^{(m)} > 0$. Then also $p_{ii}^{(a+m+b)} > 0$.

Then $d|a+b$ and $d|a+m+b$, so also $d|m$.

This demonstrates that the sets $\left\{n \geq 1 \colon p_{ii}^{(n)} > 0\right\}$ and $\left\{m \geq 1 \colon p_{jj}^{(m)} > 0\right\}$ have the same divisors, and hence the same greatest common divisor. □

In particular, if a chain is irreducible, then all states have the same period. If this period is 1, we say that the chain is aperiodic (otherwise we say the chain is periodic).

Remark. Notice that both irreducibility and periodicity are "structural properties" in the following sense: they depend only on which transition probabilities p_{ij} are positive and which are zero, not on the particular values taken by those which are positive.

Example. Look back at the three examples in Section 5.3 and consider which are irreducible and which are periodic.

The random walk on the cycle is irreducible (since every site is accessible from every other). It has period 2 if M is even, and is aperiodic if M is odd.

The random walk on $\mathbb{Z}^d$ is irreducible and has period 2 for any d.

The card-shuffling chain is irreducible (because the set of transpositions is a set of generators for the group S_{52}). It is aperiodic, since there is a positive transition probability from any state to itself.

Remark. Later we will show results about convergence to equilibrium for Markov chains. The idea will be that after a long time, a Markov chain should more or less "forget where it started". There are essentially two reasons why this might not happen: (a) periodicity; for example if a chain has period 2, then it alternates between, say, "odd" and "even" states; even an arbitrarily long time, the chain will still remember whether it started at an "odd" or "even" state. (b) lack of irreducibility. A chain with more than one closed class can never move from one to the other, and so again will retain some memory of where it started, for ever. When we prove results about convergence to equilibrium, it will be under the condition that the chain is irreducible and aperiodic.

5.7 Hitting probabilities

Let A be a subset of the state space I. Define $h_i^A = \mathbb{P}_i\left(X_n \in A \text{ for some } n \geq 0\right)$, the **hitting probability** of A starting from state i.

If A is a closed class, we might call h_i^A the **absorption probability**.

Example. Let $I = \{1, 2, 3, 4\}$ and

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

Starting from 2, what is the probability of absorption at 4?

Write $h_i = \mathbb{P}_i(X_n = 4 \text{ for some } n \geq 0)$. Then $h_4 = 1$, and $h_1 = 0$ since 1 is itself absorbing. Also by conditioning on the first jump and applying the Markov property, we have

$$\begin{aligned} h_2 =& \mathbb{P}_2(X_1 = 1)\mathbb{P}_2(X_n = 4 \text{ for some } n \geq 1 | X_1 = 1) \\ &+ \mathbb{P}_2(X_1 = 3)\mathbb{P}_2(X_n = 4 \text{ for some } n \geq 1 | X_1 = 3) \\ =& \tfrac{1}{2}h_1 + \tfrac{1}{2}h_3, \\ h_3 =& \tfrac{1}{2}h_2 + \tfrac{1}{2}h_4. \end{aligned}$$

Solving, we get $h_2 = 1/3$ and $h_3 = 2/3$.

Theorem 5.7. *The vector of hitting probabilities $\left(h_i^A, i \in I\right)$ is the minimal non-negative solution to the equations*

$$h_i^A = \begin{cases} 1 & \text{if } i \in A \\ \sum_j p_{ij} h_j^A & \text{if } i \notin A \end{cases}. \tag{5.3}$$

Here by "minimal" we mean that if $(x_i, i \in I)$ is another non-negative solution to the system (5.3), then $h_i \leq x_i$ for all i.

Proof. To see that h_i^A satisfies (5.3), we condition on the first jump of the process. If $i \notin A$, then

$$\begin{aligned} h_i^A &= \mathbb{P}_i\left(X_n \in A \text{ for some } n \geq 0\right) \\ &= \mathbb{P}_i\left(X_n \in A \text{ for some } n \geq 1\right) \\ &= \sum_j \mathbb{P}_i(X_1 = j)\mathbb{P}\left(X_n \in A \text{ for some } n \geq 1 \middle| X_0 = i, X_1 = j\right) \\ &= \sum_j p_{ij}\mathbb{P}\left(X_n \in A \text{ for some } n \geq 1 \middle| X_1 = j\right) \\ &= \sum_j p_{ij} h_j^A. \end{aligned}$$

To obtain the penultimate line we applied the Markov property. Meanwhile for $i \in A$, $h_i^A = 1$ by definition. So indeed (5.3) holds.

To prove minimality, suppose $(x_i, i \in I)$ is any non-negative solution to (5.3). We want to show that $h_i^A \leq x_i$ for all i.

We make the following claim: for any $M \in \mathbb{N}$, and for all i,

$$x_i \geq \mathbb{P}_i(X_n \in A \text{ for some } n \leq M). \tag{5.4}$$

We will prove (5.4) by induction on M.

The case $M = 0$ is easy; the LHS is 1 for $i \in A$, while the RHS is 0 for $i \notin A$.

For the induction step, suppose that for all i, $x_i \geq \mathbb{P}_i(X_n \in A \text{ for some } n \leq M - 1)$. If $i \in A$, then again $x_i = 1$ and (5.4) is clear. If $i \notin A$, then

$$\mathbb{P}_i(X_n \in A \text{ for some } n \leq M) = \sum_j p_{ij}\mathbb{P}_i(X_n \in A \text{ for some } n \in \{1, 2, \ldots, M\} | X_1 = j)$$

$$
\begin{aligned}
&= \sum_j p_{ij}\mathbb{P}_j\left(X_n \in A \text{ for some } n \in \{0,1,\dots,M-1\}\right)\\
&\leq \sum_j p_{ij}x_j\\
&= x_i,
\end{aligned}
$$

and the induction step is complete. Hence (5.4) holds for all i and M as desired. Then, using the fact that the sequence of events in (5.4) is increasing in M, we have

$$
\begin{aligned}
x_i &\geq \lim_{M\to\infty} \mathbb{P}_i(X_n \in A \text{ for some } n \leq M)\\
&= \mathbb{P}_i\big(\bigcup_M \{X_n \in A \text{ for some } n \leq M\}\big)\\
&= \mathbb{P}_i(X_n \in A \text{ for some } n)\\
&= h_i^A.
\end{aligned}
$$

□

Important example: "Gambler's ruin"

Let $I = \{0, 1, 2 \dots\}$. Let $p \in (0,1)$ and $q = 1-p$, and consider the transition probabilities given by

$$
\begin{aligned}
p_{00} &= 1\\
p_{i,i-1} &= q \text{ for } i \geq 1 \qquad (5.5)\\
p_{i,i+1} &= p \text{ for } i \geq 1.
\end{aligned}
$$

The name "gambler's ruin" comes from the interpretation where the state is the current capital of a gambler, who repeatedly bets 1 (against an infinitely rich bank). Will the gambler inevitably go broke? But chains like this come up in a wide range of settings. Chains on $\mathbb{Z}_+$ in which all transitions are steps up and down by 1 are called "birth-and-death chains" (modelling the size of a population). This is one of the simplest examples.

Let $h_i = \mathbb{P}_i(\text{hit } 0)$. To find h_i, we need the minimal non-negative solution to

$$h_0 = 1 \qquad (5.6)$$

$$h_i = ph_{i+1} + qh_{i-1} \text{ for } i \geq 1. \qquad (5.7)$$

If $p \neq q$, (5.7) has general solution

$$h_i = A + B\left(\frac{q}{p}\right)^i.$$

We look at three cases:

$\boxed{p<q}$ Jumps downwards are more likely than jumps upward. From (5.6), $A + B = 1$. Then for minimality, we take $A = 1$ and $B = 0$, since $\left(\frac{q}{p}\right)^i \geq 1$ for all i.

We obtain $h_i = 1$ for all i. So with probability 1, the chain will hit 0.

$\boxed{p > q}$ Again $A + B = 1$. Also $\left(\frac{q}{p}\right)^i \to 0$ as $i \to \infty$, so we need $A \geq 0$ for a non-negative solution. Then for a minimal solution, we will want $A = 0$, $B = 1$, since $1 \geq \left(\frac{q}{p}\right)^i$ for all i.

Hence $h_i = \left(\frac{q}{p}\right)^i$. The chain has a positive probability of "escaping to infinity".

$\boxed{p = q}$ Now the general solution of (5.7) is $h_i = A + Bi$. From $i = 0$ we get $A = 1$. For non-negativity we need $B \geq 0$, and then for minimality $B = 0$. We get $h_i = 1$ again. Now there is no drift, but still with probability 1 the chain will hit 0 eventually.

Remark. Notice that we could have seen $h_i = \alpha^i$ for some α, by a direct argument. Since the chain can only descend by one step at a time,

$$\mathbb{P}_i(\text{hit } 0) = \mathbb{P}_i(\text{hit } i-1)\mathbb{P}_{i-1}(\text{hit } i-2)\ldots\mathbb{P}_1(\text{hit } 0), \tag{5.8}$$

and all terms in the product are the same, since the transition probabilities are the same at every level.

5.8 Recurrence and transience

If the chain starts in state i, what is the chance that it returns to i at some point in the future? We can distinguish two possibilities:

(1)
$$\mathbb{P}_i(X_n = i \text{ for some } n \geq 1) = p < 1.$$

Then the total number of visits to i has geometric distribution with parameter $1-p$ (since each time we return to i, we have chance $1-p$ of never returning again). We have
$$\mathbb{P}_i(\text{hit } i \text{ infinitely often}) = 0.$$

The state i is called **transient**.

(2)
$$\mathbb{P}_i(X_n = i \text{ for some } n \geq 1) = 1.$$

Then
$$\mathbb{P}_i(\text{hit } i \text{ infinitely often}) = 1.$$

The state i is called **recurrent**.

The definition is very simple, but the concept of recurrence and transience is extremely rich (mainly for infinite chains).

There is an important criterion for recurrence and transience in terms of the transition probabilities:

Theorem 5.8. *State i is recurrent if and only if $\sum_{n=0}^{\infty} p_{ii}^{(n)} = \infty$.*

Proof. The total number of visits to i is $\sum_{n=0}^{\infty} \mathbf{1}\{X_n = i\}$ which has expectation

$$\sum_{n=0}^{\infty} \mathbb{E}\,\mathbf{1}\{X_n = i\} = \sum_{n=0}^{\infty} \mathbb{P}(X_n = i) = \sum_{n=0}^{\infty} p_{ii}^{(n)}.$$

If i is transient, the number of visits to i is geometric with parameter $1-p$, and hence with mean $\frac{1}{1-p} < \infty$.

On the other hand if i is recurrent, the number of visits to i is infinite with probability 1, and so has mean ∞.

This gives the statement of the theorem. □

Theorem 5.9. *(a) Let C be a communicating class. Either all states in C are recurrent, or all are transient (so we may refer to the whole class as transient or recurrent).*

(b) Every recurrent class is closed.

(c) Every finite closed class is recurrent.

Proof. Exercises – see example sheet 3. For part (a), use Theorem 5.8 to show that if i is recurrent and $i \leftrightarrow j$, then j is also recurrent. □

The theorem tells us that recurrence and transience are quite boring for finite chains: state i is recurrent if and only if its communicating class is closed. But infinite chains are more interesting! An infinite closed class may be either transient or recurrent.

5.9 Random walk in $\mathbb{Z}^d$

Consider a simple symmetric random walk on the d-dimensional integer lattice. This is a Markov chain with state space $\mathbb{Z}^d$ and transition probabilities $p_{xy} = 1/2d$ if $|x-y| = 1$, and $p_{xy} = 0$ otherwise. The chain is irreducible, with period 2.

In this section we will show that the random walk is recurrent when $d = 1$ or $d = 2$ but transient in higher dimensions.

5.9.1 d=1

The analysis after (5.7) (for $p = q = 1/2$) shows us that for the simple symmetric random walk on $\mathbb{Z}$, the hitting probability of 0 from any $i > 0$ is 1. By symmetry, the same is true from any negative state. This shows that starting from 0, the probability of returning to 0 is 1. Hence state 0 is recurrent (and so by irreducibility the whole chain is recurrent).

An alternative approach uses Theorem 5.8. This gives a good warm-up for the approach we will use in higher dimensions.

We need to show that $\sum_{n=0}^{\infty} p_{00}^{(n)} = \infty$. We will use Stirling's formula, which tells us that

$$n! \sim \sqrt{2\pi}\, n^{n+1/2} e^{-n} \text{ as } n \to \infty. \tag{5.9}$$

(The constant $\sqrt{2\pi}$ will not be important.)

Suppose $X_0 = 0$. If n is odd, then $\mathbb{P}_0(X_n = 0) = 0$, since the chain has period 2. For $X_{2m} = 0$ we need m "ups" and m "downs" in the first $2m$ steps. Applying Stirling's formula to the binomial probability we obtain

$$\begin{aligned} p_{00}^{(2m)} &= \binom{2m}{m}\left(\frac{1}{2}\right)^{2m} \\ &= \frac{(2m)!}{m!m!}\left(\frac{1}{2}\right)^{2m} \\ &\sim \frac{1}{\sqrt{\pi}}\frac{1}{m^{1/2}}. \end{aligned} \tag{5.10}$$

Since $\sum m^{-1/2} = \infty$, we have $\sum p_{00}^{(n)} = \infty$ and the chain is recurrent.

Exercise. *Use Stirling's formula to show that if $p \neq q$, then the chain is transient. (We could also deduce this from the hitting probability analysis after (5.7).)*

5.9.2 d=2

Why should the walk be recurrent in 1 and 2 dimensions but transient in 3 dimensions? An intuitive answer is as follows. A d-dimensional random walk behaves in some sense like d independent 1-dimensional walks. For the d-dimensional walk to be back at the origin, we require all d of the 1-dimensional walks to be at 0. From (5.10), the probability that a 1-dimensional walk is at 0 decays like $m^{-1/2}$. Hence the probability that a 2-dimensional walk is at the origin decays like m^{-1}, which sums to infinity, leading to recurrence, while the corresponding probability for a 3-dimensional walk decays like $m^{-3/2}$ which has finite sum, leading to transience.

In two dimensions we can make this precise in a very direct way. Let X_n be the walk in $\mathbb{Z}^2$ and consider its projections onto the diagonal lines $x = y$ and $x = -y$ in the plane.

Each step of the walk increases or decreases the projection onto $x = y$ by $1/\sqrt{2}$, and also increases or decreases the projection onto $x = -y$ by $1/\sqrt{2}$. All four possibilities are equally likely.

Hence if we write W_n^+ and W_n^- for the two projections of X_n, we have that the processes W_n^+ and W_n^- are independent of each other, and both of them are simple symmetric random walks on $2^{-1/2}\mathbb{Z}$.

Then we have

$$\begin{aligned} \mathbb{P}(X_{2m} = 0) &= \mathbb{P}(W_{2m}^+ = 0)\mathbb{P}(W_{2m}^- = 0) \\ &\sim \left(\frac{1}{\sqrt{\pi}}\frac{1}{m^{1/2}}\right)^2 \\ &= \frac{1}{\pi m}. \end{aligned}$$

Hence $\sum p_{00}^{(2m)} = \infty$ and the walk is recurrent.

5.9.3 d=3

The trick from the previous section does not work in $d = 3$, so we need to do a little more combinatorics. As the walk has period 2 we have a positive chance of return to the origin only when n is even. Each step is $\pm e_1, \pm e_2$ or $\pm e_3$ where $e_i, i = 1, 2, 3$ are the three unit coordinate vectors. To return to the origin after $2m$ steps, we should have made, say, i steps in each of the directions $\pm e_1$, j steps in each of the directions $\pm e_2$, and k steps in each of the directions $\pm e_3$ for some i, j, k with $i + j + k = m$. Considering all the possible orderings of these steps among the first $2m$ steps of the walk, we get

$$\begin{aligned} p_{00}^{(2m)} &= \sum_{\substack{i,j,k\geq 0\\ i+j+k=m}} \frac{(2m)!}{i!^2 j!^2 k!^2} \left(\frac{1}{6}\right)^{2m} \\ &= \binom{2m}{m} \left(\frac{1}{2}\right)^{2m} \sum_{\substack{i,j,k\geq 0\\ i+j+k=m}} \binom{m}{i,j,k}^2 \left(\frac{1}{3}\right)^{2m} \\ &\leq \binom{2m}{m} \left(\frac{1}{2}\right)^{2m} \left(\sum_{\substack{i,j,k\geq 0\\ i+j+k=m}} \binom{m}{i,j,k} \left(\frac{1}{3}\right)^{m} \right) \max_{\substack{i,j,k\geq 0\\ i+j+k=m}} \binom{m}{i,j,k} \left(\frac{1}{3}\right)^{m}. \end{aligned} \tag{5.11}$$

Here, if $i + j + k = m$, we write $\binom{m}{i,j,k} = \frac{m!}{i!j!k!}$. Note that

$$\sum_{\substack{i,j,k\geq 0\\ i+j+k=m}} \binom{m}{i,j,k} \left(\frac{1}{3}\right)^{m} = 1,$$

since it is the sum of the mass function of a "trinomial$(1/3, 1/3, 1/3)$" distribution (consider the number of ways of putting 3 balls into m boxes).

If m is divisible by 3, say $m = 3r$, then it is easy to check that the max in (5.11) is attained when $i = j = k = r$, giving

$$\begin{aligned} p_{00}^{(2m)} &\leq \binom{2m}{m} \left(\frac{1}{2}\right)^{2m} \binom{m}{m/3, m/3, m/3} \left(\frac{1}{3}\right)^{m} \\ &\sim \frac{1}{\sqrt{2\pi}} \frac{1}{m^{1/2}} \times \frac{1}{2\pi} \frac{1}{m} \\ &\sim \frac{1}{(2\pi)^{3/2}} m^{-3/2}, \end{aligned}$$

where we used Stirling's formula again for the last line. Hence we have $\sum_{r=0}^{\infty} p_{00}^{(6r)} < \infty$.

Note also that $p_{00}^{(6r)} \geq \left(\frac{1}{6}\right)^2 p_{00}^{(6r-2)}$ and $p_{00}^{(6r)} \geq \left(\frac{1}{6}\right)^4 p_{00}^{(6r-4)}$, so overall, $\sum_{n=0}^{\infty} p_{00}^{(n)} < \infty$, and the walk is transient.

5.9.4 $\mathbf{d \geq 4}$

If we have a walk on $\mathbb{Z}^d$ for $d \geq 4$, we can obtain from it a walk on $\mathbb{Z}^3$ by looking only at the first 3 coordinates, and ignoring any transitions that do not change them. Since we know that a walk on $\mathbb{Z}^3$ only visits the origin finitely often, the same must be true for the walk in higher dimensions also. Hence we have transience for all $d \geq 3$.

5.9.5 Mean hitting time

Let $H^A = \inf\{n \geq 0 \colon X_n \in A\}$, the first hitting-time of the set A, with the convention that $H^A = \infty$ if $X_n \notin A$ for all $n \geq 1$. In fact if h_i^A is the hitting probability defined above, then $h_i^A = \mathbb{P}_i(H^A < \infty)$.

Let $k_i^A = \mathbb{E}_i(H^A)$, the mean hitting time of A from i. If $h_i^A < 1$, then $\mathbb{P}_i(H^A = \infty) > 0$ and certainly $k_i^A = \infty$. Also maybe $k_i^A = \infty$ even when $h_i^A = 1$.

Theorem 5.10. *The vector of mean hitting times $k^A = (k_i^A, i \in S)$ is the minimal non-negative solution to*

$$k_i^A = \begin{cases} 0 & \text{if } i \in A \\ 1 + \sum_j p_{ij} k_j^A & \text{if } i \notin A \end{cases}.$$

Proof. "Condition on the first jump" again. For $i \notin A$,

$$\begin{aligned} k_i^A = \mathbb{E}_i(H^A) &= \sum_j \mathbb{E}_i(H^A | X_1 = j)\mathbb{P}_i(X_1 = j) \\ &= \sum_j p_{ij}(1 + k_j^A) \\ &= 1 + \sum_j p_{ij} k_j^A. \end{aligned}$$

For the minimality, one can use a similar idea to that at (5.4) in the proof of Theorem 5.7 above. Specifically, one can show by induction that if (y_i) is any non-negative solution to the recursions, then $y_i \geq \mathbb{E}_i \min\big(H^A, m\big)$ for all $m \geq 0$; we omit the details. □

5.9.6 Gambler's ruin, continued

What is the expected hitting time of 0 from state i in the gambler's ruin chain at (5.5)?

Let k_i be the mean time to hit 0 starting from i. We give brief details (of course, one can be more formal!).

If $k_i < \infty$, one can see that $k_i = \beta i$ for some β, since (compare the remark at (5.8) above),

$$\mathbb{E}_i(\text{time to } 0) = \mathbb{E}_i(\text{time to } i-1) + \mathbb{E}_{i-1}(\text{time to } i-2) + \cdots + \mathbb{E}_1(\text{time to } 0).$$

To satisfy the recursion in Theorem (5.10), we need

$$k_i = 1 + qk_{i-1} + pk_{i+1}$$

which leads to $(q-p)\beta = 1$. We obtain:

$\boxed{p < q}$ $k_i = \frac{1}{q-p} i$; the chain takes a finite time on average to hit 0.

$\boxed{p > q}$ We already know $h_i < 1$, so certainly $k_i = \infty$.

$\boxed{p = q}$ There is no suitable β, so $k_i = \infty$ here also, even though $h_i = 1$. The chain hits 0 with probability 1, but the mean time to arrive there is infinite.

5.10 Null recurrence and positive recurrence

Define m_i, the **mean return time** to a state i by

$$\begin{aligned} m_i &:= \mathbb{E}_i \left(\inf\{n \geq 1 : X_n = i\} \right) \\ &= 1 + \sum p_{ij} k_j^{\{i\}} \end{aligned} \tag{5.12}$$

(where $k_j^{\{i\}}$ is the mean hitting time of i starting from j).

This quantity will be particularly important when we consider equilibrium behaviour of a Markov chain – loosely speaking, the long-run proportion of time spent in state i ought to be the reciprocal of the mean return time.

If i is transient, then certainly $m_i = \infty$ (since the return time itself is infinite with positive probability).

If i is recurrent, then the return time is also finite, but nonetheless the mean could be infinite.

If i is recurrent but $m_i = \infty$, the state i is said to be **null recurrent**.

If $m_i < \infty$ then the state i is said to be **positive recurrent**.

For similar reasons to those in Theorem 5.9, null recurrence and positive recurrence are *class properties*; if one state in a communicating class is null (resp. positive) recurrent, then every state in the class is null (resp. positive) recurrent.

If the chain is irreducible, we can therefore call the whole chain either transient, or null recurrent, or positive recurrent.

6

Markov chains: stationary distributions and convergence to equilibrium

6.1 Stationary distributions

Let $\pi = (\pi_i, i \in I)$ be a distribution on the state space I.

We say that π is a **stationary distribution**, or **invariant distribution**, or **equilibrium distribution**, for the transition matrix P if

$$\boxed{\pi P = \pi}.$$

That is, for all j, $\pi_j = \sum_i \pi_i p_{ij}$. The row vector π is a left eigenvector for the matrix P, with eigenvalue 1.

If X_0 has distribution π, then we know that X_n has distribution πP^n. Hence if π is stationary, then X_n has distribution π for all n. It follows that the sequence

$$(X_n, X_{n+1}, X_{n+2}, \dots)$$

has the same distribution as

$$(X_0, X_1, X_2, \dots)$$

for any n.

6.2 Main theorems

Theorem 6.1 (Existence and uniqueness of stationary distributions)**.** *Let P be an irreducible transition matrix.*

(a) P has a stationary distribution if and only if P is positive recurrent.

(b) In that case, the stationary distribution π is unique, and is given by $\pi_i = 1/m_i$ for all i (where m_i is the mean return time to state i defined at (5.12)).

Theorem 6.2 (Convergence to equilibrium)**.** *Suppose P is irreducible and aperiodic, with stationary distribution π. If X_n is a Markov chain with transition matrix P and any initial distribution, then for all $j \in I$,*

$$\mathbb{P}(X_n = j) \to \pi_j \text{ as } n \to \infty.$$

In particular,

$$p_{ij}^{(n)} \to \pi_j \text{ as } n \to \infty, \text{ for all } i \text{ and } j.$$

Theorem 6.3 (Ergodic theorem)**.** *Let P be irreducible. Let $V_i(n)$ be the number of visits to state i before time n, that is*

$$V_i(n) = \sum_{r=0}^{n-1} I\left(X_r = i\right).$$

Then for any initial distribution, and for all $i \in I$,

$$\frac{V_i(n)}{n} \to \frac{1}{m_i} \text{ almost surely, as } n \to \infty.$$

That is,

$$\mathbb{P}\left(\frac{V_i(n)}{n} \to \frac{1}{m_i} \text{ as } n \to \infty\right) = 1.$$

The ergodic theorem concerns the “long-run proportion of time” spent in a state.

In the positive recurrent case, $1/m_i = \pi_i$ where π is the stationary distribution, so the ergodic theorem says that (with probability 1) the long-run proportion of time spent in a state is the stationary probability of that state.

In the null-recurrent or transient case, $1/m_i = 0$, so the ergodic theorem says that with probability 1 the long-run proportion of time spent in a state is 0.

We can see the ergodic theorem as a generalisation of the strong law of large numbers. If X_n is an i.i.d. sequence, then the strong law tells us that, with probability 1, the long run proportion of entries in the sequence which are equal to i is equal to the probability that any given entry is equal to i. The ergodic theorem can be seen as extending this to the case where X_n is not i.i.d. but is a Markov chain.

6.3 Examples of stationary distributions

Example 6.4. Let $P = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 1/2 & 1/2 \\ 1/2 & 0 & 1/2 \end{pmatrix}$. (Draw the diagram of the chain.)

For π to be stationary, we need

$$\begin{aligned} \pi_1 &= \tfrac{1}{2}\pi_3 \\ \pi_2 &= \pi_1 + \tfrac{1}{2}\pi_2 \\ \pi_3 &= \tfrac{1}{2}\pi_2 + \tfrac{1}{2}\pi_3. \end{aligned}$$

One of these equations is redundant, and we need the added relation $\pi_1 + \pi_2 + \pi_3 = 1$ to normalise the solution (so that π is a distribution).

Solving, we obtain $(\pi_1, \pi_2, \pi_3) = (1/5, 2/5, 2/5)$.

Correspondingly, the vector of mean return times is given by $(m_1, m_2, m_3) = (5, 5/2, 5/2)$.

Note that this chain is aperiodic and irreducible. Hence, from any initial state, the distribution at time n converges to π as $n \to \infty$. For example, $p_{11}^{(n)} \to 1/5$ as $n \to \infty$.

By the way, be careful to solve the equation $\pi P = \pi$ and not to solve $P\pi = \pi$ by mistake! For any transition matrix P, the equation $P\pi = \pi$ is solved by any vector π all of whose entries are the same (why is this?) which could trap you into thinking that the uniform distribution is stationary, which, of course, is not the case in general. We want the left eigenvector, rather than the right eigenvector.

Example 6.5. Recall the example of a simple symmetric random walk on a cycle of size M in Section 5.3. The distribution π with $\pi_i = 1/M$ for all i is stationary, since it solves

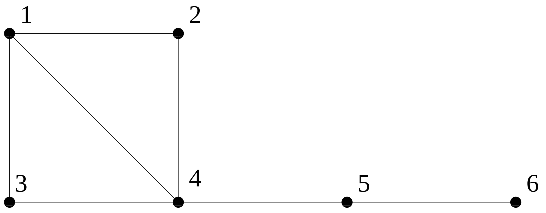
$$\pi_i = \tfrac{1}{2}\pi_{i+1} + \tfrac{1}{2}\pi_{i-1}$$

for each i. Because of the symmetry of the chain, it is not surprising that the stationary distribution is uniform.

Is it the true that $p_{00}^{(n)} \to 1/M$ as $n \to \infty$? If M is odd, then the chain is aperiodic (check this!), so the answer is yes.

However, if M is even then the chain has period 2. Then $p_{00}^{(n)} = 0$ whenever n is odd. In fact $p_{00}^{(2m)} \to 2\pi_0 = 2/M$ as $m \to \infty$ (exercise; consider the 2-step chain X_0, X_2, $X_4, \ldots$ on the subset of the state space which consists just of even sites. Is it irreducible? What is its stationary distribution?)

Example 6.6 (Random walk on a graph)**.** A "graph" in the combinatorial sense is a collection of *vertices* joined by *edges*. For example, the following graph has 6 vertices and 7 edges.



Let I be the set of vertices. Two vertices are *neighbours* in the graph if they are joined by an edge. The *degree* of a vertex is its number of neighbours. Let d_i be the degree of vertex i. In the graph above, the vector of vertex degrees is $(d_i, i \in I) = (3, 2, 2, 4, 2, 1)$.

Assume $d_i > 0$ for all i. A *random walk* on the graph is a Markov chain with state space I, evolving as follows; if i is the current vertex, then at the next step move to each of the neighbours of i with probability $1/d_i$.

Assume irreducibility of the chain (equivalently, that there is a path between any two vertices in the graph). Then the stationary distribution of the chain π is unique.

In fact, the stationary probability of a vertex is proportional to its degree. To show this, we will check that $dP = d$ where d is the vector of vertex degrees and P is the transition matrix of the chain:

$$d_j = \sum_i \mathbf{1}(i \text{ is a neighbour of } j)$$

$$\begin{aligned}&= \sum_i d_i \frac{1}{d_i} \mathbf{1}(i \text{ is a neighbour of } j)\\ &= \sum_i d_i p_{ij},\end{aligned}$$

as required.

To obtain the stationary distribution we simply need to normalise d. So we obtain $\pi_i = d_i / \sum_j d_j$.

For the graph above, $\sum_j d_j = 14$, and we obtain

$$\pi = \left(\frac{3}{14}, \frac{1}{7}, \frac{1}{7}, \frac{2}{7}, \frac{1}{7}, \frac{1}{14} \right).$$

From this we can deduce the mean return times. For example, $m_1 = 1/\pi_1 = 14/3$.

Notice that the chain is aperiodic. As a result, we also have convergence to the stationary distribution. For example, starting from any initial distribution, the probability that the walk is at vertex 1 at step n converges to $3/14$ as $n \to \infty$.

Example 6.7 (One-dimensional random walk)**.** Consider again the familiar example of a one-dimensional random walk. Let $I = \{0, 1, 2, \dots\}$ and let

$$\begin{aligned}p_{i,i+1} &= p \text{ for } i \geq 0,\\ p_{i,i-1} &= q = 1 - p \text{ for } i \geq 1,\\ p_{0,0} &= q.\end{aligned}$$

If $p > q$, we found previously that the walk is transient, so no stationary distribution will exist.

If $p = q$, the walk is recurrent, but the mean return time is infinite, so again there is no stationary distribution.

If $p < q$, the walk is positive recurrent. For stationarity, we need $\pi_i = \pi_{i-1} p + \pi_{i+1} q$ for $i \geq 1$. This is (not coincidentally) reminiscent of the hitting probability equation we previously found for the model (except the values of p and q are reversed). It has general solution $\pi_i = A + B(p/q)^i$.

We need $\sum \pi_i = 1$, which forces $A = 0$ and $B = (1 - p/q)$, giving

$$\pi_i = \left(1 - \frac{p}{q}\right) \left(\frac{p}{q}\right)^i.$$

That is, the stationary distribution of the walk is geometric with parameter $1 - \frac{p}{q}$.

Example 6.8 (A two-state chain and a non-irreducible chain)**.**

Consider the two-state chain on $\{1, 2\}$ with transition matrix $P = \begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix}$.

Solving $\pi P = \pi$ and normalising we obtain that $\pi = \left(\frac{\beta}{\alpha+\beta}, \frac{\alpha}{\alpha+\beta}\right)$.

Notice that this agrees with what we found in Example 5.4; the expression for $p_{11}^{(n)}$ given in (5.2) satisfies $p_{11}^{(n)} \to \pi_1 = \frac{\beta}{\alpha+\beta}$ as $n \to \infty$, as it should do because of the convergence to equilibrium in Theorem 6.2.

Now consider the chain on $\{1,2,3,4\}$ whose transition matrix is

$$P = \begin{pmatrix} 1-\alpha & \alpha & 0 & 0 \\ \beta & 1-\beta & 0 & 0 \\ 0 & 0 & 1-\gamma & \gamma \\ 0 & 0 & \delta & 1-\delta \end{pmatrix}.$$

This chain is not irreducible. We can view it as two separate two-state chains, on $\{1,2\}$ and $\{3,4\}$, with no communication between them. Both $\left(\frac{\beta}{\alpha+\beta}, \frac{\alpha}{\alpha+\beta}, 0, 0\right)$ and $\left(0, 0, \frac{\delta}{\gamma+\delta}, \frac{\gamma}{\gamma+\delta}\right)$ are stationary distributions. But also any mixture of these is stationary (since if $\pi^{(1)}$ and $\pi^{(2)}$ are eigenvectors of P with eigenvalue 1, then so is any linear combination of $\pi^{(1)}$ and $\pi^{(2)}$).

So any distribution

$$\left(x\frac{\beta}{\alpha+\beta}, x\frac{\alpha}{\alpha+\beta}, (1-x)\frac{\delta}{\gamma+\delta}, (1-x)\frac{\gamma}{\gamma+\delta}\right),$$

where $x \in [0,1]$, is stationary. (In fact, these are all the stationary distributions – exercise). The uniqueness result in Theorem 6.1 does not apply because the transition matrix is not irreducible.

6.4 Proof of Theorems 6.1, 6.2 and 6.3. *(non-examinable)*

The proofs below are given partly rather informally. They are not examinable; however, they are very helpful in developing your intuition about the results. The "coupling" idea used in the proof of Theorem 6.2 is particularly pretty and I certainly recommend thinking about it, but it will not be examined. (The results themselves are very much examinable!)

Proof of Theorem 6.3. This proof is essentially an application of the strong law of large numbers.

If the chain is transient, then with probability 1 there are only finitely many visits to any state, so $V_i(n)$ is bounded with probability 1. So

$$\mathbb{P}\left(\frac{V_i(n)}{n} \to 0 \text{ as } n \to \infty\right) = 1,$$

which is the result we want since $m_i = \infty$.

Suppose instead that the chain is recurrent. In this case we will visit state i infinitely often. Let R_k be the time between the kth and the $(k+1)$st visits to i. Then $R_1, R_2, R_3, \ldots$ are i.i.d. with mean m_i (which is finite in the positive recurrent case and infinite in the null recurrent case).

So by the strong law of large numbers,

$$\mathbb{P}\left(\frac{R_1 + R_2 + \cdots + R_K}{K} \to m_i \text{ as } K \to \infty\right) = 1. \tag{6.1}$$

Let T_K be the time of the Kth visit to i. Then $T_K = T_1 + X_1 + X_2 + \cdots + X_{K-1}$. It is easy to obtain that, for any c, $T_K/K \to c$ if and only if $(R_1 + \cdots + R_K)/K \to c$. Hence from (6.1) we have

$$\mathbb{P}\left(\frac{T_K}{K} \to m_i \text{ as } k \to \infty\right) = 1. \tag{6.2}$$

Notice that T_K/K is the time per visit (averaged over the first K visits) whereas $V_i(n)/n$ is the number of visits per unit time (averaged over the first n times). It is straightforward to obtain (check!) that, for any c, $T_K/K \to c$ as $K \to \infty$ if and only if $V_i(n)/n \to 1/c$ as $n \to \infty$.

Hence from (6.2) we have

$$\mathbb{P}\left(\frac{V_i(n)}{n} \to \frac{1}{m_i} \text{ as } n \to \infty\right) = 1$$

as required. □

Lemma 6.9. *If P is positive recurrent, then it has stationary distribution π with $\pi_i = 1/m_i$.*

Proof. We give an informal version of the proof, which could quite easily be made rigorous.

From the ergodic theorem, we know that (with probability 1) the long-run proportion of visits to state i is $1/m_i$.

Each time the chain visits state i, it has probability p_{ij} of jumping from there to state j. We can obtain that the long-run proportion of jumps from i to j is $\frac{1}{m_i}p_{ij}$.

First consider the case where the state space I is finite. By summing over $i \in I$, we get that the long-run proportion of jumps into state j is $\sum_i \frac{1}{m_i}p_{ij}$.

But the long-run proportion of jumps into j is the same as the long-run proportion of visits to j, which (by the ergodic theorem) is $1/m_j$.

We obtain

$$\frac{1}{m_j} = \sum_i \frac{1}{m_i} p_{ij},$$

i.e. $\pi_j = \sum_i \pi_i p_{ij}$, so that π satisfies $\pi P = \pi$ and is stationary as desired.

If i is infinite, it is not immediate that the long-run proportion of jumps into j is the sum over i of the long-run proportions of jumps from i to j. However (by considering as large a finite set of i as desired) the second quantity does give an upper bound for the first, so we get

$$\frac{1}{m_j} \leq \sum_i \frac{1}{m_i} p_{ij}$$

for all $j \in I$. But summing both sides over j gives the same (finite) amount, since $\sum_j p_{ij} = 1$ for all i. So in fact we must have equality for all j as required. □

Lemma 6.10. *If π is any stationary distribution then $\pi_i = 1/m_i$.*

Proof. Suppose π is stationary for P, and let X be a Markov chain with initial distribution π and transition matrix P. Then by stationarity, $\mathbb{P}(X_n = i) = \pi_i$ for all n, and

$$\begin{aligned}\frac{\mathbb{E}\, V_n(i)}{n} &= \frac{1}{n}\sum_{r=0}^{n-1} \mathbb{E}\left(\mathbf{1}\{X_n = i\}\right) \\ &= \frac{1}{n}\sum_{r=0}^{n-1} \mathbb{P}(X_n = i) \\ &= \pi_i. \end{aligned} \tag{6.3}$$

From the ergodic theorem, for any ϵ

$$\mathbb{P}\left(\left|\frac{V_n(i)}{n} - \frac{1}{m_i}\right| > \epsilon\right) < \epsilon \tag{6.4}$$

for large enough n (since almost sure convergence implies convergence in probability).

But since $V_n(i)/n$ is bounded between 0 and 1, it follows from (6.4) (check!) that

$$\frac{\mathbb{E}\, V_n(i)}{n} \to \frac{1}{m_i} \text{ as } n \to \infty.$$

Comparing to (6.3), we obtain $\pi_i = 1/m_i$. □

This gives uniqueness of the stationary distribution for positive recurrent chains, and shows that no stationary distribution can exist for null recurrent and transient chains. So we have proved Theorem 6.1.

Finally, we prove the result on convergence to equilibrium.

Proof of Theorem 6.2. Let P be irreducible and aperiodic, with stationary distribution π.

Let λ be any initial distribution, and let $(X_n, n \geq 0)$ be Markov(λ, P). We wish to show that $\mathbb{P}(X_n = j) \to \pi_j$ as $n \to \infty$, for any j.

Consider another chain $(Y_n, n \geq 0)$ which is Markov(π, P), and which is independent of Z. Since π is stationary, Y_n has distribution π for all n.

Let $T = \inf\{n \geq 0 : X_n = Y_n\}$. We will claim that $\mathbb{P}(T < \infty) = 1$; that is, the chains X and Y will meet at some point.

Suppose this claim is true. Then define another chain Z by

$$Z_n = \begin{cases} X_n & \text{if } n < T \\ Y_n & \text{if } n \geq T \end{cases}.$$

The idea is that Z starts in distribution λ, and evolves independently of the chain Y, until they first meet. As soon as that happens, Z copies the moves of Y exactly.

Then Z is also Markov(λ, P), since Z_n starts in distribution λ and each jump is done according to P, first by copying X up to time T, and then by copying Y after time T.

The idea is that the chain Y is "in equilibrium" (since it starts in the equilibrium distribution π) so that if there is high probability that $Y_n = Z_n$, then the distribution of Z_n must be close to π. More precisely:

$$\begin{aligned} |\mathbb{P}(Z_n = j) - \pi_j| &= |\mathbb{P}(Z_n = j) - \mathbb{P}(Y_n = j)| \\ &\leq \mathbb{P}(Z_n \neq Y_n) \\ &= \mathbb{P}(T > n) \\ &\to 0 \text{ as } n \to \infty. \end{aligned}$$

Then we have $\mathbb{P}(Z_n = j) \to \pi_j$.

But the chains X and Z have the same distribution (they are both Markov(λ, P)). So we have also shown that $\mathbb{P}(X_n = j) \to \pi_j$, as required.

It remains to prove the claim that T is finite with probability 1. Fix any state $b \in I$ and define $T_b = \inf\{n \geq 0 : X_n = Y_n = b\}$. Then $T \leq T_b$. We will show that T_b is finite with probability 1.

Consider the process $W_n = (X_n, Y_n)$. Since X_n and Y_n evolve independently, W_n is a Markov chain on the state space $I \times I$ with transition probabilities

$$\tilde{p}_{(i,k)(j,l)} = p_{ij} p_{kl},$$

and initial distribution

$$\mu_{(i,k)} = \lambda_i \pi_k.$$

P is aperiodic and irreducible, so for all i, j, k, l, we have that

$$\tilde{p}^{(n)}_{(i,k)(j,l)} = p^{(n)}_{ij} p^{(n)}_{kl} > 0$$

for all large enough n. So $\tilde{P}$ is irreducible.

$\tilde{P}$ has an invariant distribution given by

$$\tilde{\pi}_{(i,k)} = \pi_i \pi_k.$$

Hence $\tilde{P}$ is recurrent (by Theorem 6.1). But $T_b = \inf\{n \geq 0 : W_n = (b, b)\}$. Then indeed $\mathbb{P}(T_b < \infty) = 1$ (since an irreducible recurrent chain visits every state with probability 1). □

Notice where the argument above fails when P is periodic. The chain $W_n = (X_n, Y_n)$ still has the stationary distribution of the form above, but it is not irreducible, so it may never reach the state (b, b). (For example, if P has period 2, and the chains X and Y start out with "opposite parity", then they will never meet).

7

Poisson processes

A Poisson process is a natural model for a stream of events occuring one by one in continuous time, in an uncoordinated way. For example: the process of times of detections by a Geiger counter near a radioactive source (a very accurate model); the process of times of arrivals of calls at a call centre (often a good model); the process of times of arrivals of buses at a bus stop (probably an inaccurate model; different buses are not really uncoordinated, for various reasons).

Consider a random process $N_t, t \in [0, \infty)$. (Note that "time" for our process is now a continuous rather than a discrete set!)

Such a process is called a *counting process* if N_t takes values in $\{0, 1, 2, \dots\}$, and $N_s \leq N_t$ whenever $s \leq t$. We will also assume that $t \mapsto N_t$ is right-continuous.

If N_t describes an arrival process, then $N_t = k$ means that there have been k arrivals in the time interval $[0, t]$. In fact we can describe the process by the sequence of arrival times, which we might call "points" of the process. Let $T_k = \inf\{t \geq 0 : N_t \geq k\}$ for $k \geq 0$. Then $T_0 = 0$ and T_k is the "kth arrival time", for $k \geq 1$. We also define $Y_k = T_k - T_{k-1}$ for $k \geq 1$. Y_k is the "interarrival time" between arrivals $k-1$ and k.

For $s < t$, we write $N(s, t]$ for $N_t - N_s$, which we can think of as the number of points of the process which occur in the time-interval $(s, t]$. This is also called the "increment" of the process N on the interval $(s, t]$.

7.1 Poisson process: a choice of definitions

Let $\lambda > 0$. We will give two different definitions for what it means to be a **Poisson process of rate** λ. Afterwards we will show that these definitions are equivalent.

Definition 7.1 (Definition of Poisson process via exponential interarrival times)**.** $(N_t, t \geq 0)$ *is a Poisson process of rate* λ *if its interarrival times* $Y_1, Y_2, Y_3, \dots$ *are i.i.d. with* $\mathrm{Exp}(\lambda)$ *distribution.*

Definition 7.2 (Definition of Poisson process via Poisson distribution of increments)**.** $N_t, t \geq 0$ *is a Poisson process of rate* λ *if:*

(i) $N_0 = 0$.

(ii) If (s_1, t_1), $(s_2, t_2), \dots, (s_k, t_k)$ *are disjoint intervals in* $\mathbb{R}_+$, *then the increments* $N(s_1, t_1]$, $N(s_2, t_2], \dots, N(s_k, t_k]$ *are independent, where* $N(s_i, t_i] = N_{t_i} - N_{s_i}$.

(iii) For any $s < t$, *the increment* $N(s, t]$ *has Poisson distribution with mean* $\lambda(t - s)$.

Property (ii) in Definition 7.2 is called the **independent increments** property. The number of points falling in disjoint intervals is independent.

This can be seen as a version of the Markov property. For any t_0, the distribution of the process $(N(t_0, t_0 + t], t \geq 0)$, is independent of the process $(N_t, t \leq t_0)$. Put another way, the distribution of $(N_t, t > t_0)$ conditional on the process $(N_t, t \leq t_0)$ depends only on the value N_{t_0}.

7.2 Equivalence of the definitions

We wish to show that the properties listed in Definitions 7.1 and 7.2 are equivalent. The key idea is that the memoryless property for the exponential distribution and the independent increments property are telling us the same thing. The argument below is somewhat informal (but can be made completely rigorous).

Interrarival definition implies independent Poisson increments definition

Suppose we have Definition 7.1 in terms of i.i.d. exponential interarrival times. We wish to show that it implies the statements in Definition 7.2.

Property (i) is immediate: since $Y_1 = T_1 = \inf\{t \geq 0 : N_t \geq 1\}$ is strictly positive with probability 1, also $N_0 = 0$ with probability 1.

Now let us consider the distribution of the number of points in an interval. First let us take $s = 0$ in (iii), and consider $N(0, t]$. We want $N(0, t] \sim \text{Poisson}(\lambda t)$, i.e. that for all k,

$$\mathbb{P}\big(N(0, t] = k\big) = \frac{e^{-\lambda t}(\lambda t)^k}{k!}. \tag{7.1}$$

But we can rewrite the event on the LHS in terms of T_k and T_{k+1}. Since T_k is the sum of k independent exponentials of rate λ, we have $T_k \sim \text{Gamma}(k, \lambda)$, and similarly $T_{k+1} \sim \text{Gamma}(k + 1, \lambda)$. So

$$\begin{aligned} \mathbb{P}\big(N(0, t] = k\big) &= \mathbb{P}\left(T_k \leq t, T_{k+1} > t\right) \\ &= \mathbb{P}\left(T_k \leq t\right) - \mathbb{P}\left(T_{k+1} \leq t\right) \\ &= \int_0^t \frac{\lambda^k x^{k-1} e^{-\lambda x}}{(k-1)!} dx - \int_0^t \frac{\lambda^{k+1} x^k e^{-\lambda x}}{k!} dx. \end{aligned} \tag{7.2}$$

Now we can check that the RHS of (7.1) and (7.2) are the same (for example, either by integrating by parts in (7.2), or by differentiating in (7.1)). In this way we obtain that indeed $N(0, t] \sim \text{Poisson}(\lambda t)$.

Now we use the memoryless property of the exponential distribution to extend this to all intervals and to give the independent increments property.

Fix s, and suppose we condition on any outcome of the process on $[0, s]$. To be specific, condition on the event that

$$N_s = k, T_1 = t_1, T_2 = t_2, \dots, T_k = t_k.$$

Equivalently,

$$Y_1 = t_1, Y_2 = t_2 - t_1, \dots, Y_k = t_k - t_{k-1}, Y_{k+1} > s - t_k. \tag{7.3}$$

The memoryless property for Y_{k+1} tells us that conditional on $Y_{k+1} > s - t_k$, the distribution of $Y_{k+1} - (s - t_k)$ is again exponential with rate λ.

Combining this with the independence of the sequence Y_i, we have that conditional on (7.3), the sequence $Y_{k+1} - (s - t_k), Y_{k+2}, Y_{k+3}, \dots$ is i.i.d. with $\text{Exp}(\lambda)$ distribution.

But this means that, conditional on (7.3), the distribution of the process $N(s, s+u], u \geq 0$ is the same as the original distribution of the process $N_u, u \geq 0$.

So indeed, the property (iii) extends to all s. Further, the increment on $(s, t]$ is independent of the whole process on $(0, s]$, and applying this repeatedly we get independence of any set of increments on disjoint intervals. So Definition 7.2 holds as desired.

Poisson definition characterises the distribution of the process

With some work we could show the reverse implication using a direct calculation. Instead we appeal to a general (although rather subtle) property. The Poisson definition specifies the joint distribution of $N_{t_1}, N_{t_2}, \dots, N_{t_k}$ for any sequence $t_1, t_2, \dots, t_k$. It turns out that such "finite dimensional distributions", along with the assumption that the process is right-continuous, are enough to characterise completely the distribution of the entire process. We will not delve any further here into this fact from stochastic process theory. But it means that at most one process could satisfy Definition 7.2, and since we have shown that a process defined by Definition 7.1 does so, we have that Definition 7.2 implies 7.1 as desired.

7.2.1 The Poisson process as a limit of discrete-time processes

The calculation showing that (7.1) and (7.2) are the same is perhaps not very illuminating. The case $k = 0$ is easy and is illustrated for example in Example 7.4(a) below. To get more intuition for the relation between Poisson increments and exponential interarrivals, one can also think about a related discrete-time process.

Let us recall some facts from earlier in the course:

(1) If $X_n \sim \text{Binomial}(n, \lambda/n)$, then $X_n \xrightarrow{d} \text{Poisson}(\lambda)$ as $n \to \infty$. (See Example 2.9.)

(2) If $Y_n \sim \text{Geometric}(\lambda/n)$, then $Y_n/n \xrightarrow{d} \text{Exp}(\lambda)$ as $n \to \infty$. (See Example 2.3.)

Now consider a sequence of independent Bernoulli trials. In each trial (or time-slot), suppose we see a success with probability p and no event with probability $1 - p$. Then in any run of M trials, the total number of successes has $\text{Binomial}(M, p)$ distribution. Meanwhile the distances between consecutive successes are i.i.d. with $\text{Geometric}(p)$ distribution.

Now consider n large. Let $p = \lambda/n$, and rescale time by a factor of $1/n$, so that a time-interval of length t corresponds to a run of tn trials. Then the number of events in a time-interval of length t has $\text{Binomial}(tn, \lambda/n)$ distribution, which is approximately $\text{Poisson}(\lambda t)$, while the times between consecutive successes have Geometric(λ/n) distribution rescaled by $1/n$, which is approximately Exp(λ).

So indeed, as $n \to \infty$, we obtain a continuous-time process in which the interarrival times are independent exponentials, and the increments on disjoint intervals are independent Poisson random variables. So we can see this exponential/Poisson relationship in the Poisson process as a limit of the geometric/binomial relationship which is already familiar from sequences of independent trials.

7.2.2 A third definition *(non-examinable)*

Reflecting some of the ideas in the previous section, there is in fact a third natural definition of the Poisson process, which we include for completeness. This involves the independent increments property as in the case of Definition 7.2, but instead of specifying that increments have Poisson distribution, it specifies the behaviour of the increments on small time-intervals. Namely, the probability of seeing an event in a small interval should behave like λ multiplied by the length of the interval, and it should be very unlikely that two or more events occur within the interval:

Definition 7.3 (Defintion of Poisson process via infinitesimal increments)**.** *$N_t, t \geq 0$ is a Poisson process of rate λ if:*

(i) $N_0 = 0$.

(ii) If $(s_1, t_1), (s_2, t_2), \ldots, (s_k, t_k)$ are disjoint intervals in $\mathbb{R}_+$, then the increments $N(s_1, t_1]$, $N(s_2, t_2], \ldots, N(s_k, t_k]$ are independent.

(iii) The distribution of $N(s, s+h]$ is the same for all s, and as $h \to 0$,

$$\begin{aligned} \mathbb{P}(N(s, s+h] = 0) &= 1 - \lambda h + o(h) \\ \mathbb{P}(N(s, s+h] = 1) &= \lambda h + o(h) \\ \mathbb{P}(N(s, s+h] \geq 2) &= o(h). \end{aligned} \tag{7.4}$$

Note that any two of the conditions of (7.4) imply the third.

This kind of formulation is very natural when moving to the context of more general continuous-time Markov jump processes (in which the rate at which jumps occur may depend on the present state). The definition can again be shown to be equivalent to Definitions 7.1 and 7.2.

7.3 Thinning and superposition of Poisson processes

Theorem 7.1 (Superposition of Poisson processes)**.** *Let L_t and M_t be independent Poisson processes of rate λ and μ respectively. Let $N_t = L_t + M_t$. Then N_t is a Poisson process of rate $\lambda + \mu$.*

Proof. We work from the definition of a Poisson process in terms of independent Poisson increments for disjoint intervals. Clearly, $N_0 = L_0 + M_0 = 0$ for property (i), and also

N_t satisfies property (ii) (independent increments) since L_t and M_t both have independent increments and are independent of each other.

So we need to show property (iii). Since $L(s,t] \sim \text{Poisson}(\lambda t)$ and $M(s,t] \sim \text{Poisson}(\mu t)$ independently of each other, we have $N(s,t] \sim \text{Poisson}((\lambda+\mu)t)$ as required, by familiar properties of the Poisson distribution. □

Theorem 7.2 (Thinning of a Poisson process)**.** *Let N_t be a Poisson process of rate λ. "Mark" each point of the process with probability p, independently for different points. Let M_t be the counting process of the marked points. Then M_t is a Poisson process of rate $p\lambda$.*

Proof. Again we will work with the definition in terms of independent Poisson increments. Properties (i) and (ii) for M follow from the same properties for N.

Now consider any interval $(s,t]$. We have $N(s,t] \sim \text{Poisson}(\lambda(t-s))$, and conditional on $N(s,t] = n$, we have $M(s,t] \sim \text{Binomial}(n,p)$.

But if $N \sim \text{Poisson}(\mu)$, and, conditional on $N = n$, $M \sim \text{Binomial}(n,p)$, then in fact $M \sim \text{Poisson}(p\mu)$. This fact was proved in two different ways in the Prelims course. For example, it can be done using generating functions: let $M = X_1 + X_2 + \cdots + X_N$ where X_i are i.i.d. Bernoulli random variables; then $G_M(s) = G_N(G_X(s))$. Alternatively, by direct calculation:

$$\begin{aligned}
\mathbb{P}(M = k) &= \sum_n \mathbb{P}(M = k | N = n)\mathbb{P}(N = n) \\
&= \sum_{n \geq k} \frac{e^{-\mu}\mu^n}{n!} \binom{n}{k} p^k (1-p)^{n-k} \\
&\vdots \\
&= \frac{e^{-p\mu}(p\mu)^k}{k!}.
\end{aligned}$$

Hence indeed we have here that $M(s,t] \sim \text{Poisson}(p\lambda(t-s))$. So indeed property (iii) holds as desired, and M is a Poisson process of rate $p\lambda$. □

Remark 7.3. In fact, it is not too hard to prove something stronger. If L is the process of unmarked points, then L is a Poisson process of rate $(1-p)\lambda$, and the processes L and M are independent.

7.4 Poisson process examples

Example 7.4. A Geiger counter near a radioactive source detects particles at an average rate of 1 per 2 seconds. (a) What is the probability that there is no particle detected for 3 seconds after the detector is switched on? (b) What is the probability of detecting at least 3 particles in the first 4 seconds?
Solution: We model the process of detections as a Poisson process with rate $\lambda = 0.5$ (where the unit of time is 1 second).

For part (a), $\mathbb{P}(N_3 = 0) = e^{-3\lambda} = e^{-1.5}$, since N_3, the number of points up to time 3, has Poisson(3λ) distribution. Alternatively, we could calculate the same probability as $\mathbb{P}(T_1 > 3) = e^{-3\lambda}$ since T_1, the time of the first point of the process, has distribution Exp(λ).

For part (b), N_4 has Poisson distribution with mean $4\lambda = 2$. Then

$$\begin{aligned}\mathbb{P}(N_4 \geq 3) &= 1 - \mathbb{P}(N_4 = 0) - \mathbb{P}(N_4 = 1) - \mathbb{P}(N_4 = 2) \\ &= 1 - e^{-2} - 2e^{-2} - \frac{2^2 e^{-2}}{2!} \\ &= 1 - 5e^{-2}.\end{aligned}$$

Example 7.5. A call centre receives calls from existing customers at rate 1 per 20 seconds, and calls from potential new customers at rate 1 per 30 seconds. Assume that these form independent Poisson processes. (a) What is the distribution of the total number of calls in a given minute? (b) What is the probability that the next call to arrive is from a potential new customer? (c) Suppose each call from a potential new customer results in a contract with probability 1/4 independently. What is the distribution of the number of new contracts arising from calls in a given hour?

Solution: Let the unit of time be 1 minute, so that the Poisson processes in the question have rates 3 and 2.

(a) From Theorem 7.1, the combined process of all calls is a Poisson process of rate 5. The number of calls in a given minute has Poisson(5) distribution.

(b) From any given moment, the time until the next "existing" call, say U_1, is exponential with rate 3, and the time until the next "new" call, say V_1, is exponential with rate 2.

$$\begin{aligned}\mathbb{P}(U_1 < V_1) &= \int_{u=0}^{\infty} \int_{v=u}^{\infty} 3e^{-3u} \times 2e^{-2v} dv du \\ &= \int_{u=0}^{\infty} 3e^{-3u} e^{-2u} du \\ &= 3/(2+3) \\ &= 3/5.\end{aligned}$$

(In fact, it is not a coincidence that here the answer is the ratio of the rate of the "existing customer" process to the rate of the two processes combined. This fact follows from Remark 7.3; we can consider a single process of rate 5 and "mark" each point with probability 3/5, to arrive at two independent processes with rates 3 and 2. In particular, the probability that the first point is marked is then 3/5.)

(c) The process of calls resulting in contracts is a thinning of the process of calls from potential new customers. This gives us a new Poisson process of rate $1/4 \times 2 = 1/2$. So the total number of calls resulting in new contracts in a given time interval of length 60 has Poisson(30) distribution.

Example 7.6 (Genetic recombination model)**.** An illustration of genetic recombination is shown in the figure below. In most of our cells, we have two versions of each chromosome, one inherited from our mother and one from our father. Sex cells – sperm and ova – contain only one copy of each chromosome.

During *meiosis* – the process in which sperm and ova are created – the chromosomes are broken at certain random "crossover" or "recombination" points, to form new chromosomes

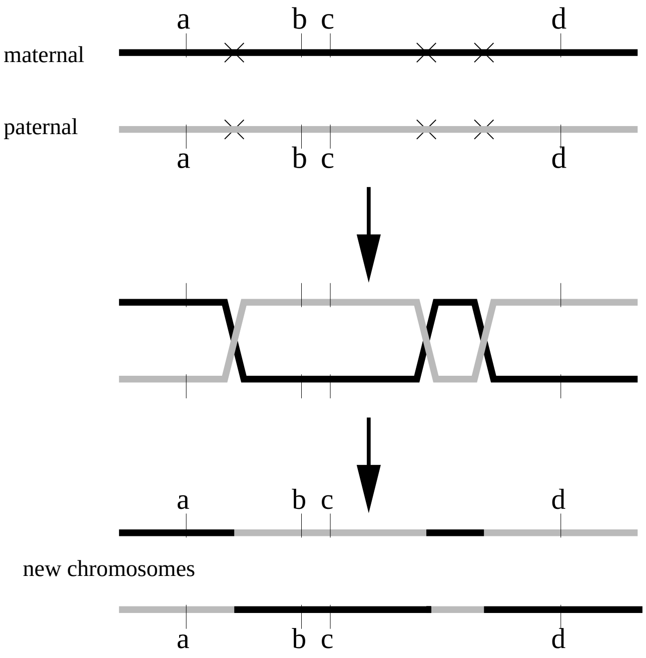


Figure 7.1: Recombination

out of pieces of the maternal and paternal chromosomes. The crossover points are shown as crosses in the top line of Figure 7.1.

Genes occur at particular positions along the chromosome. In early genetic research, biologists investigated the position of genes on chromosomes by looking at how likely the genes were to stay together, generation after generation. Genes on different chromosomes should be passed on independently. Genes that are close together on the same chromosome should almost always be passed on together, while genes that are on the same chromosome but further apart should be more likely than chance to be inherited together, but not certain.

In Figure 7.1, genes b, c and d stay together but a is separated from them.

As a simple model, we can imagine the chromosome as a continuous line, and model the recombination points as a Poisson process along it, of rate λ, say.

Consider two points a and b on the interval, representing the location of two genes. Let x be the distance between a and b. The probability of seeing no crossover at all between a and b is given by

$$\mathbb{P}\left(\text{no crossover in } (a,b)\right) = e^{-\lambda x}.$$

But what we really want to compute is the probability of seeing an *even number* of crossovers between a and b:

$$\begin{aligned} p = \mathbb{P}\left(\text{even number of crossovers in } (a,b)\right) &= \sum_{k=0}^{\infty} e^{-\lambda x} \frac{(\lambda x)^{2k}}{(2k)!} \\ &= e^{-\lambda x}\left(1 + \frac{(\lambda x)^2}{2!} + \frac{(\lambda x)^4}{4!} + \dots\right) \end{aligned}$$

$$
\begin{aligned}
&= e^{-\lambda x}\left(\frac{e^{\lambda x}+e^{-\lambda x}}{2}\right)\\
&= \frac{1+e^{-2\lambda x}}{2}.
\end{aligned}
$$

If we observe that a and b are inherited together with probability $p > 1/2$, we can invert the expression above to estimate the distance between them by

$$x = -\frac{1}{2\lambda}\log(2p-1).$$