Stochastic Processes
and
Monte-Carlo Methods

University of Massachusetts: Spring 2010

Luc Rey-Bellet

Contents

Chapter 1

Random variables and 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 is 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

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

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

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

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

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

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

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

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

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

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*** $sd(X) = \sqrt{\text{var}(X)}$.

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

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

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

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

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

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 $\text{var}(X) = M_X''(0) - (M_X'(0))^2$. 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 $X = U_{[a,b]}$ to denote this 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 $X = N_{\mu,\sigma^2}$ to denote this random variable. The ***standard normal random variable*** is the normal random variable with $\mu = 0$ and $\sigma = 1$, i.e., $N_{0,1}$

The normal random variable has the following property

$$X = N_{0,1} \quad \text{if and only if} \quad \sigma X + \mu = N_{\mu,\sigma^2}$$

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 we consider first $X = N_{0,1}$. 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 $X = Exp_\lambda$ 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 n and λ be positive numbers. The ***gamma random variable with parameters*** n ***and*** λ is the continuous random variable with p.d.f

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

The moment generating function is

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

and the mean and variance are

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

We write $X = \Gamma_{n,\lambda}$ to denote this random variable.

To show the formula for the m.g.f note that for any $\alpha > 0$

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

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

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

Note that $\Gamma_{1,\lambda} = Exp_\lambda$. Also the m.g.f of $\Gamma_{n,\lambda}$ is the m.g.f of Exp_λ to the n^{th} power. Using Theorem 1.1.1 and Proposition 1.1.2 we conclude that if $X_1, \cdots, X_n$ are n independent exponential random variables with parameters λ then $X_1 + \cdots + X_n = \Gamma_{n,\lambda}$.

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

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

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

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

We write $X = B_n, p$ to denote this random variable.

The formula for the m.g.f can be obtained directly using the binomial theorem, or simply by noting that by construction $B_{n,p}$ is a sum of n independent Bernoulli 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 $0 \le p \le 1$ 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 $X = Geo_p$ 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

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

We write $X = Poiss_\lambda$ to denote this random variable.

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 in that way 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

$$X_{n+1} = (aX_n + c) \operatorname{mod}(m)\,.$$

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*

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

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

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

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 $U = U([0,1])$.*
- **Step 2** *Set $X = F^{-1}(U)$.*

Example 1.3.4 (Simulating an exponential random variable) If $X = Exp_\lambda$ 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 $U = U([0,1])$ then

$$Exp_\lambda = -\frac{1}{\lambda}\log(1-U) = -\frac{1}{\lambda}\log(U)\,.$$

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*

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

To simulate X

- **Step 1** *Simulate Y with density g.*
- **Step 2** *Simulate a random number U.*
- **Step 3** *If*
$$U \le \frac{f(Y)}{g(Y)C}$$
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 $CP\left(U \leq \frac{f(Y)}{Cg(Y)}\right) = 1$ and thus

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

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

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

independently of the other iterations. Therefore the number of iterations needed is $Geom(\frac{1}{C})$ 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

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

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 $X = N_{\mu,\sigma^2}$ it is enough to simulate N_{μ,σ^2} and then set $X = \sigma N_{0,1} + \mu$.

Let us first consider the random variable Z whose density is

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

One can think of Z as the absolute value of $N(0,1)$.

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

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

One generates Z using the rejection method. To generate $X = N_{0,1}$ from Z one generate a discrete random variable S with takes value $+1$ and -1 with probability $\frac{1}{2}$ and then set $X = SZ$. The random variable S is $S = 2B_{1,\frac{1}{2}} - 1$.

- **Step 1** Generate a random numbers U, an exponential random variable Y and a Bernoulli random variable B.
- **Step 2** If $U \le \exp -\frac{(Y-1)^2}{2}$ set $Z = Y$ and $X = (2B-1)Z$

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

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

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 $Y = Exp_\lambda$ then

$$\lceil Y \rceil = Geom_p \qquad \text{with } p = 1 - e^{-\lambda}\,.$$

Indeed we have

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

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 $X = Geom_p$.

Example 1.3.11 (Simulating the Gamma random variable) Using the fact that $Gamma(n, \lambda)$ is a sum of n independent $Exp(\lambda)$ one immediately obtain

Algorithm 1.3.12 (Gamma random variable)

- **Step 1** *Generate n random number $U_1, \cdots, U_n$.*
- **Step 2** *Set $X_i = -\frac{1}{\lambda}\log(U_i)$*
- **Step 3** *Set $X = X_1 + \cdots + X_n$.*

Then $X = \Gamma_{n,p}$.

Finally we give an elegant algorithm which generates 2 independent normal random variables.

Example 1.3.13 (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 $S = Exp_{1/2}$ and $\Theta = U_{[0,2\pi]}$.

Therefore we obtain

Algorithm 1.3.14 (Standard normal random variable)

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

$$\begin{aligned} X &= \sqrt{-2\log(U_1)}\cos(2\pi U_2) \\ Y &= \sqrt{-2\log(U_1)}\sin(2\pi U_2) \qquad (1.4) \\ & \qquad (1.5) \end{aligned}$$

Then X and Y are 2 independent $N_{0,1}$.

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

$$I_a \leq \frac{X}{a} \qquad (1.6)$$

and that since I_a is a binomial random variable

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

Taking expectations in the inequality (1.6) gives

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

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

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

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*

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

Proof: Observe first that

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

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

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

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

$$\begin{aligned} P\left(S_n \ge \frac{3n}{4}\right) &= P\left(S_n - \frac{n}{2} \ge \frac{n}{4}\right) \\ &\le P\left(\left|S_n - \frac{n}{2}\right| \ge \frac{n}{4}\right) \\ &\le \frac{n/4}{(n/4)^2} = \frac{4}{n}. \end{aligned} \tag{1.7}$$

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*

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

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

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

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

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

Since $t > 0$ is arbitrary we obtain

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

Similarly for $t < 0$ we have

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

and thus

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

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 $P(S_n \ge 3n/4)$ we apply Chernov bound with $t > 0$ and obtain

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

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

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

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 Large deviation bounds

In this section we discuss a quantitative refinement of the weak law of large numbers. A look at the proof shows that the probability to observe a deviation from the mean is bounded by a quantity of order $1/n$ and that we have used only the fact that the variance is finite. One would expect that if we know that higher moments $E[X^n]$ are finite then sharper estimates will hold.

For this we need some preparation. Let X be a random variable with m.g.f $M_X(t) = E[e^{tX}]$. It will be useful to consider the logarithm of $M(t)$ which we denote by u

$$u(t) = u_X(t) = \log M_X(t) = \log E[e^{tX}]$$

and $u_X(t)$ to as the ***logarithmic moment generating function*** of the random variable X.

Recall that a function $f(t)$ is called ***convex*** if for any $0 \le \alpha \le 1$ we have

$$f(\alpha t_1 + (1-\alpha)t_2) \;\le\; \alpha f(t_1) + (1-\alpha) f(t_2)\,.$$

Graphically it means that for the graph $t_1 \le t \le t_2$ the graph of $f(t)$ lies below the line passing through the points $(t_1, f(t_1))$ and $(t_2, f(t_2))$. From calculus we known that f is convex iff $f'(t)$ is increasing iff $f''(t)$ is nonnegative (provided the derivatives do exist).

Lemma 1.6.1 *The logarithmic moment generating function $u(t) = \log M(t)$ is a convex function which satisfies*

$$u(0) = 0\,, \quad u'(0) = \mu\,, \quad u''(0) = \sigma^2$$

Proof: We will prove the convexity in two different ways. The first proof use Hölder inequality which states that if $1/p + 1/q = 1$ then $E[XY] \le E[X^p]^{1/p} E[Y^q]^{1/q}$. We choose $p = \frac{1}{\alpha}$ and $q = \frac{1}{1-\alpha}$ and obtain

$$E\left[e^{(\alpha t_1 + (1-\alpha)t_2)X}\right] \;=\; E\left[\left(e^{t_1 X}\right)^{\alpha}\left(e^{t_2 X}\right)^{(1-\alpha)}\right] \;\le\; E\left[e^{t_1 X}\right]^{\alpha} E\left[e^{t_2 X}\right]^{(1-\alpha)}\,.$$

Taking logarithms proves the convexity.

For our second proof note that

$$\begin{aligned} u'(t) &= \frac{M'(t)}{M(t)} \\ u''(t) &= \frac{M''(t)M(t) - M'(t)^2}{M(t)^2}\,. \end{aligned}$$

If $t = 0$ we find that

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

To prove the convexity of u we need to show that $u''(t) \ge 0$ for any t. For given t let us define the random variable Y_t to be the random variable with p.d.f

$$\frac{f_X(x)e^{tx}}{M(t)}\,,$$

if X is a continuous random variable (argue similarly if X is a discrete R.V.). Then one verifies that

$$E[Y_t] = \frac{M'(t)}{M(t)} = u'(t) \quad \mathrm{var}(Y_t) \;=\; \frac{M''(t)}{M(t)} - \left(\frac{M'(t)}{M(t)}\right)^2 \;=\; u''(t)\,.$$

Since the variance is nonnegative this proves that u is convex. ∎

Given a function $f(t)$ we define the ***Legendre transformation*** of $f(t)$ to be a new function $f^*(z)$ given by

Definition 1.6.2 *The Legendre transform of a function $f(t)$ is the function $f^*(z)$ defined by*

$$f^*(z) \;=\; \sup_t \left(zt - f(t)\right)\,. \tag{1.8}$$

Note that the supremum in Eq. (1.8) can be equal to $+\infty$. If the supremum is finite and f is differentiable then we can compute f^* using calculus, the supremum is attained at the point t^* such that the derivative of $zt - f(t)$ vanishes, i.e., at the point t^*

$$z = f'(t^*)\,.$$

Then solving for $t^*(z)$ and inserting in the l.h.s. of (1.8) gives

$$f^*(z) = zt^*(z) - f'(t^*(z))\,.$$

For future use let us compute the Legendre transform of some logarithmic moment generating functions.

Example 1.6.3 Let $M(t) = e^{\mu t + \sigma^2 t^2/2}$ be the m.g.f of $N_{0,1}$ and let $u(t) = \log M(t) = \mu t + \sigma^2 t^2/2$. Given z the maximum of $zt - \mu t - \sigma^2 t^2/2$ is attained if t^* satisfies

$$z - \mu - \sigma^2 t^* = 0\,, \quad t^* = \frac{z-\mu}{\sigma^2}$$

and thus

$$u^*(z) \;=\; zt^* - \mu t^* - \sigma^2 (t^*)^2/2 \;=\; \frac{(z-\mu)^2}{2\sigma^2}\,.$$

We see that $u^*(z)$ is a parabola centered around μ.

Example 1.6.4 Let $M(t) = (1-p) + pe^t$ be the m.g.f of $B_{1,p}$ and let $u(t) = \log M(t) = \log((1-p) + pe^t)$. We distinguish three cases

- If $z > 1$ then the function $zt - \log((1-p) + pe^t)$ is increasing since its derivative is $z - \frac{pe^t}{(1-p)+pe^t} > 0$ for all t. The maximum is attained as $t \to \infty$ and is equal to $+\infty$ since

$$\lim_{t\to\infty} zt - \log((1-p) + pe^t) \;=\; \lim_{t\to\infty} z(t-1) - \log((1-p)e^{-t} + p) \;=\; +\infty\,.$$

- If $z < 0$ then the function $zt - \log((1-p) + pe^t)$ is decreasing for all t and thus the supremum is attained as $t \to -\infty$. The supremum is $+\infty$.
- For $0 \le z \le 1$ the the maximum is attained if t^* satisfies

  $$z = \frac{pe^t}{(1-p) + pe^t}, \quad t^* = \log\left(\frac{z}{1-z}\frac{1-p}{p}\right),$$

  and we obtain

  $$u^*(z) \;=\; z\log\left(\frac{z}{p}\right) + (1-z)\log\left(\frac{1-z}{1-p}\right).$$

A simple computation shows that $u^*(z)$ is strictly convex and that $u^*(z)$ has its minimum at $z = p$.

Lemma 1.6.5 *Let $u(t)$ be the logarithmic moment generating function of the random variable X. Then the Legendre transform $u^*(z)$ of $u(t)$ is a convex function which satisfies $u(z) \ge 0$. If $\sigma^2 > 0$ then $u(z) = 0$ iff $z = \mu$, i.e. $u^*(z)$ is nonnegative and takes its unique minimum (which is equal to 0) at the mean μ of X.*

Moreover if $z > \mu$ then

$$u^*(z) = \sup_{t\ge 0}(tz - u(t)).$$

and if $z < \mu$ then

$$u^*(z) = \sup_{t\le 0}(tz - u(t)).$$

Proof:
1) The convexity of $u^*(z)$ follows from

$$\begin{aligned}\alpha u^*(z_1) + (1-\alpha)u^*(z_2) &= \sup_t(\alpha z_1 t - \alpha u(t)) + \sup_t((1-\alpha)z_2 t - (1-\alpha)u(t)) \\ &\ge \sup_t((\alpha z_1 + (1-\alpha)z_2))t - u(t)) \\ &= u^*(\alpha z_1 + (1-\alpha)z_2). \end{aligned} \tag{1.9}$$

2) Next note that $u^*(z) \ge 0z - u(0) = 0$ and thus $u^*(z)$ is nonnegative.
3) Suppose that $u^*(z_0) = 0$ for some z_0. Then $\sup_t(tz_0 - u(t)) = 0$. The supremum is attained at t^* which satisfies the equation $z_0 = u'(t^*)$ and thus we must have

$$0 \;=\; u^*(z_0) \;=\; t^*u'(t^*) - u(t^*). \tag{1.10}$$

This equation has one solution, namely take $t^* = 0$ since $u(0) = 0$. In that case $z_0 = u'(0) = \mu$. Let us show that this is the unique solution. The function $f(t) \equiv u(t) - tu'(t)$ satisfies

$$f(0) = 0, \quad f'(t) = -tu''(t).$$

If $u''(0) = \sigma^2 > 0$ then by continuity there exists $\delta > 0$ such that $u''(t) > 0$ for $t \in (-\delta, \delta)$. Thus $f'(t) > 0$ for $t \in (0, \delta)$ and $f'(t) < 0$ for $t \in (-\delta, 0)$. Therefore 0 is the only solution of $f(0) = 0$.
4) If $z > \mu$ then for $t < 0$ we have

$$zt - u(t) \;\leq\; \mu t - u(t) \;\leq\; \sup_t(\mu t - u(t)) \;=\; u^*(\mu) = 0\,.$$

Since $u^*(z) > 0$ we conclude that the supremum is attained for some $t \geq 0$. One argues similarly for $z < \mu$. ■

Theorem 1.6.6 (One-half of Cramer's theorem) *Let $X_1, X_2, \cdots$ be a sequence of independent and identically distributed random variables. Assume that the moment generating function $M(t)$ of X_i exists and is finite in a neighborhood of 0. Let $u(t) = \log \phi(t)$ and $u^*(z) = \sup_z(zt - u(t))$. Then for any $a > \mu$ we have*

$$P\left(\frac{S_n}{n} > a\right) \;\leq\; e^{-nu^*(a)}$$

and for any $a < \mu$ we have

$$P\left(\frac{S_n}{n} < a\right) \;\leq\; e^{-nu^*(a)}\,.$$

Proof: We use Chernov bounds. Let $a > \mu$, for $t > 0$ we have

$$\begin{aligned}
P\left(\frac{S_n}{n} \geq a\right) &= P\left(S_n \geq an\right) \\
&= \inf_{t \geq 0} e^{-ant} E\left[e^{tS_n}\right] \\
&= \inf_{t \geq 0} e^{-ant} M(t)^n \\
&= \inf_{t \geq 0} e^{-n(at - u(t)} \\
&= e^{-n \sup_{t \geq 0}(at - u(t)} \\
&= e^{-n \sup_t (at - u(t)} \;=\; e^{-nu^*(a)}\,.
\end{aligned}$$

One proceeds similarly for $a < 0$. ■

1.7 Monte-Carlo algorithm

The basic Monte-Carlo method uses sums of independent random variables and the law of large numbers to estimate a deterministic quantity. In order to illustrate the method let us start by an example.

Example 1.7.1 (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 $P(X = 1) = \pi/4$. We repeat this experiment n times. That is we we select n points inside the square independently. The number of points S_n within the circle can be written as $S_n = X_1 + \cdots + X_n$ where the X_i's are independent copies of X. So $S_n = B_{n,\frac{\pi}{4}}$ and $E[S_n] = n\pi/4$. By the Law of Large Numbers we can expect that S_n gives, for large enough n, a good approximation of $\pi/4$.

To estimate how good this approximation is, we will use the central limit theorem. Suppose, for example, that we perform $n = 10'000$ trials and observe $S_n = 7932$, then our estimator for π is $4\frac{7932}{10000} = 3.1728$. By the Central Limit Theorem, $\frac{S_n - n\mu}{\sqrt{n}\sigma}$ has for sufficiently large n a distribution which is close to a normal distribution $N(0,1)$. Therefore we will have

$$P\left(\frac{S_n}{n} - 1.96\frac{\sigma}{\sqrt{n}} \le \mu \le \frac{S_n}{n} + 1.96\frac{\sigma}{\sqrt{n}}\right) \cong 0.95\,,$$

for sufficiently large n. The value $x = 1.96$ is such that $P(|N(0,1)| \le x) = 0.95$. For this reason we call the interval $\left[\frac{S_n}{n} - 1.96\frac{\sigma}{\sqrt{n}}\,,\,\frac{S_n}{n} + 1.96\frac{\sigma}{\sqrt{n}}\right]$ a 95% confidence interval.

In our case a 95% confidence interval for $\pi/4$ is

$$\left[\frac{S_n}{n} - 1.96\frac{\sigma}{\sqrt{n}}\,,\,\frac{S_n}{n} + 1.96\frac{\sigma}{\sqrt{n}}\right]\,.$$

where $\sigma = \sqrt{\frac{\pi}{4}(1 - \frac{\pi}{4})}$ which we can't really evaluate since we do not know π. There are several ways to proceed

1. Use the simple bound $x(1-x) \le \frac{1}{4}$ so $\sigma \le \frac{1}{2}$ and thus

   $$1.96\frac{\sigma}{\sqrt{n}} \le 1.96\frac{1}{2\sqrt{n}} = 0.0098\,.$$

   This gives the interval $[3.1336, 3.2120]$ for a conservative 95% confidence interval.

2. We can simply use our estimate for π into the formula $\sigma = \sqrt{\frac{\pi}{4}(1 - \frac{\pi}{4})} \cong 0.405$. This gives a confidence interval of $[3.1410, 3.2046]$.

3. Another way to estimate the variance when σ^2 is unknown is to use the ***sample variance*** given by

$$V_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \frac{S_n}{n}\right)^2 .$$

The sample is an unbiased estimator since we have $E[V_n^2] = \sigma^2$. To see this note we can assume that $\mu = 0$ and then we have

$$E[V_n^2] = \frac{1}{n-1}\sum_{i=1}^{n} E\left[X_i^2 - 2X_i\frac{S_n}{n} + \left(\frac{S_n}{n}\right)^2\right] = \frac{n}{n-1}\sigma^2(1-\frac{2}{n}+\frac{1}{n^2}n) = \sigma^2 .$$

∎

This example is a particular case of the ***hit-or-miss method***. 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.

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 Example 1.7.3 below).

Many many other problems can be put in the form (maybe after some considerable work)

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

and the random variable X could also, of course be a discrete random variable.

Algorithm 1.7.2 (Monte-Carlo simple sampling) *Let X be a random variable. To estimate*

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

the simple sampling *consists in generating n i.i.d. random variables $X_1, X_2, \cdots, X_n$ and set*

$$I_n \equiv \frac{1}{n}\sum_{i=1}^{n} h(X_i) .$$

The quantity I_n gives an unbiased estimator of I, i.e., $E\left[I_n\right] = I$. By the strong law of large numbers I_n converges to I with probability 1 *as $n \to \infty$. Furthermore the variance of the simple sampling estimate is*

$$\mathrm{var}(I_n) \;=\; \frac{\mathrm{var}(h(X))}{n} .$$

Note that the variance $\mathrm{var}(I_n)$ can be used to determine the accuracy of our estimate, for example by determining a 95% confidence interval as in Example 1.7.1. If we denote $\sigma^2 = \mathrm{var}(h(X))$ then the half length of 95% confidence interval is given by $1.96\sigma/\sqrt{n}$. If we wish our confidence to half length ϵ we need to choose n such that

$$n \geq \frac{\epsilon^2}{(1.96)^2\sigma^2} .$$

So, as a rule we have

$$\textbf{accuracy of the Monte} - \textbf{Carlo method is of order } \sqrt{\mathrm{n}} .$$

which means that to imporve the accuracy of out estimate by a factor 10 we should perform 100 times more estimates. This is not very good, and the Monte-Carlo method cannot compete with numerical integration for a simple integral. However it can become competitive if the integral is over a space of very large dimension since the accuracy is dimension-independent.

We consider next another example which illustrate one technique through which one can reduce the variance considerably (***variance reduction***). The technique we will use goes under the name of ***importance sampling***. Suppose we want to compute $E[h(X)]$. We can use simple sampling by simulating i.i.d random variables with p.d.f. $f(x)$. Instead of using X we can choose another random variable Y with p.d.f $g(x)$ and write

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

We then simulate i.i.d random variables Y_i with p.d.f g and this gives a new estimator

$$J_n = \frac{1}{n}\sum_{j=1}^{n}\frac{h(Y_j)f(Y_j)}{g(Y_j)} ,$$

The variance is given by

$$\mathrm{var}(J_n) = \frac{1}{n}\mathrm{var}\left(\frac{h(Y)f(Y)}{g(Y)}\right) = \frac{1}{n}\left(\int \frac{h(x)^2f(x)^2}{g(x)}\,dx - \left(\int h(x)f(x)dx\right)^2\right) .$$

The idea of importance sampling is to choose Y such that

$$\mathrm{var}\left(\frac{h(Y)f(Y)}{g(Y)}\right) < \mathrm{var}(h(X)) ,$$

and thus to improve the efficiency of our method.

There are many other methods to reduce the variance and some are touched upon in the exercises. We illustrate the power of the importance sampling by considering a example.

Example 1.7.3 (Network reliability) Let us consider an application of simple sampling to nework reliability. Consider a connected graph as in Figure 1.1. 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 s and t and we want to compute the disconnection probability

$$p_D \equiv P\,(s \text{ is not connected to } t \text{ by working edges})$$

This can be computed by hand for very small graphs but even for the graph shown in Figure 1.1 this is hardly doable. Our graph here has 22 edges and let $\mathcal{E} = \{e_1 \cdots e_{22}\}$

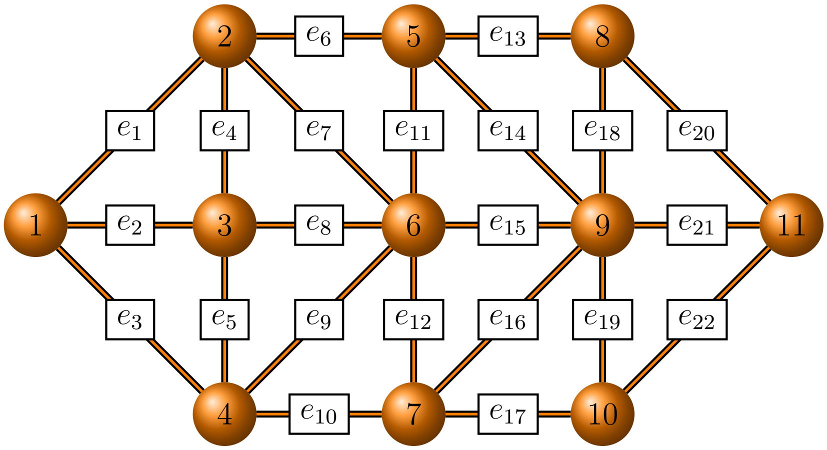

Figure 1.1: A graph with 11 vertices and 22 edges

denote the set of all edges. Let X denote the set of edges that fail, so X is a random subset of $\mathcal{E}$. So for every $B \subset \mathcal{E}$ we have

$$P(X = B) \ = \ q^{|B|}(1-q)^{|\mathcal{E}|-|B|}$$

where $|A|$ denotes the cardinality of A. If we denote by $\mathcal{S}$ the set of all subsets of $\mathcal{E}$ then X is a random variable which takes value in $\mathcal{S}$.

Let us define the function $k \, : \, \mathcal{S} \to \mathbf{R}$ by

$$k(B) \ = \ \begin{cases} 1 & \text{if } s \text{ is not connected to } t \text{ when the edges of } B \text{ fail} \\ 0 & \text{if } s \text{ is connected to } t \text{ when the edges of } B \text{ fail} \end{cases}$$

Then we have

$$p_D \ = \sum_{B\,;\,k(B)=1} P(X = B) \ = \ \sum_B k(B)P(X = B) \ = \ E[k(X)]\,.$$

The simple sampling estimator for p_D is

$$\frac{1}{n}\sum_{i=1}^{n} k(X_i)$$

where $X_1, \cdots X_n$ are i.i.d copies of X. Each 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 s and t.

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

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

since $|X| = B(22, q)$.

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

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

Therefore p_D is between 10^{-2} 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

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

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

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

A natural try is take the random variable Y to have a distribution $\phi(D) = P(y = B) = \theta^B (1-\theta)^{22-|B|}$ with a well chosen θ. Since $k(Y) = 0$ whenever $|Y| > 3$ we can for example choose θ such that $E[|Y|] = 3$. Since $|Y| = B(22, \theta)$ this gives $E[|Y|] = 22\theta$ and thus $\theta = 3/22$.

The estimator is now

$$J_n = \frac{1}{n} \sum_{i=1}^{n} \frac{k(Y_i) p(Y_i)}{\phi(Y_i)}$$

where Y_j are i.i.d with distribution $\phi(Y)$. Let us compute the variance of J_n. We have

$$\begin{aligned} \mathrm{var}(J_n) &= \frac{1}{n} \left(\sum_{B} \frac{k(B)^2 p(B)^2}{\phi(B)^2} \phi(B) - p_D^2 \right) \\ &= \frac{1}{n} \left(\sum_{B\,:\,k(B)=1} \frac{p(B)}{\phi(B)} p(B) - p_D^2 \right)\,. \end{aligned} \tag{1.11}$$

Note that

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

In Eq. (1.11) all terms with $k(B) = 1$ have $|B| \geq 3$. For those B we have

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

So we get

$$\mathrm{var}(J_n) \leq \frac{1}{n} \sum_{B \,:\, k(B)=1} 0.0053 \; p(B) = \frac{0.0053 \, p_D}{n}$$

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 $J_{n/200}$. This is pretty good!

Chapter 2

Finite Markov Chains

2.1 Introduction

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

$$\{X_n\} = X_0, X_1, X_2, \cdots ,$$

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

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

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***

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

and we have the relation

$$\begin{aligned}&P\left\{X_0 = i_0, X_1 = i_1, \cdots, X_n = i_n\right\} = \\ &P\left\{X_0 = i_0\right\} P\left\{X_1 = i_1 | X_0 = i_0\right\} \cdots P\left\{X_n = i_n | X_{n-1} = i_{n-1}, \cdots, X_0 = i_0\right\} .\end{aligned}$$

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*

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

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

The conditional probabilities $P\left\{X_n = i_n | X_{n-1} = i_{n-1}\right\}$ 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

$$\begin{aligned}&P\left\{X_0 = i_0, X_1 = i_1, \cdots, X_n = i_n\right\} = \\ &P\left\{X_0 = i_0\right\} P\left\{X_1 = i_1 | X_0 = i_0\right\} \cdots P\left\{X_n = i_n | X_{n-1} = i_{n-1}\right\}\end{aligned}$$

The transition probabilities $P\left\{X_n = j | X_{n-1} = i\right\}$ 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

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

or

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

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 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

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

(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.

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 $\mu_n(2) = 1 - \mu_n(1)$

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

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 $\mu(i) = P\{X_n = i\}$ for all n. The X_n satisfy the Markov property since all the X_n are independent

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

The stationary and limiting distribution iare μ and the transition matrix is

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

■

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 walker 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 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 holds and a maximum of N caller can be in the system. If exactly N people are in the system then a caller will be bounced back. We assume that during each time interval one new caller calls with probability p and 0 caller calls with probability $1-p$. Also during each time interval 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 $1 \le j \le N-1$

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

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 tim 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 your i^{th} 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

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

∎

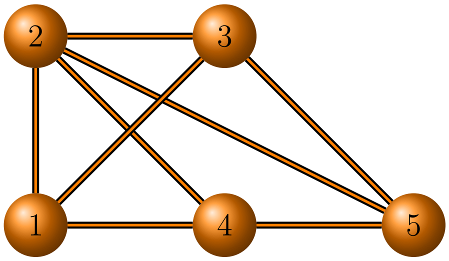

Figure 2.1: An example of a graph with vertex set {1,2,3,4,5}

Example 2.2.6 *(***Random walk on graphs)**A ***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

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

∎

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 the uniform distribution on $S = V$.

■

Example 2.2.8 (Ehrenfest urn model) Suppose K balls are distributed among two urns, A and 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 distribution 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 $\{x_n\}$ 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 $\{x_{n_k}\}$ which converges to $x \in A$.

Let us choose an arbitrary initial distribution μ and let define

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

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 $\{\nu_n\}$ 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

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

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

$$|\pi P(j) - \pi(j)| \,=\, \lim_{k\to\infty} |\nu_{n_k} P(j) - \nu_n(j)| \,\le\, \lim_{k\to\infty} \frac{2}{n_k} \,=\, 0\,.$$

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 $n \ge 0$ 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 $i \leftrightarrow j$. A Markov chain X_n is called ***irreducible*** if every state $i \in S$ communicate with every other state $j \in S$.

Lemma 2.3.2 *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 some $i \in S$. Suppose $i \to j$ then we have

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

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

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

Proposition 2.3.3 *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 $h(i_0) = \max_{i\in S} h(i) \equiv M$. Suppose $i_0 \to j$ but $h(j) < M$, then, since $P^n h = h$,

$$M = h(i_0) \,=\, P^n h(i_0) \,=\, P^n(i_0,j)h(j) + \sum_{i\neq i_0} P^n(i,j)h(j) < M \sum_j P^n(i_0,j) = M\,,$$

and this is a contradiction. ∎

Corollary 2.3.4 *Suppose X_n is irreducible then there exists a unique stationary distribution.*

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 2.3.5 (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 example motivates the following definition. 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 2.3.6 *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 2.3.7 (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 $w \in V_2$. ■

We will need

Proposition 2.3.8 *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. ■

With all this preparation we can now prove.

Theorem 2.3.9 *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.

By proposition 2.3.8 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

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

Furthermore we will need the fact that if M is any stochastic matrix then we have $M\Pi = \Pi$ (because all the rows are the same) and that if M is a stochastic matrix such that $\pi M = \pi$ then $\Pi M = \Pi$.

Next 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 $Q\Pi = \Pi$.

$$\begin{aligned} P^{r(k+1)} \ = \ P^{rk}P^r \ &= \ \left[(1-\theta^k)\Pi + \theta^k Q^k\right] P^r && (2.6) \\ &= \ (1-\theta^k)\Pi P^r + \theta^k Q^k[(1-\theta)\Pi + \theta Q] && (2.7) \\ &= \ (1-\theta^k)\Pi + \theta^k(1-\theta)\Pi + \theta^{k+1}Q^{k+1} && (2.8) \\ &= \ (1-\theta^{k+1})\Pi + \theta^{k+1}Q^{k+1} \, , && (2.9) \end{aligned}$$

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 $0 \leq l < r$. We have then

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

and thus

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

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 $\alpha = \theta^{1/r}$ ∎

So we have show that the stationary distribution is also a limiting distribution. We give first a characterization of $\pi(j)$ in terms of ***occupation times***. To do this we need a lemma from analysis

Lemma 2.3.10 *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 2.3.9 we have $P^n(i,j) \to \pi(j)$ and thus, by Lemma 2.3.10,

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

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.

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,

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

Using this we obtain

Theorem 2.3.11 (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*

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

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

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

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

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

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

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

On the other hand we know that

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

and thus

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

∎

Note that this theorem is of practical importance: it is, in principle, enough to generate one sufficiently long sample of the Markov chain to produce the stationary distribution π.

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.

Let us 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 $0 \le r < d - 1$. 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 2.3.11 of previous section and obtain

Theorem 2.4.1 *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.

We note that first that the communication relation $i \leftrightarrow j$ is an ***equivalence relation***, i.e. it is *reflexive* ($i \leftrightarrow i$), *symmetric* ($i \leftrightarrow j$ implies $j \leftrightarrow i$) and *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***.

We will distinguish between two types of communication classes. We say that a communication class $C \subset S$ is *transient* if there exists $i \in C$ and $j \in S \setminus C$ and $i \in C$ such that $P(i,j) > 0$. Otherwise we call the communication class ***closed***. If X_n start in a closed class C then X_n never leaves C. On the other hand if X_n starts in a transient class then X_n will eventually exit the transient class.

If the Markov chain has r recurrent classes $R_1, \cdots R_r$ and t transient classes $T_1, \cdots T_t$ then, after reordering the states we can put the transition matrix in the form

$$P \;=\; \begin{pmatrix} P_1 & & & & & \\ & P_2 & & 0 & & \\ & & P_3 & & & 0 \\ & 0 & & \ddots & & \\ & & & & P_k & \\ & & S & & & Q \end{pmatrix} \tag{2.10}$$

where P_i gives the transition within the class R_i, Q the transition between the transient classes and S the transistion from the transient classes into the recurrent classes.

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

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

Example 2.5.1 (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 $0, 5, 1, 2, 3, 4$

$$P = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 1/2 & 0 & 0 & 1/2 & 0 & 0 \\ 0 & 0 & 1/2 & 0 & 1/2 & 0 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 1/2 & 0 & 0 & 1/2 & 0 \end{pmatrix} \tag{2.11}$$

■

Suppose i belongs to transient class T_l, then after allowing for some time for the state to access a state which can actually exits T_l, i can exit T_l. Since T_l is finite we can find a time k and $\theta < 1$such that

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

This implies that $P\{X_{nk} \in T_l | X_0 = i\} \le \theta^n$ and so the Markov chain cannot stay in a transient class forever. So if i and j belong to transient classes we have

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

On the other hand if $i \in C_l$ belong to a class class, then the long-time behavior of X_n is entirely determined by the transition matrix of P_l restricted to the closed class C_l to which i belongs. Suppose π_{C_l} is the unique stationary distribution for the Markov chain restricted to C_l, then for $i, j \in C_l$ we have

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

If i belong to a closed class and j belongs to either another closed class or a transient class then $P^n(i,j) = 0$ for all n.

Finally if i belong to a transient class then i will eventually reach a closed class and never leaves. One question we may ask is how long does it take to exit a transient class? If there are several closed classes then one might also ask what are the probabilities to be absorbed in one or another class. We now answer these 2 questions.

The matrix Q in (2.10) is called a substochastic matrix, i.e., a matrix with non-negative entries whose row sums are less than or equal to 1. We have seen 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

$$(I + P + P^2 + \cdots P^n)(I - P) \,=\, I - P^{n+1}\,,$$

and thus

$$M = (I - P)^{-1} \,=\, \sum_{n=0}^{\infty} P^n\,.$$

There $M(j,i)$ is simply the expected number of visits to i if $X_0 = j$.

We summarize this discussion in

Proposition 2.5.2 *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 $M = (I - Q)^{-1}$.

Example 2.5.3 (Random walk with absorbing boundary conditions, cont'd) From (2.11) we have

$$Q \,=\, \left(\begin{array}{cccc} 0 & 1/2 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 \\ 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 1/2 & 0 \end{array}\right), \quad M = (I - Q)^{-1} \,=\, \left(\begin{array}{cccc} 1.6 & 1.2 & 0.8 & 0.4 \\ 1.2 & 2.4 & 1.6 & 0.8 \\ 0.8 & 1.6 & 2.4 & 1.2 \\ 0.4 & .8 & 1.2 & 1.6 \end{array}\right) \quad (2.12)$$

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 write the transition matrix in the block form

$$P \,=\, \left(\begin{array}{cc} P(i,i) & R \\ S & Q \end{array}\right) \quad (2.13)$$

Since the first visit to j starting from i does not depend on the matrix element $P(j,k)$ we can modify the transition matrix P such as to make j an absorbing state without changing the distribution of $\tau^{(j)}$. That is we set

$$\hat{P} \,=\, \left(\begin{array}{cc} 1 & 0 \\ S & Q \end{array}\right).$$

For the Markov chain with transition matrix $\hat{P}$, all states except j now form a transient class and so we can apply Proposition 2.5.2 and obtain

Proposition 2.5.4 *Let X_n be an irreducible Markov chain. We have*

$$E[\tau^{(i)}|X_0 = j] \,=\, \sum_i M(j,i)\,.$$

where $M = (I - Q)^{-1}$ and Q is given in (2.13).

Example 2.5.5 (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

$$P = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 1/2 & 0 & 0 \\ 0 & 0 & 1/2 & 0 & 1/2 & 0 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix} \qquad (2.14)$$

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

$$Q = \begin{pmatrix} 0 & 1/2 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 \\ 0 & 1/2 & 0 & 1/2 & 0 \\ 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}, \quad M = (I - Q)^{-1} = \begin{pmatrix} 2 & 2 & 2 & 2 & 1 \\ 2 & 4 & 4 & 4 & 2 \\ 2 & 4 & 6 & 6 & 3 \\ 2 & 4 & 6 & 8 & 4 \\ 2 & 4 & 6 & 8 & 5 \end{pmatrix} \qquad (2.15)$$

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 $r_1, \cdots r_k$ and that we transient states $t_1, \cdots, t_s$. 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 $\alpha(r_i, r_i) = 1$ and $\alpha(r_i, r_j) = 0$ if $i \neq j$. We condition on the first step of the Markov chain

$$\begin{aligned} A(t_i, r_j) &= P\{X_n = r_j \text{ eventually } |X_0 = t_i\} \\ &= \sum_{l \in S} P\{X_1 = l | X_0 = t_i\} P\{X_n = r_j \text{ eventually } |X_1 = l\} \\ &= \sum_{l \in S} P(t_i, l) A(l, r_j) \;=\; P(t_i, r_j) + \sum_{t_k} P(t_i, t_k) A(t_k, r_j)\,. \end{aligned} \qquad (2.16)$$

Let A be the $s \times k$ matrix with entries $A(t_i, r_j)$, then (2.16) can be written in matrix form as

$$A = S + QA$$

or

$$A = (I - Q)^{-1} S = MS\,.$$

Example 2.5.6 (Random walk with absorbing boundary conditions, cont'd) From (2.11) and (2.12) 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.... ∎

Example 2.5.7 (Gambler's ruin). Let us consider the random walk with absorbing boundary conditions on $\{0, \cdots, N\}$. Let $\alpha(j) \equiv A(j, N)$ the probability that the walker starting at j reaches N before reaching 1. Clearly we have $\alpha(0) = 0$ and $\alpha(N) = 1$. Let us condition on the first steps and we obtain

$$\alpha(j) = (1-p)\alpha(j-1) + p\alpha(j+1)\,.$$

That is we have a systems of second order difference equations for the $N-1$ unknowns $\alpha(j)$. This equations can viewed as a discretization of the linear second order equation $ay'' + by' + cy = 0$ and guided by this we solve this equation by lookin at solutions of the form $\alpha(j) = \beta^j$. This gives the equation

$$\beta = (1-p) + p\beta^2$$

whose solutions are $\frac{1-p}{p}$ and 1. If $p \neq \frac{1}{2}$ then the general solutions is

$$\alpha(j) = c_1 + c_2 \left(\frac{1-p}{p}\right)^j\,.$$

Using the boundary conditions we find

$$\alpha(j) = \frac{1 - \left(\frac{1-p}{p}\right)^j}{1 - \left(\frac{1-p}{p}\right)^N}$$

For $p = \frac{1}{2}$ we have only one solution $\beta = 1$, and inspired by differential equations we try solutions of the form $j\beta^j = j$ which is indeed a solution. For $p = 1/2$ the genera solution is

$$\alpha(j) = c_1 + c_2 j$$

and using the boundary conditions we find

$$\alpha(j) \,=\, \frac{j}{N}.$$

Note that for $p \le \frac{1}{2}$ we have $\lim_{N\to\infty} \alpha(j) = 0$ and this says that a gambler with fixed resources j who plays a game in which he wins with probability p has a very small probability to beat a house with large resources N. If $p > \frac{1}{2}$ then

$$\lim_{N\to\infty} \alpha(j) \,=\, 1 - \left(\frac{1-p}{p}\right)^j > 0\,,$$

and thus there is a postive probability the gambler will never loses all his money and will be able to play forever.

It is also instructive to compute the time until absorption, T, for $p = 1/2$, i.e., the number of (fair) 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 $G(0) = G(N) = 0$. Let us condition on the first step then we obtain

$$G(j) \,=\, 1 + \frac{1}{2}G(j-1) + \frac{1}{2}G(j+1)\,, \quad j = 1, \cdots N-1$$

This an inhomogeneous second order linear difference equation and an educated guess is to try for the particular solution $G(j) = aj^2$ 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

$$\sum_j \pi(i)P(i,j) \,=\, \sum_j \pi(j)P(j,i)\,. \tag{2.17}$$

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 (2.17) means that

$$\sum_i J(i,j) \,=\, \sum_j J(j,i)\,, \tag{2.18}$$

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 $\pi(i) \ge 0$ with $\sum_i \pi(i) = 1$ such that for all i,j we have*

$$\pi(i)P(i,j) \,=\, \pi(j)P(j,i)\,. \tag{2.19}$$

This means that for every pair i,j the probability currents $J(i,j)$ and $J(j,i)$ balance each other. Clearly (2.19) is a stronger condition than (2.17) 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*

$$P\left\{X_0 = i_0\,, X_1 = i_1\,, \cdots, X_n = i_n\right\} \,=\, P\left\{X_0 = i_n\,, X_1 = i_{n-1}\,, \cdots, X_n = i_0\right\} \tag{2.20}$$

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

$$S' = \{\sigma \in \{0,1\}^m \,;\, \sigma \cdot w \le b\}$$

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 $\sigma' \cdot v \le b$ 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 $\sum_i w_i \le b$. 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 $U \le \alpha$ 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

$$\alpha \;=\; \min\left\{1, \frac{\pi(\sigma')}{\pi(\sigma)}\right\} \;=\; \min\left\{1, e^{\beta v\cdot(\sigma'-\sigma)}\right\} \;=\; \left\{\begin{array}{cl} e^{-\beta v_j} & \text{if } \sigma_j = 1 \\ 1 & \text{if } \sigma_j = 0 \end{array}\right.$$

(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

$$\begin{aligned} P(\sigma,\sigma') &= \frac{1}{|V|}\frac{\pi(\sigma_{-v},\sigma'(v))}{\sum_{b\in\Omega}\pi(\sigma_{-v},b)} \quad \text{if } \sigma_{-v}=\sigma'_{-v} \text{ for some } v \\ P(\sigma,\sigma') &= 0 \ \text{ if } \sigma_{-v}\neq\sigma'_{-v} \text{ for all } v \\ P(\sigma,\sigma) &= 1-\sum_{\sigma'}P(\sigma,\sigma') \end{aligned}$$

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

$$\min\left\{1, \frac{\pi(\sigma_{-v}, -\sigma(v))}{\pi(\sigma_{-v}, \sigma(v))}\right\} = \min\left\{1, \frac{\pi(\sigma_{-v}, -\sigma(v))}{\pi(\sigma_{-v}, \sigma(v))}\right\} = \min\left\{1, e^{2\beta\sum_{w\,;\,w\sim v}\sigma(w)\sigma(v)}\right\}\,.$$

$\blacksquare$