Stochastic processes
Lecture Notes

Gourab Ray

Last updated: September 18, 2025

Disclaimer: These notes are a work in progress. There could be typos or small calculation mistakes. I would be grateful if you notify them to me.

Contents

1 Preliminaries

What is a Stochastic process? At it's basic form, it is simply a sequence of random variables $X_1, X_2, \ldots$, with some specified *joint distribution.* To emphasize, not only do we need to specify the distribution of each random variable $(X_i)_{i\geq 1}$, but also specify the joint distribution of any tuple $(X_{i_1}, \ldots, X_{i_k})$ where $i_j \geq 1, i_j \in \mathbb{N}$, $1 \leq j \leq k$. This is equivalent to specifying the joint distribution of $(X_1, X_2, \ldots, X_n)$ for any $n \in \mathbb{N}$.

For example, suppose we toss a fair coin independently forever. Let $X_i = 1$ if the ith toss produces a heads, and $X_i = 0$ otherwise. We can also let $Y_i = X_1$ for all $i \geq 1$. Clearly, $Y_i \sim$ Bernoulli$(1/2)$ for all $i \geq 1$ and $X_i \sim$ Bernoulli $(1/2)$ for all $i \geq 1$. However, $(X_1, X_2, \ldots)$ and $(Y_1, Y_2, \ldots)$ are clearly different as stochastic processes. Indeed, $(Y_1, Y_2, \ldots)$ either equals $(1, 1, \ldots)$ with probability $1/2$ or $(0, 0, \ldots)$ with probability $1/2$, while $(X_1, X_2, \ldots)$ can take many different combinations of 0s and 1s with positive probability.

The index set of a stochastic process can be more general. The type described above is a stochastic process **indexed by the natural numbers** $\mathbb{N}$. Sometimes, we need to deal with stochastic process $(X_t)_{t\geq 0}$ **indexed by** $\mathbb{R}_+ = [0, \infty)$. In this case, defining a stochastic process means specifying the joint distribution of $(X_{i_1}, \ldots, X_{i_k})$ where $i_j \geq 0, i_j \in \mathbb{R}$, $1 \leq j \leq k$.

Although we won't need it but a Stochastic process can be far more general, and can be indexed by an *arbitrary set* $\mathcal{I}$. For example, $\mathcal{I}$ can be the set of all smooth functions on $\mathbb{R}$ with compact support. The definition of a stochastic process do not change at all in this case. Although seems a bit contrived and too complicated for its own good, this is quite

natural, and comes up in higher level math quite often. We won't talk about it further in this notes.

One advantage of having $\mathcal{I} = \mathbb{N}$ or $\mathbb{R}_+$ is that there is a natural notion of *time* associated with the index set. Indeed, we can talk about the past of time i to be the set $\{j : j < i\}$ and the future as $\{j : j > i\}$.

1.1 Conditioning on events with positive probability

In a first course in Probability, you learnt about conditional probability of events. Recall the conditional probability of an event A given an event B with $\mathbb{P}(B) > 0$ is simply

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

Here is an example.

Example 1.1. Suppose we roll a dice twice.

- Suppose we are "given" the information that the first one is a 4. What is the probability that the next one is a 4? Does your answer change if we our 'given' information about the first roll is changed to any other number?
- What is the probability that the second roll is a 4?
- Suppose we are "given" the information that the second roll is at least 4. What is the probability that the second roll is a 6?

For the first item, let $A = \{$ first roll 4$\}$, $B = \{$ second roll 4$\}$, $C = \{$ second roll at least 4 $\}$. Therefore

$$\mathbb{P}(A) = \frac{1}{6}, \quad \mathbb{P}(B) = \frac{1}{6}, \quad \mathbb{P}(C) = \frac{3}{6} = \frac{1}{2}, \quad \mathbb{P}(A \cap B) = \frac{1}{36}, \quad \mathbb{P}(B \cap C) = \mathbb{P}(B) = \frac{1}{6}.$$

Thus we already answered the second item, it is the probability of the event B. For the first item, using the definition of conditional probability,

$$\mathbb{P}(\text{ second roll 4 } | \text{ first roll 4 }) = \mathbb{P}(B|A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)} = \frac{1/36}{1/6} = \frac{1}{6}$$

This answer is independent of the first roll being 4. No matter what the first roll is, the second roll will always have probability 1/6 of producing a 4.

$$\mathbb{P}(\text{second roll 4 } | \text{ second roll at least 4 }) = \mathbb{P}(B|C) = \frac{\mathbb{P}(C \cap B)}{\mathbb{P}(C)} = \frac{1/6}{1/2} = \frac{1}{3}$$

In Example 1.1, we conditioned on an event. But the first item is revealing something more: we can condition 'on whatever happened in the first roll' and still the probability of the second roll is 4 is unchanged. This conditioning is called 'conditioning on the information of the first roll'. The mathematically precise way of 'conditioning on the information of the first roll' is to condition on the random variable which is the output of the first roll. [1]

Suppose A is an event with $\mathbb{P}(A) > 0$. Then the conditional distribution of X conditioned on A is given by

$$F_{X;A}(t) := \mathbb{P}(X \leq t|A) = \frac{\mathbb{P}(X \leq t, A)}{\mathbb{P}(A)}, \qquad t \in \mathbb{R} \tag{1.1}$$

It is easy to check that $F_{X;A}(t)$ is indeed a cdf and hence corresponds to a probability distribution. We call this probability distribution colloquially as the **distribution of X conditioned on A.**

Sometimes, for practical reasons, it is easier to compute the complementary event:

$$1 - F_{X;A}(t) := \mathbb{P}(X > t|A) = \frac{\mathbb{P}(X > t, A)}{\mathbb{P}(A)}, \qquad t \in \mathbb{R}$$

Example 1.2. Exponential distribution has "no memory". This means that if someone tells you that $X > s$ then the conditional distribution of X conditioned on $X > s$ is the same as $s + X$. Let us see why.

$$\mathbb{P}(X > s + t|X > s) = \frac{\mathbb{P}(X > s + t)}{X > s} = \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} = e^{-\lambda t}$$

and

$$\mathbb{P}(s + X > s + t) = \mathbb{P}(X > t) = e^{-\lambda t}$$

which are the same.

In fact, Exponential distribution is the *only* continuous distribution supported on the positive real line having the memoryless property. In particular, if

$$\mathbb{P}(X > s + t|X > s) = \mathbb{P}(X > t) \text{ for all } s, t > 0$$

for some continuous random variable X with $\mathbb{P}(X > 0) = 1$, then $X \sim$ Exponential(λ) for some λ. To see this, we first claim

$$G(x) = G(1)^x. \text{ for all} x > 0. \tag{1.2}$$

where $G(x) = 1 - F_X(x)$. Let us now explain our claim. If $x = 2$, then using our condition for $s = t = 1$, we get

$$G(2) = G(1)^2.$$

[1] To be super precise, the 'information' is encoded on something called a sigma-algebra which is something we will try to avoid.

Similarly, for any integer n, we have

$$G(n) = G(n-1)G(1) \text{ and hence by induction, } G(n) = G(1)^n.$$

and similarly, for any integer n

$$G(1) = G((n-1)/n + 1/n) = G((n-1)/n)G(1/n) = \dots G(1/n)^n \implies G(1/n) = G(1)^{1/n}.$$

Next, for any rational number p/q with $p, q > 0$ and $\gcd(p, q) = 1$,

$$G(p/q) = G((p-1)/q)G(1/q), \text{ and hence by induction, } G(p/q) = G(1/q)^p = G(1)^{p/q}.$$

Thus Equation (1.2) is proved for all rationals. Now recall that G is right continuous since any cdf is right continuous. Thus for any $x \in \mathbb{R}$, take a sequence of rationals r_n converging to x from the right.

$$G(x) = \lim_{n\to\infty} G(r_n) = \lim_{n\to\infty} (G(1))^{r_n} = G(1)^x.$$

Now we conclude that that $X \sim \text{Exp}\,(-\ln(G(1)))$ simply by doing some gymnastics with exponentials and log and recalling that if $Z \sim$ Exponential λ then $\mathbb{P}(Z > t) = e^{-\lambda t}$.

Example 1.3. Suppose X_1, X_2 be i.i.d. Exponential (1) random variables. let us compute the distribution of X_1 conditioned on $A = \{X_1 > X_2\}$. It turns out, computing $\mathbb{P}(X > t|A)$ is easier here. Note $\mathbb{P}(X_1 > X_2) = 1/2$ by symmetry. Indeed, $\mathbb{P}(X_1 > X_2) + \mathbb{P}(X_2 > X_1) = 1$ and both probabilities are equal as X_1, X_2 are i.i.d. Hence, $\mathbb{P}(X_1 > X_2) = 1/2$. Here $\mathbb{P}(X_1 = X_2) = 0$ as the distributions are continuous, so we can simply ignore this term.

$$\begin{aligned}
\mathbb{P}(X_1 > t|A) &= \frac{\mathbb{P}(X_1 > t, X_1 > X_2)}{\mathbb{P}(X_1 > X_2)} \\
&= 2\int_t^\infty \int_0^{x_1} e^{-x_1} e^{-x_2} dx_2 dx_1 \\
&= 2\int_t^\infty e^{-x_1}(1 - e^{-x_1}) dx_1 \\
&= 2e^{-t} - e^{-2t}
\end{aligned}$$

Thus the cdf is

$$\mathbb{P}(X_1 \le t|A) = 1 - 2e^{-t} + e^{-2t} = (1 - e^{-t})^2.$$

Exercise 1.4. *Show that the conditional distribution of X_1 in Example 1.3 is the same as that of* $\max\{X_1, X_2\}$ *(without conditioning).*

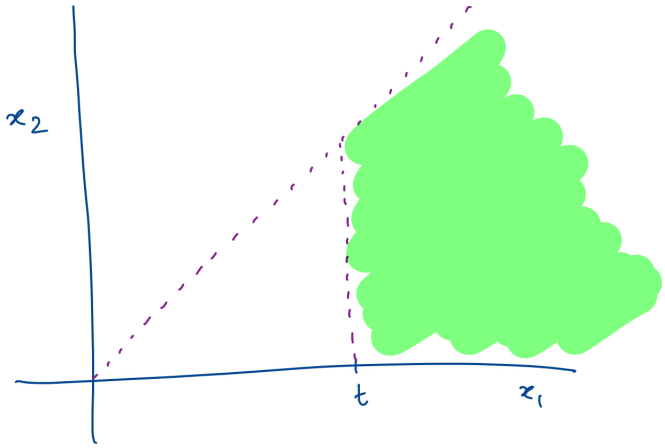


Figure 1: Example 1.3. We need to compute the integral in the green region.

1.2 Discrete case.

We now define the **conditional probability mass function (Conditional pmf)** which is what we need to compute if are random variables in question are discrete. Recall that we denote the probability mass function (pmf) of a random variable X by

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

Definition 1.5 (Conditional pmf)**.** *Let X and Y be random variables and $y \in \mathbb{R}$ is such that $\mathbb{P}(Y = y) > 0$. Then the conditional pmf of X given $Y = y$ is given by*

$$p_{X|Y}(x|y) = \mathbb{P}(X = x|Y = y) = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)} = \frac{p(x, y)}{p_Y(y)}$$

Here $p(x, y)$ is the notation for the joint pmf of X and Y.

This leads to a natural definition of **conditional distribution function** (conditional cdf)

Definition 1.6 (conditional cdf)**.** *Let X and Y be random variables and $y \in \mathbb{R}$ is such that $\mathbb{P}(Y = y) > 0$. Then the conditional cdf of X given $Y = y$ is given by*

$$F_{X|Y}(x|y) = \mathbb{P}(X \leq x|Y = y) = \frac{\mathbb{P}(X \leq x, Y = y)}{\mathbb{P}(Y = y)} = \frac{\sum_{a \leq x} p(a, y)}{p_Y(y)}$$

Remark 1.7. Note that the conditional distribution of X given $Y = y$ does NOT make sense (i.e. ill defined) if $\mathbb{P}(Y = y) = 0$. However if $\mathbb{P}(Y = y) > 0$, we can simply revert back to the old definition of conditional distribution when conditioning on an event Equation (1.1), which is what we did above

Let us do a simple example and compute the conditional pmf of Y_1 given $Y_2 = 1$ in each of the following.

$Y_1 \downarrow Y_2 \rightarrow$	0	1	Marginal(Y_1)
0	1/4	1/4	1/2
1	1/4	1/4	1/2
Marginal (Y_2)	1/2	1/2	1

Here

$$p_{Y_1|Y_2}(0|1) = \frac{1/4}{1/2} = 1/2, \qquad p_{Y_1|Y_2}(1|1) = \frac{1/4}{1/2} = 1/2.$$

Actually here we will realize later that the conditional distribution of Y_1 given $Y_2 = 0$ or $Y_2 = 1$ is the same as the distribution of Y_1. We will see later that this is equivalent to saying that Y_1 and Y_2 are *independent*.

$Y_1 \downarrow Y_2 \rightarrow$	0	1	Marginal(Y_1)
0	1/2	0	1/2
1	0	1/2	1/2
Marginal (Y_2)	1/2	1/2	1

Here

$$p_{Y_1|Y_2}(0|1) = \frac{0}{1/2} = 0, \qquad p_{Y_1|Y_2}(1|1) = \frac{1/2}{1/2} = 1.$$

Exercise 1.8. *Calculate the conditional distribution of Y_1 given $Y_2 = 0$ in each of the above two examples.*

Example 1.9. Suppose X_1, X_2 be i.i.d. Geom(p). What is the conditional distribution of X_1 given $X_1 + X_2 = n$?

Solution. Intuitively, since Geom(p) is the number of tosses needed until we get a head, X_1 should takes each value in $\{1, 2, \ldots, n-1\}$ with uniform probability. Formally a calculation gives for $i = 1, 2, \ldots, n-1$

$$\begin{aligned}
\mathbb{P}(X_1 = i | X_1 + X_2 = n) &= \frac{\mathbb{P}(X_1 = i, X_2 = n - i)}{\mathbb{P}(X_1 + X_2 = n)} \\
&= \frac{\mathbb{P}(X_1 = i)\mathbb{P}(X_2 = n - i)}{\sum_{i=1}^{n-1} \mathbb{P}(X_1 = i)\mathbb{P}(X_2 = n - i)} \\
&= \frac{(1-p)^{i-1}p(1-p)^{n-i-1}p}{\sum_{i=1}^{n-1}(1-p)^{i-1}p(1-p)^{n-i-1}p} \\
&= \frac{1}{n-1}.
\end{aligned}$$

Thus the conditional distribution is Uniform in the set $\{1, \ldots, n\}$.

Example 1.10. $X_i \sim$ Poisson(λ_i). Calculate conditional pmf of X_1 given $X_1 + X_2 = n$.

Solution. A calculation gives

$$\mathbb{P}(X_1 = k | X_1 + X_2 = n) = \binom{n}{k} \left(\frac{\lambda_1}{\lambda_1 + \lambda_2}\right)^k \left(\frac{\lambda_2}{\lambda_1 + \lambda_2}\right)^{n-k}$$

i.e. the conditional distribution of X_1 given $X_1 + X_2 = n$ is given by Binomial $(n, \frac{\lambda_1}{\lambda_1+\lambda_2})$. We leave this calculation as an exercise. (If you don't recall the pmf of Poisson (λ), look up the formula sheet at the end of the notes.)

Proposition 1.11. *Suppose (X, Y) is discrete. Then X and Y are independent if and only if*

$$p_{X|Y}(x|y) = p_X(x) \text{ for all } x, y \text{ with } p_Y(y) > 0.$$

Proof. This is a simple consequence of the fact that X, Y are independent if and only if $p_{X,Y}(x, y) = p_X(x)p_Y(y)$ for all x, y. □

Exercise 1.12. *Suppose X is a distrete random variable taking values in $\mathbb{N}$. Suppose it is memoryless, that is,*

$$\mathbb{P}(X > m + n | X > m) = \mathbb{P}(X > n).$$

for all $m, n \in \mathbb{N}$. Show that this must be a geometric random variable.

1.3 Continuous case.

As you might suspect, we need to define the conditional density function.

Definition 1.13. *Let X, Y be random variables with joint density $f(x, y)$ and marginal densities $f_X(x)$ and $f_Y(y)$ respectively. Let y be such that $f_Y(y) > 0$. Then the conditional probability density function of X given $Y = y$ is given by*

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

The conditional cdf of X given $Y = y$ is given by

$$F_{X|Y(x|y)} = \int_{-\infty}^{y} f_{X|Y}(u|y)du = \int_{-\infty}^{y} \frac{f(u, y)}{f_Y(y)}du.$$

Example 1.14. Suppose that the joint density of X and Y is given by

$$\frac{e^{-x/y}e^{-y}}{y}, \qquad 0 < x < \infty, \qquad 0 < y < \infty.$$

Calculate the conditional density of X given $Y = y$ for some $y > 0$.

To do this, we first find the marginal density of Y:

$$\int_0^\infty \frac{e^{-x/y}e^{-y}}{y}dx = e^{-y}.$$

One way is to calculate the integral directly, and another simpler (probabilistic!) way is to observe that $\frac{1}{y}e^{-x/y}$ is the density of $\exp(1/y)$, and hence $\int_0^\infty e^{-x/y} = y$. Thus the conditional density is

$$f_{X|Y=y}(x|y) = \frac{\frac{e^{-x/y}e^{-y}}{y}}{e^{-y}} = \frac{e^{-x/y}}{y}, \qquad 0 < x < \infty. \tag{1.3}$$

This is simply the density of an Exponential $(1/y)$ random variable.

Example 1.15. Let X, Y be i.i.d. Uniform $[0, 1]$. Let $U = \max(X, Y)$ and $V = \min(X, Y)$. Compute the conditional density of V given $U = u$.

Let us first compute the joint density of U, V. You can first write down the joint density of (X, Y) and then employ the Jacobian for the transformation $(x, y) \mapsto (\max(x, y), \min(x, y))$. But here is a different way. For every $u \geq v$,

$$\mathbb{P}(V > v, U \leq u) = \mathbb{P}(X \in (v, u], Y \in (v, u]) = (u - v)^2.$$

Also $\mathbb{P}(U \leq u) = \mathbb{P}(X \leq u, Y \leq u) = u^2$. Therefore for any $u \geq v$

$$\mathbb{P}(V \leq v, U \leq u) = \mathbb{P}(U \leq u) - \mathbb{P}(V > v, U \leq u) = u^2 - (u - v)^2 = v(2u - v).$$

If $u < v$, then $\mathbb{P}(V > v, U \leq u) = 0$. Thus

$$\mathbb{P}(V \leq v, U \leq u) = \mathbb{P}(U \leq u) = u^2.$$

Thus the joint cdf is

$$F_{U,V}(u, v) = \begin{cases} v(2u - v) \text{ if } u \geq v \\ u^2 \text{ if } u < v. \end{cases}$$

The joint pdf is obtained by computing the partial derivative $\frac{\partial^2}{\partial u \partial v} F_{U,V}$ which is

$$f_{U,V}(u, v) = \begin{cases} 2 \text{ if } 1 \geq u \geq v \geq 0 \\ 0 \text{ otherwise .} \end{cases}$$

The marginal density of U is given by

$$f_U(u) = \begin{cases} 2u \text{ if } u \in (0, 1) \\ 0 \text{ otherwise} \end{cases}$$

Thus the conditional density, for any $u \in (0, 1)$ is given by

$$f_{V|U}(v|u) \begin{cases} \frac{2}{2u} = \frac{1}{u} \text{ if } 0 \leq v \leq u \\ 0 \text{ otherwise} \end{cases}$$

In words, conditioned on $U = u$, $V \sim$Uniform $[0, u]$

$$f_{V|U}(v|u) = \frac{1}{u}\mathbf{1}_{0 \leq v \leq u}.$$

Proposition 1.16. *Suppose (X, Y) is jointly continuous. Then X and Y are independent if and only if*

$$f_{X|Y}(x|y) = f_X(x) \text{ for all } x, y \text{ with } f_Y(y) > 0.$$

Proof. This is a simple consequence of the fact that X, Y are independent if and only if $f_{X,Y}(x, y) = f_X(x)f_Y(y)$ for all x, y. □

Let us now introduce an important terminology. Sometimes we will talk about distributions of random variables where the parameters are also random. For example, we will talk about $X \sim$Exponential(Λ) where Λ is itself a random variable, say with some pdf f_Λ. What we mean here is that the conditional distribution of X given $\Lambda = \lambda$ is Exponential λ. So the joint density of (X, Λ) is given by

$$f_{X,\Lambda} = \lambda e^{-\lambda x} f_\Lambda(\lambda), \qquad x > 0, \lambda \in \mathbb{R}.$$

and 0 otherwise. The distribution of X is then represented by the pdf

$$f_X(x) = \int_0^\infty \lambda e^{-\lambda x} f_\Lambda(\lambda) d\lambda = \mathbb{E}(\Lambda e^{-\Lambda x})$$

for $x > 0$ and 0 otherwise.

Exercise 1.17. *Check that if $X \sim$Exponential $(1/Y)$ and $Y \sim$Exponential(1), then the joint density of (X, Y) is that given in (1.3).*

Exercise 1.18. *Let $X \sim N(\Lambda, \Sigma)$ where (Λ, Σ) are i.i.d. Exponential (1). Calculate the joint density of (X, Λ, Σ).*

Sometimes we might need to combine discrete and continuous random variables. The following example illustrates this.

Example 1.19. Let $X \sim$ Bernoulli (U) where $U \sim$Uniform $[0, 1]$, then the conditional distribution of X given $U = u$ is given by Bernoulli (u). However, we cannot write the joint density nor write the joint pmf in this case as U is continuous and X is discrete. Nevertheless, we can compute probabilities as follows. For any $z \in (0, 1)$,

$$\begin{aligned}\mathbb{P}(X = 1, U \le z) &= \int_{u=-\infty}^{z} \mathbb{P}(X = 1|U = u) f_U(u) du \\ &= \int_0^z u \cdot 1 du \\ &= \frac{z^2}{2}.\end{aligned}$$

If $z < 0$, then $\int_{u=-\infty}^{z} \mathbb{P}(X = 1|U = u) f_U(u) du = 0$ since $f_U(u) = 0$ whenever $u < 0$. If $z > 1$,

$$\int_{u=-\infty}^{z} \mathbb{P}(X = 1|U = u) f_U(u) du = \int_0^1 u du = \frac{1}{2}.$$

This makes sense as if $z > 1$ then $\{U \le z\}$ always happens, and thus $\mathbb{P}(X = 1, U \le z) = \mathbb{P}(X = 1)$. This also shows that marginally, the unconditional distributiopn of X is simply Bernoulli $(1/2)$.

Exercise 1.20. *In the above example, compute* $\mathbb{P}(X = 0, U \leq z)$ *and double check that*

$$\mathbb{P}(X = 1, U \leq z) + \mathbb{P}(X = 0, U \leq z) = \mathbb{P}(U \leq z).$$

We will deal with this in more depth in Section 1.7.1.

Example 1.21. (Using conditional density to our advantage) This is a more complicated example. Suppose $Z \sim N(0,1)$ and Y has a chi-squared distribution with n-degrees of freedom (denoted χ^2_n distribution sometimes) and is independent of Z. In other words, Y has the density [2]

$$f_Y(y) = \frac{e^{-y/2} y^{n/2-1}}{2^{n/2}\Gamma(n/2)}, \qquad y > 0.$$

Let

$$T = \frac{Z}{\sqrt{Y/n}}.$$

T is sometimes called a t-random variable with n-degrees of freedom. Calculate the density function of T.

Notice that the conditional distribution of T given $Y = y$ for some $y > 0$ is given by the distribution of $\frac{Z}{\sqrt{y/n}} = \sqrt{n/y}Z \sim N(0, n/y)$. Hence the conditional density function is

$$f_{T|Y}(t|y) = \frac{1}{\sqrt{2\pi n/y}} e^{-\frac{yt^2}{2n}}$$

Therefore the joint density of Z and Y is given by $f_{T|Y}(z|y)f_Y(y)$ and consequently the marginal density is given by

$$\int_0^\infty f_{T|Y}(t|y) f_Y(y) dy = \int_0^\infty \frac{1}{\sqrt{2\pi n/y}} e^{-\frac{yt^2}{2n}} \frac{e^{-y/2} y^{n/2-1}}{2^{n/2}\Gamma(n/2)} dy$$

We simplify

$$\int_0^\infty \frac{1}{\sqrt{2\pi n/y}} e^{-\frac{yt^2}{2n}} \frac{e^{-y/2} y^{n/2-1}}{2^{n/2}\Gamma(n/2)} dy = \int_0^\infty \frac{1}{\sqrt{2\pi n} 2^{n/2}\Gamma(n/2)} e^{-cy} y^{n/2+1/2-1} dy.$$

with $c = \frac{t^2}{2n} + \frac{1}{2}$. The integrand is the density of a Γ random variable (without the constant in front) and hence using the formula for the pdf of a Gamma random variable (see table 2.2 in book), we see that the density of t-distribution with parameter n is given by

$$f_T(t) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{\pi n}\Gamma(\frac{n}{2})} \left(1 + \frac{t^2}{n}\right)^{-(n+1)/2}, \qquad t \in \mathbb{R}.$$

[2] You might also recognize this as a $\Gamma(n/2, \frac{1}{2})$ random variable. Also one can show that if

$$Z = X_1^2 + X_2^2 + \ldots + X_n^2$$

where $X_1, \ldots, X_n$ are i.i.d. $N(0,1)$ then Z follows a chi-squared distibution with n degrees of freedom.

1.4 Conditional expectation

The definition of the conditional expectation is the same as the unconditional one, except we need to replace the pdf / pmf by conditional pdf/pmf

Definition 1.22 (Conditional expectation)**.** *Let X and Y be discrete random variables and $y \in \mathbb{R}$ is such that $\mathbb{P}(Y = y) > 0$. The conditional expectation of X given $Y = y$ is given by*

$$\mathbb{E}(X|Y = y) = \sum_x x\mathbb{P}(X = x|Y = y) = \sum_x x p_{X|Y}(x|y).$$

Also if (X, Y) are jointly continuous with $f_Y(y) > 0$ then

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

Example 1.23. We could directly compute the summation, or alternately, be a bit more lazy, and simply observe using In Example 1.10,

$$\mathbb{E}(X_1|X_1 + X_2 = n) = n\frac{\lambda_1}{\lambda_1 + \lambda_2}.$$

Indeed, Example 1.10 exactly tells us that the conditional pmf of X_1 given $X_1 + X_2 = n$ is that of a $\text{Bin}(n, \frac{\lambda_1}{\lambda_1+\lambda_2})$ random variable. We know that the expectation of a Binomial (n, p) random variable is np.

Example 1.24. To illustrate how to think of conditional expectation as a random variable, consider the following problem. Can you find a random variable X, Y such that $\mathbb{E}(X) = \infty$, $\mathbb{E}(Y) = \infty$, but $\mathbb{E}(X|Y) < \infty$?

Actually this is very easy. Take Y to be any random variable with infinite expectation, for example $Y = C/i^2, i = 1, 2, \ldots$ and $C = (\sum_{i=1}^{\infty} \frac{1}{i^2})^{-1}$. Now take $X = Y$. Of course, $\mathbb{E}(X|Y) = X < \infty$ almost surely.

Example 1.25. Suppose $X_1, X_2, \ldots$ be i.i.d. with finite mean and assume $S_n = X_1 + \ldots + X_n$. What is

$$\mathbb{E}(X_1|S_n = a)?$$

Observe that $\mathbb{E}(X_1|S_n) = \mathbb{E}(X_2|S_n) = \cdots = \mathbb{E}(X_n|S_n)$ by symmetry. Also,

$$\mathbb{E}(X_1 + \ldots X_n|S_n) = S_n.$$

Thus $\mathbb{E}(X_1|S_n) = S_n/n$.

Example 1.26. Suppose X_1, X_2 are i.i.d. with $\mathbb{E}(X_1) < \infty$. Suppose in a lake there are X_1 salmon and X_2 halibut fishes (and no other fish). Pick a fish uniformly at random. What is the chance that you pick a salmon?

Solution. By symmetry, the answer should be 1/2. Let us carefully verify it by doing the calculation.

$$\begin{aligned}\mathbb{P}(\text{pick Salmon}) &= \sum_{i,j=1}^{\infty} \mathbb{P}(\text{pick Salmon}|X_1 = i, X_2 = j)\mathbb{P}(X_1 = i, X_2 = j) \\ &= \frac{i}{i+j}\mathbb{P}(X_1 = i, X_2 = j) \\ &= \mathbb{E}(\frac{X_1}{X_1 + X_2}).\end{aligned}$$

Since the distribution is not given, there is no way to calculate this directly. However, we can use the trick in example 1.25 since X_1, X_2 are i.i.d. Using that we get

$$\mathbb{E}(\frac{X_1}{X_1 + X_2}) = \frac{1}{2}.$$

as expected.

Example 1.27. In example 1.14, the conditional expectation

$$\mathbb{E}(X|Y = y) = y$$

because the conditional pdf is that of an exponential with parameter $1/y$ and we know that expectation of an $\exp(\lambda)$ random variable is $1/\lambda$.

Just like an unconditional expectation, the conditional expectation of a function $g(X)$ of a random variable can be computed as follows.

Proposition 1.28. *Let X and Y be discrete random variables and $y \in \mathbb{R}$ is such that $\mathbb{P}(Y = y) > 0$. Then*

$$\mathbb{E}(g(X)|Y = y) = \sum_x g(x)\mathbb{P}(X = x|Y = y) = \sum_x g(x)p_{X|Y}(x|y).$$

Also if (X, Y) are jointly continuous with $f_Y(y) > 0$ then

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

Exercise 1.29. *In example 1.14, calculate*

$$\mathbb{E}(e^X|Y = y).$$

Hint: Use mgf of exponential random variable.

1.5 Expectation by conditioning.

Sometimes it is useful to think of $\mathbb{E}(X|Y)$ as a random variable (that is, when the value of the conditioned random variable Y is not fixed). The following identity is super useful.

Theorem 1.30. *For any function $g : \mathbb{R} \to \mathbb{R}$, such that $\mathbb{E}(|g(X)|) < \infty$*

$$\mathbb{E}(\mathbb{E}(g(X)|Y)) = \mathbb{E}(X).$$

Proof. Let $\mathcal{X}, \mathcal{Y}$ be the at most countable sets in which X, Y takes values. Write the joint probability mass function as $p_{X,Y}(x, y) = \mathbb{P}(X = x, Y = y)$. The marginal of Y is $p_Y(y) = \sum_{x \in \mathcal{X}} p_{X,Y}(x, y)$, and the conditional pmf of X given $Y = y$ (for those y with $p_Y(y) > 0$) is

$$p_{X|Y}(x \mid y) = \mathbb{P}(X = x \mid Y = y) = \frac{p_{X,Y}(x, y)}{p_Y(y)}.$$

By definition of conditional expectation in the discrete case,

$$\mathbb{E}[\, g(X) \mid Y = y \,] \;=\; \sum_{x \in \mathcal{X}} g(x)\, p_{X|Y}(x \mid y).$$

Therefore the iterated expectation is

$$\begin{aligned}
\mathbb{E}\big(\mathbb{E}[\, g(X) \mid Y \,]\big) &= \sum_{y \in \mathcal{Y}} \mathbb{E}[\, g(X) \mid Y = y \,]\, p_Y(y) \\
&= \sum_{y \in \mathcal{Y}} \left(\sum_{x \in \mathcal{X}} g(x)\, p_{X|Y}(x \mid y) \right) p_Y(y) \\
&= \sum_{y \in \mathcal{Y}} \sum_{x \in \mathcal{X}} g(x)\, \frac{p_{X,Y}(x, y)}{p_Y(y)}\, p_Y(y) \\
&= \sum_{y \in \mathcal{Y}} \sum_{x \in \mathcal{X}} g(x)\, p_{X,Y}(x, y).
\end{aligned}$$

Interchanging the order of summation [3] gives

$$\mathbb{E}\big(\mathbb{E}[\, g(X) \mid Y \,]\big) = \sum_{x \in \mathcal{X}} g(x) \left(\sum_{y \in \mathcal{Y}} p_{X,Y}(x, y) \right) = \sum_{x \in \mathcal{X}} g(x)\, p_X(x) = \mathbb{E}[\, g(X) \,].$$

This proves the claim. □

Example 1.31. in Example 1.14, we can think of $\mathbb{E}(X|Y) = Y$ and hence we get

$$\mathbb{E}(X) = \mathbb{E}(\mathbb{E}(X|Y)) = \mathbb{E}(Y).$$

[3] this is justified by absolute summability since $\mathbb{E}|g(X)| < \infty$, which comes from a theorem called Fubini's theorem. If this is unfamiliar, don't bother about it, and just accept that it can be done.

Let us illustrate how this can be useful.

Example 1.32. Recall the geometric random variable: $X \sim Geom(p)$ if

$$\mathbb{P}(X = k) = (1-p)^{k-1}p, \qquad k \geq 1$$

In other words, we toss a coin with $\mathbb{P}(\text{head}) = p$ until we get a heads and count the number of tosses needed. We can simply compute the expectation of X to be $1/p$ by direct calculation [4]:

$$\sum_{k=1}^{\infty} k(1-p)^{k-1}p = \frac{1}{p}.$$

Let us do this in a different way using the coin tossing experiment way of thinking about Geometric random variable. Let $Y = 1_{\text{first flip heads}}$. That is $Y = 1$ if the first flip is heads and tails otherwise. Then

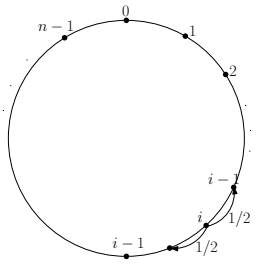
$$\begin{aligned}
\mathbb{E}(X) &= \mathbb{E}(\mathbb{E}(X|Y)) \\
&= \mathbb{E}(X|Y=1)\mathbb{P}(Y=1) + \mathbb{E}(X|Y=0)\mathbb{P}(Y=0) \\
&= 1 \cdot p + \mathbb{E}[(1+Z)|Y=0](1-p)
\end{aligned}$$

where Z is the number of tosses needed to get a head where we count from toss number 2 onwards. Note that Z is first of all independent of Y since it does not depend on toss number 1. Secondly, $Z = X$ in distribution. Therefore,

$$\mathbb{E}(X) = p + \mathbb{E}[(1+Z)](1-p) = 1 + (1-p)\mathbb{E}(X).$$

which simplifies to $\mathbb{E}(X) = 1/p$.

Exercise 1.33. *Suggested read: Example 3.15 (both editions) which computes the expected number of trials needed until we get k consecutive heads.*



Example 1.34 (Gamblers ruin)**.** Consider a *simple random walk on a cycle* in which an observer jumps to one of the neighbours in the following graph with equal probability. Starting from i, find the expected number of steps taken by the observer to hit 0.

[4] It is a good exercise to redo this summation if you forgot how to do it.

Let N_i be the required number of steps needed when the observer starts from i and let $m_i = \mathbb{E}(N_i)$. Note that $m_0 = 0$. Let ξ be $+1$ or -1 according to the event that the walker moves forward (or clockwise) as compared to backwards (or anti-clockwise). Also for each $i = 1, \ldots, n-1$

$$\begin{aligned}\mathbb{E}(N_i) &= \mathbb{E}(N_i|\xi = 1)\mathbb{P}(\xi = 1) + \mathbb{E}(N_i|\xi = -1)\mathbb{P}(\xi = -1) \\ &= (1 + \mathbb{E}(N_{i+1}))\frac{1}{2} + (1 + \mathbb{E}(N_{i-1}))\frac{1}{2} \\ &= 1 + \frac{1}{2}(m_{i+1} + m_{i-1})\end{aligned}$$

We need to solve this equation with $m_0 = 0$. Solving this, we see that

$$m_i = i(n - i)$$

Exercise 1.35. *Check the above formula. You may use induction.*

Suggested read: Example 3.16 from both editions. Can you link Example 3.16 with this example? This problem is called *Gambler's ruin* for this reason.

1.6 Variance by conditioning

We start with the definition of conditional variance.

Definition 1.36.

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

Again we can think of $\mathrm{Var}(X|Y)$ as a random variable with the randomness coming from Y.

Exercise 1.37. *What is* $\mathrm{Var}(X|X)$*?*

The formula for computing variance by conditioning is slightly complicated. Suppose we want to find the variance of X and actually conditioning by another random variable Y is useful.

Proposition 1.38 (Law of total variance)**.**

$$\mathrm{Var}(X) = \mathbb{E}(\mathrm{Var}(X|Y)) + \mathrm{Var}(E(X|Y)).$$

Proof.

$$\begin{aligned}\mathbb{E}(\mathrm{Var}(X|Y)) &= \mathbb{E}(\mathbb{E}(X^2|Y)) - (\mathbb{E}(X|Y)))^2 \\ &= \mathbb{E}(X^2) - \mathbb{E}((\mathbb{E}(X|Y))^2).\end{aligned}$$

Also

$$\begin{aligned}\mathrm{Var}(\mathbb{E}(X|Y)) &= \mathbb{E}((E(X|Y))^2) - (\mathbb{E}(\mathbb{E}(X|Y)))^2 \\ &= \mathbb{E}((\mathbb{E}(X|Y))^2) - (\mathbb{E}(X))^2.\end{aligned}$$

Adding them, $\mathbb{E}((\mathbb{E}(X|Y))^2)$ cancels and we get $\mathrm{Var}(X)$ on the right hand side, concluding the proof. □

Example 1.39. Let $X_1, X_2, \ldots,$ be i.i.d. with mean μ and Variance σ^2. Let $S_n = X_1 + \ldots + X_n$. Suppose N is a positive integer valued random variable independent of everything else. Then find $\mathrm{Var}(S_N)$.

Note that conditioning on N is a good idea. Thus we can use the formula

$$\mathrm{Var}(S_N) = \mathbb{E}(\mathrm{Var}(S_N|N)) + \mathrm{Var}(\mathbb{E}(S_N|N)).$$

Note

$$\mathrm{Var}(S_N|N=n) = \mathrm{Var}(S_n|N=n) \overset{\text{independence}}{=} \mathrm{Var}(S_n) = n\sigma^2.$$

Thus

$$\mathbb{E}(\mathrm{Var}(S_N|N)) = \mathbb{E}(N\sigma^2) = \sigma^2\mathbb{E}(N).$$

Also

$$\mathbb{E}(S_N|N=n) = \mathbb{E}(S_n|N=n) \overset{\text{independence}}{=} \mathbb{E}(S_n) = n\mu,$$

which means

$$\mathbb{E}(S_N|N) = N\mu.$$

Thus

$$\mathrm{Var}(S_N) = \mathbb{E}(\mathrm{Var}(S_N|N)) + \mathrm{Var}(\mathbb{E}(S_N|N)) = \mathbb{E}(N\sigma^2) + \mathrm{Var}(N\mu) = \sigma^2\mathbb{E}(N) + \mu^2\,\mathrm{Var}(N).$$

Remark 1.40. If $N \sim$ Poisson (λ) then S_N is called a compound Poisson random variable. Here is a practical example. Suppose the number of accidents in Victoria in a month is a Poisson random variable and an insurance company loses a random amount X of money for each accident which are i.i.d. Then the total loss of the company in a month will be given by a compound Poisson random variable.

1.7 Miscelleneous examples

Example 1.41. (Poisson thinning) Suppose a random number of points in a square are marked black. Suppose that the number of black points is Poisson(λ). Now suppose for each such mark, we independently color them red with probability p or blue with probability $(1-p)$. What is the distribution of the number of red marks in the square?

Suppose R is the number of red marks and B the number of blue marks. Then $R + B \sim$Poisson (λ). Also conditioned on $R + B = n$, R is distributed as a Bin(n, p) random variable. Thus

Now note

$$\begin{aligned}\mathbb{P}(R=r) &= \sum_{n=r}^{\infty} \mathbb{P}(R=r|R+B=n)\mathbb{P}(R+B=n) \\ &= \sum_{n=r}^{\infty} \binom{n}{r} p^r (1-p)^{n-r} \frac{e^{-\lambda}(\lambda)^n}{n!} \\ &= \frac{(p/1-p)^r}{r!} e^{-\lambda} \sum_{n=r}^{\infty} ((1-p)\lambda)^n \frac{1}{(n-r)!} \\ &= \frac{(p/1-p)^r}{r!} e^{-\lambda} e^{(1-p)\lambda} ((1-p)\lambda)^r \\ &= e^{-p\lambda} \frac{(p\lambda)^r}{r!}.\end{aligned}$$

Thus $R \sim$ Poisson $p(\lambda)$. Thus coloring lowers the mean (sometimes called the intensity) of the Poisson variables by a factor of p.

Exercise 1.42. *Show that R and B are independent and $R \sim Poisson(\lambda p)$ and $B \sim Poisson(\lambda(1-p))$.*

Suggested read: Example 3.24 (Ed 12) / Example 3.23 (Ed 11).

Example 1.43 (Order statistic)**.** Let X and Y be independent Uniform$(0,1)$ random variables. Define

$$Z := \min(X,Y), \qquad M := \max(X,Y).$$

We will compute the conditional expectation $\mathbb{E}[Z \mid M]$.

Using the conditional density deduced in Example 1.15,

$$\mathbb{E}[Z \mid M=m] = \int_0^m u\, f_{Z|M}(u \mid m)\, du = \int_0^m u \cdot \frac{1}{m}\, du = \frac{1}{m} \cdot \frac{m^2}{2} = \frac{m}{2}.$$

Therefore, as a conditional expectation (i.e. as a random variable),

$$\mathbb{E}[Z \mid M=m] = \frac{M}{2}.$$

1.7.1 Uniform Prior, Polya Urns and Bose Einstein statistics

Uniform Prior. Let us continue the example in Example 1.19. Suppose $U \sim Unif[0,1]$ and conditioned on U, $(\xi_1, \ldots, \xi_n)$ are i.i.d. Bernoulli(U). Then first of all

$$\mathbb{P}(\xi_n = 1) = \int_0^1 \mathbb{P}(\xi_n = 1|U=u)du = \int_0^1 u du = \frac{1}{2}. \tag{1.4}$$

Now suppose we want to compute the joint law of $(\xi_1, \ldots, \xi_n)$. Note that $(\xi_1, \ldots, \xi_n)$ is a vector of 0s and 1s. Thus for any vector $(\varepsilon_1, \ldots, \varepsilon_n)$ where each ε_i is either a 0 or a 1 and $\sum_{i=1}^n \varepsilon_i = k$

$$\begin{aligned}
\mathbb{P}(\xi_1 = \varepsilon_1, \ldots, \xi_n = \varepsilon_n) &= \mathbb{E}\left(\mathbb{P}(\xi_1 = \varepsilon_1, \ldots, \xi_n = \varepsilon_n | U)\right) \\
&= \mathbb{E}(U^k (1-U)^{n-k}) \\
&= \int_0^1 u^k (1-u)^{n-k} \\
&= \frac{k!(n-k)!}{(n+1)!}
\end{aligned} \tag{1.5}$$

where the second equality follows from the fact that whenever $\varepsilon_i = 1$ we need to multiply by U which is the probability of $\xi_i = 1$ conditioned on U. Otherwise, for the same reason, if $\varepsilon_i = 0$, we need to multiply by $(1-U)$. Note here we used the fact that ξ_1, ξ_2 are independent conditioned on U. The evaluation of the integral in the last line comes from the density of a Beta distribution [5] (we will assume this, you can check out the formula from the wiki link).

Thus we can now compute conditional distribution of $\xi_{n+1} = 1$ conditioned on $(\xi_1, \ldots, \xi_n) = (\varepsilon_1, \ldots, \varepsilon_n)$ where we assume $\sum_{i=1}^n \xi_i = k$. By the same logic as above,

$$\begin{aligned}
\mathbb{P}(\xi_{n+1} = 1 | \xi_1 = \varepsilon_1, \ldots, \xi_n = \varepsilon_n) &= \frac{\mathbb{P}(\xi_1 = \varepsilon_1, \ldots, \xi_n = \varepsilon_n, \xi_{n+1} = 1)}{\mathbb{P}(\xi_1 = \varepsilon_1, \ldots, \xi_n = \varepsilon_n)} \\
&= \frac{\int_0^1 u^{k+1}(1-u)^{n-k}}{\int_0^1 u^k (1-u)^{n-k}} \\
&= \frac{\frac{(k+1)!(n-k)!}{(n+2)!}}{\frac{k!(n-k)!}{(n+1)!}} \\
&= \frac{k+1}{n+2}.
\end{aligned} \tag{1.6}$$

This tells us that if more 1s are sampled in the first n draws, it is more likely that we have a 1 in the next draw, something which is not true for just i.i.d. random variables. Thus we can note already here that $(\xi_1, \xi_2, \ldots, \xi_n)$ are not unconditionally i.i.d. although they are conditionally i.i.d. given $U = u$.

An urn model. Now let us consider something which is apparantly completely different. Consider the following urn model which is the following stochastic process. Suppose initially there are r red balls and b black balls in an urn. In each step, a ball is chosen uniformly at random.

- If it is red (resp. black), the ball is put back along with an extra red (resp. black) ball.

[5] See https://en.wikipedia.org/wiki/Beta_distribution to learn more

- If it is black, the ball is put back along with an extra black ball.

Let

$$X_n = \begin{cases} 1 \text{ if } n\text{th ball drawn is red} \\ 0 \text{ otherwise} \end{cases}$$

We now make the following (perhaps surprising) claim:

Proposition 1.44. *Suppose $r = b = 1$. Let $\xi_1, \xi_2, \ldots$ be i.i.d. Bernoulli (U) (as in the uniform prior example). Let $(X_1, X_2, \ldots)$ be as described above. Then $(\xi_1, \xi_2, \ldots) = (X_1, X_2, \ldots)$ in distribution (as stochastic processes). In other words, for every $n \geq 1$,*

$$(X_1, \ldots, X_n) = (\xi_1, \ldots, \xi_n) \text{ in distribution.}$$

The claim above is surprising as the description of the two processes are very different, yet it turns out that they have the same distribution.

Proof. The proof is surprisingly easy. We already know from (1.4) that $\mathbb{P}(\xi_n = 1) = \frac{1}{2}$ for any $n \geq 1$. Of course it is easy to calculate the distribution of X_1:

$$\mathbb{P}(X_1 = 1) = \frac{r}{r+b} \qquad \mathbb{P}(X_1 = 0) = \frac{b}{r+b}.$$

If $r = b = 1$, both the ratios above are $1/2$ and hence $X_1 = \xi_1$ in distribution if $r = b = 1$.

How to compute the conditional distribution of X_n conditioned on $(X_1, \ldots, X_{n-1})$? This is actually not so hard. The vector $(X_1, \ldots, X_{n-1})$ is a sequence of 0s and 1s. The probability of $X_n = 1$ given $(X_1, \ldots, X_{n-1})$ is the number of red balls in the urn at time $n-1$ divided by the total number of balls in the urn at time $n-1$. The latter is deterministic as we always put in 1 ball in the urn, so the denominator is $r+b+n-1$. The numerator is simply $r + \sum_{i=1}^{n-1} X_i$. Thus, we have the formula

$$\mathbb{P}(X_{n+1} = 1 | X_1 = \varepsilon_1, \ldots, X_{n-1} = \varepsilon_{n-1}) = \frac{r + \sum_{i=1}^{n-1} \varepsilon_i}{r+b+n-1}.$$

Plugging in $r = b = 1$ and assuming $\sum_{i=1}^{n-1} \varepsilon_i = k$, we see that

$$\mathbb{P}(X_{n+1} = 1 | X_1 = \varepsilon_1, \ldots, X_{n-1} = \varepsilon_{n-1}) = \frac{1+k}{n+2}$$

Thus the conditional distribution of ξ_{n+1} given $(\xi_1, \ldots, \xi_n)$ is given by the same formula using (1.6). This is enough to conclude by induction that $(X_1, \ldots, X_n) = (\xi_1, \ldots, \xi_n)$ in distribution. (Exercise: Convince yourself of the last step.) □

1.8 De-Finetti's theorem

Is there a deeper reason behind Proposition 1.44 or was it a fluke? Turns out, that this phenomenon can be explained by the notions of **exchangeability** and a fascinating theorem called De-Finetti's theorem. We will attempt to explain this now. Although we already know that if $r = 1, b = 1$, $X_n = \xi_n$ in distribution, and hence $X_n \sim$ Bernoulli $(1/2)$, is there a way to prove this directly without referring to the ξ_ns? We now prove the following proposition, which computes the distribution of X_n for general initial condition r, b.

Proposition 1.45. *For any $n \geq 1$,*

$$\mathbb{P}(X_n = 1) = \frac{r}{r+b} \qquad \mathbb{P}(X_n = 0) = \frac{b}{r+b}.$$

That is the distributions of X_n and X_1 are the same!

Proof. We use induction on n, but there is a tricky step. It seems like conditioning on X_{n-1} is a good choice, but the conditional distribution of X_n given X_{n-1} is complicated as we do not know what happened in the first $n-2$ draws.On the other hand, we can condition on the first draw, and then think that the urn process 'refreshed' with a different initial condition of balls, which depends on what happened in the first draw.

Here is a full proof. The induction hypothesis is crucial, we will explain a subtlety about this assumption later. Suppose **for any two integers** u, v, and an urn initially containing u red and v black balls,

$$\mathbb{P}(X_{n-1} = \text{ red }) = \frac{u}{u+v} \qquad \mathbb{P}(X_{n-1} = \text{ black }) = \frac{v}{u+v}.$$

Then for our urn initially containing r red and b black balls,

$$\begin{aligned}\mathbb{P}(X_n = \text{red}) &= \mathbb{P}(X_n = \text{red}|X_1 = \text{red})\mathbb{P}(X_1 = \text{red}) + \mathbb{P}(X_n = \text{red}|X_1 = \text{black})\mathbb{P}(X_1 = \text{black}) \\ &= \frac{r+1}{r+b+1}\frac{r}{r+b} + \frac{r}{r+b+1}\frac{b}{r+b} \\ &= \frac{r}{r+b}.\end{aligned}$$

which completes the proof. In the induction step, we used the fact that if the first step produced a red, then we have an urn with $r+1$ red and b black balls. Thus the conditional distribution of X_n given $X_1 = 1$ is the same as the distribution of X_{n-1} when we start the process with $r+1$ red and b black balls. Since our induction step was assumed that the formula is true for for any starting configuration of red and black balls (which is critical!), induction gives us the second line from the first line. The rest is just algebra. □

In fact, more can be said. We claim that for any $n \neq m$,

$$\mathbb{P}(X_n = \text{red}, X_m = \text{black}) = \mathbb{P}(X_m = \text{red}, X_n = \text{black}).$$

This follows from the more general proposition. First we need to introduce the notion of exchangeability.

Definition 1.46. *We say $X_1, X_2, \ldots, X_n$ are **exchangeable** if for any permutation π of $\{1, 2, \ldots, n\}$,*

$$(X_1, \ldots, X_n) \overset{(d)}{=} (X_{\pi(1)}, \ldots, X_{\pi(n)}).$$

This might seem a weird notion, but is actually quite natural. For example if $(X_1, \ldots, X_n)$ are i.i.d. then they are definitely exchangeable.

Lemma 1.47. *If $X_1, X_2, \ldots, X_n$ are i.i.d. discrete random variables with common probability mass function $p_X(\cdot)$, then $(X_1, \ldots, X_n)$ is exchangeable: for every permutation π of $\{1, \ldots, n\}$ and every $x_1, \ldots, x_n$,*

$$\mathbb{P}\big(X_1 = x_1, \ldots, X_n = x_n\big) = \mathbb{P}\big(X_{\pi(1)} = x_1, \ldots, X_{\pi(n)} = x_n\big).$$

Proof. Since the X_i are independent and identically distributed with common pmf p_X, the joint pmf factors as

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^{n} p_X(x_i).$$

For any permutation π of $\{1, \ldots, n\}$ we have

$$\mathbb{P}(X_{\pi(1)} = x_1, \ldots, X_{\pi(n)} = x_n) = \prod_{i=1}^{n} p_X(x_{\pi(i)}).$$

But of course, the product of numbers is invariant under reordering, so

$$\prod_{i=1}^{n} p_X(x_{\pi(i)}) = \prod_{i=1}^{n} p_X(x_i).$$

Therefore

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) = \mathbb{P}(X_{\pi(1)} = x_1, \ldots, X_{\pi(n)} = x_n),$$

which proves exchangeability. □

Remark. The same proof works for i.i.d. continuous random variables (replace pmf by pdf); the key property is identical distribution together with independence so the joint density is a product of identical marginal densities, hence symmetric. We leave it as an exercise to complete this.

Counterexample: independent but not identically distributed. Independence alone does *not* imply exchangeability. A concrete counterexample with two discrete variables:

Let $X_1 \sim \text{Bernoulli}(0.2)$ and $X_2 \sim \text{Bernoulli}(0.8)$, with X_1 and X_2 independent. Thus

$$\mathbb{P}(X_1 = 1) = 0.2, \quad \mathbb{P}(X_1 = 0) = 0.8,$$

$$\mathbb{P}(X_2 = 1) = 0.8, \quad \mathbb{P}(X_2 = 0) = 0.2.$$

Compute

$$\mathbb{P}(X_1 = 1, X_2 = 0) = \mathbb{P}(X_1 = 1)\mathbb{P}(X_2 = 0) = 0.2 \cdot 0.2 = 0.04,$$

whereas

$$\mathbb{P}(X_1 = 0, X_2 = 1) = \mathbb{P}(X_1 = 0)\mathbb{P}(X_2 = 1) = 0.8 \cdot 0.8 = 0.64.$$

Since $\mathbb{P}(1,0) \neq \mathbb{P}(0,1)$, the joint distribution is not invariant under swapping the two coordinates; hence (X_1, X_2) is not exchangeable.

Exchangeability is a more general notion than independence. There are dependant random variables which are exchangeable.

Example 1.48 (uniform prior)**.** Suppose $U \sim Unif[0,1]$, and given U, $(\xi_1, \ldots, \xi_n)$ are i.i.d. Bernoulli (U). We claim this is exhangeable.

Fix arbitrary values $x_1, \ldots, x_n \in \{0,1\}$. Let

$$k := \sum_{i=1}^{n} x_i$$

be the number of ones among the x_i's. We already know from (1.5), that

$$\mathbb{P}(\xi_1 = x_1, \ldots, \xi_n = x_n) = \int_0^1 u^k (1-u)^{n-k}\, du. \tag{1.7}$$

Now let π be any permutation of $\{1, \ldots, n\}$. The permuted event $\{\xi_{\pi(1)} = x_1, \ldots, \xi_{\pi(n)} = x_n\}$ has the same number of ones among its coordinates as the original event; indeed the multiset of values $\{x_1, \ldots, x_n\}$ is unchanged by permutation, so the corresponding sum is again k. Repeating the same conditioning calculation yields

$$\mathbb{P}(\xi_{\pi(1)} = x_1, \ldots, \xi_{\pi(n)} = x_n) = \int_0^1 u^k (1-u)^{n-k}\, du.$$

Therefore for every permutation π we have

$$\mathbb{P}(\xi_1 = x_1, \ldots, \xi_n = x_n) = \mathbb{P}(\xi_{\pi(1)} = x_1, \ldots, \xi_{\pi(n)} = x_n),$$

which is exactly the definition of exchangeability. □

Example 1.49. If $(X_1, \ldots, X_n)$ is a Multivariate Normal $N(\boldsymbol{\mu}, \Sigma)$ with $\mathrm{Cov}(X_i, X_j) = \rho$ if $i \neq j$ and $\mathrm{Var}(X_i) = 1$ for all i then it is exchangeable. Why? This is simply because the multivariate Normal distribution is completely determined if we know the mean and the covariance matrices, which remains completely unchanged if we permute the variables.

Coming back to the urn model:

Proposition 1.50. *Let $(X_n)_{n\geq 1}$ be as in Proposition 1.45. Then $(X_1, X_2, \ldots, X_n)$ is exchangeable.*

Remark 1.51. It follows immediately from Proposition 1.50 that for any $m \neq n$

$$(X_n, X_m) = (X_m, X_n) \text{ in distribution}$$

Proof sketch of Proposition 1.50. We prove for $n = 3$.

$$\mathbb{P}(X_1 = 1, X_2 = 0, X_3 = 1) = \frac{rb(r+1)}{(r+b)(r+b+1)(r+b+2)}.$$

similarly

$$\mathbb{P}(X_1 = 1, X_2 = 1, X_3 = 0) = \frac{r(r+1)b}{(r+b)(r+b+1)(r+b+2)}.$$

Thus the probability of the sequence $(1, 1, 0)$ is the same as that of the sequence $(1, 0, 1)$. In fact any permutation of this sequence has the same probability. From this you can easily convince yourself that the probability only depends on the **number of red and black balls in the sequence chosen** which implies exchangeability.

Here is a more detailed proof. for every permutation π of $\{1, \ldots, n\}$ and every binary vector $(x_1, \ldots, x_n) \in \{0, 1\}^n$,

$$\mathbb{P}\big(X_1 = x_1, \ldots, X_n = x_n\big) = \mathbb{P}\big(X_{\pi(1)} = x_1, \ldots, X_{\pi(n)} = x_n\big).$$

Notation. Write $k = \sum_{i=1}^n x_i$ for the number of red draws among $x_1, \ldots, x_n$ (so $0 \leq k \leq n$). For an integer $m \geq 0$ and $a \in \mathbb{R}$ we denote the rising factorial (Pochhammer symbol)

$$a^{(m)} := a(a+1)\cdots(a+m-1),$$

with the convention $a^{(0)} := 1$.

Sequential probability calculation. Fix a particular sequence $(x_1, \ldots, x_n)$ with exactly k ones. Under the Pólya urn dynamics the probability of observing this particular ordered sequence is obtained by multiplying the one-step conditional probabilities. At the first draw the probability of drawing red is $r/(r+b)$ and black is $b/(r+b)$. More generally, if so far j red draws and $i-1-j$ black draws have occurred in the first $i-1$ draws, then the probability that the i-th draw is red equals

$$\frac{r+j}{r+b+i-1},$$

and the probability it is black equals

$$\frac{b+(i-1-j)}{r+b+i-1}.$$

Multiplying these stepwise probabilities along the sequence $(x_1, \ldots, x_n)$ we obtain

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^{n} \frac{r + (\text{number of reds among } x_1, \ldots, x_{i-1})}{r+b+i-1} \quad \text{for each } i \text{ with } x_i = 1,$$

and similarly the appropriate factor for $x_i = 0$. Rearranging the product by collecting the contributions of red draws and black draws gives the closed form

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) = \frac{r^{(k)}\, b^{(n-k)}}{(r+b)^{(n)}}. \tag{1.8}$$

(Indeed, the numerator is the product of the k red-factors $r, (r+1), \ldots, (r+k-1)$ times the $n-k$ black-factors $b, (b+1), \ldots, (b+n-k-1)$; the denominator is the product $(r+b), (r+b+1), \ldots, (r+b+n-1)$.)

Joint probability depends only on k. The right-hand side of (1.8) depends only on $k = \sum_i x_i$ and not on the order in which the ones and zeros appear. Hence any two sequences with the same number k of red draws have the same probability. In particular, for any permutation π of $\{1, \ldots, n\}$,

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) = \mathbb{P}(X_{\pi(1)} = x_1, \ldots, X_{\pi(n)} = x_n),$$

which is exactly the definition of exchangeability. This completes the proof. □

Exchangeable sequences satisfy the following remarkable theorem due to De Finetti:

Theorem 1.52 (De-Finetti's theorem)**.** *If $(\xi_1, \ldots, \xi_n)$ are exchangeable and takes values in $\{0, 1\}$. Then there is a random variable U defined on the interval $[0, 1]$ so that conditioned on U, $(\xi_1, \ldots, \xi_n) \sim$ i.i.d. Bernoulli (U).*

Note that the theorem does NOT imply that the law of $(\xi_1, \ldots, \xi_n)$ must be i.i.d. themselves as we saw that there are exchangeable sequences which are not i.i.d. Also the theorem does not say what U is, in fact U can be quite mysterious!

Now back to the urn problem. Let $\xi_1, \xi_2, \ldots, \xi_n$ be such that $\xi_i = 1$ if X_i is red and $\xi_i = 0$ if X_i is black. Then applying De-Finetti, there is a random variable Z supported on [0,1] such that $\xi_i \sim$ i.i.d. Bernoulli (Z). What is the law of Z?

Claim 1.53. *If we start with 1 red and 1 black ball, then $Z \sim Unif[0, 1]$. (This exactly tells us that in this case, $(\xi_1, \ldots, \xi_n)$ coming from the urn process as above has the exact same law as the uniform prior case.)*

Proof. Note that in the urn process starting with 1 black and one red, and $\xi_i = 1$ if X_i is red and $\xi_i = 0$ if X_i is black:

$$\mathbb{P}(\xi_1 = 1, \xi_2 = 1, \ldots, \xi_n = 1) = \frac{1}{2}\frac{2}{3}\frac{3}{4} \cdots \frac{n}{n+1} = \frac{1}{n+1}.$$

We also know from De-Finetti's theorem that conditioned on Z, $(\xi_1 = 1, \xi_2 = 1, \ldots, \xi_n = 1)$ are i.i.d. Bernoulli(Z). So

$$\mathbb{P}(\xi_1 = 1, \xi_2 = 1, \ldots, \xi_n = 1) = \mathbb{E}(\mathbb{P}(\xi_1 = 1, \xi_2 = 1, \ldots, \xi_n = 1 | Z)) = \mathbb{E}(Z^n).$$

Thus $\mathbb{E}(Z^n) = \frac{1}{n+1} = \int_0^1 x^n dx$ which is exactly the nth moment of the Unif[0, 1] distribution. Since Unif[0, 1] has an mgf which contains an interval around 0, we get that this completely determines the distribution of Z to be Unif[0, 1]. □

Exercise 1.54. *(hard) Show that if we start with r red and b black balls then Z (coming from De-Finetti's theorem) follows Beta(r, b) distribution.*

1.8.1 Further suggested reads (Advanced topics)

Ballot problem Suppose in an election A receives n votes and B receives m votes. The votes are counted one by one and all orderings are equally likely. Find the probability that A is always ahead. Example 3.28 (Ed 12), Example 3.27 (Ed 11).

Example 3.28 in 11th and Example 3.29 in 12th edition. Suppose $U_1, U_2, \ldots$ i.i.d. Unif[0,1]. Let

$$N = \min\{n \geq 2 : U_n > U_{n-1}\}/$$

and

$$M = \min\{n \geq 1 : U_1 + \ldots + U_n > 1\}.$$

It is shown that N and M has the same probability distribution (which might be surprising!)

Left skip-free random walk. Read Section 3.6.6 in both editions. For connections with branching processes and this walk, read the paper by M. Dwass called "The total progeny in a branching process and a related random walk".

Exercises

1. Let (X, Y) have joint pdf

$$f_{X,Y}(x, y) = \begin{cases} 4x(1-y), & 0 < x < 1,\ 0 < y < 1, \\ 0, & \text{otherwise.} \end{cases}$$

   a. Find the conditional density $f_{X|Y}(x \mid y)$ for $0 < y < 1$.

   b. Compute the conditional expectation $\mathbb{E}[X \mid Y = y]$.

   c. Use the law of total expectation to find $\mathbb{E}[X]$.

   d. Compute $\mathrm{Var}(X \mid Y = y)$ and $\mathrm{Var}(X)$.

2. Let (X, Y) be discrete random variables with joint pmf

$$\mathbb{P}(X = x, Y = y) = \frac{x + y}{21}, \qquad x \in \{1, 2\},\ y \in \{1, 2, 3\},$$

   and $\mathbb{P}(X = x, Y = y) = 0$ otherwise.

   (a) Find the marginal pmf of Y.

   (b) Compute the conditional pmf $p_{X|Y}(x \mid y)$.

(c) Find $\mathbb{E}[X \mid Y = y]$.

(d) Compute $\mathbb{E}[X]$ using the law of total expectation.

3. Suppose $X \sim$Bernoulli(P) where $P \sim$Unif$(0, 1)$.

   a. Calculate the conditional density of P given $X = 1$. Do the same, but given $X = 0$.

   b. Compute also the conditional expectations, $\mathbb{E}(P|X = 0)$ and $\mathbb{E}(P|X = 1)$. Which is bigger?

   c. Verify that $\mathbb{E}(P) = \mathbb{E}(P|X)$ using the calculations above.

4. Let $X \sim \text{Exp}(\lambda_1)$ and $Y \sim \text{Exp}(\lambda_2)$ be independent random variables (with $\lambda_1, \lambda_2 > 0$). Consider the event $X < Y$.

   (a) Compute $\mathbb{P}(X < Y)$.

   (b) Find the conditional distribution of the excess lifetime $W := Y - X$ given $X < Y$.

   (c) Show that $W \mid (X < Y) \sim \text{Exp}(\lambda_2)$.

5. Let $G(n, p)$ be the Erdös–Rényi random graph on vertex set $[n]$, where each edge is present independently with probability p. Fix two distinct vertices $u, v \in [n]$.

   (a) Compute $\mathbb{P}\big((u, v) \in E\big)$.

   (b) Compute the conditional probability

   $$\mathbb{P}\big((u, v) \in E \mid \deg(u) = k\big),$$

   where $\deg(u)$ is the degree of vertex u and $0 \le k \le n - 1$.

   (c) Interpret the result: does knowing $\deg(u) = k$ affect the probability that u is connected to v?

6. Let $(S_n)_{n\ge0}$ be a simple symmetric random walk on $\mathbb{Z}$ starting at $S_0 = 0$:

   $S_n = X_1 + X_2 + \cdots + X_n, \quad X_i \sim$ Rademacher(±1) i.i.d. (another name for Uniform $\{1, -1\}$).

   (a) Compute $\mathbb{P}(S_n = k)$ for $-n \le k \le n$ with $k \equiv n \pmod 2$.

   (b) Compute the conditional probability

   $$\mathbb{P}(X_1 = +1 \mid S_n = k).$$

   (c) Interpret the result: does knowing $S_n = k$ affect the probability of the first step being $+1$?

7. Let $(S_n)_{n\ge0}$ be as in Qn 6. Fix positive integers $a, b > 0$ and define the stopping time

   $$\tau = \min\{n \ge 0 : S_n = a \text{ or } S_n = -b\}.$$

(a) Compute the probability that the walk reaches $+a$ before $-b$: $\mathbb{P}(S_\tau = a)$.

(b) Compute the conditional probability

$$\mathbb{P}(X_1 = +1 \mid S_\tau = a),$$

where X_1 is the first step.

(c) Interpret the result: how does conditioning on eventually hitting $+a$ affect the first step?

2 Markov chains

Roughly speaking, a Markov chain is a sequence of random variables $X_0, X_1, \ldots$ such that the distribution of X_n given the *whole past* (that is, $X_0, X_1, \ldots X_{n-1}$) depends only on the immediate past (that is just X_{n-1}).

Definition 2.1. *We say $(X_n)_{n\geq 0}$ is a Markov chain if for all $n \geq 1$,*

$$\mathbb{P}(X_n = j | X_{n-1} = i, X_{n-2} = i_{n-1}, \ldots, X_0 = i_0) = \mathbb{P}(X_n = j | X_{n-1} = i).$$

*for all $i, j, i_1, \ldots, i_{n-1}$. If the above probability does not depend on n, we say the Markov chain is **time homogeneous**. We shorten the notation: $\mathbb{P}(X_n = j | X_{n-1} = i) = P_{ij}$ for a time homogeneous Markov chain.*

Example 2.2. 1. **i.i.d. sequence**. Any such sequence is clearly a Markov chain.

2. **Random walk/ Sequences of heads and tails/ biased walks**. Sometimes Markov chains can be hidden in the description of the process. For example, let $S_1, S_2, \ldots, S_n$ be defined such that

$$\mathbb{P}(S_1 = s_1, \ldots, S_n = s_n) = \prod_{i=1}^{n} (p 1_{s_i - s_{i-1} = 1} + (1-p) 1_{s_i - s_{i-1} = -1}).$$

This is actually a simple random walk. Toss a coin independently and let Let $\mathbb{P}(X_i = +1) = p = 1 - \mathbb{P}(X_i = -1)$. Let $S_k = \sum_{i=1}^{k} X_i$. Thus we have

$$\mathbb{P}(S_n = s_n | S_{n-1} = s_{n-1}, \ldots, S_1 = s_1) = \mathbb{P}(S_n = s_n | S_{n-1} = s_{n-1}) = \begin{cases} 1 \text{ if } s_n - s_{n-1} = 1 \\ -1 \text{ if } s_n - s_{n-1} = -1 \\ 0 \text{ otherwise .} \end{cases}$$

Here is another example.

Suppose $S_1, S_2, \ldots, S_n$ be random variables taking values in $\mathbb{Z}$ defined such that

$$\mathbb{P}(S_1 = s_1, \ldots, S_n = s_n) = \frac{1}{C_n} \prod_{i=1}^{n} \exp(-\lambda (s_i - s_{i-1})^2). \quad s_i \in \mathbb{Z} \quad \forall i$$

where C_n is a constant which makes the above a joint pmf.

Exercise 2.3. *Show that the above example $(S_n)_{n\geq 1}$ is a Markov chain. You need to show that*

$$\mathbb{P}(S_1 = s_1, \ldots, S_k = s_k | S_1 = s_1, \ldots, S_{k-1} = s_{k-1})$$

*does not depend on $(s_1, \ldots, s_{k-2})$ for all $k < n$. **Warning:** You cannot assume that the joint distribution for any $k < n$ is of the form given above.*

3. **Urn process** Consider the Urn process discussed in Section 1.7.1. Let $X_n = 1_{\text{ nth draw red }}$. Is X_n a Markov chian. We saw that it is NOT a Markov chain as the probability we draw red in $n+1$th draw given what we have drawn in the first n draws depend on the total number of black and red balls drawn in the previous draws. However can we consider a random variable (Obviously carrying more information), so that it is a Markov chain? The answer is Yes. For example $R_n :=$ Number of red balls after n draws is a Markov chain (exercise: check this!). Note however

4. In example 1.34, the position of the walker is a Markov chain.

Here is a precise definition. In this course, we will always *deal with Markov chains in which the random variables are always* ***discrete***, that is takes values in a finite or countable set. Unless otherwise mentioned we will take this space (called *state space*) to be the set of natural numbers $\mathbb{N}$. While the index n in X_n should be thought of as *time*. For now, we will deal with discrete time Markov chains, that is $n \in \mathbb{N}$.

Our Markov chains will be time homogeneous by default, unless stated otherwise. Notice that it is clear from the definition that

$$P_{ij} \geq 0, \quad \forall i, j \in \mathbb{N} \qquad ; \qquad \sum_j P_{ij} = 1, \forall i.$$

Sometimes it is written as a matrix $\boldsymbol{P}$ with the (i,j)th entry being P_{ij}. This is a possibly $\infty \times \infty$ matrix, but dont be alarmed, just treat it like you treat any other matrix for now. (In reality, this is an *operator* called *Markov operator*). Sometimes a Markov chain can be depicted by a picture as follows:

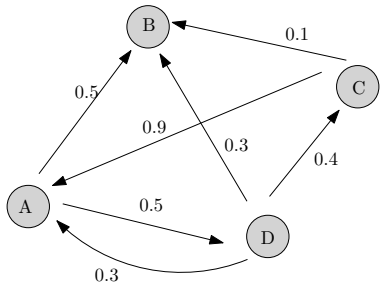


Figure 2: This Markov chain has 4 states A, B, C, D. The transition probabilities between states are given by the numbers on the arrows. Note that the sum of the probabilities of leaving a state must be 1.

Example 2.4. 1. **Two step chain** Example 4.4 from book.

2. **Self avoiding walk** Consider the set of paths of length n in the square lattice which do not intersect itself. Pick uniformly from this set a path. Let $(X_k, Y_k)_{1 \leq k \leq n}$ be the coordinates. Is this a Markov chain? Clearly not as the next step of the chain depends

on the *whole past*. THis is the notorious example of *self-avoiding walk*, something very hard to understand. [6].

3. **Hidden Markov model** Consider the previous figure in the Markov chain and color states A, C as red and B, D as blue as follows.

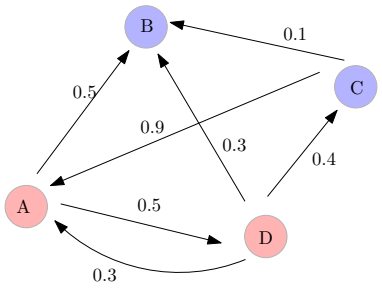


Figure 3: If we are just given the information of the colors, then this is a hidden Markov model.

Now consider the chain X_n which gives the color of the state and not the actual state. Is this a Markov chain? It is NOT a Markov chain as knowing just the color does not give us information about the transition probabilities. However, there is a Markov chain hidden underneath this chain which might be put to our use. This is called a *hidden Markov model.*

4. **Increasing state space to Markovify a chain.** In the self avoiding walk example, indeed (X_k, Y_k) is not a Markov chain. However if we take into account the whole past, then $(Z_k = ((X_j, Y_j) : 1 \leq j \leq k))_{1 \leq k \leq n}$ is a Markov chain. (Exercise: convince yourself.)

2.1 Chapman–Kolmogorov equations

We introduce the following notation for *n-step transition probabilities*:

$$\mathbb{P}(X_n = j | X_0 = i) = P_{ij}^{(n)}.$$

[6] See https://en.wikipedia.org/wiki/Self-avoiding_walk

Clearly $P_{ij}^1 = P_{ij}$ going back to the previous notation. Now let us compute P_{ij}^2.

$$\begin{aligned}
P_{ij}^{(2)} &= \mathbb{P}(X_2 = j | X_0 = i) \\
&= \sum_{k=0}^{\infty} \mathbb{P}(X_2 = j, X_1 = k | X_0 = i) \\
&= \sum_{k=0}^{\infty} \mathbb{P}(X_2 = j | X_1 = k) \mathbb{P}(X_1 = k | X_0 = i) \\
&= \sum_{k=0}^{\infty} \mathbb{P}(X_2 = j | X_1 = k) \mathbb{P}(X_1 = k | X_0 = i) \\
&= \sum_{k=0}^{\infty} P_{kj} P_{ik}
\end{aligned}$$

The point of this calculation is to note that P_{ij}^2 is the (i,j)th entry of the matrix $\boldsymbol{P}^2$ (where we multiply the $\infty \times \infty$ matrix just like we multiple a finite matrix.) Thus the matrix corresponding to the *two step chain is*

$$\boldsymbol{P}^{(2)} = \boldsymbol{P}^2$$

Actually this calculation can be generalized by iteration. One can have

$$\begin{aligned}
P_{ij}^{(n+m)} &= \mathbb{P}(X_{n+m} = j | X_0 = i) \\
&= \sum_{k=0}^{\infty} \mathbb{P}(X_{n+m} = j, X_n = k | X_0 = i) \\
&= \sum_{k=0}^{\infty} \mathbb{P}(X_{n+m} = j | X_n = k) \mathbb{P}(X_n = k | X_0 = i) \\
&= \sum_{k=0}^{\infty} \mathbb{P}(X_m = j | X_0 = k) \mathbb{P}(X_n = k | X_0 = i) \text{ (Here is where we use the Markov property)} \\
&= \sum_{k=0}^{\infty} P_{kj}^{(m)} P_{ik}^{(n)} = \sum_{k=0}^{\infty} (\boldsymbol{P}^n)_{ik} (\boldsymbol{P}^m)_{kj}
\end{aligned}$$

Using the recursion, we see that

$$P_{ij}^{(n)} = (\boldsymbol{P}^n)_{ij}$$

In other words

Proposition 2.5. *If we want to compute the transition probabilities of an n-step Markov chain, we simply compute the matrix* $\boldsymbol{P}^n$.

Example 2.6. Simpler version of Example 4.11, Example 4.12, 11th and 12th edition.

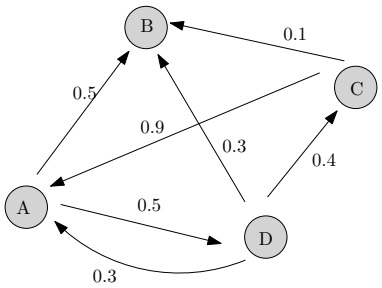


Figure 4: This Markov chain has 4 states A, B, C, D. The transition probabilities between states are given by the numbers on the arrows. Note that the sum of the probabilities of leaving a state must be 1.

2.2 Classification of states

The states of a Markov chain can be classified to our advantage so as to better understand a Stochastic process. Consider the Markov chain of the simple random walk in the cycle in example 1.34. One can consider the same problem but now with two disjoint cycles. Clearly the walk in one cycle has nothing to do with the other and one can partition the states into two subclasses one coming from each cycle. The point of this Section is to generalize this idea.

Definition 2.7. *Take two states i, j in a Markov chain. We say j is* accessible *from i if*

$$P_{ij}^{(n)} = \mathbb{P}(X_n = j | X_n = i) > 0 \text{ for some } n \geq 0.$$

This is equivalent to say that there is a positive chance that one can reach state j from i *eventually*. Said otherwise, if j is not accessible from i then

$$\mathbb{P}(X_n = j \text{ for some } n \geq 1) = 0 \text{ (Exercise: Why?)}$$

For example, if one considers simple random walk on two disjoint cycles, clearly vertices of one cycle is not accessible from the other cycle.

It is important to observe that in order for j to be accessible from i, one does not require that there is a positive chance to jump to i *in one step*, one only needs that there is a positive chance the chain *eventually* reaches j. For example let us go back to the chain in Figure 4. Observe that one cannot go from state A to state C in one step, but can do so in two steps taking a detour through D.

Coming to the next question: if i is accessible to j is j accessible to i ? The answer is not necessarily yes, take for example the chain in Figure 5. Clearly, state C is accessible from every other state, but once the chain reaches C, it stays there forever (C acts like a cemetery.)

This brings us to the following definition.

Definition 2.8. *We say i communicates with j both i is accessible from j and vice-versa.*

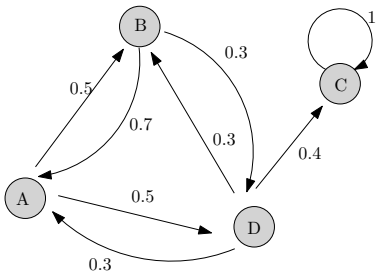


Figure 5: This Markov chain has 4 states A, B, C, D. The transition probabilities between states are given by the numbers on the arrows. Note that the sum of the probabilities of leaving a state must be 1.

Proposition 2.9. *Communication is an equivalence relation. Said otherwise:*

a. i communicates with i.

b. If i communicates with j then j communicates with i.

c. If i communicates with j and j communicates with k then i communicates with k.

Proof. Properties a. follows from the definition (note that we included $n = 0$ there). b. follows from definition. For c. apply Chapman–Kolmogorov (Exercise: convince yourself!) □

Since communication is an equivalence relation, we can divide the states into *equivalence classes.* [7]

Definition 2.10. *We say a Markov chain is irreducible if there is only one equivalence class, that is, if every state can communicate with every other state.*

Example 2.11. See Example 4.15, 4.16 from book.

2.3 Recurrence and transience

We now want to study if a Markov chain returns to the starting state or not and if so how many times. To that end define

$$f_i = \mathbb{P}(\text{ The chain starting from } i \text{ eventually returns to } i)$$

Definition 2.12. *We say a state i is recurrent if $f_i = 1$ and transient otherwise (i.e. i is transient if $f_i < 1$).*

[7] If you are unfamiliar with equivalence classes read https://en.wikipedia.org/wiki/Equivalence_relation.

Note that if a state i is recurrent, it eventually once it visits i the Markov chain "resets" as if it started from i, so it will again come back to i. This means that if a state i is recurrent,

$$\mathbb{P}(\text{ chain returns to } i \text{ infinitely often }) = 1$$

On the other hand, if $f_i < 1$ then once the chain returns to i, it has probability $1 - f_i$ to not return to i. In this case, the number of returns is a geometric random variable with paramater f_i (think of tossing a coin with success probability $1 - f_i$, each toss tells us whether the chain returns to i or not and we call it a success if the chain does not return to i eventually). Thus let N be the number of returns to i,

$$\mathbb{P}(N = n | X_0 = i) = f_i^{n-1}(1 - f_i), \quad n \geq 1$$

and thus, using the formula for the expectation of a geometric, we get

$$\mathbb{E}(N | X_0 = i) = \frac{1}{1 - f_i}.$$

Note that $\mathbb{E}(N|X_0 = i)$ is finite if and only if $f_i < 1$ (we assume $1/0 = \infty$).

Here is another way to interpret this. Let $I_n = 1_{X_n = i}$. Then $N = \sum_{n=0}^{\infty} I_n$. Thus

$$\mathbb{E}(N | X_0 = i) = \mathbb{E}(\sum_{n=0}^{\infty} I_n | X_0 = i) = \sum_{n=0}^{\infty} \mathbb{P}(X_n = i | X_0 = i) = \sum_{n=0}^{\infty} P_{ii}^{(n)}.$$

Thus we get a *dichotomy*

Proposition 2.13. *A state i is*

- *Recurrent if $\sum_{n=0}^{\infty} P_{ii}^{(n)} = \infty$*
- *Transient if $\sum_{n=0}^{\infty} P_{ii}^{(n)} < \infty$.*

Now we ask the question, if i and j communicates with each other, and if j is recurrent then is i recurrent? The answer should be "yes" as one can go from i to j in a finite number of steps and then eventually return to j and then trace back to i again using the communicating property.

Proposition 2.14. *If i communicates with j and j is recurrent, then i is recurrent. Consequently, in a communicating class either all states are transient, or all states are recurrent.*

Proof. We know from the definition of communication that there is some k, m such that $P_{ij}^k > 0$ and $P_{ji}^m > 0$ Thus using Chapman–Kolmogorov equations

$$\sum_{\ell=0}^{\infty} P_{ii}^{(\ell)} \geq \sum_{n=0}^{\infty} P_{ii}^{(k+n+m)} \geq \sum_{n=0}^{\infty} P_{ij}^{(k)} P_{jj}^{(n)} P_{ji}^{(m)} = P_{ij}^{(k)} P_{ji}^{(m)} \sum_{n=0}^{\infty} P_{jj}^{(n)} = \infty$$

since j is recurrent using Proposition 2.13. □

Proposition 2.15. *Every finite state space Markov chain must have at least one recurrent state.*

Proof. This is clear as if every state is transient then the event that

$$\{X_n \text{ eventually leaves state } i \text{ for all } i \text{ in the state space}\}.$$

has probability 1. But this set is clearly an empty set. This is a contradiction. □

We deduce the following corollaries for finite state Markov chain.

Proposition 2.16. *Every finite state Markov chain there is at least one recurrent state. Thus in every finite state, irreducible Markov chain, every state is recurrent.*

Every countably infinite state, irreducible Markov chain has either all states recurrent or all states transient. We call such chains a recurrent Markov chain or transient Markov chain in short.

2.4 Simple random walk on the $\mathbb{Z}^d$ lattice

We will consider the infinite lattice graph $\mathbb{Z}^d$ for $d \geq 1$ which consists of vertices given by d tuples of integers and two vertices are connected if exactly one of their coordinates differ by exactly ± 1. A bit of thought will show that for $d = 1$ this is simply the infinite path, $d = 2$ is the square lattice, $d = 3$ is the 3d grid and so on. The Markov chain we consider is simply the simple random walk, that is when in a state $\boldsymbol{i} = (i_1, \ldots, i_d)$, it moves to one of it's $2d$ neighbours with probability $1/2d$ independently of each other. This is clearly an irreducible chain, so the question is whether this chain is transient or not.

The story goes that Polya was taking a walk in the park in the early 1900's and saw the same couple pass by many times although it seemed to be the case that they were doing a "random walk". Polya investigated this problem and proved the amazing theorem

Theorem 2.17 (Polya)**.** *The simple random walk is recurrent if $d = 1, 2$ and transient for $d \geq 3$.*

We will prove this for $d = 1, 2, 3$ and leave the rest for you to compute at your leisure time.

$d = 1$ **Case:** This is just a simple random walk in the line. Recall that using Proposition 2.13, we only need to prove that $\sum_n P_{00}^n = \infty$. We do this via direct computation.

Note that if the walk comes back to 0 starting from 0 it needs to perform an even number of steps. Also if it does so in $2n$ steps, n steps need to be right and n to the left. The total number of such steps is $\frac{(2n)!}{n!n!}$ (think of arranging two types of objects with n objects of each type). Also each such path has probability $(\frac{1}{2})^{2n}$ (whatever the path is). Thus

$$P_{00}^{2n} = \frac{(2n)!}{n!n!}(\frac{1}{2})^{2n}.$$

Here we use Stirling's approximation

Proposition 2.18.

$$n! \sim n^{n+1/2} e^{-n} \sqrt{2\pi}$$

where $a_n \sim b_n$ *means that* $\frac{a_n}{b_n} \to 1$ *as* $n \to \infty$.

Proof. A probabilistic sketch can be found in a Remark in Pg 215 of the book (12th Edition). □

Using this formula (simply replacing the factorials by this expression), and after a whole lot of cancellation, we get that

$$P_{00}^{2n} = \frac{(2n)!}{n!n!}(\frac{1}{2})^{2n} \sim \frac{1}{\sqrt{\pi n}}.$$

We know that $\sum_n \frac{1}{n^\alpha} < \infty$ if and only if $\alpha > 1$ (If you don't remember this, recall the series chapter in calculus, one can show this by comparing with the integral $\int_1^\infty \frac{1}{x^\alpha} dx$) Thus

$$\sum_{n=0}^{\infty} P_{00}^{2n} = \infty.$$

and we infer that the random walk is recurrent in $d = 1$.

$d = 2$ **case:** We have the same calculation, but the walk can now go in four directions, left, right, up or down. Let's start the walk at $(0, 0)$. Also the walk returns to the origin if for some $0 \leq i \leq n$, the walk makes i steps to right and i steps to left and the remaining $n - i$ up steps and $n - i$ down steps. The total number of ways this is possible is (again think of arranging symbols)

$$\frac{(2n)!}{i!i!(n-i)!(n-i)!}$$

Each path has probability $\frac{1}{4^{2n}}$. Using Stirling

$$P_{(0,0),(0,0)}^{2n} = \sum_{i=0}^{n} \frac{(2n)!}{i!i!(n-i)!(n-i)!} \frac{1}{4^{2n}} = \sum_{i=0}^{n} (\binom{2n}{n})^2 \frac{1}{4^{2n}} = \frac{1}{\pi n}$$

which again sums to ∞ and hence the walk is recurrent.

$d = 3$ **case** A similar calculation gives $P_{2n}^{00} \leq \frac{1}{(\pi n)^{d/2}}$ which is summable if and only if $d \geq 3$.

2.5 Stationary distributions

We saw that a communicating class can either be recurrent or transient. We will now classify the recurrent chains into two types, positive and null recurrent. Recall that if a state i is recurrent, and

$$N_i = \text{time taken to return to } i$$

then $\mathbb{P}(N_i < \infty) = 1$. If the chain is transient then $\mathbb{P}(N_i < \infty) < 1$. For a recurrent Markov chain, we refine the question and ask whether $\mathbb{E}(N_i) < \infty$ or not. This will turn out to be equivalent to the question that if we have a recurrent Markov chain, the proportion of times it returns to i is 0 or not.

To that end, let for a state j

$$m_j = \mathbb{E}(N_j)$$

denote the mean return time.

Also note that the proportion of time spent on a state j can be written as

$$\frac{\sum_{k=0}^{n} 1_{X_k=j}}{n}.$$

Proposition 2.19. *If a Markov chain is irreducible and recurrent, then*

$$\mathbb{P}\left(\frac{\sum_{k=0}^{n} 1_{X_k=j}}{n} \to \frac{1}{m_j}\right) = 1.$$

Proof. Suppose the Markov chain starts from i. Let T_0 denote the time it takes for the chain to reach j. Suppose $T_1, T_2, \ldots$ denote the successive times it goes back to j. Then we need to compute

$$\lim_{n\to\infty} \frac{n}{T_0 + T_1 + \ldots + T_n} = \lim_{n\to\infty} \frac{1}{\frac{T_0}{n} + \frac{T_1+\ldots+T_n}{n}}.$$

Note $T_0/n \to 0$ almost surely. By strong law of large numbers $\frac{T_1+\ldots+T_n}{n} \to \mathbb{E}(N_j) = m_j$ almost surely. Thus we have that the limit is $1/m_j$ almost surely. □

We cheated a little bit in the above proof. Can you identify where and rectify it?

Definition 2.20. *We denote by*

$$\pi_j = \frac{1}{m_j}$$

to denote the long run proportion of times a Markov chain stays in state j. Here if $m_j = \infty$ we take $\pi_i = 0$.

Note that

Proposition 2.21. *If $\pi_j > 0$ and i communicates with j then $\pi_i > 0$*

Proof. The proof needs a certain input from Martingale theory which we need, called *Wald's identity* which we will use (but not prove). Let $Y_0, Y_1, \ldots, Y_n$ be the successive return times to j when started from j. Let

$$p = \mathbb{P}(X_1, X_2, \ldots \text{ hits } i \text{ before } j | X_0 = j).$$

Note since j communicates with i, $p > 0$. Let N be the smallest k such that i is hit between times Y_{k-1} and Y_k. Clearly N is Geometric(p). Then, if we let $\tau_{j,i}$ denote the expected time to hit i started from j, then

$$\mathbb{E}(\tau_{j,i}) \leq \mathbb{E}(Y_1 + \ldots + Y_N) = \text{ (by Wald's identity)} \quad \mathbb{E}(N)\mathbb{E}(Y_1) = \mathbb{E}(\tau_{j,j})/p < \infty$$

since $\mathbb{E}(\tau_{j,j}) = 1/\pi_j < \infty$ Also let n be such that $P^n_{ji} > 0$ and let A be the event that started from j, the chain hits i in *exactly* n steps. Then

$$\mathbb{E}(\tau_{j,j}) \geq \mathbb{E}(\tau_{j,j}|A)\mathbb{P}(A) = (n + \mathbb{E}(\tau_{i,j}))\mathbb{P}(A).$$

Since $\mathbb{E}(\tau_{j,j}) = 1/\pi_j$ is finite as j is positive, $\mathbb{E}(\tau_{i,j})$ is finite also from the above inequality. Finally

$$\tau_{i.i} \leq \tau_{i,j} + \tau_{j,i} \implies \mathbb{E}(\tau_{i.i}) \leq \mathbb{E}(\tau_{i,j}) + \mathbb{E}(\tau_{j,i}) < \infty.$$

□

Corollary 2.22. *Either $\pi_i > 0$ for all i in a communicating class OR $\pi_i = 0$ for all i in a communicating class. We say a class is null recurrent if the latter occurs and positive recurrent if the former occurs.*

If a Markov chain is irreducible, there is only one communicating class. If this class is positive recurrent or null recurrent, we say the Markov chain is positive recurrent or null recurrent respectively.

Corollary 2.23. *A finite, irreducible Markov chain is always positive recurrent.*

Proof. This is simply because if it is null recurrent, then $\pi_i = 0$ for all i. But π_i denotes the proportion of times the chain spends in a state i and hence $\sum_i \pi_i = 1$ which is impossible. □

But how do we compute π_j? It is usually not so easy to compute the expected return times. Luckily, we have the following way out.

Note that

$$\frac{\sum_{k=0}^{n} 1_{X_k=j}}{n} = \frac{1_{X_0=j} + \sum_{k=0}^{n-1}\sum_i 1_{X_k=i, X_{k+1}=j}}{n}$$

Taking Expectation and using the bounded convergence theorem,

$$\mathbb{E}(\frac{\sum_{k=0}^{n} 1_{X_k=j}}{n}) = \frac{\sum_{k=0}^{n}\mathbb{P}(X_k = j)}{n} \to \pi_j$$

and for the right hand side, exchanging the summation,

$$\begin{aligned}\frac{\sum_i\sum_{k=0}^{n-1}\mathbb{P}(X_k = i, X_{k+1} = j)}{n} &= \frac{\sum_i\sum_{k=0}^{n-1}\mathbb{P}(X_{k+1} = j|X_k = i)\mathbb{P}(X_k = i)}{n} \\ &= \frac{\sum_i\sum_{k=0}^{n-1}\mathbb{P}(X_k = i)}{n}P_{ij} \to \sum_i \pi_i P_{ij}.\end{aligned}$$

Thus we obtain the equation

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

In vector notation, let $\boldsymbol{\pi} = (\pi_1, \pi_2, \dots,)$ be the row vector. Then we have

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

where P is the transition matrix. Note that we still treat $\infty \times \infty$ matrices like finite matrices.

Note that $\boldsymbol{\pi}$ equal to the row vector of $0's$ is always a solution of the above equation. However we are interested in a solution such that $\sum_i \pi_i = 1$. Such a solution can only exist if and only if the chain is positive-recurrent as is clear from all the above analysis.

Theorem 2.24. *Suppose we are given an irreducible Markov chain. This Markov chain is positive recurrent if and only if there exists a unique solution of the following equation system*

$$\boldsymbol{\pi} P = \boldsymbol{\pi}., \qquad \sum_i \pi_i = 1$$

exists. Furthermore, if such a solution exists, it is given by $\pi_j = \frac{1}{m_j}$.

Consequently if the above equation system has no solution, then the Markov chain is either transient or null-recurrent.

One consequence of Theorem 2.24 is that if we can find a solution to the equation system given there, and the Markov chain is irreducible, then we are guaranteed that the chain is positive recurrent, and hence in particular recurrent. This is one cheap way to show that a Markov chain is positive recurrent.

Remark 2.25. Note that Theorem 2.24 says that if the MC has no solution, then the Markov chain in particular cannot have a stationary distribution. Furthermore, this theorem says nothing about the case when the chain is not irreducible. In that case, there could be multiple solutions to the equations, and is more complicated.

Definition 2.26. *Note that we can treat* $\boldsymbol{\pi}$ *as a probability mass function on the state space. When an unique solution to the equation in Theorem 2.24 exists, we say* $\boldsymbol{\pi}$ *is a stationary distribution of the Markov chain.*

Why stationary? Well stationary means something which does not change, and the object which is stationary here is the distribution of the individual random variables $X_0, X_1, \ldots$. To be more precise, suppose we pick $X_0 \sim \pi$. Then

$$\mathbb{P}(X_1 = i) = \sum_j \mathbb{P}(X_1 = i | X_0 = j)\mathbb{P}(X_0 = j) = \sum_j \pi_j P_{ij} = \pi_i = \mathbb{P}(X_0 = i).$$

That is X_0 and X_1 has the same distribution. Iterating, X_n and X_0 have the same distribution, which is given by $\boldsymbol{\pi}$.

Example 2.27 (Null recurrent chain)**.** Simple random walk in $\mathbb{Z}$ is null-recurrent. We will show this by comparing it with the walk in the cycle (a direct computation is possible, but we take a more elegant route). Recall the finite version, which is the cycle example a.k.a. Gambler's ruin problem (example 1.34). Let $m_{i,j}$ denote the expected time to hit j starting from i. Let m_{i,j,C_n} denote the expected time to hit j started from i in the cycle C_n and let $m_{i,j,\mathbb{Z}}$ be the same quantity for $\mathbb{Z}$. Note that

$$m_{0,0} = 1 + \frac{1}{2}m_{1,0,\mathbb{Z}} + \frac{1}{2}m_{-1,0,\mathbb{Z}}$$

Also note $m_{i,0,\mathbb{Z}} \geq m_{i,0,C_n}$ for any n. This is simply because the walks have the same distribution until either 0 or n is hit. When such an event occurs, 0 may not be hit in $\mathbb{Z}$ but it is hit in C_n, thus it takes longer to hit 0 for the walker in $\mathbb{Z}$. Now recall that $m_{i,0,C_n}$ was computed to be $i(n-i)$ in the Gamber's ruin problem. Thus, we get

$$m_{0,0,\mathbb{Z}} \geq 1 + (n-1) + (n-1) \text{for all } n \geq 1 \ .$$

Letting $n \to \infty$, on the right hand side, we get $m_{0,0} = \infty$.

Example 2.28 (Positive recurrent chain)**.** Consider a Markov chain with drift to the left but having a "wall" at 0 where it "bounces off".

Consider the state space is $\{0, 1, 2, \ldots\}$ and

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

We solve now for

$$\boldsymbol{\pi} P = \boldsymbol{\pi}., \qquad \sum_i \pi_i = 1$$

using Theorem 2.24. This gives

$$\begin{aligned} \pi_0 &= \pi_0 P_{00} + \pi_1 P_{10} = \pi_0(1-p) + \pi_1(1-p) \implies \pi_1 = \frac{p}{1-p}\pi_0 \\ \pi_1 &= \pi_0 P_{01} + \pi_2 P_{21} = \pi_0 p + \pi_2(1-p) \implies \pi_2 = (\frac{p}{1-p})^2 \pi_0 \end{aligned}$$

Iterating and using induction, we get (exercise)

$$\pi_k = (\frac{p}{1-p})^k \pi_0$$

Now using $\sum_i \pi_i = 1$,

$$\sum_{k=0}^{\infty} (\frac{p}{1-p})^k \pi_0 = 1 \implies \pi_0 = \frac{1-2p}{1-p}$$

which is strictly positive if and only if $p < 1/2$.

2.6 Limiting probabilities

We learnt that $\mathbb{P}(X_n = j | X_0 = i) = P_{ij}^{(n)}$. Note that as $n \to \infty$, this is simply a sequence of numbers between 0 and 1 (as they are probabilities). Does this sequence have a limit? If so what is it? We will answer these questions in this subsection.

Let us go back to the cycle example example 1.34 a.k.a. Gamblers ruin. Suppose we want to understand the sequence

$$\mathbb{P}(X_n \text{ is even } | X_0 = 0).$$

Does this sequence converge? Well, clearly it does NOT. Simply because $\mathbb{P}(X_n \text{ is even } | X_0 = 0) = 0$ if n is odd, and is 1 if n is even. So this sequence is simply the sequence $\{0, 1, 0, 1, \ldots\}$ which does not converge.

Therefore we need an extra assumption. Suppose we have a Markov chain when there is a state i where the chain can reach in only a multiple of d many steps. If $d > 1$ for some such state, we say the chain is *periodic*. Otherwise, (that is if for all states, $d = 1$) we say the chain is aperiodic. More precisely, let

$$\mathcal{Q}_j = \{n : P_{jj}^n > 0\}.$$

If $\mathcal{Q}_j = \{kd : k \geq 1\}$ for some $d > 1$ we say the state j is periodic. On the other hand if $d = 1$, the state is aperiodic. The smallest number d we can find is simply gcd $(\mathcal{Q}_j)$. It can also be shown that

Lemma 2.29. *Periodicity is a class property.*

Proof. Take i, j in a communicating class. By definition of communication, find $r, \ell > 0$ such that $P_{ij}^r > 0$ and $P_{ji}^\ell > 0$. Let $m = r + \ell$. Note that $\mathcal{Q}_i + m \subset \mathcal{Q}_j$ where $\mathcal{Q}_i + m$ is obtained by adding m to every element in $\mathcal{Q}_i$. Also $m \in \mathcal{Q}_i \cap \mathcal{Q}_j$. Thus if every element of $\mathcal{Q}_j$ is kd for some $d > 1$, every element of $\mathcal{Q}_i$ must also be kd for the same $d > 1$. Thus gcd $(\mathcal{Q}_j) \geq$ gcd $(\mathcal{Q}_i)$. Re applying the same argument, reversing the roles of i and j, we get gcd $(\mathcal{Q}_i) \geq$ gcd $(\mathcal{Q}_j)$. Thus gcd $(\mathcal{Q}_i) =$ gcd $(\mathcal{Q}_j)$ concluding the proof. □

Thus, as before, we can talk about an irreducible Markov chain being periodic or aperiodic.

For Markov chains, aperiodicity is simply a nuisance. one can make the chain aperiodic by modifying the transition probabilities as follows. Toss a fair coin independent of everything else. If heads comes, the chain stays put does not move. If tail comes, the chain moves according to the Markov chain. Let $X_0, X_1, \ldots$ denote the original chain and $\tilde{X}_0, \tilde{X}_1, \ldots$ denote the modified chain. Let us compute the new transition probabilities $\tilde{P}_{ij}$: of the modified chain. If $j \neq i$, then the coin toss must produce tails. Thus

$$\begin{aligned}\tilde{P}_{ij} = \mathbb{P}(\tilde{X}_1 = j | X_0 = i) = \mathbb{P}(X_1 = j, \text{ coin toss tails} | X_0 = i) \\ = \mathbb{P}(\text{ coin toss tails}) \mathbb{P}(X_1 = j | X_0 = i) = \frac{1}{2} P_{ij}.\end{aligned}$$

On the other hand if $i = j$ then two things can happen. Either the coin toss is heads, or the coin toss is tails and the chain moves from i to i. Thus

$$\begin{aligned}\tilde{P}_{ii} = \mathbb{P}(\tilde{X}_1 = i | X_0 = i) = \mathbb{P}(X_1 = i, \text{ coin toss tails} | X_0 = i) + \mathbb{P}(\text{ coin toss heads}) \\ = \frac{1}{2} P_{ii} + \frac{1}{2}.\end{aligned}$$

Check that the new matrix can be written as

$$\tilde{P} = \frac{1}{2} I + \frac{1}{2} P$$

were I is the identity matrix (with the usual interpretation of a matrix when we have infinitely many states).

Exercise 2.30. *Going back to the cycle example example 1.34, suppose we have a two cycle (i.e. with two vertices). Then make the chain lazy and compute P^n and its limit as $n \to \infty$.*

Another way to make the chain lazy is to take a coin with probability of heads $= p$ for any $p \in (0,1)$. Show that the transition probability matrix becomes:

$$\tilde{P} = pI + (1-p)P.$$

Do the above exercise when we make the chain lazy by tossing coin with $p \in (0,1)$ and $p \neq 1/2$.

What is the stationary distribution of the lazy chain?

Theorem 2.31. *An irreducible positive recurrent Markov chain converges to its stationary distribution π if and only if the Markov chain is aperiodic.*

We will not prove the theorem but argue only about the part where we show that if the limit exists, then the limit must be π.

Suppose

$$\alpha_j = \lim_{n\to\infty} \mathbb{P}(X_n = j)$$

exists for all j. Then

$$\mathbb{P}_{X_{n+1}=j} = \sum_i P_{ij}\mathbb{P}(X_n = i)$$

taking limits,

$$\alpha_j = \sum_i \alpha_i P_{ij}$$

Also $\sum_i \mathbb{P}(X_n = i) = 1$, thus taking limits, $\sum_i \alpha_i = 1$ Since the chain is irreducible, positive recurrent, the chain must have a unique solution π which means $\alpha_i = \pi_i$ for all i.

2.7 Time reversal

Suppose we show a simulation of a simple random walk. Can you tell whether the movie is going backward or forwards? If we run the movie backwards, do we still get a Markov chain? In this section, we answer these questions.

Proposition 2.32. *Running backwards a Markov chain gives us a Markov chain, meaning that if $X_0, X_1, \ldots,$ is a Markov chain, $(X_n, X_{n-1}, \ldots,)$ is also a Markov chain.*

Proof. We have to show that

$$\mathbb{P}(X_n = i | X_{n+1}, X_{n+2}, \ldots) = \mathbb{P}(X_n = i | X_{n+1})$$

That is conditioned on X_{n+1}, X_n is independent of $X_{n+2}, X_{n+3}, \ldots$. Said otherwise, we need to prove conditioned on X_{n+1}, $X_{n+2}, X_{n+3}, \ldots$ is independent of X_n (if A is independent of B, then B is independent of A). But this is true since X is a Markov chain. □

Suppose we start an irreducible, aperiodic Markov chain from stationarity (or "equivalently" start the chain after running it for a long time). Are we able to tell if it runs forward or backward? For example consider the simple random walk on a cycle example 1.34 and start at stationarity : $X_0 \sim Unif\{0, 1, \ldots, n-1\}$. A simple calculation shows that for any sequence $i_0, i_1, \ldots, i_n$ where $i_k = i_{k-1} \pm 1 \mod n$ (a sequence which the random walk can possibly take),

$$\mathbb{P}(X_0 = i_0, \ldots, X_n = i_n) = \frac{1}{n}\frac{1}{2^n} = \mathbb{P}(X_n = i_n, X_{n-1} = i_{n-1}, \ldots X_0 = i_0).$$

In other words, the reverse walk has the same distribution as the forward walk, hence it is impossible to tell whether the movie has run forward or backward.

But in general how can we tell? Let us find calculate the transition probability matrix Q of the reverse process.

$$Q_{ij} = \mathbb{P}(X_n = j | X_{n+1} = i) = \frac{\mathbb{P}(X_{n+1} = i, X_n = j)}{\mathbb{P}(X_{n+1} = i)} = \frac{\mathbb{P}(X_{n+1} = i | X_n = j)\mathbb{P}(X_n = j)}{\mathbb{P}(X_{n+1} = i)} = \frac{\pi_j P_{ji}}{\pi_i}$$

Exercise 2.33. *Show that Q_{ij} is a stochastic matrix, that is $\sum_j Q_{ij} = 1$ for all i and $Q_{ij} \geq 0$.*

Definition 2.34. *We say a Markov chain with stationary distribution π is reversible if*

$$Q_{ij} = P_{ij} \implies \pi_i P_{ij} = \pi_j P_{ji}. \text{ for all } i, j.$$

Example 2.35. Consider the following transition matrix

$$P = \begin{bmatrix} 0 & 0.1 & 0.9 \\ 0.9 & 0 & 0.1 \\ 0.1 & 0.9 & 0 \end{bmatrix}$$

One can check that the chain is simply a like a cycle with a bias in one direction. This should not be reversible as the chain makes more moves in one direction than the other (and hence one can guess that whether the movie of the chain is running backwards or forwards). One can easily calculate that $\pi_0 = \pi_1 = \pi_2 = \frac{1}{3}$ and this gives $Q_{ij} = P_{ji}$ from the formula. Thus the reverse chain matrix $Q = P^T$.

Is there a way to check reversibility without computing π_i ?

Theorem 2.36 (Kolmogorov's criterion)**.** *A stationary Markov chain for which $P_{ij} = 0$ whenever $P_{ji} = 0$, is time reversible if and only if starting from state i, any path back to i has the same probability as the reversed path, that is,*

$$P_{ii_1}P_{i_1i_2}\ldots P_{i_ki} = P_{ii_k}P_{i_ki_{k-1}}\ldots P_{i_1i}$$

for all $i, i_1, \ldots, i_k$.

Proof. Let $i_0 = i_{k+1} = i$ If the chain is reversible, we know $\pi_{i_j} P_{i_j i_{j+1}} = \pi_{i_{j+1}} P_{i_{j+1} i_j}$ multiplying this equation for all $0 \leq j \leq k$ we see that the π_{i_j}s cancel and we are left with the above equation. For the reverse, summing over all possible values of $i_1, i_2, \ldots, i_{k-1}$, we get using Chapman Kolmogorov that $P^{(k)}_{ii_k} P_{i_k i} = P_{ii_k} P^{(k)}_{i_k i}$. Keeping $i_k = j$ for some fixed j for all k and taking $k \to \infty$, we see that $\pi_j P_{ji} = \pi_i P_{ij}$ and we are done. □

Sometimes, computing stationary distributions can be problematic. One trick is to guess the reverse chain and this helps in computing stationary distributions.

Proposition 2.37. *Consider an irreducible Markov chain with transition probabilities P_{ij}. Suppose we can find positive numbers π_i summing to 1 and transition probability matrix Q such that*

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

then Q is the transition probability matrix of the reversed chain and π_i is the stationary distribution for both the forward and the reverse chain.

Proof. Sum over i, the left hand side. We get $\pi_j = \sum_i \pi_i P_{ij}$. There is a unique solution to this equation which is the stationary distribution by Theorem 2.24. Once we have this, $Q = P$ and the Markov chain is reversible. □

Example 2.38. Read the lightbulb example 4.40 (11th edition).

2.8 Branching processes

Branching processes are used to model the growth of a population. Suppose we start with a single individual of a certain specie who gives rise to Z many offsprings where Z has some pmf given by

$$\mathbb{P}(Z = i) = p_i \text{ for } i \geq 0.$$

Call this offsprings members of generation 1. Next, each offspring of generation 1 gives rise to a certain number of offsprings distributed as Z and these are independent of each other. Let X_n be the number of offsprings in the nth generation for $n \geq 0$ with $X_0 = 1$. This is clearly a Markov chain with state space being the space of natural numbers $\mathbb{N}$.

We are interested in the question: does the specie die out? If so, can we compute/estimate its probability?

Notice that the Markov chain is not irreducible, because once X_n hits 0, it remains at 0. Assume $p_0 > 0$ (since if $p_0 = 0$, the specie never die out for sure.) In this setup, 0 is a recurrent state and every other state is transient (check!).

We want to compute

$$\mathbb{P}(X_n = 0 \text{ for some } n \geq 1 | X_0 = 1).$$

A good way to approach this problem is analytically. Consider the following generating function of Z

$$p_Z(s) = \mathbb{E}(s^Z) = \sum_{n=0}^{\infty} p_n s^n.$$

Clearly, the radius of convergence R satisfies $R \geq 1$ as for any s with $|s| < 1$,

$$p_Z(s) \leq \sum_{n=0}^{\infty} p_n |s|^n \leq \sum_{n=0}^{\infty} p_n = 1.$$

Fact. Using Abel's theorem [8], one can show that $p_Z(s)$ satisfies

$$\lim_{s \to 1-} p_Z(s) = p_Z(1) = 1.$$

In particular, $p_Z(s)$ is continuous inside $[0, 1]$.

Just like moment generating functions, generating functions are very useful when adding independent random variables. Let $\varphi(s) = p_Z(s)$. if X, Y are i.i.d $\sim Z$, then,

$$p_{X+Y}(s) = \mathbb{E}(s^{X+Y}) = \mathbb{E}(s^X s^Y) = \mathbb{E}(s^X)\mathbb{E}(s^Y) = \varphi^2(s).$$

We used independence of X and Y in the third equality.

Ok now what is happening in the branching process at hand? Note that simply

$$p_{X_1}(s) = p_Z(s).$$

[8] see Wikipedia or google 'Abels theorem'.

But we can write

$$X_2 = \sum_{i=1}^{X_1} Z_{1i}$$

where Z_{1i} are i.i.d. and distributed as Z by the defnitiion of the branching process. Notice that the number of terms is random. But we can compute the generating function by conditioning on X_1. For each $|s| < 1$,

$$p_{X_2}(s) \;=\; \mathbb{E}(s^{X_2}) \;=\; \mathbb{E}(s^{\sum_{i=1}^{X_1} Z_{1i}}) \;=\; \mathbb{E}(\mathbb{E}(s^{\sum_{i=1}^{X_1} Z_{1i}} \;\mid\; X_1)) \;=\; \mathbb{E}(\varphi(s)^{X_1}) \;=\; \varphi \circ \varphi(s).$$

Iterating this, we can get

$$p_{X_n}(s) = \underbrace{\varphi \circ \varphi \circ \ldots \circ \varphi(s)}_{n \text{ times}}$$

What happens to $\mathbb{P}(X_n = 0)$ as n increases? Clearly if the population has died out in step n, $X_{n+1} = 0$ is trivially true. Thus $\{X_n = 0\} \subseteq \{X_{n+1} = 0\}$. Thus

$$\{\text{Population eventually dies }\} = \{X_n = 0 \text{ for some } n \geq 1\} = \cup_{n\geq 1}\{X_n = 0\} = \lim_{n\to\infty}\{X_n = 0\}.$$

Said otherwise, $\mathbb{P}(X_n = 0)$ is non-decreasing, therefore must have a limit. Let $d_n = \mathbb{P}(X_n = 0)$ and $d = \lim_{n\to\infty} d_n$. Clearly $d_n \in [0, 1]$ and hence so does d.

Let $\mu > 0$ be $\mathbb{E}(Z)$. Clearly,

$$\mathbb{E}(X_2) = \mathbb{E}(\sum_{i=1}^{X_1} Z_{1i}) = \mathbb{E}(\mathbb{E}(\sum_{i=1}^{X_1} Z_{1i}|X_1)) = \mathbb{E}(\mu X_1) = \mu^2.$$

Iterating, we can see that

$$\mathbb{E}(X_n) = \mu^n.$$

So if $\mu < 1$, $\mathbb{E}(X_n) \to 0$ and

$$\mathbb{E}(X_n) = \sum_{j\geq 1} j\mathbb{P}(X_n = j) \geq \sum_{j\geq 1} \mathbb{P}(X_n = j) = 1 - \mathbb{P}(X_n = 0)$$

Thus $\mathbb{P}(X_n = 0) \to 1$ since $\mathbb{E}(X_n) \to 0$. In other words,

Proposition 2.39. *If $\mu < 1$, the population dies out with prob 1.*

In fact it can be shown (but is not straightforward) that if $\mu = 1$, the population also dies out, and if $\mu > 1$, the population has a positive chance to survive. But we will not show these and actually argue how to explicitly compute the probabiliity of survival (which is very surprising at first glance.)

Note that by continuity of φ inside $[0,1]$,

$$\varphi(d_n) \to \varphi(d).$$

But

$$d = \lim_{n\to\infty} d_n = \lim_{n\to\infty} d_{n+1} = \lim_{n\to\infty} \underbrace{\varphi \circ \varphi \circ \ldots \circ \varphi(0)}_{n+1 \text{ times}} = \lim_{n\to\infty} \varphi(\underbrace{\varphi \circ \varphi \circ \ldots \circ \varphi(0)}_{n \text{ times}}) = \lim_{n\to\infty} \varphi(d_n) = \varphi(d).$$

Theorem 2.40. *The extinction probability of a branching process is given by the smallest positive solution of*

$$d = \varphi(d).$$

Thus by plotting $\varphi(s)$, once can find the extinction probability.
Note that φ must satisfy:

- $\varphi'(1) = \mathbb{E}(X)$,
- $\varphi(1) = 1$
- $\varphi(s) \geq 0$ for all s,
- $\varphi''(s) \geq 0$ for all $0 \leq s \leq 1$ (so φ is convex).

It is now a calculus exercise to convince yourself that if $\varphi'(1) > 1$, then $d^* < 1$ and when $\varphi'(1) < 1$ then $d^* = 1$ where d^* is the solution to the fixed point equation $\varphi(d) = d$.

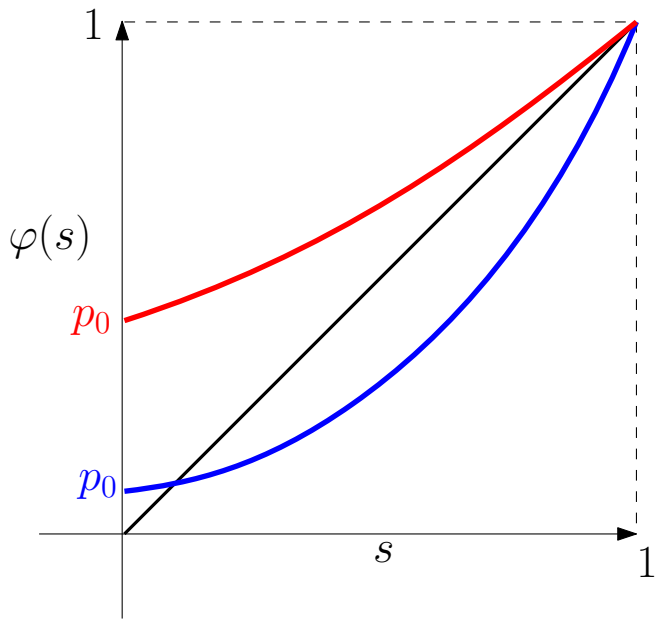


Figure 6: The red curve is a branching process which has extinction probability 1. For the blue one, the extinction probability can be calculated by computing the x-coordinate of the point of intersection of the blue curve with the black straight line $y = x$.

3 Poisson processes.

We learnt before that Binomial and Poisson distribution are intimately related. In particular if we have a $\text{Bin}(n, \lambda/n)$ then it is approximately $\text{Poisson}(\lambda)$. As it turns out this approximation is very useful and makes the math intricate. Here is another observation.

Lemma 3.1. *Let $X \sim$ Geom (λ/n). Then for all $x > 0$,*

$$\lim_{n\to\infty} \mathbb{P}(X/n > x) = e^{-\lambda x} = \mathbb{P}(Y > x).$$

where $Y \sim$ Exp(λ). In other words, X/n converges to an Exponential (λ) random variable (the proper sense of convergence is convergence in distribution.)

Proof.

$$\begin{aligned}\mathbb{P}(X/n > x) = \mathbb{P}(X > nx) &= \sum_{k=\lfloor nx\rfloor+1}^{\infty} (1-\lambda/n)^k \frac{\lambda}{n} \\ &= (1-\lambda/n)^{nx+1} \\ &\to e^{-\lambda x}\end{aligned}$$

□

Imagine dividing $[0,1]$ into n equal subintervals and then coloring each endpoint of the interval red with probability λ/n independently. Then if we start from 0 and look at the endpoints from left to right, what is the number of points must we look until we find a red point? It is equivalent to tossing i.i.d. coins with probability of heads being λ/n and then waiting for heads. The waiting time we know is Geom (λ/n). As $n \to \infty$, the endpoints of all intervals "converge to he continuum interval [0,1]" in some intuitive sense and by the above lemma, the interspacing between the red points are simply Exp(λ). On the other hand, the number of red points in an interval is simply $Poisson(\lambda)$ and the number of red points in two disjoint intervals are independent. Thus process is simply going to be the Poisson process, and this two different ways of looking at the same process is going to be very important in crunching out the math.

3.1 Properties of Exponential random variable

It is now hopefully clear that Exponential random variable plays a crucial role in the study of Poisson processes. Let us list some properties of exponential random variable. Let $X \sim$ Exp(λ).

a. First of all recall the density

$$f_X(t) = \begin{cases} \lambda e^{-\lambda x} \text{ when } x \geq 0 \\ 0 \text{ when } x < 0. \end{cases}$$

and the cdf

$$F_X(x) = \mathbb{P}(X \leq x) = \begin{cases} 1 - e^{-\lambda x} \text{ when } x \geq 0 \\ 0 \text{ when } x < 0. \end{cases}$$

Or in other words,

$$1 - F_X(x) = \mathbb{P}(X > x) = \begin{cases} e^{-\lambda x} \text{ when } x \geq 0 \\ 0 \text{ when } x < 0. \end{cases}$$

b. $\mathbb{E}(X) = \frac{1}{\lambda}$. $\text{Var}(X) = \frac{1}{\lambda^2}$.

c. (Memoryless property) Exponential distribution has "no memory". This means that if someone tells you that $X > s$ then the conditional distribution of X conditioned on $X > s$ is the same as $s + X$. Let us see why.

$$\mathbb{P}(X > s + t | X > s) = \frac{\mathbb{P}(X > s + t)}{X > s} = \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} = e^{-\lambda t}$$

and

$$\mathbb{P}(s + X > s + t) = \mathbb{P}(X > t) = e^{-\lambda t}$$

which are the same.

Exercise 3.2. *In fact, Exponential distribution is the* only *continuous distribution having the memoryless property. In particular, show that if* $G(x) = 1 - F_X(x)$*, for some continuous random variable* X *so that* $\mathbb{P}(X > 0) = G(0) = 1$ *then*

$$G(x) = G(1)^x. \text{ for all} x > 0.$$

Conclude that $X \sim$ *Exp* $(-\ln(G(1)))$.

Exercise 3.3. *Suppose* X *is a distrete random variable taking values in* $\mathbb{N}$*. Suppose it is memoryless, that is,*

$$\mathbb{P}(X > m + n | X > m) = \mathbb{P}(X > n).$$

for all $m, n \in \mathbb{N}$*. Show that this must be a geometric random variable.*

d. If $X_1, \ldots, X_n$ are i.i.d. Exp (λ) then

$$X_1 + \ldots + X_n \sim Gamma(n, \lambda)$$

where Gamma(n, λ) has the density

$$f(t) = \begin{cases} \lambda e^{-\lambda t} \frac{(\lambda t)^{n-1}}{(n-1)!} \text{ if } t \geq 0 \\ 0 \text{ if } t < 0. \end{cases} \tag{3.1}$$

Exercise 3.4. *Show this using moment generating functions of* $Exp(\lambda)$ *and* $\Gamma(n, \lambda)$ *and use the fact that adding independent random variables correspond to multiplying the moment generating functions.*

e. Suppose $X_1, X_2, \ldots, X_n$ be independent with $X_i \sim \text{Exp}(\lambda_i)$ then

$$Y = \min\{X_1, \ldots, X_n\} \sim Exp(\lambda_1 + \ldots + \lambda_n).$$

Exercise 3.5. *Show this by writing*

$$\mathbb{P}(Y > x) = \mathbb{P}(min\{X_1, \ldots, X_n\} > x) = \mathbb{P}(X_1 > x, X_2 > x, \ldots, X_n > x).$$

f. Suppose $X_1, X_2, \ldots, X_n$ be independent with $X_i \sim \text{Exp}(\lambda_i)$. Then

$$\mathbb{P}(X_i = \min\{X_1, \ldots, X_n\}) = \frac{\lambda_i}{\sum_{k=1}^n \lambda_k}.$$

Show the above as follows. First, suppose $n = 2$. Then

$$\mathbb{P}(X_1 = \min\{X_1, X_2\}) = \mathbb{P}(X_1 < X_2)$$
$$= \int_0^\infty \int_0^v \lambda_1 e^{-\lambda_1 u} \lambda_2 e^{-\lambda_2 v} du dv = \int_0^\infty (1 - e^{-\lambda_1 v}) \lambda_2 e^{-\lambda_2 v} dv = 1 - \frac{\lambda_2}{\lambda_1 + \lambda_2} = \frac{\lambda_1}{\lambda_1 + \lambda_2}.$$

For the general case, note

$$\mathbb{P}(X_i = \min\{X_1, \ldots, X_n\}) = \mathbb{P}(X_i < \min_{j \neq i} X_j)$$

and then use the fact that X_i and $\min_{j \neq i} X_j$ are independent exponential with parameters λ_i and $\sum_{j \neq i} \lambda_j$ respectively using property e.

3.2 Definition of Poisson process

Recall the intuition at the beginning of the section which gives us the description of the Poisson process, namely we have a clock which rings after exponential (λ) amount of time and the intervals between successive rings are independent. Then if we *count* the *number of rings* then we get a function $(N(t))_{t\geq 0}$ which simply gives us the number of rings which have occurred up to time t.

In mathematics, when setting up one likes to come up with a general, compact and abstract definition of the process we care about and start with that to build the theory . We will do exactly this here.

So FORGET what has happened so far and we start fresh with the following definition. We will see later that how all the intuition that we have built so far will fall into place under this few lines (very similar to poetry).

Definition 3.6. *A stochastic process $\{N(t) : t \geq 0\}$ is said to be a counting process if it is non-negative integer valued, non-decreasing and for $s \leq t$, $N(t) - N(s)$ counts the number of "events" which happen between times s and t.*

Some examples of counting process:

- Number of cars crossing an intersection.
- Number of people born up to time t since the beginning of time.
- Number of tweets by Trump up to time t with $t = 0$ being the start of his presidency.
- Number of people in a bus is at time t is *NOT* a counting process (since it might decrease.)

Definition 3.7. *A function $(N(t))_{t\geq 0}$ is a Poisson process with rate λ if the following holds:*

(i) $N(0) = 0$

(ii) $\{N(t) : t \geq 0\}$ *has independent increments, meaning that the number of events that occur in disjoint intervals of times are independent. More precisely, if $[s_1, t_1], [s_2, t_2], \ldots, [s_n, t_n]$ are disjoint intervals then $(N(t_1) - N(s_1)), \ldots, N(t_n) - N(s_n)$ are independent.*

(iii) For all $t \geq 0$,

$$\mathbb{P}(N(t + h) - N(t) = 1) = \lambda h + o(h).$$

Here $o(h)$ is a function $f(h)$ so that

$$\lim_{h\to 0} \frac{f(h)}{h} = 0.$$

For example, $f(h) = h^2$ is $o(h)$, but $f(h) = h$ or $f(h) = \log(h)$ is not.

(iv) $\mathbb{P}(N(t+h) - N(t) \geq 2) = o(h)$. *This tells two events occurring in a small interval is much smaller than the length of the interval.*

A few words about the $o(h)$ notation. If two functions are $o(h)$ then their sum or differences or constant multiples are also $o(h)$. For example, if $f(h) = o(h)$ and $g(h) = o(h)$, then for some constant $c \in \mathbb{R}$,

$$\lim_{h\to 0} \frac{f(h) + g(h)}{h} = 0 \text{ and } \lim_{h\to 0} \frac{cf(h)}{h} = 0.$$

Lemma 3.8. *If we start counting at any time $s > 0$, then what we get is also a Poisson process, namely*

$$\{N_s(t) : t \geq 0\} = \{N(t+s) - N(s) : t \geq 0\},$$

is a Poisson process.

Proof. Check that all the axioms in Definition 3.18 are satisfied. □

Now let us check how to get back that the distribution of the intervals between the events are i.i.d. exponential. Let us start with the first interval:

$$T_1 : \min\{t \geq 0 : N(t) = 1\}.$$

Proposition 3.9. $T_1 \sim Exp(\lambda)$, *that is,*

$$\mathbb{P}(T_1 > t) = e^{-\lambda t}.$$

Proof. Note

$$\{T_1 > t\} = \{N(t) = 0\} \implies \mathbb{P}(T_1 > t) = \mathbb{P}(N(t) = 0)$$

Now let us analyze the function

$$P_0(t) = \mathbb{P}(N(t) = 0)$$

We want to understand $P_0(t+h) - P_0(t)$ for small h as that will give us the rate at which this function is decreasing (why is this function non-increasing). Note

$$\begin{aligned} P_0(t+h) &= \mathbb{P}(N(t+h) = 0) \\ &= \mathbb{P}(N(t+h) - N(t) = 0, N(t) = 0) \\ &= \mathbb{P}(N(t+h) - N(t) = 0)P_0(t) \\ &= P_0(t)(1 - \lambda h + o(h)). \end{aligned}$$

Thus

$$P_0(t+h) - P_0(t) = -P_0(t)(\lambda h + o(h)) \implies \frac{P_0(t+h) - P_0(t)}{h} = -P_0(t)(\lambda + \frac{o(h)}{h}).$$

Taking limits as $h \to 0$, the left hand side converges to $P_0'(t)$ and the right hand side converges to $-P_0(t)\lambda$. Thus overall, we have

$$P_0'(t) = -P_0(t)\lambda.$$

This is an ODE, which we can solve to get

$$\frac{P_0'(t)}{P_0(t)} = -\lambda \implies \int_0^s \frac{d(P_0(t))}{P_0(t)} = -\lambda \int_0^s dt \implies \ln(P_0(s)/P_0(0)) = -\lambda s.$$

But $P_0(0) = \mathbb{P}(N(0) = 0) = 1$. Thus exponentiating,

$$P_0(s) = e^{-\lambda s}, \qquad s \geq 0$$

which is what we want. □

Now let us look at the other intervals between events. To that end, define

$$T_n = \text{ time interval between the } n-1\text{th and } n\text{th event.}$$

Since given $T_1 = s$, $\{T_2 > t\}$ is simply describing the situation when no event has occurred in $(s, s+t]$ which is independent of what has happened in the interval $[0, s)$. Also since $N_s(t)$ is also a Poisson process with the same rate (Lemma 3.8), T_2 has the same distribution as T_1. This shows that

Proposition 3.10. *$T_1, T_2, \ldots$ are i.i.d. Exp(λ)*

For those of you with a rigorous bent of mind, this might seem a bit "hand-wavy". However, this can be made completely rigorous.

Now, how about the number of events occurring in an interval. If our intuition was correct, this must be Poisson(λ). Let us show this.

Proposition 3.11. *Let $(N(t))_{t\geq 0}$ be a Poisson process with rate λ. Then*

$$\mathbb{P}(N(t) = n) = \frac{e^{-\lambda t}(\lambda t)^n}{n!}$$

or, said otherwise, $N(t) \sim$ Poisson (λt).

Proof. Let $T_1, T_2, \ldots$ be the intervals between events and we know from Proposition 3.10 that they are i.i.d. Exp(λ). Let

$$S_n = T_1 + T_2 + \ldots + T_n.$$

By (3.1), we know that S_n has density of a Gamma (n, λ) random variable:

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

Note that $N(t) = n$ means that $S_n < t$ and $T_{n+1} > t - S_n$. Thus we can condition on S_n and write

$$\mathbb{P}(N(t) = n) = \int_0^t \mathbb{P}(N(t) = n | S_n = s) \lambda e^{-\lambda s} \frac{(\lambda s)^{n-1}}{(n-1)!} ds$$

But now,

$$\mathbb{P}(N_t = n | S_n = s) = \mathbb{P}(T_{n+1} > t - s) = e^{-\lambda(t-s)}$$

since the increments are independent and T_{n+1} follows an exponential (λ) distribution. Plugging this back,

$$\int_0^t e^{-\lambda(t-s)} \lambda e^{-\lambda s} \frac{(\lambda s)^{n-1}}{(n-1)!} ds = \frac{e^{-\lambda t} \lambda^n}{(n-1)!} \int_0^t s^{n-1} ds = \frac{e^{-\lambda t} \lambda^n}{(n-1)!} \frac{t^n}{n} = \frac{e^{-\lambda t} (\lambda t)^n}{(n)!}$$

as desired. □

Summary and properties of Poisson processes For a Poisson process with rate λ,

- The number of "events" in an interval $(s, s+t)$ is distributed as Poisson(λt).
- the intervals between events are i.i.d. Exp(λ).
- Total time elapsed until n events is Gamma(n, λ).
- For any interval (s, t), $N(s+t) - N(s) \sim$ Poisson(λt). Note that this denotes the number of "events" in an interval and only depends upon the length of the interval and not the location (think of the analogy of colorings of intervals again given at the beginnning of the section and it will be clear why this is true.) Such a property is called *stationary increments*.

Example 3.12. Suppose that people arrive into a store according to a Poisson process with rate 5 per day.

1. What is the probability that exactly 10 people arrive in 2 days?
2. What is the probability that no one arrives for the first 3 days?

Solution. For the first one: $\mathbb{P}(N(2) = 10) = e^{-10} \frac{(10)^{10}}{10!}$ (probabiliy that a poisson with $\lambda = 2 \times 5 = 10$ is equal to 10) Second one: $\mathbb{P}(T_1 > 3) = e^{-5 \times 3} = e^{-15}$.

Poisson thinning. Suppose $N(t)$ is a Poisson process with rate λ. Now suppose for each "event", we color it red with prob p and blue with prob $1 - p$ independently. Let $N_1(t)$ be the number of red events and $N_2(t)$ be the number of blue events up until time t. Clearly $N_1(t) + N_2(t) = N(t)$.

Proposition 3.13. *We have $(N_1(t)$ and $N_2(t)$ are independent Poisson processes with rates λp and $\lambda(1-p)$.*

Proof. First we show that $N_1(t)$ is a Poisson process, and the way to do it is to check the axioms. Clearly $N_1(t) \geq 0$ and has independent increments, as the number of red events in disjoint intervals can be obtained by first conditioning on the number of events in the intervals and then coloring them red or blue independently. To be more precise, take intervals (s,t) and (s',t') and then the number of red points in (s,t) and (s',t') are simply $Bin(N(s+t)-N(s),p)$ and $Bin(N(s'+t')-N(s'),p)$ where the Binomials are obtained by independent coin tosses. Thus they are independent. Finally, we show the third property for $t=0$, without loss of generality to make the notations less heavy

$$\begin{aligned}\mathbb{P}(N_1(h)=1) &= \mathbb{P}(N(h)=1, N_1(h)=1) + \mathbb{P}(N(h)=1, N_1(h)\geq 2)\\ &= \mathbb{P}(N_1(h)=1|N(h)=1)\mathbb{P}(N(h)=1) + \mathbb{P}(N_1(h)=1|N(h)\geq 2)\mathbb{P}(N(h)\geq 2))\\ &= p(\lambda h + o(h)) + o(h)\\ &= p\lambda h + o(h).\end{aligned}$$

Also

$$\mathbb{P}(N_1(h)\geq 2) \leq \mathbb{P}(N(h)\geq 2) = o(h).$$

Thus $N_1(t)$ is a Poisson process with rate λp and exactly for a similar reason $N_2(t)$ is a Poisson process with rate $\lambda(1-p)$. Recall here that to prove $(N_i(t))_{t\geq 0}$ for $i \in \{1,2\}$ are independent as processes, we need to show that for any $t_1,\ldots,t_k$, the joint distribution of $(N_1(t_1), N_1(t_2),\ldots,N_1(t_k))$ and $(N_2(t_1), N_2(t_2),\ldots,N_2(t_k))$ are independent. To that end, since $N_1(t') - N_1(t)$ is independent of the 'events' $(N(t))_{t\geq 0}$ outside the interval (t,t') it is enough to prove that $N_1(t') - N_1(t)$ is independent of $N_2(t') - N_2(t)$ (convince yourself that this is the case). To prove the latter, go back to Exercise 1.42 and example 1.41. □

Example 3.14. Suppose that people arrive into security check in an airport rate 5 per minute. Suppose each person is given a random special security check by tossing a coin with prob 0.2

1. What is the probability that exactly 10 people were given the random special check in 30 mins?
2. What is the probability that no one is given the random special security check in 15 mins?

Solution. The number of people getting security checks is a Poisson process with rate $5 \times 0.2 = 1$. Call it $N_1(t)$. For Thus we want $\mathbb{P}(N_1(30)=10) = e^{-30}\frac{(30)^{10}}{10!}$. Second one: $\mathbb{P}(T_1 > 15) = e^{-15}$.

This has some cool applications.

Exercise 3.15. *Read example 5.17 (10th and 11th edition)*

3.3 Conditional distribution of interarrival times

Suppose we condition on $N(t) = 1$. This simply means that there is one "event" before time t. What is the conditional distribution of this time? The interval coloring coin flip will lead us to guess that this is uniform in $(0, t)$. This is the content of the next proposition.

Proposition 3.16. $\mathbb{P}(T_1 < s | N(t) = 1) = \frac{s}{t}$.

Proof.

$$\begin{aligned}
\mathbb{P}(T_1 < s | N(t) = 1) &= \frac{\mathbb{P}(T_1 < s, N(t) = 1)}{\mathbb{P}(N(t) = 1)} \\
&= \frac{\mathbb{P}(\text{ 1 event in } [0, s), 0 \text{ events in } [s, t))}{\mathbb{P}(N(t) = 1)} \\
&= \frac{e^{-\lambda s} \lambda s e^{-\lambda(t-s)}}{\lambda t e^{-\lambda t}} \\
&= \frac{s}{t}
\end{aligned}$$

□

Now what if we condition on $N(t) = n$? It can be shown that they behave like "order statistics" of i.i.d. Uniform random variables. Suppose $X_1, X_2, \ldots,$ are i.i.d. Uniform$[0, t]$. Let $X_{(1)}$ be the minimum of them, $X_{(2)}$ be the second minimum and so on. This defines what is called the *order statistics*:

$$X_{(1)} < X_{(2)} < \ldots < X_{(n)}.$$

What is the joint density of order statistics? The density of order statistics is given by

$$n! f(x_1) f(x_2) \ldots f(x_n).$$

where f is the density of X_1. Roughly, for any $x_1 < x_2 < \ldots < x_n$,

$$\text{“}\mathbb{P}(X_{(1)} = x_1, X_{(2)} = x_2, \ldots, X_{(n)} = x_n)\text{”} = \sum_{\pi} \text{“}\mathbb{P}(X_{\pi(1)}) = x_1, \ldots, X_{\pi(n)} = x_n)\text{”} = n! f(x_1) f(x_2) \ldots f(x_n).$$

where $\pi : \{1, \ldots, n\} \to \{1, \ldots, n\}$ is a permutation. The quotation marks need justification, but we skip it here.

Using the above idea, here is a proposition.

Proposition 3.17. *The joint distribution of $(T_1, T_1 + T_2, \ldots, T_1 + \ldots + T_n)$ conditional on $N(t) = n$ is given by order statistics of i.i.d. uniform $[0, t]$ random variable. In other words, we have the joint density:*

$$f(t_1, \ldots, t_n) = \frac{n!}{t^n} \text{ if } 0 < t_1 < \ldots < t_n < t, \text{ and 0 otherwise.}$$

3.4 Non-homogeneous Poisson processes

The only difference between a non-homogeneous Poisson process and a homogeneous Poisson process is that the rate λ is now a function of t. So we have a function $\lambda(t) : \mathbb{R}_+ \to \mathbb{R}_+$, which is usually called the rate function.

Definition 3.18. *A function $(N(t))_{t\geq 0}$ is a **non-homoegeneous Poisson process** with rate function or **intensity function** or simply **intensity** $\lambda(t)$ if the following holds:*

(i) $N(0) = 0$

(ii) $\{N(t) : t \geq 0\}$ *has independent increments, meaning that the number of events that occur in disjoint intervals of times are independent. More precisely, if $[s_1, t_1], [s_2, t_2], \ldots, [s_n, t_n]$ are disjoint intervals then $(N(t_1) - N(s_1)), \ldots, N(t_n) - N(s_n)$ are independent.*

(iii) For all $t \geq 0$,

$$\mathbb{P}(N(t + h) - N(t) = 1) = \lambda(t)h + o(h).$$

Here $o(h)$ is a function $f(h)$ so that

$$\lim_{h\to 0} \frac{f(h)}{h} = 0.$$

For example, $f(h) = h^2$ is $o(h)$, but $f(h) = h$ or $f(h) = \log(h)$ is not.

(iv) $\mathbb{P}(N(t + h) - N(t) \geq 2) = o(h)$. *This tells two events occurring in a small interval is much smaller than the length of the interval.*

Recall that the number of events occurring in a homogeneous Poisson process in an interval (s, t) was a Poisson random variable with mean $\lambda(t - s) = \int_s^t \lambda dy$. Here, it will be the integral of a function $\lambda(t)$ via almost the same calculations as before. We define

$$m(t) = \int_0^t \lambda(t)dt.$$

to be the mean value function. We summarize the properties again here. You can try to prove them yourselves or consult the book, but really the calculations are almost exactly the same.

Suppose $N(t)$ is a non-homogeneous Poisson process with rate function $\lambda(t)$.

- $\{N(s + t) - N(s) : t \geq 0\}$ is another non-homogeneous Poisson process with rate $\{\lambda(s + t) : t \geq 0\}$.
- $N(t) \sim$ Poisson $(\int_0^t \lambda(u)du) =$ Poisson $(m(t))$. Similarly for $0 < s < t$, $N(t) - N(s) \sim$ Poisson $(\int_s^t \lambda(u)du)$. Note that this process does NOT have stationary increments (unless of course $\lambda(t)$ is the constant function and we are back to the homogeneous case).

- Note
$$\mathbb{P}(T_1 > t) = \mathbb{P}(N(t) = 0) = e^{-m(t)}$$
- Here the time gaps between events will be independent but may NOT have the same distribution.

Example 3.19. Suppose that the number of arrivals of students into the university of Victoria is given by the following funcion

$$\lambda(t) = \begin{cases} t^2 \text{ per hour for } t \text{ between 7 am to 10 am} \\ t \text{ per hour for } t \text{ between 10 am and 2 pm} \\ 0 \text{ for other times.} \end{cases}$$

What is the probability that total number of students arriving between 8 am and 11 am is 100? What is the covariance between the number of students arriving between 8 am and 10 am and 9 am and 11 am?

Solution. The number of students is Poisson with mean given by

$$\int_8^{11} \lambda(u)du = \int_8^{10} u^2 du + \int_{10}^{11} u du = \frac{10^3 - 8^3}{3} + \frac{11^2 - 10^2}{2} := A$$

So, the required probability is

$$e^{-A}\frac{A^{100}}{100!}$$

By independent increments, the covariance is simply the variance of the number of arrivals in the interval $(10, 11)$. The number of students arriving in that interval is Poisson with mean

$$\int_9^{10} u^2 du = \frac{10^3 - 9^3}{3}$$

Thus the variance of the number of students arriving in the interval $(9, 10)$ is also $\frac{10^3-9^3}{3}$.

We now present a proposition without proof related to Poisson thinning but in the inhomogeneous setup.

Suppose $N(t)$ is a Poisson process with rate λ. Now suppose for each "event" which occurs at time s, we color it red with prob $p(s)$ and blue with prob $1 - p(s)$ independently for some function $p : [0, \infty) \to [0, 1]$. Let $N_1(t)$ be the number of red events and $N_2(t)$ be the number of blue events up until time t. Clearly $N_1(t) + N_2(t) = N(t)$.

Proposition 3.20. *We have $(N_1(t)$ and $N_2(t)$ are independent inhomogenous Poisson processes with rate functions $(\lambda p(s))_{s\geq 0}$ and $(\lambda(1 - p(s)))_{s\geq 0}$.*

Example 3.21. Read the infinite server Poisson queue, Example 5.25 from Ross. (Editions 10 and 11)

4 Continuous time Markov chain

Suppose $(X_n)_{n\in\mathbb{N}}$ be a discrete time Markov chain and for each state i assign a parameter $\nu_i > 0$ and assume $P_{ii} = 0$ for all $i \in \mathcal{S}$. A continuous time Markov chain is a process $(X(t))_{t\geq 0}$ obtained from the discrete time chain $(X_n)_{n\in\mathbb{N}}$ by waiting independent Exponential (ν_i) amount of time to jump if the chain is at state i. Note that by memoryless property of Exponential,

$$\mathbb{P}(X_{t+s} = j | X(s) = i, (X(u))_{0\leq u<s}) = \mathbb{P}(X_t = j | X(0) = i) \tag{4.1}$$

Indeed, if we know the history of the process up to time s and the chain is at state i and suppose the jump to i occured at time $s' < s$, then the exponential random variable used to wait at state i in the last step is conditioned to be larger than $s - s'$. But by memoryless property of exponential random variables, the conditional distribution of this waiting time has the same distribution as $s - s'$+Exponential(ν_i). Thus starting to count time at $s - s'$, and since the probability of jumping between states is given by a discrete time Markov chain to begin with, we arrive at the right hand side.

4.1 Birth and death process

Now we study a very important continuous time Markov chain which is called the **birth and death process** or **birth and death chain**. Here the process is defined in continous time, and the embedded discrete time chain can be derived from it. Suppose we are studying the population size of a species. If there are n individuals (for some $n \geq 0$), new individual arrives after exponential (λ_n) amount of time, for some $\lambda_n > 0$. On the other hand, if there are $n \geq 1$ many individuals, one of the organisms die after Exponential (μ_n) amount of time, independent of everything else. We can now conclude the following.

- The waiting time when there are n individuals is Exponential with parameter $\nu_n := \lambda_n + \mu_n$ with $\mu_0 = 0$. This follows from property e. in Section 3.1.
- For the embedded discrete time chain, the total number of individuals can change by ± 1 with transition probabilities:

$$P_{i,i+1} = \frac{\lambda_i}{\lambda_i + \mu_i} = 1 - P_{i,i-1}.$$

  This follows from property f. in Section 3.1.

We now look at several special cases.

Poisson process. If we take $\lambda_n = \lambda$ for all n and $\mu_n = 0$ for all n, then we get back a Poisson process with rate λ.

Yule process. Suppose each organism is immortal and they come with a Poisson clock with rate λ, which when rings gives birth to a new individual, independently of each other. In this case, when there are n individuals, the amount of time we need to wait is Exponential with parameter $n\lambda$, by applying property $f.$ in Section 3.1. Since nobody dies, this is a birth and death process with $\lambda_n = n\lambda$ and $\mu_n = 0$ for all n. This is called the **Yule process**. We will study a more general process below, which will yield the Yule process as a special case.

Population model with immigration Suppose each individual has two independent Poisson clocks, one *birth clock* with rate $\lambda > 0$ and another *death clock* with rate $\mu > 0$. Also, *immigration* occurs according to an independent Poisson process with constant rate θ. Using item e. of Section 3.1, we can frame this as a birth and death process with

$$\lambda_n = n\lambda + \theta \text{ for } \quad n \geq 0; \qquad \mu_n = n\mu \text{ for } \quad n \geq 1.$$

Let $X(t)$ denote the population size at time t. We will compute

$$M(t) = \mathbb{E}(X(t))$$

Suppose we condition on $X(t) = k$. Then in the next jump time, the population size changes to $k + 1$ if one of the birth clocks of the k individuals or the immigration clock rings, and none of the death clocks ring. We observe that if $X \sim$Exponential (a) then for small values of $h > 0$,

$$\mathbb{P}(X < h) = 1 - e^{-ah} = ah + o(h), \text{ and hence } \mathbb{P}(X \geq h) = 1 - ah + o(h).$$

Also if we have m clocks with rates $a_1, a_2, \ldots, a_m$ then the probability that at least two rings occur in a small interval of time is

$$\mathbb{P}(N(h) \geq 2) = o(h).$$

Here $(N(t))_{t\geq 0}$ is a Poisson process with rate $a_1 + \ldots + a_k$. The last two observations are essentially reverse engineering to get back item (iii),(iv) in the definition of a Poisson process with rate a. Using this we compute

$$\begin{aligned}
&\mathbb{P}(X(t+h) - X(t) = 1 | X(t) = k) \\
&\qquad = \mathbb{P}(\text{one of the birth or immigration clocks ring and no death clock rings}) \\
&\qquad\qquad + \mathbb{P}(\text{at least 2 rings}).
\end{aligned}$$

Using the two observations above, we see that

$$\begin{aligned}
\mathbb{P}(X(t+h) - X(t) = 1 | X(t) = k) &= (k\lambda + \theta)h(1 - k\mu h) + o(h) = (k\lambda + \theta)h + o(h) \\
\mathbb{P}(X(t+h) - X(t) = -1 | X(t) = k) &= \mu k h + o(h) \\
\mathbb{P}(X(t+h) - X(t) = 0 | X(t) = k) &= 1 - (k\lambda + \theta + k\mu)h + o(h)
\end{aligned}$$

where the second equation above has similar reasoning as the first one. Thus we get the following equation for $M(t+h) - M(t)$:

$$\mathbb{E}(X(t+h) - X(t)|X(t) = k) = (k\lambda + \theta - k\mu)h + o(h)$$

which yields

$$\mathbb{E}(X(t+h)|X(t)) = X(t) + (\lambda - \mu)X(t)h + \theta h + o(h)$$

Taking expectation again,

$$M(t+h) - M(t) = (\lambda - \mu)M(t)h + \theta h + o(h)$$

Dividing by h and letting $h \to 0$, we get

$$M'(t) = (\lambda - \mu)M(t) + \theta$$

with initial value $M(0) = i$.

Let $\lambda \neq \mu$. Then Solving this ODE (substitute $f(t) = (\lambda - \mu)M(t) + \theta$), we get that

$$M(t) = \frac{\theta}{\lambda - \mu}(e^{(\lambda-\mu)t} - 1) + ie^{(\lambda-\mu)t}$$

Thus as $t \to \infty$ $M(t) \to \infty$ if $\lambda > \mu$ and $M(t) \to \frac{\theta}{\mu-\lambda}$ if $\mu > \theta$ (there is an equilibrium population). If $\mu = \lambda$, then the ODE is $M'(t) = \theta$ which yields

$$M(t) = \theta t + i.$$

4.2 Transition probabilities

How to compute transition probabilities in this setup. Transition probabilities are the quantities

$$P_{ij}(t) := \mathbb{P}(X(t) = j \mid X(0) = i)$$

for all i, j in the state space. We have to turn to Calculus and find out ODEs for these functions.

Proposition 4.1 (Chapman Kolmogorov)**.** *We have for all $s, t > 0$*

$$P_{ij}(s+t) = \sum_{k \in \mathcal{S}} P_{ik}(s)P_{kj}(t).$$

Proof. This follows exactly in the same way as in the discrete case using the Markov property of Continuous time Markov chains (4.1). □

We now introduce two notations. Let

$$q_{ij} = \nu_i P_{ij}; \text{ for all } i, j \in \mathcal{S}.$$

This quantity heuristically denotes the 'rate' at which, when a chain is in a state i moves to state j: ν_i is the rate at which it moves to some state, and when it does, it moves to j with probability P_{ij}. The following lemma justifies this:

Lemma 4.2. *We have for all i, j and $t > 0$,*

$$\lim_{h\to 0} \frac{P_{ij}(h)}{h} = q_{ij} \text{ for all } i \neq j$$
$$\lim_{h\to 0} \frac{1 - P_{ii}(h)}{h} = \nu_i.$$

Proof. Since we assumed $P_{ii} = 0$ for the embedded discrete time Markov chain, $1 - P_{ii}(h)$ is the probability that either exactly one transition occurs in $(0, h)$, which has probability $\nu_i h$ or more than one transition occurs, which has probability $o(h)$. Thus

$$1 - P_{ii}(h) = \nu_i h + o(h)$$

Dividing by h and letting $h \to 0$, we get the first equality. For the second, note that for the same reason as above,

$$P_{ij}(h) = \nu_i P_{ij} h + o(h) = q_{ij} h + o(h),$$

Dividing by h and letting $h \to 0$, we are done. □

Using this, we derive two differential equations for $P_{ij}(t)$.

Proposition 4.3. *We have the following two ODEs for all $t, i \neq j$.*

- $P'_{ij}(t) = \sum_{k\neq i} q_{ik} P_{kj}(t) - \nu_i P_{ij}(t)$ ***(Kolmogorov's backward equations)***
- $P'_{ij}(t) = \sum_{k\neq j} q_{kj} P_{ik}(t) - \nu_j P_{ij}(t)$ ***(Kolmogorov's forward equations)***

Proof. We will prove the backward equations.

$$\begin{aligned} P_{ij}(h + t) - P_{ij}(t) &= \sum_k P_{ik}(h) P_{kj}(t) - P_{ij}(t) \\ &= \sum_{k\neq i} P_{ik}(h) P_{kj}(t) - (1 - P_{ii}(h)) P_{ij}(t) \\ &= \sum_{k\neq i} (q_{ik} h + o(h)) P_{kj}(t) - (\nu_i h + o(h)) P_{ij}(t). \end{aligned}$$

Dividing by h and letting $h \to 0$, we are done.

For the forward equation,

$$\begin{aligned} P_{ij}(t + h) - P_{ij}(t) &= \sum_k P_{ik}(t) P_{kj}(h) - P_{ij}(t) \\ &= \sum_{k\neq j} P_{ik}(t) P_{kj}(h) - (1 - P_{jj}(h)) P_{ij}(t) \\ &= \sum_{k\neq j} (q_{kj} h + o(h)) P_{ik}(t) - (\nu_j h + o(h)) P_{ij}(t). \end{aligned}$$

Dividing by h and letting $h \to 0$, we are done. □

Example 4.4 (Birth and death process)**.** For a birth and death process with rates λ_i and μ_i, we have the following equations

$$P'_{ij}(t) = \lambda_i P_{i+1,j}(t) + \mu_i P_{i-1,j}(t) - (\lambda_i + \mu_i) P_{ij}(t)$$
$$P'_{0j}(t) = \lambda_0 P_{1j}(t) - \lambda_0 P_{0j}(t).$$

Example 4.5 (Pure birth process)**.** For a pure birth process, $\mu_i = 0$ for all i. Thus we get from Example 4.4

$$P'_{ij}(t) = \lambda_i P_{i+1,j}(t) - \lambda_i P_{ij}(t)$$
$$P'_{0j}(t) = \lambda_0 P_{1j}(t) - \lambda_0 P_{0j}(t).$$

But also $P_{ij}(t) = 0$ if $j < i$, so the system of ODE becomes *triangular* in the following sense

$$P'_{ii}(t) = -\lambda_i P_{ii}(t)$$
$$P'_{ij}(t) = \lambda_i P_{i+1,j}(t) - \lambda_i P_{ij}(t) \text{ for } j > i.$$

This can be explicitly solved using ODE techniques.

Limiting probabilities

If the ODE becomes hard to solve, we turn to long run proportions. Let us assume

$$\lim_{t\to\infty} P_{ij}(t) \text{ exists and is independent of } i. \text{ Let us denote this limit by } P_j.$$

We note (and won't prove) that a sufficient condition for the above limit to exist is that the underlying Markov chain is irreducible and positive recurrent.

Note the similarities with the discrete time Markov chain result about long run proportion spent at j converging to stationary distributions. Note this also means that

$$\lim_{t\to\infty} P'_{ij}(t) = 0.$$

Taking $t \to \infty$ in the backward equation, we get

$$0 = \sum_{k\neq j} q_{kj} P_k - \nu_j P_j \implies \sum_{k\neq j} q_{kj} P_k = \nu_j P_j$$

The left hand side above can be interpreted as the rate at which the chain leaves j and the left hand side is the rate in which it enters j. At equilibrium, these two should be equal, which is exactly what the equation states.

Exercise 4.6. *Read example 6.15 from Ross (11th edition)*

5 Brownian motion

Brownian motion, in one dimension, can be simply thought of as a simple random walk, viewed from far away (or in other words, rescaled). Consider a simple random walk $(S_n)_{n\geq 0}$ on $\mathbb{Z}$ and recall that we can write

$$S_n = X_1 + X_2 + \ldots + X_n$$

where $X_1, X_2, \ldots$ are i.i.d. so that

$$\mathbb{P}(X_i = 1) = \frac{1}{2}$$
$$\mathbb{P}(X_i = -1) = \frac{1}{2}.$$

However, instead of moving $+1$ or -1, let us consider moving $+\varepsilon$ or $-\varepsilon$ (think of taking ε small, later we will let $\varepsilon \to 0$). Also, instead of thinking of doing the simple random walk step in each unit interval of time, suppose we make the step after each Δ unit of time (we will also let $\Delta \to 0$ later). In effect, we are shrinking the whole picture and at the same time *speeding up* the process. Questions to address:

- What happens if we let $\Delta \to 0$ and $\varepsilon \to 0$?
- Does it matter if we let $\Delta \to 0$ and $\varepsilon \to 0$ at the same rate?
- If yes, what are the correct rates?

Call this modified rescaled and sped up process $X_{\varepsilon,\Delta}$. We will not answer all these questions in rigorous forms as that involves looking at the path in the correct space (which is $C[0,1]$, the space of all continuous functions in $[0,1]$). Rather we will do some back of the envelope calculations to guess correctly. Let $t = \Delta n$ for some positive integer n. That is n is the number of steps of the walk in time t (once we have sped up the process).

$$\mathbb{E}(X(t)) = 0, \qquad \mathrm{Var}(X(t)) = \mathrm{Var}(\varepsilon(X_1 + \ldots + X_n)) = \varepsilon^2 n \,\mathrm{Var}(X_1) = \varepsilon^2 \frac{t}{\Delta}.$$

since $\mathrm{Var}(X_1) = 1$ here. This suggests that if in the limit, we want the variance to be finite, we must take $\Delta = \varepsilon^2$. Anything much smaller (i.e. faster), will make the whole process happen *instantly in the limit.* Anything much bigger (i.e. slower), and we will see nothing happen for eternity.

Using the central limit theorem, it is natural to conclude that if we

$$\lim_{\varepsilon \to 0} X_{\varepsilon,\varepsilon^2}(t) = \lim_{\varepsilon \to 0}(\varepsilon(X_1 + \ldots + X_{t/\varepsilon^2})) = N(0,t)$$

where the limit above is in distribution. Also, it is easy to see that if we replace $X_1, X_2, \ldots$ by any i.i.d. sequence with $\mathbb{E}(X_1) = 0$ and $\mathrm{Var}(X_1) = \sigma^2$, the result does not change.

In fact, in the proper space, the limit can be actually computed for the whole path, and the Stochastic process we get is something called a **Brownian motion**, which we now define.

Definition 5.1. *A stochastic process $(X(t))_{t\geq 0}$ is a Brownian motion with variance σ^2 if it satisfies the following properties:*

(i) $X(0) = 0$

(ii) $\{X(t) : t \geq 0\}$ *has independent increments, that the if* $[s_1, t_1], [s_2, t_2], \ldots, [s_n, t_n]$ *are disjoint intervals then* $(X(t_1) - X(s_1)), \ldots, X(t_n) - X(s_n)$ *are independent.*

(iii) $\{X(t) : t \geq 0\}$ *is stationary, meaning that for every* $s \geq 0$,

$$(X(s+t) - X(s))_{t\geq 0}$$

has the same distribution as $(X(t))_{t\geq 0}$.

(iv) For each $t \geq 0$, $X(t) \sim N(0, \sigma^2 t)$.

If $\sigma = 1$, *this process is called a* ***standard Brownian motion****. See* [9] *for a history of Brownian motion.*

We will now assume the following which we do not prove:

Theorem 5.2. *One can construct a Stochastic process which satisfies the properties above, so that almost surely,* $X(t)$ *is a continuous function.*

We remark that the above statement is quite complicated as one needs to specify the joint distribution of *uncountably many* random variables together, satisfying the above three properties. This is highly non-trivial (but can be done).

We will now see how this definition allows us to compute the joint distribution of any set $X(t_1), X(t_2), \ldots, X(t_n)$ of random variables. This is simply because of stationarity and independent increments, the joint distribution of

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

is the same as independent random variables distributed as $X(t_1)$, $X(t_2 - t_1)$, $\ldots$, $X(t_n - t_{n-1})$.

Example 5.3. Suppose $X(t)$ is a Brownian motion with mean 0 and Variance σ^2. Compute

- $\mathrm{Var}(X(t))$.
- $\mathrm{Cov}(X(t), X(s))$ for $s < t$.
- $\mathbb{E}(X(t)|X(s) = 5)$, $\mathrm{Var}(X(t)|X(s) = 5)$,

[9] https://en.wikipedia.org/wiki/Brownian_motion

Solution. First one is $\sigma^2 t$ by definition. For the second one use independent increments, to conclude that $\text{Cov}(X(t), X(s)) = \text{Var}(X(s)) = \sigma^2 s$. For the third one, use independent increments again to conclude that

$$\mathbb{E}(X(t)|X(s) = 5) = \mathbb{E}(X(s) + X(t) - X(s)|X(s) = 5) = 5 + 0 = 5.$$

Also

$$\text{Var}(X(t)|X(s) = 5) = \text{Var}(X(s) + X(t) - X(s)|X(s) = 5) = \text{Var}(X(t) - X(s)) = \sigma^2(t - s).$$

Example 5.4. Show that if $s < t$,

$$E(X(s)|X(t) = 10)) = 10\frac{s}{t}, \qquad \text{Var}(X(s)|X(t) = 10) = \frac{s}{t}(t - s).$$

Solution. To compute the conditional expectation, we compute the conditional density. However we note that computing the joint density of $X(t) - X(s)$ and $X(s)$ is easier: they are independent normal with mean 0 and variance $t - s$ and s respectively. Thus note that $\{X(t) = 10, X(s) = x\}$ is the same as $\{X(s) = x, X(t) - X(s) = 10 - x\}$. Thus, we can write (exercise: argue more rigorously usign properties of multivariate Normal)

$$f_{X(s)|X(t)(x|10)} = \frac{f_{X(s)}(x) f_{X(t)-X(s)}(10 - x)}{f_{X(t)}(10)}$$

One way to simplify computations is to not care about the constants which do not depend upon x as we know this is a density hence we can always recover the constants by integrating over x and equating to 1. Using this idea, we write

$$\frac{f_{X(s)}(x) f_{X(t)-X(s)}(10 - x)}{f_{X(t)}(10)} = C_1 \exp(-\frac{(x - 10)^2)}{2(t - s)} - \frac{x^2}{2s}) = C_2 \exp(-\frac{(x - 10s/t)^2}{2s(t - s)/t})$$

Comparing this with the density of a Normal random variable, we find that

$$E(X(s)|X(t) = 10)) = 10\frac{s}{t} \qquad \text{and } \text{Var}(X(s)|X(t) = 10) = \frac{s}{t}(t - s).$$

The above example shows that conditioning on the future makes the mean of the Brownian motion to be a *straight line* joining 0 and the point where the Brownian motion will be in the future because of the conditioning.

Example 5.5. See example 10.1 in book.

What happens if we rescale multiply a Brownian motion $X(t)$ with variance σ by a constant c? All the properties are preserved, except the last one, namely,

$$cX(t) \sim N(0, c^2\sigma^2 t)$$

But this is the same as the distribution of $X(c^2 t)$. Thus we reach the important conclusion

Proposition 5.6.

$$(\frac{1}{c} X(c^2 t))_{t \geq 0} = (X(t))_{t \geq 0}$$

in distribution.

In particular for $c = -1$, we see $X(t)$ is equal in distribution to $-X(t)$, which means Brownian motion is symmetric.

5.1 Higher dimensions

In higher dimensions, say in dimension d, Brownian motion is simply defined as a vector consisting of d independent Brownian motions.

$$(\boldsymbol{X}(t))_{t\geq 0} := (X_1(t), X_2(t), \ldots, X_d(t))_{t\geq 0}$$

where

$$X_1(t), X_2(t), \ldots$$

are i.i.d. Brownian motions (the usual one dimensional ones). Covariances are computed very similarly and we have the following similar scaling

$$(\boldsymbol{X}(c^2t))_{t\geq 0} = (X_1(c^2t), X_2(c^2t), \ldots, X_d(c^2t))_{t\geq 0} \overset{(d)}{=} c((X_1(t), X_2(t), \ldots, X_d(t))_{t\geq 0}) = c\boldsymbol{X}(t).$$

5.2 Gamblers ruin and hitting times

As usual let $X(t)$ be a Brownian motion. Let T_a be the first hitting time of $a > 0$ and suppose $b > 0$ and T_{-b} be the first time the Brownian motion hits $-b$. We will show

Proposition 5.7. $\mathbb{P}(T_a < T_{-b}) = \frac{b}{a+b}$. *That is, the probability that a Brownian motion hits a before b is given by the above formula. In particular, if $-a = b$, the Brownian motion has an equal chance to hit a or $-a$ before the other.*

Proof. We will again use a " hand wavy argument" using the fact that Brownian motion can be thought of as a limit of a simple random walk. For a simple random walk, when the steps are of size ε or $-\varepsilon$ with equal probability, suppose the probability that starting from 0 it hits $i = \varepsilon x$ before $j = -\varepsilon y$ is given by $p_{x,y}$. Then

$$p_{x,y} = \frac{1}{2}p_{x-1,y} + \frac{1}{2}p_{x,y-1} \text{ for } 0 \leq x \leq a, -b \leq y \leq 0.$$

Solving this with boundary conditions $p_{0,y} = 1$ for all $x \neq 0$ and $p_{x,0} = 0$ for all $x \neq 0$, we see that the probability is exactly

$$p_{a,b} = \frac{\varepsilon b}{\varepsilon(a+b)} = \frac{b}{a+b}$$

Thus since the expression does not depend on ε, letting $\varepsilon \to 0$, we get the same formula. □

Corollary 5.8. *Fix $k \in \mathbb{R}$. $\mathbb{P}(T_k < \infty) = 1$, or in other words, Brownian motion hits 1 at some point of time.*

Proof. $\mathbb{P}(T_k < T_{-n}) = \frac{n}{n+k} \to 1$ as $n \to \infty$ (since k is fixed). But the events $\{T_1 < T_{-n}\}$ are getting larger as n grows and eventually reaches $\{T_1 < \infty\}$. Hence so does their probabilities. □

Ok, so we now Brownian motion hits 1 with probability 1. After hitting 1, Brownian motion is the same as $1 + X(t)$ where $X(t)$ is another Brownian motion. So the Brownian motion will hit -1 with probability 1 at some point. Iterating this argument, we see that a Brownian motion will hit 1 and -1 infinitely often. By continuity (intermediate value theorem), Brownian motion will hit 0 in between hitting 1 and -1. Thus we have established a form of *recurrence* of Brownian motion:

Theorem 5.9. *With probability 1, a Brownian motion $X(t)$ hits 0 infinitely often.*

We will now see that this actually gives us the following (remarkable at first) property of Brownian motion: for for any $\varepsilon > 0$, $X(t)_{0 \le t \le \varepsilon}$ infinitely often with probability 1. To see this we first make a claim:

Proposition 5.10.

$$Y(t) = (tX(1/t)) \text{ for } t > 0 \text{ and } Y(0) = 0$$

is another Brownian motion.

Proof. We will not prove this completely, rather wave our hands a bit. Recall that for a multivariate normal distribution, to know the distribution, one only needs to specify the mean vector and the covariance matrix. This is also the case if we have a "vector" consisting of uncountably many random variables, which is the case here, $X(t)$ is Normal for each $t \ge 0$. Note that $\mathbb{E}(tX(1/t)) = 0$ and for $s < t$,

$$\operatorname{Cov}(sX(1/s), tX(1/t)) = st \operatorname{Cov}(X(1/s), X(1/t)) = st\frac{1}{t} = s = \operatorname{Cov}(X(s), X(t)).$$

Showing that $tX(1/t) \to 0$ as $t \to 0$ can also be done, but we skip that here. □

But note that $Y(t) = tX(1/t)$ hits 0 infinitely often as $t \to \infty$. However, this means that $X(1/t)$ hits 0 infinitely often as $1/t$ goes to 0 (since we are reversing time in some sense). This shows that $X(t)_{0 \le t \le \varepsilon}$ must hit 0 infintely often with probability 1.

5.3 Reflection principle and law of the maximum

Let $X(t)$ be a Brownian motion and suppose that $M(t) = \sup\{X(s) : 0 \le s \le t\} = \max\{X(s) : 0 \le s \le t\}$ (since Brownian motion is a continuous function on a compact set, its supremum is its maximum). We will show the following nice proposition

Proposition 5.11. *For all $a > 0$,*

$$\mathbb{P}(M(t) > a) = 2\mathbb{P}(X(t) > a) = \mathbb{P}(|X(t)| > a).$$

Proof. The second equality is obvious

$$\mathbb{P}(|X(t)| > a) = \mathbb{P}(X(t) > a) + \mathbb{P}(\mathbb{P}(X(t) < -a)) = 2\mathbb{P}(X(t) > a).$$

which follows from the symmetry of Brownian motion since $X(t)$ and $-X(t)$ have the same distribution.

For the first part, we will "hand wave" a little and assume the following. Let $T = \inf\{t \geq 0 : X(t) = a\}$. We will use a consequence of something called the *strong Markov property* of Brownian motion:

$$(X(T+t) - X(T))_{t\geq 0} \text{ has the same distribution as } (X(t))_{t\geq 0}.$$

Note that T is a random variable so this is not a consequence of stationarity, but is still true [10]. But if this is the case, we can use the symmetry of the Brownian motion, to reflect it at T. This is still a Brownian motion. This gives us the process

$$X^*(t) = X(t)1_{0\leq t\leq T} + (2X(T) - X(t))1_{T>t}.$$

which has the same distribution as $(X(t))_{t\geq 0}$.

Using this, we can say the following

$$\mathbb{P}(M(t) > a) = \mathbb{P}(X(t) > a) + \mathbb{P}(M(t) > a, X(t) \leq a) = .\mathbb{P}(X(t) > a) + \mathbb{P}(X*(t) \geq a) = 2\mathbb{P}(X(t) > a)$$

since if $X(t)$ is lower than a, the process reflected at T must be bigger than a. Thus since $X^*(t)$ has the same distribution as $X(t)$, we have the final equality.

Example 5.12. Let $M(t) = \max\{X(s) : 0 \leq s \leq t\}$ for a Brownian motion X. Show that $\mathbb{E}(|M(t)|) < \infty$.

Solution We have using the

$$\mathbb{P}(M(t) > a) = 2\mathbb{P}(B(t) > a) = 2\int_a^\infty \frac{1}{\sqrt{2\pi t}} e^{-\frac{x^2}{2t}} \leq 2\int_a^\infty \frac{1}{\sqrt{2\pi t}} e^{-\frac{ax}{2t}} = 2\frac{2t}{a} e^{-a^2} \frac{1}{\sqrt{2\pi t}} = 2\frac{\sqrt{2t}}{a\sqrt{\pi}} e^{-a^2}$$

Now recall,

$$\mathbb{E}(|M(t)|) = \int_0^\infty \mathbb{P}(|M(t)| > a) da$$

But

$$\int_1^\infty 2\mathbb{P}(M(t) > a) da = \int_1^\infty 4\frac{\sqrt{2t}}{a\sqrt{\pi}} e^{-a^2} da < \infty$$

and

$$\int_0^1 \mathbb{P}(|M(t)| > a) da \leq \int_0^1 \frac{1}{\sqrt{2\pi t}} e^{-\frac{x^2}{2t}} dx$$

which is clearly finite.

□

[10] in particular, this is true for any *stopping time*, if you are interested you can look it up.

5.4 Gaussian processes

We studied multivariate Normal distribution in Math 352/ Stat 350. Here is a quick recap:

5.4.1 Multivariate Normal

The joint pdf of a multivariate Normal distribution as follows.

Definition 5.13. *We say* $(X_1, \ldots, X_n)$ *follows multivariate Normal with mean vector* $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^T$ *and Covariance matrix* Σ *is*

$$f_{X_1,\ldots,X_n}(\boldsymbol{x}) = \frac{1}{(2\pi)^{n/2}\sqrt{\det(\Sigma)}} \exp((\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})), \quad \boldsymbol{x} \in \mathbb{R}^n$$

The above expression is quite complicated. There is a natural interpretation of the parameters as in the one variable case: $\mu_i = \mathbb{E}(X_i)$. The variance of the one variable case is replace by Σ which is an $n \times n$ matrix and the (i,j)th entry of Σ is $Cov(X_i, X_j)$.

We will learn a different way to define multivariate normal through Moment generating functions, which avoid all the nasty multivariate integrals in the above definition.

Alternate equivalent definition of multivariate Normal:

Consider the matrix

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \ldots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \ldots & a_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \ldots & a_{nn} \end{bmatrix}$$

and

$$\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^T.$$

Definition 5.14 (Alternate definition equivalent to Definition 5.13)**.** *Let* $Z_1, \ldots, Z_n$ *be i.i.d. Normal* $(0,1)$*. Let* $Z = (Z_1, \ldots, Z_n)$*. A multivariate Normal with mean* $\boldsymbol{\mu}$ *and covariance matrix* AA^T *is the joint distribution of the vector*

$$X = AZ + \boldsymbol{\mu}.$$

We write

$$X \sim N(\boldsymbol{\mu}, AA^T).$$

Important facts:

1. If we know the mean vector $\boldsymbol{\mu}$ and the covariance matrix Σ, then we know the joint density of multivariate Normal.

2. The sub-vector obtained by removing some of the entries of a multivariate Normal is again a multivariate Normal distribution.

3. If $\boldsymbol{X}$ is a multivariate Normal distribution with n entries and A is a $m \times n$ matrix then $A\boldsymbol{X}$ is another multivariate Normal distribution with m entries. That is linear combinations of multivariate Normal is another multivariate Normal. Exercise: If $\boldsymbol{X}$ has Covariance matrix Σ, find the covariance matrix of $A\boldsymbol{X}$.

4. In a multivariate Normal, if we condition on the values of some of the coordinates, the law of the remaining is still a multivariate Normal.

Now back to Gaussian processes.

Definition 5.15. *A stochastic process $(X(t))_{t\geq 0}$ is a Gaussian process if for all $t_1 < t_2 < \ldots < t_n$,*

$$(X(t_1), X(t_2), \ldots, X(t_n))$$

is a multivariate Normal.

Note that we did not specify the mean vector and the covariance matrix which opens up a wide range of possibilities for Gaussian processes. For example,

Proposition 5.16. *Brownian motion is a Gaussian process.*

Proof. Let $(X(t))_{t\geq 0}$ be a Brownian motion. We know

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

are independent Normal distributions and hence is a multivariate Normal. Furthermore,

$$\begin{bmatrix} X(t_1) \\ X(t_2) \\ \vdots \\ X(t_n) \end{bmatrix} = \begin{pmatrix} 1 & 0 & \ldots & 0 & 0 \\ 1 & 1 & \ldots & 0 & 0 \\ & & \ldots & & \\ 1 & 1 & \ldots & 1 & 1 \end{pmatrix} \begin{bmatrix} X(t_1) \\ X(t_2) - X(t_1) \\ \vdots \\ X(t_n) - X(t_{n-1}) \end{bmatrix}$$

which implies using the fact that linear combination of multivariate Normal is a multivariate normal that

$$(X(t_1), X(t_2), \ldots, X(t_n))$$

is a multivariate Normal. □

In fact since multivariate Normal distributions are completely determined by mean and covariance matrices, we also have the following

Proposition 5.17. *Let $(X(t))_{t\geq 0}$ and $(Y(t))_{t\geq 0}$ be Gaussian processes such that*

$$\mathbb{E}(X(t)) = \mathbb{E}(Y(t)) \text{ for all } t \geq 0$$

and

$$\mathrm{Cov}(X(s), X(t)) = \mathrm{Cov}(Y(s), Y(t)) \text{ for all } s, t \geq 0$$

then $(X(t))_{t\geq 0}$ and $(Y(t))_{t\geq 0}$ have the same distribution (i.e. are the same Gaussian processes).

As a corollary we obtain

Corollary 5.18. *Brownian motion is the only Gaussian process* $(X(t))_{t\geq 0}$ *such that*

$$\mathbb{E}(X(t)) = 0 \text{ and } \operatorname{Cov}(X(s), X(t)) = min(s,t).$$

Warning: This is a very special property of Gaussian processes and for any other process this is certainly not true. For example consider the following Compound Poisson process. Let $N(t)$ be a Poisson process with rate 1 and let $X_1, \ldots, X_n$ be i.i.d. with $\mathbb{E}(X_1) = 0$ and $\operatorname{Var}(X_1) = 1$. Let

$$Y(t) = \sum_{i=1}^{N(t)} X_i$$

Then

$$\mathbb{E}(Y(t)) = 0 \text{ (exercise)}$$

and

$$\operatorname{Cov}(Y(s), Y(t)) = \min(s,t) \text{ (exercise)}$$

This is the same as that of a Brownian motion, however a Poisson process is not Gaussian (i.e. $N(t) \sim Poisson(t)$) hence there is no contradiction to Proposition 5.17.

5.5 Brownian bridge

We want to define an object like a Brownian motion, but defined on a finite interval with both endpoints taking value 0. To that end, we simply define

Definition 5.19. *A* ***Brownian bridge*** *in the interval* $[0,1]$ *is a stochastic process* $(Y(t))_{0\leq t\leq 1}$ *which defined as a Brownian motion* $X(t)$ *conditioned on* $X(1) = 0$.

Using item 4 of the properties of multivariate Normal, we see that a Brownian bridge is a Gaussian process (exercise: prove it in details). Thus all we need to characterize it is to compute the mean and the covariance. Using example 5.4, we see that

$$\mathbb{E}(Y(t)) = \mathbb{E}(X(t)|X(1) = 0) = \frac{t}{1} \times 0 = 0$$

and for $s < t$,

$$\begin{aligned}
\mathrm{Cov}(Y(s), Y(t)) &= \mathbb{E}(Y(s)Y(t)) \\
&= \mathbb{E}(X(s)X(t)|X(1) = 0) \\
&= \mathbb{E}(\mathbb{E}(X(s)X(t)|X(t), X(1) = 0)) \\
&= \mathbb{E}(X(t)\mathbb{E}(X(s)|X(t), X(1) = 0)) \\
&= \mathbb{E}(X(t)\frac{s}{t}X(t)|X(1) = 0) \text{ using example 5.4} \\
&= \mathbb{E}(X^2(t)\frac{s}{t}|X(1) = 0) \\
&= \frac{s}{t}\,\mathrm{Var}(X(t)|X(1) = 0) \\
&= \frac{s}{t}t(1-t) = s(1-t) \text{ using example 5.4.}
\end{aligned}$$

Now we prove that Brownian bridge can be alternatively represented as follows.

Proposition 5.20. *Let $(X(t))_{t\geq 0}$ be a Brownian motion. Then*

$$Z(t) := X(t) - tX(1), \qquad 0 \leq t \leq 1$$

is a Brownian bridge.

Proof. Using Proposition 5.17, we only need to check whether the mean and the covariances are the same as that we computed above. Note

$$\mathbb{E}(Z(t)) = \mathbb{E}(X(t)) - t\mathbb{E}(X(1)) = 0$$

and for $s < t$, it is a straightforward calculation to show that

$$\mathrm{Cov}(Z(s), Z(t)) = s(1-t) \qquad (exercise).$$

□

6 Solutions

6.1 Section 1

1. (a) *Marginal of Y.* For $0 < y < 1$,

$$f_Y(y) = \int_0^1 4x(1-y)\,dx = 4(1-y)\int_0^1 x\,dx = 4(1-y)\cdot\frac{1}{2} = 2(1-y).$$

Conditional density. For $0 < x < 1$, $0 < y < 1$,

$$f_{X|Y}(x \mid y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{4x(1-y)}{2(1-y)} = 2x, \qquad 0 < x < 1.$$

Thus $X \mid (Y = y) \sim \text{Beta}(2, 1)$.

(b) *Conditional expectation.*

$$\mathbb{E}[X \mid Y = y] = \int_0^1 x \cdot (2x)\,dx = 2\int_0^1 x^2\,dx = \frac{2}{3}.$$

(c) *Unconditional expectation.* By the law of total expectation,

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X \mid Y]] = \int_0^1 \frac{2}{3}\,f_Y(y)\,dy = \frac{2}{3}\int_0^1 2(1-y)\,dy = \frac{2}{3}\cdot 1 = \frac{2}{3}.$$

(d) *Variances* First,

$$\mathbb{E}[X^2 \mid Y = y] = \int_0^1 x^2 \cdot (2x)\,dx = 2\int_0^1 x^3\,dx = \frac{1}{2}.$$

So

$$\text{Var}(X \mid Y = y) = \frac{1}{2} - \left(\frac{2}{3}\right)^2 = \frac{1}{18}.$$

By the law of total variance,

$$\text{Var}(X) = \mathbb{E}\big[\,\text{Var}(X \mid Y)\big] + \text{Var}\big(\mathbb{E}[X \mid Y]\big) = \frac{1}{18} + 0 = \frac{1}{18}.$$

2. (a) *Marginal of Y.* For $y \in \{1, 2, 3\}$,

$$\mathbb{P}(Y = y) = \sum_{x=1}^{2} \mathbb{P}(X = x, Y = y) = \frac{(1+y)+(2+y)}{21} = \frac{3+2y}{21}.$$

(Checking: $y = 1 \mapsto 5/21$, $y = 2 \mapsto 7/21$, $y = 3 \mapsto 9/21$, which sum to 1.)

(b) *Conditional pmf.* For $y \in \{1,2,3\}$ and $x \in \{1,2\}$,

$$\mathbb{P}(X = x \mid Y = y) = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)} = \frac{(x+y)/21}{(3+2y)/21} = \frac{x+y}{3+2y}.$$

(c) *Conditional expectation.* For $y \in \{1,2,3\}$,

$$\mathbb{E}[X \mid Y = y] = \sum_{x=1}^{2} x \cdot \mathbb{P}(X = x \mid Y = y) = \frac{1 \cdot (1+y) + 2 \cdot (2+y)}{3+2y} = \frac{5+3y}{3+2y}.$$

(d) *Unconditional expectation.* By the law of total expectation,

$$\mathbb{E}[X] = \sum_{y=1}^{3} \mathbb{E}[X \mid Y = y]\, \mathbb{P}(Y = y) = \sum_{y=1}^{3} \frac{5+3y}{3+2y} \cdot \frac{3+2y}{21} = \frac{1}{21} \sum_{y=1}^{3} (5+3y).$$

Compute the sum:

$$\sum_{y=1}^{3} (5+3y) = (5+3\cdot 1) + (5+3\cdot 2) + (5+3\cdot 3) = 8 + 11 + 14 = 33,$$

hence

$$\mathbb{E}[X] = \frac{33}{21} = \frac{11}{7}.$$

3. For $p \in (0,1)$

$$\mathbb{P}(P \le p | X = 1) = \frac{\mathbb{P}(X = 1, P \le p)}{\mathbb{P}(X = 1)} = \frac{\int_0^p \mathbb{P}(X = 1 | P = u) du}{1/2} = 2\int_0^p u du = p^2$$

Differentiating, we get the conditional density

$$f_{P|X}(p \mid 1) = 2p.$$

Since $\mathbb{P}(X = 1 \mid P = p) = p$, $f_P(p) = 1$ for $0 < p < 1$ (uniform prior), and $\mathbb{P}(X = 1) = \frac{1}{2}$, we obtain

$$f_{P|X}(p \mid 1) = \frac{p \cdot 1}{1/2} = 2p, \qquad 0 < p < 1.$$

If $p \le 0$ then $\mathbb{P}(P \le p | X = 1) = 0$ and $p \ge 1$ then $\mathbb{P}(P \le p | X = 1) = 1$. Thus, overall

$$f_{P|X}(p \mid 1) = \begin{cases} 2p, & 0 < p < 1. \\ 0 \text{ otherwise.} \end{cases}$$

By same calculations, we get

$$f_{P|X}(p \mid 0) = \begin{cases} 2(1-p), & 0 < p < 1. \\ 0 \text{ otherwise.} \end{cases}$$

The conditional expectations are $\mathbb{E}(P|X=1) = \int_0^1 2p^2 dp = \frac{2}{3}$ and $\mathbb{E}(P|X=0) = \int_0^1 2p(1-p)dp = \frac{1}{3}$.

Also

$$\mathbb{E}(P) = \mathbb{E}(P|X) = \mathbb{E}(P|X=0)\mathbb{P}(X=0) + \mathbb{E}(P|X=1)\mathbb{P}(X=1) = \frac{2}{3}\frac{1}{2} + \frac{1}{3}\frac{1}{2} = \frac{1}{2}.$$

as it should be.

4. **(1)** $\mathbb{P}(X < Y)$**.** Using independence,

$$\mathbb{P}(X < Y) = \int_0^\infty \mathbb{P}(Y > x)\, f_X(x)\, dx = \int_0^\infty e^{-\lambda_2 x}\, \lambda_1 e^{-\lambda_1 x}\, dx = \lambda_1 \int_0^\infty e^{-(\lambda_1+\lambda_2)x}\, dx = \frac{\lambda_1}{\lambda_1+\lambda_2}.$$

(2) Conditional law of the excess $W := Y - X$ **given** $X < Y$**.** Fix $x > 0$. Condition on the event $\{X = x\}$. Under this conditioning,

$$\mathbb{P}\big(W > w \mid X = x,\ X < Y\big) = \mathbb{P}\big(Y - x > w \mid Y > x\big).$$

as X is independent of Y. By the memoryless property of the exponential,

$$\mathbb{P}\big(Y - x > w \mid Y > x\big) = \mathbb{P}(Y > w) = e^{-\lambda_2 w}, \qquad w > 0,$$

so for each fixed x the conditional distribution of W given $X = x$ and $X < Y$ is $\text{Exp}(\lambda_2)$ (and does not depend on x). Hence after integrating over x we obtain

$$\int_{-\infty}^{\infty} \mathbb{P}\big(W > w \mid X = x,\ X < Y\big) f_X(x) dx = \int_{-\infty}^{\infty} e^{-\lambda_2 w} f_X(x) dx = e^{-\lambda_2 w}$$

i.e. W conditioned on $X < Y$ follows Exponential (λ_2).

5. **(1)** By definition of $G(n,p)$, each possible edge appears independently with probability p. In particular,

$$\mathbb{P}\big((u,v) \in E\big) = p.$$

(2) Let A denote the event that $(u,v) \in E$. Write

$$\deg(u) = \sum_{w \neq u} I_{uw},$$

where I_{uw} is the indicator that u is adjacent to w. There are $n-1$ potential neighbours of u, one of which is v and the remaining $n-2$ are the other vertices.

Compute the joint probability that A occurs and $\deg(u) = k$. This is the event that the edge (u, v) is present and exactly $k-1$ of the other $n-2$ possible edges from u are present. Thus

$$\mathbb{P}\big(A,\ \deg(u) = k\big) = p \cdot \binom{n-2}{k-1} p^{k-1}(1-p)^{(n-2)-(k-1)} = \binom{n-2}{k-1} p^{k}(1-p)^{n-1-k}.$$

The marginal probability that $\deg(u) = k$ (without conditioning on A) is

$$\mathbb{P}\big(\deg(u) = k\big) = \binom{n-1}{k} p^{k}(1-p)^{n-1-k},$$

since $\deg(u) \sim \mathrm{Bin}(n-1, p)$.

Therefore the conditional probability is

$$\mathbb{P}\big(A \mid \deg(u) = k\big) = \frac{\mathbb{P}(A, \deg(u) = k)}{\mathbb{P}(\deg(u) = k)} = \frac{\binom{n-2}{k-1} p^{k}(1-p)^{n-1-k}}{\binom{n-1}{k} p^{k}(1-p)^{n-1-k}} = \frac{\binom{n-2}{k-1}}{\binom{n-1}{k}}.$$

A simple combinatorial simplification gives

$$\frac{\binom{n-2}{k-1}}{\binom{n-1}{k}} = \frac{k}{n-1}.$$

Hence

$$\boxed{\mathbb{P}\big((u, v) \in E \mid \deg(u) = k\big) = \frac{k}{n-1}\,.}$$

(3) Interpretation. Knowledge of the degree $\deg(u) = k$ *does* affect the probability that u is adjacent to a particular vertex v: conditioned on $\deg(u) = k$, the probability is $k/(n-1)$, which in general differs from the unconditional probability p. Intuitively, conditioning on degree k says "u has exactly k neighbours chosen uniformly at random from the $n-1$ other vertices," so the chance that v is among those k neighbours is $k/(n-1)$. Note that taking expectation over k recovers the unconditional edge probability:

$$\mathbb{E}\left[\frac{\deg(u)}{n-1}\right] = \frac{\mathbb{E}[\deg(u)]}{n-1} = \frac{(n-1)p}{n-1} = p.$$

Thus on average the conditional result agrees with p, but for a given observed degree the conditional probability is $k/(n-1)$.

6. **(1) Unconditional probability of $S_n = k$.** For a symmetric random walk, $S_n = k$ if and only if exactly

$$m = \frac{n+k}{2}$$

of the steps are $+1$ (and the remaining $n-m=(n-k)/2$ steps are -1). Hence

$$\mathbb{P}(S_n=k)=\binom{n}{(n+k)/2}2^{-n}, \qquad k\equiv n \pmod 2.$$

(2) Conditional probability of the first step. Write

$$\mathbb{P}(X_1=+1 \mid S_n=k)=\frac{\mathbb{P}(X_1=+1, S_n=k)}{\mathbb{P}(S_n=k)}.$$

If $X_1=+1$, then the remaining $n-1$ steps must sum to

$$S_{n-1}=k-1.$$

Therefore,

$$\mathbb{P}(X_1=+1, S_n=k)=\mathbb{P}(X_1=+1)\,\mathbb{P}(S_{n-1}=k-1)=\frac{1}{2}\binom{n-1}{\frac{n+k}{2}-1}2^{-(n-1)}=\binom{n-1}{\frac{n+k}{2}-1}2^{-n}.$$

Similarly, by part (1),

$$\mathbb{P}(S_n=k)=\binom{n}{\frac{n+k}{2}}2^{-n}.$$

Thus

$$\mathbb{P}(X_1=+1 \mid S_n=k)=\frac{\binom{n-1}{\frac{n+k}{2}-1}}{\binom{n}{\frac{n+k}{2}}}.$$

A simple combinatorial simplification gives

$$\frac{\binom{n-1}{(n+k)/2-1}}{\binom{n}{(n+k)/2}}=\frac{(n+k)/2}{n}.$$

Hence

$$\boxed{\mathbb{P}(X_1=+1 \mid S_n=k)=\frac{n+k}{2n}.}$$

(3) Interpretation. The probability that the first step is $+1$ *does depend* on the final position $S_n=k$. Intuitively, if the walk ends up far to the right (k large), the first step is more likely to have been $+1$; if the walk ends up far to the left (k negative), the first step is more likely to have been -1.

Unconditionally, $\mathbb{P}(X_1=+1)=1/2$, but conditioning on $S_n=k$ skews this probability linearly:

$$\mathbb{P}(X_1=+1 \mid S_n=k)=\frac{1}{2}+\frac{k}{2n}.$$

7. **(1) Probability of reaching $+a$ before $-b$.**

For a symmetric random walk starting at 0, let

$$u_0 = \mathbb{P}(S_\tau = a \mid S_0 = 0).$$

The probability satisfies the standard difference equation for gambler's ruin:

$$u_0 = \frac{1}{2}u_1 + \frac{1}{2}u_{-1}, \quad \text{with } u_a = 1,\ u_{-b} = 0.$$

The solution is linear:

$$u_k = \frac{k+b}{a+b}, \quad -b \le k \le a.$$

In particular,

$$\mathbb{P}(S_\tau = a) = u_0 = \frac{b}{a+b}.$$

(2) Conditional probability of the first step.

By the Markov property,

$$\mathbb{P}(X_1 = +1 \mid S_\tau = a) = \frac{\mathbb{P}(X_1 = +1, S_\tau = a)}{\mathbb{P}(S_\tau = a)}.$$

The numerator can be computed as:

$$\mathbb{P}(X_1 = +1, S_\tau = a) = \mathbb{P}(X_1 = +1)\,\mathbb{P}(S_\tau = a \mid S_1 = +1) = \frac{1}{2} \cdot \mathbb{P}(S_\tau = a \mid S_1 = +1).$$

After the first step, the walk is at $S_1 = +1$. The probability that it eventually reaches $+a$ before $-b$ from $+1$ is

$$\mathbb{P}(S_\tau = a \mid S_1 = +1) = u_1 = \frac{1+b}{a+b}.$$

Hence

$$\mathbb{P}(X_1 = +1, S_\tau = a) = \frac{1}{2} \cdot \frac{1+b}{a+b} = \frac{1+b}{2(a+b)}.$$

Dividing by $\mathbb{P}(S_\tau = a) = \frac{b}{a+b}$ gives

$$\mathbb{P}(X_1 = +1 \mid S_\tau = a) = \frac{1+b}{2b}.$$

(3) Interpretation.

Conditioning on eventually hitting $+a$ biases the first step toward $+1$. Indeed, unconditionally $\mathbb{P}(X_1 = +1) = 1/2$, but conditioned on success (hitting $+a$ before $-b$), the first step has probability $\frac{b+1}{2b} > 1/2$. The closer b is to 0 (i.e., the closer the lower barrier), the stronger this bias; the first step is more likely to point toward the barrier we wish to reach.

Table 2.1

Discrete probability distribution	Probability mass function, $p(x)$	Moment generating function, $\phi(t)$	Mean	Variance
Binomial with parameters n, p, $0 \le p \le 1$	$\binom{n}{x} p^x (1-p)^{n-x}$, $x = 0, 1, \ldots, n$	$(pe^t + (1-p))^n$	np	$np(1-p)$
Poisson with parameter $\lambda > 0$	$e^{-\lambda}\frac{\lambda^x}{x!}$, $x = 0, 1, 2, \ldots$	$\exp\{\lambda(e^t - 1)\}$	λ	λ
Geometric with parameter $0 \le p \le 1$	$p(1-p)^{x-1}$, $x = 1, 2, \ldots$	$\frac{pe^t}{1-(1-p)e^t}$	$\frac{1}{p}$	$\frac{1-p}{p^2}$

Table 2.2

Continuous probability distribution	Probability density function, $f(x)$	Moment generating function, $\phi(t)$	Mean	Variance
Uniform over (a, b)	$f(x) = \begin{cases} \frac{1}{b-a}, & a < x < b \\ 0, & \text{otherwise} \end{cases}$	$\frac{e^{tb} - e^{ta}}{t(b-a)}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Exponential with parameter $\lambda > 0$	$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x > 0 \\ 0, & x < 0 \end{cases}$	$\frac{\lambda}{\lambda - t}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Gamma with parameters (n, λ), $\lambda > 0$	$f(x) = \begin{cases} \frac{\lambda e^{-\lambda x}(\lambda x)^{n-1}}{(n-1)!}, & x \ge 0 \\ 0, & x < 0 \end{cases}$	$\left(\frac{\lambda}{\lambda - t}\right)^n$	$\frac{n}{\lambda}$	$\frac{n}{\lambda^2}$
Normal with parameters (μ, σ^2)	$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \times \exp\{-(x-\mu)^2/2\sigma^2\}$, $-\infty < x < \infty$	$\exp\left\{\mu t + \frac{\sigma^2 t^2}{2}\right\}$	μ	σ^2