Stochastic Processes
and
Monte-Carlo Methods

University of Massachusetts: Spring 2018 version

Luc Rey-Bellet

April 5, 2018

Contents

Chapter 1

Simulation and the Monte-Carlo method

1.1 Review of probability

In this section we briefly review some basic terminology of probability, see any elementary probability book for reference.

Any real-valued random variable X can described by its ***cumulative distribution function*** (abbreviated ***c.d.f***) of X, i.e., the function $F_X : \mathbf{R} \to [0,1]$ defined by

$$F_X(x) \;=\; P\{X \leq x\}\,.$$

which is a right continuous with $0 \leq F_X(x) \leq 1$, $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to+\infty} F(x) = 1$. Conversely any function $F(x)$ with such properties define a random variable X.

If there exists a function $f : \mathbf{R} \to [0,\infty)$ such that $F_X(x) = \int_{-\infty}^{x} f_X(y)\,dy$ then X is said to be *continuous* with ***probability density function*** (abbreviated ***p.d.f***) f_X. By the fundamental theorem of calculus the p.d.f of X is obtained from the c.d.f of X by differentiating, i.e.,

$$f_X(x) = F_X'(x)\,.$$

On the other hand if X takes values in the set of integers, or more generally in some countable or finite subset S of the real numbers, then the random variable X and its c.d.f. are completely determined by its ***probability distribution function*** (also abbreviated p.d.f), i.e., by $p : S \to [0,1]$ where

$$p(i) \;=\; P\{X = i\}\,, \quad i \in S\,.$$

In this case X is called a ***discrete*** random variable.

The p.d.f. f of a continuous random variable satisfies $\int_{-\infty}^{\infty} f(x)\,dx = 1$ and the p.d.f of a discrete random variable satisfies $\sum_{i\in S} p_i = 1$. Either the c.d.f or p.d.f describes

the distribution of X and we compute the probability of any ***event*** $A \subset \mathbf{R}$ by

$$P\{X \in A\} = \begin{cases} \int_A f_X(x)\,dx & \text{if } X \text{ is continuous}\,, \\ \sum_{i\in A} p(i)\,dx & \text{if } X \text{ is discrete}\,. \end{cases}$$

Let $\mathbf{X} = (X_1, \cdots, X_d)$ be a ***random vector***, i.e., $X_1, \cdots X_d$ are a collection of d real-valued random variables with a joint distribution. Often the joint distribution can be described by the multi-parameter analogue of the p.d.f. For example if there is a function $f_{\mathbf{X}} : \mathbf{R}^d \to [0, \infty)$ such that

$$P(\mathbf{X} \in A) = \int \cdots \int_A f_{\mathbf{X}}(x_1, \cdots, x_d) dx_1 \cdots dx_d$$

then $\mathbf{X}$ is called a continuous random vector with p.d.f $f_{\mathbf{X}}$. Similarly a discrete random vector $\mathbf{X}$ taking values $\mathbf{i} = (i_1, \cdots, i_d)$ is described by

$$p(i_1, \cdots, i_d) = P\{X_1 = i_1, \cdots X_d = i_d\}\,.$$

A collection of random variables $X_1, \cdots, X_d$ are ***independent*** if the joint p.d.f satisfies

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}) &= f_{X_1}(x_1) \cdots f_{X_d}(x_d)\,, \quad \text{continuous case} \\ p_{\mathbf{X}}(\mathbf{i}) &= p_{X_1}(i_1) \cdots p_{X_d}(i_d)\,, \quad \text{discrete case} \end{aligned} \tag{1.1}$$

If $\mathbf{X}$ is a random vector and $g : \mathbf{R}^d \to \mathbf{R}$ is a function then $Y = g(\mathbf{X})$ is a real random variable. The ***mean*** or ***expectation*** of a real random variable X is defined by

$$E[X] = \begin{cases} \int_{-\infty}^{\infty} x f_X(x)\,dx & \text{if } X \text{ is continuous} \\ \sum_{i\in S} i\, p_X(i) & \text{if } X \text{ is discrete} \end{cases}$$

More generally if $Y = g(\mathbf{X})$ then

$$E[Y] = E[g(\mathbf{X})] = \begin{cases} \int_{\mathbf{R}^d} g(\mathbf{x}) f_{\mathbf{X}}(\mathbf{x})\,dx & \text{if } X \text{ is continuous} \\ \sum_{\mathbf{i}} g(\mathbf{i})\, p_{\mathbf{x}}(\mathbf{i}) & \text{if } X \text{ is discrete} \end{cases}$$

The ***variance*** of a random variable X, denoted by $\mathrm{var}(X)$, is given by

$$\mathrm{var}(X) = E\left[(X - E[X])^2\right] = E[X^2] - E[X]^2\,.$$

The mean of a random variable X measures the average value of X while its variance is a measure of the spread of the distribution of X. Also commonly used is the ***standard deviation*** $sd(X) = \sqrt{\mathrm{var}(X)}$.

Let X and Y be two random variables then we have

$$E[X + Y] = E[X] + E[Y]\,.$$

For the variance a simple computation shows that

$$\mathrm{var}(X+Y) = \mathrm{var}(X) + 2\mathrm{cov}(X,Y) + \mathrm{var}(Y)$$

where $\mathrm{cov}(X,Y)$ is the ***covariance*** of X and Y and is defined by

$$\mathrm{cov}(X,Y) = E\left[(X - E[X])(Y - E[Y])\right] .$$

In particular if X and Y are independent then $E[XY] = E[X]E[Y]$ and so $\mathrm{cov}(X,Y) = 0$ and thus $\mathrm{var}(X_1 + X_2) = \mathrm{var}(X_1) + \mathrm{var}(X_2)$.

Another important and useful object is the ***moment generating function*** (abbreviated ***m.g.f.***) of a random variable X and is given by

$$M_X(t) \;=\; E\left[e^{tX}\right] .$$

Whenever we use a m.g.f we will always assume that $M_X(t)$ is finite, at least in an interval around 0. Note that this is not always the case. If the moment generating function of X is known then one can compute all ***moments*** of X, $E[X^n]$, by repeated differentiation of the function $M_X(t)$ with respect to t. The n^{th} derivative of $M_x(t)$ is given by

$$M_x^{(n)}(t) \;=\; E\left[X^n e^{tX}\right]$$

and therefore

$$E[X^n] \;=\; M^{(n)}(0) .$$

In particular $E[X] = M_X'(0)$ and $\mathrm{var}(X) = M_X''(0) - (M_X'(0))^2$. It is often very convenient to compute the mean and variance of X using these formulas (see the examples below).

An important fact is the following (its proof is not easy!)

Theorem 1.1.1 *Let X and Y be two random variables and suppose that $M_X(t) = M_Y(t)$ for all $t \in (-\delta, \delta)$ then X and Y have the same distribution.*

Another easy and important property of the m.g.f is

Proposition 1.1.2 *If X and Y are independent random variable then the m.g.f of $X+Y$ satisfies*

$$M_{X+Y}(t) = M_X(t) M_Y(t) ,$$

i.e., the m.g.f of a sum of independent random variable is the product of the m.g.f.

Proof: We have

$$E\left[e^{t(X+Y)}\right] \;=\; E\left[e^{tX} e^{tY)}\right] \;=\; E\left[e^{tX}\right] E\left[e^{tY)}\right] ,$$

since e^{tX} and e^{tY} are independent. ∎

1.2 Some common random variables

We recall some important distributions together with their basic properties. The following facts are useful to remember.

Proposition 1.2.1 *We have*

1. *Suppose X is a continuous random variable with p.d.f $f(x)$. For any real number a the p.d.f of $X+a$ is $f(x-a)$.*
2. *Suppose X is a continuous random variable with p.d.f $f(x)$. For any non zero real number b the p.d.f of bX is $\frac{1}{|b|}f\left(\frac{x}{b}\right)$.*
3. *If X is a random variable, then for any real number a and b we have $M_{bX+a}(t) = e^{at}M_X(bt)$.*

Proof: The c.d.f of $X+a$ is

$$F_{X+a}(x) \;=\; P(X+a \le x) \;=\; P(X \le x-a) \;=\; F_X(x-a)\,.$$

Differentiating with respect to x gives

$$f_{X+a}(x) = F'_{X+a}(x) = f_X(x-a)\,.$$

This shows (i).

To prove (ii) one proceeds similarly. For $b>0$

$$F_{bX}(x) \;=\; P(bX \le x) \;=\; P(X \le x/b) \;=\; F_X(x/b)\,.$$

Differentiating gives $f_{bX}(x) = \frac{1}{b}f\left(\frac{x}{b}\right)$. The case $b<0$ is left to the reader.

To prove (iii) note that

$$M_{bX+a}(t) \;=\; E\left[e^{t(bX+a)}\right] \;=\; e^{ta}E\left[e^{tbX}\right] \;=\; e^{ta}M_X(bt)\,.$$

■

We recall the basic random variables and their properties.

1) Uniform Random Variable

Consider real numbers $a<b$. The ***uniform random variable on*** $[a,b]$ is the continuous random variable with p.d.f

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

The moment generating function is

$$E\left[e^{tX}\right] = \int_a^b e^{tx}\,dx \;=\; \frac{e^{tb}-e^{ta}}{t(b-a)}\,.$$

and the mean and variance are

$$E[X] = \frac{b-a}{2}\,, \quad \mathrm{var}(X) \;=\; \frac{(b-a)^2}{12}\,.$$

We write $X \sim \mathcal{U}([a,b])$ to denote such a random variable.

2) Normal Random Variable

Let μ be a real number and σ be a positive number. The ***normal random variable with mean*** μ ***and variance*** σ^2 is the continuous random variable with p.d.f

$$f(x) \;=\; \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,.$$

The moment generating function is (see below for a proof)

$$E\left[e^{tX}\right] = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tx} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,dx \;=\; e^{\mu t + \frac{\sigma^2 t^2}{2}}\,. \tag{1.2}$$

and the mean and variance are, indeed,

$$E[X] = \mu\,, \quad \mathrm{var}(X) \;=\; \sigma^2\,.$$

We write $X \sim \mathcal{N}(\mu,\sigma^2)$ to a normal random variable. The ***standard normal random variable*** is the normal random variable with $\mu = 0$ and $\sigma = 1$, i.e., $X \sim \mathcal{N}(0,1)$

The normal random variable has the following property

$$X \sim \mathcal{N}(0,1) \quad \text{ if and only if } \quad \sigma X + \mu \sim \mathcal{N}(\mu,\sigma^2)$$

To see this one applies Proposition 1.2.1 (i) and (ii) and this tells us that the density of $\sigma X + \mu$ is $\frac{1}{\sigma} f(\frac{x-\mu}{\sigma})$.

To show the formula for the moment generating function consider first $X \sim \mathcal{N}(0,1)$. Then by completing the square we have

$$\begin{aligned} M_X(t) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tx} e^{-\frac{x^2}{2}}\,dx \;=\; \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\frac{t^2}{2}} e^{-\frac{(x-t)^2}{2}}\,dx \\ &= e^{\frac{t^2}{2}} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{(x-t)^2}{2}}\,dx \;=\; e^{\frac{t^2}{2}} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{y^2}{2}}\,dy \;=\; e^{\frac{t^2}{2}} \end{aligned} \tag{1.3}$$

This proves the formula for $N(0,1)$. Since $N(\mu,\sigma^2) = \sigma N_{0,1} + \mu$, by Proposition 1.2.1, (iii) the moment generating function of N_{μ,σ^2} is $e^{t\mu} e^{\frac{\sigma^2 t^2}{2}}$ as claimed.

3) Exponential Random Variable
Let λ be a positive number. The ***exponential random variable with parameter*** λ is the continuous random variable with p.d.f

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases} .$$

The moment generating function is

$$E\left[e^{tX}\right] = \lambda \int_0^\infty e^{tx} e^{-\lambda x} = \begin{cases} \frac{\lambda}{\lambda - t} & \text{if } \lambda < t \\ +\infty & \text{otherwise} \end{cases}$$

and the mean and variance are

$$E[X] = \frac{1}{\lambda}, \quad \text{var}(X) = \frac{1}{\lambda^2}.$$

We write $X \sim Exp(\lambda)$ to denote this random variable. This random variable will play an important role in the construction of continuous-time Markov chains. It often has the interpretation of a waiting time until the occurrence of an event.

4) Gamma Random Variable
Let α and λ be positive numbers. The ***gamma random variable with parameters*** α ***and*** λ is the continuous random variable with p.d.f

$$f(x) = \begin{cases} \lambda e^{-\lambda x} \frac{(\lambda x)^{\alpha - 1}}{\Gamma(\alpha)} & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases} .$$

where

$$\Gamma(\alpha) = \int_0^\infty e^{-x} x^{\alpha - 1} \, dx$$

is the gamma function. If $\alpha = n$ is an integer we have $\Gamma(n) = (n-1)!$ (proof by induction).

The moment generating function is

$$E\left[e^{tX}\right] = \lambda \int_0^\infty e^{tx} \lambda e^{-\lambda x} \frac{(\lambda x)^{\alpha - 1}}{\Gamma(\alpha)} = \begin{cases} \left(\frac{\lambda}{\lambda - t}\right)^\alpha & \text{if } t < \lambda \\ +\infty & \text{otherwise} \end{cases} .$$

and the mean and variance are

$$E[X] = \frac{\alpha}{\lambda}, \quad \text{var}(X) = \frac{\alpha}{\lambda^2}.$$

We write $X \sim \Gamma(n, \lambda)$ to denote a random variable with his distribution.

To show the formula for the m.g.f (in the case of $\alpha = n$ an integer) note that for any $\beta > 0$

$$\int_0^\infty e^{-\beta x}\, dx \;=\; \frac{1}{\beta}\,.$$

and differentiating repeatedly w.r.t. β gives the formula

$$\int_0^\infty e^{-\beta x} x^{n-1}\, dx \;=\; \frac{(n-1)!}{\beta^n}\,,$$

and then set $\beta = \lambda - t$.

Note that $\Gamma(1,\lambda) = Exp(\lambda)$. From the form of the m.g.f., Theorem 1.1.1 and Proposition 1.1.2 we see that if $X_i \sim \Gamma(\alpha_i,\lambda)$ $i = 1, 2$ are independent, then $X_1 + X_2 \sim \Gamma(\alpha_1 + \alpha_2, \lambda)$. In particular $X_1, \cdots, X_n$ are n independent with $X_i \sim Exp(\lambda)$ then $X_1 + \cdots + X_n \sim \Gamma(n,\lambda)$.

5) Bernoulli Random Variable

A Bernoulli random variable models the toss a (possibly unfair coin), or more generally any random experiment with exactly two outcomes. Let p be a number with $0 \le p \le 1$. The ***Bernoulli random variable with parameter*** p is the discrete random variable taking value in $\{0, 1\}$ with

$$p(0) = 1 - p\,, \quad p(1) = p$$

The moment generating function is

$$E\left[e^{tX}\right] \;=\; 1 - p + pe^t\,,$$

and the mean and the variance are

$$E[X] = p\,, \quad \mathrm{var}(X) \;=\; p(1-p)\,.$$

We denote such a random variable by $X \sim \mathcal{B}(1,p)$.

A typical example where Bernoulli random variable occur is the following. Let Y be any random variable, let A be any event, the indicator random variable $\mathbf{1}_A(Y)$ is defined by

$$\mathbf{1}_A(Y) \;=\; \begin{cases} 1 & \text{if } Y \in A \\ 0 & \text{if } Y \notin A \end{cases}$$

Then $\mathbf{1}_A(Y)$ is a Bernoulli random variable with $p = P\{Y \in A\}$.

6) Binomial Random Variable

Consider an experiment which has exactly two outcomes 0 or 1 and is repeated n times, each time independently of each other (i.e., n ***independent trials***). The binomial random variable is the random variable which counts the number of 1 obtained during the n trials. Let p be a number with $0 \le p \le 1$ and let n be a positive integer. The

Bernoulli random variable with parameters n ***and*** p is the random variable which counts the number of 1 occurring in the n outcomes. The p.d.f is

$$p(i) = \binom{n}{i} p^i (1-p)^{n-i}\,, \quad i = 0, 1, \cdots, n\,.$$

The moment generating function is

$$E\left[e^{tX}\right] \;=\; ((1-p) + pe^t)^n\,,$$

and the mean and the variance are

$$E[X] = np\,, \quad \text{var}(X) \;=\; np(1-p)\,.$$

We write $X \sim \mathcal{B}(n, p)$ to denote such a random variable.

The formula for the m.g.f can be obtained directly using the binomial theorem, or simply by noting that by construction $X \sim \mathcal{B}(n, p)$ can be written as a sum of n independent $\mathcal{B}(1, p)$ random variables.

7) Geometric Random Variable

Consider an experiment which has exactly two outcomes 0 or 1 and is repeated as many times as needed until a 1 occurs. The geometric random describes the probability that the first 1 occurs at exactly the n^{th} trial. Let p be a number with $0 \leq p \leq 1$ and let n be a positive integer. The ***geometric random variable with parameter*** p is the random variable with p.d.f

$$p(n) = (1-p)^{n-1} p\,, \quad n = 1, 2, 3, \cdots$$

The moment generating function is

$$E\left[e^{tX}\right] \;=\; \sum_{n=1}^{\infty} e^{tn}(1-p)^{n-1}p \;=\; \begin{cases} \frac{pe^t}{1-e^t(1-p)} & \text{if } e^t(1-p) < 1 \\ 0 & \text{otherwise} \end{cases}\,,$$

The mean and the variance are

$$E[X] = \frac{1}{p}\,, \quad \text{var}(X) \;=\; \frac{1-p}{p^2}\,.$$

We write $X \sim Geo(p)$ to denote this random variable.

8) Poisson Random Variable

Let λ be a positive number. The ***Poisson random variable with parameter*** λ is the discrete random variable which takes values in $\{0, 1, 2, \cdots\}$ and with p.d.f

$$p(n) \;=\; e^{-\lambda}\frac{\lambda^n}{n!} \quad n = 0, 1, 2, \cdots\,.$$

The moment generating function is

$$E\left[e^{tX}\right] = \sum_{n=0}^{\infty} e^{tn}\frac{\lambda^n}{n!}e^{-\lambda} = e^{\lambda(e^t-1)}\,.$$

The mean and the variance are

$$E[X] = \lambda\,, \quad \mathrm{var}(X) = \lambda\,.$$

We write $X \sim \mathcal{P}(\lambda)$ to denote this random variable.

Poisson, Exponential and Gamma random variables are intimately related. Let T_i be a sequence of independent exponential random variables with $T_i \sim Exp(\lambda)$. In particular we have $P(T_i > t) = e^{-\lambda t}$. We denote by $S_n = T_1 + \cdots + T_n$ and we have $S_n \sim \Gamma(n, \lambda)$.

The following results show that we can interpret a Poisson random variable X as the number of events occurring in a certain time interval (say of length 1) and T_i the time elapsing between the occurrence of successive events. Then S_n is the time until the n^{th} event occurs.

Proposition 1.2.2 (Poisson and exponential) *If $X \sim \mathcal{P}(\lambda)$ and $T_i \sim Exp(\lambda)$, $i = 1, 2, cdots...$ are independent with $S_n = T_1 + \cdots + T_n$ we have*

$$P(X = n) = P(n \text{ events occur in } [0,1]) = P(S_n \leq 1, S_{n+1} > 1)$$

Proof: To warm-up note first that

$$P(X = 0) = P(T_1 > 1) = e^{-\lambda}\,.$$

Next we have, by conditioning on T_1,

$$\begin{aligned} P(X=1) &= P(T_1 + T_2 > 1, T_1 \leq 1) \\ &= \int_0^{\infty} P(T_1 + T_2 > 1, T_1 \leq 1 \,|\, T_1 = s) f_{T_1}(s)\, ds \\ &= \int_0^1 P(s + T_2 > 1)\lambda e^{-\lambda s}\, ds \\ &= \int_0^1 e^{-\lambda(1-s)}\lambda e^{-\lambda s}\, ds = \lambda e^{-\lambda} \end{aligned} \tag{1.4}$$

and similarly, by conditionning on the value of S_n

$$\begin{aligned} P(X=n) &= P(S_n + T_{n+1} > 1, S_n \leq 1) \\ &= \int_0^1 P(s + T_{n+1} > 1)\lambda e^{-\lambda s}\frac{(\lambda s)^{n-1}}{(n-1)!}\, ds \\ &= \int_0^1 e^{-\lambda(1-s)}\lambda e^{-\lambda s}\frac{(\lambda s)^{n-1}}{(n-1)!}\, ds = \frac{\lambda^n}{n!}e^{-\lambda} \end{aligned} \tag{1.5}$$

■

1.3 Simulation of random variables

In this section we discuss a few techniques to simulate a given random variable on a computer. The first step which is built-in in any computer is the simulation of a *random number*, i.e., the simulation of a uniform random variable $U([0,1])$, rounded off to the nearest $\frac{1}{10^n}$.

In principle this is not difficult: take ten slips of paper numbered $0, 1, \cdots, 9$, place them in a hat and select successively n slips, with replacement, from the hat. The sequence of digits obtained (with a decimal point in front) is the value of a uniform random variable rounded off to the nearest $\frac{1}{10^n}$. In pre-computer times, tables of random numbers were produced and still can be found. This is of course not the way a actual computer generates a random number. A computer will usually generates a random number by using a deterministic algorithm which produce a pseudo random number which "looks like" a random number For example choose positive integers a, c and m and set

$$X_{n+1} = (aX_n + c) \bmod(m) .$$

The number X_n is either $0, 1, \cdots, m-1$ and the quantity X_n/m is taken to be an approximation of a uniform random variable. One can show that for suitable a, C and m this is a good approximation. This algorithm is just one of many possibles and used in practice. The issue of actually generating a good random number is a nice, interesting, and classical problem in computer science. For our purpose we will simply content ourselves with assuming that there is a "black box" in your computer which generates $U([0,1])$ in a satisfying manner.

We start with a very easy example, namely simulating a discrete random variable X.

Algorithm 1.3.1 (Discrete random variable) *Let X be a discrete random variable taking the values $x_1, x_2, \cdots$ with p.d.f. $p(j) = P\{X = x_j\}$. To simulate X,*

- *Generate a random number $U = U([0,1])$.*
- *Set*

$$X = \begin{cases} x_1 & \text{if } U < p(1) \\ x_2 & \text{if } p(1) \le U < p(1) + p(2) \\ \vdots & \vdots \\ x_n & \text{if } p(1) + \cdots + p(n-1) \le U < p(1) + \cdots p(n) \\ \vdots & \vdots \end{cases}$$

Then X has the desired distribution.

We discuss next two general methods simulating continuous random variable. The first is called the ***inverse transformation method*** and is based on the following

Proposition 1.3.2 *Let $U = U([0,1])$ and let $F = F_X$ be the c.d.f of the continuous random variable X. Then*

$$X = F^{-1}(U)\,,$$

and also

$$X = F^{-1}(1-U)\,.$$

Proof: By definition the c.d.f of the random variable X is a continuous increasing function of F, therefore the inverse function F^{-1} is well-defined and we have

$$P\{F^{-1}(U) \leq a\} \;=\; P\{U \leq F(a)\} = F(a)\,.$$

and this shows that the c.d.f of $F^{-1}(U)$ is F and thus $X = F^{-1}(U)$. To prove the second formula simply note that U and $1-U$ have the same distribution. ∎

So we obtain

Algorithm 1.3.3 (Inversion method for continuous random variable) *Let X be a random variable with c.d.f $F = F_X$. To simulate X*

- **Step 1** *Generate a random number $U \sim \mathcal{U}([0,1])$.*
- **Step 2** *Set $X = F^{-1}(U)$.*

Example 1.3.4 (Simulating an exponential random variable) If $X \sim Exp(\lambda)$ then its c.d.f if

$$F(x) = 1 - e^{-\lambda x}\,.$$

The inverse function F^{-1} is given by

$$1 - e^{-\lambda x} = u \quad \text{iff } u = -\frac{1}{\lambda}\log(1-u)\,.$$

Therefore we have $F^{-1}(u) = -\frac{1}{\lambda}\log(1-u)$. So if $U \sim \mathcal{U}([0,1])$ then

$$X = -\frac{1}{\lambda}\log(1-U) \sim Exp(\lambda)$$

Note also that that U and $1-U$ have the same distribution so

$$X = -\frac{1}{\lambda}\log(U) \sim Exp(\lambda)\,,$$

is also an exponential random variable.

The inversion method is most straightforward when there is an explicit formula for the inverse function F^{-1}. In many examples however a such a nice formula is not available. Possible remedies to that situation is to solve $F(X) = U$ numerically for example by Newton method.

Another method for simulating a continuous random variable is the ***rejection method***. Suppose we have a method to simulate a random variable with p.d.f $g(x)$ and that we want to simulate the random variable with p.d.f $f(x)$. The following algorithm is due to Von Neumann.

Algorithm 1.3.5 (Rejection method for continuous random variable). *Let X be a random variable with p.d.f $f(x)$ and let Y be a random variable with p.d.f $g(x)$. Furthermore assume that there exists a constant C such that*

$$\frac{f(y)}{g(y)} \leq C\,, \quad \text{for all } y\,.$$

To simulate X

- **Step 1** *Simulate Y with density g.*
- **Step 2** *Simulate a random number U.*
- **Step 3** *If*

$$U \leq \frac{f(Y)}{g(Y)C}$$

set $X = Y$. Otherwise return to Step 1.

That the algorithm does the job is the object of the following proposition.

Proposition 1.3.6 *The random variable X generated by the rejection method has p.d.f $f(x)$. If N is the number of times the algorithm is run until one value is accepted then N is a geometric random variable with parameter $\frac{1}{C}$.*

Proof: To obtain a value of X we will need in general to iterate the algorithm a random number of times We generate random variables $Y_1, \cdots, Y_N$ until Y_N is accepted and then set $X = Y_N$. We need to verify that the p.d.f of X is actually $f(x)$.

Then we have

$$\begin{aligned} P\{X \le x\} &= P\{Y_N \le x\} = P\left\{Y \le x \,|\, U \le \frac{f(Y)}{Cg(Y)}\right\} \\ &= \frac{P\left\{Y \le x\,, U \le \frac{f(Y)}{Cg(Y)}\right\}}{P\left\{U \le \frac{f(Y)}{Cg(Y)}\right\}} \\ &= \frac{\int_{-\infty}^{\infty} P\left\{Y \le x\,, U \le \frac{f(Y)}{Cg(Y)} \,|\, Y = y\right\} g(y)\, dy}{P\left\{U \le \frac{f(Y)}{Cg(Y)}\right\}} \\ &= \frac{\int_{-\infty}^{x} P\left(U \le \frac{f(y)}{Cg(y)}\right) g(y)\, dy}{P\left(U \le \frac{f(Y)}{Cg(Y)}\right)} \\ &= \frac{\int_{-\infty}^{x} \frac{f(y)}{Cg(y)} g(y)\, dy}{P\left(U \le \frac{f(Y)}{Cg(Y)}\right)} = \frac{\int_{-\infty}^{x} f(y)\, dy}{CP\left(U \le \frac{f(Y)}{Cg(Y)}\right)}\,. \end{aligned}$$

If we let $x \to \infty$ we obtain that $CP\left(U \le \frac{f(Y)}{Cg(Y)}\right) = 1$ and thus

$$P(X \le x) \;=\; \int_{-\infty}^{x} f(x)\, dx\,.$$

and this shows that X has p.d.f $f(x)$.

In addition the above argument that at each iteration of the algorithm the value for X is accepted with probability

$$P\left(U \le \frac{f(Y)}{Cg(Y)}\right) = \frac{1}{C}$$

independently of the other iterations. Therefore the number of iterations needed is $Geo(\frac{1}{C})$ with mean C. ∎

In order to decide whether this method is efficient of not, we need to ensure that rejections occur with small probability. Therefore the ability to choose a reasonably small C will ensure that the method is efficient.

Example 1.3.7 Let X be the random variable with p.d.f

$$f(x) \;=\; 20x(1-x)^3\,, \quad 0 < x < 1\,.$$

Since the p.d.f. is concentrated on $[0,1]$ let us take

$$g(x) = 1 \quad 0 < x < 1\,.$$

To determine C such that $f(x)/g(x) \leq C$ we need to maximize the function $h(x) \equiv f(x)/g(x) = 20x(1-x)^3$. Differentiating gives $h'(x) = 20\left((1-x)^3 - 3x(1-x)^2\right)$ and thus the maximum is attained at $x = 1/4$. Thus

$$\frac{f(x)}{g(x)} \leq 20\frac{1}{4}\left(\frac{3}{4}\right)^3 = \frac{135}{64} \equiv C\,.$$

We obtain

$$\frac{f(x)}{Cg(x)} \;=\; \frac{256}{27}x(1-x)^3$$

and the rejection method is

- **Step 1** Generate random numbers U_1 and U_2.
- **Step 2** If $U_2 \leq \frac{256}{27}U_1(1-U_1)^3$, stop and set $X = U_1$. Otherwise return to step 1.

The average number of accepted iterations is 135/64.

Example 1.3.8 (Simulating a normal random variable) Note first that to simulate a normal random variable $Y \sim \mathcal{N}(\mu, \sigma^2)$ it is enough to simulate $X \sim \mathcal{N}(0,1)$ and then set $Y = \sigma X + \mu$.

Let us first consider the random variable Z whose density is

$$f(x) \;=\; \frac{2}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\,, \quad 0 \leq x \leq \infty\,.$$

and we have $Z = |X|$. We simulate Z by using the rejection method with

$$g(x) = e^{-x} \quad 0, x < \infty\,,$$

i.e., $Y = Exp(1)$. To find C we note that

$$\frac{f(x)}{g(x)} \;=\; \sqrt{\frac{2e}{\pi}}e^{-\frac{(x-1)^2}{2}} \;\leq\; \frac{2e}{\pi} \equiv C\,.$$

One generates Z using the rejection method. To generate $X \sim \mathcal{N}(0,1)$ from Z one generate a discrete random variable S with takes value $+1$ and -1 with probability $\frac{1}{2}$ and then set $X = SZ$. For example $S = 2\lfloor 2U \rfloor - 1$ if $U \sim \mathcal{N}(0,1)$.

- **Step 1** Generate two random numbers U and V and an exponential random variable Y.
- **Step 2** If $U \leq \exp -\frac{(Y-1)^2}{2}$ set $Z = Y$ and $X = (2\lfloor 2V \rfloor - 1)Z$. Otherwise return to step 1.

For particular random variables many special techniques have been devised. We give here some examples.

Example 1.3.9 (Simulating a geometric random variable) The c.d.f of the geometric random variable $X = Geom_p$ is given by

$$F(n) = P(X \leq n) = 1 - P(X > n) = 1 - \sum_{k=n+1}^{\infty} (1-p)^{n-1}p = 1 - (1-p)^n$$

The exponential random variable $Y = Exp_\lambda$ has c.d.f $1 - e^{-\lambda x}$.

For any positive real number let $\lceil x \rceil$ denote the smallest integer greater than or equal to x, e.g. $\lceil 3.72 \rceil = 4$. Then we claim that if $Y \sim Exp(\lambda)$ then

$$\lceil Y \rceil \sim Geo(p) \qquad \text{with } p = 1 - e^{-\lambda}\,.$$

Indeed we have

$$P(\lceil Y \rceil \leq n) = P(Y \leq n) = 1 - e^{-\lambda n}\,.$$

Thus we obtain

Algorithm 1.3.10 (Geometric random variable)

- **Step 1** *Generate a random number U.*
- **Step 2** *Set $X = \lceil \frac{\log(U)}{\log(1-p)} \rceil$*

Then $X \sim Geo(p)$.

Example 1.3.11 (Simulating a Poisson random variable) To simulate a Poisson random variable the Algorithm 1.3.1 is not very convenient since it requires storing the number $e^{-\lambda}\sum_{k=1}^{n} \frac{\lambda^k}{k!}$.

Alternatively we can use the relation between exponential random variable and Poisson random variable spelled out in Proposition 1.2.2. We generate independent exponential RV until they sum up to more than 1. That is we set $X = n$ if

$$\begin{aligned}
\sum_{k=1}^{n} -\frac{1}{\lambda}\ln(U_i) \leq \quad 1 \quad &< \sum_{k=1}^{n} -\frac{1}{\lambda}\ln(U_k) \\
\ln(\prod_{k=1}^{n} U_k) \geq \quad -\lambda \quad &> \ln(\prod_{k=1}^{n+1} U_k) \\
\prod_{k=1}^{n} U_k \geq \quad e^{-\lambda} \quad &> \prod_{k=1}^{n+1} U_k
\end{aligned}$$

and thus we have

Algorithm 1.3.12 (Poisson random variable)

- **Step 1** *Generate random number* $U_1, U_2, U_3, \cdots$.
- **Step 2** *Set* $X = n$ *is* $n+1 = \inf_j\{j \,:\, \prod_{k=1}^{j} U_j < e^{-\lambda}\}$

Then $X \sim \mathcal{P}(\lambda)$.

Example 1.3.13 (Simulating a Gamma random variable). Using the fact that $\Gamma(n,\lambda)$ is a sum of n independent $Exp(\lambda)$ one immediately obtain

Algorithm 1.3.14 (Gamma random variable)

- **Step 1** *Generate* n *random number* $U_1, \cdots, U_n$.
- **Step 2** *Set* $X = \sum_{k=1}^{n} -\frac{1}{\lambda}\log(U_k)$.

Then $X \sim \Gamma(n,p)$.

An alternative method which is more efficient if n large (and works for non-integer parameters) is to use the rejection method with $g(x)$ the pdf of $Y \sim Exp(\lambda/n)$ (see exercises).

Finally we give an elegant algorithm which generates 2 independent normal random variables without rejection.

Example 1.3.15 (Simulating a normal random variable: Box-Müller). We show a simple way to generate 2 independent standard normal random variables X and Y. The joint p.d.f. of X and Y is given by

$$f(x,y) \,=\, \frac{1}{2\pi} e^{-\frac{(x^2+y^2)}{2}} \,.$$

Let us change into polar coordinates (r,θ) with $r^2 = x^2+y^2$ and $\tan(\theta) = y/x$. The change of variables formula gives

$$f(x,y)\,dxdy \,=\, re^{-\frac{r^2}{2}}\,dr\frac{1}{2\pi}d\theta\,.$$

Consider further the change of variables set $s = r^2$ so that

$$f(x,y)\,dxdy \,=\, \frac{1}{2}e^{-\frac{s}{2}}\,ds\frac{1}{2\pi}d\theta\,.$$

The right-hand side is to be the joint p.d.f of the two independent random variables $S \sim Exp(1/2)$ and $\Theta \sim \mathcal{U}([0,2\pi])$.

Therefore we obtain

Algorithm 1.3.16 (Standard normal random variable)

- **Step 1** *Generate two random number U_1 and U_2*
- **Step 2** *Set*

$$\begin{aligned} X &= \sqrt{-2\log(U_1)}\cos(2\pi U_2) \\ Y &= \sqrt{-2\log(U_1)}\sin(2\pi U_2) \end{aligned} \tag{1.6}$$
$$\tag{1.7}$$

Then X and Y are independent and $\sim N_{0,1}$.

We give next a couple examples of how to ***generate random vectors*** $\mathbf{X} \in \mathbb{R}^d$. We start with a simple rejection method for vectors which are uniformly distributed on some region $G \subset \mathbb{R}^d$. The pdf of $\mathbf{X}$ is then

$$f_{\mathbf{X}}(\mathbf{x}) = \begin{cases} \frac{1}{|G|} & \mathbf{x} \in G \\ 0 & \text{otherwise} \end{cases}$$

If we do not know the volume $|G|$ then the p.d.f is known only up to a (unknown) constant. The idea consists in enclosing the set G into a set V on which we know how to generate a uniform distribution (for example a cube or a rectangle). This lead to

Algorithm 1.3.17 (Generating a random vector on set $G \subset \mathbb{R}^d$) *Suppose $G \subset V \subset \mathbb{R}^d$.*

- **Step 1** *Generate a random vector $\mathbf{Y}$ uniformly distributed on V.*
- **Step 2** *If $\mathbf{Y} \in G$ set $\mathbf{X} = \mathbf{Y}$. Otherwise return to step 1.*

Then X is uniformly distributed on G and the acceptance rate is $\frac{|G|}{|V|}$.

Example 1.3.18 (Uniform distribution on the unit d-dimensional ball). Consider the ball of radius 1, i.e., $B_d = \{x \in \mathbb{R}^d\,;\, x_1^2 + \cdots + x_d^2 \leq 1\}$ and enclose it in the cube of side length 2 centered at 0, i.e. $C_d = \{x \in \mathbb{R}^d\,;\, -1 \leq x_1 \leq 1, \cdots, -1 \leq x_d \leq 1\}$.

Algorithm 1.3.19 (Generating a random vector on the unit ball).

- **Step 1** *Let $\mathbf{Y} = (2U_1 - 1, \cdots 2U_d - 1)$ with independent $U_i \sim \mathcal{U}([0,1])$.*
- **Step 2** *If $Y_1^2 + \cdots + Y_d^2 \leq 1$ set $\mathbf{X} = \mathbf{Y}$, otherwise return to step 1.*

Then $\mathbf{X}$ has uniform distribution on the ball of radius 1.

It is interesting to study the acceptance rate as a function of the dimension d. We have

$$\text{Acceptance probability } = \frac{\text{volume of the sphere}}{\text{volume of the cube}} = \frac{\frac{\pi^{d/2}}{d\Gamma(d/2)}}{2^d}$$

To see how it behaves for large d takes $d = 2l$ even and then the acceptance rate is

$$\frac{\frac{\pi^l}{2l(l-1)!}}{2^{2l}} = \frac{1}{2l!}\left(\frac{\pi}{4}\right)^l \longrightarrow 0 \ \ (\text{as } l \to \infty).$$

Note that this converges to 0 extremely fast! For example for a 6-dimensional ball the acceptance rate is $\frac{1}{12}\left(\frac{\pi}{4}\right)^3 = 0.0403$ and only one in about 25 vector is accepted. So for large d this algorithm is quite useless!

Example 1.3.20 (Uniform distribution on a d-hypersphere). Rather than a rejection method the idea is to use a random vector whose distribution is spherically symmetric and then project it on the sphere of radius 1. For example we can use d independent normal random $Y_i \sim \mathcal{N}(0,1)$ with joint p.d.f

$$f_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{(2\pi)^{d/2}} e^{-\frac{\sum_{i=1}^d y_i^2}{2}} = \frac{1}{(2\pi)^{d/2}} e^{-\frac{r^2}{2}}$$

where $r = \|\mathbf{y}\|$ is the norm of $\mathbf{y}$. The p.d.f of $\mathbf{Y}$ is obviously spherically symmetric since its distribution depends only on $\|y\|$ and thus have

Algorithm 1.3.21 (Generating a random vector on $\{x \in \mathbb{R}^d\,;\, x_1^2 + \cdots + x_d^2 = 1\}$.

- **Step 1** *Generate* $\mathbf{Y} = (Y_1, \cdots, Y_d)$ *with independent* $Y_i \sim \mathcal{N}(0,1)$.
- **Step 2** *Set* $X = \frac{\mathbf{Y}}{\|\mathbf{Y}\|}$.

Then X has uniform distribution on the sphere of radius 1.

We can then use this to provide a rejection free method to generate a random vector on the unit ball. To do this note that if $\mathbf{X}$ is uniformly distributed on the unit ball and $R = \|\mathbf{X}\|$ then using spherical coordinates in d dimensions we can prove that

$$P(R \le r) = r^d, \quad 0 \le r \le 1$$

and we can generate R by the inverse method. This leads to

Algorithm 1.3.22 (Generating a random vector on $\{x \in \mathbb{R}^d\,;\, x_1^2 + \cdots + x_d^2 \le 1\}$.

- **Step 1** *Generate a uniform vector* $\mathbf{Y}$ *of the unit d -hypersphere and* $U \sim \mathcal{U}([0,1])$.
- **Step 2** *Set* $\mathbf{X} = U^{1/n}\mathbf{Y}$.

Then $\mathbf{X}$ *has uniform distribution on the ball of radius* 1.

1.4 Markov, Chebyshev, and Chernov

We recall simple techniques for bounding the ***tail distribution*** of a random variable, i.e., bounding the probability that the random variable takes value far from the its mean.

Our first inequality, called ***Markov's inequality*** simply assumes that we know the mean of X.

Proposition 1.4.1 (Markov's Inequality) *Let X be a random variable which assumes only nonnegative values, i.e. $P(X \geq 0) = 1$. Then for any $a > 0$ we have*

$$P(X \geq a) \leq \frac{E[X]}{a}.$$

Proof: For $a > 0$ let us define the random variable

$$I_a = \begin{cases} 1 & \text{if } X \geq a \\ 0 & \text{otherwise} \end{cases}.$$

Note that, since $X \geq 0$ we have

$$I_a \leq \frac{X}{a} \tag{1.8}$$

and that since I_a is a binomial random variable

$$E[I_a] = P(X \geq a).$$

Taking expectations in the inequality (1.8) gives

$$P(X \geq a) = E[I_a] \leq E\left[\frac{X}{a}\right] = \frac{E[X]}{a}. \quad \blacksquare$$

Example 1.4.2 (Flipping coins) Let us flip a fair coin n times and let us define the random variables X_i, $i = 1, 2, \cdots, n$ by

$$X_i = \begin{cases} 1 & \text{if the } i^{th} \text{ coin flip is head} \\ 0 & \text{otherwise} \end{cases}.$$

Then each X_i is a Bernoulli random variable and $S_n = X_1 + \cdots X_n = B_{n,\frac{1}{2}}$ is a binomial random variable.

Let us the Markov inequality to estimate the probability that at least 75% of the n coin flips are head. Since $E[S_n] = \frac{n}{2}$ the markov's inequality tells us that

$$P(S_n \geq \frac{3n}{4}) \leq \frac{E[S_n]}{3n/4} = \frac{n/2}{3n/4} = \frac{2}{3}.$$

As we will see later this is an extremely lousy bound but note that we obtained it using only the value of the mean and nothing else.

Our next inequality, which we can derive from Markov's inequality, involves now the variance of X. This is called ***Chebyshev's inequality***.

Proposition 1.4.3 (Chebyshev's Inequality) *Let X be a random variable with $E[X] = \mu$ and $Var(X) = \sigma^2$. Then for any $a > 0$ we have*

$$P(|X - \mu| \geq a) \leq \frac{\sigma^2}{a^2}.$$

Proof: Observe first that

$$P(|X - \mu| \geq a) = P((X - \mu)^2 \geq a^2).$$

Since $(X - \mu)^2$ is a nonnegative random variable we can apply Markov's inequality and obtain

$$P(|X - \mu| \geq a) \leq \frac{E[(X - \mu)^2]}{a^2} = \frac{\text{var}(X)}{a^2}. \quad \blacksquare$$

Let us apply this result to our coin flipping example

Example 1.4.4 (Flipping coins, cont'd) Since S_n has mean $n/2$ and variance $n/4$ Chebyshev's inequality tells us that

$$\begin{aligned} P\left(S_n \geq \frac{3n}{4}\right) &= P\left(S_n - \frac{n}{2} \geq \frac{n}{4}\right) \\ &\leq P\left(\left|S_n - \frac{n}{2}\right| \geq \frac{n}{4}\right) \\ &\leq \frac{n/4}{(n/4)^2} = \frac{4}{n}. \end{aligned} \tag{1.9}$$

This is significantly better that the bound provided by Markov's inequality! Note also that we can do a bit better by noting that the distribution of S_n is symmetric around its mean and thus we can replace $4/n$ by $2/n$.

We can do better if we know all moments of the random variable X, for example if we know the moment generating function $M_X(t)$ of the random variable X. The inequalities in the following theorems are usually called ***Chernov bounds*** or ***exponential Markov inequality***.

Proposition 1.4.5 (Chernov's bounds) *Let X be a random variable with moment generating function $M_X(t) = E[e^{tX}]$.*

- *For any a and any $t > 0$ we have*

$$P(X \geq a) \leq \min_{t \geq 0} \frac{E[e^{tX}]}{e^{ta}}.$$

- *For any a and any $t < 0$ we have*

$$P(X \leq a) \leq \min_{t<0} \frac{E[e^{tX}]}{e^{ta}}.$$

Proof: This follows from Markov inequality. For $t > 0$ we have

$$P(X \geq a) = P(e^{tX} > e^{ta}) \leq \frac{E[e^{tX}]}{e^{ta}}.$$

Since $t > 0$ is arbitrary we obtain

$$P(X \geq a) \leq \min_{t\geq 0} \frac{E[e^{tX}]}{e^{ta}}.$$

Similarly for $t < 0$ we have

$$P(X \leq a) = P(e^{tX} > e^{ta}) \leq \frac{E[e^{tX}]}{e^{ta}},$$

and thus

$$P(X \geq a) \leq \min_{t\leq 0} \frac{E[e^{tX}]}{e^{ta}}. \blacksquare$$

Let us consider again our flipping coin examples

Example 1.4.6 (Flipping coins, cont'd) Since S_n is a binomial $B_{n,\frac{1}{2}}$ random variable its moment generating function is given by $M_{S_n}(t) = (\frac{1}{2} + \frac{1}{2}e^t)^n$. To estimate $P(S_n \geq 3n/4)$ we apply Chernov bound with $t > 0$ and obtain

$$P\left(S_n \geq \frac{3n}{4}\right) \leq \frac{(\frac{1}{2} + \frac{1}{2}e^t)^n}{e^{\frac{3nt}{4}}} = \left(\frac{1}{2}e^{-\frac{3t}{4}} + \frac{1}{2}e^{\frac{t}{4}}\right)^n.$$

To find the optimal bound we minimize the function $f(t) = \frac{1}{2}e^{-\frac{3t}{4}} + \frac{1}{2}e^{\frac{t}{4}}$. The mimimum is at $t = \log 3$ and

$$f(\log(3)) = \frac{1}{2}(e^{-\frac{3}{4}\log(3)} + e^{\frac{1}{4}\log(3)}) = \frac{1}{2}e^{\frac{1}{4}\log(3)}(e^{-\log 3} + 1) = \frac{2}{3}3^{\frac{1}{4}} \simeq 0.877$$

and thus we obtain

$$P\left(S_n \geq \frac{3n}{4}\right) \leq 0.877^n.$$

This is course much better than $2/n$. For $n = 100$ Chebyshev inequality tells us that the probability to obtain 75 heads is not bigger than 0.02 while the Chernov bounds tells us that it is actually not greater than 2.09×10^{-6}.

1.5 Limit theorems

In this section we study the behavior, for large n of a ***sum of independent identically distributed variables*** (abbreviated ***i.i.d.***). Let $X_1, X_2, \cdots$ be a sequence of independent random variables where all X_i's have the same distribution. Then we denote by S_n the sum

$$S_n = X_1 + \cdots + X_n \,.$$

The random variable $\frac{S_n}{n}$ is called the ***empirical average***. You can imagine that X_i represent the output of some experiment and then S_n/n is the random variable obtained by averaging the outcomes of n successive experiments, performed independently of each other.

Under suitable conditions S_n will exhibit a universal behavior which does not depend on all the details of the distribution of the X_i's but only on a few of its charcteristics, like the mean or the variance.

The first result is the ***weak law of large numbers***. It tells us that if we perform a large number of independent trials the average value of our trials is close to the mean with probability close to 1. The proof is not very difficult, but it is a very important result!

Theorem 1.5.1 (The weak Law of Large Numbers) *Let $X_1, X_2, \cdots$ be a sequence of independent identically distributed random variables with mean μ and variance σ^2. Let*

$$S_n = X_1 + \cdots + X_n \,.$$

Then for any $\epsilon > 0$

$$\lim_{n\to\infty} P\left(\left|\frac{S_n}{n} - \mu\right| \geq \epsilon\right) \,=\, 0\,.$$

Proof: : By the linearity of expectation we have

$$E\left[\frac{S_n}{n}\right] \,=\, \frac{1}{n}E[X_1 + \cdots + X_n] \,=\, \frac{n\mu}{n} \,=\, \mu\,.$$

i.e. the mean of S_n/n is μ. Furthermore by the independence of $X_1, \cdots, X_n$ we have

$$\mathrm{var}\left(\frac{S_n}{n}\right) \,=\, \frac{1}{n^2}\mathrm{var}(S_n) \,=\, \frac{1}{n^2}\mathrm{var}(X_1 + \cdots + X_n) \,=\, \frac{n\sigma^2}{n^2} \,=\, \frac{\sigma^2}{n}\,.$$

Applying Chebyshev's inequality we obtain

$$P\left\{\left|\frac{S_n}{n} - \mu\right| \geq \epsilon\right\} \,\leq\, \frac{\mathrm{var}\left(\frac{S_n}{n}\right)}{\epsilon^2} \,=\, \frac{\sigma^2}{\epsilon^2}\frac{1}{n}\,,$$

and for any $\epsilon > 0$ the right hand sides goes to 0 as n goes to ∞. ∎

The weak law of large numbers tells us that if we perform a large number of independent trials then the average value of our trials is close to the mean with probability close to 1. The proof is not very difficult, but it is a very important result. There is a strengthening of the weak law of large numbers called the ***strong law of large numbers***

Theorem 1.5.2 (Strong Law of Large Numbers) *Let $X_1, X_2, \cdots$ be a sequence of independent identically distributed random variables with mean μ. Then S_n/n converges to μ with probability 1, i.e.,*

$$P\left\{\lim_{n\to\infty}\frac{S_n}{n}=\mu\right\} = 1\,.$$

The strong law of large numbers is useful in many respects. Imagine for example that you are simulating a sequence of i.i.d random variables and that you are trying to determine the mean μ. The strong law of large numbers tells you that, in principle, it is enough to do 1 simulation for a sufficiently long time to produce the mean. The weak law of large numbers tells you something a little weaker: with very large probability you will obtain the mean. Based on the weak law of large numbers only you might want to repeat your experiment a number of times to make sure you were not unlucky and hit an event of small probability. The strong law of large numbers tells you not to worry.

Proof: The proof of the strong law of large numbers use more advanced tools that we are willing to use here. ∎

Finally we discuss the ***central limit theorem***. The law of large number and cramer's theorem deals with large fluctuations for S_n/n, that is with the probability that S_n/n is at a distance away from the mean which is of order 1. In particular these fluctuations vanish when $n \to \infty$. For example we can ask if there are non trivial fluctuations of order $\frac{1}{n^\alpha}$ for some $\alpha > 0$. One can easily figure out which power α has to be chosen. Since $E[S_n] = n\mu$ $\mathrm{var}(S_n) = n\sigma^2$ we see that the ratio

$$\frac{S_n - n\mu}{\sqrt{n}\sigma}$$

has mean 0 and variance 1 for all n. This means that fluctuation of order $1/\sqrt{n}$ may be non trivial. The Central limit theorem shows not the fluctuation of order $1/\sqrt{n}$ of S_n are in fact universal: for large n they behave like a normal random variable, that is

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

or

$$\frac{S_n}{n} \sim \mu + \frac{1}{\sqrt{n}}N(0,\sigma^2)\,.$$

What we exactly mean by $\sim$ is given in

Theorem 1.5.3 (Central Limit Theorem) *Let $X_1, X_2, \cdots$ be a sequence of independent identically distributed random variables with mean μ and variance $\sigma^2 > 0$. Then for any $-\infty \le a \le b \le \infty$ we have*

$$\lim_{n\to\infty} P\left(a \le \frac{S_n - n\mu}{\sqrt{n}\sigma} \le b\right) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{x^2}{2}}\, dx\,.$$

Proof: We will not give the complete proof here but we will prove that the moment generating function of $\frac{S_n - n\mu}{\sqrt{n}\sigma}$ converges to the moment generating of $N(0,1)$ as $n \to \infty$.

Let by $X_i^* = \frac{X_i - \mu}{\sigma}$ then $E[X_i^*] = 0$ and $\mathrm{var}(X_i^*) = 1$. If $S_n^* = X_1^* + \cdots X_n^*$ then

$$\frac{S_n - n\mu}{\sqrt{n}\sigma} \;=\; \frac{S_n^*}{\sqrt{n}}\,.$$

Therefore without loss of generality we can assume that $\mu = 0$ and $\sigma = 1$.

Let $M(t) = M_{X_i}(t)$ denote the moment generating function of the R.V. X_i then we have $M(0) = 1$, $M'(0) = E[X_i] = \mu = 0$ and $M''(0) = \mathrm{var}(X) = 1$. Using independence we have

$$M_{\frac{S_n}{\sqrt{n}}}(t) \;=\; E\left[e^{t\frac{S_n}{\sqrt{n}}}\right] \;=\; E\left[e^{\frac{t}{\sqrt{n}}(X_1+\cdots X_n)}\right] \;=\; \left(M\left(\frac{t}{\sqrt{n}}\right)\right)^n\,.$$

Recall that the m.g.f. of $N(0,1)$ is given by $e^{t^2/2}$, so we need to show that $M_{\frac{S_n}{\sqrt{n}}}(t) \to e^{t^2/2}$ as $n \to \infty$. Let

$$u(t) = \log M(t)\,, \quad u_n(t) \;=\; \log M_{\frac{S_n}{\sqrt{n}}}(t)$$

and we will show that $u_n(t) \to t^2/2$ as $n \to \infty$. We have

$$u_n(t) \;=\; \log \phi_n(t) \;=\; n \log \phi\left(\frac{t}{\sqrt{n}}\right)\,.$$

Note that

$$\begin{aligned} u(0) &= \log M(0) \;=\; 0 \\ u'(0) &= \frac{M'(0)}{M(0)} \;=\; \mu \;=\; 0 \\ u''(0) &= \frac{M''(0)M(0) - M'(0)^2}{M(0)^2} \;=\; \sigma^2 = 1\,. \end{aligned}$$

By using L'Hospital rule twice we obtain

$$\begin{aligned} \lim_{n\to\infty} u_n(t) &= \lim_{s\to\infty} \frac{\phi(t/\sqrt{s})}{s^{-1}} \\ &= \lim_{s\to\infty} \frac{\phi'(t/\sqrt{s})t}{2s^{-1/2}} \\ &= \lim_{s\to\infty} \phi''(t/\sqrt{s})\frac{t^2}{2} \;=\; \frac{t^2}{2}\,. \end{aligned}$$

Therefore $\lim_{n\to\infty}\phi_n(t) = e^{t^2}2$. One can show with a non-negligible amount of work that this implies that the c.d.f of $S_n/\sqrt{n}$ converges to the c.d.f of $N(0,1)$. ∎

1.6 Monte-Carlo algorithm

1.6.1 Introduction and examples

The ***simple Monte-Carlo method*** is a probabilistic algorithm using sums of independent random variables the law of large numbers to estimate a deterministic quantity which we will call μ. The basic idea is to write μ as the expectation of some random variable

$$\mu = E[h(X)]\,,$$

and then use the law of large numbers.

Algorithm 1.6.1 (Simple Monte-Carlo sampling) *Let X be a random variable. Simulate N copies X_i, $i = 1, \cdots, N$ of X, then*

$$I_N \;=\; \frac{1}{N}\sum_{k=1}^{N} h(X_k)$$

*is an **unbiased estimator** for μ, that is we have*

1. *For all N we have $E[I_N] = \mu$ (unbiased).*
2. *By the strong law of large numbers, with probability 1, $\lim_{N\to\infty} I_N = \mu$.*

Let us give some illustrative examples.

Example 1.6.2 (Estimating the number π). We construct a random algorithm to generate the number π. Consider a circle of radius 1 that lies inside a 2×2 square. The square has area 4 and the circle has area π. Suppose we pick a point at random within the square and define

$$X \;=\; \begin{cases} 1 & \text{if the point is inside the circle} \\ 0 & \text{otherwise} \end{cases}$$

and $E[X] = P(X = 1) = \pi/4$ and

This example is a particular case of the ***hit-or-miss method*** that we already encountered when learning how to generate random vectors.: Suppose you want to estimate the volume of the set B in $\mathbf{R}^d$ and that you know the volume of a set A which contains B. The hit-or-miss method consists in choosing n points in A uniformly at random and use the fraction of the points that land in B as an estimate for the volume of B.

Example 1.6.3 (Computing integrals.) Another class of examples where Monte-Carlo methods can be applied is the computation of integrals. Suppose you want to compute the integral

$$I_1 = \int_0^1 \frac{e^{\sqrt{x}} - e^{cos(x^3)}}{3 + \cos(x)}\, dx\,.$$

or more generally

$$I_2 = \int_S h(\mathbf{x}) d\mathbf{x}$$

where S is a subset of $\mathbf{R}^d$ and h is a given real-valued function on S. A special example is the function $h = 1$ on S in which case you are simply trying to compute the volume of S. Another example is

$$I_3 = \int_{\mathbf{R}^d} h(\mathbf{x}) f(\mathbf{x})\, d\mathbf{x}\,.$$

where h is a given real-valued function and f is a p.d.f of some random vector on $\mathbf{R}^d$. All these examples can be written as expectations of a suitable random variable. Indeed we have

$$I_3 = E[h(\mathbf{X})] \quad \text{where } \mathbf{X} \text{ has p.d.f } f(\mathbf{x})\,.$$

We have also

$$I_1 = E[h(U)] \quad \text{where } U = U_{[0,1]}\,.$$

To write I_2 has an expectation choose a random vector such that its p.d.f f satisfies $f(\mathbf{x}) > 0$ for every $\mathbf{x} \in S$. Extend h to $\mathbf{R}^d$ by setting $k = 0$ if $x \notin S$. Then

$$I_2 = \int_{\mathbf{R}^d} h(\mathbf{x})\, dx = \int_{\mathbf{R}^d} \frac{h(\mathbf{x})}{f(\mathbf{x})} f(\mathbf{x})\, d\mathbf{x} = E\left[\frac{h(\mathbf{X})}{f(\mathbf{X})}\right].$$

Note that you have a considerable freedom in choosing f and this is what lies behind the idea of importance sampling, see below.

Another class of examples where Monte-Carlo are useful comes from ***reliability theory*** which is branch of operation research. We will see several class of models of that type later on.

Imagine a system consisting of multiple components, where each component has a certain probability to function (or to fail). The basic question is compute the probability that the system itself works (or fail).

To describe such systems we associate to each component i a Bernoulli random variable $X_i \sim \mathcal{B}(1, p_i)$ where

$$X_i = \begin{cases} 1 & \text{if the } i^{th} \text{ component is functioning} \\ 0 & \text{if the } i^{th} \text{ component has failed} \end{cases}.$$

The state of a system with n component is then described by the state vector $\mathbf{X} = (X_1, \cdots, X_d)$ and usually we will assume that the X_i are independent.

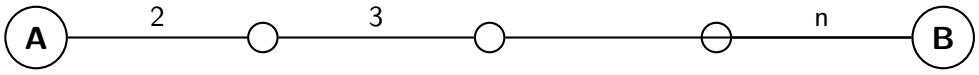
To describe if the system works or not we introduce the ***structure function*** $H(\mathbf{X})$ which is defined by

$$H(\mathbf{X}) = \begin{cases} 1 & \text{if the system is functioning} \\ 0 & \text{if the system has failed} \end{cases},$$

and we have

$$E[H(\mathbf{X})] \;=\; P(H(\mathbf{X}) = 1) \;=\; P(\text{ the system is functioning}) \,.$$

For example consider the ***series structure***



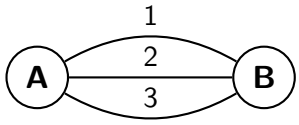
functions only if every component functions and thus its structure function is

$$H(\mathbf{X}) = X_1 X_2 \cdots X_n \,.$$

and

$$P(\text{the system is functioning}) = E[H(\mathbf{X})] = p_1 p_2 \cdots p_n \,.$$

The ***parallel structure*** consists of n component and the system functions if at least one component functions



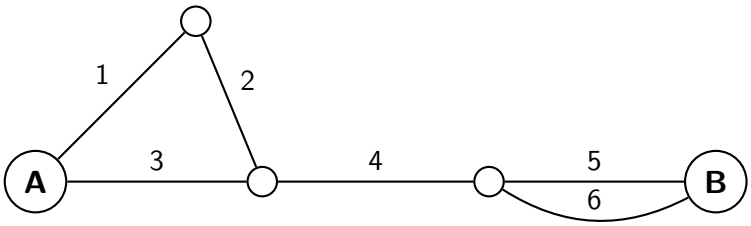
and therefore its structure function has the form

$$\begin{aligned} H(\mathbf{X}) &= \max(X_1, \cdots X_n) \\ &= 1 - (1 - X_1)(1 - X_2) \cdots (1 - X_n) \,. \end{aligned}$$

and so

$$P(\text{the system is functioning}) = 1 - (1 - p_1)(1 - p_2) \cdots (1 - p_n)$$

If the system can be decomposed in a combined series/parallel structure, for example



has the structure function

$$H((\mathbf{X}) = [1 - (1 - X_1X_2)(1 - X_3)]\, X_4\, [1 - (1 - X_5)(1 - X_6)]$$

and

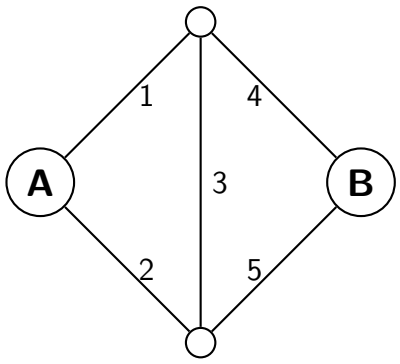
$$P(\text{the system is functioning}) = [1 - (1 - p_1p_2)(1 - p_3)]\, p_4\, [1 - (1 - p_5)(1 - p_6)]\,.$$

In general, however, the structure function is not easy to write. There is a general, principled, way to find the structure function which involves the concept of minimal path. A ***path*** A is a collection of components, $A \subset \{1, \cdots, n\}$ such that if the components in A are functioning (i.e. $X_i = 1$ for $i \in A$), then the system functions (i.e., $H(\mathbf{X}) = 1$). A ***minimal path*** A is a path such that if you remove any component from A the system does not function. Since for the system to function at least one minimal path must function we obtain

Theorem 1.6.4 (Structure function formula)

$$\begin{aligned} H(\mathbf{X}) &= \max\left\{\prod_{j\in A} X_j\,;A \textit{ minimal path}\right\} \\ &= 1 - \prod_{A \textit{ minimal path}} (1 - \prod_{j\in A} X_j) \end{aligned}$$

But it should be noted this formula of limited interest since it requires determining all the minimal paths, not an easy task. For example for the bridge structure



the minimal paths are

$$\{1, 4\}, \{2, 5\}, \{1, 3, 5\}, \{2, 3, 4\}$$

and thus we have

$$H(X) = 1 - (1 - X_1X_4)(1 - X_2X_5)(1 - X_1X_3X_5)(1 - X_2X_3X_4)$$

Note that is not straightforward anymore to compute $E[H(\mathbf{X})]$ since the minimal path are not independent. To compute this expectation it is necessary to multiply out all the factors and using the relation $X_i^2 = X_i$ which holds for binary variable one obtains after some algebra

$$\begin{aligned} P(\text{the system is functioning}) &= p_1p_4 + p_2p_5 + p_1p_3p_5 + p_2p_3p_4 - p_1p_2p_3p_4 - p_1p_2p_3p_5 \\ &\quad -p_1p_2p_4p_5 - p_1p_3p_4p_5 - p_2p_3p_4p_5 + 2p_1p_2p_3p_4p_5 \end{aligned}$$

Another way to obtain this formula is use the inclusion-exclusion formula (see exercises).

Example 1.6.5 (Network reliability). If the system has "many" components it is very tedious to compute the structure function $H(\mathbf{X})$ and the probability that the system is functioning and a Monte-Carlo method may be a decent alternative.

- **Step 1** Generate N independent random vector $\mathbf{X^{(i)}}$ where $X_k^{(i)} \sim \mathcal{B}(1, p_k)$ are independent and evaluate $H(\mathbf{X^{(i)}})$.
- **Step 2** $I_N = \frac{1}{N}\sum_{k=1}^{N} H(\mathbf{X^{(i)}})$ is an estimator for the probability that the system functions.

It is important to note that often it much easier to evaluate $H(\mathbf{X})$ (that is check if the system functions) than evaluate the structure function.

1.6.2 Confidence intervals for the Monte-Carlo method

We consider here two ways of building confidence intervals using either the central limit theorem or concentration inequalities.

The first main message to remember is that

$$\boxed{\text{For a tolerance } \epsilon \text{ one needs } N \sim \frac{1}{\epsilon^2} \text{ samples.}}$$

which gives the order of the Monte-Carlo method. This is not necessarily good news: to compute an integral deterministic algorithms may be much more efficient than a Monte-Carlo algorithm. On the other hand some piece of good news is that the scaling is "independent" of the dimension so there is some hope that Monte-Carlo algorithms could be useful in high dimension.

The second main message is that to compare different Monte-Carlo estimator for some given quantity μ a good piece of information is obtained by comparing the variances of different algorithms to determine which one is the most efficient.

$$\boxed{\text{If } I_N \to \mu \text{ and } \hat{I}_N \to \mu \text{ compare } \operatorname{var}(I_N) \text{ versus } \operatorname{var}(\hat{I}_N)}.$$

Asymptotic confidence intervals and CLT: First we use the Central Limit Theorems to build ***asymptotic confidence intervals*** for Monte-Carlo estimators. Recall that

$$\text{var}\left(\frac{\sqrt{N}}{\sigma}(I_N-\mu)\right) \;=\; \text{var}\left(\frac{1}{\sqrt{N}\sigma}\sum_{k=1}^{N}(X_i-\mu)\right) \;=\; \frac{1}{\sigma^2 N}\text{var}\left(\sum_{k=1}^{N}(X_i-\mu)\right) = 1\,.$$

and by the Central limit theorem the random $\frac{\sqrt{N}}{\sigma}(I_N-\mu)$ is asymptotically normal. To build a α-confidence interval we consider z_α the number defined by

$$\alpha = \frac{1}{\sqrt{2\pi}}\int_{-z_\alpha}^{z_\alpha} e^{-\frac{x^2}{2}}\,dx$$

For example $z_{.90} = 1.645$, $z_{.95} = 1.96$ and $z_{.99} = 2.576$. Then by the CLT

$$\lim_{N\to\infty} P\left(-z_\alpha \le \frac{\sqrt{N}}{\sigma}(I_N-\mu) \le z_\alpha\right) \;=\; \frac{1}{\sqrt{2\pi}}\int_a^b e^{-\frac{x^2}{2}}\,dx = \alpha$$

or

$$P\left(\mu \in \left[I_N - z_\alpha\frac{\sigma}{\sqrt{N}}\,,\, I_N + z_\alpha\frac{\sigma}{\sqrt{N}}\right]\right) \approx \alpha$$

and thus we obtain

Asymptotic α-confidence interval (assuming X has finite variance.)

$$\boxed{P\left(\mu \in [I_N-\epsilon, I_N+\epsilon]\right) \approx \alpha \quad \text{provided} \quad N \ge z_\alpha\frac{\sigma^2}{\epsilon^2}}$$

To be useful in practice, such a confidence interval requires us to know σ^2. But, in general there is reason we should know $\sigma^2 = \text{var}(X)$ if our goal of to compute $\mu = E[X]$. If doable we may use a bound on σ^2 if we have some information on the random variable. For example if we know that X is Bernoulli then $\text{var}(X) = p(1-p) \le \frac{1}{4}$. A more practical way to track the variance is to use the ***sample variance***

$$V_N^2 \;=\; \frac{1}{N-1}\sum_{i=1}^{N}\left(X_i - I_N\right)^2\,,$$

which is an unbiased estimator for σ^2, i.e., we have $E[V_N^2] = \sigma^2$ for every $N \ge 1$ and $\lim_{N\to\infty} V_N^2 = \sigma^2$ with probability 1.

There is a generalization of the CLT (based on Slutzky's Theorem) which states that we can replace the true variance σ^2 by the sample variance V_N^2 in the CLT:

$$\lim_{N\to\infty} P\left(-z_\alpha \le \frac{\sqrt{N}}{V_N}(I_N-\mu) \le z_\alpha\right) \;=\; \frac{1}{\sqrt{2\pi}}\int_a^b e^{-\frac{x^2}{2}}\,dx = \alpha$$

and thus we can build an asymptotic confidence interval using V_N instead of σ since we have

$$P\left(\mu \in \left[I_N - z_\alpha \frac{V_N}{\sqrt{N}}\,,\, I_N + z_\alpha \frac{V_N}{\sqrt{N}}\right]\right) \approx \alpha\,.$$

Nonasymptotic confidence intervals and concentration inequalities: Confidence interval based on the CLT are convenient to use in practice, and usually quite accurate but they are asymptotic in the sense that they use asymptotic normality and thus requires large N in the first place, and it is not clear a-priori how large N should be in the first place!

We explain another more rigorous approach based on Chernov bounds , Theorem 1.4.5, which provides bounds valid for all N and are thus called ***non-asymptotic confidence intervals***. Compared to the central limit theorem which requires only finite mean and variance establishing such bounds require knowledge of some more refined properties (tail behavior) of the random variable X. We will prove such a theorem (Hoeffding's theorem) here only for bounded random variables but many other results in that spirit exists and go under the general name of ***concentration inequalities***.

Theorem 1.6.6 (Hoeffding's theorem) *Suppose $X_1, X_2, \cdots X_N$ are I.I.D. random variables such that*

$$a \leq X_i \leq b\,,$$

and with mean μ. Then, for any $N \geq 1$ we have

$$P\left(|I_N - \mu| \geq \epsilon\right) \;\leq\; 2e^{-\frac{2N\epsilon^2}{(b-a)^2}} \tag{1.10}$$

This provides immediately a α confidence interval:

Non-asymptotic α-confidence interval (assuming $a \leq X \leq b$.)

$$\boxed{P\left(\mu \in [I_N - \epsilon, I_N + \epsilon]\right) \geq \alpha \quad \text{provided} \quad N \geq \log\left(\frac{2}{1-\alpha}\right)\frac{(b-a)^2}{2\epsilon^2}}$$

The key to prove Hoeffdings theorem is the following proposition.

Proposition 1.6.7 *Suppose Y satisfies $E[Y] = 0$ and $c \leq Y \leq d$. Then*

$$E[e^{tY}] \leq e^{\frac{t^2(d-c)^2}{8}}\,.$$

Proof: Since the function e^{ty} is convex for $y \in [c,d]$ write $y = \frac{d-y}{d-c}c + \frac{y-c}{d-c}d$ as a convex combination and thus

$$e^{ty} \leq \frac{d-y}{d-c}e^{tc} + \frac{y-c}{d-c}e^{td}\,,$$

and taking expectations and using $E[Y] = 0$ gives

$$E[e^{tY}] \;\leq\; \frac{d}{d-c}e^{tc} + \frac{-c}{d-c}e^{td}\,, \tag{1.11}$$

Set

$$p = \frac{d}{d-c}\,, \quad 1-p = \frac{-c}{d-c}\,, \quad u = t(d-c)\,.$$

Then we have

$$\begin{aligned} \log E[e^{tY}] &\leq \log(pe^{tc} + (1-p)e^{td}) \\ &= tc + \log(p + (1-p)e^{t(d-c)}) \\ &= (p-1)u + \log(p + (1-p)e^{u}) \\ &\equiv \varphi(u) \end{aligned} \tag{1.12}$$

To bound $\varphi(u)$ we use Taylor theorem with reminder

$$\varphi(u) = \varphi(0) + \varphi'(0)u + \frac{1}{2}\varphi''(\xi)u^2\,, \qquad \text{for some } \xi \in [0,u]$$

and we have $\varphi(0) = 0$ and

$$\varphi'(u) = (p-1) + \frac{(1-p)e^u}{p + (1-p)e^u} \text{ and so } \varphi'(0) = 0\,,$$

$$\varphi''(u) \;=\; \frac{p(1-p)e^u}{(p + (1-p)e^u)^2}\,.$$

Using the inequality $xy \leq (\frac{x+y}{2})^2$ we see that $\varphi''(u) \leq \frac{1}{4}$ and so

$$\log E[e^{tY}] = \varphi(u) = \frac{1}{2}\varphi''(\xi)u^2 \leq \frac{u^2}{8} = \frac{t^2(d-c)^2}{8}\,.$$

■

Proof of Theorem 1.6.6: This is an application of Chernov bound. Note first that $Y_i = X_i - \mu$ has mean 0 and $c \equiv a - \mu \leq Y \leq b - \mu \equiv d$ and so $(d-c) = (b-a)$. Note also

$$E[e^{tN(I_N - \mu)}] \;=\; E[e^{t\sum_{k=1}^{N}(X_k - \mu)}] \;=\; \prod_{k=1}^{N} E[e^{t(X_k-\mu)}] \;\leq\; e^{N\frac{t^2(b-a)^2}{8}}$$

We have then for any $t \geq 0$

$$P\left(I_N - \mu \geq \epsilon\right) = P\left(N(I_N - \mu) \geq N\epsilon\right) \leq \frac{E[e^{tN(I_N-\mu)}]}{e^{tN\epsilon}} \leq e^{N(\frac{1}{8}t^2(b-a)^2 - t\epsilon)}\,.$$

Optimizing over t we find the optimal $t^* = \frac{4\epsilon}{(b-a)^2}$ and thus

$$P\left(I_N - \mu \geq \epsilon\right) \leq e^{-\frac{2N\epsilon^2}{(b-a)^2}}$$

The estimate for $P\left(I_N - \mu \leq -\epsilon\right)$ is very similar and has the same bound and this proves the bound (1.10). ∎

Example 1.6.8 (Estimating the number π) Continuing Example 1.6.2 we sample N points in a cube of side length 2 and accept them if they land in the circle.

Suppose, for example, that we perform $N = 10'000$ trials and observe $S_N = 7932$, then our estimator for π is $4\frac{7932}{10000} = 3.1728$. If we use that $\sigma^2 \leq 1/4$ we have a tolerance

$$\epsilon = 1.96\frac{\sigma}{\sqrt{N}} \leq 1.96\frac{1}{2\sqrt{N}} = 0.0098\,.$$

and $[3.1336, 3.2120]$ for a 95% confidence interval.

If we use Hoeffding's bound with $a = 0, b = 1$ we obtain a tolerance

$$\epsilon = \sqrt{\log(40)}\frac{1}{\sqrt{2}\sqrt{N}} = 0.0135$$

and thus we have a slighly worse but not asymptotic confidence interval $[3.1184, 3.2271]$.

1.6.3 Variance reduction

We give two examples on how to reduce the variance of an algorithm. There are many other technique to perform ***variance reduction***, e.g. conditional MC, control variates, stratified sampling, etc.....

Antithetic Monte Carlo and common random numbers: These techniques use ***dependent random variables*** or dependent random numbers to decrease the variance. Note that for dependent random X and Y we have

$$\mathrm{var}(X + Y) = \mathrm{var}(X) + \mathrm{var}(Y) + 2\mathrm{cov}(X, Y)\,.$$

For the antithetic Monte-Carlo estimator one needs to have X and Y ***negatively correlated*** i.e. $\mathrm{cov}(X, Y) < 0$. The following theorem is key to use this techniques

Theorem 1.6.9 *Suppose $X_1, \cdots, X_n$ are independent random variables and f and g are increasing function of n variables. Then*

$$\mathrm{cov}(f(X_1, \cdots X_n), g(X_1, \cdots X_n)) \geq 0\,. \tag{1.13}$$

Proof: The proof is by induction on n. To prove it for $n = 1$ note that that if f and g are increasing we have for any x, y

$$(f(x) - f(y))(g(x) - g(y)) \geq 0$$

So for any random variable X, Y we have

$$(f(X) - f(Y))(g(X) - g(Y)) \geq 0\,,$$

implying that

$$E(f(X) - f(Y))(g(X) - g(Y))] \geq 0\,,$$

or

$$E[f(X)(g(X)] + E[f(Y)g(Y))] \geq E[f(X)g(Y)] + E[(f(Y)g(X))]\,.$$

If we suppose that X and Y are two independent copies of the same random variables then this implies

$$2E[f(X)g(X)] \geq 2E[f(X)]E[g(X)]$$

which proves (1.13) for the case for $n = 1$.

Assuming now that (1.13) is true for $n - 1$ we condition on the value of X_n and

$$\begin{aligned}
&E[f(X_1, \cdots X_n)g(X_1, \cdots X_n)|X_n = x_n] \\
&= E[f(X_1, \cdots, X_{n-1}, x_n)g(X_1, \cdots X_{n-1}, x_n)] \quad \text{by independence} \\
&\geq E[f(X_1, \cdots, X_{n-1}, x_n)]E[g(X_1, \cdots X_{n-1}, x_n)] \quad \text{by induction hypothesis} \\
&= E[f(X_1, \cdots, X_n)|X_n = x_n]E[g(X_1, \cdots, X_n)|X_n = x_n] \qquad (1.14)
\end{aligned}$$

Now taking expectation on both sides one finds

$$\begin{aligned}
E[f(X_1, \cdots X_n)g(X_1, \cdots X_n)] &= E[E[f(X_1, \cdots X_n)g(X_1, \cdots X_n)|X_n]] \\
&\geq E[E[f(X_1, \cdots X_n)|X_n]E[g(X_1, \cdots X_n)|X_n]]
\end{aligned}$$

Since $E[f(X_1, \cdots X_n)|X_n]$ and $E[g(X_1, \cdots X_n)|X_n]$ are increasing in X_n, by the result for $n = 1$ we have

$$\begin{aligned}
&E[E[f(X_1, \cdots X_n)|X_n]E[g(X_1, \cdots X_n)|X_n]] \\
&\geq E[E[f(X_1, \cdots X_n)|X_n]]E[E[g(X_1, \cdots X_n)|X_n]] \\
&= E[f(X_1, \cdots X_n)]E[g(X_1, \cdots X_n)] \qquad (1.15)
\end{aligned}$$

and this proves the theorem. ∎

From this we obtain the following useful corollary

Corollary 1.6.10 *If $U_1, \cdots, U_n$ are independent with $U_i \sim \mathcal{U}([0,1])$ and h is increasing or decreasing, then*

$$\operatorname{cov}\left(h(U_1, \cdots, U_n), h(1-U_1, \cdots, 1-U_n\right) \leq 0$$

Proof: If h is increasing then $-h(1-U_1, \cdots, 1-U_n)$ is increasing and so

$$\operatorname{cov}\left(h(U_1, \cdots, U_n), -h(1-U_1, \cdots, 1-U_n)\right) \geq 0$$

∎

Algorithm 1.6.11 (Antithetic Monte-Carlo sampling) *Assume h is increasing or decreasing. To estimate*

$$\mu = E\left[h(U)\right],$$

the antithetic sampling *consists in generating $\frac{N}{2}$ i.i.d. random variables $U_1, U_2, \cdots, U_{N/2}$ and set*

$$\tilde{I}_N = \frac{1}{N}\sum_{i=1}^{N/2}(h(U_i) + h(1-U_i)).$$

The quantity $\tilde{I}_n$ gives an unbiased estimator of μ with smaller variance

Comparing the variances we have $\operatorname{var}(I_N) = \frac{1}{N}\operatorname{var}(h(U))$ for the simple Monte-Carlo algorithm while for the antithetic variables we have

$$\begin{aligned} \operatorname{var}(\tilde{I}_N) &= \frac{1}{N}\frac{1}{2}\left[\operatorname{var}(h(U)) + \operatorname{var}(h(1-U)) + 2\operatorname{cov}(h(U), h(1-U))\right] \\ &= \frac{1}{N}\left[\operatorname{var}(h(U)) + \operatorname{cov}(h(U), h(1-U))\right] \end{aligned} \tag{1.16}$$

which is smaller than $\operatorname{var}(I_N)$ since $\operatorname{cov}(h(U), h(1-U)) < 0$.

Example 1.6.12 (Network reliability (continued)). To estimate the probability that a system functions recall that we have to compute

$$E[H(\mathbf{X})],$$

where $\mathbf{X} = (X_1, \cdots, X_n)$ is vector with independent components $X_i \sim \mathcal{B}(1, p_i)$ and H is the structure function, i.e. $H(\mathbf{X}) = 1$ is the system works and 0 otherwise.

Clearly $H(\mathbf{X})$ is increasing in each X_i since adding a functioning component to the system can only increase the structure function. Moreover we can write X_i as a function of $U_i \sim \mathcal{U}([0,1])$ as

$$X_i = \begin{cases} 1 & \text{if } U_i < p_i \\ 0 & \text{otherwise} \end{cases}$$

which is a decreasing function of U_i. Hence, we see that

$$H(X_1, \cdots, X_n) = k(U_1, \cdots, U_n)$$

is decreasing in $U_1, \cdots U_n$ and so we can use antithetic variables.

Importance sampling: We consider next another technique through which one can reduce the variance, sometimes by a considerable factor as we will demonstrate in an example below. We explain the method for continuous random variable but the same techniques works for discrete random variable as well. If X has density $f(x)$ to compute

$$E[h(X)] = \int h(x)f(x)dx$$

we can always use the simple MC algorithm $I_N = \frac{1}{N}\sum_{k=1}^{N} h(X_k)$. The idea of ***importance sampling*** is to sample from another random variable Y with density g and write

$$E[h(X)] = \int h(x)f(x)dx = \int h(x)\frac{f(x)}{g(x)}g(x)dx = E\left[\frac{h(Y)f(Y)}{g(Y)}\right].$$

For the integral to make sense we at least need to pick g such that

$$g(x) = 0 \text{ is allowed only if } \; f(x) = 0 \;\text{ or }\; h(x)f(x) = 0\,.$$

Often in practice if X belongs to some family, say $Exp(\lambda)$ then we may choose Y in the same family say $Exp(\beta)$ of $Gamma(\alpha, \beta)$ and so f anf g have the same support.

We have then

Algorithm 1.6.13 (Importance sampling Monte-Carlo estimator) *To estimate*

$$\mu = E\left[h(X)\right],$$

the importance sampling *use the random variable Y and the estimator*

$$\hat{I}_N = \frac{1}{N}\sum_{k=1}^{N}\frac{h(Y_k)f(Y_k)}{g(Y_k)}$$

The variance of the simple Monte-Carlo estimator I_N is

$$\text{var}(I_N) = \frac{1}{N}(E[h(X)^2] - E[h(X)]^2)$$

while the variance of the importance sampling estimator is given by

$$\begin{aligned}\mathrm{var}(\hat{I}_N) \;=\; \frac{1}{N}\mathrm{var}\left(\frac{h(Y)f(Y)}{g(Y)}\right) &= \frac{1}{N}\left[\int \frac{h(x)^2 f(x)^2}{g^2(x)} g(x)\,dx - \left(\int h(x)\frac{f(x)}{g(x)} g(x) dx\right)^2\right] \\ &= E\left[h(X)^2\frac{f(X)}{g(X)}\right] - E[h(X)]^2\end{aligned}$$

For the importance sampling method to be useful we need to have $\mathrm{var}(\hat{I}_N) < \mathrm{var}(I_N)$ and thus we need, roughly speaking,

$$\frac{f(x)}{g(x)} \text{ to be small where } h(x) \text{ is "big"},.$$

We will consider a couple examples but let us start with some theoretical (if practically not very useful) considerations.

Theorem 1.6.14 (Optimal importance sampling)

- *If $h(x) \geq 0$ then the optimal importance sampling distribution is*

$$g_*(x) \;=\; \frac{h(x)f(x)}{\mu} = \frac{h(x)f(x)}{E[h(X)]}\,,$$

  and the corresponding importance sampling estimator has 0 variance.

- *For general h the optimal importance sampling distribution is*

$$g_*(x) \;=\; \frac{|h(x)|f(x)}{E[|h|(X)]}\,.$$

Proof: If $h(x) \geq 0$ then with $g_*(x) \;=\; \frac{h(x)f(x)}{\mu}$ we have

$$\mathrm{var}(\hat{I}_N) = E\left[h(X)^2\frac{f(X)}{g_*(X)}\right] - E[h(X)]^2 = E\left[\frac{h(X)^2 f(X)\mu}{h(X)f(X)}\right] - E[h(X)]^2 \;=\; 0\,.$$

and thus it is optimal.

For general h the optimal variance does not vanish but we have (using Cauchy-Schwartz inequality (in the form $E[Z]^2 \leq E[Z^2]$) that for any choice of g

$$\begin{aligned}E\left[h(X)^2\frac{f(X)}{g_*(X)}\right] &= E\left[\frac{h(X)^2 f(X)}{|h|(X)f(X)}\right] E[|h|(X)] \\ &= E\left[|h|(X)\right]^2 \\ &= E\left[|h|(Y)\frac{f(Y)}{g(Y)}\right]^2 \\ &\leq E\left[h(Y)^2\frac{f(Y)^2}{g(Y)^2}\right] \\ &= E\left[h(X)^2\frac{f(X)}{g(X)}\right]\end{aligned}$$

which shows that g_* gives the optimal variance. ∎

This theorem is not very useful in practice since we cannot use the optimal importance sampling since it requires the knowledge of $E[h(x)]$ (or $E[|h|(X)]$) which is precisely what we are trying to compute in the first place! Let us consider first two benchmarking example where everything can be computed exactly.

Example 1.6.15 Suppose we want to sample the mean of $X \sim \mathcal{N}(0,1)$ using an importance sampling estimator, i.e. $h(x) = x$ and we use $Y \sim \mathcal{N}(0,\sigma)$ for importance sampling. Then

$$\begin{aligned} N\text{var}(\hat{I}_N)) &= \int_{-\infty}^{\infty} x^2 \frac{(e^{-x^2/2}/\sqrt{2\pi})^2}{e^{-x^2/2\sigma}/\sqrt{2\pi}\sigma}\, dx \\ &= \sigma \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} x^2 e^{-x^2(2-\sigma^{-2})/2}\, dx \\ &= \begin{cases} \frac{\sigma}{(2-\sigma^{-2})^{3/2}} & \text{if } \sigma^2 > 1/2 \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

An elementary computation shows that the optimal choice is $\sigma^2 = 5/2$.

Example 1.6.16 Suppose $X \sim Exp(\lambda)$ and $h = \mathbf{I}_{[a,\infty)}$ so that we are trying to estimate

$$\mu = E[h(X)] = P(X \geq a) = e^{-\lambda a}\,.$$

Note that $e^{-\lambda a}$ can be a very small number and since the variance is $e^{-\lambda a}(1 - e^{-\lambda a}) \approx e^{-\lambda a}$ we will need at n to be at the very least of order $e^{\lambda a}$ to have meaningful estimates.

The optimal importance sampling is

$$g_*(x) = \frac{h(x)f(x)}{e^{-\lambda a}} = \mathbf{I}_{[a,\infty)} \lambda e^{-\lambda(x-a)}$$

which is a shifted exponential random variable. Note that if sample from X we will mostly obtain $h(X) = 0$ and most samples are "useless" while if we sample from the importance sampling distribution every single samples contribute a non-zero term to the estimator.

We conclude with a (slightly) more realistic example from Network reliability.

Example 1.6.17 (Network reliability) Consider the connected graph as in Figure 1.1. We assume that each edge as a probability q of failing and all edges are independent. Think of q as a very small number, to fix the idea let $q = 10^{-2}$. Fix two vertices 1 and 11 in the graph and we want to compute the disconnection probability

$$\mu \equiv P\,(1 \text{ is not connected to } 11 \text{ by working edges})$$

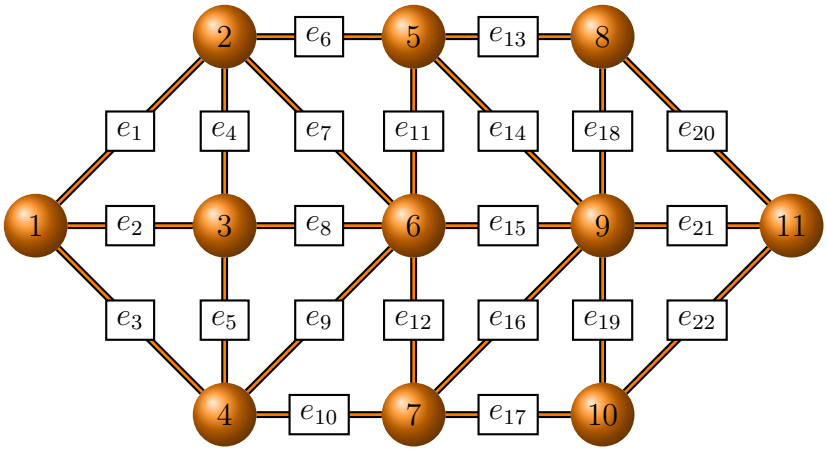


Figure 1.1: A graph with 11 vertices and 22 edges

This cannot be computed by hand for the graph in Figure 1.1 even though it is not that big.

The graph here has 22 edges numbered as in Figure 1.1 and let $\mathbf{X} = (X_1, \cdots X_{22})$ where $X_i = 1$ if the i^{th} edge fails and $X_i = 0$. We set $|\mathbf{x}| = \sum_{i=1}^{22} x_i$ and note that $|\mathbf{x}|$ is the number of failing edges in the graph. We have

$$p(\mathbf{x}) \equiv P(\mathbf{X} = \mathbf{x}) = \prod_{i=1}^{22} q^{x_i}(1-q)^{1-x_i} = q^{|\mathbf{x}|}(1-q)^{22-|\mathbf{x}|}$$

If we define the function k

$$k(\mathbf{x}) = \begin{cases} 1 & \text{if node 1 is not connected to node 11 through working edges} \\ 0 & \text{if node 1 is connected to node 11 through working edges} \end{cases}$$

Then we have

$$\mu = \sum_{\mathbf{x}\,;\,k(\mathbf{x})=\mathbf{1}} P(\mathbf{X} = \mathbf{x}) = \sum_{\mathbf{x}} k(\mathbf{x}) P(\mathbf{X} = \mathbf{x}) = E[k(\mathbf{X})]\,.$$

The simple MC sampling estimator for μ is

$$I_N = \frac{1}{N} \sum_{k=1}^{N} k(\mathbf{X}^{(k)})$$

where $\mathbf{X}^{(1)}, \cdots \mathbf{X}^{(N)}$ are i.i.d copies of $\mathbf{X}$. Each $\mathbf{X}^{(i)}$ can be generated by tossing an unfair coin 22 times. Then our estimator is simply the fraction of those simulated networks that fail to connect edges 1 and 11.

In order to get an idea of the number involved let us give a rough estimate of μ. It is easy to see that at least 3 nodes must fail for 1 not to be connected to 11. So we have

$$\mu \leq P(|\mathbf{X}| \geq 3) = 1 - \sum_{j=0}^{2} \binom{22}{j} q^j (1-q)^{22-j} \cong 0.00136\,,$$

since $|\mathbf{X}| \sim \mathcal{B}(22, q)$.

On the other hand we can get a lower bound for μ by noting that

$$\mu \geq P(e_1, e_2, e_3 \text{ fail}) = q^3 \;=\; 10^{-6}\,.$$

Therefore p_D is between 10^{-3} and 10^{-6} which is very small. We will thus need very tight confidence intervals. To compute $\mathrm{var}(I_n)$ note that $k(X)$ is a Bernoulli random variable with parameter p_D. Hence

$$\mathrm{var}(I_n) \;=\; \frac{1}{n}\mu(1-\mu) \cong \mu\,,$$

since p_D is small. To get a meaningful confidence interval we need its half length $2\sqrt{\mu/n}$ to be at the very least less than $\mu/2$. This implies however that we must choose $n > 16/\mu$, and thus we need millions of iterations for a network which is not particularly big.

Let us use importance sampling here. Note that $E[k(\mathbf{X})]$ is very small which means that typical X have $k(\mathbf{X}) = 0$. The basic idea is to choose the sampling variable in such a way that we sample more often these $\mathbf{X}$ for which $k(\mathbf{X}) = 1$ (i.e., large in our case).

A natural guess is take the random variable $\mathbf{Y}$ with

$$\phi(\mathbf{y}) \equiv P(\mathbf{Y} = \mathbf{y}) = \theta^{|\mathbf{y}|}(1-\theta)^{22-|\mathbf{y}|}$$

with a well chosen θ. Since $k(\mathbf{Y}) = 0$ whenever $|\mathbf{Y}| < 3$ we can for example choose θ such that $E[|\mathbf{Y}|] = 3$. Since $|\mathbf{Y}| = B(22, \theta)$ this gives $E[|\mathbf{Y}|] = 22\theta$ and thus $\theta = 3/22$.

The importance sampling estimator is now

$$\hat{I}_N = \frac{1}{N}\sum_{i=1}^{N} \frac{k(\mathbf{Y}^{(i)})p(\mathbf{Y}^{(i)})}{\phi(\mathbf{Y}^{(i)})}$$

where $\mathbf{Y}^{(i)}$ are i.i.d with distribution $\phi(\mathbf{y})$. Let us compute the variance of J_n. We have

$$\begin{aligned}
\mathrm{var}(\hat{I}_N) &= \frac{1}{N}\left(\sum_{\mathbf{y}} \frac{k(\mathbf{y})^2 p(\mathbf{y})^2}{\phi(\mathbf{y})^2}\phi(\mathbf{y}) - \mu^2\right) \\
&= \frac{1}{N}\left(\sum_{\mathbf{y}\,:\,k(\mathbf{y})=1} \frac{p(\mathbf{y})}{\phi(\mathbf{y})}p(\mathbf{y}) - \mu^2\right)\,. \qquad (1.17)
\end{aligned}$$

Note that

$$\frac{p(\mathbf{y})}{\phi(\mathbf{y})} = \frac{q^{|\mathbf{y}|}(1-q)^{22-[\mathbf{y}]}}{\theta^{|\mathbf{y}|}(1-\theta)^{22-|\mathbf{y}|}} = \left(\frac{1-q}{1-\theta}\right)^{22}\left(\frac{q(1-\theta)}{\theta(1-q)}\right)^{|\mathbf{y}|} = 20.2 \times (0.064)^{|\mathbf{y}|}\,.$$

In Eq. (1.17) all terms with $k(\mathbf{y}) = 1$ have $|\mathbf{y}| \geq 3$. For those $\mathbf{y}$ we have

$$\frac{p(\mathbf{y})}{\phi(\mathbf{y})} \leq 20.2 \times (0.064)^3 \leq 0.0053$$

So we get

$$\mathrm{var}(\hat{I}_N) \leq \frac{1}{N} \sum_{\mathbf{y}\,:\,k(\mathbf{y})=1} 0.0053\; p(\mathbf{y}) = \frac{0.0053\,\mu}{N}$$

This means that we have reduced the variance by a factor approximately of 200. So for the same n the confidence interval is going to be about $\sqrt{200} \cong 14$ times smaller. Alternatively a given confidence interval for I_N can be obtained for $\hat{I}_{N/200}$. This is not too bad.

Chapter 2

Markov Chains with Finite State Space

2.1 Introduction

A ***discrete-time stochastic processes*** is a sequence of random variables

$$\{X_n\}_{n=1}^{\infty} = \{X_0, X_1, X_2, \cdots, \}$$

where each of the random variables X_n takes value in the ***state space*** S (the same S for all n). Throughout this chapter we assume that

$$\boxed{S \text{ is finite.}}$$

so that X_n are discrete random variables. More general state S will be considered later.

Usually we will think of n as "time" and X_n describe the state of some system at time n. The simplest example of a stochastic process is to take the X_n as a sequence of i.i.d random variables. In that case one simply sample a random variable again and again. In general however the X_n are not independent and to describe a stochastic process we need to specify all the ***joint probability density functions***

$$P\{X_0 = i_0, X_1 = i_1, \cdots, X_n = i_n\}$$

for all $n = 0, 1, \cdots$ and for all $i_0 \in S, \cdots i_n \in S$. Instead of the joint p.d.f we can specify instead the ***conditional probability density functions***

$$\begin{aligned}
&P\{X_0 = i_0\} \\
&P\{X_1 = i_1 | X_0 = i_0\} \\
&P\{X_2 = i_2 | X_1 = i_1, X_0 = i_0\} \\
&\qquad \vdots \\
&P\{X_n = i_n | X_{n-1} = i_{n-1}, \cdots, X_1 = i_1, X_0 = i_0\}
\end{aligned} \tag{2.1}$$

and by conditioning we have the relation

$$P\{X_0 = i_0, X_1 = i_1, \cdots, X_n = i_n\} = \\ P\{X_0 = i_0\} P\{X_1 = i_1 | X_0 = i_0\} \cdots P\{X_n = i_n | X_{n-1} = i_{n-1}, \cdots, X_0 = i_0\} .$$

To obtain a Markov chain ones makes a special assumption on the conditional p.d.f.: Imagine you are at time $n-1$ (your "present"), think of time n as your "future" and $1, 2, \cdots, n-2$ as your "past". For a ***Markov chain*** one assumes that the future state depends only on the present state but not on the past states. Formally we have

Definition 2.1.1 *A stochastic process $\{X_n\}$ with a discrete state space S is called a* ***Markov chain*** *if*

$$P\{X_n = i_n | X_{n-1} = i_{n-1}, \cdots, X_1 = i_1, X_0 = i_0\} = P\{X_n = i_n | X_{n-1} = i_{n-1}\}$$

for all n and for all $i_0 \in S, \cdots i_n \in S$.

The conditional probabilities $P\{X_n = i_n | X_{n-1} = i_{n-1}\}$ are called the ***transition probabilities*** of the Markov chain $\{X_n\}$. In order to specify a Markov chain we need to specify in addition the ***initial distribution*** $P\{X_0 = i_0\}$ and we have then

$$P\{X_0 = i_0, X_1 = i_1, \cdots, X_n = i_n\} = \\ P\{X_0 = i_0\} P\{X_1 = i_1 | X_0 = i_0\} \cdots P\{X_n = i_n | X_{n-1} = i_{n-1}\}$$

The transition probabilities $P\{X_n = j | X_{n-1} = i\}$ are the probability that the chain moves from i to j at time n. In general these probabilities might depend on n. If they are independent of n then we call the Markov chain ***time homogeneous***. Unless explicitly stated we will always assume in the sequel that the Markov chain is time homogeneous. Such a Markov chain is specified by

$$\mu(i) \equiv P\{X_0 = i\}, \quad \text{initial distribution}$$

and

$$P(i,j) \equiv P\{X_n = j | X_{n-1} = i\} \quad \text{transition probabilities}$$

All quantities of interest for the Markov chain can be computed using these two objects. For example we have

$$P\{X_0 = i_0, X_1 = i_1, X_2 = i_2\} = \mu(i_0) P(i_0, i_1) P(i_1, i_2) .$$

or

$$P\{X_2 = i\} = \sum_{i_0 \in S} \sum_{i_1 \in S} \mu(i_0) P(i_0, i_1) P(i_1, i)$$

and so on.

If S is a finite set with N elements, without loss of generality, we can relabel the state so that $S = \{1, 2, \cdots, N\}$. It will be convenient to set

$$\mu = (\mu(1), \cdots, \mu(N))$$

that is μ is a row vector whose entries are the initial distribution. The vector μ is called a ***probability vector***, i.e., μ is a vector such that $\mu(i) \geq 0$ and $\sum_i \mu(i) = 1$.

Also we will write P for the $N \times N$ matrix whose entries are $P(i, j)$,

$$P = \begin{pmatrix} P(1,1) & P(1,2) & \cdots & P(1,n) \\ P(2,1) & P(2,2) & \cdots & P(2,n) \\ \vdots & \vdots & & \vdots \\ P(n,1) & P(n,2) & \cdots & P(n,n) \end{pmatrix}$$

The matrix P is called a ***stochastic matrix***, i.e., P is a matrix with nonnegative entries $P(i, j) \geq 0$ and the sum of every row is equal to 1, $\sum_{j=1}^{N} P(i, j) = 1$ for all i.

Lemma 2.1.2 *(a) The n-step transition probabilities are given by*

$$P\{X_n = j | X_0 = i\} = P^n(i, j)$$

where P^n is the matrix product $\underbrace{P \cdots P}_{\text{n times}}$.

(b) If $\mu(i) = P\{X_0 = i\}$ then

$$P\{X_n = i\} = \mu P^n(i)\,.$$

(c) If $f = (f(1), \cdots, f(n))^T$ is a column vector then we have

$$P^n f(i) = E[f(X_n) | X_0 = i]\,.$$

Proof: (a) By induction it is true for $n = 1$ and so let assume the formula is true for $n - 1$. We condition on the state at time $n - 1$, use the formula

$$P(AB|C) = P(A|BC)P(B|C)$$

for conditional probabilities, the Markov property, and the induction hypothesis. We obtain

$$\begin{aligned} P\{X_n = j | X_0 = i\} &= \sum_{k \in S} P\{X_n = j\,, X_{n-1} = k | X_0 = i\} \\ &= \sum_{k \in S} P\{X_n = j | X_{n-1} = k\,, X_0 = i\}\, P\{X_{n-1} = k | X_0 = i\} \\ &= \sum_{k \in S} P\{X_n = j | X_{n-1} = k\}\, P\{X_{n-1} = k | X_0 = i\} \\ &= \sum_{k \in S} P^{n-1}(i, k) P(k, j) = P^n(i, j)\,. \end{aligned} \tag{2.2}$$

(b) Note that if μ is a probability vector and P is a stochastic matrix then μP is a probability vector since

$$\sum_i \mu P(i) \;=\; \sum_i \sum_j \mu(j) P(j,i) \;=\; \sum_j \mu(j) \sum_i P(j,i) \;=\; \sum_j \mu(j)\,.$$

Furthermore by the formula for conditional probabilities and (a)

$$P\{X_n = j\} \;=\; \sum_{k\in S} P\{X_n = j | X_0 = k\} P\{X_0 = k\} \;=\; \sum_k \mu(k) P^n(k,j) \;=\; \mu P^n(j)\,.$$

(c) We have

$$P^n f(i) \;=\; \sum_k P^n(i,k) f(k) \;=\; \sum_k f(k) P\left\{X_n = k | X_0 = i\right\} \;=\; E[f(X_n)\,|\,X_0 = i]\,.$$

∎

A basic question in Markov chain is to understand the distribution of $\{X_n\}$ for large n, for example we want to know whether the limit

$$\lim_{n\to\infty} P\{X_n = i\} \;=\; \lim_{n\to\infty} \mu P^n(i)$$

exists or not, whether it depends on the choice of initial distribution π and how to compute it.

Definition 2.1.3 *A probability vector π is called a **limiting distribution** if the limit*

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

exists.

It could well be that the limit depend on the choice of the initial distribution μ.

Definition 2.1.4 *A probability vector π is called a **stationary distribution** if the limit*

$$\pi P = \pi$$

exists.

Limiting distributions are always stationary distributions:

Lemma 2.1.5 *If π is a limiting distribution then π is a stationary distribution.*

Proof: Suppose $\lim_{n\to\infty} \mu P^n = \pi$. Then

$$\pi P \;=\; (\lim_{n\to\infty} \mu P^n) P \;=\; \lim_{n\to\infty} \mu P^{n+1} \;=\; \lim_{n\to\infty} \mu P^n \;=\; \pi\,.$$

and thus π is stationary. ∎

Later in this chapter we will derive conditions under which stationary distributions are unique and are limiting distributions.

2.2 Examples

We give here a fairly long list of classical and useful Markov chains that we will meet again and again in the sequel.

Example 2.2.1 (2-state Markov chain) Let us consider a Markov chain with two states, i.e. $S = \{1, 2\}$. The transition matrix has the general form

$$P = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix}. \tag{2.3}$$

The equation for the stationary distribution $\pi P = \pi$ is

$$\begin{aligned} \pi(1)(1-p) + \pi(2)q &= \pi(1) \\ \pi(1)q + \pi(2)(1-p) &= \pi(1) \end{aligned}$$

or $p\pi(1) = q\pi(2)$. Normalizing to a probability vector gives $\pi = \left(\frac{q}{p+q}, \frac{p}{p+q}\right)$.

We show that π is also a limiting distribution. Let us set $\mu_n \equiv \mu P^n$ and let us look at the difference between μ_n and π. We have using $\mu_n(2) = 1 - \mu_n(1)$

$$\begin{aligned} \mu_n(1) - \pi(1) &= \mu_{n-1}P(1) - \pi(1) = \mu_{n-1}(1)(1-p) + (1-\mu_{n-1}(1))q - \frac{q}{p+q} \\ &= \mu_{n-1}(1)(1-p-q) - \frac{q}{p+q}(1-p-q) = (1-p-q)(\mu_{n-1}(1) - \pi(1)) \end{aligned}$$

By induction we have $\mu_n(1) - \pi(1) = (1-p-q)^n(\mu_0(1) - \pi(1))$. If either $p > 0$ or $q > 0$ then $-1 < 1-p-q < 1$ and so $\lim_{n\to\infty} \mu_n(1) = \pi(1)$. Clearly we have also $\lim_{n\to\infty} \mu_n(2) = \pi(2)$.

If either p or q does not vanish then $\mu_n = \mu P^n$ converges to a stationary distribution. ∎

The next example is a very simple example of Markov chain.

Example 2.2.2 (i.i.d random variables) Let X_n, $n = 0, 1, 2, \cdots$ be a sequence of i.i.d random variables with common distribution $\pi(i) = P\{X_n = i\}$ for all n. The X_n satisfy the Markov property since all the X_n are independent

$$P\{X_n = i_n | X_{n-1} = i_{n-1} \cdots X_0 = i_0\} = P\{X_n = i_n\} = P\{X_n = i_n | X_{n-1} = i_{n-1}\}$$

The stationary and limiting distribution are π and the transition matrix is

$$P = \begin{pmatrix} \pi(1) & \pi(2) & \cdots & \pi(N) \\ \pi(1) & \pi(2) & \cdots & \pi(N) \\ \vdots & \vdots & & \vdots \\ \pi(1) & \pi(2) & \cdots & \pi(N) \end{pmatrix}$$

■

Example 2.2.3 (Random walks on $\{0, 1, \cdots, N\}$) In the random walk Markov chain if $X_n = j$ and $j \neq 0, N$ then the next step consist in jumping to the right to $j+1$ with probability p and jumping to the right with probability $1-p$, i.e.,

$$P(j, j+1) = p\,, P(j, j-1) = 1-p\,, \quad j = 1, \cdots, N-1$$

If j is at "the boundary", i.e., either 0 or N there are several variants of the random walks

(a) (**Absorbing boundary conditions:**) Upon hitting the boundary the random walk stays there, i.e.

$$P(0,0) = 1\,, \quad P(N,N) = 1$$

The states 0 or N are called absorbing states: if the Markov chain reaches 0 at some time n then $X_{n+k} = 0$ for all $k \geq 0$. For Markov chain with absorbing states the questions of interests which we will be study later are

• How long does it take to reach an absorbing state?

• What it the probability to reach one absorbing state (say 0) before reaching another one (say N). (In the context of the random walk this is called the **gambler's ruin problem**).

(b) (**Reflecting boundary conditions**) Upon hitting the boundary the random walks bounces back, i.e.,

$$P(0,1) = 1\,, \quad P(N,N-1) = 1$$

(c) (**Partially reflecting boundary conditions**) The following intermediate case has nice properties, in particular an easy formula for the invariant measure.

$$P(0,0) = (1-p)\,, P(0,1) = p\,, \quad P(N,N-1) = (1-p)\,, P(N,N) = p\,.$$

(d) (**Periodic boundary conditions**) In the periodic case we imagine that 0 and N are "neighbors" or we identify 0 with $N+1$. We have

$$P(0,1) = p\,, P(0,N) = (1-p)\,, \quad P(N,0) = p\,, P(N,N-1) = (1-p)\,.$$

■

Example 2.2.4 (Finite queueing models) Imagine the following phone system. An operator answers calls and if the operator is busy answering a call incoming calls can be put on hold but a maximum of N caller can be in the system at any time. If exactly N people are in the system then a caller will be bounced back. We assume that during each time interval exactly one new caller calls with probability p and 0 caller calls with probability $1-p$. Also during each time interval exactly one call is completed with probability q and 0 call is completed with probability $1-q$.

If X_n is the number of people in the system then the state space of the system is $\{0, 1, 2, \cdots, N\}$ and the transition probabilities are

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

and for $1 \leq j \leq N-1$

$$P(j, j-1) = q(1-p)\,, \;\; P(j,j) = pq + (1-p)(1-q)\,, \;\; P(j, j+1) \;=\; p(1-q)\,,$$

and

$$P(N, N-1) = q\,, \quad P(N,N) = 1-q\,.$$

■

Example 2.2.5 (Coupon collecting problem)A company offers toys in breakfast cereal boxes. There are N different toys available and each toy is equally likelky to be found in any cereal box. Let X_n be the number of distinct toys that you collect after buying n boxes and is natural to set $X_0 = 0$. Then X_n is a Markov chain, it has a simple structure since X_n either stays the same of increase by 1. The transition probabilities are

$$P(j, j+1) \;=\; P\{ \text{ new toy } \mid \text{ already j toys}\} \;=\; \frac{N-j}{N}\,.$$

and

$$P(j,j) \;=\; P\{ \text{ no new toy } \mid \text{ already j toys}\} \;=\; \frac{j}{N}\,.$$

Clearly after a random finite time τ, the Markov chain X_N reaches the absorbing state N. To compute $E[\tau]$ let us write

$$\tau = T_1 + \cdots + T_N\,,$$

where T_i is the time needed to get a new toy after you have gotten your $(i-1)^{th}$ toy. The T_i's are independent and have T_i has a geometric distribution with $p_i = (N-i)/N$. Thus

$$E[\tau] \;=\; \sum_{i=1}^{N} E[T_i] \;=\; \sum_{i=1}^{N} \frac{N}{N-i} \;=\; N \sum_{i=1}^{N} \frac{1}{i} \approx N \ln(N)\,.$$

■

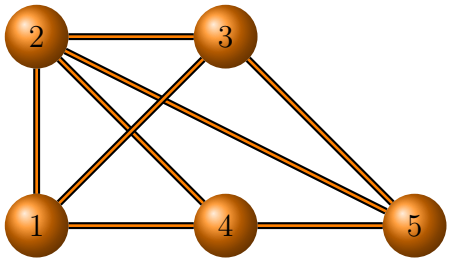


Figure 2.1: An example of a graph with vertex set {1,2,3,4,5}

Example 2.2.6 (Random walk on graphs) An ***undirected graph*** G consists of a ***vertex set*** V and a ***edge set*** E where the elements of E are (unordered) pairs of vertices. Think of the graph G as a collection of dots (the vertices) and lines joining two dots v and w if and only if the pair $\{v, w\}$ is an edge. We say that the vertex v is a ***neighbor*** of the vertex w, and write $v \sim w$, if $\{v, w\}$ is an edge. The ***degree*** of a vertex v, denoted $\deg(v)$, is the number of neighbor of v.

Given a graph $G = (V, E)$ the simple random walk on G is the Markov chain with state space V and transition matrix

$$P(v, w) = \begin{cases} \frac{1}{\deg(v)} & \text{if } w \sim v \\ 0 & \text{otherwise} \end{cases} .$$

For example if the graph is the one given in figure 2.1 then the transition matrix is

$$P = \begin{pmatrix} 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 \\ \frac{1}{4} & 0 & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{3} & \frac{1}{3} & 0 & 0 & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & 0 & 0 & \frac{1}{3} \\ 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 \end{pmatrix}$$

The invariant distribution for the random walk on graph is given by

$$\pi(v) = \frac{\deg(v)}{2|E|}$$

where $|E|$ is the cardinality of the set E, i.e., the number of edges. First note that $\sum_v \pi(v) = 1$ since each edge connects two vertices. To show that it is invariant note that

$$\pi P(v) = \sum_{w; w \sim v} \frac{\deg(w)}{2|E|} \frac{1}{\deg(w)} = \frac{1}{2|E|} \sum_{w; w \sim v} 1 = \pi(v)$$

■

Example 2.2.7 (Random walk on the N-dimensional hypercube) The ***K-dimensional hypercube*** is a graph whose vertices are the binary K-tuples $\{0,1\}^K$. Two vertices are connected by an edge when they differ in exactly one coordinate. The simple random walk on the hypercube moves from one vertex $x = (x_1, \cdots, x_K)$ by choosing a coordinate $j \in \{1, 2, \cdots, K\}$ uniformly at random and setting the new state equal to $x' = (x_1, \cdots, 1 - x_j, \cdots, x_K)$. That is the j^{th} bit is flipped.

The degree of each vertex is k, the number of vertices is 2^K and the number of edges is $2^k k/2$ so we have for any x

$$\pi(x) \;=\; \frac{1}{2^k}$$

which is, unsurprisingly, the uniform distribution on $S = V$.

■

Example 2.2.8 (Ehrenfest urn model) Suppose K balls are distributed among two urns called urn A and urn B. At each move one ball is selected uniformly at random among the K balls and is transferred from its current urn to the other urn. If X_n is the number of balls in urn A then the state space is $S' = \{0, 1, \cdots, K\}$ and the transition probabilities

$$P(j, j+1) \;=\; \frac{K-j}{K}\,, \quad P(j, j-1) \;=\; \frac{j}{K}\,.$$

We will show that the invariant distribution is

$$\pi(j) \;=\; \binom{K}{j}\frac{1}{2^K}\,.$$

Indeed we have

$$\begin{aligned}
\pi P(j) &= \sum_k \pi(k)P(k,j) \\
&= \pi(j-1)P(j-1,j) + \pi(j+1)P(j+1,j) \\
&= \frac{1}{2^K}\left[\binom{K}{j-1}\frac{K-(j-1)}{K} + \binom{K}{j+1}\frac{j+1}{K}\right] \\
&= \binom{K}{j}\frac{1}{2^K}\,.
\end{aligned}$$

This Markov chain is closely related to the simple random walk on the hypercube. Let S be the state space of the random walk Y_n and S' the state space of the urn model Markov chain X_n. Let us define the map $F : S \mapsto S'$ given by

$$F(x) = j \quad \text{iff } j = \#\{l, x_l = 0\}\,.$$

that is we just count the number of 0 in x. The transition of the random walks corresponds then exactly to the transition for the urn model and we have for any x with $F(x) = j$

$$P(j, j+1) \;=\; \sum_{y;F(y)=j+1} P(x,y) \;=\; \binom{K}{j}\frac{1}{2^K}$$

We obtain the urn model by lumping together the states of the random walk on the hypercube. This does not always lead to a Markov chain, but if does the Markov chain is called ***lumpable***.

■

2.3 Existence and uniqueness of stationary distribution

We first show that stationary distributions always exist for finite state Markov chains. This will not be the case if the state space is countable.

Theorem 2.3.1 *Let X_n be a Markov chain on a finite state space S. Then X_n has at least one stationary distribution.*

Proof: We prove this using the Boltzano Weierstrass theorem which asserts that if the sequence $\{x_n\}_{n=1}^{\infty}$ is a bounded sequence in $\mathbf{R}^n$ (i.e., there exists M such that $\|x_n\| \le M$ for all n) then we can find a convergent subsequence $\{x_{n_k}\}_{k=1}^{\infty}$ which converges to x.

Let us consider an initial distribution μ and the distribution μP^n of X_n for $n = 1, 2, \cdots$. We could try to apply Bolzano-Weierstrass on the sequence $\{\mu P^n\}$ but this will not work. Instead we consider another sequence

$$\nu_n \;=\; \frac{1}{n}\left(\mu + \mu P + \cdots \mu P^{n-1}\right)$$

i.e., we average the distribution of X_n over the first n steps. Note that ν_n is a probability vector, in particular $0 \le \nu_n(j) \le 1$ for all $j \in S$ and thus the sequence $\{\nu_n\}_{n=1}^{\infty}$ is bounded.

Note further that

$$\nu_n P - \nu_n \;=\; \frac{1}{n}\left(\mu P + \cdots \mu P^n - \mu - \cdots - \mu P^{n-1}\right) \;=\; \frac{1}{n}(\mu P^n - \mu)\,.$$

and thus for any $j \in S$

$$|\nu_n P(j) - \nu_n(j)| \;\le\; \frac{1}{n} \tag{2.4}$$

Using Boltzano-Weierstrass Theorem we pick an increasing sequence n_k with $\lim_{k\to\infty} n_k = \infty$ such that the sequnce $\{\nu_{n_1}, \nu_{n_2}, \cdots\}$ converges, i.e.,

$$\lim_{k\to\infty} \nu_{n_k}(j) \,=\, \pi(j)\,.$$

and π is a probability vector.

Finally we have

$$|\pi P(j) - \pi(j)| \,=\, \lim_{k\to\infty} |\nu_{n_k} P(j) - \nu_{n_k}(j)| \,\leq\, \lim_{k\to\infty} \frac{1}{n_k} \,=\, 0\,.$$

and thus $\pi P = \pi$ and π is invariant. ∎

In general we can have several stationary distribution for a Markov chains, in particular if the state space can be partitioned in at least two different classes which do not communicate (see later for more details). We say that j is ***accessible*** from i and write $i \to j$ if there exists $n \geq 0$ such that $P^n(i,j) > 0$. We say that i and j ***communicate*** if $i \to j$ and $j \to i$ in which case we write $i \leftrightarrow j$.

Definition 2.3.2 A Markov chain X_n is called ***irreducible*** if every state $i \in S$ communicate with every other state $j \in S$ that is for any pair of states i, j in S there exists n such that $P^n(i,j) > 0$.

Lemma 2.3.3 *Let X_n be an irreducible Markov chain and let π be a stationary distribution. Then $\pi(i) > 0$ for any $i \in S$.*

Proof: If π is stationary distribution then $\pi(i) > 0$ for at least one i. Now take $j \in S$ such that $i \to j$ then there exists a time r such that $P^r(i,j) > 0$ and we have

$$\pi(j) \,=\, \sum_k \pi(k) P^r(k,j) \,\geq\, \pi(i) P^r(i,j) > 0\,.$$

and thus $\pi(j) > 0$. Since X_n is irreducible $i \to j$ for all $j \in S$ and so $\pi(j) > 0$ for any $j \in S$. ∎

We can now prove that the stationary distribution is unique. Note that π is a left eigenvector of P corresponding to the eigenvalue 1, ($\pi P = \pi$) or equivalently a right eigenvector for the transpose matrix $P^T(i,j) = P(j,i)$. (i.e. $P^T \pi^T = \pi^T$). To prove uniqueness we are going to study right eigenvectors of P instead.

Proposition 2.3.4 *Suppose X_n is irreducible and h is a column vector such that $Ph = h$ then $h = c(1, 1, \cdots, 1)$ is a constant vector.*

Proof: Suppose $Ph = h$, then there exists i_0 such that $h(i_0) = \max_{i\in S} h(i) \equiv M$. If h is not a constant vector there exists j with $i_0 \to j$ but $h(j) < M$. Since $P^n h = h$,

$$M = h(i_0) \;=\; P^n h(i_0) \;=\; P^n(i_0, j)\underbrace{h(j)}_{<M} + \sum_{l\neq i_0} P^n(i_0, l)\underbrace{h(l)}_{\leq M} < M \sum_l P^n(i_0, l) = M\,,$$

and this is a contradiction. ∎

Corollary 2.3.5 *Suppose X_n is irreducible then there exists a unique stationary distribution π.*

Proof: The previous proposition show that the kernel of $P - I$ has dimension one. From linear algebra we know that the dimension of the kernel of a matrix A is the same as the dimension of the kernel of the transpose matrix A^T. So the kernel of $P^T - I$ has dimension 1, this space contains exactly one vector whose entries sum to 1, namely π. ∎

Example 2.3.6 (Random walks, cont'd) The random walks of Example 2.2.3 are irreducible for reflecting, partially reflecting, and periodic boundary conditions. For absorbing boundary conditions no states are accesible from 0 or N. In that case the Markov chain is not irreducible and we can find two stationary distribution, $\pi = (1, 0, \cdots, 0)$ and $\pi' = (0, \cdots, 0, 1)$. ∎

For an irreducible Markov chain X_n, we have a unique stationary distribution π and it is natural to ask whether μP^n converges to π. This is however in general not true. To see what can go wrong let us consider the random walk on $\{0, \cdots, N\}$ with periodic boundary conditions and let us assume that N is odd so that the state space has an even number of elements. The stationary distribution is the uniform distribution

$$\pi \;=\; \left(\frac{1}{N+1}, \cdots, \frac{1}{N+1}\right)\,.$$

On the other hand let us suppose that the initial distribution is $X_0 = 0$, then for odd n, X_n will be on an odd site $j \in S$ and will be on an even site for even times n. In that case the distribution of X_n at time n alternates between even and odd states and thus certainly cannot converges to π.

This motivates the following definition;

Definition 2.3.7 A Markov chain X_n is called ***irreducible and aperiodic*** if there exists an integer n such that $P^n(i, j) > 0$ for all pair i, j in S.

Theorem 2.3.8 *Let X_n be an irreducible and aperiodic Markov chain with stationary distribution π. There exists a constant $C > 0$ and number α with $0 \leq \alpha < 1$ such that for any initial distribution μ we have*

$$|\mu P^n(j) - \pi(j)| \leq C\alpha^n \,, \tag{2.5}$$

i.e., the distribution of X_n converges, exponentially fast, to π.

Proof: Since the Markov chain is irreducible and aperiodic we can find an integer r such that P^r has strictly positive entries. Let Π be the stochastic matrix

$$\Pi = \begin{pmatrix} \pi(1) & \pi(2) & \cdots & \pi(N) \\ \pi(1) & \pi(2) & \cdots & \pi(N) \\ \vdots & \vdots & & \vdots \\ \pi(1) & \pi(2) & \cdots & \pi(N) \end{pmatrix}$$

where every row is the stationary distribution π. Note that this corresponds to independent sampling from the stationary distribution.

Since all elements of $P^r(i,j)$ are strictly positive we can pick $\delta > 0$ sufficiently small such that

$$P^r(i,j) \geq \delta\Pi(i,j) = \delta\pi(j) \,.$$

for all $i, j \in S$. Let us set $\theta = 1 - \delta$ and define define a stochastic matrix Q through the equation

$$P^r = (1-\theta)\Pi + \theta Q \,.$$

(check that Q is stochastic). We have the following two elementary facts

• Since $\pi P = \pi$ we have

$$\Pi P^n = \Pi$$

for any $n \geq 1$.

• For any stochastic matrix M we have

$$M\Pi = \Pi \,,$$

because all rows of Π are identical.

Using this we show, by induction, that any integer $k \geq 1$,

$$P^{kr} = (1-\theta^k)\Pi + \theta^k Q^k \,.$$

This is true for $k = 1$ and so let us assume it is true for k. We have then using $\Pi P^r = \Pi$ and $Q\Pi = \Pi$.

$$\begin{aligned} P^{r(k+1)} = P^{rk}P^r &= \left[(1-\theta^k)\Pi + \theta^k Q^k\right] P^r \\ &= (1-\theta^k)\Pi P^r + \theta^k Q^k[(1-\theta)\Pi + \theta Q] \\ &= (1-\theta^k)\Pi + \theta^k(1-\theta)\Pi + \theta^{k+1}Q^{k+1} \\ &= (1-\theta^{k+1})\Pi + \theta^{k+1}Q^{k+1} \,, \end{aligned}$$

and this concludes the induction step. From this we see that $P^{rk} \to \Pi$ as $k \to \infty$. An arbitrary integer n can be written as $n = kr + l$ where $0 \le l < r$. We have then

$$P^n \;=\; P^{kr}P^l \;=\; \Pi + \theta^k \left[Q^k P^l - \Pi\right]$$

and thus

$$|P^n(i,j) - \pi(j)| \;=\; \theta^k \left|Q^k P^l(i,j) - \Pi(i,j)\right| \;\le\; \theta^k \le \frac{1}{\theta}(\theta^{1/r})^n \,.$$

Finally if μ is an arbitrary initial distribution we obtain the desired bound by multiplying $P^n(i,j) - \pi(j)$ by $\mu(i)$ and summing over i. So we obtain (2.5) with $C = \theta^{-1}$ and $\alpha = \theta^{1/r}$ ∎

We have just showed that the stationary distribution is also a limiting distribution. A very important chararcterization of $\pi(j)$ is in terms of ***occupation times***. To do this we need a lemma from analysis

Lemma 2.3.9 *Suppose $\{a_n\}$ is a sequence of number converging to a. Let*

$$b_n = \frac{1}{n}(a_0 + \cdots a_{n-1}) \;=\; \frac{1}{n}\sum_{k=0}^{n-1} a_k$$

then $\lim_{n\to\infty} b_n = a$.

Proof: (see exercise). ∎

Note that the converse statement is not always true. If b_n converges then a_n needs not converge. (Take e.g. $\{a_n\} = \{0, 1, 0, 1, 0, \cdots\}$).

From Theorem 2.3.8 we have $P^n(i,j) \to \pi(j)$ and thus, by Lemma 2.3.9,

$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} P^k(i,j) \;=\; \pi(j)\,.$$

In order to interpret this quantity let us introduce the random variable

$$Y_n^{(j)} \equiv \sum_{k=1}^{n} \mathbf{1}_{\{X_k = j\}}$$

where $\mathbf{1}_A$ is the indicator function of the event A. The random variables $Y_n^{(j)}$ counts the number of visits to the state j up to time n. Note that if $X_0 = i$ then

$$E[\mathbf{1}_{\{X_k=j\}} | X_0 = i] = P^k(i,j)\,,$$

and thus we conclude that

$$\pi(j) \;=\; \lim_{n\to\infty} \frac{1}{n} E\left[\sum_{k=1}^{n} \mathbf{1}_{\{X_k=j\}}\right],$$

that is $\pi(j)$ represents the expected fraction of time that the Markov chain spends in i, in the long run.

We also need another random variable which is the ***first return time to state*** j. It is defined as

$$\tau^{(j)} \;=\; \min\{n>0, X_n = j\}\,.$$

i.e., $\tau^{(j)}$ is the first time the Markov chain returns to j.

We can also consider k^{th} return to state j. By the Markov property, once the Markov chain reaches j it forgets about the past, and therefore the k^{th} return to j will occur at the time

$$T_k^{(j)} \;=\; \tau_1^{(j)} + \cdots + \tau_k^{(j)}\,,$$

where $\tau_l^{(j)}$ are independent copies of the return times $\tau^{(j)}$. For $l \geq 2$ $\tau_l^{(j)}$ is conditioned on starting at j while for $l = 1$ it depends on the initial condition. Note that by the strong LLN for IID random variables we have

$$\lim_{k\to\infty} \frac{T_k}{k} \;=\; \lim_{k\to\infty} \frac{1}{k}\left(\tau_1^{(j)} + \cdots + \tau_k^{(j)}\right) \;=\; E[\tau^{(j)}|X_0 = j]\,.$$

Using this we obtain

Theorem 2.3.10 (Ergodic Theorem for Markov chain).

Let X_n be an irreducible aperiodic Markov chain with arbitrary initial condition μ, then, with probability 1 we have

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \mathbf{1}_{\{X_n=j\}} \;=\; \pi(j)\,,$$

this is the ***ergodic theorem*** *for Markov chain.*

Moreover if $\tau^{(j)}$ is the first return time to j we have the ***Kac's formula***

$$\pi(j) \;=\; \frac{1}{E[\tau^{(j)}|X_0 = j]}\,.$$

Proof: Given n consider a sample of the Markov chain $X_0, X_1, X_2, \cdots, ...X_n$. Let us denote $Y_n = Y_n^{(j)}$ the number of times that the Markov chain visits j up to time n. By definition if $Y_n = k$ then we have

$$T_k^{(j)} \leq n < T_{k+1}^{(j)}$$

So we obtain

$$\frac{T^{(j)}_{Y_n}}{Y_n} < \frac{n}{Y_n} \leq \frac{T^{(j)}_{Y_n+1}}{Y_n+1}\frac{Y_n+1}{Y_n}$$

Now taking $n \to \infty$ both extremes of the inequality converge to $E[\tau^{(j)}|X_0 = j]$ with probability 1 and thus we conclude that with probability 1

$$\lim_{n\to\infty} \frac{Y_n}{n} \;=\; \frac{1}{E[\tau^{(j)}|X_0 = j]}\,.$$

On the other hand we know that

$$\lim_{n\to\infty} \frac{E[Y_n]}{n} \;=\; \pi(j)\,,$$

and thus

$$\pi(j) \;=\; \frac{1}{E[\tau^{(j)}|X_0 = 1]}\,.$$

Putting all pieces together gives the theorem. ∎

Note that this theorem is of great practical importance: it is, in principle, enough to generate one sufficiently long sample of the Markov chain to produce the stationary distribution π. Generate a "path" $X_0, X_1, \cdots, x_n$ of the Markov chain then we have an estimator for $\pi(j)$

$$\boxed{\pi(j) \approx \frac{\#\text{ of visits to } j \text{ between 0 and } n}{n+1}}$$

The paths of the Markov chain will be generated as follows:

Algorithm 2.3.11 (Algorithm to generate a Markov chain

1. Set $X_0 = i$. Generate a random number U.
2. - If $U < P(i,1)$ set $X_1 = 1$
   - If $P(i,1) \leq U < P(i,1) + P(i,2)$ set $X_1 = 2$.
   - If $P(i,1) + P(i,2) \leq U < P(i,1) + P(i,2) + P(i,3)$ set $X_1 = 3$.
   - $\cdots$
3. Return to step 1.

We generalize slightly the ergodic theorem by considering a function $f : S \to \mathbf{R}$ or equivalently a a column vector $f = (f(1), \cdots, f(N))^T$. You may think of $f(j)$ as being the reward for being in state j.

Corollary 2.3.12 *Let X_n is an irreducible Markov chain with stationary distribution π. Then for any initial distribution μ we have, with probability* 1*,*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=0}^{n-1} f(X_k) \;=\; \sum_j \pi(j) f(j)\,.$$

Proof: We write

$$f(X_k) \;=\; \sum_{j\in S} f(j)\mathbf{1}_{\{X_k=j\}}\,.$$

and thus

$$\frac{1}{n}\sum_{k=0}^{n-1} f(X_k) \;=\; \sum_{j\in S} f(j)\left[\frac{1}{n}\sum_{k=0}^{n-1}\mathbf{1}_{\{X_k=j\}}\right] \to \sum_{j\in S} f(j)\pi(j)\,.$$

∎

2.4 Periodicity

In this section we discuss, briefly, the behavior of periodic and irreducible Markov chains. For a state j, let us consider the set

$$\mathcal{T}(j) \;=\; \{n \geq 1\,,\; P^n(j,j) > 0\}$$

of times when the chain can return to the starting position j. The ***period*** of the state j is the greatest common divisor of $\mathcal{T}(j)$.

We have

Lemma 2.4.1 *Suppose $i \leftrightarrow j$, then the period of i and the period of j coincide.*

Proof: Since $i \leftrightarrow j$ there exists integers r and l such that $P^r(i,j) > 0$ and $P^l(j,i) > 0$. Set $m = r + l$. Then we have

$$P^m(i,i) \geq P^r(i,j)P^l(j,i) > 0$$

and

$$P^m(j,j) \geq P^l(j,i)P^r(i,j) > 0$$

and thus $m \in \mathcal{T}(i) \cap \mathcal{T}(j)$. Furthermore assume that $t \in \mathcal{T}(i)$ then

$$P^{t+m}(j,j) \geq P^l(j,i)P^t(i,i)P^r(i,j) > 0\,,$$

and thus $t + m \in \mathcal{T}(j)$. If d_j is the gcd of $\mathcal{T}(j)$ then, by the above we have

$$m = kd_j\,, \quad m + t = \tilde{k}d_j$$

and thus $t = (\tilde{k} - k)d_j$ that $t \in \mathcal{T}(j)$. This implies that $\mathcal{T}(i) \subset \mathcal{T}(j)$ and thus gcd $\mathcal{T}(j) \leq$ gcd $\mathcal{T}(i)$. By reversing the roles of i and j we have gcd $\mathcal{T}(j) =$ gcd $\mathcal{T}(j)$. ∎

We say that a Markov chain is ***aperiodic*** if the period of every state is 1.

Example 2.4.2 (Random walks on graph, cont'd) The random walks on a graph, see Example 2.2.6, is irreducible if and only if the graph is connected. The random walk is aperiodic if and only if the graph is not ***bipartite*** (a graph is bipartite if there exists a partition $V = V_1 \cup V_2$ of the set off all vertices that $v \sim w$ if and only if $v \in V_1$ and $w \in V_2$. ∎

To connect our notation with the one of the previous section we prove

Proposition 2.4.3 *If X_n is irreducible and aperiodic then there exists n_0 such that $P^n(i,j) > 0$ for all $n \geq n_0$ and all $i, j \in S$.*

Proof: The proof relies on a number-theoretic fact (whose proof is omitted): suppose $\mathcal{A}$ is a subset of the integers which is closed under addition and whose gcd is 1, then $\mathcal{A}$ contain all but finitely many integers.

For $j \in S$, if $m, n \in \mathcal{T}(j)$ then $m + n \in m + n \in \mathcal{T}(j)$ since we have $P^{n+m}(j,j) \geq P^n(j,j)P^m(j,j) > 0$. This shows that $\mathcal{T}(j)$ is closed under addition and thus there exists $n(j)$ such that $P^n(j,j) > 0$ for all $n \geq n(j)$. Since $i \to j$ there exists $k = k(j,i)$ such that $P^{n+k}(j,i) \geq P^n(j,j)P^k(j,i) > 0$ if $n > n(j)$. Since S is finite we can find a n_0 such that $P^n(i,j) > 0$ for all $n \geq n_0$ and all i and j. ∎

Let us now assume that X_n is irreducible and has period $d > 1$. Let us pick two states i and j. By irreducibility there exists m and r with $P^m(i,j) > 0$ and $P^l(j,i) > 0$ and so a return to i is possible in $n = m + l$ steps. So d divides $n + l$. Therefore if j can be reached from i in m_1 steps and in m_2 steps then $m_2 - m_1$ must be divisible by d so we can write $m_1 = k_1 d + r$ and $m_2 = k_2 d + r$ for some $0 \leq r < d - 1$. So j can be reached from i only in r, $d + r$, $2d + r, \cdots$ steps. This implies that we can decompose the state space

$$S = G_1 \cup \cdots \cup G_d$$

and the only transitions that can occur are from G_l to G_{l+1} (and we set that $1 \equiv d+1$).

Note also that, in d steps the Markov chain moves from G_l back to G_l and since X_n is irreducible the Markov chain with state space G_l and transition matrix P^d is irreducible and aperiodic.

Relabelling the state space we can assume that the transition matrix in the block form

$$P = \begin{pmatrix} 0 & P_{G_1G_2} & 0 & 0 & \cdots & 0 \\ 0 & 0 & P_{G_2G_3} & 0 & \cdots & 0 \\ \vdots & \vdots & & & \vdots & \vdots \\ 0 & 0 & \cdots & & 0 & P_{G_{d-1}G_d} \\ P_{G_dG_1} & 0 & \cdots & & 0 & 0 \end{pmatrix}$$

and we have

$$P^d = \begin{pmatrix} P^d_{G_1} & 0 & 0 & \cdots & 0 \\ 0 & P^d_{G_2} & 0 & \cdots & 0 \\ \vdots & \vdots & & & \vdots \\ 0 & 0 & \cdots & 0 & P^d_{G_d} \end{pmatrix}$$

We can now use our results on aperiodic irreducibles chains to deduce the behavior of period chains. Let us denote by π_{G_l} the stationary distribution for the Markov chain with transition matrix $P^d_{G_l}$.

If $i \in G_l$ and $j \in G_l$ then we have

$$\lim_{n\to\infty} P^{nd}(i,j) = \pi_{G_l}(j)\,,$$

and thus for $i \in G_l$ and $j \in G_{l+1}$

$$\lim_{n\to\infty} P^{nd+1}(i,j) = \lim_{n\to\infty} \sum_{k\in G_{l+1}} P(ik)P^{nd}(k,j) = \sum_{k\in G_{l+1}} P(ik)\pi_{G_{l+1}}(j) = \pi_{G_{l+1}}(j)\,,$$

and so $i \in G_l$ and $j \in G_{(l+r)\mathrm{mod}(d)}$ we have

$$\lim_{n\to\infty} P^{nd+r}(i,j) = \pi_{G_{l+r}}(j)\,.$$

So for a given $i \in S$ and $j \in G_l$ the sequence $P^n(i,j)$ is asymptotically periodic where a sequence of $d-1$ successive 0 alternates with a number eventually very close to $\pi_{G_l}(k)$.

Let us define now

$$\pi \equiv \frac{1}{d}(\pi_{G_1}, \cdots, \pi_{G_d})\,.$$

The distribution π is normalized, stationary (you should check this) and furthermore we have

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} P^k(i,j) = \pi(k)$$

since the time spend in state k is asymptotically equal to $\frac{1}{d}\pi_{G_l}(k)$.

At this point we can also repeat, word for word, the argument of the Theorem 2.3.10 of previous section and obtain

Theorem 2.4.4 *Assume that X_n is irreducible of period d. Then with probability 1 we have*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \mathbf{1}_{\{X_n=j\}} = \pi(j)\,.$$

Moreover if $\tau^{(j)}$ the first return time to j

$$\pi(j) = \frac{1}{E[\tau^{(j)}|X_0 = j]}\,.$$

In particular for any initial distribution μ we have

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \mu P^k(j) \;=\; \pi(j)\,.$$

2.5 Decomposition of state space and transient behavior

In this section we drop the assumption of irreducibility and develop a number of tool to study the transient behavior of Markov chains.

We note that first that the communication relation $i \leftrightarrow j$ is an ***equivalence relation***. We use the convention $P^0 = I$ (the identity matrix) and then also that $i \leftrightarrow i$ (i communicates with itself $P^0(i,i) = 1$). We have

- It is *reflexive* ($i \leftrightarrow i$).
- It is *symmetric* ($i \leftrightarrow j$ implies $j \leftrightarrow i$).
- It is *transitive* ($i \leftrightarrow j$ and $j \leftrightarrow l$ implies $i \leftrightarrow l$).

Using this equivalence relation we can decompose the state space S into mutually disjoint ***communication classes***,

$$S = C_1 \cup C_2 \cup C_M\,.$$

It will be useful to distinguish two kinds of communication classes: ***transient*** classes and ***closed*** classes.

Definition 2.5.1 *1. A class C is called **transient** if there exists $i \in C$ and $j \in S \setminus C$ with $i \to j$ (that is j is accessible from i.)*

*2. A class C is called **closed** if it is not transient that is, for any pair $i \in C$ and $j \in S \setminus C$ we have $i \not\to j$.*

Note that for if the Markov chain X_n starts a transient class and we have $i \to j$ for $i \in C$ and $j \in S \setminus C$ then $j \not\to i$ because if it were it would imply that $j \in C$ as well. Therefore if the Markov chain exits a transient class C, it will never return to it. Actually the next lemma says that you always exits a transient class.

Lemma 2.5.2 *Let C be a communication class.*

1. If C is closed and $X_0 \in C$ then $X_n \in C$ for all $n = 1, 2, 3, \cdots$.

2. *If C is transient and $X_0 \in C$ then with probability 1, X_n exits C after a finite time and thus for any $i, j \in C$ we have*

$$\lim_{n\to\infty} P^n(i,j) = 0\,.$$

Proof: Part 1. is immediate. To prove part 2 note i is a state in a transient class C, then after allowing for some time for the Markov chain to access a state which can actually exits C, the Markov chain can exit C (and never return). Since C is a finite set we can thus find a time k and $\theta < 1$ such that for any $i \in C$ we have

$$P\{X_k \in C | X_0 = i\} \le \theta\,, \qquad \text{for all } i \in \ .$$

Repeating the argument implies that $P\{X_{nk} \in C | X_0 = i\} \le \theta^n \to 0$ and so the probability to stay in transient class goes to 0 as time goes by. As a consequence if i and j both belong to a transient class C we must have $\lim_{n\to\infty} P^n(i,j) = 0$. This proves the lemma. ∎

If the Markov chain has L recurrent classes let us label them $R_1, \cdots R_L$ and K transient classes let us label them $T_1, \cdots T_K$ and set also $T = T_1 \cup \cdots T_K$. After reordering the states we can put the transition matrix in the form

$$P = \begin{pmatrix} P_1 & & & & \\ & P_2 & & 0 & \\ & & P_3 & & 0 \\ & 0 & & \ddots & \\ & & & & P_L \\ & & S & & Q \end{pmatrix} \tag{2.6}$$

where P_l gives the transition probabilities within the class R_l, Q the transition iwthin the transient classes and S the transition from the transient classes into the recurrent classes.

It is easy to see that P^n has the form, for some matrix S_n

$$P^n = \begin{pmatrix} P_1^n & & & & \\ & P_2^n & & 0 & \\ & & P_3^n & & 0 \\ & 0 & & \ddots & \\ & & & & P_L^n \\ & & S_n & & Q^n \end{pmatrix}$$

Example 2.5.3 (Random walk with absorbing boundary conditions, cont'd) The Markov chain has three classes, 2 closed ones $\{0\}$, $\{N\}$ and 1 transient one

$\{1, \cdots, N-1\}$ (of period 2) We can write P as with $N = 5$ and the states ordered as $0, 5, 1, 2, 3, 4$

$$P = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 1/2 & 0 & 0 & 1/2 & 0 & 0 \\ 0 & 0 & 1/2 & 0 & 1/2 & 0 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 1/2 & 0 & 0 & 1/2 & 0 \end{pmatrix} \tag{2.7}$$

■

In order to understand the behavior of a reducible Markov chains we make the observations and questions

- If $X_0 = i \in R_l$ belongs to some closed class R_l then the behavior of X_n is entirely determined by the transition matrix of P_l restricted to the class R_l. If we denote by π_{R_l} the stationary measure $\pi_{R_l} P_l = \pi_{R_l}$ then we have for $i, j \in R_l$

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} P^k(i,j) \;=\; \pi_{C_l}(j) \quad \text{or} \quad \lim_{n\to\infty} P^n(i,j) \;=\; \pi_{C_l}(j)$$

  depending on whether X_n is periodic or not.

- If $X_0 = i \in T$ belongs to some transient class then X_n will exit T after some finite time. Let us define an ***absorption time***

$$T_{abs} = \min\{n \geq 0\,;\, X_n \;\text{ does not belong to } T\}$$

  and we may ask

$$\text{Can we compute the expected tim until absorption } E[T_{abs} \,|\, X_0 = i]\,?$$

- If $X_0 = i \in T$ belongs to some transient class and if there are more than one closed class, say the closed class $R_1, R_2, \cdots, R_L$ then the Markov may be absorbed in distinct closed class and se we may ask

$$\text{Compute } P\{X_n \;\text{ reaches class } R_l \,|\, X_0 = i\} \text{ for each class } R_l,\; l = 1, \cdots, L$$

The matrix Q in (2.6) is called a substochastic matrix, i.e., a matrix with nonnegative entries whose row sums are less than or equal to 1. We have seen in Lemma 2.5.2 that $Q^n(i,j) \to 0$ for all i, j and thus all eiegenvalues of Q have absolute values strictly less than 1. Therefore $I - Q$ is an invertible matrix and we can define

$$M = (I - Q)^{-1}$$

We give next a probabilistic interpretation of the matrix M. Let i be a transient state and consider the random variables $Y^{(i)}$ the total number of visits to i, i.e.,

$$Y^{(i)} = \sum_{n=0}^{\infty} \mathbf{I}_{\{X_n=i\}} .$$

Since i is transient $Y^{(i)} < \infty$ with probability 1. Suppose j is another transient state and $X_0 = j$. Then we have

$$\begin{aligned} E[Y^{(i)} \,|\, X_0 = j] &= E\left[\sum_{n=0}^{\infty} \mathbf{I}_{\{X_n=i\}} \,|\, X_0 = j\right] \\ &= \sum_{n=0}^{\infty} P\left\{X_n = i \,|\, X_0 = j\right\} \\ &= \sum_{n=0}^{\infty} P^n(i,j) . \end{aligned}$$

That is

$$\begin{aligned} E[Y^{(i)} \,|\, X_0 = j] &= I(i,j) + P(i,j) + P^2(i,j) + \cdots \\ &= I(i,j) + Q(i,j) + Q^2(i,j) + \cdots \end{aligned}$$

But we have

$$(I + Q + Q^2 + \cdots Q^n)(I - Q) = I - Q^{n+1} ,$$

and thus since $Q^n \to 0$ as $n \to \infty$

$$M = (I - P)^{-1} = \sum_{n=0}^{\infty} Q^n .$$

There $M(j,i)$ is simply the expected number of visits to i if $X_0 = j$.

We summarize this discussion in

Proposition 2.5.4 *Let j be a transient state and let T_{abs} to be the time until the Markov chain reaches some closed class. Then we have*

$$E[T_{abs}|X_0 = j] = \sum_i M(j,i) .$$

where

$$M = (I - Q)^{-1} = 1 + Q + Q^2 + \cdots$$

Example 2.5.5 (Random walk with absorbing boundary conditions, cont'd) From (2.7) we have

$$Q = \begin{pmatrix} 0 & 1/2 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 \\ 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 1/2 & 0 \end{pmatrix}, \quad M = (I-Q)^{-1} = \begin{pmatrix} 1.6 & 1.2 & 0.8 & 0.4 \\ 1.2 & 2.4 & 1.6 & 0.8 \\ 0.8 & 1.6 & 2.4 & 1.2 \\ 0.4 & .8 & 1.2 & 1.6 \end{pmatrix} \tag{2.8}$$

and thus the expected time until absorption are 4 for states 1 and 4 and 6 for states 2 and 3.

■

This technique can also be used if we want to compute the expected number of steps that an irreducible Markov chain needs to reach one state j from a state i, i.e., $E[\tau^{(i)}|X_0 = j]$. First we reorder the states to write the transition matrix in the block form

$$P = \begin{pmatrix} P(i,i) & R \\ S & Q \end{pmatrix} \tag{2.9}$$

Since the first visit to i starting from j does not depend on the matrix element $P(i,k)$ we can modify the transition matrix P such as to make j an absorbing state without changing the distribution of $\tau^{(i)}$. That is we set

$$\hat{P} = \begin{pmatrix} 1 & 0 \\ S & Q \end{pmatrix}.$$

For the Markov chain with transition matrix $\hat{P}$, all states except i now form a transient class and so we can apply Proposition 2.5.4 and obtain

Proposition 2.5.6 *Let X_n be an irreducible Markov chain. For inotj we*

$$E[\tau^{(i)}|X_0 = j] = \sum_i M(j,i)\,.$$

where $M = (I-Q)^{-1}$ and Q is given in (2.9) and obtained by deleting the i^{th} row and i^{th} column from P.

Example 2.5.7 (Random walk with reflecting boundary conditions, cont'd) Suppose we have reflecting boundary conditions $N = 5$, and we want to compute

$$E[\tau^{(1)}|X_0 = i]\,.$$

The transition matrix is

$$P = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 1/2 & 0 & 0 \\ 0 & 0 & 1/2 & 0 & 1/2 & 0 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix} \tag{2.10}$$

To compute $E[\tau^{(1)}|X_0 = 1] = \pi(1)^{-1}$ we need the stationary distribution which is $\pi = \left(\frac{1}{10}, \frac{2}{10}, \frac{2}{10}, \frac{2}{10}, \frac{2}{10}, \frac{1}{10}\right)$ To compute the other return times we have

$$Q = \begin{pmatrix} 0 & 1/2 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 \\ 0 & 1/2 & 0 & 1/2 & 0 \\ 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}, \quad M = (I-Q)^{-1} = \begin{pmatrix} 2 & 2 & 2 & 2 & 1 \\ 2 & 4 & 4 & 4 & 2 \\ 2 & 4 & 6 & 6 & 3 \\ 2 & 4 & 6 & 8 & 4 \\ 2 & 4 & 6 & 8 & 5 \end{pmatrix} \tag{2.11}$$

and so the expected return times to 1 are $10, 9, 16, 21, 24, 25$ respectively ∎

Let us suppose now that there exists at least two different closed classes and we ask the question: starting in a transient state j what is the probability that the Markov chain ends up in a particular closed class. To answer this question we can assume, without loss of generality, that every closed class is an absorbing state $r_1, \cdots r_L$ (we can always collapse a closed class into a absorbing state) and and we denote tha transient states by $t_1, \cdots, t_M$. By reordering the states we have

$$P = \begin{pmatrix} I & 0 \\ S & Q \end{pmatrix}$$

Let $A(t_i, r_j)$ be the probability that the chain starting at t_i eventually ends up in state r_j and we also set $A(r_i, r_i) = 1$ and $A(r_i, r_j) = 0$ if $i \neq j$. We condition on the first step of the Markov chain

$$\begin{aligned} A(t_i, r_j) &= P\{X_n = r_j \text{ eventually } | X_0 = t_i\} \\ &= \sum_{l \in S} P\{X_1 = l | X_0 = t_i\} P\{X_n = r_j \text{ eventually } | X_1 = l\} \\ &= \sum_{l \in S} P(t_i, l) A(l, r_j) = P(t_i, r_j) + \sum_{t_k} P(t_i, t_k) A(t_k, r_j). \end{aligned} \tag{2.12}$$

Let A be the $L \times M$ matrix with entries $A(t_i, r_j)$, then (2.12) can be written in matrix form as

$$A = S + QA$$

or

$$A = (I - Q)^{-1}S = MS\,.$$

We summarize the discussion in

Proposition 2.5.8 *For a Markov chain with transition probabilities*

$$P = \begin{matrix} s_1 \\ \vdots \\ s_L \\ t_1 \\ \vdots \\ t_M \end{matrix} \begin{pmatrix} & \\ I & 0 \\ & \\ & \\ S & Q \\ & \end{pmatrix}$$

where the states $t_1, \cdots t_M$ are transient we have

$$P\{X_n \text{ reaches class } s_l \,|\, X_0 = t_m\} = A(s_l, t_m) \text{ with } A = (I - Q)^{-1}S$$

Example 2.5.9 (Random walk with absorbing boundary conditions, cont'd) From (2.7) and (2.8) we have

$$A = MS = \begin{pmatrix} 1.6 & 1.2 & 0.8 & 0.4 \\ 1.2 & 2.4 & 1.6 & 0.8 \\ 0.8 & 1.6 & 2.4 & 1.2 \\ 0.4 & .8 & 1.2 & 1.6 \end{pmatrix} \begin{pmatrix} 1/2 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 1/2 \end{pmatrix} = \begin{pmatrix} .8 & .2 \\ .6 & .4 \\ .4 & .6 \\ .2 & .8 \end{pmatrix}$$

For example from state 2 the probability to be absorbed in 0 is .6, and so on.... ∎

2.5.1 Gambler's ruin

To compute absorption probabilities and absorption time, we need to inverse the matrix $(I - Q)$ which is doable only for small matrices. We use a different method to study here the classical ***Gambler's ruin problem*** which is nothing but a random walk on $\{0, 1, \cdots, N\}$ with absorbing boundary conditions and transition probabilities:

$$p(j, j+1) = p\,,\ p(j, j-1) = q\,,\ p(j, j) = r\,, \quad \text{with } p + q + r = 1$$

and

$$p(0, 0) = p(N, N) = 1\,.$$

The gambling interpretation consists of a series of bets: if betting \$1 one wins \$1 with probability p, loses \$1 with probability q, and loses \$0 with probability r. For example

in the American Roulette if you bet \$1 on red you have $p = 18/38 = 0.47336...$ and $q = 20/38 = .5263...$ while in the French Roulette you have $p = 18/37 = 0.4864$, $q = 18/37 + 19/37^2 = 0.5003...$ and $r = 18/37^2 = 0.0131....$

We are interested in computing $\alpha(j) \equiv A(j, N)$ the ***absorption probabilities*** that, starting with fortune j, you fortune reaches N ("your goal") before reaching 0 ("the ruin"). Clearly we have $\alpha(0) = 0$ and $\alpha(N) = 1$, and condition on the first step (i.e. the first bet) we obtain

$$\alpha(j) \;=\; q\alpha(j-1) + r\alpha(j) + p\alpha(j+1)\,. \tag{2.13}$$

which is a ***second order linear difference equation***.

One can view second order linear difference equations as discretization of second order differential equations and thus for linear equations we can try and use the same guessing techniques. The basic idea is to try solutions of the form of an exponential

$$\alpha(j) = x^j$$

and to use the linearity principle: if $\alpha_1(j)$ and $\alpha_2(j)$ are two solutions a second order linear difference equation then so is $c_1\alpha_1(j) + c_2\alpha_2(j)$ for any constants c_1, c_2.

For the gambler's ruin in (2.13) we find

$$x^j = qx^{j-1} + rx^j + px^{j+1}$$

which simplifies to

$$px^2 - (p+q)x + q \;=\; 0 \implies x = 1\,,\; x = \frac{q}{p}$$

If $p \neq q$ (i.e. if the game is not fair) we have two distinct roots and the general solution is

$$\alpha(j) \;=\; c_1 + c_2 \left(\frac{q}{p}\right)^j\,.$$

Using the boundary conditions we find

$$\boxed{\alpha(j) \;=\; \frac{1 - \left(\frac{q}{p}\right)^j}{1 - \left(\frac{q}{p}\right)^N}} \quad \text{Gambler's ruin for } q \neq p$$

For $p = q$ we have only one solution $x = 1$, and inspired by differential equations we try solutions of the form $jx^j = j$ which is easily seen to be indeed a solution. For $p = q$ the general solution is

$$\alpha(j) = c_1 + c_2 j$$

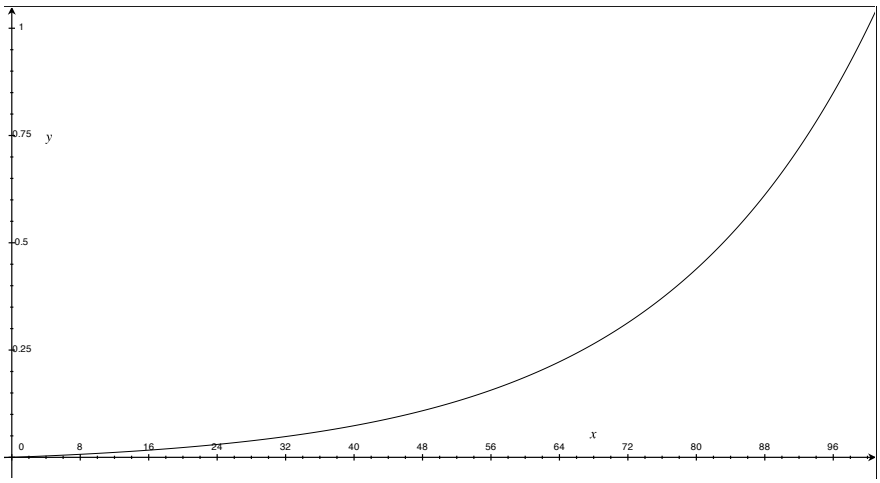


Figure 2.2: Gambler's ruin probabilities for $n = 100$, $p = 0.49$, $q = 0.51$, $r = 0$

and using the boundary conditions we find

$$\boxed{\alpha(j) = \frac{j}{N}} \quad \text{Gambler's ruin for } q = p$$

How bad does it get? To get an idea on how the gambler's ruin formula looks like let us take a *subfair game* with $q = .51$ and $p = 0.49$ and $N = 100$. Starting with a fortune of j we wish to reach a fortune of 100 and we have

$$\alpha(10) = 0.0091, \quad \alpha(50) = 0.1191 \quad \alpha(83) = 0.4973.$$

That is starting with \$10 the probability to win \$100 before losing all you money is less than one in a hundred and to reach a fifty-fifty chance to win you need to start with \$83! For a *superfair game* with $q = .49$ and $p = 0.51$ and $N = 100$ we have

$$\alpha(10) = 0.3358, \quad \alpha(50) = 0.8808, \quad \alpha(75) = 0.9679.$$

Bold or cautious? Using the formula for the gambler's ruin we can investigate whether there is a better gambling strategy than betting \$1 repeatedly. For example if we start with \$10 and our goal is to reach \$100 we could choose between

- Play \$1 bets?
- Play \$10 bets?

Using the gamblers ruin formula for $N = 100$ and $N = 10$ respectively we find

$$P\{\text{ Win \$100 in \$1 bets starting with \$10}\} = \frac{1 - (51/49)^{10}}{1 - (51/49)^{100}} = 0.0091$$

$$P\{\text{ Win \$100 in \$10 bets starting with \$10 }\} = \frac{1-(51/49)}{1-(51/49)^{10}} = 0.0829$$

that is, your probability to win is about nine time better in \$10 increments than in \$1 increments.

If on the contrary the odds are in your favor, even so slightly, say q=.49, and p=.51 then the opposite is true. We find

$$P\{\text{ Win \$100 in \$1 bets starting with \$10}\} = \frac{1-(49/51)^{10}}{1-(49/51)^{100}} = 0.3358$$

$$P\{\text{ Win \$100 in \$10 bets starting with \$10 is }\} = \frac{1-(49/51)}{1-(49/51)^{10}} = 0.1189.$$

In summary we have

If the odds are in your favor be cautious but if the odds are against you be bold!

Two limiting cases: To get a better handle on the formula let us look at 2 limiting cases and we slightly rephrase the problem:

- We start at 0.
- We stop whenever we reach W (W stands for our desired gain) and when we reach $-L$ (L stands for how much money we are willing to lose).

Now $j \in \{-L, -L+1, \cdots, \cdots, W-1, W\}$. We have simply changed variables and so obtain

$$P(-L, W) \equiv P(\text{Reach } W \text{ before reaching } -L \text{ starting from 0}) = \frac{1-(q/p)^L}{1-(q/p)^{L+W}}.$$

We consider the two limiting cases where W and L go to ∞.

• $L \to \infty$ means that the player is willing to lose an infinite amount of money, i.e. he has infinite resources and he is trying to reach a gain of W units. We find

$$P\{\text{ A gambler with unlimited ressources gains } W\} = \begin{cases} 1 & \text{if } q < p \\ (p/q)^W & \text{if } p < q \end{cases}$$

This is *bad news*: even with infinite resources the probability to ever win a certain given amount in a casino is exponential small!

• $W \to \infty$ means that the player has no limit and he will be playing either forever or until he looses his original fortune of L. We have

$$P\{\text{A gambler with fortune L plays for ever }\} = \begin{cases} 1-(q/p)^L & \text{if } q < p \\ 0 & \text{if } p < q \end{cases}$$

This is *bad news* again: in a casino the probability to play forever is 0.

How much is free money worth in casino? Imagine that one of your friend is an extremely rich owner of a casino and make you the following present. You can go his casino and he will give you an *infinite credit*, provided that you play at a table of craps ($p = 244/495$, $q = 251/495$) which happens to have a house limit of \$15,000 (maximal bet allowed). You of course then decide to be bold and bet the maximum of \$15,000 every time. Based on your knowledge of the gambler's ruin formula, you fix yourself a goal of W (in units of \$15000) and deduce that you will reach W with probability $(p/q)^W$ and never reach W with probability $(1-(p/q)^W)$. So on average you will make

$$15,000\, W(p/q)^W$$

money with your free credit. Since you are free to fix W and the optimal choice W^* is obtained my maximization, that is $W^* = \operatorname{argmax} W(p/q)^W$. Differentiating with respect to W gives

$$0 = \frac{d}{dW} W(p/q)^W \;=\; (p/q)^W + \ln(p/q) W(p/q)^W \,.$$

or $W^* = \frac{1}{\ln(q/p)} = 35.35$ and the expected amount of money you can extract out of this infinite credit line is a (paltry)

$$15,000\, W^*(p/q)^{W^*} \;=\; 15,000\, \frac{1}{\ln(q/p)e} \;\approx 195,000\,.$$

Time until absorption. It is also instructive to compute the time until absorption, T, the number of games that a gambler with resources j can play before losing (or winning). To do this let us define

$$G(j) \;=\; E[T \,|\, X_0 = j]\,.$$

Clearly we have $G(0) = G(N) = 0$. Conditioning on the first step then we find the ***second order linear inhomogeneous difference equation***

$$G(j) \;=\; 1 + qG(j-1) + rG(j) + pG(j+1)\,.$$

We will solve it here for $p = q = 1/2$. An educated guess is to try for the particular solution $G(j) = aj^2$ (since 1 and j solve the homogenerous equation) which yields $a = -1$. Therefore the general solution has the form

$$G(j) = c_1 + c_2 j - j^2$$

and using the boundary conditions we find

$$E[T \,|\, X_0 = j] \;=\; j(N-j)\,.$$

2.6 Reversible Markov chains

Let us consider a Markov chain with transition probabilities $P(i,j)$ and stationary distribution $\pi(i)$. The equation for π is

$$\pi(i) \;=\; \sum_j \pi(j)P(j,i)\,,$$

which we can rewrite as

$$\sum_j \pi(i)P(i,j) \;=\; \sum_j \pi(j)P(j,i)\,. \tag{2.14}$$

It is useful to interpret this equation as ***balance equation***. Let us set

$$J(i,j) \equiv \pi(i)P(i,j)$$

and we can interpret $J(i,j)$ as the ***probability current*** from i to j. The equation (2.14) means that

$$\sum_i J(i,j) \;=\; \sum_j J(j,i)\,, \tag{2.15}$$

i.e., to be stationary the total probability current from i must be equal to the total probability current into i.

A stronger condition for stationarity can be expressed in terms of the balance between the currents $J(i,j)$ and this called ***detailed balance***.

Definition 2.6.1 *A Markov chain X_n satisfies detailed balance if there exists $\pi(i) \geq 0$ with $\sum_i \pi(i) = 1$ such that for all i,j we have*

$$\pi(i)P(i,j) \;=\; \pi(j)P(j,i)\,. \tag{2.16}$$

This means that for every pair i,j the probability currents $J(i,j)$ and $J(j,i)$ balance each other. Clearly (2.16) is a stronger condition than (2.14) and thus we have

Lemma 2.6.2 *If the Markov chain satisfies detailed balance for a probability distribution π then π is a stationary distribution.*

But it is easy to see that detailed balance is a stronger condition than stationarity. The property of detailed balance is often called ***(time)-reversibility*** since we have

Lemma 2.6.3 *Suppose the Markov chain X_n satisfies detailed balance and assume that the initial distribution is the stationary distribution π. Then for any sequnce of states $i_0, \cdots i_n$ we have*

$$P\left\{X_0 = i_0\,, X_1 = i_1\,, \cdots\,, X_n = i_n\right\} \;=\; P\left\{X_0 = i_n\,, X_1 = i_{n-1}\,, \cdots\,, X_n = i_0\right\} \tag{2.17}$$

Proof: Using the detailed balance equation repeatedly we have

$$\begin{aligned} P\{X_0 = i_0, X_1 = i_1, \cdots, X_n = i_n\} &= \pi(i_0)P(i_0,i_1)P(i_1,i_2)\cdots P(i_{n-1},i_n) \\ &= P(i_1,i_0)\pi(i_1)P(i_1,i_2)\cdots P(i_{n-1},i_n) \\ &= P(i_1,i_0)P(i_2,i_1)\pi(i_2)\cdots P(i_{n-1},i_n) \\ &\cdots \\ &= P(i_1,i_0)P(i_2,i_1)\cdots\pi(i_{n-1})P(i_{n-1},i_n) \\ &= P(i_1,i_0)P(i_2,i_1)\cdots P(i_n,i_{n-1})\pi(i_n) \\ &= P\{X_0 = i_n, X_1 = i_{n-1}, \cdots, X_n = i_0\} \end{aligned}$$

∎

The next result is very easy and very useful.

Proposition 2.6.4 *Suppose X_n is a Markov chain with state space S and with a* ***symmetric*** *transition matrix, i.e, $P(i,j) = P(j,i)$. Then X_n satisfies detailed balance with $\pi(j) = \text{const} = 1/|S|$, i.e., the stationary distribution is uniform on S.*

Proof: obvious. ∎

Example 2.6.5 Let us consider the ***random walk on the hypercube*** $\{0,1\}^m$. The state space

$$S = \{0,1\}^m = \{\sigma = (\sigma_1, \cdots, \sigma_m);\ \sigma_i \in \{0,1\}\}$$

To define the move of the random walk, just pick one coordinate $j \in \{1, \cdots, m\}$ and flip the j^{th} coordinate, i.e., $\sigma_j \to 2\sigma_j - 1$. We have thus

$$P(\sigma,\sigma') = \begin{cases} \frac{1}{m} & \text{if } \sigma \text{ and } \sigma' \text{ differ by one coordinate} \\ 0 & \text{otherwise} \end{cases}$$

Clearly P is symmetric and thus $\pi(\sigma) = 1/2^m$. ∎

Example 2.6.6 Let us consider a ***simple random walk on the graph*** $G = (E,V)$ with the transition probabilities $p(v,w) = \frac{1}{\deg(v)}$. Let us check that this Markov chain is satifies detailed balance with the unnormalized $\mu(v) = deg(v)$. Indeed we have $P(v,w) > 0$ if and only if $P(w,v) > 0$ and thus if $P(v,w) > 0$ we have

$$\mu(v)P(v,w) = \deg(v)\frac{1}{\deg(v)} = 1 = \mu(w)P(w,v).$$

This is slightly easier to verify that the stationary equation $\pi P = \pi$. After normalization we find $\pi(v) = \deg(v)/2|E|$.

For example for the simple random walk on $\{0, 1, \cdots, N\}$ with reflecting boundary conditions we obtain in this way

$$\pi \;=\; \left(\frac{1}{2N}, \frac{2}{2N}, \cdots, \frac{1}{2N}\right) .$$

■

Example 2.6.7 (Network) The previous example can be generalized as follows. For a given graph $G = (E, V)$ let us assign a positive weight $c(e) > 0$ to each edge $e = \{v, w\}$, that is we choose numbers $c(v, w) = c(w, v)$ with $c(v, w) = 0$ if v and w are not connected by an edge. If the transition probabilities are given by

$$P(v, w) \;=\; \frac{c(v, w)}{c(v)}\,, \quad \text{with } c(v) \;=\; \sum_{w} c(v, w)\,,$$

then it is easy to verify that the Markov chain satisfies detailed balance with

$$\pi(v) \;=\; \frac{c(v)}{c_G}\,, \quad \text{with } c_G \;=\; \sum_{v} c(v)\,.$$

■

Example 2.6.8 (Birth-Death Processes) Let us consider a Markov chain on the state space $S = \{0, \cdots, N\}$ with transition probabilities

$$\begin{aligned} P(j, j) &= r_j\,, \quad j = 0, \cdots, N\,, \\ P(j, j+1) &= p_j\,, \quad j = 0, \cdots, N_1\,, \\ P(j, j-1) &= q_j\,, \quad j = 1, \cdots, N\,, \end{aligned}$$

and all the other $P(i, j)$ vanish. This is called a ***birth and death process*** since the only possible transition are to move up or down by unit or stay unchanged.

These Markov chains always satisfy detailed balance. Indeed the non trivial detailed balance conditions are

$$\pi(j)p_j \;=\; \pi(j+1)q_{j+1}\,, \quad j = 0, \cdots, N-1\,.$$

and this can be solved recursively. We obtain

$$\begin{aligned} \pi(1) &= \pi(0)\frac{p_0}{q_1} \\ \pi(2) &= \pi(1)\frac{p_1}{q_2} \;=\; \pi(0)\frac{p_0 p_1}{q_1 q_2} \\ &\vdots \\ \pi(N) &= \pi(0)\frac{p_0 p_1 \cdots p_{N-1}}{q_1 q_2 \cdots q_{N-1}} \end{aligned}$$

and with normalization

$$\pi(j) \;=\; \frac{\prod_{k=1}^{j} \frac{p_{k-1}}{q_k}}{\sum_{l=0}^{N} \prod_{k=1}^{l} \frac{p_{k_1}}{q_k}}$$

For example the Ehrenfest urn in Example 2.2.8 model has

$$p_j \;=\; \frac{N-j}{N}\,, \quad q_j \;=\; \frac{j}{N}$$

and thus we obtain

$$\pi(j) \;=\; \pi(0) \frac{\frac{N}{N}\frac{N-1}{N}\cdots\frac{N-(j-1)}{N}}{\frac{1}{N}\frac{2}{N}\cdots\frac{j}{N}} \;=\; \pi(0)\binom{N}{j}$$

and the normalization is $\sum_{j=0}^{N}\binom{N}{j} \;=\; 2^N$. ∎

2.7 Monte-Carlo Markov chains

Suppose you are given a certain probability distribution π on a set S and you goal is to generate a sample from this distribution. The ***Monte-Carlo Markov chain*** method consists in constructing an irreducible Markov chain X_n whose stationary distribution is π. Then to generate π one simply runs the Markov chains X_n long enough such that it is close to its equilibrium distribution. It turns out that using the detailed balance condition is a very useful tool to construct the Markov chain in this manner.

A-priori this method might seems an unduly complicated way to sample from π. Indeed why not simply simulate from π directly using one of the method of Section 1? To dispel this impression let us consider some concrete examples.

Example 2.7.1 (Proper q-coloring of a graph) Let $G = (E, V)$ be a graph. A ***proper q-coloring*** of a graph consist of assigning to each vertex v of the graph one of q colors subject to the constraint that if 2 vertices are linked by an edge they should have different colors. Let S' be the set of all such proper q-coloring which is a subset of $S \;=\; \{1, \cdots, q\}^V$. Let us denote the elements of S by $\sigma = \{\sigma(v)\}_{v\in V}$ with $\sigma(v) \in \{1, \cdots, q\}$. Let π be the uniform distribution on all such proper coloring, i.e., $\pi(\sigma) = 1/|S'|$ for all $\sigma \in S'$. Even for moderately complicated graph it can be very difficult to compute to $|S'|$!

A Monte-Carlo method can be used to generate π even without an explicit knowledge of $|S'|$. Suppose $X_n = \sigma$, then the transition probabilites are generated by the algorithm

(i) Choose a vertex v of at random and choose a color a at random.

(ii) Set $\sigma'(v) = a$ and $\sigma'(w) = \sigma(w)$ for $w \neq v$.

(iii) If σ' is a proper q-coloring then set $X_{n+1} = \sigma'$. Otherwise set $X_n = \sigma$.

The transition probabilities are given by

$$\begin{aligned} P(\sigma, \sigma') &= \frac{1}{q|V|} \text{ if } \sigma \text{ and } \sigma' \text{ differ at exactly one vertex} \\ P(\sigma, \sigma') &= 0 \text{ if } \sigma \text{ and } \sigma' \text{ differ at more than one vertex} \\ P(\sigma, \sigma) &= 1 - \sum_{\sigma'} P(\sigma, \sigma') \end{aligned}$$

Note that $|S'|$ does not enter in the transition probabilities. Note further that $P(\sigma, \sigma)$ is not known explicitly either but is also not used to run the algorithm.

In order to check that the uniform distribution is stationary for this Markov chain it is enough to note that P is a symmetric matrix. Indeed if one can change σ into σ' by changing one color then one can do the reverse transformation too. ∎

Let us considering another example which is a fairly classical optimization problem.

Example 2.7.2 (Knapsack problem). Suppose you own m books and the i^{th} book has weight w_i lb and is worth \$ v_i. In your knapsack you can put at most a total of b pounds and you are looking to pack the most valuable knapsack possible.

To formulate the problem mathematically we introduce

$$\begin{aligned} w &= (w_1, \cdots w_m) \in \mathbf{R}^m, \quad \text{weight vector} \\ v &= (v_1, \cdots v_m) \in \mathbf{R}^m, \quad \text{value vector} \\ \sigma &= (\sigma_1, \cdots \sigma_m) \in \{0,1\}^m, \quad \text{decision vector} \end{aligned}$$

where we think that $\sigma_i = 1$ is the i^{th} item is in the knapsack. The state space is

$$S' = \{\sigma \in \{0,1\}^m ; \sigma \cdot w \leq b\}$$

and the optimization problem is

$$\text{Maximize } v \cdot \sigma \text{ subject to } \sigma \in S'.$$

As a first step we discuss the problem of generating a random element in S' using a simple algorithm. If $X_n = \sigma$ then

(i) Choose $j \in \{1, \cdots m\}$ at random.

(ii) Set $\sigma' = (\sigma_1, \cdots, 1 - \sigma_j, \cdots, \sigma_m)$.

(iii) If $\sigma' \in S'$, i.e., if $\sigma' \cdot v \leq b$ then let $X_{n+1} = \sigma'$. Otherwise $X_{n+1} = \sigma$.

In other words, choose a random book. If it is in the sack already remove it. If it is not in the sack add it provided you do not exceed the the maximum weight. Note

that the Markov chain X_n is irreducible, since each state communicates with the state $\sigma = (0, \cdots, 0)$. It is aperiodic except in the uninteresting case where $\sum_i w_i \leq b$. Finally the transition probabilities are symmetric and thus the uniform distribution the unique stationary distribution. ∎

In the knapsack problem we want to maximize a function f on the state space. One possible algorithm would be to generate an uniform distribution on the state space and then to look for the maximum value of the function. But it would be a better idea to sample from a distribution which assign higher probabilities to the state with a high value of f.

Let S be the state space and let $f : S \to \mathbb{R}$ be a function. It is convenient to introduce the probability distributions define for $\beta > 0$ by

$$\pi_\beta(i) = \frac{e^{\beta f(i)}}{Z_\beta} \quad \text{with} \quad Z_\beta = \sum_{j \in S} e^{\beta f(j)} .$$

Clearly π_β assign higher weights to the the i with bigger values of $f(i)$. Let us define

$$S^* = \left\{ i \in S \,;\, f(i) = f^* \equiv \max_{j \in S} f(j) \right\} .$$

If $\beta = 0$ then π_0 is simply the uniform distribution on S. For $\beta \to \infty$ we have

$$\lim_{\beta \to \infty} \pi_\beta(i) = \lim_{\beta \to \infty} \frac{e^{\beta(f(i)-f^*)}}{|S^*| + \sum_{j \in S \setminus S^*} e^{\beta(f(j)-f^*)}} = \begin{cases} \frac{1}{|S^*|} & \text{if } j \in S^* \\ 0 & \text{if } j \notin S^* \end{cases} ,$$

i.e., for large β π_β is concentrated on the global maxima of f.

A fairly general method to generate a distribution π on the state space S is given by the ***Metropolis algorithm***. This algorithm assumes that you already know how to generate the uniform distribution on S by using a symmetric transition matrix Q.

Algorithm 2.7.3 (Metropolis algorithm with proposal matrix Q) Let Q be a symmetric transition matrix. If $X_n = i$ then
(i) Choose $Y \in S$ according to Q, i.e.,

$$P\{Y = j \,|\, X_n = i\} = Q(i, j) .$$

(ii) Define the acceptance probability

$$\alpha = \min \left\{ 1 \,, \frac{\pi(Y)}{\pi(i)} \right\} .$$

(iii) Accept Y with probability α. That is generate a random number U. If $U \leq \alpha$ then $X_{n+1} = Y$ (i.e., accept the move) and if $U > \alpha$ then $X_{n+1} = X_n$ (i.e., reject the move). ∎

The general case with non-symmetric proposal matrix is called the ***Metropolis-Hastings algorithm*** and is discussed in Exercise **??**. We have

Proposition 2.7.4 *Suppose Q is an irreducible transition probability matrix on S and suppose π is a probability distribution on S with $\pi(i) > 0$. Then the Metropolis algorithm defines an irreducible Markov chain on S which satisfies detailed balance with stationary distribution π.*

Proof: Let $P(i,j)$ be the transition probabilities for the Metropolis Markov chain. Then we have

$$P(i,j) \;=\; Q(i,j)\alpha \;=\; Q(i,j)\min\left\{1\,,\;\frac{\pi(j)}{\pi(i)}\right\}\,.$$

Since $\pi(i) > 0$ the acceptance probability α never vanishes. Thus if $P(i,j) > 0$ whenever $Q(i,j) > 0$ and thus P is irreducible if Q is.

In order to check the reversibility we note that

$$\pi(i)P(i,j) \;=\; Q(i,j)\pi(i)\min\left\{1,\frac{\pi(j)}{\pi(i)}\right\} \;=\; Q(i,j)\min\left\{\pi(i)\,,\;\pi(j)\right\}$$

and the r.h.s is symmetric in i,j and thus $\pi(i)P(i,j) = \pi(j)P(j,i)$. ∎

Note that only the ratio $\pi(i)/pi(j)$ are needed to run the algorithm, in particular we do not need the normalization constant.

Example 2.7.5 (Knapsack problem) Let us consider the probability distribution

$$\pi_\beta(\sigma) \;=\; e^{\beta v\cdot\sigma}Z_\beta\,.$$

The normalization constant $Z_\beta = \sum_{\sigma\in S'} e^{\beta v\cdot\sigma}$ is almost always impossible to compute. However we have

$$\frac{\pi(\sigma')}{\pi(\sigma)} = e^{\beta v\cdot(\sigma'-\sigma)}$$

which does not involve Z_β.

For this distribution we take as the Q matrix constructed in Example 2.7.2 and the Metropolis algorithm is

If $X_n = \sigma$ then
(i) Choose $j \in \{1,\cdots m\}$ at random.
(ii) Set $\sigma' \;=\; (\sigma_1,\cdots,1-\sigma_j,\cdots,\sigma_m)$.
(iii) If $\sigma' \notin S'$, i.e., then $X_{n+1} = \sigma$.
(iv) If $\sigma' \in S'$, i.e., then let

$$\alpha \;=\; \min\left\{1,\frac{\pi(\sigma')}{\pi(\sigma)}\right\} \;=\; \min\left\{1,e^{\beta v\cdot(\sigma'-\sigma)}\right\} \;=\; \begin{cases} e^{-\beta v_j} & \text{if } \sigma_j = 1 \\ 1 & \text{if } \sigma_j = 0 \end{cases}$$

(v) Generate a random number U, If $U \leq \alpha$ then $X_{n+1} = \sigma'$. Otherwise $X_{n+1} = \sigma$.

If you can add a book to your knapsack you always do while you remove a book with a probability which is exponentially related to the weight of the book. ∎

Another algorithm which is widely used for Monte-Carlo Markov chain is the ***Glauber algorithm*** which appear in the literature under a variety of other names such as ***Gibbs sampler*** in statistical applications, ***logit rule*** in economics and social sciences, ***heat bath*** in physics, and undoubtedly under various other names.

The Glauber algorithm is not quite as general as the Metropolis algorithm. We assume that the state space S has the following structure

$$S \subset \Omega^V$$

where both Ω and V are finite sets. For example $S \subset \{0,1\}^m$ in the case of the knapsack problem or $S \subset \{1, \cdots, q\}^V$ for the case of the proper q-coloring of a graph. We denote by

$$\sigma \,=\, \{\sigma(v)\}_{v \in V}\,, \quad \sigma(v) \in \Omega\,.$$

the elements of S.

It is useful to introduce the notation

$$\sigma_{-v} \,=\, \{\sigma(w)\}_{w \in V, w \neq v}$$

and we write

$$\sigma \,=\, (\sigma_{-v}, \sigma(v))\,.$$

Algorithm 2.7.6 (Glauber algorithm) Let π be a probability distribution on $S \subset \Omega^V$. Extend π to Ω^V by setting $\pi(\sigma) = 0$ if $\sigma \in \Omega^V \setminus S$. If $X_n \,=\, \sigma$ then

(i) Choose $v \in V$ at random.

(ii) Replace $\sigma(v)$ by a new value $a \in \Omega$ (provided $(\sigma_{-v}, a) \in S$) with probability

$$\frac{\pi(\sigma_{-v}, a)}{\sum_{b \in \Omega} \pi(\sigma_{-v}, b)}\,.$$

∎

The irreducibility of the algorithm is not guaranteed a-priori and needs to be checked on a case-by-case basis. We have

Proposition 2.7.7 *The Glauber algorithm defines a Markov chain on S which satisfies detailed balance with stationary distribution π.*

Proof: The transition probabilities are given by

$$\begin{aligned} P(\sigma,\sigma') &= \frac{1}{|V|}\frac{\pi(\sigma_{-v},\sigma'(v))}{\sum_{b\in\Omega}\pi(\sigma_{-v},b)} \text{ if } \sigma_{-v}=\sigma'_{-v} \text{ for some } v \\ P(\sigma,\sigma') &= 0 \text{ if } \sigma_{-v}\neq\sigma'_{-v} \text{ for all } v \\ P(\sigma,\sigma) &= 1-\sum_{\sigma'} P(\sigma,\sigma') \end{aligned}$$

To check detailed balance we note that if $P(\sigma,\sigma')\neq 0$

$$\pi(\sigma)P(\sigma,\sigma') \;=\; \frac{\pi(\sigma)\pi(\sigma')}{\sum_{b\in\Omega}\pi(\sigma_{-v},b)}\,,$$

and this is symmetric in σ and σ'. ∎

Example 2.7.8 (Ising Model on a graph) Let $G=(E,V)$ be a graph and let $S=\{-1,1\}^V$. That is to each vertex assign the value ± 1, you can think of a magnet at each vertex pointing either upward $(+1)$ or downward -1). To each $\sigma\in S$ we assign an "energy" $H(\sigma)$ given by

$$H(\sigma) \;=\; -\sum_{e=(v,w)\in E}\sigma(v)\sigma(w)\,.$$

The energy σ is minimal if $\sigma(v)\sigma(w)=1$ i.e., if the magnets at v and w are aligned. Let us consider the probability distribution

$$\pi_\beta(\sigma) \;=\; \frac{e^{-\beta H(\sigma)}}{Z_\beta}\,,\quad Z_\beta \;=\; \sum_\sigma e^{-\beta H(\sigma)}\,.$$

The distribution π_β is concentrated around the minima of $H(\sigma)$. To describe the Glauber dynamics note that

$$H(\sigma_{-v},1)-H(\sigma_{-v},-1) \;=\; -2\sum_{w\,;\,w\sim v}\sigma(w)$$

and this can be computed simply by looking at the vertices connected to v and not at all the graph. So the transition probabilities for the Glauber algorithm are given by picking a vertex at random and then updating with probabilities

$$\frac{\pi(\sigma_{-v},\pm 1)}{\pi(\sigma_{-v},1)+\pi(\sigma_{-v},-1)} \;=\; \frac{1}{1+e^{\pm\beta[H(\sigma_{-v},1)-H(\sigma_{-v},-1)]}} \;=\; \frac{1}{1+e^{\mp 2\beta\sum_{w\,;\,w\sim v}\sigma(w)}}\,.$$

By comparison for the Metropolis algorithm we pick a vertex at random and switch $\sigma(v)$ to $-\sigma(v)$ and accept the move with probability

$$\min\left\{1,\frac{\pi(\sigma_{-v},-\sigma(v))}{\pi(\sigma_{-v},\sigma(v))}\right\} \;=\; \min\left\{1,\frac{\pi(\sigma_{-v},-\sigma(v))}{\pi(\sigma_{-v},\sigma(v))}\right\} \;=\; \min\left\{1,e^{2\beta\sum_{w\,;\,w\sim v}\sigma(w)\sigma(v)}\right\}\,.$$

Chapter 3

Markov Chains with Countable State Space

3.1 Definitions and examples

We consider now Markov chains with a ***countable state space*** S, i.e., the set S can be put in 1-to-1 correspondence with the set of positive integers $\mathbf{N}$, e.g $S = \mathbf{Z}$ of $S = \mathbf{Z^d}$, and so on...

Many definitions for finite Markov chains carry over in a straightforward manner to countable Markov chains and we will mostly insist on what is different for countable state spaces. A Markov chain X_n on S is defined by the transition probabilities

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

and as before the k-step transition probabilities

$$P\{X_{n+k} = j \,|\, X_n = i\} = P^k(i,j)$$

satisfies the Chapman-Kolmogorov equations

$$P^{m+n}(i,j) \;=\; \sum_{k\in S} P^m(i,k)P^n(k,j)\,.$$

which we may know think of as an "infinite matrix" whose row sum up to 1, i.e. $\sum_{j\in S} P(i,j) = 1$. A probability distribution for X_n can be represented by an (infinite) row vector μ and as before if X_0 has distribution μ then X_n has distribution μP^n where $\mu P(j) \sum_i \mu(i)P(i,j)$. We cannot appeal to linear algebra considerations anymore, or it has been done with the much more sophisticated tools of functional analysis.

First note that the concepts of ***stationary distribution***, **em limiting distribution**, **em communications**, ***irreducibility***, ***period of a state***, ***closed classes***,

and ***transient classes***, ***reversibility*** all extend without difficulty to countable state space.

There are notable differences though, that we shall explore in the next sections. For a finite state irreducible Markov chain there is a unique stationary distribution π, or more generally there is a stationary distribution associated to any single closed class if the Markov chain is reducible. This is *not true anymore* in general for countable state space Markov chain. There is the possiblity that the Markov chain "escapes to infinity" after some time and never comes back even though the chain is irreducible.

Let us start with some simple examples

Example 3.1.1 (Random walk on$\{0,1,2,3,\cdots,\}$) Let us consider a random walk on the set of nonnegative integers with partially reflecting boundary conditions at 0. The transition probabilities are given by

$$P = \begin{array}{c} 0 \\ 1 \\ 2 \\ \vdots \end{array} \begin{pmatrix} q & p & 0 & 0 & \dots \\ q & 0 & p & 0 & \dots \\ 0 & q & 0 & p & \\ \vdots & \ddots & \ddots & \ddots & \ddots \end{pmatrix}$$

with $q = 1 - p$.

Example 3.1.2 (Discrete-time queueing model) Imagine a service station, say a single cash register at a store. During each time period there is a probability p that a an additional comes in the queue. The first person in the queue is being served and during each time period there is a probability queue that this person exits the queue.

We denote by X_n the number of people in the queue (either in being served or waiting in line). The state space is $S = \{0,1,2,3,\cdots\}$ and the transition probabilities are

$$P = \begin{array}{c} 0 \\ 1 \\ 2 \\ \vdots \end{array} \begin{pmatrix} 1-p & p & 0 & 0 & \dots \\ q(1-p) & qp+(1-p)(1-q) & p(1-q) & 0 & \dots \\ 0 & q(1-p) & qp+(1-p)(1-q) & p(1-q) & \\ \vdots & \ddots & \ddots & \ddots & \ddots \end{pmatrix}$$

Example 3.1.3 (Repair shop) A repair shop is able to repair one item one any single. During day n Z_n machines break down are brought for repair to the shop, we assume that Z_n are IID random variables with p.d.f $P\{Z_n = k\} = a_k$ for $k = 0,1,2,\cdots$. If X_n denotes the number of item in the shop waiting to be repaired we have

$$X_{n+1} = \max\{(X_n - 1), 0\} + Z_n$$

The state space is $S = \{0, 1, 2, 3, \cdots\}$ and the transition probabilities are

$$P = \begin{array}{c} 0 \\ 1 \\ 2 \\ 3 \\ \vdots \end{array} \begin{pmatrix} a_0 & a_1 & a_2 & a_3 & \cdots \\ a_0 & a_1 & a_2 & a_3 & \cdots \\ 0 & a_0 & a_1 & a_2 & \cdots \\ 0 & 0 & a_0 & a_1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \end{pmatrix}$$

Example 3.1.4 (Success run chain) The state space is the set of nonnegative integers $\mathbf{N}$. Imagine a player taking a series of bet where the probability of winning the j^{th} bet is p_j. We let X_n denotes the number of successive winning bets. The Markov chain X_n has state space $S = \{0, 1, 2, 3, \cdots\}$ and transition probabilities

$$P = \begin{array}{c} 0 \\ 1 \\ 2 \\ \vdots \end{array} \begin{pmatrix} q_0 & p_0 & 0 & 0 & \dots \\ q_1 & 0 & p_1 & 0 & \dots \\ q_2 & 0 & 0 & p_2 & \\ \vdots & \vdots & & \ddots & \ddots \end{pmatrix}$$

with $q_i = 1 - p_i$. This Markov chain has allows for nice analytical computations. ∎

Example 3.1.5 (Simple d-dimensional random walk) The state space of the Markov chain in $\mathbb{Z}^d$. We denote by $\mathbf{e}_i$, $i = 1, \cdots, d$ the standard orthonormal basis in $\mathbb{R}^d$. We view $\mathbb{Z}^d$ as the vertex set of a graph and any point $\mathbf{x} = (\mathbf{x_1}, \cdots, \mathbf{x_d})$ is connected by edges to $2d$ neighbors $\mathbf{x} \pm \mathbf{e}_i$. For the simple random walk we have

$$p(\mathbf{x}, \mathbf{x} \pm \mathbf{e}_i) = \frac{1}{2d}$$

and all the others $p(\mathbf{x}, \mathbf{y}) = 0$.

Example 3.1.6 (Branching process) The ***branching process***, also known as the ***Galton-Watson process*** model the evolution over time of population. In a unit of time every individual in a population and leave behind a random number of descendents. To describe the Markov chain we will use IID random variables $Z_n^{(k)}$ indexed by $n = 0, 1, 2 \cdots$ and $k = 0, 1, 2, \cdots$. We have

$$X_{n+1} = \sum_{k=1}^{X_n} Z_n^{(k)}$$

which simply says that the each X_n individuals in the population each has a random number $Z_n^{(k)}$ of descendents. It is not convenient to write down the transition probabilities but we will study this process later using its moment generating function.

3.2 Recurrence and transience

We define first the two fundamental concepts of ***recurrence*** and ***transience*** of a state. Recall the ***return time to state*** j, $\tau^{(j)}$, is given by

$$\tau^{(i)} = \min\{n \geq 1\,;\, X_n = i\}\,.$$

Definition 3.2.1 (Recurrence-Transience)

1. *A state i is recurrent if the Markov chain starting in i will eventually return to i with probability 1, i.e. if*

$$P\left\{\tau^{(i)} < \infty | X_0 = i\right\} = 1$$

2. *A state i is transient if it is not recurrent, that is starting in i the Markov chain return to i with probability $q < 1$, i.e., if*

$$P\left\{\tau^{(i)} < \infty | X_0 = i\right\} = q < 1$$

Let consider the random variable $Y^{(j)}$ ***which counts the number of visits to state*** j, i.e.,

$$Y^{(i)} = \sum_{k=0}^{\infty} I_{\{X_k = i\}}\,.$$

and recall that

$$E[Y^{(i)} | X_0 = j] = \sum_{k=0}^{\infty} P^k(j,i)$$

If j is recurrent, X_n, starting from j, will return to j with probability 1, and then by the Markov property will return a second time with probability 1. Repeating the argument we find that

$$\text{The state } i \text{ is recurrent } \Leftrightarrow P\left\{Y^{(i)} = \infty \,|\, X_0 = i\right\} = 1\,.$$

On the other hand if i is transient then the probability of returning to i is $q < 1$ and by the Markov property the number of returns $Y^{(j)}$ is a geometric random variables with mean $1/q$

Summarizing we have

Proposition 3.2.2 (Transience/recurrence dichotomy)

1. *i is recurrent $\Leftrightarrow$ $Y^{(i)} = +\infty$ with probability 1 (conditioned on $X_0 = 1$) $\Leftrightarrow$ $\sum_k P^k(i,i) = +\infty$.*

2. *i is recurrent $\Leftrightarrow$ $Y^{(i)} < \infty$ with probability 1 (conditioned on $X_0 = 1$) $\Leftrightarrow$ $\sum_k P^k(i,i) = +\infty$.*

The next proposition shows, in particular, that transience and recurrence are properties of communication classes: all the states in a class are either recurrent or transient. If a MC is irreducible then we can call the Markov chain transient or recurrent.

Proposition 3.2.3 *The following are equivalent*

(i) $\sum_{k=0}^{\infty} P^k(j_0, j_0) = \infty$ *for some state* j_0

(ii) $\sum_{k=0}^{\infty} P^k(i,j) = \infty$ *for all states* i, j *in the communication class of* j_0.

(iii) $P\left\{\tau^{(j_0)} < \infty | X_0 = j_0\right\} = 1$

(iv) $P\left\{\tau^{(j)} < \infty | X_0 = i\right\} = 1$ *for all states* i, j *in the communication class of* j_0.

Proof: The equivalence between (i) and (iii) has already been established.

Suppose that (i) holds, since i, j communicates with j_0 there exists l and m such that $P^l(i, j_0) > 0$ and $P^m(j_0, j) > 0$. Then since

$$P^{k+l+m}(i,j) \geq P^l(i,j_0) P^k(j_0,j_0) P^m(j_0,j) > 0$$

we have

$$\sum_{k=0}^{\infty} P^k(i,j) \geq \sum_{k=0}^{\infty} P^{k+l+m}(i,j) \geq P^l(i,j_0) \sum_{k=0}^{\infty} P^k(j_0,j_0) P^m(j_0,j) = \infty\,,$$

which establishes the equivalence of ((i) and (ii).

Suppose that (iii) holds. By irreducibility we must have $P\left\{\tau^{(i)} < \tau^{(j_0)} | X_0 = j_0\right\} > 0$ (argue by contradiction, if this probability were 0 by the Markov property, the chain would never visits k starting from j). Therefore

$$0 \;=\; P\left\{\tau^{(j_0)} = \infty | X_0 = j_0\right\} \;\geq\; P\left\{\tau^{(i)} < \tau^{(j_0)} | X_0 = j_0\right\} P\left\{\tau^{(j_0)} = \infty | X_0 = i\right\}$$

and therefore $P\left\{\tau^{(j_0)} < \infty | X_0 = i\right\} = 1$. On the other hand, starting from j_0, by irreducibility if X_n has a positive probability to visit j, independently of the past. Since X_n visits j_0 infinitely often this implies that X_n must eventually visit the state j. To summarize starting from i the Markov chain visits j_0 with probability 1 and starting from j_0 the Markov chain visits j with probability 1. This implies that $P\left\{\tau^{(j)} < \infty | X_0 = j_0\right\} = 1$. This shows that (iii) and (iv) are equivalent. ∎

If the state space is finite, an irreducible Markov chain is always recurrent but this is not case for countable state space as it is possible for the Markov chain can wander away to infinity and never come back.

Example 3.2.4 (transience for the assymetric random walk) Consider a random walk on $\mathbb{Z}$ with $P(j, j+1) = p$ and $P(j, j-1) = 1-p$. Starting the Markov chain at 0 we have

$$X_n = Y_1 + \cdots Y_n$$

where the Y_i are IID random variable with $P(Y_i = 1) = p$ and $P(Y_i = 1-p)$ and $E[Y_i] = 2p-1$. By the Law of Large Nubers we have

$$\lim_{n\to\infty} \frac{X_n}{n} = 2p - 1$$

almost surely and thus if $p \neq \frac{1}{2}$ we must have $\lim_n X_n \to \infty$ with probability 1 and thus X_n is transient. If $p = \frac{1}{2}$ this argument does not work and we must do something else.

Example 3.2.5 (recurence and transience for the simple random walk). Consider the simple random walk of Example 3.1.5

- For $d = 1$ we consider the state 0 and note that the chain has period 2 and $P^{2k+1}(0,0) = 0$. To return to 0 in $2n$ steps the Markov chain must take exactly n steps to the left and n steps to the right and thus we have

  $$P^{2n}(0,0) = \binom{2n}{n} \frac{1}{2^{2n}}$$

  By Stirling's formula we have $n! \sim \sqrt{2\pi n} e^{-n} n^n$ where $a_n \sim b_n$ means that $\lim a_n/b_n = 1$. Thus we have

  $$\binom{2n}{n} \frac{1}{2^{2n}} \sim \frac{1}{2^{2n}} \frac{\sqrt{2\pi 2n} e^{-2n} (2n)^{2n}}{2\pi n e^{-2n} n^{2n}} = \frac{1}{\sqrt{\pi n}}.$$

  If $a_n \sim b_n$ then $\sum a_n$ converges if and only if $\sum b_n$ converges and thus the random walk in $d = 1$ is recurrent.

- For $d = 2$ to return to 0 in $2n$ steps the Markov chain must take exactly k steps to the left and k steps to the right and $n-k$ steps up and $n-k$ steps down. Therefore

  $$P^{2n}(0,0) = \sum_{k=0}^{n} \frac{2n!}{k!k!(n-k)!(n-k)!} \frac{1}{4^{2n}} = \frac{1}{4^{2n}} \binom{2n}{n} \sum_{k=0}^{n} \binom{n}{k} \binom{n}{n-k}$$

  Now we claim that $\sum_{k=0}^{n} \binom{n}{k}\binom{n}{n-k} = \binom{2n}{n}$ as can be seen by the counting the number of ways that a team of n can be formed out of n boys and n girls. Therefore

  $$P^{2n}(0,0) = \binom{2n}{n}^2 \frac{1}{4^{2n}} \sim \frac{1}{\pi n}$$

  and thus the simple random walk is recurrent for $d = 2$ since $\sum \frac{1}{n}$ diverges.

- For $d = 3$ we proceed in a similar way and obtain

$$\begin{aligned} P^{2n}(0,0) &= \sum_{k,j:k+j\le n} \frac{2n!}{j!j!k!k!(n-k-j)!(n-k-j)!}\frac{1}{6^{2n}} \\ &= \frac{1}{2^{2n}}\binom{2n}{n}\sum_{k,j:k+j\le n}\left(\frac{1}{3^n}\frac{n!}{j!k!(n-j-k)!}\right)^2 \end{aligned}$$

Note that $\sum_{k,j:k+j\le n}\frac{1}{3^n}\frac{n!}{j!k!(n-j-k)!} = 1$ (trinomial coefficients) and that if $\sum q_i = 1$ we have $\sum q_i^2 \le \max_i q_i$. To find that maximum of $\frac{n!}{j!k!(n-j-k)!}$ we note that if the maximum occurs at k_0, j_0 we must have that

$$\frac{n!}{(j_0-1)!k_0!(n-j_0-k_0+1)!}, \quad \frac{n!}{(j_0+1)k_0!(n-j_0-k_0-1)!}$$

$$\frac{n!}{(j_0)!(k_0-1)!(n-j_0-k_0+1)!}, \quad \frac{n!}{j_0(k_0+1)!(n-j_0-k_0-1)!}$$

are all bounded by $\frac{n!}{(j_0)!k_0!(n-j_0-k_0)!}$. This gives 4 inequalities which reduce to

$$n-j_0-1 \le 2k_0 \le n-j_0+1 \quad \text{and} \quad n-k_0-1 \le 2j_0 \le n-k_0+1$$

which itself show that k_0 and j_0 are both of the order of $n/3$. So we find, using Stirling's formula

$$P^{2n}(0,0) \le \frac{1}{2^{2n}}\binom{2n}{n}\frac{1}{3^n}\frac{n!}{(n/3)!(n/3)!(n/3)!} \sim \frac{3\sqrt{3}}{2}\frac{1}{(\pi n)^{3/2}}$$

which shows that the random walk is transient in dimension 3.

Example 3.2.6 (recurence and transience for the succcess run chain). Continuing with Example 3.1.4, we consider the return time to state 0 $\tau^{(0)}$ and we compute explicitly its p.d.f since to return to 0 the only possible paths are $0 \to 0$, $0 \to 1 \to 0$, $0 \to 1 \to 2 \to 0$, and so on. Therefore we find that

$$P(\tau^{(0)} = k | X_0 = 0) = p_0 p_1 p_{k-2} q_{k-1} = p_0 p_1 p_{k-2}(1 - p_{k-1})$$

If we set $u_n \equiv p_0 p_1 \cdots p_{n-1}$ we have then

$$P(\tau^{(0)} = k | X_0 = 0) = u_{k-1} - u_k$$

and ttherefore

$$\sum_{k=1}^{n} P(\tau^{(0)} = k | X_0 = 0) = (1 - u_0) + (u_o - u_1) + \cdots + u_{n-1} - u_n = 1 - u_n$$

So

$$P(\tau^{(0)} < \infty | X_0 = 0) = \sum_{k=1}^{\infty} P(\tau^{(0)} = k | X_0 = 0) = 1 - \lim_{n\to\infty} u_n$$

and we obtain that

$$X_n \text{ is recurrent if and only if } \lim_{n\to\infty} u_n = \lim_{n\to\infty} p_0 \cdots p_{n-1} = 0$$

So for example the Markov chain is recurrent if $p_n = p$ is constant. More generally we have

Lemma 3.2.7 *We have*

$$\lim_{n\to\infty} u_n = \lim_{n\to\infty} \prod_{k=0}^{n-1} p_k = 0 \quad \textit{if and only if} \quad \sum_{k=0}^{\infty} q_k = \infty$$

$$\lim_{n\to\infty} u_n = \lim_{n\to\infty} \prod_{k=0}^{n-1} p_k > 0 \quad \textit{if and only if} \quad \sum_{k=0}^{\infty} q_k < \infty$$

Proof: We have $\prod_{k=0}^{n-1} p_k > 0$ if and only if $\infty > -\log(\prod_{k=0}^{n-1} p_k) = -\sum_{k=1}^{n} \log p_k = -\sum_{k=1}^{n} \log(1 - q_k)$. Taking $n \to \infty$ shows that $\lim_{n\to\infty} \prod_{k=0}^{n-1} p_k > 0$ if only if $\sum_{k=1}^{n} \log(1 - q_k)$ converges. For this to happen we must have $\lim_{k\to\infty} q_k = 0$. But since $\lim_{x\to 0} \log(1 - x)/x = 1$ by L'Hospital rule, we have that $\sum_{k=1}^{n} \log(1 - q_k)$ converges if and only if $\sum_{k=1}^{n} q_k$ converges. ∎

For the chain to be transient we must then have that q_k goes to 0 fast enough. If say $q_i = \frac{1}{i}$ then the chain is recurrent while if $q_i = \frac{1}{i^2}$ then it is transient.

We add one more method to establish transience. For this pick reference state j_0 and consider the quantity

$$\alpha(i) \;=\; P\left(X_n \text{ visits } j_0 \text{ for some } n \geq 0 | X_0 = i\right)$$

By definition we have

$$\alpha(j_0) = 1$$

since then $X_0 = j_0$. On the other hand if the chain is transient then we must have

$$\alpha(i) < 1 \text{ for } i \neq j_0\,.$$

(see Proposition 3.2.3) since for $i \neq j_0$ we have $\alpha(i) = P\left\{\tau^{(j_0)} < \infty | X_0 = i\right\}$.

Let us derive an equation for $\alpha(i)$ by conditioning of the fist step. We have for $i \neq j_0$

$$\begin{aligned}
\alpha(i) &= P\left(X_n \text{ visits } j_0 \text{ for some } n \geq 0 | X_0 = i\right) \\
&= \sum_{j \in S} P\left(X_n \text{ visits } j_0 \text{ for some } n \geq 0\,,\ X_1 = j | X_0 = i\right) \\
&= \sum_{j \in S} P\left(X_n \text{ visits } j_0 \text{ for some } n \geq 0 | X_1 = j\right) P(i,j) \\
&= \sum_{j \in S} P\left(X_n \text{ visits } j_0 \text{ for some } n \geq 1 | X_1 = j\right) P(i,j) \\
&= \sum_{j \in S} P(i,j)\alpha(j)
\end{aligned}$$

and thus $\alpha(i)$ satisfies the equation

$$P\alpha(i) = \alpha(i)\,, \quad i \neq j_0\,.$$

We have the folllowing criterion for transience

Theorem 3.2.8 *An irreducible Markov chain X_n is transient if and only if for some state j_0 there exists a solution for the equation*

$$P\alpha(i) = \alpha(i) \text{ for } i \neq j_0 \tag{3.1}$$

with the conditions

$$\alpha(j_0) = 1 \quad \text{and} \quad 0 \leq \alpha(i) < 1 \text{ for } i \neq j_0 \tag{3.2}$$

Proof: We have already established the necessity. In order to show the sufficiency we assume that we have found a solution for the equations (3.1) and (3.2). Then for $i \neq j_0$ we have, using repeatedly the equation $P\alpha(i) = \alpha(i)$

$$\begin{aligned}
1 > \alpha(i) &= P\alpha(i) \\
&= P(i,j_0)\alpha(j_0) + \sum_{j \neq j_0} P(i,j)\alpha(j) \\
&= P(i,j_0) + \sum_{j \neq j_0} P(i,j)P\alpha(j) \\
&= P(i,j_0) + \sum_{j \neq j_0} P(i,j)P(j,j_0) + \sum_{j \neq j_0, k \neq j_0} P(i,j)P(j,k)\alpha(k) \\
&= P(i,j_0) + \sum_{j \neq j_0} P(i,j)P(j,j_0) + \sum_{j,k \neq j_0} P(i,j)P(j,k)P(j,j_0) + \cdots \\
&= P(\tau^{(j_0)} = 1 | X_0 = i) + P(\tau^{(j_0)} = 2 | X_0 = i) + P(\tau^{(j_0)} = 3 | X_0 = i) + \cdots \\
&= P(\tau^{(j_0)} < \infty | X_0 = i)
\end{aligned}$$

which establishes transience. ∎

Example 3.2.9 (transience for the Random walk on $\{0,1,2,\cdots\}$) Continuing with Example 3.1.1 we use Theorem 3.2.8. We pick the reference state $j_0 = 0$ and for $j \neq 0$ solve the equation

$$P\alpha(j) = P(j,j-1)\alpha(j-1) + P(j,j+1)\alpha(j+1) = (1-p)\alpha(j-1) + p\alpha(j) = \alpha(j)$$

whose solution is

$$\alpha(j) = \begin{cases} C_1 + C_2\left(\frac{1-p}{p}\right)^j & \text{if } p \neq \frac{1}{2} \\ C_1 + C_2 j & \text{if } p = \frac{1}{2} \end{cases} .$$

Using that $\alpha(0) = 0$ we find

$$\alpha(j) = \begin{cases} (1-C_2) + C_2\left(\frac{1-p}{p}\right)^j & \text{if } p \neq \frac{1}{2} \\ (1-C_2) + C_2 j & \text{if } p = \frac{1}{2} \end{cases} .$$

and we see the condition $0 \leq \alpha(i) < 1$ is possible only if $(1-p)/p < 1$ (that is $p > 1/2$) and by choosing $C_2 = 1$. Thus we conclude

$$\text{The random walk on } \{0,1,2,\cdots\} \text{ is } \begin{cases} \text{transient for } p > \frac{1}{2} \\ \text{recurrent for } p \leq \frac{1}{2} \end{cases} .$$

3.3 Positive recurrent Markov chains

A finite state irreducible Markov chain is always recurrent and we have Kac's formula for the invariant measure $\pi(i) = E[\tau^{(i)}|X_0 = i]^{-1}$, that is the random variable $\tau^{(i)}$ has finite expectation. For a countable state space it is possible for a Markov chain to be recurrent that is $\tau^{(j)} < \infty$ with probability one but that $\tau^{(j)}$ does not have finite expectation. This motivates the following definition of ***positive recurrence***.

Definition 3.3.1 (Null recurrence-Positive recurrence)

1. *As state i is positive recurrent if $E[\tau^{(i)}|X_0 = i] < \infty$.*
2. *As state i is null recurrent if it is recurrent but not positive recurrent.*

We first investigate the relation between recurrence and existence of invariant measures. We first show that if one state j is positive recurrent then there exists a stationary distribution. The basic idea is to decompose any path of the Markov chain into successive visits to the state j. To build up our intuition if a stationary distribution

were to exists it should measure the amount of time spent in state i and to measure this we introduce the

$$\mu(i) = E[\sum_{n=0}^{\tau^{(j)}-1} \mathbf{1}_{\{X_n=i\}} | X_0 = j] \tag{3.3}$$

which measure counts the number of visits to i between two successive visits to j. Note that if j is positive recurrent then we have

$$\sum_i \mu(i) = \sum_i E[\sum_{n=0}^{\tau^{(j)}-1} \mathbf{1}_{\{X_n=i\}} | X_0 = j] \; = \; E[\tau^{(j)} \,|\, X_0 = j]$$

and thus

$$\pi(i) = \frac{\mu(i)}{E[\tau^{(i)} | X_0 = i]}$$

Proposition 3.3.2 *For a recurrent irreducible Markov chain X_n and a reference state j let*

$$\mu(i) = E[\sum_{n=0}^{\tau^{(j)}-1} \mathbf{1}_{\{X_n=i\}} | X_0 = j]$$

counts the expected number of visits to i between two successive visits to j. Then μ satisfies

$$\mu P = \mu$$

and if j is positive recurrent $\sum_{i \in S} \mu(i) < \infty$ so that μ can be normalized to a stationary distribution π.

Proof: Note first that we have

$$\mu(i) = E\left[\sum_{n=0}^{\tau^{(j)}-1} \mathbf{1}_{\{X_n=i\}} | X_0 = j\right] = \sum_{n=0}^{\infty} P(X_n = k, \tau^{(j)} > n | X_0 = j)$$

and since the chain visits j at time 0 and then only again at time $\tau^{(j)}$ we have $\mu(j) = 1$ and

$$\mu(i) = E\left[\sum_{n=1}^{\tau^{(j)}} \mathbf{1}_{\{X_n=i\}} | X_0 = j\right] = \sum_{n=1}^{\infty} P(X_n = i, \tau^{(j)} \geq n | X_0 = j)$$

We have then, by conditioning on the last step, and using $\mu(j) = 1$

$$\begin{aligned}
\mu(i) &= \sum_{n=1}^{\infty} P(X_n = i, \tau^{(j)} \geq n | X_0 = j) \\
&= P(j,i) + \sum_{n=2}^{\infty} P(X_n = i, \tau^{(j)} \geq n | X_0 = j) \\
&= P(j,i) + \sum_{k \in S, k \neq j} \sum_{n=2}^{\infty} P(X_n = i, X_{n-1} = k, \tau^{(j)} \geq n | X_0 = j) \\
&= P(j,i) + \sum_{k \in S, k \neq j} \sum_{n=2}^{\infty} P(k,i) P(X_{n-1} = k, \tau^{(j)} \geq n | X_0 = j) \\
&= P(j,i) + \sum_{k \in S, k \neq j} \sum_{n=2}^{\infty} P(k,i) P(X_{n-1} = k, \tau^{(j)} > n-1 | X_0 = j) \\
&= P(j,i) + \sum_{k \in S, k \neq j} \sum_{m=1}^{\infty} P(k,i) P(X_m = k, \tau^{(j)} > m | X_0 = j) \\
&= \mu(j) P(j,i) + \sum_{k \neq j} E\left[\sum_{m=1}^{\tau^{(j)}-1} \mathbf{1}_{\{X_m = k\}} | X_0 = j \right] P(k,j) \\
&= \mu(j) P(j,i) + \sum_{k \neq j} \mu(k) P(k,j) \; = \; \sum_{k} \mu(k) P(k,j) \qquad (3.4)
\end{aligned}$$

which proves the invariance of μ.

If the chain is positive recurrent, we have

$$\sum_{i \in S} \mu(i) \; = \; \sum_{i \in S} E\left[\sum_{n=0}^{\tau^{(j)}-1} \mathbf{1}_{\{X_n = i\}} | X_0 = j \right] \; = \; E\left[\tau^{(j)} | X_0 = j \right] < \infty$$

and thus $\mu(i)$ can be normalized to a stationary distribution. ∎

We next prove a converse statement.

Proposition 3.3.3 *Assume the irreducible Markov chain has a stationary distribution $\pi(i)$ then $\pi(i) > 0$ for any i and we have Kac's formula*

$$\pi(i) E\left[\tau^{(i)} | X_0 = i \right] \; = \; 1 \, .$$

Proof: Let us assume that π is invariant. First we show that the chain must be recurrent. If the chain were transient then we have $P^n(i,j) \to 0$ as $n \to \infty$ and so by dominated convergence

$$\pi(i) = \sum_{j} \pi(j) P^n(i,j) \to 0 \text{ as } n \to \infty \, .$$

which is impossible. The fact that $\pi(i) > 0$ is proved as in the case of finite state space.

We will also consider the time-reversed chain Y_n whose transition probabilities are $\widehat{P}(i,j) = \frac{\pi(j)P(j,i)}{\pi(i)}$ and which has π has stationary distribution (see exercise for details). By the same argument we see that the chain Y_n is recurrent.

Next we write

$$\pi(i)E[\tau^{(i)}|X_0 = i] \;=\; \pi(i)\sum_{n=1}^{\infty} P(\tau^{(i)} \geq n|X_0 = i)$$

The event $\{\tau^{(i)} \geq n\}$ conditioned on $\{X_0 = i\}$ correspond to a sequence of states $i_0 = i, i_1, \cdots, \cdots, i_{n-1}, i_n = j$ where $i_1, \cdots, i_{n-1}$ are not equal to i. The probability of such event is

$$\pi(i)P(i,i_1)\cdots P(i_{n-1},i_n) \;=\; \pi(j)\widehat{P}(j,i_{n-1})\cdots\widehat{P}(i_1,i)$$

using time-reversal. Summing over all such sequences we find that

$$\begin{aligned}
\pi(i)E[\tau^{(i)}|X_0 = i] &= \pi(i)\sum_{n=1}^{\infty} P(\tau^{(i)} \geq n|X_0 = i) \\
&= \sum_{j\in S}\pi(j)\sum_{n=1}^{\infty} P(\tau^{(i)} = n|Y_0 = j) \\
&= \sum_{j\in S}\pi(j)P(\tau^{(i)} < \infty|Y_0 = j) = 1\,. \qquad (3.5)
\end{aligned}$$

using the recurrence of the time reversed chain.

■

As a corollary of these two results we obtain the fact that positive recurrence is communication class property

Corollary 3.3.4 *Suppose j is positive recurrent and $j \leftrightarrow i$. Then i is positive recurrent.*

Proof: By considering every communication class separately we can assume without loss of generality that X_n is irreducible. (If the communication class is not closed then i can not be recurrent.)

Then by Proposition if j is positive recurrent then there exists a stationary distribution π and then by Proposition every state must be positive recurrent. ■

Using these results we now obtain

Theorem 3.3.5 (Ergodic theorem for countable state space Markov chains)

1. *A irreducible positive recurrent Markov chain has a unique stationary distribution* π *and for any initial distribution* ν *of* X_0 *we have*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \mathbf{1}_{\{X_k=j\}} \;=\; \pi(j)\,,$$

with probability 1. *In particular*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \nu P^k(j) \;=\; \pi(j)\,.$$

and π *satisfies the Kac's formula*

$$\pi(j) = \frac{1}{E[\tau^{(j)}|X_0 = j]}\,.$$

2. *Conversely if an irreducible Markov has a stationary distribution then it is positive recurrent.*

Proof: We have already proved most of it. Part 2. is is Proposition which also establishes Kac's formula. By Proposition positive recurrence implies the existence of a stationary distribution and thus repeating the proof of Theorem 2.3.10 we show that if $X_0 = i$ with i arbitrary we have

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \mathbf{1}_{\{X_k=j\}} \;=\; \frac{1}{E[\tau^{(j)}|X_0 = j]} = \pi(j)\,. \tag{3.6}$$

Note that the proof in Theorem 2.3.10 only use positive recurrence condition and not that fact the state space is fine.

Taking now expectation of (3.6) we find that for any $i \in S$ we have

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} P^k(i,j) \;=\; \pi(j)\,,$$

and summing of the initial condition i we find that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \mu P^k(j) \;=\; \pi(j)\,.$$

Finally to prove uniqueness we assume that there exists another stationary distribution $\widetilde{\pi}$. If we use $\widetilde{\pi}$ as an initial distribution we get, using that $\widetilde{\pi} = \widetilde{\pi}P$ that $\sum_{k=1}^{n} \widetilde{\pi}P^k(j) = n\widetilde{\pi}(j)$ and thus

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \widetilde{\pi}P^k(j) \;=\; \lim_{n\to\infty} \widetilde{\pi}(j) = \widetilde{\pi} = \pi(j)\,.$$

and this concludes the proof. ∎

We continue our theoretical consideration by proving that if the chain is aperiodic then the distribution of X_n converges to $\pi(j)$.

Theorem 3.3.6 *Suppose X_n is an irreducible positive recurrent aperiodic Markov chain. Then for any initial distribution we have*

$$\lim_{n\to\infty} \nu P^n(j) \;=\; \pi(j)\,.$$

Proof: We will use a ***coupling argument***: we take two independent copies X_n and Y_n where X_n is starting in the initial distribution $X_0 = j$ while Y_n is starting in the stationary distribution π.

The idea is to consider the coupling time

$$\sigma = \inf\{n \geq 1; X_n = Y_n\}\,.$$

By the Markov property at the (random) time σ X_n and Y_n are in the same state. After the time τ X_n and Y_n must have the same distribution by the Markov property (independence on the past). But since Y_n is distributed according to π so must X_n be and thus at the coupling time X_n has reached its stationary distribution.

Let us now consider the chain $Z_n = (X_n, Y_n)$ which has transition probabilities

$$P(Z_{n+1} = (k,l)\,|\,Z_n = (i,j)) = P(i,k)P(j,l)$$

and stationary distribution $\pi(i,j) = \pi(i)\pi(j)$. Since X_n and Y_n are aperiodic, by Proposition 2.4.3, we can find n_0 such that for every $n \geq n_0$ we have $P^n(i,k) > 0$ and $P^n(j,l) > 0$. This implies that Z_n is irreducible and thus, since a staionary measure exists, by Theorem 3.3.5 the chain Z_n is (positive recurrent). This implies that $P(\sigma < \infty) = 1$ and thus $P(\sigma > n) \to 0$.

We have then

$$\begin{aligned}
|P^n(i,j) - \pi(j)| &= |P(X_n = j) - P(Y_n = j)| \\
&\leq |P(X_n = j, \tau \leq n) - P(Y_n = j, \tau \leq n)| \\
&\quad + |P(X_n = j, \tau > n) - P(Y_n = j, \tau > n)| \\
&= |P(X_n = j, \tau > n) - P(Y_n = j, \tau > n)| \\
&= \left|E[(\mathbf{1}_{\{X_n = j\}} - \mathbf{1}_{\{X_n = j\}}), \mathbf{1}_{\{\tau > n\}}]\right| \\
&\leq E[\mathbf{1}_{\{\tau > n\}}] \\
&= P(\tau \geq n) \to 0, \text{ as } n \to \infty
\end{aligned} \tag{3.7}$$

and this prove the convergence.

Let us conclude with a few examples

Example 3.3.7 (Positive recurrence for the random walk on $\{0,1,2,\cdots\}$) Continuing with Example 3.1.1 we establish positive recurrence by computing the stationary distribution. We can use detailed balance and obtain

$$\pi(j)p = \pi(j+1)(1-p)$$

which we can solve

$$\pi(n) = \left(\frac{p}{1-p}\right)^n \pi(0)\,.$$

This is normalizable only if $p/1-p<1$ or $p<1/2$ in which case pi is a geometric random variable. Thus together with the results in Example 3.1.1 we obtain

$$\text{The random walk on } \{0,1,2,\cdots\} \text{ is } \begin{cases} \text{transient for } p>\frac{1}{2} \\ \text{null recurrent for } p=\frac{1}{2} \\ \text{positive recurrent for } p<\frac{1}{2} \end{cases}.$$

Example 3.3.8 (Positive recurrence for the success run chain) Continuing with Examples 3.1.4 and 3.2.6 we determine recurrence by solving $\pi P = \pi$. This gives the equations

$$\begin{aligned} \pi(0) &= \pi(0)q_0 + \pi(1)q_1 + \cdots = \sum_{n=0}^{\infty} \pi(n)q_n \\ \pi(1) &= \pi(0)p_0 \\ \pi(2) &= \pi(1)p_1 \\ \vdots &\quad \vdots \end{aligned}$$

We find that

$$\pi(n) = \pi(0)p_0p_1\cdots p_{n-1}$$

and inserting into the first equation

$$\pi(0) = \pi(0)\left[(1-p_0) + p_0(1-p_1) + p_0p_1(1-p_2) + \cdots\right] = \pi(0)\left[1 + \lim_{n\to\infty} p_0p_1\cdots p_{n-1}\right]$$

Recall that we have recurrence, by Example 3.2.6 , provided $\lim_{n\to\infty}\prod_{j=0}^{n-1} p_j = 0$ and so if X_n is recurrent there exists a solution of $\pi P = \pi$ and we can normalize π provided

$$\sum_{n=1}^{\infty}\prod_{j=0}^{n-1} p_j < \infty$$

Therefore we obtain

$$\text{The success run chain is } \begin{cases} \text{transient if } \lim_n \prod_{j=0}^{n-1} p_j > 0 \\ \text{recurrent if } \lim_n \prod_{j=0}^{n-1} p_j = 0 \\ \text{positive recurrent if } \sum_n \prod_{j=0}^{n-1} p_j < \infty \end{cases}.$$

3.4 Branching processes

Before we discuss branching processes we recall a number of facts about moment generating functions. The moment generating function of a random variable X is given by $M_X(t) = E[e^{tX}]$ and if $M_X(t)$ is finite in a neighborhood of 0 we have that $M_X^{(n)}(0) = E[X^n]$. In this section it will be more convenient to use the variable $s = e^t$ and so we have $s \geq 0$. We have the following easy proposition

Proposition 3.4.1 *Let X be a random variable taking values in $\{0, 1, 2, 3, \cdots\}$ and let*

$$\phi(s) = E[s^X] = \sum_{k=0}^{\infty} s^k P\{X = k\}$$

. We have

1. *$\phi(1) = 1$ and $\phi(0) = P\{X = 0\}$.*
2. *$\phi'(s) = \sum_{k=1}^{\infty} k s^{k-1} P\{X = k\}$ and thus $\phi'(1) = E[X]$. Unless $X = 0$, $\phi(s)$ is strictly increasing.*
3. *$\phi''(s) = \sum_{k=2}^{\infty} k(k-1) s^{k-1} P\{X = k\}$. If $P\{X \geq 2\} > 0$, $\phi(s)$ is strictly convex.*

Moment generating function for the sum of independent random variable $X_1, \cdots, X_n$ satisfies

$$\phi_{X_1+\cdots+X_n} = E[s^{X_1+\cdots+X_n}] = \phi_{X_1}(s) \cdots \phi_{X_n}(s)$$

and for the branching process we will to consider the ***compound sum***

$$S_N = X_1 + \cdots + X_N$$

where N is itself a random variable taking value in $\{0, 1, 2, 3, \cdots\}$ (we set $S_0 = 0$).

Proposition 3.4.2 *Let N be a random variable taking value in $\{0, 1, 2, 3, \cdots\}$. Let $X_1, X_2, X_3, \cdots$ be IID copies of the random variable X with moment generating function $\phi_X(s)$. Then the compound random variable*

$$S_N = \sum_{k=1}^{N} X_k$$

has moment generating function

$$\phi_{S_N} = \phi_N(\phi_X(s)) \,.$$

In particular we have Wald's formula

$$E[S_N] = E[N]E[X] \,.$$

Proof: By conditioning

$$\begin{aligned}\phi_{S_N}(s) &= E[s^{S_N}] = E[s^{X_1+\cdots X_N}] \\ &= E[E[s^{S_N}\,|N] \\ &= \sum_{n=0}^{\infty} E[s^{X_1+\cdots X_n}]P\{N=n\} \\ &= \sum_{n=0}^{\infty} (\phi_X(s))^n P\{N=n\} \\ &= \phi_N(\phi_X(s))\,.\end{aligned}$$

The differentiating we find

$$\phi'_{S_N}(s) = \phi'_N(\phi_X(s))\phi'(s)$$

and thus by Proposition 3.4.1 we have

$$E[S_N] = \phi'_{S_N}(1) = \phi'_N(\phi_X(1))\phi'(1) = \phi'_N(1)\phi'(1) = E[N]E[X]\,.$$

Branching process