Bruce K. Driver

180B Lecture Notes, W2011

February 16, 2011 *File:180Lec.tex*

Contents

Part

180B Notes

0

Basic Probability Facts / Conditional Expectations

0.1 Course Notation

1. (Ω, P) will denote a probability spaces and S will denote a set which is called **state space**.
2. If S is a discrete set, i.e. finite or countable and $X : \Omega \to S$ we let

$$\rho_X(s) := P(X = s).$$

   More generally if $X_i : \Omega \to S_i$ for $1 \le i \le n$ we let

$$\rho_{X_1,\dots,X_n}(\mathbf{s}) := P(X_1 = s_1, \dots, X_n = s_n)$$

   for all $\mathbf{s} = (s_1, \dots, s_n) \in S_1 \times \cdots \times S_n$.
3. If S is $\mathbb{R}$ or $\mathbb{R}^n$ and $X : \Omega \to S$ is a continuous random variable, we let $\rho_X(x)$ be the operability density function of X, namely,

$$\mathbb{E}[f(X)] = \int_S f(x)\,\rho_X(x)\,dx.$$

4. Given random variables X and Y we let;
   a) $\mu_X := \mathbb{E}X$ be the mean of X.
   b) $\operatorname{Var}(X) := \mathbb{E}\left[(X-\mu_X)^2\right] = \mathbb{E}X^2 - \mu_X^2$ be the variance of X.
   c) $\sigma_X = \sigma(X) := \sqrt{\operatorname{Var}(X)}$ be the standard deviation of X.
   d) $\operatorname{Cov}(X,Y) := \mathbb{E}[(X-\mu_X)(Y-\mu_Y)] = \mathbb{E}[XY] - \mu_X\mu_Y$ be the covariance of X and Y.
   e) $Corr(X,Y) := \operatorname{Cov}(X,Y)/(\sigma_X\sigma_Y)$ be the **correlation** of X and Y.

0.2 Some Discrete Distributions

Definition 0.1 (Generating Function). *Suppose that $N : \Omega \to \mathbb{N}_0$ is an integer valued random variable on a probability space, $(\Omega, \mathcal{B}, P)$. The generating function associated to N is defined by*

$$G_N(z) := \mathbb{E}\left[z^N\right] = \sum_{n=0}^{\infty} P(N=n)\,z^n \text{ for } |z| \le 1. \tag{0.1}$$

By Corollary **??**, it follows that $P(N=n) = \frac{1}{n!}G_N^{(n)}(0)$ so that G_N can be used to completely recover the distribution of N.

Proposition 0.2 (Generating Functions). *The generating function satisfies,*

$$G_N^{(k)}(z) = \mathbb{E}\left[N(N-1)\dots(N-k+1)\,z^{N-k}\right] \text{ for } |z| < 1$$

and

$$G^{(k)}(1) = \lim_{z\uparrow 1} G^{(k)}(z) = \mathbb{E}\left[N(N-1)\dots(N-k+1)\right],$$

where it is possible that one and hence both sides of this equation are infinite. In particular, $G'(1) := \lim_{z\uparrow 1} G'(z) = \mathbb{E}N$ and if $\mathbb{E}N^2 < \infty$,

$$\operatorname{Var}(N) = G''(1) + G'(1) - [G'(1)]^2. \tag{0.2}$$

Proof. By Corollary **??** for $|z| < 1$,

$$\begin{aligned} G_N^{(k)}(z) &= \sum_{n=0}^{\infty} P(N=n)\cdot n(n-1)\dots(n-k+1)\,z^{n-k} \\ &= \mathbb{E}\left[N(N-1)\dots(N-k+1)\,z^{N-k}\right]. \end{aligned} \tag{0.3}$$

Since, for $z \in (0,1)$,

$$0 \le N(N-1)\dots(N-k+1)\,z^{N-k} \uparrow N(N-1)\dots(N-k+1) \text{ as } z \uparrow 1,$$

we may apply the MCT to pass to the limit as $z \uparrow 1$ in Eq. (0.3) to find,

$$G^{(k)}(1) = \lim_{z\uparrow 1} G^{(k)}(z) = \mathbb{E}\left[N(N-1)\dots(N-k+1)\right].$$

■

Exercise 0.1 (Some Discrete Distributions). Let $p \in (0,1]$ and $\lambda > 0$. In the four parts below, the distribution of N will be described. You should work out the generating function, $G_N(z)$, in each case and use it to verify the given formulas for $\mathbb{E}N$ and $\operatorname{Var}(N)$.

1. Bernoulli(p) : $P(N=1) = p$ and $P(N=0) = 1-p$. You should find $\mathbb{E}N = p$ and $\operatorname{Var}(N) = p - p^2$.

2. Binomial(n,p) : $P(N=k) = \binom{n}{k}p^k(1-p)^{n-k}$ for $k = 0,1,\dots,n$. ($P(N=k)$ is the probability of k successes in a sequence of n independent yes/no experiments with probability of success being p.) You should find $\mathbb{E}N = np$ and $\operatorname{Var}(N) = n\left(p-p^2\right)$.
3. Geometric(p) : $P(N=k) = p(1-p)^{k-1}$ for $k \in \mathbb{N}$. ($P(N=k)$ is the probability that the k^{th} – trial is the first time of success out a sequence of independent trials with probability of success being p.) You should find $\mathbb{E}N = 1/p$ and $\operatorname{Var}(N) = \frac{1-p}{p^2}$.
4. Poisson(λ) : $P(N=k) = \frac{\lambda^k}{k!}e^{-\lambda}$ for all $k \in \mathbb{N}_0$. You should find $\mathbb{E}N = \lambda = \operatorname{Var}(N)$.

Solution to Exercise (0.1).

1. $G_N(z) = pz^1+(1-p)z^0 = pz+1-p$. Therefore, $G_N'(z) = p$ and $G_N''(z) = 0$ so that $\mathbb{E}N = p$ and $\operatorname{Var}(N) = 0+p-p^2$.
2. $G_N(z) = \sum_{k=0}^n z^k\binom{n}{k}p^k(1-p)^{n-k} = (pz+(1-p))^n$. Therefore,

$$G_N'(z) = n(pz+(1-p))^{n-1}p,$$
$$G_N''(z) = n(n-1)(pz+(1-p))^{n-2}p^2$$

and

$$\mathbb{E}N = np \text{ and } \operatorname{Var}(N) = n(n-1)p^2+np-(np)^2 = n\left(p-p^2\right).$$

3. For the geometric distribution,

$$G_N(z) = \mathbb{E}\left[z^N\right] = \sum_{k=1}^{\infty} z^k p(1-p)^{k-1} = \frac{zp}{1-z(1-p)} \text{ for } |z| < (1-p)^{-1}.$$

Differentiating this equation in z implies,

$$\mathbb{E}\left[Nz^{N-1}\right] = G_N'(z) = \frac{p[1-z(1-p)]+(1-p)pz}{(1-z(1-p))^2}$$
$$= \frac{p}{(1-z(1-p))^2} \text{ and}$$
$$\mathbb{E}\left[N(N-1)z^{N-2}\right] = G_N''(z) = \frac{2(1-p)p}{(1-z(1-p))^3}.$$

Therefore,

$$\mathbb{E}N = G_N'(1) = 1/p,$$
$$\mathbb{E}[N(N-1)] = \frac{2(1-p)p}{p^3} = \frac{2(1-p)p}{p^2},$$

and

$$\operatorname{Var}(N) = 2\frac{1-p}{p^2}+\frac{1}{p}-\frac{1}{p^2} = \frac{1}{p^2}-\frac{1}{p} = \frac{1-p}{p^2}.$$

Alternative method. Starting with $\sum_{n=0}^{\infty} z^n = \frac{1}{1-z}$ for $|z|<1$ we learn that

$$\frac{1}{(1-z)^2} = \frac{d}{dz}\frac{1}{1-z} = \sum_{n=0}^{\infty} nz^{n-1} = \sum_{n=1}^{\infty} nz^{n-1} \text{ and}$$
$$\sum_{n=0}^{\infty} n^2 z^{n-1} = \frac{d}{dz}\frac{z}{(1-z)^2} = \frac{(1-z)^2+2z(1-z)}{(1-z)^4} = \frac{1+z}{(1-z)^3}.$$

Taking $z = 1-p$ in these formulas shows,

$$\mathbb{E}N = p\sum_{n=1}^{\infty} n(1-p)^{n-1} = p\frac{1}{p^2} = \frac{1}{p}$$

and

$$\mathbb{E}N^2 = p\sum_{n=1}^{\infty} n^2(1-p)^{n-1} = p\cdot\frac{2-p}{p^3} = \frac{2-p}{p^2}$$

and therefore,

$$\operatorname{Var}(N) = \frac{2-p}{p^2}-\frac{1}{p^2} = \frac{1-p}{p^2}.$$

4. In the Poisson case,

$$G_N(z) = \mathbb{E}\left[z^N\right] = \sum_{k=0}^{\infty} z^k\frac{\lambda^k}{k!}e^{-\lambda} = e^{-\lambda}e^{\lambda z} = e^{\lambda(z-1)}$$

and $G_N^{(k)}(z) = \lambda^k e^{\lambda(z-1)}$. Therefore, $\mathbb{E}N = \lambda$ and $\mathbb{E}[N\cdot(N-1)] = \lambda^2$ so that $\operatorname{Var}(N) = \lambda^2+\lambda-\lambda^2 = \lambda$.

Remark 0.3 (Memoryless property of the geometric distribution). Suppose that $\{X_i\}$ are i.i.d. Bernoulli random variables with $P(X_i=1) = p$ and $P(X_i=0) = 1-p$ and $N = \inf\{i \geq 1 : X_i = 1\}$. Then $P(N=k) = P(X_1=0,\dots,X_{k-1}=0,X_k=1) = (1-p)^{k-1}p$, so that N is geometric with parameter p. Using this representation we easily and intuitively see that

$$P(N=n+k|N>n) = \frac{P(X_1=0,\dots,X_{n+k-1}=0,X_{n+k}=1)}{P(X_1=0,\dots,X_n=0)}$$
$$= P(X_{n+1}=0,\dots,X_{n+k-1}=0,X_{n+k}=1)$$
$$= P(X_1=0,\dots,X_{k-1}=0,X_k=1) = P(N=k).$$

This can be verified by first principles as well;

$$\begin{aligned} P(N=n+k|N>n) &= \frac{P(N=n+k)}{P(N>n)} = \frac{p(1-p)^{n+k-1}}{\sum_{k>n} p(1-p)^{k-1}} \\ &= \frac{p(1-p)^{n+k-1}}{\sum_{j=0}^{\infty} p(1-p)^{n+j}} = \frac{(1-p)^{n+k-1}}{(1-p)^n \sum_{j=0}^{\infty}(1-p)^j} \\ &= \frac{(1-p)^{k-1}}{\frac{1}{1-(1-p)}} = p(1-p)^{k-1} = P(N=k). \end{aligned}$$

Exercise 0.2. Let $S_{n,p} \overset{d}{=} \text{Binomial}(n,p)$, $k \in \mathbb{N}$, $p_n = \lambda_n/n$ where $\lambda_n \to \lambda > 0$ as $n \to \infty$. Show that

$$\lim_{n\to\infty} P(S_{n,p_n} = k) = \frac{\lambda^k}{k!} e^{-\lambda} = P(\text{Poisson}(\lambda) = k).$$

Thus we see that for $p = O(1/n)$ and k not too large relative to n that for large n,

$$P(\text{Binomial}(n,p) = k) \cong P(\text{Poisson}(pn) = k) = \frac{(pn)^k}{k!} e^{-pn}.$$

(We will come back to the Poisson distribution and the related Poisson process later on.)

Solution to Exercise (0.2). We have,

$$\begin{aligned} P(S_{n,p_n} = k) &= \binom{n}{k} (\lambda_n/n)^k (1-\lambda_n/n)^{n-k} \\ &= \frac{\lambda_n^k}{k!} \frac{n(n-1)\dots(n-k+1)}{n^k} (1-\lambda_n/n)^{n-k}. \end{aligned}$$

The result now follows since,

$$\lim_{n\to\infty} \frac{n(n-1)\dots(n-k+1)}{n^k} = 1$$

and

$$\begin{aligned} \lim_{n\to\infty} \ln(1-\lambda_n/n)^{n-k} &= \lim_{n\to\infty} (n-k)\ln(1-\lambda_n/n) \\ &= -\lim_{n\to\infty} [(n-k)\lambda_n/n] = -\lambda. \end{aligned}$$

1

Course Overview and Plan

This course is an introduction to some basic topics in the theory of stochastic processes. After finishing the discussion of multivariate distributions and conditional probabilities initiated in Math 180A, we will study Markov chains in discrete time. We then begin our investigation of stochastic processes in continuous time with a detailed discussion of the Poisson process. These two topics will be combined in Math 180C when we study Markov chains in continuous time and renewal processes.

In the next two quarters we will study some aspects of Stochastic Processes. Stochastic (from the Greek $\sigma\tau\acute{o}\chi o\xi$ for aim or guess) means random. A stochastic process is one whose behavior is non-deterministic, in that a system's subsequent state is determined both by the process's predictable actions and by a random element. However, according to M. Kac[1] and E. Nelson[2], any kind of time development (be it deterministic or essentially probabilistic) which is analyzable in terms of probability deserves the name of stochastic process.

Mathematically we will be interested in collection of random variables or vectors $\{X_t\}_{t\in T}$ with $X_t : \Omega \to S$ (S is the **state space**) on some probability space, (Ω, P). Here T is typically in $\mathbb{R}_+$ or $\mathbb{Z}_+$ but not always.

Example 1.1. 1. X_t is the value of a spinner at times $t \in \mathbb{Z}_+$.
2. X_t denotes the prices of a stock (or stocks) on the stock market.
3. X_t denotes the value of your portfolio at time t.
4. X_t is the position of a dust particle like in Brownian motion.
5. X_A is the number of stars in a region A contained in space or the number of raisins in a region of a cake, etc.
6. $X_n \in S = Perm(\{1,\dots,52\})$ is the ordering of cards in a deck of cards after the n^{th} shuffle.

Our goal in this course is to introduce and analyze models for such random objects. This is clearly going to require that we make assumptions on $\{X_t\}$ which will typically be some sort of dependency structures. This is where we will begin our study – namely heading towards conditional expectations and related topics.

[1] M. Kac & J. Logan, in Fluctuation Phenomena, eds. E.W. Montroll & J.L. Lebowitz, North-Holland, Amsterdam, 1976.
[2] E. Nelson, Quantum Fluctuations, Princeton University Press, Princeton, 1985.

1.1 180B Course Topics:

1. Review the linear algebra of orthogonal projections in the context of least squares approximations in the context of Probability Theory.
2. Use the least squares theory to interpret covariance and correlations.
3. Review of conditional probabilities for discrete random variables.
4. Introduce conditional expectations as least square approximations.
5. Develop conditional expectation relative to discrete random variables.
6. Give a short introduction to martingale theory.
7. Study in some detail discrete time Markov chains.
8. Review of conditional probability densities for continuous random variables.
9. Develop conditional expectations relative to continuous random variables.
10. Begin our study of the Poisson process.

The bulk of this quarter will involve the study of Markov chains and processes. These are processes for which the past and future are independent given the present. This is a typical example of a dependency structure that we will consider in this course. For an example of such a process, let $S = \mathbb{Z}$ and place a coin at each site of S (perhaps the coins are biased with different probabilities of heads at each site of S.) Let $X_0 = s_0$ be some point in S be fixed and then flip the coin at s_0 and move to the right on step if the result is heads and to left one step if the result is tails. Repeat this process to determine the position X_{n+1} from the position X_n along with a flip of the coin at X_n. This is a typical example of a Markov process.

Before going into these and other processes in more detail we are going to develop the extremely important concept of **conditional expectation.** The idea is as follows. Suppose that X and Y are two random variables with $\mathbb{E}|Y|^2 < \infty$. We wish to find the function h such that $h(X)$ is the minimizer of $\mathbb{E}(Y - f(X))^2$ over all functions f such that $\mathbb{E}\left[f(X)^2\right] < \infty$, that is $h(X)$ is a least squares approximation to Y among random variables of the form $f(X)$, i.e.

$$\mathbb{E}(Y - h(X))^2 = \min_f \mathbb{E}(Y - f(X))^2. \tag{1.1}$$

Fact: a minimizing function h always exist and is "essentially unique." We denote $h(X)$ as $\mathbb{E}[Y|X]$ and call it the **conditional expectation of Y given**

X. We are going to spend a fair amount of time filling in the details of this construction and becoming familiar with this concept.

As a warm up to conditional expectation, we are going to consider a simpler problem of best linear approximations. The goal now is to find $a_0, b_0 \in \mathbb{R}$ such that

$$\mathbb{E}\left(Y - a_0 X + b_0\right)^2 = \min_{a,b\in\mathbb{R}} \mathbb{E}\left(Y - aX + b\right)^2. \tag{1.2}$$

This is the same sort of problem as finding conditional expectations except we now only allow consider functions of the form $f(x) = ax + b$. (You should be able to find a_0 and b_0 using the first derivative test from calculus! We will carry this out using linear algebra ideas below.) It turns out the answer to finding (a_0, b_0) solving Eq. (1.2) only requires knowing the first and second moments of X and Y and $\mathbb{E}[XY]$. On the other hand finding $h(X)$ solving Eq. (1.1) require full knowledge of the joint distribution of (X, Y).

By the way, you are asked to show on your first homework that $\min_{c\in\mathbb{R}} \mathbb{E}\left(Y - c\right)^2 = \mathrm{Var}(Y)$ which occurs for $c = \mathbb{E}Y$. Thus $\mathbb{E}Y$ is the least squares approximation to Y by a constant function and $\mathrm{Var}(Y)$ is the least square error associated with this problem.

Covariance and Correlation

Suppose that (Ω, P) is a probability space. We say that $X : \Omega \to \mathbb{R}$ is **integrable** if $\mathbb{E}|X| < \infty$ and X is **square integrable** if $\mathbb{E}|X|^2 < \infty$. We denote the set of integrable random variables by $L^1(P)$ and the square integrable random variables by $L^2(P)$. When X is integrable we let $\mu_X := \mathbb{E}X$ be the **mean** of X. If Ω is a finite set, then

$$\mathbb{E}[|X|^p] = \sum_{\omega \in \Omega} |X(\omega)|^p P(\{\omega\}) < \infty$$

for any $0 < p < \infty$. So when the sample space is finite requiring integrability or square integrability is no restriction at all. On the other hand when Ω is infinite life can become a little more complicated.

Example 2.1. Suppose that N is a geometric with parameter p so that $P(N = k) = p(1-p)^{k-1}$ for $k \in \mathbb{N} = \{1, 2, 3, \dots\}$. If $X = f(N)$ for some function $f : \mathbb{N} \to \mathbb{R}$, then

$$\mathbb{E}[f(N)] = \sum_{k=1}^{\infty} p(1-p)^{k-1} f(k)$$

when the sum makes sense. So if $X_\lambda = \lambda^N$ for some $\lambda > 0$ we have

$$\mathbb{E}\left[X_\lambda^2\right] = \sum_{k=1}^{\infty} p(1-p)^{k-1} \lambda^{2k} = p\lambda^2 \sum_{k=1}^{\infty} \left[(1-p)\lambda^2\right]^{k-1} < \infty$$

iff $(1-p)\lambda^2 < 1$, i.e. $\lambda < 1/\sqrt{1-p}$. Thus we see that $X_\lambda \in L^2(P)$ iff $\lambda < 1/\sqrt{1-p}$.

Lemma 2.2. *$L^2(P)$ is a subspace of the vector space of random variables on (Ω, P). Moreover if $X, Y \in L^2(P)$, then $XY \in L^1(P)$ and in particular (take $Y = 1$) it follows that $L^2(P) \subset L^1(P)$.*

Proof. If $X, Y \in L^2(P)$ and $c \in \mathbb{R}$ then $\mathbb{E}|cX|^2 = c^2\mathbb{E}|X|^2 < \infty$ so that $cX \in L^2(P)$. Since

$$0 \le (|X| - |Y|)^2 = |X|^2 + |Y|^2 - 2|X||Y|,$$

it follows that

$$|XY| \le \frac{1}{2}|X|^2 + \frac{1}{2}|Y|^2 \in L^1(P).$$

Moreover,

$$(X+Y)^2 = X^2 + Y^2 + 2XY \le X^2 + Y^2 + 2|XY| \le 2\left(X^2 + Y^2\right)$$

from which it follows that $\mathbb{E}(X+Y)^2 < \infty$, i.e. $X + Y \in L^2(P)$. ∎

Definition 2.3. *The **covariance,** $\mathrm{Cov}(X, Y)$, of two square integrable random variables, X and Y, is defined by*

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

*where $\mu_X := \mathbb{E}X$ and $\mu_Y := \mathbb{E}Y$. The **variance** of X,*

$$\mathrm{Var}(X) = \mathrm{Cov}(X, X) = \mathbb{E}\left[X^2\right] - (\mathbb{E}X)^2 \tag{2.1}$$

$$= \mathbb{E}\left[(X - \mu_X)^2\right] \tag{2.2}$$

*We say that X and Y are **uncorrelated** if $\mathrm{Cov}(X, Y) = 0$, i.e. $\mathbb{E}[XY] = \mathbb{E}X \cdot \mathbb{E}Y$. More generally we say $\{X_k\}_{k=1}^n \subset L^2(P)$ are **uncorrelated** iff $\mathrm{Cov}(X_i, X_j) = 0$ for all $i \neq j$.*

Definition 2.4 (Correlation). *Given two non-constant random variables we define $Corr(X, Y) := \frac{\mathrm{Cov}(X,Y)}{\sigma(X)\cdot\sigma(Y)}$ to be the **correlation** of X and Y.*

It follows from Eqs. (2.1) and (2.2) that

$$0 \le \mathrm{Var}(X) \le \mathbb{E}\left[X^2\right] \text{ for all } X \in L^2(P). \tag{2.3}$$

Exercise 2.1. Let X, Y be two random variables on $(\Omega, \mathcal{B}, P)$;

1. Show that X and Y are independent iff $\mathrm{Cov}(f(X), g(Y)) = 0$ (i.e. $f(X)$ and $g(Y)$ are **uncorrelated**) for bounded measurable functions, $f, g : \mathbb{R} \to \mathbb{R}$. (In this setting X and Y may take values in some arbitrary state space, S.)
2. If $X, Y \in L^2(P)$ and X and Y are independent, then $\mathrm{Cov}(X, Y) = 0$. Note well: we will see in examples below that $\mathrm{Cov}(X, Y) = 0$ does **not** necessarily imply that X and Y are independent.

Solution to Exercise (2.1). (Only roughly sketched the proof of this in class.)

1. Since

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

it follows that $\operatorname{Cov}(f(X), g(Y)) = 0$ iff

$$\mathbb{E}[f(X)g(Y)] = \mathbb{E}[f(X)]\mathbb{E}[g(Y)]$$

from which item 1. easily follows.

2. Let $f_M(x) = x1_{|x|\le M}$, then by independence,

$$\mathbb{E}[f_M(X)g_M(Y)] = \mathbb{E}[f_M(X)]\mathbb{E}[g_M(Y)]. \tag{2.4}$$

Since

$$|f_M(X)g_M(Y)| \le |XY| \le \frac{1}{2}\left(X^2+Y^2\right) \in L^1(P),$$
$$|f_M(X)| \le |X| \le \frac{1}{2}\left(1+X^2\right) \in L^1(P), \text{ and}$$
$$|g_M(Y)| \le |Y| \le \frac{1}{2}\left(1+Y^2\right) \in L^1(P),$$

we may use the DCT three times to pass to the limit as $M \to \infty$ in Eq. (2.4) to learn that $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$, i.e. $\operatorname{Cov}(X,Y) = 0$. (These technical details were omitted in class.)

End of 1/3/2011 Lecture.

Example 2.5. Suppose that $P(X \in dx, Y \in dy) = e^{-y}1_{0<x<y}dxdy$. Recall that

$$\int_0^\infty y^k e^{-\lambda y}dy = \left(-\frac{d}{d\lambda}\right)^k \int_0^\infty e^{-\lambda y}dy = \left(-\frac{d}{d\lambda}\right)^k \frac{1}{\lambda} = k!\frac{1}{\lambda^{k+1}}.$$

Therefore,

$$\mathbb{E}Y = \int\int ye^{-y}1_{0<x<y}dxdy = \int_0^\infty y^2e^{-y}dy = 2,$$
$$\mathbb{E}Y^2 = \int\int y^2e^{-y}1_{0<x<y}dxdy = \int_0^\infty y^3e^{-y}dy = 3! = 6$$

$$\mathbb{E}X = \int\int xe^{-y}1_{0<x<y}dxdy = \frac{1}{2}\int_0^\infty y^2e^{-y}dy = 1,$$
$$\mathbb{E}X^2 = \int\int x^2e^{-y}1_{0<x<y}dxdy = \frac{1}{3}\int_0^\infty y^3e^{-y}dy = \frac{1}{3}3! = 2$$

and

$$\mathbb{E}[XY] = \int\int xye^{-y}1_{0<x<y}dxdy = \frac{1}{2}\int_0^\infty y^3e^{-y}dy = \frac{3!}{2} = 3.$$

Therefore $\operatorname{Cov}(X,Y) = 3 - 2\cdot 1 = 1$, $\sigma^2(X) = 2 - 1^2 = 1$, $\sigma^2(Y) = 6 - 2^2 = 2$,

$$Corr(X,Y) = \frac{1}{\sqrt{2}}.$$

Lemma 2.6. *The covariance function,* $\operatorname{Cov}(X,Y)$ *is bilinear in* X *and* Y *and* $\operatorname{Cov}(X,Y) = 0$ *if either* X *or* Y *is constant. For any constant* k, $\operatorname{Var}(X+k) = \operatorname{Var}(X)$ *and* $\operatorname{Var}(kX) = k^2\operatorname{Var}(X)$. *If* $\{X_k\}_{k=1}^n$ *are uncorrelated* $L^2(P)$ *– random variables, then*

$$\operatorname{Var}(S_n) = \sum_{k=1}^n \operatorname{Var}(X_k).$$

Proof. We leave most of this simple proof to the reader. As an example of the type of argument involved, let us prove $\operatorname{Var}(X+k) = \operatorname{Var}(X)$;

$$\begin{aligned}\operatorname{Var}(X+k) &= \operatorname{Cov}(X+k, X+k) = \operatorname{Cov}(X+k, X) + \operatorname{Cov}(X+k, k)\\ &= \operatorname{Cov}(X+k, X) = \operatorname{Cov}(X,X) + \operatorname{Cov}(k,X)\\ &= \operatorname{Cov}(X,X) = \operatorname{Var}(X),\end{aligned}$$

wherein we have used the bilinearity of $\operatorname{Cov}(\cdot,\cdot)$ and the property that $\operatorname{Cov}(Y,k) = 0$ whenever k is a constant. ∎

Example 2.7. Suppose that X and Y are distributed as follows;

$$\begin{array}{cccccc} & \rho_Y & 1/4 & \frac{1}{2} & 1/4 \\ \rho_X & X\backslash Y & -1 & 0 & 1 \\ 1/4 & 1 & 0 & 1/4 & 0 \\ 3/4 & 0 & 1/4 & 1/4 & 1/4 \end{array}$$

so that $\rho_{X,Y}(1,-1) = P(X=1,Y=-1) = 0$, $\rho_{X,Y}(1,0) = P(X=1,Y=0) = 1/4$, etc. In this case $XY = 0$ a.s. so that $\mathbb{E}[XY] = 0$ while

$$\mathbb{E}[X] = 1 \cdot \frac{1}{4} + 0 \cdot \frac{3}{4} = \frac{1}{4}, \text{ and}$$
$$\mathbb{E}Y = (-1)\,1/4 + 0\frac{1}{2} + 1\frac{1}{4} = 0$$

so that $\operatorname{Cov}(X,Y) = 0 - \frac{1}{4} \cdot 0 = 0$. Again X and Y are not independent since $\rho_{X,Y}(x,y) \neq \rho_X(x)\,\rho_Y(y)$.

Example 2.8. Let X have an even distribution and let $Y = X^2$, then

$$\operatorname{Cov}(X,Y) = \mathbb{E}\left[X^3\right] - \mathbb{E}\left[X^2\right] \cdot \mathbb{E}X = 0$$

since,

$$\mathbb{E}\left[X^{2k+1}\right] = \int_{-\infty}^{\infty} x^{2k+1} \rho(x)\,dx = 0 \text{ for all } k \in \mathbb{N}.$$

On the other hand $\operatorname{Cov}\left(Y, X^2\right) = \operatorname{Cov}(Y,Y) = \operatorname{Var}(Y) \neq 0$ in general so that Y is not independent of X.

Example 2.9 (Not done in class.). Let X and Z be independent with $P(Z = \pm 1) = \frac{1}{2}$ and take $Y = XZ$. Then $\mathbb{E}Z = 0$ and

$$\begin{aligned} \operatorname{Cov}(X,Y) &= \mathbb{E}\left[X^2 Z\right] - \mathbb{E}[X]\,\mathbb{E}[XZ] \\ &= \mathbb{E}\left[X^2\right] \cdot \mathbb{E}Z - \mathbb{E}[X]\,\mathbb{E}[X]\,\mathbb{E}Z = 0. \end{aligned}$$

On the other hand it should be intuitively clear that X and Y are not independent since knowledge of X typically will give some information about Y. To verify this assertion let us suppose that X is a discrete random variable with $P(X=0) = 0$. Then

$$P(X=x, Y=y) = P(X=x, xZ=y) = P(X=x) \cdot P(X=y/x)$$

while

$$P(X=x)\,P(Y=y) = P(X=x) \cdot P(XZ=y).$$

Thus for X and Y to be independent we would have to have,

$$P(xX=y) = P(XZ=y) \text{ for all } x, y.$$

This is clearly not going to be true in general. For example, suppose that $P(X=1) = \frac{1}{2} = P(X=0)$. Taking $x = y = 1$ in the previously displayed equation would imply

$$\frac{1}{2} = P(X=1) = P(XZ=1) = P(X=1, Z=1) = P(X=1)\,P(Z=1) = \frac{1}{4}$$

which is false.

Presumably you saw the following exercise in Math 180A.

Exercise 2.2 (A Weak Law of Large Numbers). Assume $\{X_n\}_{n=1}^{\infty}$ is a sequence if uncorrelated square integrable random variables which are identically distributed, i.e. $X_n \overset{d}{=} X_m$ for all $m, n \in \mathbb{N}$. Let $S_n := \sum_{k=1}^n X_k$, $\mu := \mathbb{E}X_k$ and $\sigma^2 := \operatorname{Var}(X_k)$ (these are independent of k). Show;

$$\mathbb{E}\left[\frac{S_n}{n}\right] = \mu,$$
$$\mathbb{E}\left(\frac{S_n}{n} - \mu\right)^2 = \operatorname{Var}\left(\frac{S_n}{n}\right) = \frac{\sigma^2}{n}, \text{ and}$$
$$P\left(\left|\frac{S_n}{n} - \mu\right| > \varepsilon\right) \leq \frac{\sigma^2}{n\varepsilon^2}$$

for all $\varepsilon > 0$ and $n \in \mathbb{N}$.

3

Geometric aspects of $L^2(P)$

Definition 3.1 (Inner Product). *For $X, Y \in L^2(P)$, let $(X,Y) := \mathbb{E}[XY]$ and $\|X\| := \sqrt{(X,X)} = \sqrt{\mathbb{E}|X^2|}$.*

Example 3.2 (This was already mentioned in Lecture 1 with $N = 4$.). Suppose that $\Omega = \{1, \dots, N\}$ and $P(\{i\}) = \frac{1}{N}$ for $1 \le i \le N$. Then

$$(X,Y) = \mathbb{E}[XY] = \frac{1}{N}\sum_{i=1}^{N} X(i)\,Y(i) = \frac{1}{N}\mathbf{X}\cdot\mathbf{Y}$$

where

$$\mathbf{X} := \begin{bmatrix} X(1) \\ X(2) \\ \vdots \\ X(N) \end{bmatrix} \text{ and } \mathbf{Y} := \begin{bmatrix} Y(1) \\ Y(2) \\ \vdots \\ Y(N) \end{bmatrix}.$$

Thus the inner product we have defined in this case is essentially the dot product that you studied in math 20F.

Remark 3.3. The inner product on $H := L^2(P)$ satisfies,

1. $(aX + bY, Z) = a(X,Z) + b(Y,Z)$ i.e. $X \to (X,Z)$ is linear.
2. $(X,Y) = (Y,X)$ (symmetry).
3. $\|X\|^2 := (X,X) \ge 0$ with $\|X\|^2 = 0$ iff $X = 0$.

Notice that combining properties (1) and (2) that $X \to (Z,X)$ is linear for fixed $Z \in H$, i.e.

$$(Z, aX + bY) = a(Z,X) + b(Z,Y).$$

The following identity will be used frequently in the sequel without further mention,

$$\begin{aligned} \|X+Y\|^2 &= (X+Y, X+Y) = \|X\|^2 + \|Y\|^2 + (X,Y) + (Y,X) \\ &= \|X\|^2 + \|Y\|^2 + 2(X,Y). \end{aligned} \tag{3.1}$$

Theorem 3.4 (Schwarz Inequality). *Let $(H, (\cdot,\cdot))$ be an inner product space, then for all $X, Y \in H$*

$$|(X,Y)| \le \|X\|\|Y\|$$

and equality holds iff X and Y are linearly dependent. Applying this result to $|X|$ and $|Y|$ shows,

$$\mathbb{E}[|XY|] \le \|X\| \cdot \|Y\|.$$

Proof. If $Y = 0$, the result holds trivially. So assume that $Y \neq 0$ and observe; if $X = \alpha Y$ for some $\alpha \in \mathbb{C}$, then $(X,Y) = \alpha\|Y\|^2$ and hence

$$|(X,Y)| = |\alpha|\,\|Y\|^2 = \|X\|\|Y\|.$$

Now suppose that $X \in H$ is arbitrary, let $Z := X - \|Y\|^{-2}(X,Y)Y$. (So $\|Y\|^{-2}(X,Y)Y$ is the "orthogonal projection" of X along Y, see Figure 3.1.)

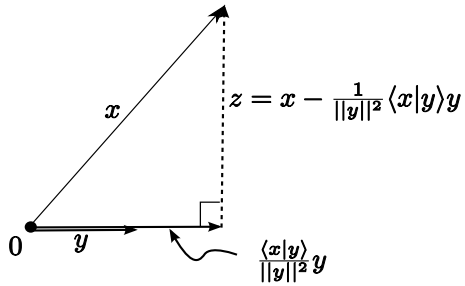


Fig. 3.1. The picture behind the proof of the Schwarz inequality.

Then

$$\begin{aligned} 0 \le \|Z\|^2 &= \left\| X - \frac{(X,Y)}{\|Y\|^2}Y \right\|^2 = \|X\|^2 + \frac{|(X|Y)|^2}{\|Y\|^4}\|Y\|^2 - 2(X|\frac{(X|Y)}{\|Y\|^2}Y) \\ &= \|X\|^2 - \frac{|(X|Y)|^2}{\|Y\|^2} \end{aligned}$$

from which it follows that $0 \le \|Y\|^2\|X\|^2 - |(X|Y)|^2$ with equality iff $Z = 0$ or equivalently iff $X = \|Y\|^{-2}(X|Y)Y$.

Alternative argument: Let $c \in \mathbb{R}$ and $Z := X - cY$, then

$$0 \le \|Z\|^2 = \|X - cY\|^2 = \|X\|^2 - 2c(X,Y) + c^2\|Y\|^2.$$

The right side of this equation is minimized at $c = (X,Y)/\|Y\|^2$ and for this valued of c we find,

$$0 \le \|X - cY\|^2 = \|X\|^2 - (X,Y)^2/\|Y\|^2$$

with equality iff $X = cY$. Solving this last inequality for $|(X,Y)|$ gives the result. ∎

Corollary 3.5. *The norm, $\|\cdot\|$, satisfies the triangle inequality and $(\cdot,\cdot)$ is continuous on $H \times H$.*

Proof. If $X, Y \in H$, then, using Schwarz's inequality,

$$\begin{aligned}\|X+Y\|^2 &= \|X\|^2 + \|Y\|^2 + 2(X,Y)\\ &\le \|X\|^2 + \|Y\|^2 + 2\|X\|\|Y\| = (\|X\| + \|Y\|)^2.\end{aligned}$$

Taking the square root of this inequality shows $\|\cdot\|$ satisfies the triangle inequality. (The rest of this proof may be skipped.)

Checking that $\|\cdot\|$ satisfies the remaining axioms of a norm is now routine and will be left to the reader. If $X, Y, \Delta X, \Delta Y \in H$, then

$$\begin{aligned}|(X+\Delta X, Y+\Delta Y) - (X,Y)| &= |(X,\Delta Y) + (\Delta X, Y) + (\Delta X, \Delta Y)|\\ &\le \|X\|\|\Delta Y\| + \|Y\|\|\Delta X\| + \|\Delta X\|\|\Delta Y\|\\ &\to 0 \text{ as } \Delta X, \Delta Y \to 0,\end{aligned}$$

from which it follows that $(\cdot,\cdot)$ is continuous. ∎

Definition 3.6. *Let $(H,(\cdot,\cdot))$ be an inner product space, we say $X, Y \in H$ are **orthogonal** and write $X \perp Y$ iff $(X,Y) = 0$. More generally if $A \subset H$ is a set, $X \in H$ is **orthogonal to** A (write $X \perp A$) iff $(X,Y) = 0$ for all $Y \in A$. Let $A^\perp = \{X \in H : X \perp A\}$ be the set of vectors orthogonal to A. A subset $S \subset H$ is an **orthogonal set** if $X \perp Y$ for all distinct elements $X, Y \in S$. If S further satisfies, $\|X\| = 1$ for all $X \in S$, then S is said to be an **orthonormal set.***

Proposition 3.7. *Let $(H,(\cdot,\cdot))$ be an inner product space then*

1. ***(Pythagorean Theorem)** If $S \subset\subset H$ is a finite orthogonal set, then*

$$\left\|\sum_{X\in S} X\right\|^2 = \sum_{X\in S}\|X\|^2. \tag{3.2}$$

2. ***(Parallelogram Law)** (Skip this one.) For all $X, Y \in H$,*

$$\|X+Y\|^2 + \|X-Y\|^2 = 2\|X\|^2 + 2\|Y\|^2 \tag{3.3}$$

Proof. Items 1. and 2. are proved by the following elementary computations;and

$$\begin{aligned}\left\|\sum_{X\in S} X\right\|^2 &= \left(\sum_{X\in S} X, \sum_{Y\in S} Y\right) = \sum_{X,Y\in S}(X,Y)\\ &= \sum_{X\in S}(X,X) = \sum_{X\in S}\|X\|^2\end{aligned}$$

and

$$\begin{aligned}&\|X+Y\|^2 + \|X-Y\|^2\\ &\quad= \|X\|^2 + \|Y\|^2 + 2(X,Y) + \|X\|^2 + \|Y\|^2 - 2(X,Y)\\ &\quad= 2\|X\|^2 + 2\|Y\|^2.\end{aligned}$$

∎

Theorem 3.8 (Least Squares Approximation Theorem). *Suppose that V is a subspace of $H := L^2(P)$, $X \in V$, and $Y \in L^2(P)$. Then the following are equivalent;*

1. *$\|Y - X\| \ge \|Y - Z\|$ for all $Z \in V$ (i.e. X is a least squares approximation to Y by an element from V) and*
2. *$(Y - X) \perp V$.*

*Moreover there is "essentially" at most one $X \in V$ satisfying 1. or equivalently 2. We denote random variable by $Q_V Y$ and call it **orthogonal projection of Y along V.***

Proof. 1 $\implies$ 2. If 1. holds then $f(t) := \|Y - (X + tZ)\|^2$ has a minimum at $t = 0$ and therefore $\dot f(0) = 0$. Since

$$f(t) := \|Y - X - tZ\|^2 = \|Y - X\|^2 + t^2\|Z\|^2 - 2t(Y - X, Z),$$

we may conclude that

$$0 = \dot f(0) = -2(Y - X, Z).$$

As $Z \in V$ was arbitrary we may conclude that $(Y - X) \perp V$.

2 $\implies$ 1. Now suppose that $(Y - X) \perp V$ and $Z \in V$, then $(Y - X) \perp (X - Z)$ and so

$$\|Y - Z\|^2 = \|Y - X + X - Z\|^2 = \|Y - X\|^2 + \|X - Z\|^2 \ge \|Y - X\|^2. \tag{3.4}$$

Moreover if Z is another best approximation to Y then $\|Y - Z\|^2 = \|Y - X\|^2$ which happens according to Eq. (3.4) iff

$$\|X - Z\|^2 = \mathbb{E}(X - Z)^2 = 0,$$

i.e. iff $X = Z$ a.s. ∎

End of Lecture 3: 1/07/2011 (Given by Tom Laetsch)

Corollary 3.9 (Orthogonal Projection Formula). *Suppose that V is a subspace of $H := L^2(P)$ and $\{X_i\}_{i=1}^N$ is an orthogonal basis for V. Then*

$$Q_V Y = \sum_{i=1}^{N} \frac{(Y, X_i)}{\|X_i\|^2} X_i \text{ for all } Y \in H.$$

Proof. The best approximation $X \in V$ to Y is of the form $X = \sum_{i=1}^N c_i X_i$ where $c_i \in \mathbb{R}$ need to be chosen so that $(Y - X) \perp V$. Equivalently put we must have

$$0 = (Y - X, X_j) = (Y, X_j) - (X, X_j) \text{ for } 1 \leq j \leq N.$$

Since

$$(X, X_j) = \sum_{i=1}^{N} c_i (X_i, X_j) = c_j \|X_j\|^2,$$

we see that $c_j = (Y, X_j) / \|X_j\|^2$, i.e.

$$Q_V Y = X = \sum_{i=1}^{N} \frac{(Y, X_i)}{\|X_i\|^2} X_i.$$

■

Example 3.10. Given $Y \in L^2(P)$ the best approximation to Y by a constant function c is given by

$$c = \frac{\mathbb{E}[Y1]}{\mathbb{E}1^2} 1 = \mathbb{E}Y.$$

You already proved this on your first homework by a direct calculus exercise.

Linear prediction and a canonical form

Corollary 4.1 (Correlation Bounds). *For all square integrable random variables, X and Y,*

$$|\mathrm{Cov}(X,Y)| \le \sigma(X)\cdot\sigma(Y)$$

or equivalently,

$$|Corr(X,Y)| \le 1.$$

.

Proof. This is a simply application of Schwarz's inequality (Theorem 3.4);

$$|\mathrm{Cov}(X,Y)| = |\mathbb{E}[(X-\mu_X)(Y-\mu_Y)]| \le \|X-\mu_X\|\cdot\|Y-\mu_Y\| = \sigma(X)\cdot\sigma(Y).$$

■

Since $Corr(X,Y) > 0$ iff $\mathrm{Cov}(X,Y) > 0$ iff $\mathbb{E}[(X-\mu_X)(Y-\mu_Y)] > 0$, we see that X and Y are positively correlated iff $X-\mu_X$ and $Y-\mu_Y$ tend to have the same sign more often than not. While X and Y are negatively correlated iff $X-\mu_X$ and $Y-\mu_Y$ tend to have opposite signs more often than not. This description is of course rather crude given that it ignores size of $X-\mu_X$ and $Y-\mu_Y$ but should however give the reader a little intuition into the meaning of correlation. (See Corollary 4.4 below for the special case where $Corr(X,Y)=1$ or $Corr(X,Y)=-1$.)

Theorem 4.2 (Linear Prediction Theorem). *Let X and Y be two square integrable random variables, then*

$$\sigma(Y)\sqrt{1-Corr^2(X,Y)} = \min_{a,b\in\mathbb{R}} \|Y-(aX+b)\| = \|Y-W\| \tag{4.1}$$

where

$$W = \mu_Y + \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}(X-\mu_X) = \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}X + \left(\mathbb{E}Y - \mu_X\frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}\right).$$

Proof. Let $\mu = \mathbb{E}X$ and $\bar{X} = X-\mu$. Then $\{1,\bar{X}\}$ is an orthogonal set and $V := \mathrm{span}\{1,X\} = \mathrm{span}\{1,\bar{X}\}$. Thus best approximation of Y by random variable of the form $aX+b$ is given by

$$W = (Y,1)\,1 + \frac{(Y,\bar{X})}{\|\bar{X}\|^2}\bar{X} = \mathbb{E}Y + \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}(X-\mu_X).$$

The **root mean square error** of this approximation is

$$\begin{aligned}\|Y-W\|^2 &= \left\|\bar{Y} - \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}\bar{X}\right\|^2 = \sigma^2(Y) - \frac{\mathrm{Cov}^2(X,Y)}{\sigma^2(X)}\\ &= \sigma^2(Y)\left(1-Corr^2(X,Y)\right),\end{aligned}$$

so that

$$\|Y-W\| = \sigma(Y)\sqrt{1-Corr^2(X,Y)}.$$

■

Example 4.3. Suppose that $P(X\in dx, Y\in dy) = e^{-y}1_{0<x<y}dxdy$. Recall from Example 2.5 that

$$\begin{aligned}\mathbb{E}X = 1, &\qquad \mathbb{E}Y = 2,\\ \mathbb{E}X^2 = 2, &\qquad \mathbb{E}Y^2 = 6\\ \sigma(X) = 1, &\qquad \sigma(Y) = \sqrt{2},\\ \mathrm{Cov}(X,Y) = 1, &\text{ and } Corr(X,Y) = \frac{1}{\sqrt{2}}.\end{aligned}$$

So in this case

$$W = 2 + \frac{1}{1}(X-1) = X+1$$

is the best linear predictor of Y and the root mean square error in this prediction is

$$\|Y-W\| = \sqrt{2}\sqrt{1-\frac{1}{2}} = 1.$$

Corollary 4.4. *If $Corr(X,Y) = \pm 1$, then*

$$Y = \mu_Y \pm \frac{\sigma(Y)}{\sigma(X)}(X-\mu_X),$$

i.e. $Y-\mu_Y$ is a positive (negative) multiple of $X-\mu_X$ if $Corr(X,Y)=1$ ($Corr(X,Y)=-1$).

Proof. According to Eq. (4.1) of Theorem 4.2, if $Corr(X,Y) = \pm 1$ then

$$\begin{aligned} Y &= \mu_Y + \frac{\operatorname{Cov}(X,Y)}{\operatorname{Var}(X)}(X-\mu_X) \\ &= \mu_Y \pm \frac{\sigma_X\sigma_Y}{\sigma_X^2}(X-\mu_X) = \mu_Y \pm \frac{\sigma_Y}{\sigma_X}(X-\mu_X), \end{aligned}$$

wherein we have used $\operatorname{Cov}(X,Y) = \operatorname{Cov}(X,Y)\,\sigma_X\sigma_Y = \pm 1\sigma_X\sigma_Y$. ∎

Theorem 4.5 (Canonical form). *If $X, Y \in L^2(P)$, then there are two mean zero uncorrelated Random variables $\{Z_1, Z_2\}$ such that $\|Z_1\| = \|Z_2\| = 1$ and*

$$\begin{aligned} X &= \mu_X + \sigma(X) Z_1, \text{ and} \\ Y &= \mu_Y + \sigma(Y)[\cos\theta \cdot Z_1 + \sin\theta \cdot Z_2], \end{aligned}$$

where $0 \le \theta \le \pi$ is chosen such that $\cos\theta := Corr(X,Y)$.

Proof. (Just sketch the main ideal in class!). The proof amounts to applying the Gram-Schmidt procedure to $\{\bar{X} := X - \mu_X, \bar{Y} := Y - \mu_Y\}$ to find Z_1 and Z_2 followed by expressing X and Y in uniquely in terms of the linearly independent set, $\{1, Z_1, Z_2\}$. The details follow.

Performing Gram-Schmidt on $\{\bar{X}, \bar{Y}\}$ gives $Z_1 = \bar{X}/\sigma(X)$ and

$$\tilde{Z}_2 = \bar{Y} - \frac{(\bar{Y}, \bar{X})}{\sigma(X)^2}\bar{X}.$$

To get Z_2 we need to normalize $\tilde{Z}_2$ using;

$$\begin{aligned} \mathbb{E}\tilde{Z}_2^2 &= \sigma(Y)^2 - 2\frac{(\bar{Y}, \bar{X})}{\sigma(X)^2}(\bar{X}, \bar{Y}) + \frac{(\bar{Y}, \bar{X})^2}{\sigma(X)^4}\sigma(X)^2 \\ &= \sigma(Y)^2 - \frac{(\bar{X}, \bar{Y})^2}{\sigma(X)^2} = \sigma(Y)^2\left(1 - Corr^2(X,Y)\right) \\ &= \sigma(Y)^2 \sin^2\theta. \end{aligned}$$

Therefore $Z_1 = \bar{X}/\sigma(X)$ and

$$\begin{aligned} Z_2 := \frac{\tilde{Z}_2}{\left\|\tilde{Z}_2\right\|} &= \frac{\bar{Y} - \frac{(\bar{Y},\bar{X})}{\sigma(X)^2}\bar{X}}{\sigma(Y)\sin\theta} = \frac{\bar{Y} - \frac{\sigma(X)\sigma(Y)Corr(X,Y)}{\sigma(X)^2}\bar{X}}{\sigma(Y)\sin\theta} \\ &= \frac{\bar{Y} - \frac{\sigma(Y)}{\sigma(X)}\cos\theta \cdot \bar{X}}{\sigma(Y)\sin\theta} = \frac{\bar{Y} - \sigma(Y)\cos\theta \cdot Z_1}{\sigma(Y)\sin\theta} \end{aligned}$$

Solving for $\bar{X}$ and $\bar{Y}$ shows,

$$\bar{X} = \sigma(X) Z_1 \text{ and } \bar{Y} = \sigma(Y)[\sin\theta \cdot Z_2 + \cos\theta \cdot Z_1]$$

which is equivalent to the desired result. ∎

Remark 4.6. It is easy to give a second proof of Corollary 4.4 based on Theorem 4.5. Indeed, if $Corr(X,Y) = 1$, then $\theta = 0$ and $\bar{Y} = \sigma(Y) Z_1 = \frac{\sigma(Y)}{\sigma(X)}\bar{X}$ while if $Corr(X,Y) = -1$, then $\theta = \pi$ and therefore $\bar{Y} = -\sigma(Y) Z_1 = -\frac{\sigma(Y)}{\sigma(X)}\bar{X}$.

Exercise 4.1 (A correlation inequality). Suppose that X is a random variable and $f, g : \mathbb{R} \to \mathbb{R}$ are two increasing functions such that both $f(X)$ and $g(X)$ are square integrable, i.e. $\mathbb{E}|f(X)|^2 + \mathbb{E}|g(X)|^2 < \infty$. Show $\operatorname{Cov}(f(X), g(X)) \ge 0$. **Hint:** let Y be another random variable which has the same law as X and is independent of X. Then consider

$$\mathbb{E}[(f(Y) - f(X)) \cdot (g(Y) - g(X))].$$

5

Conditional Expectation

Notation 5.1 (Conditional Expectation 1) *Given $Y \in L^1(P)$ and $A \subset \Omega$ let*

$$\mathbb{E}[Y:A] := \mathbb{E}[1_A Y]$$

and

$$\mathbb{E}[Y|A] = \begin{cases} \mathbb{E}[Y:A]/P(A) & \text{if } P(A) > 0 \\ 0 & \text{if } P(A) = 0 \end{cases}. \tag{5.1}$$

(In point of fact, when $P(A) = 0$ we could set $\mathbb{E}[Y|A]$ to be any real number. We choose 0 for definiteness and so that $Y \to \mathbb{E}[Y|A]$ is always linear.)

Example 5.2 (Conditioning for the uniform distribution). Suppose that Ω is a finite set and P is the uniform distribution on P so that $P(\{\omega\}) = \frac{1}{\#(\Omega)}$ for all $\omega \in W$. Then for non-empty any subset $A \subset \Omega$ and $Y : \Omega \to \mathbb{R}$ we have $\mathbb{E}[Y|A]$ is the expectation of Y restricted to A under the uniform distribution on A. Indeed,

$$\begin{aligned} \mathbb{E}[Y|A] &= \frac{1}{P(A)}\mathbb{E}[Y:A] = \frac{1}{P(A)}\sum_{\omega\in A} Y(\omega) P(\{\omega\}) \\ &= \frac{1}{\#(A)/\#(\Omega)}\sum_{\omega\in A} Y(\omega)\frac{1}{\#(\Omega)} = \frac{1}{\#(A)}\sum_{\omega\in A} Y(\omega). \end{aligned}$$

Lemma 5.3. *If $P(A) > 0$ then $\mathbb{E}[Y|A] = \mathbb{E}_{P(\cdot|A)}Y$ for all $Y \in L^1(P)$.*

Proof. I will only prove this lemma when Y is a discrete random variable, although the result does hold in general. So suppose that $Y : \Omega \to S$ where S is a finite or countable subset of $\mathbb{R}$. Then taking expectation relative to $P(\cdot|A)$ of the identity, $Y = \sum_{y\in S} y 1_{Y=y}$, gives

$$\begin{aligned} \mathbb{E}_{P(\cdot|A)}Y &= \mathbb{E}_{P(\cdot|A)}\sum_{y\in S} y1_{Y=y} = \sum_{y\in S} y\mathbb{E}_{P(\cdot|A)}1_{Y=y} = \sum_{y\in S} yP(Y=y|A) \\ &= \sum_{y\in S} yP(Y=y|A) = \sum_{y\in S} y\frac{P(Y=y,A)}{P(A)} = \frac{1}{P(A)}\sum_{y\in S} y\mathbb{E}[1_A 1_{Y=y}] \\ &= \frac{1}{P(A)}\mathbb{E}\left[1_A\sum_{y\in S} y1_{Y=y}\right] = \frac{1}{P(A)}\mathbb{E}[1_A Y] = \mathbb{E}[Y|A]. \end{aligned}$$

■

Lemma 5.4. *No matter whether $P(A) > 0$ or $P(A) = 0$ we always have,*

$$|\mathbb{E}[Y|A]| \le \mathbb{E}[|Y||A] \le \sqrt{\mathbb{E}\left[|Y|^2|A\right]}. \tag{5.2}$$

Proof. If $P(A) = 0$ then all terms in Eq. (5.2) are zero and so the inequalities hold. For $P(A) > 0$ we have, using the Schwarz inequality in Theorem 3.4), that

$$|\mathbb{E}[Y|A]| = \left|\mathbb{E}_{P(\cdot|A)}Y\right| \le \mathbb{E}_{P(\cdot|A)}|Y| \le \sqrt{\mathbb{E}_{P(\cdot|A)}|Y|^2 \cdot \mathbb{E}_{P(\cdot|A)}1} = \sqrt{\mathbb{E}_{P(\cdot|A)}|Y|^2}.$$

This completes that proof as $\mathbb{E}_{P(\cdot|A)}|Y| = \mathbb{E}[|Y||A]$ and $\mathbb{E}_{P(\cdot|A)}|Y|^2 = \mathbb{E}\left[|Y|^2|A\right]$. ■

Notation 5.5 *Let S be a set (often $S = \mathbb{R}$ or $S = \mathbb{R}^N$) and suppose that $X : \Omega \to S$ is a function. (So X is a random variable if $S = \mathbb{R}$ and a random vector when $S = \mathbb{R}^N$.) Further let V_X denote those random variables $Z \in L^2(P)$ which may be written as $Z = f(X)$ for some function $f : S \to \mathbb{R}$. (This is a subspace of $L^2(P)$ and we let $\mathcal{F}_X := \{f : S \to \mathbb{R} : f(X) \in L^2(P)\}$.)*

Definition 5.6 (Conditional Expectation 2). *Given a function $X : \Omega \to S$ and $Y \in L^2(P)$, we define $\mathbb{E}[Y|X] := Q_{V_X}Y$ where Q_{V_X} is orthogonal projection onto V_X. (**Fact:** $Q_{V_X}Y$ always exists. The proof requires technical details beyond the scope of this course.)*

Remark 5.7. By definition, $\mathbb{E}[Y|X] = h(X)$ where $h \in \mathcal{F}_X$ is chosen so that $[Y - h(X)] \perp V_X$, i.e. $\mathbb{E}[Y|X] = h(X)$ iff $(Y - h(X), f(X)) = 0$ for all $f \in \mathcal{F}_X$. So in summary, $\mathbb{E}[Y|X] = h(X)$ iff

$$\mathbb{E}[Yf(X)] = \mathbb{E}[h(X)f(X)] \text{ for all } f \in \mathcal{F}_X. \tag{5.3}$$

Corollary 5.8 (Law of total expectation). *For all random variables $Y \in L^2(P)$, we have $\mathbb{E}Y = \mathbb{E}(\mathbb{E}(Y|X))$.*

Proof. Take $f = 1$ in Eq. (5.3). ■

This notion of conditional expectation is rather abstract. It is now time to see how to explicitly compute conditional expectations. (In general this can be quite tricky to carry out in concrete examples!)

5.1 Conditional Expectation for Discrete Random Variables

Recall that if A and B are events with $P(A) > 0$, then we define $P(B|A) := \frac{P(B\cap A)}{P(A)}$. By convention we will set $P(B|A) = 0$ if $P(A) = 0$.

Example 5.9. If Ω is a finite set with N elements, P is the uniform distribution on Ω, and A is a non-empty subset of Ω, then $P(\cdot|A)$ restricted to events contained in A is the uniform distribution on A. Indeed, $a = \#(A)$ and $B \subset A$, we have

$$P(B|A) = \frac{P(B\cap A)}{P(A)} = \frac{P(B)}{P(A)} = \frac{\#(B)/N}{\#(A)/N} = \frac{\#(B)}{\#(A)} = \frac{\#(B)}{a}.$$

Theorem 5.10. *Suppose that S is a finite or countable set and $X : \Omega \to S$, then $\mathbb{E}[Y|X] = h(X)$ where $h(s) := \mathbb{E}[Y|X = s]$ for all $s \in S$.*

Proof. First Proof. Our goal is to find $h(s)$ such that

$$\mathbb{E}[Yf(X)] = \mathbb{E}[h(X)f(X)] \text{ for all bounded } f.$$

Let $S' = \{s \in S : P(X = s) > 0\}$, then

$$\begin{aligned}\mathbb{E}[Yf(X)] &= \sum_{s\in S}\mathbb{E}[Yf(X) : X = s] = \sum_{s\in S'}\mathbb{E}[Yf(X) : X = s]\\ &= \sum_{s\in S'} f(s)\,\mathbb{E}[Y|X = s]\cdot P(X = s)\\ &= \sum_{s\in S'} f(s)\,h(s)\cdot P(X = s)\\ &= \sum_{s\in S} f(s)\,h(s)\cdot P(X = s) = \mathbb{E}[h(X)f(X)]\end{aligned}$$

where $h(s) := \mathbb{E}[Y|X = s]$.

Second Proof. If S is a finite set, such that $P(X = s) > 0$ for all $s \in S$. Then

$$f(X) = \sum_{s\in S} f(s)\,1_{X=s}$$

which shows that $V_X = \operatorname{span}\{1_{X=s} : s \in S\}$. As $\{1_{X=s}\}_{s\in S}$ is an orthogonal set, we may compute

$$\begin{aligned}\mathbb{E}[Y|X] &= \sum_{s\in S}\frac{(Y, 1_{X=s})}{\|1_{X=s}\|^2}1_{X=s} = \sum_{s\in S}\frac{\mathbb{E}[Y : X = s]}{P(X = s)}1_{X=s}\\ &= \sum_{s\in S}\mathbb{E}[Y|X = s]\cdot 1_{X=s} = h(X).\end{aligned}$$

■

Example 5.11. Suppose that X and Y are discrete random variables with joint distribution given as;

	ρ_Y	1/4	$\frac{1}{2}$	1/4
ρ_X	$X\backslash Y$	−1	0	1
1/4	1	0	1/4	0
3/4	0	1/4	1/4	1/4

.

We then have

$$\mathbb{E}[Y|X = 1] = \frac{1}{1/4}\left(-1\cdot 0 + 0\cdot\frac{1}{4} + 1\cdot 0\right) = 0 \text{ and}$$
$$\mathbb{E}[Y|X = 0] = \frac{1}{3/4}\left(-1\cdot 1/4 + 0\cdot\frac{1}{4} + 1\cdot 1/4\right) = 0$$

and therefore $\mathbb{E}[Y|X] = 0$. On the other hand,

$$\mathbb{E}[X|Y = -1] = \frac{1}{1/4}\left(1\cdot 0 + 0\cdot\frac{1}{4}\right) = 0,$$
$$\mathbb{E}[X|Y = 0] = \frac{1}{1/2}\left(1\cdot 1/4 + 0\cdot\frac{1}{4}\right) = \frac{1}{2}, \text{ and}$$
$$\mathbb{E}[X|Y = 1] = \frac{1}{1/4}\left(1\cdot 0 + 0\cdot\frac{1}{4}\right) = 0.$$

Therefore

$$\mathbb{E}[X|Y] = \frac{1}{2}1_{Y=0}.$$

Example 5.12. Let X and Y be discrete random variables with values in $\{1, 2, 3\}$ whose joint distribution and marginals are given by

	ρ_X	.3	.35	.35
ρ_Y	$Y\backslash X$	1	2	3
.6	1	.1	.2	.3
.3	2	.15	.15	0
.1	3	.05	0	.05

.

Then

$$\rho_{X|Y}(1, 3) = P(X = 1|Y = 3) = \frac{.05}{.1} = \frac{1}{2},$$
$$\rho_{X|Y}(2, 3) = P(X = 2|Y = 3) = \frac{0}{.1} = 0, \text{ and}$$
$$\rho_{X|Y}(3, 3) = P(X = 3|Y = 3) = \frac{.05}{.1} = \frac{1}{2}.$$

Therefore,

$$\mathbb{E}[X|Y=3] = 1\cdot\frac{1}{2} + 2\cdot 0 + 3\cdot\frac{1}{2} = 2$$

or

$$h(3) := \mathbb{E}[X|Y=3] = \frac{1}{.1}(1\cdot.05 + 2\cdot 0 + 3\cdot.05) = 2$$

Similarly,

$$h(1) := \mathbb{E}[X|Y=1] = \frac{1}{.6}(1\cdot.1 + 2\cdot.2 + 3\cdot.3) = 2\frac{1}{3},$$
$$h(2) := \mathbb{E}[X|Y=2] = \frac{1}{.3}(1\cdot.15 + 2\cdot.15 + 3\cdot 0) = 1.5$$

and so

$$\mathbb{E}[X|Y] = h(Y) = 2\frac{1}{3}\cdot 1_{Y=1} + 1.5\cdot 1_{Y=2} + 2\cdot 1_{Y=3}.$$

Example 5.13 (Number of girls in a family). Suppose the number of children in a family is a random variable X with mean μ, and given $X=n$ for $n\geq 1$, each of the n children in the family is a girl with probability p and a boy with probability $1-p$. Problem. What is the expected number of girls in a family?

Solution. Intuitively, the answer should be $p\mu$. To show this is correct let G be the random number of girls in a family. Then,

$$\mathbb{E}[G|X=n] = p\cdot n$$

as $G = 1_{A_1} + \cdots + 1_{A_n}$ on $X=n$ where A_i is the event the i^{th} – child is a girl. We are given $P(A_i|X=n) = p$ so that $\mathbb{E}[1_{A_i}|X=n] = p$ and so $\mathbb{E}[G|X=n] = p\cdot n$. Therefore, $\mathbb{E}[G|X] = p\cdot X$ and

$$\mathbb{E}[G] = \mathbb{E}\mathbb{E}[G|X] = \mathbb{E}[p\cdot X] = p\mu.$$

Example 5.14. Suppose that X and Y are i.i.d. random variables with the geometric distribution,

$$P(X=k) = P(Y=k) = (1-p)^{k-1}p \text{ for } k\in\mathbb{N}.$$

We compute, for $n>m$,

$$P(X=m|X+Y=n) = \frac{P(X=m, X+Y=n)}{P(X+Y=n)} = \frac{P(X=m, Y=n-m)}{\sum_{k+l=n}P(X=k,Y=l)}$$

where

$$P(X=m, Y=n-m) = p^2(1-p)^{m-1}(1-p)^{n-m-1} = p^2(1-p)^{n-2}$$

and

$$\sum_{k+l=n}P(X=k,Y=l) = \sum_{k+l=n}(1-p)^{k-1}p(1-p)^{l-1}p = \sum_{k+l=n}p^2(1-p)^{n-2} = p^2(1-p)^{n-2}\sum_{k=1}^{n-1}1.$$

Thus we have shown,

$$P(X=m|X+Y=n) = \frac{1}{n-1} \text{ for } 1\leq m<n.$$

From this it follows that

$$\mathbb{E}[f(X)|X+Y=n] = \frac{1}{n-1}\sum_{m=1}^{n-1}f(m)$$

and so

$$\mathbb{E}[f(X)|X+Y] = \frac{1}{X+Y-1}\sum_{m=1}^{X+Y-1}f(m).$$

As a check if $f(m)=m$ we have

$$\mathbb{E}[X|X+Y] = \frac{1}{X+Y-1}\sum_{m=1}^{X+Y-1}m = \frac{1}{X+Y-1}\frac{1}{2}(X+Y-1)(X+Y-1+1) = \frac{1}{2}(X+Y)$$

as we will see hold in fair generality, see Example 5.24 below.

Example 5.15 (Durrett Example 4.6.2, p. 205). Suppose we want to determine the expected value of

$$Y = \#\text{ of rolls to complete one game of craps.}$$

Let X be the sum we obtain on the first roll. In this game, if;

$$X\in\{2,3,12\} =: L \implies \text{game ends and you loose,}$$
$$X\in\{7,11\} =: W \implies \text{game ends and you win,and}$$
$$X\in\{4,5,6,8,9,10\} =: P \implies X \text{ is your "point."}$$

In the last case, you roll your dice again and again until you either throw until you get X (your point) or 7. (If you hit X before the 7 then you win.) We are going to compute $\mathbb{E}Y$ as $\mathbb{E}[\mathbb{E}[Y|X]]$.

Clearly if $x \in L \cup W$ then $\mathbb{E}[Y|X=x] = 1$ while if $x \in P$, then $\mathbb{E}[Y|X=x] = 1 + \mathbb{E}N_x$ where N_x is the number of rolls need to hit either x or 7. This is a geometric random variable with parameter p_x (probability of rolling an x or a 7) and so $\mathbb{E}N_x = \frac{1}{p_x}$. For example if $x=4$, then $p_x = \frac{3+6}{36} = \frac{9}{36}$ (3 is the number of ways to roll a 4 and 6 is the number of ways to roll as 7) and hence $1 + \mathbb{E}N_x = 1 + 4 = 5$. Similar calculations gives us the following table;

$$\begin{array}{ccccc} x \in & \{2,3,7,11,12\} & \{4,10\} & \{5,9\} & \{6,8\} \\ \mathbb{E}[Y|X=x] & 1 & \frac{45}{9} & \frac{46}{10} & \frac{47}{11} \\ P(\text{set}) & \frac{12}{36} & \frac{6}{36} & \frac{8}{36} & \frac{10}{36} \end{array}.$$

(For example, there are 5 ways to get a 6 and 6 ways to get a 7 so when $x=6$ we are waiting for an event with probability $11/36$ and the mean of this geometric random variables is $36/11$ and adding the first roll to this implies, $\mathbb{E}[Y|X=6] = 47/11$. Similarly for $x=8$ and $P(X=6 \text{ or } 8) = (5+5)/36$.) Putting the pieces together and using the law of total expectation gives,

$$\begin{aligned} \mathbb{E}Y = \mathbb{E}[\mathbb{E}[Y|X]] &= 1 \cdot \frac{12}{36} + \frac{45}{9} \cdot \frac{6}{36} + \frac{46}{10} \cdot \frac{8}{36} + \frac{47}{11} \cdot \frac{10}{36} \\ &= \frac{557}{165} \cong 3.376 \text{ rolls.} \end{aligned}$$

The following two facts are often helpful when computing conditional expectations.

Proposition 5.16 (Bayes formula). *Suppose that $A \subset \Omega$ and $\{A_i\}$ is a partition of A, then*

$$\mathbb{E}[Y|A] = \frac{1}{P(A)} \sum_i \mathbb{E}[Y|A_i] P(A_i) = \frac{\sum_i \mathbb{E}[Y|A_i] P(A_i)}{\sum_i P(A_i)}.$$

If we further assume that $\mathbb{E}[Y|A_i] = c$ independent of i, then $\mathbb{E}[Y|A] = c$.

The proof of this proposition is straight forward and is left to the reader.

Proposition 5.17. *Suppose that $X_i : \Omega \to S_i$ for $1 \le i \le n$ are independent random functions with each S_i being discrete. Then for any $T_i \subset S_i$ we have,*

$$\mathbb{E}[u(X_1,\dots,X_n) | X_1 \in T_1, \dots, X_n \in T_n] = \mathbb{E}[u(Y_1,\dots,Y_n)]$$

where $Y_i : \Omega \to T_i$ for $1 \le i \le n$ are independent random functions such that $P(Y_i = t) = P(X_i = t | X_i \in T_i)$ for all $t \in T_i$.

Proof. The proof is contained in the following computation,

$$\begin{aligned} &\mathbb{E}[u(X_1,\dots,X_n) | X_1 \in T_1, \dots, X_n \in T_n] \\ &= \frac{\mathbb{E}[u(X_1,\dots,X_n) : X_1 \in T_1, \dots, X_n \in T_n]}{P(X_1 \in T_1, \dots, X_n \in T_n)} \\ &= \frac{1}{P(X_1 \in T_1, \dots, X_n \in T_n)} \sum_{t_i \in T_i} u(t_1,\dots,t_n) P(X_1 = t_1, \dots, X_n = t_n) \\ &= \frac{1}{\prod_i P(X_i \in T_i)} \sum_{(t_1,\dots,t_n) \in T_1 \times \cdots \times T_n} u(t_1,\dots,t_n) \prod_i P(X_i \in t_i) \\ &= \sum_{(t_1,\dots,t_n) \in T_1 \times \cdots \times T_n} u(t_1,\dots,t_n) \prod_i \frac{P(X_i \in t_i)}{P(X_i \in T_i)} \\ &= \sum_{(t_1,\dots,t_n) \in T_1 \times \cdots \times T_n} u(t_1,\dots,t_n) \prod_i P(X_i = t | X_i \in T_i) \\ &= \sum_{(t_1,\dots,t_n) \in T_1 \times \cdots \times T_n} u(t_1,\dots,t_n) P(Y_1 = t_1, \dots, Y_n = t_n) \\ &= \mathbb{E}[u(Y_1,\dots,Y_n)]. \end{aligned}$$

■

Here is an example of how to use these two propositions.

Example 5.18. Suppose we roll a die n – times with results $\{X_i\}_{i=1}^n$ where $X_i \in \{1,2,3,4,5,6\}$ for each i. Let

$$\begin{aligned} Y &= \sum_{i=1}^n 1_{\{1,3,5\}}(X_i) = \text{ number of odd rolls and} \\ Z &= \sum_{i=1}^n 1_{\{3,4,6\}}(X_i) \\ &= \text{ number of times 3, 4, or 6 are rolled.} \end{aligned}$$

We wish to compute $\mathbb{E}[Z|Y]$. So let $0 \le y \le n$ be given and let A be the event where X_i is odd for $1 \le i \le y$ and X_i is even for $y < i \le n$. Then

$$\mathbb{E}[Z|A] = y\frac{1}{3} + (n-y) \cdot \frac{2}{3}$$

where $\frac{1}{3} = P(X_1 \in \{3,4,6\} | X_1 \text{ is odd})$ and $\frac{2}{3} = P(X_1 \in \{3,4,6\} | X_1 \text{ is even})$. Now it is clear that $\{Y = y\}$ can be partitioned into events like the one above being labeled by which of the y – slots are even and the results are the same for all such choices by symmetry, therefore by Proposition 5.16 we may conclude

$$\mathbb{E}[Z|Y=y] = y\frac{1}{3} + (n-y)\cdot\frac{2}{3}$$

and therefore,

$$\mathbb{E}[Z|Y] = Y\frac{1}{3} + (n-Y)\cdot\frac{2}{3}.$$

As a check notice that

$$\begin{aligned}\mathbb{E}\mathbb{E}[Z|Y] &= \mathbb{E}Y\frac{1}{3} + (n-\mathbb{E}Y)\cdot\frac{2}{3} = \frac{n}{2}\frac{1}{3} + \left(n-\frac{n}{2}\right)\cdot\frac{2}{3}\\ &= \frac{n}{6}+\frac{n}{3} = \frac{1}{2}n = \mathbb{E}Z.\end{aligned}$$

The next lemma generalizes this result.

Lemma 5.19. *Suppose that $X_i : \Omega \to S$ for $1 \le i \le n$ are i.i.d. random functions into a discrete set S. Given a subset $A \subset S$ let*

$$Z_A := \sum_{i=1}^{n} 1_A(X_i) = \#(\{i : X_i \in A\}).$$

If B is another subset of S, then

$$\mathbb{E}[Z_A|Z_B] = Z_B \cdot P(X_1 \in A|X_1 \in B) + (n-Z_B)\cdot P(X_1 \in A|X_1 \notin B). \quad (5.4)$$

Proof. Intuitively, for a typical trial there are Z_B of the X_i in B and for these i we have $\mathbb{E}[1_A(X_i)|X_i \in B] = P(X_1 \in A|X_1 \in B)$. Likewise there are $n - Z_B$ of the X_i in $S \setminus B$ and for these i we have $\mathbb{E}[1_A(X_i)|X_i \notin B] = P(X_1 \in A|X_1 \notin B)$. On these grounds we are quickly lead to Eq. (5.4).

To prove Eq. (5.4) rigorously we will compute $\mathbb{E}[Z_A|Z_B = m]$ by partitioning $\{Z_B = m\}$ as $\cup Q_\Lambda$ where Λ runs through subsets of k elements of S and

$$Q_\Lambda = (\cap_{i\in\Lambda}\{X_i \in B\}) \cap (\cap_{i\in\Lambda^c}\{X_i \notin B\}).$$

Then according to Proposition 5.17,

$$\mathbb{E}[Z_A|Q_\Lambda] = \mathbb{E}\left[\sum_{i=1}^{n} 1_A(Y_i)\right]$$

where $\{Y_i\}$ are independent and

$$P(Y_i = s) = P(X_i = s|X_i \in B) = P(X_1 = s|X_1 \in B) \text{ for } i \in \Lambda$$

and

$$P(Y_i = s) = P(X_i = s|X_i \notin B) = P(X_1 = s|X_1 \notin B) \text{ for } i \notin \Lambda.$$

Therefore,

$$\begin{aligned}\mathbb{E}[Z_A|Q_\Lambda] &= \mathbb{E}\left[\sum_{i=1}^{n} 1_A(Y_i)\right] = \sum_{i=1}^{n}\mathbb{E}1_A(Y_i)\\ &= \sum_{i\in\Lambda} P(X_1 \in A|X_1 \in B) + \sum_{i\notin\Lambda} P(X_1 \in A|X_1 \notin B)\\ &= m\cdot P(X_1 \in A|X_1 \in B) + (n-m)\cdot P(X_1 \in A|X_1 \notin B).\end{aligned}$$

As the result is independent of the choice of Λ with $\#(\Lambda) = m$ we may use Proposition 5.16 to conclude that

$$\mathbb{E}[Z_A|Z_B = m] = m\cdot P(X_1 \in A|X_1 \in B) + (n-m)\cdot P(X_1 \in A|X_1 \notin B).$$

As $0 \le m \le n$ is arbitrary Eq. (5.4) follows.

As a check notice that $\mathbb{E}Z_A = n\cdot P(X_1 \in A)$ while

$$\begin{aligned}\mathbb{E}\mathbb{E}[Z_A|Z_B] =& \mathbb{E}Z_B\cdot P(X_1 \in A|X_1 \in B) + \mathbb{E}(n-Z_B)\cdot P(X_1 \in A|X_1 \notin B)\\ =& n\cdot P(X_1 \in B)\cdot P(X_1 \in A|X_1 \in B)\\ &+ (n - n\cdot P(X_1 \in B))\cdot P(X_1 \in A|X_1 \notin B)\\ =& n\cdot\begin{bmatrix} P(X_1 \in B)\cdot P(X_1 \in A|X_1 \in B)\\ +(1-P(X_1 \in B))\cdot P(X_1 \in A|X_1 \notin B)\end{bmatrix}\\ =& n\cdot[P(X_1 \in A|X_1 \in B)P(X_1 \in B) + P(X_1 \in A|X_1 \notin B)P(X_1 \notin B)]\\ =& n\cdot[P(X_1 \in A, X_1 \in B) + P(X_1 \in A, X_1 \notin B)]\\ =& n\cdot P(X_1 \in A) = \mathbb{E}Z_A.\end{aligned}$$

■

5.2 General Properties of Conditional Expectation

Let us pause for a moment to record a few basic general properties of conditional expectations.

Proposition 5.20 (Contraction Property). *For all $Y \in L^2(P)$, we have $\mathbb{E}|\mathbb{E}[Y|X]| \le \mathbb{E}|Y|$. Moreover if $Y \ge 0$ then $\mathbb{E}[Y|X] \ge 0$ (a.s.).*

Proof. Let $\mathbb{E}[Y|X] = h(X)$ (with $h : S \to \mathbb{R}$) and then define

$$f(x) = \begin{cases} 1 & \text{if } h(x) \ge 0\\ -1 & \text{if } h(x) < 0\end{cases}.$$

Since $h(x)f(x) = |h(x)|$, it follows from Eq. (5.3) that

$$\mathbb{E}\left[|h(X)|\right] = \mathbb{E}\left[Yf(X)\right] = \left|\mathbb{E}\left[Yf(X)\right]\right| \le \mathbb{E}\left[|Yf(X)|\right] = \mathbb{E}|Y|.$$

For the second assertion take $f(x) = 1_{h(x)<0}$ in Eq. (5.3) in order to learn

$$\mathbb{E}\left[h(X)\,1_{h(X)<0}\right] = \mathbb{E}\left[Y1_{h(X)<0}\right] \ge 0.$$

As $h(X)\,1_{h(X)<0} \le 0$ we may conclude that $h(X)\,1_{h(X)<0} = 0$ a.s. ∎

Because of this proposition we may extend the notion of conditional expectation to $Y \in L^1(P)$ as stated in the following theorem which we do not bother to prove here.

Theorem 5.21. *Given $X : \Omega \to S$ and $Y \in L^1(P)$, there exists an "essentially unique" function $h : S \to \mathbb{R}$ such that Eq. (5.3) holds for all bounded functions, $f : S \to \mathbb{R}$. (As above we write $\mathbb{E}[Y|X]$ for $h(X)$.) Moreover the contraction property, $\mathbb{E}\left|\mathbb{E}[Y|X]\right| \le \mathbb{E}|Y|$, still holds.*

Theorem 5.22 (Basic properties). *Let Y, Y_1, and Y_2 be integrable random variables and $X : \Omega \to S$ be given. Then:*

1. $\mathbb{E}(Y_1 + Y_2|X) = \mathbb{E}(Y_1|X) + \mathbb{E}(Y_2|X)$.
2. $\mathbb{E}(aY|X) = a\mathbb{E}(Y|X)$ *for all constants* a.
3. $\mathbb{E}(g(X)Y|X) = g(X)\mathbb{E}(Y|X)$ *for all bounded functions* g.
4. $\mathbb{E}(\mathbb{E}(Y|X)) = \mathbb{E}Y$. ***(Law of total expectation.)***
5. *If Y and X are independent then* $\mathbb{E}(Y|X) = \mathbb{E}Y$.

Proof. 1. Let $h_i(X) = \mathbb{E}[Y_i|X]$, then for all bounded f,

$$\mathbb{E}\left[Y_1 f(X)\right] = \mathbb{E}\left[h_1(X) f(X)\right] \text{ and}$$
$$\mathbb{E}\left[Y_2 f(X)\right] = \mathbb{E}\left[h_2(X) f(X)\right]$$

and therefore adding these two equations together implies

$$\begin{aligned}\mathbb{E}\left[(Y_1+Y_2) f(X)\right] &= \mathbb{E}\left[(h_1(X)+h_2(X)) f(X)\right] \\ &= \mathbb{E}\left[(h_1+h_2)(X) f(X)\right] \\ \mathbb{E}\left[Y_2 f(X)\right] &= \mathbb{E}\left[h_2(X) f(X)\right]\end{aligned}$$

for all bounded f. Therefore we may conclude that

$$\mathbb{E}(Y_1+Y_2|X) = (h_1+h_2)(X) = h_1(X) + h_2(X) = \mathbb{E}(Y_1|X) + \mathbb{E}(Y_2|X).$$

2. The proof is similar to 1 but easier and so is omitted.

3. Let $h(X) = \mathbb{E}[Y|X]$, then $\mathbb{E}\left[Yf(X)\right] = \mathbb{E}\left[h(X) f(X)\right]$ for all bounded functions f. Replacing f by $g \cdot f$ implies

$$\mathbb{E}\left[Yg(X) f(X)\right] = \mathbb{E}\left[h(X) g(X) f(X)\right] = \mathbb{E}\left[(h\cdot g)(X) f(X)\right]$$

for all bounded functions f. Therefore we may conclude that

$$\mathbb{E}\left[Yg(X)|X\right] = (h\cdot g)(X) = h(X) g(X) = g(X)\,\mathbb{E}(Y|X).$$

4. Take $f \equiv 1$ in Eq. (5.3).

5. If X and Y are independent and $\mu := \mathbb{E}[Y]$, then

$$\mathbb{E}\left[Yf(X)\right] = \mathbb{E}[Y]\,\mathbb{E}\left[f(X)\right] = \mu\mathbb{E}\left[f(X)\right] = \mathbb{E}\left[\mu f(X)\right]$$

from which it follows that $\mathbb{E}[Y|X] = \mu$ as desired. ∎

The next theorem says that conditional expectations essentially only depends on the distribution of (X, Y) and nothing else.

Theorem 5.23. *Suppose that (X,Y) and $\left(\tilde{X},\tilde{Y}\right)$ are random vectors such that $(X,Y) \overset{d}{=} \left(\tilde{X},\tilde{Y}\right)$, i.e. $\mathbb{E}\left[f(X,Y)\right] = \mathbb{E}\left[f\left(\tilde{X},\tilde{Y}\right)\right]$ for all bounded (or non-negative) functions f. If $h(X) = \mathbb{E}\left[u(X,Y)|X\right]$, then $\mathbb{E}\left[u\left(\tilde{X},\tilde{Y}\right)|\tilde{X}\right] = h\left(\tilde{X}\right)$.*

Proof. By assumption we know that

$$\mathbb{E}\left[u(X,Y) f(X)\right] = \mathbb{E}\left[h(X) f(X)\right] \text{ for all bounded } f.$$

Since $(X,Y) \overset{d}{=} \left(\tilde{X},\tilde{Y}\right)$, this is equivalent to

$$\mathbb{E}\left[u\left(\tilde{X},\tilde{Y}\right) f\left(\tilde{X}\right)\right] = \mathbb{E}\left[h\left(\tilde{X}\right) f\left(\tilde{X}\right)\right] \text{ for all bounded } f$$

which is equivalent to $\mathbb{E}\left[u\left(\tilde{X},\tilde{Y}\right)|\tilde{X}\right] = h\left(\tilde{X}\right)$. ∎

Example 5.24. Let $\{X_i\}_{i=1}^{\infty}$ be i.i.d. random variables with $\mathbb{E}|X_i| < \infty$ for all i and let $S_m := X_1 + \cdots + X_m$ for $m = 1, 2, \ldots$. We wish to show,

$$\mathbb{E}\left[S_m|S_n\right] = \frac{m}{n} S_n \text{ for all } m \le n.$$

for all $m \le n$. To prove this first observe by symmetry[1] that

$$\mathbb{E}\left(X_i|S_n\right) = h(S_n) \text{ independent of } i.$$

Therefore

$$S_n = \mathbb{E}\left(S_n|S_n\right) = \sum_{i=1}^{n} \mathbb{E}\left(X_i|S_n\right) = \sum_{i=1}^{n} h(S_n) = n\cdot h(S_n).$$

[1] Apply Theorem 5.23 using $(X_1, S_n) \overset{d}{=} (X_i, S_n)$ for $1 \le i \le n$.

Thus we see that

$$\mathbb{E}\left(X_i|S_n\right) = \frac{1}{n}S_n$$

and therefore

$$\mathbb{E}\left(S_m|S_n\right) = \sum_{i=1}^{m}\mathbb{E}\left(X_i|S_n\right) = \sum_{i=1}^{m}\frac{1}{n}S_n = \frac{m}{n}S_n.$$

If $m > n$, then $S_m = S_n + X_{n+1} + \cdots + X_m$. Since X_i is independent of S_n for $i > n$, it follows that

$$\begin{aligned}\mathbb{E}\left(S_m|S_n\right) &= \mathbb{E}\left(S_n + X_{n+1} + \cdots + X_m|S_n\right)\\ &= \mathbb{E}\left(S_n|S_n\right) + \mathbb{E}\left(X_{n+1}|S_n\right) + \cdots + \mathbb{E}\left(X_m|S_n\right)\\ &= S_n + (m-n)\,\mu \text{ if } m \geq n\end{aligned}$$

where $\mu = \mathbb{E}X_i$.

Example 5.25 (See Durrett, #8, p. 213). Suppose that X and Y are two integrable random variables such that

$$\mathbb{E}\left[X|Y\right] = 18 - \frac{3}{5}Y \text{ and } \mathbb{E}\left[Y|X\right] = 10 - \frac{1}{3}X.$$

We would like to find $\mathbb{E}X$ and $\mathbb{E}Y$. To do this we use the law of total expectation to find,

$$\begin{aligned}\mathbb{E}X &= \mathbb{E}\mathbb{E}\left[X|Y\right] = \mathbb{E}\left(18 - \frac{3}{5}Y\right) = 18 - \frac{3}{5}\mathbb{E}Y \text{ and}\\ \mathbb{E}Y &= \mathbb{E}\mathbb{E}\left[Y|X\right] = \mathbb{E}\left(10 - \frac{1}{3}X\right) = 10 - \frac{1}{3}\mathbb{E}X.\end{aligned}$$

Solving this pair of linear equations shows $\mathbb{E}X = 15$ and $\mathbb{E}Y = 5$.

5.3 Conditional Expectation for Continuous Random Variables

(This section will be covered later in the course when first needed.)

Suppose that Y and X are continuous random variables which have a joint density, $\rho_{(Y,X)}\left(y,x\right)$. Then by definition of $\rho_{(Y,X)}$, we have, for all bounded or non-negative, U, that

$$\mathbb{E}\left[U\left(Y,X\right)\right] = \int\int U\left(y,x\right)\rho_{(Y,X)}\left(y,x\right)dydx. \tag{5.5}$$

The marginal density associated to X is then given by

$$\rho_X\left(x\right) := \int \rho_{(Y,X)}\left(y,x\right)dy \tag{5.6}$$

and recall from Math 180A that the conditional density $\rho_{(Y|X)}\left(y,x\right)$ is defined by

$$\rho_{(Y|X)}\left(y,x\right) = \begin{cases}\frac{\rho_{(Y,X)}(y,x)}{\rho_X(x)} & \text{if } \rho_X\left(x\right) > 0\\ 0 & \text{if } \rho_X\left(x\right) = 0\end{cases}. \tag{5.7}$$

Observe that if $\rho_{(Y,X)}\left(y,x\right)$ is continuous, then

$$\rho_{(Y,X)}\left(y,x\right) = \rho_{(Y|X)}\left(y,x\right)\rho_X\left(x\right) \text{ for all } (x,y). \tag{5.8}$$

Indeed, if $\rho_X\left(x\right) = 0$, then

$$0 = \rho_X\left(x\right) = \int \rho_{(Y,X)}\left(y,x\right)dy$$

from which it follows that $\rho_{(Y,X)}\left(y,x\right) = 0$ for all y. If $\rho_{(Y,X)}$ is not continuous, Eq. (5.8) still holds for "a.e." (x,y) which is good enough.

Lemma 5.26. *In the notation above,*

$$\rho\left(x,y\right) = \rho_{(Y|X)}\left(y,x\right)\rho_X\left(x\right) \text{ for a.e. } (x,y). \tag{5.9}$$

Proof. By definition Eq. (5.9) holds when $\rho_X\left(x\right) > 0$ and $\rho\left(x,y\right) \geq \rho_{(Y|X)}\left(y,x\right)\rho_X\left(x\right)$ for all (x,y). Moreover,

$$\begin{aligned}\int\int \rho_{(Y|X)}\left(y,x\right)\rho_X\left(x\right)dxdy &= \int\int \rho_{(Y|X)}\left(y,x\right)\rho_X\left(x\right)1_{\rho_X(x)>0}dxdy\\ &= \int\int \rho\left(x,y\right)1_{\rho_X(x)>0}dxdy\\ &= \int \rho_X\left(x\right)1_{\rho_X(x)>0}dx = \int \rho_X\left(x\right)dx\\ &= 1 = \int\int \rho\left(x,y\right)dxdy,\end{aligned}$$

or equivalently,

$$\int\int \left[\rho\left(x,y\right) - \rho_{(Y|X)}\left(y,x\right)\rho_X\left(x\right)\right]dxdy = 0$$

which implies the result. ∎

Theorem 5.27. *Keeping the notation above,for all or all bounded or non-negative, U, we have $\mathbb{E}[U(Y,X)|X] = h(X)$ where*

$$h(x) = \int U(y,x)\,\rho_{(Y|X)}(y,x)\,dy \tag{5.10}$$

$$= \begin{cases} \frac{\int U(y,x)\rho_{(Y,X)}(y,x)dy}{\int \rho_{(Y,X)}(y,x)dy} & \text{if} \quad \int \rho_{(Y,X)}(y,x)\,dy > 0 \\ 0 & \text{otherwise} \end{cases}. \tag{5.11}$$

In the future we will usually denote $h(x)$ informally by $\mathbb{E}[U(Y,x)|X=x]$,[2] so that

$$\mathbb{E}[U(Y,x)|X=x] := \int U(y,x)\,\rho_{(Y|X)}(y,x)\,dy. \tag{5.12}$$

Proof. We are looking for $h : S \to \mathbb{R}$ such that

$$\mathbb{E}[U(Y,X)\,f(X)] = \mathbb{E}[h(X)\,f(X)] \text{ for all bounded } f.$$

Using Lemma 5.26, we find

$$\begin{aligned}
\mathbb{E}[U(Y,X)\,f(X)] &= \int\int U(y,x)\,f(x)\,\rho_{(Y,X)}(y,x)\,dydx \\
&= \int\int U(y,x)\,f(x)\,\rho_{(Y|X)}(y,x)\,\rho_X(x)\,dydx \\
&= \int \left[\int U(y,x)\,\rho_{(Y|X)}(y,x)\,dy\right] f(x)\,\rho_X(x)\,dx \\
&= \int h(x)\,f(x)\,\rho_X(x)\,dx \\
&= \mathbb{E}[h(X)\,f(X)]
\end{aligned}$$

where h is given as in Eq. (5.10). ∎

Example 5.28 (Durrett 8.15, p. 145). Suppose that X and Y have joint density $\rho(x,y) = 8xy \cdot 1_{0<y<x<1}$. We wish to compute $\mathbb{E}[u(X,Y)|Y]$. To this end we compute

$$\rho_Y(y) = \int_{\mathbb{R}} 8xy \cdot 1_{0<y<x<1}dx = 8y\int_{x=y}^{x=1} x\cdot dx = 8y\cdot\frac{x^2}{2}|_y^1 = 4y\cdot\left(1-y^2\right).$$

Therefore,

$$\rho_{X|Y}(x,y) = \frac{\rho(x,y)}{\rho_Y(y)} = \frac{8xy\cdot 1_{0<y<x<1}}{4y\cdot(1-y^2)} = \frac{2x\cdot 1_{0<y<x<1}}{(1-y^2)}$$

and so

$$\mathbb{E}[u(X,Y)|Y=y] = \int_{\mathbb{R}} \frac{2x\cdot 1_{0<y<x<1}}{(1-y^2)}u(x,y)\,dx = 2\frac{1_{0<y<1}}{1-y^2}\int_y^1 u(x,y)\;xdx$$

and so

$$\mathbb{E}[u(X,Y)|Y] = 2\frac{1}{1-Y^2}\int_Y^1 u(x,Y)\,xdx.$$

is the best approximation to $u(X,Y)$ be a function of Y alone.

Proposition 5.29. *Suppose that X, Y are independent random functions, then*

$$\mathbb{E}[U(Y,X)|X] = h(X)$$

where

$$h(x) := \mathbb{E}[U(Y,x)].$$

Proof. I will prove this in the continuous distribution case and leave the discrete case to the reader. (The theorem is true in general but requires measure theory in order to prove it in full generality.) The independence assumption is equivalent to $\rho_{(Y,X)}(y,x) = \rho_Y(y)\,\rho_X(x)$. Therefore,

$$\rho_{(Y|X)}(y,x) = \begin{cases} \rho_Y(y) & \text{if } \rho_X(x) > 0 \\ 0 & \text{if } \rho_X(x) = 0 \end{cases}$$

and therefore $\mathbb{E}[U(Y,X)|X] = h_0(X)$ where

$$\begin{aligned}
h_0(x) &= \int U(y,x)\,\rho_{(Y|X)}(y,x)\,dy \\
&= 1_{\rho_X(x)>0}\int U(y,x)\,\rho_Y(y)\,dy = 1_{\rho_X(x)>0}\mathbb{E}[U(Y,x)] \\
&= 1_{\rho_X(x)>0}h(x).
\end{aligned}$$

If f is a bounded function of x, then

$$\begin{aligned}
\mathbb{E}[h_0(X)\,f(X)] &= \int h_0(x)\,f(x)\,\rho_X(x)\,dx = \int_{\{x:\rho_X(x)>0\}} h_0(x)\,f(x)\,\rho_X(x)\,dx \\
&= \int_{\{x:\rho_X(x)>0\}} h(x)\,f(x)\,\rho_X(x)\,dx = \int h(x)\,f(x)\,\rho_X(x)\,dx \\
&= \mathbb{E}[h(X)\,f(X)].
\end{aligned}$$

So for all practical purposes, $h(X) = h_0(X)$, i.e. $h(X) = h_0(X)$ – a.s. (Indeed, take $f(x) = \operatorname{sgn}(h(x) - h_0(x))$ in the above equation to learn that $\mathbb{E}|h(X) - h_0(X)| = 0$. ∎

[2] **Warning:** this is **not** consistent with Eq. (5.1) as $P(X=x) = 0$ for continuous distributions.

5.4 Conditional Variances

Definition 5.30 (Conditional Variance). *Suppose that $Y \in L^2(P)$ and $X : \Omega \to S$ are given. We define*

$$\operatorname{Var}(Y|X) = \mathbb{E}\left[Y^2|X\right] - \left(\mathbb{E}[Y|X]\right)^2 \tag{5.13}$$

$$= \mathbb{E}\left[\left(Y - \mathbb{E}[Y|X]\right)^2 |X\right] \tag{5.14}$$

*to be the **conditional variance of** Y **given** X.*

Theorem 5.31. *Suppose that $Y \in L^2(P)$ and $X : \Omega \to S$ are given, then*

$$\operatorname{Var}(Y) = \mathbb{E}[\operatorname{Var}(Y|X)] + \operatorname{Var}(\mathbb{E}[Y|X]).$$

Proof. Taking expectations of Eq. (5.13) implies,

$$\begin{aligned}\mathbb{E}[\operatorname{Var}(Y|X)] &= \mathbb{E}\mathbb{E}\left[Y^2|X\right] - \mathbb{E}\left(\mathbb{E}[Y|X]\right)^2 \\ &= \mathbb{E}Y^2 - \mathbb{E}\left(\mathbb{E}[Y|X]\right)^2 = \operatorname{Var}(Y) + (\mathbb{E}Y)^2 - \mathbb{E}\left(\mathbb{E}[Y|X]\right)^2.\end{aligned}$$

The result follows from this identity and the fact that

$$\operatorname{Var}(\mathbb{E}[Y|X]) = \mathbb{E}\left(\mathbb{E}[Y|X]\right)^2 - (\mathbb{E}\mathbb{E}[Y|X])^2 = \mathbb{E}\left(\mathbb{E}[Y|X]\right)^2 - (\mathbb{E}Y)^2.$$

■

5.5 Summary on Conditional Expectation Properties

Let Y and X be random variables such that $\mathbb{E}Y^2 < \infty$ and h be function from the range of X to $\mathbb{R}$. Then the following are equivalent:

1. $h(X) = \mathbb{E}(Y|X)$, i.e. $h(X)$ is the conditional expectation of Y given X.
2. $\mathbb{E}(Y - h(X))^2 \le \mathbb{E}(Y - g(X))^2$ for all functions g, i.e. $h(X)$ is the best approximation to Y among functions of X.
3. $\mathbb{E}(Y \cdot g(X)) = \mathbb{E}(h(X) \cdot g(X))$ for all functions g, i.e. $Y - h(X)$ is orthogonal to all functions of X. Moreover, this condition uniquely determines $h(X)$.

The methods for computing $\mathbb{E}(Y|X)$ are given in the next two propositions.

Proposition 5.32 (Discrete Case). *Suppose that Y and X are discrete random variables and $p(y,x) := P(Y = y, X = x)$. Then $\mathbb{E}(Y|X) = h(X)$, where*

$$h(x) = \mathbb{E}(Y|X = x) = \frac{\mathbb{E}(Y : X = x)}{P(X = x)} = \frac{1}{p_X(x)} \sum_y y p(y,x) \tag{5.15}$$

and $p_X(x) = P(X = x)$ is the marginal distribution of X which may be computed as $p_X(x) = \sum_y p(y,x)$.

Proposition 5.33 (Continuous Case). *Suppose that Y and X are random variables which have a joint probability density $\rho(y,x)$ (i.e. $P(Y \in dy, X \in dx) = \rho(y,x)dydx$). Then $\mathbb{E}(Y|X) = h(X)$, where*

$$h(x) = \mathbb{E}(Y|X = x) := \frac{1}{\rho_X(x)} \int_{-\infty}^{\infty} y\rho(y,x)dy \tag{5.16}$$

and $\rho_X(x)$ is the marginal density of X which may be computed as

$$\rho_X(x) = \int_{-\infty}^{\infty} \rho(y,x)dy.$$

Intuitively, in all cases, $\mathbb{E}(Y|X)$ on the set $\{X = x\}$ is $\mathbb{E}(Y|X = x)$. This intuitions should help motivate some of the basic properties of $\mathbb{E}(Y|X)$ summarized in the next theorem.

Theorem 5.34. *Let Y, Y_1, Y_2 and X be random variables. Then:*

1. $\mathbb{E}(Y_1 + Y_2|X) = \mathbb{E}(Y_1|X) + \mathbb{E}(Y_2|X)$.
2. $\mathbb{E}(aY|X) = a\mathbb{E}(Y|X)$ *for all constants a.*
3. $\mathbb{E}(f(X)Y|X) = f(X)\mathbb{E}(Y|X)$ *for all functions f.*
4. $\mathbb{E}(\mathbb{E}(Y|X)) = \mathbb{E}Y$.
5. *If Y and X are independent then $\mathbb{E}(Y|X) = \mathbb{E}Y$.*
6. *If $Y \ge 0$ then $\mathbb{E}(Y|X) \ge 0$.*

Remark 5.35. Property 4 in Theorem 5.34 turns out to be a very powerful method for computing expectations. I will finish this summary by writing out Property 4 in the discrete and continuous cases:

$$\mathbb{E}Y = \sum_x \mathbb{E}(Y|X = x)p_X(x) \qquad (\textbf{Discrete Case})$$

where

$$\mathbb{E}(Y|X = x) = \begin{cases} \frac{\mathbb{E}(Y1_{X=x})}{P(X=x)} & \text{if} \quad P(X = x) > 0 \\ 0 & \text{otherwise} \end{cases}$$

$$\mathbb{E}[U(Y,X)] = \int \mathbb{E}(U(Y,X)|X = x)\rho_X(x)dx, \qquad (\textbf{Continuous Case})$$

where

$$\mathbb{E}[U(Y,x)|X = x] := \int U(y,x)\,\rho_{(Y|X)}(y,x)\,dy$$

and

$$\rho_{(Y|X)}(y,x) = \begin{cases} \frac{\rho_{(Y,X)}(y,x)}{\rho_X(x)} & \text{if } \rho_X(x) > 0 \\ 0 & \text{if } \rho_X(x) = 0 \end{cases}.$$

6

Random Sums

Suppose that $\{X_i\}_{i=1}^{\infty}$ is a collection of random variables and let

$$S_n := \begin{cases} X_1 + \cdots + X_n & \text{if } n \geq 1 \\ 0 & \text{if } n = 0 \end{cases}.$$

Given a $\mathbb{Z}_+$ – valued random variable, N, we wish to consider the random sum;

$$S_N = X_1 + \cdots + X_N.$$

We are now going to suppose for the rest of this subsection that N is independent of $\{X_i\}_{i=1}^{\infty}$ and for $f \geq 0$ we let

$$Tf(n) := \mathbb{E}[f(S_n)] \text{ for all } n \in \mathbb{N}_0.$$

Theorem 6.1. *Suppose that N is independent of $\{X_i\}_{i=1}^{\infty}$ as above. Then for any positive function f, we have,*

$$\mathbb{E}[f(S_N)] = \mathbb{E}[Tf(N)].$$

Moreover this formula holds for any f such that

$$\mathbb{E}[|f(S_N)|] = \mathbb{E}[T|f|(N)] < \infty.$$

Proof. If $f \geq 0$ we have,

$$\begin{aligned} \mathbb{E}[f(S_N)] &= \sum_{n=0}^{\infty} \mathbb{E}[f(S_N) : S_N = n] = \sum_{n=0}^{\infty} \mathbb{E}[f(S_n) : S_N = n] \\ &= \sum_{n=0}^{\infty} \mathbb{E}[f(S_n)] P(S_N = n) = \sum_{n=0}^{\infty} (Tf)(n) P(S_N = n) \\ &= \mathbb{E}[Tf(N)]. \end{aligned}$$

The moreover part follows from general non-sense not really covered in this course. ∎

Theorem 6.2. *Suppose that $\{X_i\}_{i=1}^{\infty}$ are uncorrelated $L^2(P)$ – random variables with $\mu = \mathbb{E}X_i$ and $\sigma^2 = \operatorname{Var}(X_i)$ independent of i. Assuming that $N \in L^2(P)$ is independent of the $\{X_i\}$, then*

$$\mathbb{E}[S_N] = \mu \cdot \mathbb{E}N \tag{6.1}$$

and

$$\operatorname{Var}(S_N) = \sigma^2 \mathbb{E}[N] + \mu^2 \operatorname{Var}(N). \tag{6.2}$$

Proof. Taking $f(x) = x$ in Theorem 6.1 using $Tf(n) = \mathbb{E}[S_n] = n \cdot \mu$ we find,

$$\mathbb{E}[S_N] = \mathbb{E}[\mu \cdot N] = \mu \cdot \mathbb{E}N$$

as claimed. Next take $f(x) = x^2$ in Theorem 6.1 using

$$Tf(n) = \mathbb{E}\left[S_n^2\right] = \operatorname{Var}(S_n) + (\mathbb{E}S_n)^2 = \sigma^2 n + (n \cdot \mu)^2,$$

we find that

$$\begin{aligned} \mathbb{E}\left[S_N^2\right] &= \mathbb{E}\left[\sigma^2 N + \mu^2 N^2\right] \\ &= \sigma^2 \mathbb{E}[N] + \mu^2 \mathbb{E}\left[N^2\right]. \end{aligned}$$

Combining these results shows,

$$\begin{aligned} \operatorname{Var}(S_N) &= \sigma^2 \mathbb{E}[N] + \mu^2 \mathbb{E}\left[N^2\right] - \mu^2 (\mathbb{E}N)^2 \\ &= \sigma^2 \mathbb{E}[N] + \mu^2 \operatorname{Var}(N). \end{aligned}$$

∎

Example 6.3 (Karlin and Taylor E.3.1. p77). A six-sided die is rolled, and the number N on the uppermost face is recorded. Then a fair coin is tossed N times, and the total number Z of heads to appear is observed. Determine the mean and variance of Z by viewing Z as a random sum of N Bernoulli random variables. Determine the probability mass function of Z, and use it to find the mean and variance of Z.

We have $Z = S_N = X_1 + \cdots + X_N$ where $X_i = 1$ if heads on the ith toss and zero otherwise. In this case

$$\mathbb{E}X_1 = \frac{1}{2},$$

$$\operatorname{Var}(X_1) = \frac{1}{2} - \left(\frac{1}{2}\right)^2 = \frac{1}{4},$$

$$\mathbb{E}N = \frac{1}{6}(1 + \cdots + 6) = \frac{1}{6}\frac{7 \cdot 6}{2} = \frac{7}{2},$$

$$\mathbb{E}N^2 = \frac{1}{6}\left(1^2 + 2^2 + 3^2 + 4^2 + 5^2 + 6^2\right) = \frac{91}{6}$$

$$\operatorname{Var}(N) = \frac{91}{6} - \left(\frac{7}{2}\right)^2 = \frac{35}{12}.$$

Therefore,

$$\mathbb{E}Z = \mathbb{E}X_1 \cdot \mathbb{E}N = \frac{1}{2} \cdot \frac{7}{2} = \frac{7}{4}$$
$$\operatorname{Var}(Z) = \frac{1}{4} \cdot \frac{7}{2} + \left(\frac{1}{2}\right)^2 \cdot \frac{35}{12} = \frac{77}{48} = 1.604\,2.$$

Alternatively, we have

$$\begin{aligned} P(Z=k) &= \sum_{n=1}^{6} P(Z=k|N=n)\, P(N=n) \\ &= \frac{1}{6} \sum_{n=k\vee 1}^{6} P(Z=k|N=n) \\ &= \frac{1}{6} \sum_{n=k\vee 1}^{6} \binom{n}{k} \left(\frac{1}{2}\right)^n. \end{aligned}$$

where

$$\begin{aligned} \mathbb{E}Z &= \sum_{k=0}^{6} kP(Z=k) = \sum_{k=1}^{6} kP(Z=k) \\ &= \sum_{k=1}^{6} k \frac{1}{6} \sum_{n=k}^{6} \binom{n}{k} \left(\frac{1}{2}\right)^n = \frac{7}{4} \end{aligned}$$

and

$$\mathbb{E}Z^2 = \sum_{k=0}^{6} k^2 P(Z=k) = \sum_{k=1}^{6} k^2 \frac{1}{6} \sum_{n=k}^{6} \binom{n}{k} \left(\frac{1}{2}\right)^n = \frac{14}{3}$$

so that

$$\operatorname{Var}(Z) = \frac{14}{3} - \left(\frac{7}{4}\right)^2 = \frac{77}{48}.$$

We have,

$$\begin{aligned} P(Z=0) &= \frac{1}{6} \sum_{n=1}^{6} \binom{n}{0} \left(\frac{1}{2}\right)^n = \frac{21}{128} \\ P(Z=1) &= \frac{1}{6} \sum_{n=1}^{6} \binom{n}{1} \left(\frac{1}{2}\right)^n = \frac{5}{16} \\ P(Z=2) &= \frac{1}{6} \sum_{n=2}^{6} \binom{n}{2} \left(\frac{1}{2}\right)^n = \frac{33}{128} \\ P(Z=3) &= \frac{1}{6} \sum_{n=3}^{6} \binom{n}{3} \left(\frac{1}{2}\right)^n = \frac{1}{6} \\ P(Z=4) &= \frac{1}{6} \sum_{n=4}^{6} \binom{n}{4} \left(\frac{1}{2}\right)^n = \frac{29}{384} \\ P(Z=5) &= \frac{1}{6} \sum_{n=5}^{6} \binom{n}{5} \left(\frac{1}{2}\right)^n = \frac{1}{48} \\ P(Z=6) &= \frac{1}{6} \sum_{n=6}^{6} \binom{n}{6} \left(\frac{1}{2}\right)^n = \frac{1}{384}. \end{aligned}$$

Remark 6.4. If the $\{X_i\}$ are i.i.d., we may work out the moment generating function, $mgf_{S_N}(t) := \mathbb{E}\left[e^{tS_N}\right]$ as follows. Conditioning on $N = n$ shows,

$$\begin{aligned} \mathbb{E}\left[e^{tS_N}|N=n\right] &= \mathbb{E}\left[e^{tS_n}|N=n\right] = \mathbb{E}\left[e^{tS_n}\right] \\ &= \left[\mathbb{E}e^{tX_1}\right]^n = \left[mgf_{X_1}(t)\right]^n \end{aligned}$$

so that

$$\mathbb{E}\left[e^{tS_N}|N\right] = \left[mgf_{X_1}(t)\right]^N = e^{N\ln\left(mgf_{X_1}(t)\right)}.$$

Taking expectations of this equation using the law of total expectation gives,

$$mgf_{S_N}(t) = mgf_N\left(\ln\left(mgf_{X_1}(t)\right)\right).$$

Exercise 6.1 (Karlin and Taylor II.3.P2). For each given p, let Z have a binomial distribution with parameters p and N. Suppose that N is itself binomially distributed with parameters q and M. Formulate Z as a random sum and show that Z has a binomial distribution with parameters pq and M.

Solution to Exercise (Karlin and Taylor II.3.P2). Let $\{X_i\}_{i=1}^{\infty}$ be i.i.d. Bernoulli random variables with $P(X_i = 1) = p$ and $P(X_i = 0) = 1 - p$. Then $Z \overset{d}{=} X_1 + \cdots + X_N$. We now compute

$$
\begin{aligned}
P(Z=k) &= \sum_{n=k}^{M} P(Z=k|N=n)\,P(N=n) \\
&= \sum_{l=0}^{M-k} P(Z=k|N=k+l)\,P(N=k+l) \\
&= \sum_{l=0}^{M-k} P(Z=k|N=k+l)\,P(N=k+l) \\
&= \sum_{l=0}^{M-k} p^k(1-p)^{k+l-k}\binom{k+l}{k}\cdot\binom{M}{k+l}q^{k+l}(1-q)^{M-(k+l)} \\
&= (pq)^k \sum_{l=0}^{M-k}(1-p)^l \frac{M!}{k!l!(M-k-l)!}q^l(1-q)^{M-k-l} \\
&= \binom{M}{k}(pq)^k \sum_{l=0}^{M-k}\frac{(M-k)!}{l!(M-k-l)!}[(1-p)q]^l(1-q)^{M-k-l} \\
&= \binom{M}{k}(pq)^k \sum_{l=0}^{M-k}\binom{M-k}{l}[(1-p)q]^l(1-q)^{M-k-l} \\
&= \binom{M}{k}(pq)^k[(1-p)q+(1-q)]^{M-k} \\
&= \binom{M}{k}(pq)^k[1-pq]^{M-k}
\end{aligned}
$$

as claimed. See page 58-59 of the notes where this is carried out.

Alternatively. Let $\{\xi_i\}$ be i.i.d. Bernoulli random variables with parameter q and $\{\eta_i\}$ be i.i.d. Bernoulli random variables with parameter p independent of the $\{\xi_i\}$. Then let $N=\eta_1+\cdots+\eta_M$ and $Z=\xi_1\eta_1+\cdots+\xi_M\eta_M$. Notice that $\{\xi_i\eta_i\}_{i=1}^M$ are Bernoulli random variables with parameter pq so that Z is Binomial with parameters pq and M. Further N is binomial with parameters p and M. Let $B(i_1,\ldots,i_n)$ be the event where $\eta_{i_1}=\eta_{i_2}=\cdots=\eta_{i_n}=1$ with all others being zero, then

$$\{N=n\}=\cup_{i_1<\cdots<i_n}B(i_1,\ldots,i_n)$$

so that

$$
\begin{aligned}
P(Z=k|N=n) &= \frac{\sum_{i_1<\cdots<i_n}P(\{Z=k\}\cap B(i_1,\ldots,i_n))}{\sum_{i_1<\cdots<i_n}P(B(i_1,\ldots,i_n))} \\
&= \frac{\sum_{i_1<\cdots<i_n}P(Z=k|B(i_1,\ldots,i_n))\,P(B(i_1,\ldots,i_n))}{\sum_{i_1<\cdots<i_n}P(B(i_1,\ldots,i_n))} \\
&= \frac{\sum_{i_1<\cdots<i_n}\binom{n}{k}q^k(1-q)^{n-k}P(B(i_1,\ldots,i_n))}{\sum_{i_1<\cdots<i_n}P(B(i_1,\ldots,i_n))} \\
&= \binom{n}{k}q^k(1-q)^{n-k}
\end{aligned}
$$

and this gives another more intuitive proof of the result.

Part I

Discrete Time Markov Chains

7

Markov Chains Basics

For this chapter, let S be a finite or at most countable **state space** and $p : S \times S \to [0,1]$ be a **Markov kernel,** i.e.

$$\sum_{y \in S} p(x,y) = 1 \text{ for all } i \in S. \tag{7.1}$$

A **probability** on S is a function, $\pi : S \to [0,1]$ such that $\sum_{x \in S} \pi(x) = 1$. Further, let $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$,

$$\Omega := S^{\mathbb{N}_0} = \{\omega = (s_0, s_1, \dots) : s_j \in S\},$$

and for each $n \in \mathbb{N}_0$, let $X_n : \Omega \to S$ be given by

$$X_n(s_0, s_1, \dots) = s_n.$$

Notation 7.1 *We will denote $(X_0, X_1, X_2, \dots)$ by X.*

Definition 7.2 (Markov probabilities). *A (time homogeneous) **Markov probability**[1], P, on Ω with transition kernel, p, is probability on Ω such that*

$$\begin{aligned} &P(X_{n+1} = x_{n+1} | X_0 = x_0, X_1 = x_1, \dots, X_n = x_n) \\ &\quad = P(X_{n+1} = x_{n+1} | X_n = x_n) = p(x_n, x_{n+1}) \end{aligned} \tag{7.2}$$

where $\{x_j\}_{j=1}^{n+1}$ are allowed to range over S and n over $\mathbb{N}_0$. The identity in Eq. (7.2) is only to be checked on for those $x_j \in S$ such that $P(X_0 = x_0, X_1 = x_1, \dots, X_n = x_n) > 0$. (Poetically, a Markov chain does not remember its past, its future moves are determined only by its present location and not how it got there.)

[1] The set Ω is sufficiently big that it is no longer so easy to give a rigorous definition of a probability on Ω. For the purposes of this class, a **probability on** Ω should be taken to mean an assignment, $P(A) \in [0,1]$ for all subsets, $A \subset \Omega$, such that $P(\emptyset) = 0$, $P(\Omega) = 1$, and

$$P(A) = \sum_{n=1}^{\infty} P(A_n)$$

whenever $A = \cup_{n=1}^{\infty} A_n$ with $A_n \cap A_m = \emptyset$ for all $m \neq n$. (There are technical problems with this definition which are addressed in a course on "measure theory." We may safely ignore these problems here.)

If a Markov probability P is given we will often refer to $\{X_n\}_{n=0}^{\infty}$ as a Markov chain. The condition in Eq. (7.2) may also be written as,

$$\mathbb{E}[f(X_{n+1}) \mid X_0, X_1, \dots, X_n] = \mathbb{E}[f(X_{n+1}) \mid X_n] = \sum_{y \in S} p(X_n, y) f(y) \tag{7.3}$$

for all $n \in \mathbb{N}_0$ and any bounded function, $f : S \to \mathbb{R}$.

Proposition 7.3 (Markov joint distributions). *If P is a Markov probability as in Definition 7.2 and $\pi(x) := P(X_0 = x)$, then for all $n \in \mathbb{N}_0$ and $\{x_j\} \subset S$,*

$$P(X_0 = x_0, \dots, X_n = x_n) = \pi(x_0) p(x_0, x_1) \dots p(x_{n-1}, x_n). \tag{7.4}$$

Conversely if $\pi : S \to [0,1]$ is a probability and $\{X_n\}_{n=0}^{\infty}$ is a sequence of random variables satisfying Eq. (7.4) for all n and $\{x_j\} \subset S$, then $(\{X_n\}, P, p)$ satisfies Definition 7.2.

Proof. ($\Longrightarrow$) This formal proof is by induction on n. I will do the case $n = 1$ and $n = 2$ here. For $n = 1$, if $\pi(x_0) = P(X_0 = x_0) = 0$ then both sides of Eq. (7.4) are zero and there is nothing to prove. If $\pi(x_0) = P(X_0 = x_0) > 0$, then

$$\begin{aligned} P(X_0 = x_0, X_1 = x_1) &= P(X_1 = x_1 | X_0 = x_0) P(X_0 = x_0) \\ &= \pi(x_0) \cdot p(x_0, x_1). \end{aligned}$$

Now for the case $n = 2$. Let $p := P(X_0 = x_0, X_1 = x_1) = \pi(x_0) \cdot p(x_0, x_1)$. If $p = 0$ then again both sides of Eq. (7.4) while if $p > 0$ we have by assumption and the case $n = 1$ that

$$\begin{aligned} &P(X_0 = x_0, X_1 = x_1, X_2 = x_2) \\ &\quad = P(X_2 = x_2 | X_0 = x_0, X_1 = x_1,) \cdot P(X_0 = x_0, X_1 = x_1) \\ &\quad = P(X_2 = x_2 | X_1 = x_1) \cdot P(X_0 = x_0, X_1 = x_1) \\ &\quad = p(x_1, x_2) \cdot \pi(x_0) p(x_0, x_1) = \pi(x_0) p(x_0, x_1) p(x_1, x_2). \end{aligned}$$

The formal induction argument is now left to the reader.

($\Longleftarrow$) If

$$\pi(x_0) p(x_0, x_1) \dots p(x_{n-1}, x_n) = P(X_0 = x_0, \dots, X_n = x_n) > 0,$$

then by Eq. (7.4) and the definition of conditional probabilities we find,

$$\begin{aligned} &P(X_{n+1}=x_{n+1}|X_0=x_0,X_1=x_1,\dots,X_n=x_n)\\ &=\frac{P(X_0=x_0,X_1=x_1,\dots,X_n=x_n,X_{n+1}=x_{n+1})}{P(X_0=x_0,\dots,X_n=x_n)}\\ &=\frac{\pi(x_0)p(x_0,x_1)\dots p(x_{n-1},x_n)p(x_n,x_{n+1})}{\pi(x_0)p(x_0,x_1)\dots p(x_{n-1},x_n)}=p(x_n,x_{n+1}) \end{aligned}$$

as desired. ∎

Fact 7.4 *To each probability π on S there is a unique Markov probability, P_π, on Ω such that $P_\pi(X_0=x)=\pi(x)$ for all $x\in X$. Moreover, P_π is uniquely determined by Eq. (7.4).*

Notation 7.5 *We will abbreviate the expectation ($\mathbb{E}_{P_\pi}$) with respect to P_π by $\mathbb{E}_\pi$. Moreover if*

$$\pi(y)=\delta_x(y):=\begin{cases}1 \text{ if } x=y\\ 0 \text{ if } x\neq y\end{cases}, \tag{7.5}$$

we will write P_x for $P_\pi=P_{\delta_x}$ and $\mathbb{E}_x$ for $\mathbb{E}_{\delta_x}$

For a general probability, π, on S, it follows from Proposition 7.3 and Corollary 7.6 that

$$P_\pi=\sum_{x\in S}\pi(x)P_x \text{ and } \mathbb{E}_\pi=\sum_{x\in S}\pi(x)\mathbb{E}_x. \tag{7.6}$$

Corollary 7.6. *If π is a probability on S and $u:S^{n+1}\to\mathbb{R}$ is a bounded or non-negative function, then*

$$\mathbb{E}_\pi[u(X_0,\dots,X_n)]=\sum_{x_0,\dots,x_n\in S}u(x_0,\dots,x_n)\pi(x_0)p(x_0,x_1)\dots p(x_{n-1},x_n).$$

Definition 7.7 (Matrix multiplication). *If $q:S\times S\to[0,1]$ is another Markov kernel we let $p\cdot q:S\times S\to[0,1]$ be defined by*

$$(p\cdot q)(x,y):=\sum_{z\in S}p(x,z)q(z,y). \quad (!) \tag{7.7}$$

We also let

$$p^n:=\overbrace{p\cdot p\cdot\dots\cdot p}^{n\text{ - times}}.$$

If $\pi:S\to[0,1]$ is a probability we let $(\pi\cdot q):S\to[0,1]$ be defined by

$$(\pi\cdot q)(y):=\sum_{x\in S}\pi(x)q(x,y).$$

As the definition suggests, $p\cdot q$ is the multiplication of matrices and $\pi\cdot q$ is the multiplication of a row vector π with a matrix q. It is easy to check that $\pi\cdot q$ is still a probability and $p\cdot q$ and p^n are Markov kernels. A key point to keep in mind is that a Markov process is completely specified by its transition kernel, $p:S\times S\to[0,1]$. For example we have the following method for computing $P_x(X_n=y)$.

Lemma 7.8. *Keeping the above notation, $P_x(X_n=y)=p^n(x,y)$ and more generally,*

$$P_\pi(X_n=y)=\sum_{x\in S}\pi(x)p^n(x,y)=(\pi\cdot p^n)(y).$$

Proof. We have from Eq. (7.4) that

$$\begin{aligned} P_x(X_n=y)&=\sum_{x_0,\dots,x_{n-1}\in S}P_x(X_0=x_0,X_1=x_1,\dots,X_{n-1}=x_{n-1},X_n=y)\\ &=\sum_{x_0,\dots,x_{n-1}\in S}\delta_x(x_0)p(x_0,x_1)\dots p(x_{n-2},x_{n-1})p(x_{n-1},y)\\ &=\sum_{x_1,\dots,x_{n-1}\in S}p(x,x_1)\dots p(x_{n-2},x_{n-1})p(x_{n-1},y)=p^n(x,y). \end{aligned}$$

The formula for $P_\pi(X_n=y)$ easily follows from this formula. ∎

To get a feeling for Markov chains, I suggest the reader play around with the simulation provided by Stefan Waner and Steven R. Costenoble at `www.zweigmedia.com/RealWorld/markov/markov.html` – see Figure 7.1 below.

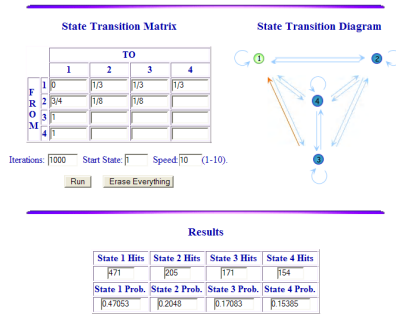


Fig. 7.1. See `www.zweigmedia.com/RealWorld/markov/markov.html` for a Markov chain simulator for chains with a state space of 4 elements or less. The user describes the chain by filling in the transition matrix **P**.

7.1 Examples

Notation 7.9 *Associated to a transition kernel, p, is a **jump graph (or jump diagram)** gotten by taking S as the set of vertices and then for $x, y \in S$, draw an arrow from x to y if $p(x,y) > 0$ and label this arrow by the value $p(x,y)$.*

Example 7.10. The transition matrix,

$$\mathbf{P} = \begin{array}{c} \begin{matrix} 1 & 2 & 3 \end{matrix} \\ \begin{bmatrix} 1/4 & 1/2 & 1/4 \\ 1/2 & 0 & 1/2 \\ 1/3 & 1/3 & 1/3 \end{bmatrix} \end{array} \begin{matrix} 1 \\ 2 \\ 3 \end{matrix}$$

is represented by the jump diagram in Figure 7.2.

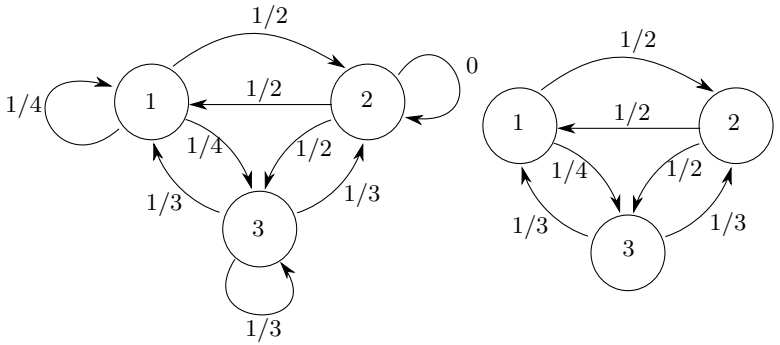


Fig. 7.2. A simple 3 state jump diagram. We typically abbreviate the jump diagram on the left by the one on the right. That is we infer by conservation of probability there has to be probability 1/4 of staying at 1, 1/3 of staying at 3 and 0 probability of staying at 2.

Example 7.11. The jump diagram for

$$\mathbf{P} = \begin{array}{c} \begin{matrix} 1 & 2 & 3 \end{matrix} \\ \begin{bmatrix} 1/4 & 1/2 & 1/4 \\ 1/2 & 0 & 1/2 \\ 1/3 & 1/3 & 1/3 \end{bmatrix} \end{array} \begin{matrix} 1 \\ 2 \\ 3 \end{matrix}$$

is shown in Figure 7.3.

Example 7.12. Suppose that $S = \{1,2,3\}$, then

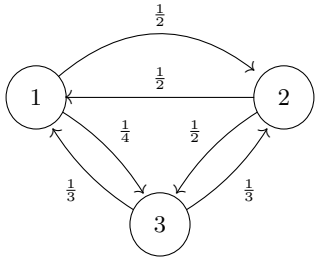


Fig. 7.3. In the above diagram there are jumps from 1 to 1 with probability 1/4 and jumps from 3 to 3 with probability 1/3 which are not explicitly shown but must be inferred by conservation of probability.

$$\mathbf{P} = \begin{array}{c} \begin{matrix} 1 & 2 & 3 \end{matrix} \\ \begin{bmatrix} 0 & 1 & 0 \\ 1/2 & 0 & 1/2 \\ 1 & 0 & 0 \end{bmatrix} \end{array} \begin{matrix} 1 \\ 2 \\ 3 \end{matrix}$$

has the jump graph given by 7.2.

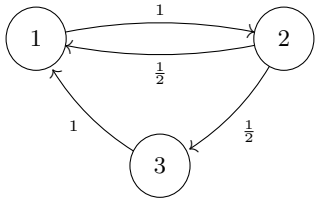


Fig. 7.4. A simple 3 state jump diagram.

Example 7.13 (Ehrenfest Urn Model). Let a beaker filled with a particle fluid mixture be divided into two parts A and B by a semipermeable membrane. Let $X_n = (\#$ of particles in $A)$ which we assume evolves by choosing a particle at random from $A \cup B$ and then replacing this particle in the opposite bin from which it was found. Modeling $\{X_n\}$ as a Markov process we find,

$$P(X_{n+1} = j \mid X_n = i) = \begin{cases} 0 & \text{if } j \notin \{i-1, i+1\} \\ \frac{i}{N} & \text{if } \quad j = i-1 \\ \frac{N-i}{N} & \text{if } \quad j = i+1 \end{cases} =: q(i,j)$$

As these probabilities do not depend on n, $\{X_n\}$ is a time homogeneous Markov chain.

Exercise 7.1. Consider a rat in a maze consisting of 7 rooms which is laid out as in the following figure.

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & & \end{bmatrix}$$

In this figure rooms are connected by either vertical or horizontal adjacent passages only, so that 1 is connected to 2 and 4 but not to 5 and 7 is only connected to 4. At each time $t \in \mathbb{N}_0$ the rat moves from her current room to one of the adjacent rooms with equal probability (the rat always changes rooms at each time step). Find the one step 7×7 transition matrix, q, with entries given by $\mathbf{P}_{ij} := P(X_{n+1} = j | X_n = i)$, where X_n denotes the room the rat is in at time n.

Solution to Exercise (7.1). The rat moves to an adjacent room from nearest neighbor locations probability being $1/D$ where D is the number of doors in the room where the rat is currently located. The transition matrix is therefore,

$$\mathbf{P} = \begin{array}{c} \begin{matrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{matrix} \\ \begin{bmatrix} 0 & 1/2 & 0 & 1/2 & 0 & 0 & 0 \\ 1/3 & 0 & 1/3 & 0 & 1/3 & 0 & 0 \\ 0 & 1/2 & 0 & 0 & 0 & 1/2 & 0 \\ 1/3 & 0 & 0 & 0 & 1/3 & 0 & 1/3 \\ 0 & 1/3 & 0 & 1/3 & 0 & 1/3 & 0 \\ 0 & 0 & 1/2 & 0 & 1/2 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix} \end{array} \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \end{matrix}. \tag{7.8}$$

and the corresponding jump diagram is given in Figure 7.5.

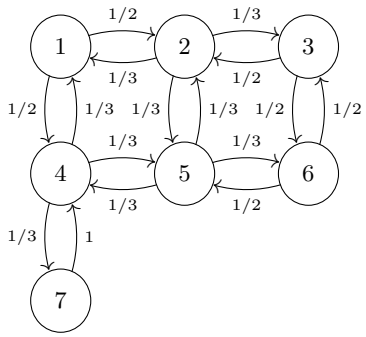


Fig. 7.5. The jump diagram for our rat in the maze.

Exercise 7.2 (2 - step MC). Consider the following simple (i.e. no-brainer) two state "game" consisting of moving between two sites labeled 1 and 2. At each site you find a coin with sides labeled 1 and 2. The probability of flipping a 2 at site 1 is $a \in (0,1)$ and a 1 at site 2 is $b \in (0,1)$. If you are at site i at time n, then you flip the coin at this site and move or stay at the current site as indicated by coin toss. We summarize this scheme by the "jump diagram" of Figure 9.8. It is reasonable to suppose that your location, X_n, at time n is modeled by a

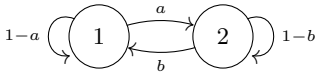


Fig. 7.6. The generic jump diagram for a two state Markov chain.

Markov process with state space, $S = \{1,2\}$. Explain (briefly) why this is a time homogeneous chain and find the one step transition probabilities,

$$p(i,j) = P(X_{n+1} = j | X_n = i) \text{ for } i,j \in S.$$

Use your result and basic linear (matrix) algebra to compute, $\lim_{n\to\infty} P(X_n = 1)$. Your answer should be independent of the possible starting distributions, $\pi = (\pi_1, \pi_2)$ for X_0 where $\pi_i := P(X_0 = i)$.

Solution to Exercise (7.2). Writing q as a matrix with entry in the i^{th} row and j^{th} column being $q(i,j)$, we have

$$q = \begin{bmatrix} 1-a & a \\ b & 1-b \end{bmatrix}.$$

If $P(X_0 = i) = \nu_i$ for $i = 1,2$ then

$$P(X_n = 1) = \sum_{k=1}^{2} \nu_k q_{k,1}^n = [\nu q^n]_1$$

where we now write $\nu = (\nu_1, \nu_2)$ as a row vector. A simple computation shows that

$$\begin{aligned} \det\left(q^{\text{tr}} - \lambda I\right) &= \det(q - \lambda I) \\ &= \lambda^2 + (a+b-2)\lambda + (1-b-a) \\ &= (\lambda - 1)(\lambda - (1-a-b)). \end{aligned}$$

Note that

$$q \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

since $\sum_j q(i,j) = 1$ – this is a general fact. Thus we always know that $\lambda_1 = 1$ is an eigenvalue of q. The second eigenvalue is $\lambda_2 = 1 - a - b$. We now find the eigenvectors of q^{tr};

$$\mathrm{Nul}\left(q^{\mathrm{tr}} - \lambda_1 I\right) = \mathrm{Nul}\left(\begin{bmatrix} -a & b \\ a & -b \end{bmatrix}\right) = \mathbb{R} \cdot \begin{bmatrix} b \\ a \end{bmatrix}$$

while

$$\mathrm{Nul}\left(q^{\mathrm{tr}} - \lambda I_2\right) = \mathrm{Nul}\left(\begin{bmatrix} b & b \\ a & a \end{bmatrix}\right) = \mathbb{R} \cdot \begin{bmatrix} 1 \\ -1 \end{bmatrix}.$$

Thus we may write

$$\nu = \alpha(b,a) + \beta(1,-1)$$

where

$$1 = \nu \cdot (1,1) = \alpha(b,a) \cdot (1,1) = \alpha(a+b).$$

Thus $\beta = \nu_1 - b = -(\nu_2 - a)$, we have

$$\nu = \frac{1}{a+b}(b,a) + \beta(1,-1).$$

and therefore,

$$\nu q^n = (b,a)\, q^n + \beta(1,-1)\, q^n = \frac{1}{a+b}(b,a) + \beta(1,-1)\, \lambda_2^n.$$

By our assumptions on and $a, b \in (0,1)$ it follows that $|\lambda_2| < 1$ and therefore

$$\lim_{n\to\infty} \nu q^n = \frac{1}{a+b}(b,a)$$

and we have shown

$$\lim_{n\to\infty} P(X_n = 1) = \frac{b}{a+b} \text{ and } \lim_{n\to\infty} P(X_n = 2) = \frac{a}{a+b}$$

independent of the starting distribution ν. Also observe that the convergence is exponentially fast.

Example 7.14. As we will see in concrete examples (see the homework and the text), many Markov chains arise in the following general fashion. Let S and T be discrete sets, $\alpha : S \times T \to S$ be a function, $\{\xi_n\}_{n=1}^{\infty}$ be i.i.d. random functions with values in T. Then given a random function, X_0 independent of the $\{\xi_n\}_{n=1}^{\infty}$ with values in S define X_n inductively by $X_{n+1} = \alpha(X_n, \xi_{n+1})$ for $n = 0, 1, 2, \ldots$. We will see that $\{X_n\}_{n=0}^{\infty}$ satisfies the Markov property with

$$p(x,y) = P(\{\alpha(x,\xi) = y\})$$

where $\xi \overset{d}{=} \xi_n$. To verify this is a Markov process first observe that notice that ξ_{n+1} is independent of $\{X_k\}_{k=0}^n$ as X_k depends on $(X_0, \xi_1, \ldots, \xi_k)$ for all k. Therefore

$$\begin{aligned} P[X_{n+1} = x_{n+1} \mid X_0 = x_0, \ldots, X_n = x_n] &= P[\alpha(X_n, \xi_{n+1}) = x_{n+1} \mid X_0 = x_0, \ldots, X_n = x_n] \\ &= P[\alpha(x_n, \xi_{n+1}) = x_{n+1} \mid X_0 = x_0, \ldots, X_n = x_n] \\ &= P(\alpha(x_n, \xi_{n+1}) = x_{n+1}) = p(x_n, x_{n+1}). \end{aligned}$$

Example 7.15 (Random Walks on the line). Suppose we have a walk on the line with probability of jumping to the right (left) is p ($q = 1 - p$). In this case we have

$$\mathbf{P} = \begin{array}{c} \begin{matrix} \cdots & -1 & 0 & 1 & 2 & \cdots \end{matrix} \\ \begin{bmatrix} \ddots & \ddots & & & & \\ \ddots & 0 & p & & & \\ & q & 0 & p & & \\ & & q & 0 & p & \\ & & & q & 0 & \ddots \\ & & & & \ddots & \ddots \end{bmatrix} \end{array} \begin{matrix} \vdots \\ -1 \\ 0 \\ 1 \\ 2 \\ \vdots \end{matrix},$$

i.e.

$$\mathbf{P}_{ij} = \begin{cases} p & \text{if} \quad j = i+1 \\ q & \text{if} \quad j = i-1 \\ 0 & \text{otherwise} \end{cases}$$

The jump diagram for such a walk is given in Figure 7.7.This fits into Example 7.14 by taking $S = \mathbb{Z}$, $T = \{\pm 1\}$, $F(s,t) = s + t$, and $\xi_n \overset{d}{=} \xi$ where $P(\xi = +1) = p$ and $P(\xi = -1) = q = 1 - p$.

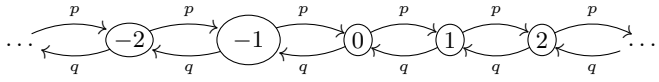
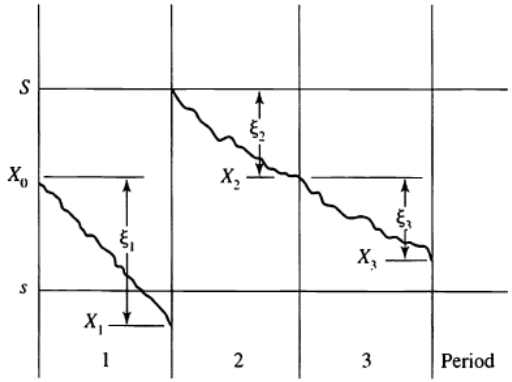


Fig. 7.7. The jump diagram for a possibly biassed simple random walk on the line.

Example 7.16 (See III.3.1 of Karlin and Taylor). Let ξ_n denote the demand of a commodity during the n^{th} – period. We will assume that $\{\xi_n\}_{n=1}^{\infty}$ are i.i.d. with $P(\xi_n = k) = a_k$ for $k \in \mathbb{N}_0$. Let X_n denote the quantity of stock on hand at the end of the n^{th} – period which is subject to the following replacement policy. We choose $s, S \in \mathbb{N}_0$ with $s < S$, if $X_n \le s$ we immediately replace the



stock to have S on hand at the beginning of the next period while if $X_n > s$ we do not add any stock. Thus,

$$X_{n+1} = \begin{cases} X_n - \xi_{n+1} \text{ if } s < X_n \leq S \\ S - \xi_{n+1} \text{ if } \quad X_n \leq s, \end{cases}$$

see Figure 3.1 on p. 106 of the book (also repeated below). Notice that we allow the stock to go negative indicating the demand is not met. It now follows that

$$\begin{aligned} P(X_{n+1} = y | X_n = x) &= \begin{cases} P(\xi_{n+1} = x - y) \text{ if } s < x \leq S \\ P(\xi_{n+1} = S - y) \text{ if } \quad x \leq s \end{cases} \\ &= \begin{cases} a_{x-y} \text{ if } s < x \leq S \\ a_{S-y} \text{ if } \quad x \leq s \end{cases} \end{aligned}$$

Example 7.17 (Discrete queueing model). Let $X_n = \#$ of people in line at time n, $\{\xi_n\}$ be i.i.d. be the number of customers arriving for service in a period and assume one person is served if there are people in the queue (think of a taxi stand). Therefore, $X_{n+1} = (X_n - 1)_+ + \xi_n$ and assuming that $P(\xi_n = k) = a_k$ for all $k \in \mathbb{N}_0$ we have,

$$P(X_{n+1} = j \mid X_n = i) = \begin{cases} 0 & \text{if } j < i - 1 \\ P(\xi_n = 0) = a_0 & \text{if } j = i - 1 \\ P(\xi_n = j - (i-1)) = a_{j-i+1} & \text{if } j \geq i \end{cases}$$

$$\mathbf{P} = \begin{array}{c} \begin{matrix} 0 & 1 & 2 & 3 & 4 & \cdots \end{matrix} \\ \begin{bmatrix} a_0 & a_1 & a_2 & a_3 & \cdots & \cdots \\ a_0 & a_1 & a_2 & \cdots & \cdots & \cdots \\ 0 & a_0 & a_1 & a_2 & \cdots & \cdots \\ 0 & 0 & a_0 & a_1 & a_2 & \cdots \\ \vdots & \vdots & & \ddots & \ddots & \vdots \end{bmatrix} \end{array} \begin{matrix} 0 \\ 1 \\ 2 \\ 3 \\ \vdots \end{matrix}.$$

Remark 7.18 (Memoryless property of the geometric distribution). Suppose that $\{X_i\}$ are i.i.d. Bernoulli random variables with $P(X_i = 1) = p$ and $P(X_i = 0) = 1 - p$ and $N = \inf\{i \geq 1 : X_i = 1\}$. Then $P(N = k) = P(X_1 = 0, \ldots, X_{k-1} = 0, X_k = 1) = (1-p)^{k-1} p$, so that N is geometric with parameter p. Using this representation we easily and intuitively see that

$$\begin{aligned} P(N = n + k | N > n) &= \frac{P(X_1 = 0, \ldots, X_{n+k-1} = 0, X_{n+k} = 1)}{P(X_1 = 0, \ldots, X_n = 0)} \\ &= P(X_{n+1} = 0, \ldots, X_{n+k-1} = 0, X_{n+k} = 1) \\ &= P(X_1 = 0, \ldots, X_{k-1} = 0, X_k = 1) = P(N = k). \end{aligned}$$

This can be verified by first principles as well;

$$\begin{aligned} P(N = n + k | N > n) &= \frac{P(N = n + k)}{P(N > n)} = \frac{p(1-p)^{n+k-1}}{\sum_{k>n} p(1-p)^{k-1}} \\ &= \frac{p(1-p)^{n+k-1}}{\sum_{j=0}^{\infty} p(1-p)^{n+j}} = \frac{(1-p)^{n+k-1}}{(1-p)^n \sum_{j=0}^{\infty} (1-p)^j} \\ &= \frac{(1-p)^{k-1}}{\frac{1}{1-(1-p)}} = p(1-p)^{k-1} = P(N = k). \end{aligned}$$

Exercise 7.3 (III.3.P4. (Queueing model)). Consider the queueing model of Section 3.4. of Karlin and Taylor. Now suppose that at most a single customer arrives during a single period, but that the service time of a customer is a random variable Z with the geometric probability distribution

$$P(Z = k) = \alpha(1-\alpha)^{k-1} \text{ for } k \in \mathbb{N}.$$

Specify the transition probabilities for the Markov chain whose state is the number of customers waiting for service or being served at the start of each period. Assume that the probability that a customer arrives in a period is β and that no customer arrives with probability $1 - \beta$.

Solution to Exercise (III.3.P4). Notice that the probability that the service of customer currently being served is finished at the end of the current period

is $\alpha = P(Z = m+1 | Z > m)$; this is the memoryless property of the geometric distribution. A $k \to k$ transition can happen in two ways: (i) a new customer arrives and the customer being served finishes, or (ii) no new customer arrives and the customer in service does not finish. The total probability of a $k \to k$ transition is therefore

$$\beta \cdot \alpha + (1-\beta)(1-\alpha) = 1 - \alpha - \beta + 2\alpha\beta.$$

(If $k = 0$ this formula must be emended; the probability of a $0 \to 0$ transition is simply $1 - \beta$.) A $k \to k+1$ transition occurs if a new customer arrives but the customer in service does not finish; this has probability $(1-\alpha)\beta$ (β if $k = 0$). Finally, for $k \geq 1$, the probability of a $k \to k-1$ transition is $\alpha(1-\beta)$, see Figure 7.8 for the jump diagram.

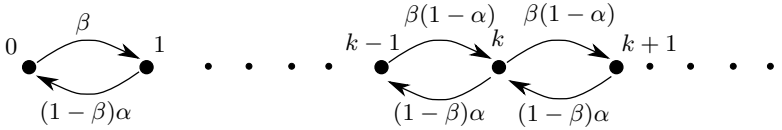


Fig. 7.8. A jump diagram for a simple queueing model.

Proposition 7.19 (Historical MC). *Suppose that $\{X_n\}_{n=0}^{\infty}$ is a Markov chain with transition probabilities, $p(x,y)$ for $x, y \in S$. Then for any $m \in \mathbb{N}$,*

$$Y_n := (X_n, X_{n+1}, \ldots, X_{n+m})$$

is a Markov chain with values in S^{m+1} whose transition kernel, q, is given by

$$q((a_0,\ldots,a_m),(b_0,\ldots,b_m)) = \delta(b_0,a_1)\ldots\delta(b_{m-1},a_m)\,p(a_m,b_m).$$

Proof. Let me give the proof for $m = 2$ only as this should suffice to explain the ideas. We have,

$$\begin{aligned}
&P(Y_{n+1} = (b_0,b_1,b_2) | Y_n = (a_0,a_1,a_2), Y_{n-1} = *, \ldots, Y_0 = *) = \\
&= P\left((X_{n+1},X_{n+2},X_{n+3}) = (b_0,b_1,b_2) \middle| \begin{matrix}(X_n,X_{n+1},X_{n+2}) = (a_0,a_1,a_2)\\ Y_{n-1} = *,\ldots,Y_0 = *\end{matrix}\right) \\
&= P\left((X_{n+1},X_{n+2},X_{n+3}) = (b_0,b_1,b_2) \middle| \begin{matrix}(X_n,X_{n+1},X_{n+2}) = (a_0,a_1,a_2)\\ X_{n-1} = *,\ldots,X_0 = *\end{matrix}\right) \\
&= P\left((a_1,a_2,X_{n+3}) = (b_0,b_1,b_2) \middle| \begin{matrix}(X_n,X_{n+1},X_{n+2}) = (a_0,a_1,a_2)\\ X_{n-1} = *,\ldots,X_0 = *\end{matrix}\right) \\
&= \delta(b_0,a_1)\,\delta(b_1,a_2)\,P(X_{n+3} = b_2 | X_{n+2} = a_2, X_{n+1} = *, \ldots, X_0 = *) \\
&= \delta(a_0,b_1)\,\delta(a_2,b_1)\,p(a_2,b_2).
\end{aligned}$$

■

Example 7.20. Suppose we flip a fair coin repeatedly and would like to find the first time the pattern HHT appears. To do this we will later examine the Markov chain, $Y_n = (X_n, X_{n+1}, X_{n+2})$ where $\{X_n\}_{n=0}^{\infty}$ is the sequence of unbiased independent coin flips with values in $\{H,T\}$. The state space for Y_n is

$$S = \{\, TTT\ \ THT\ \ TTH\ \ THH\ HHH\ HTT\ HTH\ HHT \,\}.$$

The transition matrix for recording three flips in a row of a fair coin is

$$\mathbf{P} = \frac{1}{2}\left[\begin{array}{c|cccccccc}
 & TTT & THT & TTH & THH & HHH & HTT & HTH & HHT \\ \hline
TTT & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
THT & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\
TTH & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\
THH & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\
HHH & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\
HTT & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
HTH & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\
HHT & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0
\end{array}\right].$$

7.2 Hitting Times

Skip this section. It is redone better later

We assume the $\{X_n\}_{n=0}^{\infty}$ is a Markov chain with values in S and transition kernel $\mathbf{P}$. I will often write $p(x,y)$ for $\mathbf{P}_{xy}$. We are going to further assume that $B \subset S$ is non-empty proper subset of S and $A = S \setminus B$.

Definition 7.21 (Hitting times). *Given a subset $B \subset S$ we let T_B be the first time $\{X_n\}$ hits B, i.e.*

$$T_B = \min\{n : X_n \in B\}$$

*with the convention that $T_B = \infty$ if $\{n : X_n \in B\} = \emptyset$. We call T_B the **first hitting time** of B by $X = \{X_n\}_n$.*

Observe that

$$\begin{aligned}
\{T_B = n\} &= \{X_0 \notin B, \ldots, X_{n-1} \notin B, X_n \in B\} \\
&= \{X_0 \in A, \ldots, X_{n-1} \in A, X_n \in B\}
\end{aligned}$$

and

$$\{T_B > n\} = \{X_0 \in A, \ldots, X_{n-1} \in A, X_n \in A\}$$

so that $\{T_B = n\}$ and $\{T_B > n\}$ only depends on $(X_0, \ldots, X_n)$. A random time, $T : \Omega \to \mathbb{N} \cup \{0, \infty\}$, with either of these properties is called a **stopping time.**

Lemma 7.22. *For any random time $T:\Omega\to\mathbb{N}\cup\{0,\infty\}$ we have*

$$P(T=\infty)=\lim_{n\to\infty}P(T>n)\ \text{ and }\mathbb{E}T=\sum_{k=0}^{\infty}P(T>k).$$

Proof. The first equality is a consequence of the continuity of P and the fact that

$$\{T>n\}\downarrow\{T=\infty\}.$$

The second equality is proved as follows;

$$\begin{aligned}\mathbb{E}T&=\sum_{m>0}mP(T=m)=\sum_{0<k\le m<\infty}P(T=m)\\&=\sum_{k=1}^{\infty}P(T\ge k)=\sum_{k=0}^{\infty}P(T>k).\end{aligned}$$

■

Notation 7.23 *Let $\mathbf{Q}$ be $\mathbf{P}$ restricted to A, i.e. $\mathbf{Q}_{x,y}=\mathbf{P}_{x,y}$ for all $x,y\in A$. In particular we have*

$$\mathbf{Q}^N_{x,y}:=\sum_{x_1,\dots,x_{N-1}\in A}Q_{x,x_1}Q_{x_1,x_2}\dots Q_{x_{N-1},y}\text{ for all }x,y\in A.$$

Corollary 7.24. *Continuing the notation introduced above, for any $x\in A$ we have*

$$P_x(T_B=\infty)=\lim_{N\to\infty}\sum_{y\in A}\mathbf{Q}^N_{x,y}$$

and

$$\mathbb{E}_x[T_B]=\sum_{N=0}^{\infty}\sum_{y\in A}\mathbf{Q}^N_{x,y}$$

with the convention that

$$\mathbf{Q}^0_{x,y}=\delta_{x,y}=\begin{cases}1\text{ if }x=y\\0\text{ if }x\neq y\end{cases}.$$

Proof. The results follow from Lemma 7.22 after observing that

$$\begin{aligned}P_x(T_B>N)&=P_x(X_0\in A,\dots,X_N\in A)\\&=\sum_{x_1,\dots,x_N\in A}p(x,x_1)p(x_1,x_2)\dots p(x_{N-1},x_N)=\sum_{y\in A}\mathbf{Q}^N_{x,y}.\end{aligned}\tag{7.9}$$

■

Proposition 7.25. *Suppose that $B\subset S$ is non-empty proper subset of S and $A=S\setminus B$. Further suppose there is some $\alpha<1$ such that $P_x(T_B=\infty)\le\alpha$ for all $x\in A$, then $P_x(T_B=\infty)=0$ for all $x\in A$. [In words; if there is a "uniform" chance that X hits B starting from any site, then X will surely hit B.]*

Proof. Taking $N=m+n$ in Eq. (7.9) shows

$$P_x(T_B>m+n)=\sum_{y,z\in A}\mathbf{Q}^m_{x,y}\mathbf{Q}^n_{y,z}=\sum_{y\in A}\mathbf{Q}^m_{x,y}P_y(T_B>n).\tag{7.10}$$

Letting $n\to\infty$ (using D.C.T.) in this equation shows,

$$\begin{aligned}P_x(T_B=\infty)&=\sum_{y\in A}\mathbf{Q}^m_{x,y}P_y(T_B=\infty)\\&\le\alpha\sum_{y\in A}\mathbf{Q}^m_{x,y}=\alpha P_x(T_B>n).\end{aligned}$$

Finally letting $n\to\infty$ shows $P_x(T_B=\infty)\le\alpha P_x(T_B=\infty)$, i.e. $P_x(T_B=\infty)=0$ for all $x\in A$. ■

We will see in examples later that it is possible for $P_x(T_B=\infty)=0$ while $\mathbb{E}_xT_B=\infty$. The next theorem gives a criteria which avoids this scenario.

Theorem 7.26. *Suppose that $B\subset S$ is non-empty proper subset of S and $A=S\setminus B$. Further suppose there is some $\alpha<1$ and $n<\infty$ such that $P_x(T_B>n)\le\alpha$ for all $x\in A$, then*

$$\mathbb{E}_x(T_B)\le\frac{n}{1-\alpha}<\infty$$

for all $x\in A$. [In words; if there is a "uniform" chance that X hits B starting from any site within a fixed number of steps, then the expected hitting time of B is finite and bounded independent of the starting point.]

Proof. From Eq. (7.10) for any $m\in\mathbb{N}$ we have

$$P_x(T_B>m+n)=\sum_{y\in A}\mathbf{Q}^m_{x,y}P_y(T_B>n)\le\alpha\sum_{y\in A}\mathbf{Q}^m_{x,y}=\alpha P_x(T_B>m).$$

One easily uses this relationship to show inductively that

$$P_x(T_B>kn)\le\alpha^k\text{ for all }k=0,1,2\dots.$$

We then have,

$$\begin{aligned}\mathbb{E}_xT_B&=\sum_{k=0}^{\infty}P(T_B>k)\le\sum_{k=0}^{\infty}nP(T_B>kn)\\&\le\sum_{k=0}^{\infty}n\alpha^k=\frac{n}{1-\alpha}<\infty,\end{aligned}$$

wherein we have used,

$$P(T_B > kn+m) \le P(T_B > kn) \text{ for } m = 0, \dots, n-1.$$

■

Corollary 7.27. *If $A = S \setminus B$ is a finite set and $P_x(T_B = \infty) < 1$ for all $x \in A$, then $\mathbb{E}_x T_B < \infty$ for all $x \in A$.*

Proof. Let $\alpha_0 = \max_{x \in A} P_x(T = \infty) < 1$. Now fix $\alpha \in (\alpha_0, 1)$. Using

$$\alpha_0 \ge P_x(T = \infty) =\downarrow \lim_{n \to \infty} P_x(T > n)$$

we will have $P_x(T > m) \le \alpha$ for $m \ge N_x$ for some $N_x < \infty$. Taking $n := \max\{N_x : x \in A\} < \infty$ (A is a finite set), we will have $P_x(T > n) \le \alpha$ for all $x \in A$ and we may now apply Theorem 7.26. ■

Definition 7.28 (First return time). *For any $x \in S$, let $R_x := \min\{n \ge 1 : X_n = x\}$ where the minimum of the empty set is defined to be ∞.*

On the event $\{X_0 \neq x\}$ we have $R_x = T_x := \min\{n \ge 0 : X_n = x\}$ – the first hitting time of x. So R_x is really manufactured for the case where $X_0 = x$ in which case $T_x = 0$ while R_x is the *first return time* to x.

Exercise 7.4. Let $x \in X$. Show;

a) for all $n \in \mathbb{N}_0$,

$$P_x(R_x > n+1) \le \sum_{y \neq x} p(x,y) P_y(T_x > n). \tag{7.11}$$

b) Use Eq. (7.11) to conclude that if $P_y(T_x = \infty) = 0$ for all $y \neq x$ then $P_x(R_x = \infty) = 0$, i.e. $\{X_n\}$ will return to x when started at x.

c) Sum Eq. (7.11) on $n \in \mathbb{N}_0$ to show

$$\mathbb{E}_x[R_x] \le P_x(R_x > 0) + \sum_{y \neq x} p(x,y) \mathbb{E}_y[T_x]. \tag{7.12}$$

d) Now suppose that S is a finite set and $P_y(T_x = \infty) < 1$ for all $y \neq x$, i.e. there is a positive chance of hitting x from any $y \neq x$ in S. Explain how Eq. (7.12) combined with Corollary 7.27 shows that $\mathbb{E}_x[R_x] < \infty$.

Solution to Exercise (7.4). a) Using the first step analysis we have,

$$\begin{aligned} P_x(R_x > n+1) &= \mathbb{E}_x[1_{R_x > n+1}] = \mathbb{E}_{p(x,\cdot)}\left[1_{R_x(x,X)>n+1}\right] \\ &= p(x,x)\mathbb{E}_x\left[1_{R_x(x,X)>n+1}\right] + \sum_{y \neq x} p(x,y)\mathbb{E}_y\left[1_{R_x(x,X)>n+1}\right]. \end{aligned}$$

On the event $X_0 = x$ we have $R_x(x,X) = 1$ which is not greater than $n+1$ so that $\mathbb{E}_x\left[1_{R_x(x,X)>n+1}\right] = 0$ while on the event $X_0 \neq x$ we have $R_x(x,X) = T_x(X) + 1$ so that for $y \neq x$,

$$\mathbb{E}_y\left[1_{R_x(x,X)>n+1}\right] = \mathbb{E}_y\left[1_{T_x(X)+1>n+1}\right] = P_y(T_x > n).$$

Putting these comments together prove Eq. (7.11).

b) Let $n \to \infty$ in Eq. (7.11) using DCT in order to conclude,

$$P_x(R_x = \infty) \le \sum_{y \neq x} p(x,y) P_y(T_x = \infty) = 0.$$

c) Using Lemma[2] 7.22 twice along with Fubini's theorem for sums we have,

$$\begin{aligned} \mathbb{E}_x[R_x] &= P_x(R_x > 0) + \sum_{n=0}^{\infty} P_x(R_x > n+1) \\ &\le P_x(R_x > 0) + \sum_{n=0}^{\infty} \sum_{y \neq x} p(x,y) P_y(T_x > n) \\ &= P_x(R_x > 0) + \sum_{y \neq x} p(x,y) \sum_{n=0}^{\infty} P_y(T_x > n) \\ &= P_x(R_x > 0) + \sum_{y \neq x} p(x,y) \mathbb{E}_y[T_x]. \end{aligned}$$

d) From Corollary 7.27 with $B = \{x\}$, we know that $\mathbb{E}_y[T_x] < \infty$ for all $y \neq x$. Thus the right side of Eq. (7.12) is a finite sum of finite terms and therefore is finite. This then implies $\mathbb{E}_x R_x < \infty$.

[2] That is $\mathbb{E}T = \sum_{k=0}^{\infty} P(T > k)$.

8

Markov Conditioning

We assume the $\{X_n\}_{n=0}^{\infty}$ is a Markov chain with values in S and transition kernel $\mathbf{P}$ and $\pi : S \to [0,1]$ is a probability on S. As usual we write P_π for the unique probability satisfying Eq. (7.4) and we will often write $p(x,y)$ for $\mathbf{P}_{xy}$.

Theorem 8.1 (Markov conditioning). *Let π be a probability on S, $F(X) = F(X_0, X_1, \dots)$ be a random variable[1] depending on X. Then for each $m \in \mathbb{N}$ we have*

$$\mathbb{E}_\pi [F(X_0, X_1, \dots)] = \mathbb{E}_\pi \left[\mathbb{E}^{(Y)}_{X_m} F(X_0, X_1, \dots X_{m-1}, Y_0, Y_1, \dots) \right] \tag{8.1}$$

where $\mathbb{E}^{(Y)}_x$ denotes the expectation with respect to an independent copy, Y, of the chain X which starts at $x \in S$. To be more explicit,

$$\mathbb{E}_\pi [F(X_0, X_1, \dots)] = \mathbb{E}_\pi [h(X_0, \dots, X_m)]$$

where for all $x_0, \dots, x_m \in S$,

$$h(x_0, \dots, x_m) := \mathbb{E}_{x_m} [F(x_0, \dots, x_{m-1}, X_0, X_1, \dots)].$$

[In words, given $X_0, \dots, X_m$, $(X_m, X_{m+1}, \dots)$ has the same distribution as independent copy $(Y_0, Y_1, \dots)$ of the chain X where Y required to start at X_m.]

Alternatively stated: *if $x_0, x_1, \dots, x_m \in S$ with $P_\pi(X_0 = x_0, \dots, X_m = x_m) > 0$, then*

$$\begin{aligned} &\mathbb{E}_\pi [F(X_0, X_1, \dots) | X_0 = x_0, \dots, X_m = x_m] \\ &\quad = \mathbb{E}_{x_m} [F(x_0, x_1, \dots, x_{m-1}, X_0, X_1, \dots)] \end{aligned} \tag{8.2}$$

or equivalently put,

$$\mathbb{E}_\pi ([F(X_0, X_1, \dots) | X_0, \dots, X_m]) = \mathbb{E}^{(Y)}_{X_m} [F(X_0, X_1, \dots, X_{m-1}, Y_0, Y_1, \dots)]. \tag{8.3}$$

Proof. Fact: by "limiting" arguments beyond the scope of this course it suffices to prove Eq. (8.1) for $F(X)$ of the form, $F(X) = F(X_0, X_1, \dots, X_N)$ with $N < \infty$. Now for such a function we have,

[1] In this theorem we assume that F is either bounded or non-negative.

$$\begin{aligned} &\mathbb{E}_\pi [F(X_0, X_1, \dots, X_N) : X_0 = x_0, \dots, X_m = x_m] \\ &= \sum_{x_{m+1}, \dots, x_N \in S} F(x_0, \dots, x_m, x_{m+1}, \dots, x_N) \begin{bmatrix} \pi(x_0) p(x_0, x_1) \dots p(x_{m-1}, x_m) \cdot \\ p(x_m, x_{m+1}) \dots p(x_{N-1}, x_N) \end{bmatrix} \\ &= P_\pi(X_0 = x_0, \dots, X_m = x_m) \cdot \\ &\quad \cdot \sum_{x_{m+1}, \dots, x_N \in S} F(x_0, \dots, x_m, x_{m+1}, \dots, x_N) p(x_m, x_{m+1}) \dots p(x_{N-1}, x_N) \\ &= P_\pi(X_0 = x_0, \dots, X_m = x_m) \\ &\quad \cdot \sum_{y_1, \dots, y_{N-m} \in S} F(x_0, \dots, x_m, y_1, y_2, \dots, y_{N-m}) p(x_m, y_1) \dots p(y_{N-m-1}, y_{N-m}) \\ &= P_\pi(X_0 = x_0, \dots, X_m = x_m) h(x_0, \dots, x_m). \end{aligned} \tag{8.4}$$

Summing this equation on $x_0, \dots, x_m$ in S gives Eq. (8.1) and dividing this equation by $P_\pi(X_0 = x_0, \dots, X_m = x_m)$ proves Eq. (8.2). ∎

To help cement the ideas above, let me pause to write out the above argument in the special case where $m = 2$ and $N = 5$. In this case we have;

$$\begin{aligned} &\mathbb{E}_\pi [F(X_0, X_1, \dots, X_5) : X_0 = x_0, X_1 = x_1, X_2 = x_2] \\ &= \sum_{x_3, x_4, x_5 \in S} F(x_0, x_1, x_2, x_3, x_4, x_5) \begin{bmatrix} \pi(x_0) p(x_0, x_1) p(x_1, x_2) \cdot \\ p(x_2, x_3) p(x_3, x_4) p(x_4, x_5) \end{bmatrix} \\ &= P_\pi(X_0 = x_0, X_1 = x_1, X_2 = x_2) \cdot \\ &\quad \cdot \sum_{x_3, x_4, x_5 \in S} F(x_0, x_1, x_2, x_3, x_4, x_5) [p(x_2, x_3) p(x_3, x_4) p(x_4, x_5)] \\ &= P_\pi(X_0 = x_0, X_1 = x_1, X_2 = x_2) \\ &\quad \cdot \sum_{y_1, y_2, y_3 \in S} F(x_0, x_1, x_2, y_1, y_2, y_3) [p(x_2, y_1) p(y_1, y_2) p(y_2, y_3)] \\ &= P_\pi(X_0 = x_0, X_1 = x_1, X_2 = x_2) \cdot \mathbb{E}^{(Y)}_{x_2} [F(x_0, x_1, Y_0, Y_1, Y_2, Y_3)]. \end{aligned}$$

8.1 Hitting Time Estiamtes

We assume the $\{X_n\}_{n=0}^{\infty}$ is a Markov chain with values in S and transition kernel $\mathbf{P}$. I will often write $p(x,y)$ for $\mathbf{P}_{xy}$. We are going to further assume that $B \subset S$ is non-empty proper subset of S and $A = S \setminus B$.

Definition 8.2 (Hitting times). *Given a subset $B \subset S$ we let T_B be the first time $\{X_n\}$ hits B, i.e.*

$$T_B = \min\{n : X_n \in B\}$$

*with the convention that $T_B = \infty$ if $\{n : X_n \in B\} = \emptyset$. We call T_B the **first hitting time** of B by $X = \{X_n\}_n$.*

Observe that

$$\begin{aligned}\{T_B = n\} &= \{X_0 \notin B, \dots, X_{n-1} \notin B, X_n \in B\} \\ &= \{X_0 \in A, \dots, X_{n-1} \in A, X_n \in B\}\end{aligned}$$

and

$$\{T_B > n\} = \{X_0 \in A, \dots, X_{n-1} \in A, X_n \in A\}$$

so that $\{T_B = n\}$ and $\{T_B > n\}$ only depends on $(X_0, \dots, X_n)$. A random time, $T : \Omega \to \mathbb{N} \cup \{0, \infty\}$, with either of these properties is called a **stopping time.**

Lemma 8.3. *For any random time $T : \Omega \to \mathbb{N} \cup \{0, \infty\}$ we have*

$$P(T = \infty) = \lim_{n\to\infty} P(T > n) \text{ and } \mathbb{E}T = \sum_{k=0}^{\infty} P(T > k).$$

Proof. The first equality is a consequence of the continuity of P and the fact that

$$\{T > n\} \downarrow \{T = \infty\}.$$

The second equality is proved as follows;

$$\begin{aligned}\mathbb{E}T &= \sum_{m>0} mP(T = m) = \sum_{0<k\le m<\infty} P(T = m) \\ &= \sum_{k=1}^{\infty} P(T \ge k) = \sum_{k=0}^{\infty} P(T > k).\end{aligned}$$

■

Let us now use Theorem 8.1 to give variants of the proofs of our hitting time results above. In what follows π will denote a probability on S.

Corollary 8.4. *Let $B \subset S$ and T_B be as above, then for $n, m \in \mathbb{N}$ we have*

$$P_\pi(T_B > m + n) = \mathbb{E}_\pi\left[1_{T_B>m} P_{X_m}[T_B > n]\right]. \tag{8.5}$$

Proof. Using Theorem 8.1,

$$\begin{aligned}P_\pi(T_B > m+n) &= \mathbb{E}_\pi\left[1_{T_B(X)>m+n}\right] \\ &= \mathbb{E}_\pi\left[\mathbb{E}^{(Y)}_{X_m}\left[1_{T_B(X_0,\dots,X_{m-1},Y_0,Y_1,\dots)>m+n}\right]\right] \\ &= \mathbb{E}_\pi\left[\mathbb{E}^{(Y)}_{X_m}\left[1_{T_B(X)>m} \cdot 1_{T_B(Y)>n}\right]\right] \\ &= \mathbb{E}_\pi\left[1_{T_B(X)>m}\mathbb{E}^{(Y)}_{X_m}\left[1_{T_B(Y)>n}\right]\right] = \mathbb{E}_\pi\left[1_{T_B>m} P_{X_m}[T_B > n]\right].\end{aligned}$$

■

Corollary 8.5. *Suppose that $B \subset S$ is non-empty proper subset of S and $A = S \setminus B$. Further suppose there is some $\alpha < 1$ such that $P_x(T_B = \infty) \le \alpha$ for all $x \in A$, then $P_\pi(T_B = \infty) = 0$. [In words; if there is a "uniform" chance that X hits B starting from any site, then X will surely hit B from any point in A.]*

Proof. Since $T_B = 0$ on $\{X_0 \in B\}$ we in fact have $P_x(T_B = \infty) \le \alpha$ for all $x \in S$. Letting $n \to \infty$ in Eq. (8.5) shows,

$$P_\pi(T_B = \infty) = \mathbb{E}_\pi\left[1_{T_B>m} P_{X_m}[T_B = \infty]\right] \le \mathbb{E}_\pi\left[1_{T_B>m}\alpha\right] = \alpha P_\pi(T_B > m).$$

Now letting $m \to \infty$ in this equation shows $P_\pi(T_B = \infty) \le \alpha P_\pi(T_B = \infty)$ from which it follows that $P_\pi(T_B = \infty) = 0$. ■

Corollary 8.6. *Suppose that $B \subset S$ is non-empty proper subset of S and $A = S \setminus B$. Further suppose there is some $\alpha < 1$ and $n < \infty$ such that $P_x(T_B > n) \le \alpha$ for all $x \in A$, then*

$$\mathbb{E}_\pi(T_B) \le \frac{n}{1-\alpha} < \infty$$

for all $x \in A$. [In words; if there is a "uniform" chance that X hits B starting from any site within a fixed number of steps, then the expected hitting time of B is finite and bounded independent of the starting distribution.]

Proof. Again using $T_B = 0$ on $\{X_0 \in B\}$ we may conclude that $P_x(T_B > n) \le \alpha$ for all $x \in S$. Letting $m = kn$ in Eq. (8.5) shows

$$P_\pi(T_B > kn + n) = \mathbb{E}_\pi\left[1_{T_B>kn} P_{X_m}[T_B > n]\right] \le \mathbb{E}_\pi\left[1_{T_B>kn} \cdot \alpha\right] = \alpha P_\pi(T_B > kn).$$

Iterating this equation using the fact that $P_\pi(T_B > 0) \le 1$ shows $P_\pi(T_B > kn) \le \alpha^k$ for all $k \in \mathbb{N}_0$. Therefore with the aid of Lemma 8.3 and the observation,

$$P(T_B > kn + m) \le P(T_B > kn) \text{ for } m = 0, \dots, n-1,$$

we find,

$$\mathbb{E}_x T_B = \sum_{k=0}^{\infty} P(T_B > k) \le \sum_{k=0}^{\infty} nP(T_B > kn)$$
$$\le \sum_{k=0}^{\infty} n\alpha^k = \frac{n}{1-\alpha} < \infty.$$

■

Corollary 8.7. *If $A = S \setminus B$ is a finite set and $P_x(T_B = \infty) < 1$ for all $x \in A$, then $\mathbb{E}_\pi T_B < \infty$.*

Proof. Since

$$P_x(T > m) \downarrow P_x(T = \infty) < 1 \text{ for all } x \in A$$

we can find $M_x < \infty$ such that $P_x(T > M_x) < 1$. Using the fact that A is a finite set we let $n := \max_{x\in A} M_x < \infty$ and then take $\alpha := \max_{x\in A} P_x(T > n) < 1$. Corollary 8.6 now applies to complete the proof. ■

8.2 First Step Analysis

The next theorem (which is a special case of Theorem 8.1) is the basis of the first step analysis developed in this section.

Theorem 8.8 (First step analysis). *Let $F(X) = F(X_0, X_1, \dots)$ be some function of the paths $(X_0, X_1, \dots)$ of our Markov chain, then for all $x, y \in S$ with $p(x,y) > 0$ we have*

$$\mathbb{E}_x[F(X_0, X_1, \dots) | X_1 = y] = \mathbb{E}_y[F(x, X_0, X_1, \dots)] \tag{8.6}$$

and

$$\mathbb{E}_x[F(X_0, X_1, \dots)] = \mathbb{E}_{p(x,\cdot)}[F(x, X_0, X_1, \dots)]$$
$$= \sum_{y\in S} p(x,y)\, \mathbb{E}_y[F(x, X_0, X_1, \dots)]. \tag{8.7}$$

Proof. Equation (8.6) follows directly from Theorem 8.1,

$$\mathbb{E}_x[F(X_0, X_1, \dots) | X_1 = y] = \mathbb{E}_x[F(X_0, X_1, \dots) | X_0 = x, X_1 = y]$$
$$= \mathbb{E}_y[F(x, X_0, X_1, \dots)].$$

Equation (8.7) now follows from Eq. (8.6), the law of total expectation, and the fact that $P_x(X_1 = y) = p(x,y)$. ■

Let us now suppose for until further notice that B is a non-empty proper subset of S, $A = S \setminus B$, and $T_B = T_B(X)$ is the first hitting time of B by X.

Notation 8.9 *Given a transition matrix $\mathbf{P} = (p(x,y))_{x,y\in S}$ we let $Q = (p(x,y))_{x,y\in A}$ and $\mathbf{R} := (p(x,y))_{x\in A, y\in B}$ so that, schematically,*

$$\mathbf{P} = \begin{matrix} \\ A \\ B \end{matrix} \begin{matrix} \begin{matrix} A & B \end{matrix} \\ \begin{bmatrix} \mathbf{Q} & \mathbf{R} \\ * & * \end{bmatrix} \end{matrix}.$$

Remark 8.10. To construct the matrix $\mathbf{Q}$ and $\mathbf{R}$ from $\mathbf{P}$, let $\mathbf{P}'$ be $\mathbf{P}$ with the rows corresponding to B omitted. To form $\mathbf{Q}$ from $\mathbf{P}'$, remove the columns of $\mathbf{P}'$ corresponding to B and to form $\mathbf{R}$ from $\mathbf{P}'$, remove the columns of $\mathbf{P}'$ corresponding to A.

Example 8.11. If $S = \{1,2,3,4,5,6,7\}$, $A = \{1,2,4,5,6\}$, $B = \{3,7\}$, and

$$\mathbf{P} = \begin{matrix} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \end{matrix} \begin{matrix} \begin{matrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{matrix} \\ \begin{bmatrix} 0 & 1/2 & 0 & 1/2 & 0 & 0 & 0 \\ 1/3 & 0 & 1/3 & 0 & 1/3 & 0 & 0 \\ 0 & 1/2 & 0 & 0 & 0 & 1/2 & 0 \\ 1/3 & 0 & 0 & 0 & 1/3 & 0 & 1/3 \\ 0 & 1/3 & 0 & 1/3 & 0 & 1/3 & 0 \\ 0 & 0 & 1/2 & 0 & 1/2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \end{matrix},$$

then

$$\mathbf{P}' = \begin{matrix} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{matrix} \begin{matrix} \begin{matrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{matrix} \\ \begin{bmatrix} 0 & 1/2 & 0 & 1/2 & 0 & 0 & 0 \\ 1/3 & 0 & 1/3 & 0 & 1/3 & 0 & 0 \\ 1/3 & 0 & 0 & 0 & 1/3 & 0 & 1/3 \\ 0 & 1/3 & 0 & 1/3 & 0 & 1/3 & 0 \\ 0 & 0 & 1/2 & 0 & 1/2 & 0 & 0 \end{bmatrix} \end{matrix}.$$

Deleting the 3 and 7 columns of $\mathbf{P}'$ gives

$$\mathbf{Q} = \mathbf{P}_{A,A} = \begin{matrix} \\ 1 \\ 2 \\ 4 \\ 5 \\ 6 \end{matrix} \begin{matrix} \begin{matrix} 1 & 2 & 4 & 5 & 6 \end{matrix} \\ \begin{bmatrix} 0 & 1/2 & 1/2 & 0 & 0 \\ 1/3 & 0 & 0 & 1/3 & 0 \\ 1/3 & 0 & 0 & 1/3 & 0 \\ 0 & 1/3 & 1/3 & 0 & 1/3 \\ 0 & 0 & 0 & 1/2 & 0 \end{bmatrix} \end{matrix}$$

and deleting the $1, 2, 4, 5$, and 6 columns of $\mathbf{P}'$ gives

$$\mathbf{R}=\mathbf{P}_{A,B}=\begin{array}{c} \begin{matrix} 3 & 7 \end{matrix} \\ \begin{bmatrix} 0 & 0 \\ 1/3 & 0 \\ 0 & 1/3 \\ 0 & 0 \\ 1/2 & 0 \end{bmatrix} \end{array}\begin{matrix} 1 \\ 2 \\ 4 \\ 5 \\ 6 \end{matrix}.$$

Before continuing on you may wish to first visit Example 8.14 below.

Theorem 8.12 (Hitting distributions). *Let $h : B \to \mathbb{R}$ be a bounded or non-negative function and let $u : S \to \mathbb{R}$ be defined by*

$$u(x) := \mathbb{E}_x\left[h\left(X_{T_B}\right) : T_B < \infty\right] \text{ for } x \in A.$$

Then $u = h$ on B and

$$u(x) = \sum_{y\in A} p(x,y)\, u(y) + \sum_{y\in B} p(x,y)\, h(y) \text{ for all } x \in A. \tag{8.8}$$

In matrix notation this becomes

$$\mathbf{u} = \mathbf{Q}u + \mathbf{Rh} \implies \mathbf{u} = (I-\mathbf{Q})^{-1}\mathbf{Rh},$$

i.e.

$$\mathbb{E}_x\left[h\left(X_{T_B}\right) : T_B < \infty\right] = \left[(I-\mathbf{Q})^{-1}\mathbf{Rh}\right]_x \text{ for all } x \in A. \tag{8.9}$$

As a special case if $h(s) = \delta_y(s)$ for some $y \in B$, then Eq. (8.9) becomes,

$$P_x\left(X_{T_B} = y : T_B < \infty\right) = \left[(I-\mathbf{Q})^{-1}\mathbf{R}\right]_{x,y}. \tag{8.10}$$

Proof. To shorten the notation we will use the convention that $h\left(X_{T_B}\right) = 0$ if $T_B = \infty$ so that we may simply write $u(x) := \mathbb{E}_x\left[h\left(X_{T_B}\right)\right]$. Let

$$F(X_0, X_1, \dots) = h\left(X_{T_B(X)}\right) = h\left(X_{T_B(X)}\right) 1_{T_B(X)<\infty},$$

then for $x \in A$ we have $F(x, X_0, X_1, \dots) = F(X_0, X_1, \dots)$. Therefore by the first step analysis (Theorem 8.8) we learn

$$\begin{aligned} u(x) &= \mathbb{E}_x h\left(X_{T_B(X)}\right) = \mathbb{E}_x F(x, X_1, \dots) = \sum_{y\in S} p(x,y)\, \mathbb{E}_y F(x, X_0, X_1, \dots) \\ &= \sum_{y\in S} p(x,y)\, \mathbb{E}_y F(X_0, X_1, \dots) = \sum_{y\in S} p(x,y)\, \mathbb{E}_y\left[h\left(X_{T_B(X)}\right)\right] \\ &= \sum_{y\in A} p(x,y)\, \mathbb{E}_y\left[h\left(X_{T_B(X)}\right)\right] + \sum_{y\in B} p(x,y)\, h(y) \\ &= \sum_{y\in A} p(x,y)\, u(y) + \sum_{y\in B} p(x,y)\, h(y). \end{aligned}$$

■

Theorem 8.13 (Travel averages). *Given $g : A \to [0,\infty]$, let $w(x) := \mathbb{E}_x\left[\sum_{n<T_B} g\left(X_n\right)\right]$. Then $w(x)$ satisfies*

$$w(x) = \sum_{y\in A} p(x,y)\, w(y) + g(x) \text{ for all } x \in A. \tag{8.11}$$

In matrix notation this becomes,

$$\mathbf{w} = \mathbf{Q}\mathbf{w} + \mathbf{g} \implies \mathbf{w} = (I-\mathbf{Q})^{-1}\mathbf{g}$$

so that

$$\mathbb{E}_x\left[\sum_{n<T_B} g\left(X_n\right)\right] = \left[(I-\mathbf{Q})^{-1}\mathbf{g}\right]_x.$$

The following two special cases are of most interest;

1. *Suppose $g(x) = \delta_y(x)$ for some $y \in A$, then $\sum_{n<T_B} g\left(X_n\right) = \sum_{n<T_B} \delta_y\left(X_n\right)$ is the number of visits of the chain to y and*

$$\begin{aligned} &\mathbb{E}_x\left(\#\text{ visits to } y \text{ before hitting } B\right) \\ &= \mathbb{E}_x\left[\sum_{n<T_B} \delta_y\left(X_n\right)\right] = (I-\mathbf{Q})^{-1}_{x,y}. \end{aligned}$$

2. *Suppose that $g(x) = 1$, then $\sum_{n<T_B} g\left(X_n\right) = T_B$ and we may conclude that*

$$\mathbb{E}_x\left[T_B\right] = \left[(I-\mathbf{Q})^{-1}\mathbf{1}\right]_x$$

where $\mathbf{1}$ is the column vector consisting of all ones.

Proof. Let $F(X_0, X_1, \dots) = \sum_{n<T_B(X_0,X_1,\dots)} g\left(X_n\right)$ be the sum of the values of g along the chain before its first exit from A, i.e. entrance into B. With this interpretation in mind, if $x \in A$, it is easy to see that

$$\begin{aligned} F(x, X_0, X_1, \dots) &= \begin{cases} g(x) & \text{if } X_0 \in B \\ g(x) + F(X_0, X_1, \dots) & \text{if } X_0 \in A \end{cases} \\ &= g(x) + 1_{X_0\in A} \cdot F(X_0, X_1, \dots). \end{aligned}$$

Therefore by the first step analysis (Theorem 8.8) it follows that

$$\begin{aligned} w(x) &= \mathbb{E}_x F(X_0, X_1, \dots) = \sum_{y\in S} p(x,y)\, \mathbb{E}_y F(x, X_0, X_1, \dots) \\ &= \sum_{y\in S} p(x,y)\, \mathbb{E}_y\left[g(x) + 1_{X_0\in A} \cdot F(X_0, X_1, \dots)\right] \\ &= g(x) + \sum_{y\in A} p(x,y)\, \mathbb{E}_y\left[F(X_0, X_1, \dots)\right] \\ &= g(x) + \sum_{y\in A} p(x,y)\, w(y). \end{aligned}$$

8.3 Finite state space examples

Example 8.14. Consider the Markov chain determined by

$$\mathbf{P} = \begin{bmatrix} 1 & 2 & 3 & 4 \\ \hline 0 & 1/3 & 1/3 & 1/3 \\ 3/4 & 1/8 & 1/8 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \end{matrix}$$

whose hitting diagram is given in Figure 8.1.Notice that 3 and 4 are absorb-

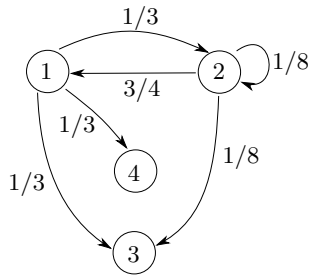


Fig. 8.1. For this chain the states 3 and 4 are absorbing.

ing states. Let $h_i = P_i(X_n \text{ hits } 3) = P(X_n \text{ hits 3 before 4})$ for $i = 1, 2, 3, 4$. Clearly $h_3 = 1$ while $h_4 = 0$ and by the first step analysis we have

$$h_i = P_i(X_n \text{ hits } 3) = \sum_{j=1}^{4} P_i(X_n \text{ hits } 3 | X_1 = j)\, p(i,j) = \sum_{j=1}^{4} p(i,j)\, h_j,$$

and hence

$$\begin{aligned} h_1 &= \frac{1}{3}h_2 + \frac{1}{3}h_3 + \frac{1}{3}h_4 = \frac{1}{3}h_2 + \frac{1}{3} \\ h_2 &= \frac{3}{4}h_1 + \frac{1}{8}h_2 + \frac{1}{8}h_3 = \frac{3}{4}h_1 + \frac{1}{8}h_2 + \frac{1}{8}. \end{aligned} \tag{8.12}$$

Solving

$$h_1 = \frac{1}{3}h_2 + \frac{1}{3} \text{ and } h_2 = \frac{3}{4}h_1 + \frac{1}{8}h_2 + \frac{1}{8}$$

for h_1 and h_2 shows,

$$\begin{aligned} P_1(X_n \text{ hits } 3) = h_1 &= \frac{8}{15} \cong 0.533\,33 \\ P_2(X_n \text{ hits } 3) = h_2 &= \frac{3}{5}. \end{aligned}$$

Similarly if we let $h_i = P_i(X_n \text{ hits } 4)$ instead, from Eqs. (**??** with $h_3 = 0$ and $h_4 = 1$, we find

$$\begin{aligned} h_1 &= \frac{1}{3}h_2 + \frac{1}{3} \\ h_2 &= \frac{3}{4}h_1 + \frac{1}{8}h_2 \end{aligned}$$

which has solutions,

$$\begin{aligned} P_1(X_n \text{ hits } 4) = h_1 &= \frac{7}{15} = 0.466\,67 \text{ and} \\ P_2(X_n \text{ hits } 4) = h_2 &= \frac{2}{5} = 0.4. \end{aligned}$$

Of course we did not really need to compute these, since

$$\begin{aligned} P_1(X_n \text{ hits } 3) + P_1(X_n \text{ hits } 4) &= 1 \text{ and} \\ P_2(X_n \text{ hits } 3) + P_2(X_n \text{ hits } 4) &= 1. \end{aligned}$$

Similarly, if $T = T_{\{3,4\}}$ is the first hitting time of $\{3,4\}$ and $u_i := \mathbb{E}_i T$, we have,

$$u_i = \sum_{j=1}^{4} \mathbb{E}_i[T|X_1 = j]\, p(i,j)$$

where

$$\mathbb{E}_i[T|X_1 = j] = \begin{cases} 1 & \text{if } j \in \{3,4\} \\ 1 + \mathbb{E}_j T = 1 + u_j & \text{if } j \in \{1,2\} \end{cases}.$$

Therefore it follows that

$$u_i = \sum_{j=1}^{4} 1 p(i,j) + \sum_{j=1}^{2} p(i,j)\, u_j = 1 + \sum_{j=1}^{2} p(i,j)\, u_j$$

and this leads to the equations,

$$\begin{aligned} u_1 &= 1 + \frac{1}{3}u_2 \\ u_2 &= 1 + \frac{3}{4}u_1 + \frac{1}{8}u_2 \end{aligned}$$

which has solutions

$$\mathbb{E}_1[T] = u_1 = \frac{29}{15} \text{ and}$$
$$\mathbb{E}_2[T] = u_2 = \frac{14}{5}.$$

Example 8.15 (Example 8.14 revisited). We may also consider Example 8.14 using the matrix formalism. For this we have

$$\mathbf{P}' = \begin{matrix} & 1 & 2 & 3 & 4 \\ \end{matrix}\begin{bmatrix} 0 & 1/3 & 1/3 & 1/3 \\ 3/4 & 1/8 & 1/8 & 0 \end{bmatrix}\begin{matrix} 1 \\ 2 \end{matrix},$$

$$\mathbf{Q} = \begin{matrix} & 1 & 2 \\ \end{matrix}\begin{bmatrix} 0 & 1/3 \\ 3/4 & 1/8 \end{bmatrix}\begin{matrix} 1 \\ 2 \end{matrix}, \text{ and } \mathbf{R} = \begin{matrix} & 3 & 4 \\ \end{matrix}\begin{bmatrix} 1/3 & 1/3 \\ 1/8 & 0 \end{bmatrix}\begin{matrix} 1 \\ 2 \end{matrix}.$$

Matrix manipulations now show,

$$\mathbb{E}_i(\#\text{ visits to } j \text{ before hitting } \{3,4\}) = (I-\mathbf{Q})^{-1} = \begin{matrix} i\backslash j & 1 & 2 \\ 1 \\ 2 \end{matrix}\begin{bmatrix} \frac{7}{5} & \frac{8}{15} \\ \frac{6}{5} & \frac{8}{5} \end{bmatrix} = \begin{bmatrix} 1.4 & 0.53333 \\ 1.2 & 1.6 \end{bmatrix},$$

$$\mathbb{E}_i T_{\{3,4\}} = (I-\mathbf{Q})^{-1}\begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{matrix} i \\ 1 \\ 2 \end{matrix}\begin{bmatrix} \frac{29}{15} \\ \frac{14}{5} \end{bmatrix} = \begin{bmatrix} 1.933\,3 \\ 2.8 \end{bmatrix} \text{ and}$$

$$P_i\left(X_{T_{\{3,4\}}} = j\right) = (I-\mathbf{Q})^{-1}\mathbf{R} = \begin{matrix} i\backslash j & 3 & 4 \\ 1 \\ 2 \end{matrix}\begin{bmatrix} \frac{8}{15} & \frac{7}{15} \\ \frac{3}{5} & \frac{2}{5} \end{bmatrix}.$$

The output of one simulation from `www.zweigmedia.com/RealWorld/markov/markov.html` is in Figure 8.2 below.

Example 8.16. Let us continue the rat in the maze Exercise 7.1 and now suppose that room 3 contains food while room 7 contains a mouse trap.

$$\begin{bmatrix} 1 & 2 & 3\text{ (food)} \\ 4 & 5 & 6 \\ 7\text{ (trap)} & & \end{bmatrix}.$$

Recall that the transition matrix for this chain with sites 3 and 7 absorbing is given by,

$$\mathbf{P} = \begin{matrix} & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ \end{matrix}\begin{bmatrix} 0 & 1/2 & 0 & 1/2 & 0 & 0 & 0 \\ 1/3 & 0 & 1/3 & 0 & 1/3 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1/3 & 0 & 0 & 0 & 1/3 & 0 & 1/3 \\ 0 & 1/3 & 0 & 1/3 & 0 & 1/3 & 0 \\ 0 & 0 & 1/2 & 0 & 1/2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}\begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \end{matrix},$$

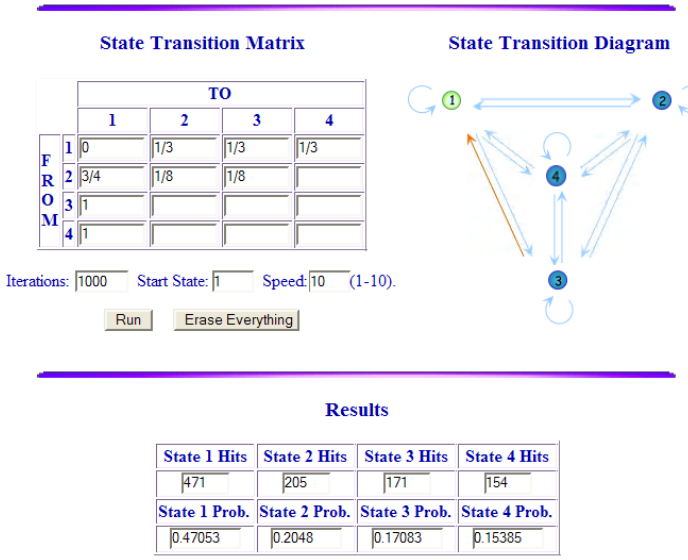


Fig. 8.2. In this run, rather than making sites 3 and 4 absorbing, we have made them transition back to 1. I claim now to get an approximate value for $P_1(X_n \text{ hits } 3)$ we should compute: (State 3 Hits)/(State 3 Hits + State 4 Hits). In this example we will get $171/(171+154) = 0.526\,15$ which is a little lower than the predicted value of 0.533. You can try your own runs of this simulator.

see Figure 8.3 for the corresponding jump diagram for this chain.

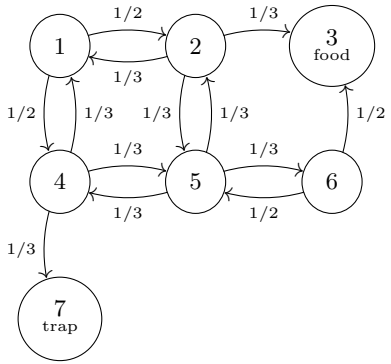


Fig. 8.3. The jump diagram for our proverbial rat in the maze. Here we assume the rat is "absorbed" at sites 3 and 7

We would like to compute the probability that the rat reaches the food before he is trapped. To answer this question we let $A = \{1,2,4,5,6\}$, $B = \{3,7\}$, and $T := T_B$ be the first hitting time of B. Then deleting the 3 and 7 rows of $\mathbf{P}$ leaves the matrix,

$$\mathbf{P}' = \begin{array}{c} \begin{matrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{matrix} \\ \begin{bmatrix} 0 & 1/2 & 0 & 1/2 & 0 & 0 & 0 \\ 1/3 & 0 & 1/3 & 0 & 1/3 & 0 & 0 \\ 1/3 & 0 & 0 & 0 & 1/3 & 0 & 1/3 \\ 0 & 1/3 & 0 & 1/3 & 0 & 1/3 & 0 \\ 0 & 0 & 1/2 & 0 & 1/2 & 0 & 0 \end{bmatrix} \end{array} \begin{matrix} 1 \\ 2 \\ 4 \\ 5 \\ 6 \end{matrix}.$$

Deleting the 3 and 7 columns of $\mathbf{P}'$ gives

$$\mathbf{Q} = \mathbf{P}_{A,A} = \begin{array}{c} \begin{matrix} 1 & 2 & 4 & 5 & 6 \end{matrix} \\ \begin{bmatrix} 0 & 1/2 & 1/2 & 0 & 0 \\ 1/3 & 0 & 0 & 1/3 & 0 \\ 1/3 & 0 & 0 & 1/3 & 0 \\ 0 & 1/3 & 1/3 & 0 & 1/3 \\ 0 & 0 & 0 & 1/2 & 0 \end{bmatrix} \end{array} \begin{matrix} 1 \\ 2 \\ 4 \\ 5 \\ 6 \end{matrix}$$

and deleting the $1,2,4,5$, and 6 columns of $\mathbf{P}'$ gives

$$\mathbf{R} = \mathbf{P}_{A,B} = \begin{array}{c} \begin{matrix} 3 & 7 \end{matrix} \\ \begin{bmatrix} 0 & 0 \\ 1/3 & 0 \\ 0 & 1/3 \\ 0 & 0 \\ 1/2 & 0 \end{bmatrix} \end{array} \begin{matrix} 1 \\ 2 \\ 4 \\ 5 \\ 6 \end{matrix}.$$

Therefore,

$$I - \mathbf{Q} = \begin{bmatrix} 1 & -\frac{1}{2} & -\frac{1}{2} & 0 & 0 \\ -\frac{1}{3} & 1 & 0 & -\frac{1}{3} & 0 \\ -\frac{1}{3} & 0 & 1 & -\frac{1}{3} & 0 \\ 0 & -\frac{1}{3} & -\frac{1}{3} & 1 & -\frac{1}{3} \\ 0 & 0 & 0 & -\frac{1}{2} & 1 \end{bmatrix},$$

and using a computer algebra package we find

$$\mathbb{E}_i\left[\#\text{ visits to } j \text{ before hitting } \{3,7\}\right] = (I-\mathbf{Q})^{-1} = \begin{array}{c} \begin{matrix} 1 & 2 & 4 & 5 & 6 & j \end{matrix} \\ \begin{bmatrix} \frac{11}{6} & \frac{5}{4} & \frac{5}{4} & 1 & \frac{1}{3} \\ \frac{5}{6} & \frac{7}{4} & \frac{3}{4} & 1 & \frac{1}{3} \\ \frac{5}{6} & \frac{3}{4} & \frac{7}{4} & 1 & \frac{1}{3} \\ \frac{2}{3} & 1 & 1 & 2 & \frac{2}{3} \\ \frac{1}{3} & \frac{1}{2} & \frac{1}{2} & 1 & \frac{4}{3} \end{bmatrix} \end{array} \begin{matrix} i \\ 1 \\ 2 \\ 4 \\ 5 \\ 6 \end{matrix}.$$

In particular we may conclude,

$$\begin{bmatrix} \mathbb{E}_1 T \\ \mathbb{E}_2 T \\ \mathbb{E}_4 T \\ \mathbb{E}_5 T \\ \mathbb{E}_6 T \end{bmatrix} = (I-\mathbf{Q})^{-1}\mathbf{1} = \begin{bmatrix} \frac{17}{3} \\ \frac{14}{3} \\ \frac{14}{3} \\ \frac{16}{3} \\ \frac{11}{3} \end{bmatrix},$$

and

$$\begin{bmatrix} P_1(X_T=3) & P_1(X_T=7) \\ P_2(X_T=3) & P_2(X_T=3) \\ P_4(X_T=3) & P_4(X_T=3) \\ P_5(X_T=3) & P_5(X_T=3) \\ P_6(X_T=3) & P_6(X_T=7) \end{bmatrix} = (I-\mathbf{Q})^{-1}\mathbf{R} = \begin{array}{c} \begin{matrix} 3 & 7 \end{matrix} \\ \begin{bmatrix} \frac{7}{12} & \frac{5}{12} \\ \frac{3}{4} & \frac{1}{4} \\ \frac{5}{12} & \frac{7}{12} \\ \frac{2}{3} & \frac{1}{3} \\ \frac{5}{6} & \frac{1}{6} \end{bmatrix} \end{array} \begin{matrix} 1 \\ 2 \\ 4 \\ 5 \\ 6 \end{matrix}.$$

Since the event of hitting 3 before 7 is the same as the event $\{X_T = 3\}$, the desired hitting probabilities are

$$\begin{bmatrix} P_1(X_T=3) \\ P_2(X_T=3) \\ P_4(X_T=3) \\ P_5(X_T=3) \\ P_6(X_T=3) \end{bmatrix} = \begin{bmatrix} \frac{7}{12} \\ \frac{3}{4} \\ \frac{5}{12} \\ \frac{2}{3} \\ \frac{5}{6} \end{bmatrix}.$$

We can also derive these hitting probabilities from scratch using the first step analysis. In order to do this let

$$h_i = P_i(X_T = 3) = P_i(X_n \text{ hits 3 (food) before 7(trapped)}).$$

By the first step analysis we will have,

$$\begin{aligned} h_i &= \sum_j P_i(X_T = 3 | X_1 = j) P_i(X_1 = j) \\ &= \sum_j p(i,j) P_i(X_T = 3 | X_1 = j) \\ &= \sum_j p(i,j) P_j(X_T = 3) \\ &= \sum_j p(i,j) h_j \end{aligned}$$

where $h_3 = 1$ and $h_7 = 0$. Looking at the jump diagram in Figure 8.3 we easily find

$$h_1 = \frac{1}{2}(h_2 + h_4)$$
$$h_2 = \frac{1}{3}(h_1 + h_3 + h_5) = \frac{1}{3}(h_1 + 1 + h_5)$$
$$h_4 = \frac{1}{3}(h_1 + h_5 + h_7) = \frac{1}{3}(h_1 + h_5)$$
$$h_5 = \frac{1}{3}(h_2 + h_4 + h_6)$$
$$h_6 = \frac{1}{2}(h_3 + h_5) = \frac{1}{2}(1 + h_5)$$

and the solutions to these equations are (as seen before) given by

$$\left[h_1 = \frac{7}{12}, h_2 = \frac{3}{4}, h_4 = \frac{5}{12}, h_5 = \frac{2}{3}, h_6 = \frac{5}{6}\right]. \tag{8.13}$$

Similarly, if

$$k_i := P_i(X_T = 7) = P_i(X_n \text{ is trapped before dinner}),$$

we need only use the above equations with h replaced by k and now taking $k_3 = 0$ and $k_7 = 1$ to find,

$$k_1 = \frac{1}{2}(k_2 + k_4)$$
$$k_2 = \frac{1}{3}(k_1 + k_5)$$
$$k_4 = \frac{1}{3}(k_1 + k_5 + 1)$$
$$k_5 = \frac{1}{3}(k_2 + k_4 + k_6)$$
$$k_6 = \frac{1}{2}k_5$$

and then solve to find,

$$\left[k_1 = \frac{5}{12}, k_2 = \frac{1}{4}, k_4 = \frac{7}{12}, k_5 = \frac{1}{3}, k_6 = \frac{1}{6}\right]. \tag{8.14}$$

Notice that the sum of the hitting probabilities in Eqs. (8.13) and (8.14) add up to 1 as they should.

Example 8.17 (A modified rat maze). Here is the modified maze,

$$\begin{bmatrix} 1 & 2 & 3(\text{food}) \\ 4 & 5 & \\ 6(\text{trap}) & & \end{bmatrix}.$$

We now let $T = T_{\{3,6\}}$ be the first time to absorbtion – we assume that 3 and 6 made are absorbing states.[2] The transition matrix is given by

$$\mathbf{P} = \begin{array}{c} \begin{matrix} 1 & 2 & 3 & 4 & 5 & 6 \end{matrix} \\ \begin{bmatrix} 0 & 1/2 & 0 & 1/2 & 0 & 0 \\ 1/3 & 0 & 1/3 & 0 & 1/3 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 1/3 & 0 & 0 & 0 & 1/3 & 1/3 \\ 0 & 1/2 & 0 & 1/2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \end{array} \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{matrix}.$$

The corresponding $\mathbf{Q}$ and $\mathbf{R}$ matrices in this case are;

$$\mathbf{Q} = \begin{array}{c} \begin{matrix} 1 & 2 & 4 & 5 \end{matrix} \\ \begin{bmatrix} 0 & 1/2 & 1/2 & 0 \\ 1/3 & 0 & 0 & 1/3 \\ 1/3 & 0 & 0 & 1/3 \\ 0 & 1/2 & 1/2 & 0 \end{bmatrix} \end{array} \begin{matrix} 1 \\ 2 \\ 4 \\ 5 \end{matrix}, \text{ and } \quad \mathbf{R} = \begin{array}{c} \begin{matrix} 3 & 6 \end{matrix} \\ \begin{bmatrix} 0 & 0 \\ 1/3 & 0 \\ 0 & 1/3 \\ 0 & 0 \end{bmatrix} \end{array} \begin{matrix} 1 \\ 2 \\ 4 \\ 5 \end{matrix}.$$

After some matrix manipulation we then learn,

$$\mathbb{E}_i[\#\text{ visits to } j] = (I_4 - \mathbf{Q})^{-1} = \begin{array}{c} \begin{matrix} 1 & 2 & 4 & 5 \end{matrix} \\ \begin{bmatrix} 2 & \frac{3}{2} & \frac{3}{2} & 1 \\ 1 & 2 & 1 & 1 \\ 1 & 1 & 2 & 1 \\ 1 & \frac{3}{2} & \frac{3}{2} & 2 \end{bmatrix} \end{array} \begin{matrix} 1 \\ 2 \\ 4 \\ 5 \end{matrix},$$

$$P_i[X_T = j] = (I_4 - \mathbf{Q})^{-1}\mathbf{R} = \begin{array}{c} \begin{matrix} 3 & 6 \end{matrix} \\ \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{2}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{2}{3} \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix} \end{array} \begin{matrix} 1 \\ 2 \\ 4 \\ 5 \end{matrix},$$

$$\mathbb{E}_i[T] = (I_4 - \mathbf{Q})^{-1} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 6 \\ 5 \\ 5 \\ 6 \end{bmatrix} \begin{matrix} 1 \\ 2 \\ 4 \\ 5 \end{matrix}.$$

So for example, $P_4(X_T = 3(\text{food})) = 1/3$, $\mathbb{E}_4(\text{Number of visits to } 1) = 1$, $\mathbb{E}_5(\text{Number of visits to } 2) = 3/2$ and $\mathbb{E}_1 T = \mathbb{E}_5 T = 6$ and $\mathbb{E}_2 T = \mathbb{E}_4 T = 5$.

[2] It is not necessary to make states 3 and 6 absorbing. In fact it does matter at all what the transition probabilities are for the chain for leaving either of the states 3 or 6 since we are going to stop when we hit these states. This is reflected in the fact that the first thing we will do in the first step analysis is to delete rows 3 and 6 from P. Making 3 and 6 absorbing simply saves a little ink.

For practice let us compute $h_i = P_i(X_n \text{ hits 3 before 6}) = P_i(X_T = 3(food))$. By the first step analysis we have,

$$\begin{aligned} h_6 &= 0 \\ h_3 &= 1 \\ h_5 &= \frac{1}{2}(h_2 + h_4) \\ h_4 &= \frac{1}{3}(h_1 + h_5 + h_6) \\ h_2 &= \frac{1}{3}(h_1 + h_3 + h_5) \\ h_1 &= \frac{1}{2}(h_2 + h_4) \end{aligned}$$

which have solutions

$$\left[h_1 = \frac{1}{2}, h_2 = \frac{2}{3}, h_3 = 1, h_4 = \frac{1}{3}, h_5 = \frac{1}{2}, h_6 = 0\right]. \tag{8.15}$$

Similarly if $h_i = P_i(X_n \text{ hits 6 before 3}) = P_i(X_T = 6)$ we have

$$\begin{aligned} h_6 &= 1 \\ h_3 &= 0 \\ h_5 &= \frac{1}{2}(h_2 + h_4) \\ h_4 &= \frac{1}{3}(h_1 + h_5 + h_6) \\ h_2 &= \frac{1}{3}(h_1 + h_3 + h_5) \\ h_1 &= \frac{1}{2}(h_2 + h_4) \end{aligned}$$

which have solutions

$$\left[h_1 = \frac{1}{2}, h_2 = \frac{1}{3}, h_3 = 0, h_4 = \frac{2}{3}, h_5 = \frac{1}{2}, h_6 = 1\right]. \tag{8.16}$$

Notice that the sum of the hitting probabilities in Eqs. (8.15) and (8.16) add up to 1 as they should. These results are in agreement with our previous results using the matrix method as well.

Exercise 8.1 (III.4.P11 on p.132). An urn contains two red and two green balls. The balls are chosen at random, one by one, and removed from the urn. The selection process continues until all of the green balls have been removed from the urn. What is the probability that a single red ball is in the urn at the time that the last green ball is chosen?

Solution to Exercise (III.4.P11 on p.132). Let's choose the states to be $(G, R) = (i, j)$ with $i, j = 0, 1, 2$ so that $(1, 2)$ implies that there is one green ball and two red balls in the urn. Let $B = \{(0,0), (0,1), (0,2\}$,

$$T = T_B = \min\{n \geq 0 : X_n = (0,0) \text{ or } (0,1) \text{ or } (0,2)\}.$$

We wish to compute $P(X_T = (0,1) | X_0 = (2,2))$. The transition matrix for this chain is given by;

$$\mathbf{P} = \begin{array}{c} \begin{array}{ccccccccc} (0,0) & (0,1) & (0,2) & (1,0) & (1,1) & (1,2) & (2,0) & (2,1) & (2,2) \end{array} \\ \left[\begin{array}{ccccccccc} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 1/2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/3 & 0 & 2/3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2/3 & 0 & 1/3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1/2 & 0 & 1/2 & 0 \end{array}\right] \end{array} \begin{array}{c} (0,0) \\ (0,1) \\ (0,2) \\ (1,0) \\ (1,1) \\ (1,2) \\ (2,0) \\ (2,1) \\ (2,2) \end{array}.$$

Using the matrix method. First we remove the $\{(0,0), (0,1), (0,2)\}$ - row of $\mathbf{P}$;

$$\mathbf{P}' = \begin{array}{c} \begin{array}{ccccccccc} (0,0) & (0,1) & (0,2) & (1,0) & (1,1) & (1,2) & (2,0) & (2,1) & (2,2) \end{array} \\ \left[\begin{array}{ccccccccc} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 1/2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/3 & 0 & 2/3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2/3 & 0 & 1/3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1/2 & 0 & 1/2 & 0 \end{array}\right] \end{array} \begin{array}{c} (1,0) \\ (1,1) \\ (1,2) \\ (2,0) \\ (2,1) \\ (2,2) \end{array}$$

and now form $\mathbf{Q}$ by removing the $\{(0,0), (0,1), (0,2)\}$ columns of $\mathbf{P}'$ and $\mathbf{R}$ by keeping the $\{(0,0), (0,1), (0,2)\}$ columns of $\mathbf{P}'$;

$$\mathbf{Q}=\left[\begin{array}{cccccc|c}(1,0)&(1,1)&(1,2)&(2,0)&(2,1)&(2,2)&\\\hline 0&0&0&0&0&0&(1,0)\\1/2&0&0&0&0&0&(1,1)\\0&2/3&0&0&0&0&(1,2)\\1&0&0&0&0&0&(2,0)\\0&2/3&0&1/3&0&0&(2,1)\\0&0&1/2&0&1/2&0&(2,2)\end{array}\right]$$

$$\mathbf{R}=\left[\begin{array}{c|ccc}&(0,0)&(0,1)&(0,2)\\\hline(1,0)&1&0&0\\(1,1)&0&1/2&0\\(1,2)&0&0&1/3\\(2,0)&0&0&0\\(2,1)&0&0&0\\(2,2)&0&0&0\end{array}\right].$$

So

$$P_{(a,b)}\left[X_{T_B}=(c,d)\right]=(I-\mathbf{Q})^{-1}\mathbf{R}=\begin{array}{c|c}(a,b)\setminus(c,d)&\begin{matrix}(0,0)&(0,1)&(0,2)\end{matrix}\\\hline(1,0)&\begin{bmatrix}1&0&0\end{bmatrix}\\(1,1)&\begin{bmatrix}\frac{1}{2}&\frac{1}{2}&0\end{bmatrix}\\(1,2)&\begin{bmatrix}\frac{1}{3}&\frac{1}{3}&\frac{1}{3}\end{bmatrix}\\(2,0)&\begin{bmatrix}1&0&0\end{bmatrix}\\(2,1)&\begin{bmatrix}\frac{2}{3}&\frac{1}{3}&0\end{bmatrix}\\(2,2)&\begin{bmatrix}\frac{1}{2}&\frac{1}{3}&\frac{1}{6}\end{bmatrix}\end{array}$$

and therefore,

$$P_{(2,2)}(X_T=(0,1))=P(X_T=(0,1)|X_0=(2,2))=1/3.$$

Theorem 8.18. *Let $h:B\to[0,\infty]$ and $g:A\to[0,\infty]$ be given and for $x\in S$. If we let* [3]

$$w(x):=\mathbb{E}_x\left[h\left(X_{T_B}\right)\cdot\sum_{n<T_B}g\left(X_n\right):T_B<\infty\right]\text{ and}$$
$$g_h(x)=g(x)\,\mathbb{E}_x\left[h\left(X_{T_B}\right):T_B<\infty\right],$$

then

[3] Recall from Theorem 8.12 that $\mathbf{u}_h=(I-\mathbf{Q})^{-1}\mathbf{Rh}$, i.e. $u=h$ on B and u satisfies

$$u(x)=\sum_{y\in A}p(x,y)\,u(y)+\sum_{y\in B}p(x,y)\,h(y)\text{ for all }x\in A.$$

$$w(x)=\mathbb{E}_x\left[\sum_{n<T_B}g_h\left(X_n\right):T_B<\infty\right].\tag{8.17}$$

Remark 8.19. Recall that we can find $u_h(x):=\mathbb{E}_x\left[h\left(X_{T_B}\right):T_B<\infty\right]$ using Theorem 8.12 and then we can solve for $w(x)$ using Theorem 8.13 with g replaced by $g_h(x)=g(x)\,u_h(x)$. So in the matrix language we solve for $w(x)$ as follows;

$$\begin{aligned}\mathbf{u}_h&:=(I-\mathbf{Q})^{-1}\mathbf{Rh},\\\mathbf{g_h}&:=\mathbf{g}*\mathbf{u}_h,\text{ and}\\\mathbf{w}&=(I-\mathbf{Q})^{-1}\mathbf{g}_h,\end{aligned}$$

where $[\mathbf{a}*\mathbf{b}]_x:=\mathbf{a}_x\cdot\mathbf{b}_x$ – the entry by entry product of column vectors.

Proof. First proof. Let $H(X):=h\left(X_{T_B}\right)1_{T_B<\infty}$, then using $1_{n<T_B(X)}=1_{X_0\in A,\dots,X_n\in A}$ and

$$H(X_0,\dots,X_{n-1},X_n,\dots)=H(X_n,X_{n+1},\dots)\text{ when }X_0,\dots,X_n\in A$$

along with the Markov property in Theorem 8.1 shows;

$$\begin{aligned}w(x)&=\sum_{n=0}^{\infty}\mathbb{E}_x\left[H(X)\cdot1_{n<T_B}g\left(X_n\right)\right]\\&=\sum_{n=0}^{\infty}\mathbb{E}_x\left[H(X)\cdot1_{X_0\in A,\dots,X_n\in A}g\left(X_n\right)\right]\\&=\sum_{n=0}^{\infty}\mathbb{E}_x\left[\mathbb{E}_{X_n}^{(Y)}\left[H\left(X_0,\dots,X_{n-1},Y\right)\right]\cdot1_{X_0\in A,\dots,X_n\in A}g\left(X_n\right)\right]\\&=\sum_{n=0}^{\infty}\mathbb{E}_x\left[\mathbb{E}_{X_n}^{(Y)}H(Y)\cdot1_{X_0\in A,\dots,X_n\in A}g\left(X_n\right)\right]\\&=\sum_{n=0}^{\infty}\mathbb{E}_x\left[u_h\left(X_n\right)\cdot1_{X_0\in A,\dots,X_n\in A}g\left(X_n\right)\right]\\&=\sum_{n=0}^{\infty}\mathbb{E}_x\left[u_h\left(X_n\right)\cdot g\left(X_n\right)1_{n<T_B}\right]\\&=\mathbb{E}_x\left[\sum_{n<T_B}u_h\left(X_n\right)\cdot g\left(X_n\right)\right]=\mathbb{E}_x\left[\sum_{n<T_B}g_h\left(X_n\right)\right].\end{aligned}$$

Second proof. Let $G(X):=\sum_{n<T_B}g\left(X_n\right)$ and observe that

$$H(x,Y)=\begin{cases}H(Y) \text{ if } x\in A\\ h(x) \text{ if } x\in B\end{cases} \text{ and}$$
$$G(x,Y)=g(x)+G(Y)$$

and so by the first step analysis we find,

$$\begin{aligned}w(x)&=\mathbb{E}_x\left[H(X)G(X)\right]=\mathbb{E}_{p(x,\cdot)}\left[H(x,Y)G(x,Y)\right]\\&=\mathbb{E}_{p(x,\cdot)}\left[H(x,Y)(g(x)+G(Y))\right]\\&=g(x)\,\mathbb{E}_{p(x,\cdot)}\left[H(x,Y)\right]+\mathbb{E}_{p(x,\cdot)}\left[H(x,Y)G(Y)\right].\end{aligned}$$

The first step analysis also shows (see the proof of Theorem 8.12)

$$u_h(x):=\mathbb{E}_x\left[h(X_{T_B})1_{T_B<\infty}\right]=\mathbb{E}_x\left[H(X)\right]=\mathbb{E}_{p(x,\cdot)}\left[H(x,Y)\right].$$

and therefore,

$$w(x)=g(x)u_h(x)+\mathbb{E}_{p(x,\cdot)}\left[H(x,Y)G(Y)\right]$$

Since $G(Y)=0$ if $Y_0\in B$ and $H(x,Y)=H(Y)$ if $Y_0\in A$ we find,

$$\begin{aligned}\mathbb{E}_{p(x,\cdot)}\left[H(x,Y)G(Y)\right]&=\sum_{x\in S}p(x,y)\,\mathbb{E}_y\left[H(x,Y)G(Y)\right]\\&=\sum_{x\in A}p(x,y)\,\mathbb{E}_y\left[H(x,Y)G(Y)\right]\\&=\sum_{x\in A}p(x,y)\,\mathbb{E}_y\left[H(Y)G(Y)\right]\\&=\sum_{x\in A}p(x,y)\,w(y)\end{aligned}$$

and hence

$$w(x)=g(x)u_h(x)+\sum_{x\in A}p(x,y)\,w(y)=g_h(x)+\sum_{x\in A}p(x,y)\,w(y).$$

But Theorem 8.13 with g replaced by g_h then shows w is given by Eq. (8.17). ■

Example 8.20 (A possible carnival game). Suppose that B is the disjoint union of L and W and suppose that you win $\sum_{n<T_B}g(X_n)$ if you end in W and win nothing when you end in L. What is the least we can expect to have to pay to play this game and where in $A:=S\setminus B$ should we choose to start the game. To answer these questions we should compute our expected winnings $(w(x))$ for each starting point $x\in A$;

$$w(x)=\mathbb{E}_x\left[1_W(X_{T_B})\sum_{n<T_B}g(X_n)\right].$$

Once we find w we should expect to pay at least $C:=\max_{x\in A}w(x)$ and we should start at a location $x_0\in A$ where $w(x_0)=\max_{x\in A}w(x)=C$. As an application of Theorem 8.18 we know that

$$w(x)=\left[(\mathbf{I}-\mathbf{Q})^{-1}\mathbf{g}_h\right]_x$$

where[4]

$$g_h(x)=g(x)\,\mathbb{E}_x\left[1_W(X_{T_B})\right]=g(x)\,P_x\left(X_{T_B}\in W\right).$$

Let us now specialize these results to the chain in Example 8.14 where

$$\mathbf{P}=\left[\begin{array}{cccc|c}1&2&3&4&\\\hline 0&1/3&1/3&1/3&1\\3/4&1/8&1/8&0&2\\0&0&1&0&3\\0&0&0&1&4\end{array}\right]$$

Let us make 4 the winning state and 3 the losing state (i.e. $h(3)=0$ and $h(4)=1$) and let $g=(g(1),g(2))$ be the payoff function. We have already seen that

$$\begin{bmatrix}u_h(1)\\u_h(2)\end{bmatrix}=\begin{bmatrix}P_1(X_{T_B}=4)\\P_2(X_{T_B}=4)\end{bmatrix}=\begin{bmatrix}\frac{7}{15}\\\frac{2}{5}\end{bmatrix}$$

so that $\mathbf{g}*\mathbf{u}_h=\begin{bmatrix}\frac{7}{15}g_1\\\frac{2}{5}g_2\end{bmatrix}$ and therefore

$$\begin{aligned}\begin{bmatrix}w(1)\\w(2)\end{bmatrix}&=(I-\mathbf{Q})^{-1}\begin{bmatrix}\frac{7}{15}g_1\\\frac{2}{5}g_2\end{bmatrix}\\&=\begin{bmatrix}\frac{7}{5}&\frac{8}{15}\\\frac{6}{5}&\frac{8}{5}\end{bmatrix}\begin{bmatrix}\frac{7}{15}g_1\\\frac{2}{5}g_2\end{bmatrix}=\begin{bmatrix}\frac{49}{75}g_1+\frac{16}{75}g_2\\\frac{14}{25}g_1+\frac{16}{25}g_2\end{bmatrix}.\end{aligned}$$

Let us examine a few different choices for g.

1. When $g(1)=32$ and $g(2)=7$, we have

$$\begin{bmatrix}w(1)\\w(2)\end{bmatrix}=\begin{bmatrix}\frac{7}{5}&\frac{8}{15}\\\frac{6}{5}&\frac{8}{5}\end{bmatrix}\begin{bmatrix}\frac{7}{15}32\\\frac{2}{5}7\end{bmatrix}=\begin{bmatrix}\frac{112}{5}\\\frac{112}{5}\end{bmatrix}=\begin{bmatrix}22.4\\22.4\end{bmatrix}$$

   and so it does not matter where we start and we are going to have to pay at least \$22.40 to play.

[4] Intuitively, the effective pay off for a visit to site x is $g(x)\cdot P_x(\text{ we win})+0\cdot P_x(\text{we loose})$.

2. When $g(1)=10=g(2)$, then

$$\begin{bmatrix} w(1) \\ w(2) \end{bmatrix} = \begin{bmatrix} \frac{7}{5} & \frac{8}{15} \\ \frac{6}{5} & \frac{8}{5} \end{bmatrix} \begin{bmatrix} \frac{7}{15}10 \\ \frac{2}{5}10 \end{bmatrix} = \begin{bmatrix} \frac{26}{3} \\ 12 \end{bmatrix} = \begin{bmatrix} 8.666\,7 \\ 12.0 \end{bmatrix}$$

and we should enter the game at site 2. We are going to have to pay at least \$12 to play.

3. If $g(1)=20$ and $g(2)=7$,

$$\begin{bmatrix} w(1) \\ w(2) \end{bmatrix} = \begin{bmatrix} \frac{7}{5} & \frac{8}{15} \\ \frac{6}{5} & \frac{8}{5} \end{bmatrix} \begin{bmatrix} \frac{7}{15}20 \\ \frac{2}{5}7 \end{bmatrix} = \begin{bmatrix} \frac{364}{25} \\ \frac{392}{25} \end{bmatrix} = \begin{bmatrix} 14.56 \\ 15.68 \end{bmatrix}$$

and again we should enter the game at site 2. We are going to have to pay at least \$15.68 to play.

8.4 Random Walk Exercises

Exercise 8.2 (Uniqueness of solutions to 2nd order recurrence relations). Let a,b,c be real numbers with $a\neq 0\neq c$, $\alpha,\beta\in\mathbb{Z}\cup\{\pm\infty\}$ with $\alpha<\beta$, and $g:\mathbb{Z}\cap(\alpha,\beta)\to\mathbb{R}$ be a given function. Show that there is exactly one function $u:[\alpha,\beta]\cap\mathbb{Z}\to\mathbb{R}$ with prescribed values on two two consecutive points in $[\alpha,\beta]\cap\mathbb{Z}$ which satisfies the second order recurrence relation:

$$au(x+1)+bu(x)+cu(x-1)=f(x) \text{ for all } x\in\mathbb{Z}\cap(\alpha,\beta). \tag{8.18}$$

are for $\alpha<x<\beta$. Show; if u and w both satisfy Eq. (8.18) and $u=w$ on two consecutive points in $(\alpha,\beta)\cap\mathbb{Z}$, then $u(x)=w(x)$ for all $x\in[\alpha,\beta]\cap\mathbb{Z}$.

Solution to Exercise (8.2). Suppose we are given $u(x_0)=\mu$ and $u(x_0+1)=\nu$ for some $\mu,\nu\in\mathbb{R}$ and $x_0\in\mathbb{Z}$ such that $\{x_0,x_0+1\}\subset[\alpha,\beta]\cap\mathbb{Z}$. If $x_0-1\in[\alpha,\beta]\cap\mathbb{Z}$, then according to Eq. (8.18),

$$au(x_0+1)+bu(x_0)+cu(x_0-1)=f(x_0)$$

from which we may uniquely determine $u(x_0-1)$. Repeating this procedure we see that we can uniquely determine $u(x)$ for all $x\in[\alpha,x_0]\cap\mathbb{Z}$. Similarly if $x_0+2\in[\alpha,\beta]\cap\mathbb{Z}$ Eq. (8.18) implies,

$$au(x_0+2)+bu(x_0+1)+cu(x_0)=f(x_0+1)$$

from which we may uniquely determine $u(x_0+2)$. Repeating this procedure we see that we can uniquely determine $u(x)$ for all $x\in[x_0+1,\beta]\cap\mathbb{Z}$ and hence there is exactly one function satisfying Eq. (8.18) with $u(x_0)=\mu$ and $u(x_0+1)=\nu$.

Exercise 8.3 (General homogeneous solutions). Let a,b,c be real numbers with $a\neq 0\neq c$, $\alpha,\beta\in\mathbb{Z}\cup\{\pm\infty\}$ with $\alpha<\beta$, and suppose $\{u(x):x\in[\alpha,\beta]\cap\mathbb{Z}\}$ solves the second order homogeneous recurrence relation

$$au(x+1)+bu(x)+cu(x-1)=0 \text{ for all } x\in\mathbb{Z}\cap(\alpha,\beta), \tag{8.19}$$

i.e. Eq. (8.18) with $f(x)\equiv 0$. Show:

1. for any $\lambda\in\mathbb{C}$,
$$a\lambda^{x+1}+b\lambda^x+c\lambda^{x-1}=\lambda^{x-1}p(\lambda) \tag{8.20}$$
where $p(\lambda)=a\lambda^2+b\lambda+c$ is the **characteristic polynomial** associated to Eq. (8.18).
Let $\lambda_\pm=\frac{-b\pm\sqrt{b^2-4ac}}{2a}$ be the roots of $p(\lambda)$ and suppose for the moment that $b^2-4ac\neq 0$. From Eq. (8.18) it follows that for any choice of $A_\pm\in\mathbb{R}$, the function,
$$w(x):=A_+\lambda_+^x+A_-\lambda_-^x, \tag{8.21}$$
solves Eq. (8.18) for all $x\in\mathbb{Z}$.
2. Show there is a unique choice of constants, $A_\pm\in\mathbb{R}$, such that the function $u(x)$ is given by
$$u(x):=A_+\lambda_+^x+A_-\lambda_-^x \text{ for all } \alpha\le x\le\beta.$$
3. Now suppose that $b^2=4ac$ and $\lambda_0:=-b/(2a)$ is the double root of $p(\lambda)$. Show for any choice of A_0 and A_1 in $\mathbb{R}$ that
$$w(x):=(A_0+A_1x)\lambda_0^x \tag{8.22}$$
solves Eq. (8.18) for all $x\in\mathbb{Z}$. **Hint:** Differentiate Eq. (8.20) with respect to λ and then set $\lambda=\lambda_0$.
4. Again show that any function u solving Eq. (8.18) is of the form $u(x)=(A_0+A_1x)\lambda_0^x$ for $\alpha\le x\le\beta$ for some unique choice of constants $A_0,A_1\in\mathbb{R}$.

Solution to Exercise (8.3). Items 1. and 3. follows by a direct verification. Indeed if $u(x)=\lambda^x$, then

$$\begin{aligned} au(x+1)+bu(x)+cu(x-1) &= a\lambda^{x+1}+b\lambda^x+c\lambda^{x-1} \\ &= \lambda^{x-1}\left[a\lambda^2+b\lambda+c\right]=\lambda^{x-1}p(\lambda). \end{aligned}$$

Since the equation is linear it now follows w given in Eq. (8.21) solves Eq. (8.19). It can be directly verified that $x\lambda_0^x$ solves Eq. (8.19) when $b^2=4ac$ and then linearity again shows that Eq. (8.22) solves Eq. (8.19) in this case. Alternatively, if we differentiate Eq. (8.20) to find

$$a(x+1)\lambda^x + bx\lambda^{x-1} + c(x-1)\lambda^{x-2} = (x-1)\lambda^{x-2}p(\lambda) + \lambda^{x-1}p'(\lambda).$$

Since λ_0 is a double root for p, $p(\lambda_0) = 0 = p'(\lambda_0)$ so if we evaluate the previous equation at $\lambda = \lambda_0$ and multiply the result by λ_0 we find,

$$0 = a(x+1)\lambda_0^{x+1} + bx\lambda_0^x + c(x-1)\lambda_0^{x-1},$$

i.e. $u(x) = x\lambda_0^x$ solves Eq. (8.19).

According to Exercises 8.2, to prove items 2. and 4 we have to show that we may adjust the constants $A_\pm$ or A_0 and A_1 so that $u(x) = \mu$ and $u(x+1) = \nu$ where μ and ν are given numbers in $\mathbb{R}$ and $x \in [\alpha, \beta) \cap \mathbb{Z}$. In the case of item 2. $\left(b^2 - 4ac \neq 0\right)$ this amounts to solving,

$$\begin{aligned} A_+\lambda_+^x + A_-\lambda_-^x &= \mu \text{ and} \\ A_+\lambda_+^{x+1} + A_-\lambda_-^{x+1} &= \nu \end{aligned}$$

which can always be done since

$$\det\begin{bmatrix} \lambda_+^x & \lambda_-^x \\ \lambda_+^{x+1} & \lambda_-^{x+1} \end{bmatrix} = \lambda_+^x\lambda_-^x(\lambda_- - \lambda_+) \neq 0.$$

Here we have used $\lambda_\pm \neq 0$ since $p(0) = c \neq 0$.[5]

Similarly in the case of item 4. $\left(b^2 - 4ac = 0\right)$, we must show there exists $A_0, A_1 \in \mathbb{R}$ such that

$$\begin{aligned} (A_0 + A_1 x)\lambda_0^x &= \mu \\ (A_0 + A_1(x+1))\lambda_0^{x+1} &= \nu \end{aligned}$$

which is again the case since

$$\det\begin{bmatrix} \lambda_0^x & x\lambda_0^x \\ \lambda_0^{x+1} & (x+1)\lambda_0^{x+1} \end{bmatrix} = \lambda_0^{2x+1} \neq 0.$$

Again we have used $\lambda_0 \neq 0$ since $p(0) = c \neq 0$. In fact in this case,

$$|\lambda_0| := |b/(2a)| = \frac{\sqrt{ac}}{|a|} = \sqrt{\frac{c}{a}} \neq 0.$$

In the next group of exercises you are going to use first step analysis to show that a simple unbiased random walk on $\mathbb{Z}$ is null recurrent. We let $\{X_n\}_{n=0}^\infty$ be the Markov chain with values in $\mathbb{Z}$ with transition probabilities given by

$$P(X_{n+1} = x \pm 1 | X_n = x) = 1/2 \text{ for all } n \in \mathbb{N}_0 \text{ and } x \in \mathbb{Z}.$$

Further let $a, b \in \mathbb{Z}$ with $a < 0 < b$ and

$$T_{a,b} := \min\{n : X_n \in \{a, b\}\} \text{ and } T_b := \inf\{n : X_n = b\}.$$

We know by Corollary 8.7 that $\mathbb{E}_0[T_{a,b}] < \infty$ from which it follows that $P(T_{a,b} < \infty) = 1$ for all $a < 0 < b$. For these reasons we will ignore the event $\{T_{a,b} = \infty\}$ in what follows below.

Exercise 8.4. Let $w(x) := P_x\left(X_{T_{a,b}} = b\right) := P\left(X_{T_{a,b}} = b | X_0 = x\right)$.

1. Use first step analysis to show for $a < x < b$ that
$$w(x) = \frac{1}{2}(w(x+1) + w(x-1)) \tag{8.23}$$
provided we define $w(a) = 0$ and $w(b) = 1$.
2. Use the results of Exercises 8.2 and 8.3 to show
$$P_x\left(X_{T_{a,b}} = b\right) = w(x) = \frac{1}{b-a}(x-a). \tag{8.24}$$
3. Let
$$T_b := \begin{cases} \min\{n : X_n = b\} & \text{if } \{X_n\} \text{ hits } b \\ \infty & \text{otherwise} \end{cases}$$
be the first time $\{X_n\}$ hits b. Explain why, $\{T_{a,b} < \infty\} \subset \{T_b < \infty\}$ and use this along with Eq. (8.24) to conclude that $P_x(T_b < \infty) = 1$ for all $x < b$. (By symmetry this result holds true for all $x \in \mathbb{Z}$.)

Exercise 8.5. The goal of this exercise is to give a second proof of the fact that $P_x(T_b < \infty) = 1$. Here is the outline:

1. Let $w(x) := P_x(T_b < \infty)$. Again use first step analysis to show that $w(x)$ satisfies Eq. (8.23) for all x with $w(b) = 1$.
2. Use Exercises 8.2 and 8.3 to show that there is a constant, c, such that
$$w(x) = c(x-b) + 1 \text{ for all } x \in \mathbb{Z}.$$
3. Explain why c must be zero to again show that $P_x(T_b < \infty) = 1$ for all $x \in \mathbb{Z}$.

Exercise 8.6. Let $T = T_{a,b}$ and $u(x) := \mathbb{E}_x T := \mathbb{E}[T | X_0 = x]$.

1. Use first step analysis to show for $a < x < b$ that
$$u(x) = \frac{1}{2}(u(x+1) + u(x-1)) + 1 \tag{8.25}$$
with the convention that $u(a) = 0 = u(b)$.

[5] In fact,
$$\lambda_+ \cdot \lambda_- = \lambda_\pm = \frac{b^2 - \left(b^2 - 4ac\right)}{4a^2} = c/a \neq 0.$$

2. Show that
$$u(x) = A_0 + A_1 x - x^2 \tag{8.26}$$
solves Eq. (8.25) for any choice of constants A_0 and A_1.
3. Choose A_0 and A_1 so that $u(x)$ satisfies the boundary conditions, $u(a) = 0 = u(b)$. Use this to conclude that
$$\mathbb{E}_x T_{a,b} = -ab + (b+a)x - x^2 = -a(b-x) + bx - x^2. \tag{8.27}$$

Remark 8.21. Notice that $T_{a,b} \uparrow T_b = \inf\{n : X_n = b\}$ as $a \downarrow -\infty$, and so passing to the limit as $a \downarrow -\infty$ in Eq. (8.27) shows
$$\mathbb{E}_x T_b = \infty \text{ for all } x < b.$$
Combining the last couple of exercises together shows that $\{X_n\}$ is "null - recurrent."

Exercise 8.7. Let $T = T_b$. The goal of this exercise is to give a second proof of the fact and $u(x) := \mathbb{E}_x T = \infty$ for all $x \neq b$. Here is the outline. Let $u(x) := \mathbb{E}_x T \in [0,\infty] = [0,\infty) \cup \{\infty\}$.

1. Note that $u(b) = 0$ and, by a first step analysis, that $u(x)$ satisfies Eq. (8.25) for all $x \neq b$ – allowing for the possibility that some of the $u(x)$ may be infinite.
2. Argue, using Eq. (8.25), that if $u(x) < \infty$ for some $x < b$ then $u(y) < \infty$ for all $y < b$. Similarly, if $u(x) < \infty$ for some $x > b$ then $u(y) < \infty$ for all $y > b$.
3. If $u(x) < \infty$ for all $x > b$ then $u(x)$ must be of the form in Eq. (8.26) for some A_0 and A_1 in $\mathbb{R}$ such that $u(b) = 0$. However, this would imply, $u(x) = \mathbb{E}_x T \to -\infty$ as $x \to \infty$ which is impossible since $\mathbb{E}_x T \geq 0$ for all x. Thus we must conclude that $\mathbb{E}_x T = u(x) = \infty$ for all $x > b$. (A similar argument works if we assume that $u(x) < \infty$ for all $x < b$.)

For the remaining exercises in this section we will assume that $p \in (1/2, 1)$ and $q = 1 - p$ so that $p/q > 1$.

Exercise 8.8 (Biased random walks I). Let $p \in (1/2, 1)$ and consider the biased random walk $\{X_n\}_{n\geq 0}$ on the $S = \mathbb{Z}$ where $X_n = \xi_0 + \xi_1 + \cdots + \xi_n$, $\{\xi_i\}_{i=1}^\infty$ are i.i.d. with $P(\xi_i = 1) = p \in (0,1)$ and $P(\xi_i = -1) = q := 1 - p$, and $\xi_0 = x$ for some $x \in \mathbb{Z}$. Let $T = T_{\{0\}}$ be the first hitting time of $\{0\}$ and $u(x) := P_x(T < \infty)$.

a) Use the first step analysis to show
$$u(x) = pu(x+1) + qu(x-1) \text{ for } x \neq 0 \text{ and } u(0) = 1. \tag{8.28}$$

b) Use Eq. (8.28) along with Exercises 8.2 and 8.3 to show for some $a_\pm \in \mathbb{R}$ that
$$u(x) = (1 - a_+) + a_+ (q/p)^x \text{ for } x \geq 0 \text{ and} \tag{8.29}$$
$$u(x) = (1 - a_-) + a_- (q/p)^x \text{ for } x \leq 0. \tag{8.30}$$

c) By considering the limit as $x \to -\infty$ conclude that $a_- = 0$ and $u(x) = 1$ for all $x < 0$, i.e. $P_x(T_0 < \infty) = 1$ for all $x \leq 0$.

Exercise 8.9 (Biased random walks II). The goal of this exercise is to evaluate $P_x(T_0 < \infty)$ for $x \geq 0$. To do this let $B_n := \{0, n\}$ and $T_n := T_{\{0,n\}}$. Let $h(x) := P_x(X_{T_n} = 0)$ where $\{X_{T_n} = 0\}$ is the event of hitting 0 before n.

a) Use the first step analysis to show
$$h(x) = ph(x+1) + qh(x-1) \text{ with } h(0) = 1 \text{ and } h(n) = 0.$$

b) Show the unique solution to this equation is given by
$$P_x(X_{T_n} = 0) = h(x) = \frac{(q/p)^x - (q/p)^n}{1 - (q/p)^n}.$$

c) Argue that
$$P_x(T < \infty) = \lim_{n\to\infty} P_x(\{X_{T_n} = 0\}) = (q/p)^x < 1 \text{ for all } x \geq 0.$$

The following formula summarizes Exercises 8.8 and 8.9; for $\frac{1}{2} < p < 1$,
$$P_x(T < \infty) = \begin{cases} (q/p)^x & \text{if } x \geq 0 \\ 1 & \text{if } x < 0 \end{cases}. \tag{8.31}$$

Example 8.22 (Biased random walks III). Continue the notation in Exercise 8.8. Let us start to compute $\mathbb{E}_x T$. Since $P_x(T = \infty) > 0$ for $x > 0$ we already know that $\mathbb{E}_x T = \infty$ for all $x > 0$. Nevertheless we will deduce this fact again here. Letting $u(x) = \mathbb{E}_x T$ it follows by the first step analysis that, for $x \neq 0$,
$$\begin{aligned} u(x) &= p[1 + u(x+1)] + q[1 + u(x-1)] \\ &= pu(x+1) + qu(x-1) + 1 \end{aligned} \tag{8.32}$$
with $u(0) = 0$. Notice $u(x) = \infty$ is a solution to this equation while if $u(n) < \infty$ for some $n \neq 0$ then Eq. (8.32) implies that $u(x) < \infty$ for all $x \neq 0$ with the same sign as n. A particular solution to this equation may be found by trying $u(x) = \alpha x$ to learn,
$$\alpha x = p\alpha(x+1) + q\alpha(x-1) + 1 = \alpha x + \alpha(p - q) + 1$$

which is valid for all x provided $\alpha = (q-p)^{-1}$. The general **finite** solution to Eq. (8.32) is therefore,

$$u(x) = (q-p)^{-1} x + a + b(q/p)^x. \tag{8.33}$$

Using the boundary condition, $u(0) = 0$ allows us to conclude that $a + b = 0$ and therefore,

$$u(x) = (q-p)^{-1} x + a\left[1 - (q/p)^x\right]. \tag{8.34}$$

Notice that $u(x) \to -\infty$ as $x \to +\infty$ no matter how a is chosen and therefore we must conclude that the desired solution to Eq. (8.32) is $u(x) = \infty$ for $x > 0$ as we already mentioned. In the next exercise you will compute $\mathbb{E}_x T$ for $x < 0$.

Exercise 8.10 (Biased random walks IV). Continue the notation in Example 8.22. Using the outline below, show

$$\mathbb{E}_x T = \frac{|x|}{p-q} \text{ for } x \le 0. \tag{8.35}$$

In the following outline n is a negative integer, T_n is the first hitting time of n so that $T_{\{n,0\}} = T_n \wedge T = \min\{T, T_n\}$ is the first hitting time of $\{n, 0\}$. By Corollary 8.7 we know that $u(x) := \mathbb{E}_x\left[T_{\{n,0\}}\right] < \infty$ for all $n \le x \le 0$ and by a first step analysis one sees that $u(x)$ still satisfies Eq. (8.32) for $n < x < 0$ and has boundary conditions $u(n) = 0 = u(0)$.

a) From Eq. (8.34) we know that, for some $a \in \mathbb{R}$,

$$\mathbb{E}_x\left[T_{\{n,0\}}\right] = u(x) = (q-p)^{-1} x + a\left[1 - (q/p)^x\right].$$

Use $u(n) = 0$ in order to show

$$a = a_n = \frac{n}{\left(1 - (q/p)^n\right)(p-q)}$$

and therefore,

$$\mathbb{E}_x\left[T_{\{n,0\}}\right] = \frac{1}{p-q}\left[|x| + n\frac{1-(q/p)^x}{1-(q/p)^n}\right] \text{ for } n \le x \le 0.$$

b) Argue that $\mathbb{E}_x T = \lim_{n\to-\infty} \mathbb{E}_x\left[T_n \wedge T\right]$ and use this and part a) to prove Eq. (8.35).

8.5 Computations avoiding the first step analysis

You may (SHOULD) skip the rest of this chapter!!

Theorem 8.23. *Let n denote a non-negative integer. If $h : B \to \mathbb{R}$ is measurable and either bounded or non-negative, then*

$$\mathbb{E}_x\left[h(X_n) : T_B = n\right] = \left(Q_A^{n-1} Q\left[1_B h\right]\right)(x)$$

and

$$\mathbb{E}_x\left[h(X_{T_B}) : T_B < \infty\right] = \left(\sum_{n=0}^{\infty} Q_A^n Q\left[1_B h\right]\right)(x). \tag{8.36}$$

If $g : A \to \mathbb{R}_+$ is a measurable function, then for all $x \in A$ and $n \in \mathbb{N}_0$,

$$\mathbb{E}_x\left[g(X_n)\, 1_{n<T_B}\right] = \left(Q_A^n g\right)(x).$$

In particular we have

$$\mathbb{E}_x\left[\sum_{n<T_B} g(X_n)\right] = \sum_{n=0}^{\infty}\left(Q_A^n g\right)(x) =: u(x), \tag{8.37}$$

where by convention, $\sum_{n<T_B} g(X_n) = 0$ when $T_B = 0$.

Proof. Let $x \in A$. In computing each of these quantities we will use;

$$\begin{aligned}\{T_B > n\} &= \{X_i \in A \text{ for } 0 \le i \le n\} \text{ and}\\ \{T_B = n\} &= \{X_i \in A \text{ for } 0 \le i \le n-1\} \cap \{X_n \in B\}.\end{aligned}$$

From the second identity above it follows that for

$$\begin{aligned}\mathbb{E}_x\left[h(X_n) : T_B = n\right] &= \mathbb{E}_x\left[h(X_n) : (X_1, \dots, X_{n-1}) \in A^{n-1},\, X_n \in B\right]\\ &= \sum_{n=1}^{\infty}\int_{A^{n-1}\times B}\prod_{j=1}^{n} Q(x_{j-1}, dx_j)\, h(x_n)\\ &= \left(Q_A^{n-1} Q\left[1_B h\right]\right)(x)\end{aligned}$$

and therefore

$$\begin{aligned}\mathbb{E}_x\left[h(X_{T_B}) : T_B < \infty\right] &= \sum_{n=1}^{\infty}\mathbb{E}_x\left[h(X_n) : T_B = n\right]\\ &= \sum_{n=1}^{\infty} Q_A^{n-1} Q\left[1_B h\right] = \sum_{n=0}^{\infty} Q_A^n Q\left[1_B h\right].\end{aligned}$$

Similarly,

$$\begin{aligned}\mathbb{E}_x\left[g\left(X_n\right)1_{n<T_B}\right] &= \int_{A^n} Q\left(x,dx_1\right)Q\left(x_1,dx_2\right)\ldots Q\left(x_{n-1},dx_n\right)g\left(x_n\right)\\ &= \left(Q_A^n g\right)(x)\end{aligned}$$

and therefore,

$$\begin{aligned}\mathbb{E}_x\left[\sum_{n=0}^{\infty} g\left(X_n\right)1_{n<T_B}\right] &= \sum_{n=0}^{\infty}\mathbb{E}_x\left[g\left(X_n\right)1_{n<T_B}\right]\\ &= \sum_{n=0}^{\infty}\left(Q_A^n g\right)(x).\end{aligned}$$

■

In practice it is not so easy to sum the series in Eqs. (8.36) and (8.37). Thus we would like to have another way to compute these quantities. Since $\sum_{n=0}^{\infty} Q_A^n$ is a geometric series, we expect that

$$\sum_{n=0}^{\infty} Q_A^n = \left(I-Q_A\right)^{-1}$$

which is basically correct at least when $(I-Q_A)$ is invertible. This suggests that if $u(x)=\mathbb{E}_x\left[h\left(X_{T_B}\right):T_B<\infty\right]$, then (see Eq. (8.36))

$$u = Q_A u + Q\left[1_B h\right] \text{ on } A, \tag{8.38}$$

and if $u(x)=\mathbb{E}_x\left[\sum_{n<T_B} g\left(X_n\right)\right]$, then (see Eq. (8.37))

$$u = Q_A u + g \text{ on } A. \tag{8.39}$$

That these equations are valid was the content of Corollary **??** and Theorem 8.13 above. below which we will prove using the "first step" analysis in the next theorem. We will give another direct proof in Theorem 8.28 below as well.

Lemma 8.24. *Keeping the notation above we have*

$$\mathbb{E}_x T = \sum_{n=0}^{\infty}\sum_{y\in A} Q^n(x,y) \text{ for all } x\in A, \tag{8.40}$$

where $\mathbb{E}_x T=\infty$ is possible.

Proof. By definition of T we have for $x\in A$ and $n\in\mathbb{N}_0$ that,

$$\begin{aligned}P_x(T>n) &= P_x\left(X_1,\ldots,X_n\in A\right)\\ &= \sum_{x_1,\ldots,x_n\in A} p(x,x_1)\,p(x_1,x_2)\ldots p(x_{n-1},x_n)\\ &= \sum_{y\in A} Q^n(x,y).\end{aligned} \tag{8.41}$$

Therefore Eq. (8.40) now follows from Lemma 8.3 and Eq. (8.41). ■

Proposition 8.25. *Let us continue the notation above and let us further assume that A is a finite set and*

$$P_x(T<\infty) = P\left(X_n\in B \text{ for some } n\right) > 0 \;\forall\; x\in A. \tag{8.42}$$

Under these assumptions, $\mathbb{E}_x T<\infty$ for all $x\in A$ and in particular $P_x(T<\infty)=1$ for all $x\in A$. In this case we may may write Eq. (8.40) as

$$\left(\mathbb{E}_x T\right)_{x\in A} = \left(I-Q\right)^{-1}\mathbf{1} \tag{8.43}$$

where $\mathbf{1}(x)=1$ for all $x\in A$.

Proof. Since $\{T>n\}\downarrow\{T=\infty\}$ and $P_x(T=\infty)<1$ for all $x\in A$ it follows that there exists an $m\in\mathbb{N}$ and $0\le\alpha<1$ such that $P_x(T>m)\le\alpha$ for all $x\in A$. Since $P_x(T>m)=\sum_{y\in A}Q^m(x,y)$ it follows that the row sums of Q^m are all less than $\alpha<1$. Further observe that

$$\begin{aligned}\sum_{y\in A} Q^{2m}(x,y) &= \sum_{y,z\in A} Q^m(x,z)\,Q^m(z,y) = \sum_{z\in A} Q^m(x,z)\sum_{y\in A} Q^m(z,y)\\ &\le \sum_{z\in A} Q^m(x,z)\,\alpha \le \alpha^2.\end{aligned}$$

Similarly one may show that $\sum_{y\in A}Q^{km}(x,y)\le\alpha^k$ for all $k\in\mathbb{N}$. Therefore from Eq. (8.41) with m replaced by km, we learn that $P_x(T>km)\le\alpha^k$ for all $k\in\mathbb{N}$ which then implies that

$$\sum_{y\in A} Q^n(x,y) = P_x(T>n) \le \alpha^{\left\lfloor\frac{n}{k}\right\rfloor} \text{ for all } n\in\mathbb{N},$$

where $\lfloor t\rfloor = m\in\mathbb{N}_0$ if $m\le t<m+1$, i.e. $\lfloor t\rfloor$ is the nearest integer to t which is smaller than t. Therefore, we have

$$\mathbb{E}_x T = \sum_{n=0}^{\infty}\sum_{y\in A} Q^n(x,y) \le \sum_{n=0}^{\infty}\alpha^{\left\lfloor\frac{n}{m}\right\rfloor} \le m\cdot\sum_{l=0}^{\infty}\alpha^l = m\frac{1}{1-\alpha}<\infty.$$

So it only remains to prove Eq. (8.43). From the above computations we see that $\sum_{n=0}^{\infty}Q^n$ is convergent. Moreover,

$$(I-Q)\sum_{n=0}^{\infty}Q^n=\sum_{n=0}^{\infty}Q^n-\sum_{n=0}^{\infty}Q^{n+1}=I$$

and therefore $(I-Q)$ is invertible and $\sum_{n=0}^{\infty}Q^n=(I-Q)^{-1}$. Finally,

$$(I-Q)^{-1}\mathbf{1}=\sum_{n=0}^{\infty}Q^n\mathbf{1}=\left(\sum_{n=0}^{\infty}\sum_{y\in A}Q^n(x,y)\right)_{x\in A}=(\mathbb{E}_xT)_{x\in A}$$

as claimed. ∎

Remark 8.26. Let $\{X_n\}_{n=0}^{\infty}$ denote the fair random walk on $\{0,1,2,\dots\}$ with 0 being an absorbing state. Using the first homework problems, see Remark **??**, we learn that $\mathbb{E}_iT=\infty$ for all $i>0$. This shows that we can not in general drop the assumption that A ($A=\{1,2,\dots\}$ in this example) is a finite set the statement of Proposition 8.25.

8.5.1 General facts about sub-probability kernels

Definition 8.27. *Suppose $(A,\mathcal{A})$ is a measurable space. A **sub-probability kernel** on $(A,\mathcal{A})$ is a function $\rho:A\times\mathcal{A}\to[0,1]$ such that $\rho(\cdot,C)$ is $\mathcal{A}/\mathcal{B}_{\mathbb{R}}$ – measurable for all $C\in\mathcal{A}$ and $\rho(x,\cdot):\mathcal{A}\to[0,1]$ is a measure for all $x\in A$.*

As with probability kernels we will identify ρ with the linear map, $\rho:\mathcal{A}_b\to\mathcal{A}_b$ given by

$$(\rho f)(x)=\rho(x,f)=\int_A f(y)\,\rho(x,dy).$$

Of course we have in mind that $\mathcal{A}=\mathcal{S}_A$ and $\rho=Q_A$. In the following lemma let $\|g\|_{\infty}:=\sup_{x\in A}|g(x)|$ for all $g\in\mathcal{A}_b$.

Theorem 8.28. *Let ρ be a sub-probability kernel on a measurable space $(A,\mathcal{A})$ and define $u_n(x):=(\rho^n1)(x)$ for all $x\in A$ and $n\in\mathbb{N}_0$. Then;*

1. *u_n is a decreasing sequence so that $u:=\lim_{n\to\infty}u_n$ exists and is in $\mathcal{A}_b$. (When $\rho=Q_A$, $u_n(x)=P_x(T_B>n)\downarrow u(x)=P(T_B=\infty)$ as $n\to\infty$.)*
2. *The function u satisfies $\rho u=u$.*
3. *If $w\in\mathcal{A}_b$ and $\rho w=w$ then $|w|\le\|w\|_{\infty}u$. In particular the equation, $\rho w=w$, has a non-zero solution $w\in\mathcal{A}_b$ iff $u\neq0$.*
4. *If $u=0$ and $g\in\mathcal{A}_b$, then there is at most one $w\in\mathcal{A}_b$ such that $w=\rho w+g$.*
5. *Let*

$$U:=\sum_{n=0}^{\infty}u_n=\sum_{n=0}^{\infty}\rho^n1:A\to[0,\infty] \tag{8.44}$$

and suppose that $U(x)<\infty$ for all $x\in A$. Then for each $g\in\mathcal{S}_b$,

$$w=\sum_{n=0}^{\infty}\rho^ng \tag{8.45}$$

is absolutely convergent,

$$|w|\le\|g\|_{\infty}U, \tag{8.46}$$

$\rho(x,|w|)<\infty$ for all $x\in A$, and w solves $w=\rho w+g$. Moreover if v also solves $v=\rho v+g$ and $|v|\le CU$ for some $C<\infty$ then $v=w$.
Observe that when $\rho=Q_A$,

$$U(x)=\sum_{n=0}^{\infty}P_x(T_B>n)=\sum_{n=0}^{\infty}\mathbb{E}_x(1_{T_B>n})=\mathbb{E}_x\left(\sum_{n=0}^{\infty}1_{T_B>n}\right)=\mathbb{E}_x[T_B].$$

6. *If $g:A\to[0,\infty]$ is any measurable function then*

$$w:=\sum_{n=0}^{\infty}\rho^ng:A\to[0,\infty]$$

is a solution to $w=\rho w+g$. (It may be that $w\equiv\infty$ though!) Moreover if $v:A\to[0,\infty]$ satisfies $v=\rho v+g$ then $w\le v$. Thus w is the minimal non-negative solution to $v=\rho v+g$.
7. *If there exists $\alpha<1$ such that $u\le\alpha$ on A then $u=0$. (When $\rho=Q_A$, this states that $P_x(T_B=\infty)\le\alpha$ for all $x\in A$ implies $P_x(T_A=\infty)=0$ for all $x\in A$.)*
8. *If there exists an $\alpha<1$ and an $n\in\mathbb{N}$ such that $u_n=\rho^n1\le\alpha$ on A, then there exists $C<\infty$ such that*

$$u_k(x)=\left(\rho^k1\right)(x)\le C\beta^k \text{ for all } x\in A \text{ and } k\in\mathbb{N}_0$$

where $\beta:=\alpha^{1/n}<1$. In particular, $U\le C(1-\beta)^{-1}$ and $u=0$ under this assumption.
(When $\rho=Q_A$ this assertion states; if $P_x(T_B>n)\le\alpha$ for all $\alpha\in A$, then $P_x(T_B>k)\le C\beta^k$ and $\mathbb{E}_xT_B\le C(1-\beta)^{-1}$ for all $k\in\mathbb{N}_0$.)

Proof. We will prove each item in turn.

1. First observe that $u_1(x)=\rho(x,A)\le1=u_0(x)$ and therefore,

$$u_{n+1}=\rho^{n+1}1=\rho^nu_1\le\rho^n1=u_n.$$

   We now let $u:=\lim_{n\to\infty}u_n$ so that $u:A\to[0,1]$.
2. Using DCT we may let $n\to\infty$ in the identity, $\rho u_n=u_{n+1}$ in order to show $\rho u=u$.

3. If $w \in \mathcal{A}_b$ with $\rho w = w$, then
$$|w| = |\rho^n w| \le \rho^n |w| \le \|w\|_\infty \rho^n 1 = \|w\|_\infty \cdot u_n.$$
Letting $n \to \infty$ shows that $|w| \le \|w\|_\infty u$.
4. If $w_i \in \mathcal{A}_b$ solves $w_i = \rho w_i + g$ for $i = 1, 2$ then $w := w_2 - w_1$ satisfies $w = \rho w$ and therefore $|w| \le Cu = 0$.
5. Let $U := \sum_{n=0}^\infty u_n = \sum_{n=0}^\infty \rho^n 1 : A \to [0, \infty]$ and suppose $U(x) < \infty$ for all $x \in A$. Then $u_n(x) \to 0$ as $n \to \infty$ and so bounded solutions to $\rho u = u$ are necessarily zero. Moreover we have, for all $k \in \mathbb{N}_0$, that
$$\rho^k U = \sum_{n=0}^\infty \rho^k u_n = \sum_{n=0}^\infty u_{n+k} = \sum_{n=k}^\infty u_n \le U. \tag{8.47}$$
Since the tails of convergent series tend to zero it follows that $\lim_{k\to\infty} \rho^k U = 0$.
Now if $g \in \mathcal{S}_b$, we have
$$\sum_{n=0}^\infty |\rho^n g| \le \sum_{n=0}^\infty \rho^n |g| \le \sum_{n=0}^\infty \rho^n \|g\|_\infty = \|g\|_\infty \cdot U < \infty \tag{8.48}$$
and therefore $\sum_{n=0}^\infty \rho^n g$ is absolutely convergent. Making use of Eqs. (8.47) and (8.48) we see that
$$\sum_{n=1}^\infty \rho |\rho^n g| \le \|g\|_\infty \cdot \rho U \le \|g\|_\infty U < \infty$$
and therefore (using DCT),
$$\begin{aligned} w &= \sum_{n=0}^\infty \rho^n g = g + \sum_{n=1}^\infty \rho^n g \\ &= g + \rho \sum_{n=1}^\infty \rho^{n-1} g = g + \rho w, \end{aligned}$$
i.e. w solves $w = g + \rho w$.
If $v : A \to \mathbb{R}$ is measurable such that $|v| \le CU$ and $v = g + \rho v$, then $y := w - v$ solves $y = \rho y$ with $|y| \le (C + \|g\|_\infty) U$. It follows that
$$|y| = |\rho^n y| \le (C + \|g\|_\infty) \rho^n U \to 0 \text{ as } n \to \infty,$$
i.e. $0 = y = w - v$.
6. If $g \ge 0$ we may always define w by Eq. (8.45) allowing for $w(x) = \infty$ for some or even all $x \in A$. As in the proof of the previous item (with DCT being replaced by MCT), it follows that $w = \rho w + g$. If $v \ge 0$ also solves $v = g + \rho v$, then
$$v = g + \rho(g + \rho v) = g + \rho g + \rho^2 v$$
and more generally by induction we have
$$v = \sum_{k=0}^n \rho^k g + \rho^{n+1} v \ge \sum_{k=0}^n \rho^k g.$$
Letting $n \to \infty$ in this last equation shows that $v \ge w$.
7. If $u \le \alpha < 1$ on A, then by item 3. with $w = u$ we find that
$$u \le \|u\|_\infty \cdot u \le \alpha u$$
which clearly implies $u = 0$.
8. If $u_n \le \alpha < 1$, then for any $m \in \mathbb{N}$ we have,
$$u_{n+m} = \rho^m u_n \le \alpha \rho^m 1 = \alpha u_m.$$
Taking $m = kn$ in this inequality shows, $u_{(k+1)n} \le \alpha u_{kn}$. Thus a simple induction argument shows $u_{kn} \le \alpha^k$ for all $k \in \mathbb{N}_0$. For general $l \in \mathbb{N}_0$ we write $l = kn + r$ with $0 \le r < n$. We then have,
$$u_l = u_{kn+r} \le u_{kn} \le \alpha^k = \alpha^{\frac{l-r}{n}} = C\alpha^{l/n}$$
where $C = \alpha^{-\frac{n-1}{n}}$.

■

Corollary 8.29. *If $h : B \to [0, \infty]$ is measurable, then $u(x) := \mathbb{E}_x[h(X_{T_B}) : T_B < \infty]$ is the unique minimal non-negative solution to Eq. (8.38) while if $g : A \to [0, \infty]$ is measurable, then $u(x) = \mathbb{E}_x\left[\sum_{n<T_B} g(X_n)\right]$ is the unique minimal non-negative solution to Eq. (8.39).*

Exercise 8.11. Keeping the notation of Exercise 8.8 and 8.10. Use Corollary 8.29 to show again that $P_x(T_B < \infty) = (q/p)^x$ for all $x > 0$ and $\mathbb{E}_x T_0 = x/(q-p)$ for $x < 0$. You should do so without making use of the extraneous hitting times, T_n for $n \ne 0$.

Solution to Exercise (8.11). From Eq. (8.28) of Exercise 8.8 we have seen for $x > 1$ that
$$P_x(T_0 < \infty) = a + (1 - a)(q/p)^x$$

for some $a \in [0,1]$. Since

$$\frac{d}{da}\left[a + (1-a)(q/p)^x\right] = 1 - (q/p)^x > 0,$$

the right side will be smallest when $a = 0$ and therefore we may (Corollary 8.29) conclude that

$$P_x(T_0 < \infty) = (q/p)^x \text{ for all } x > 0.$$

Similarly from Eq. (8.34) of Exercise 8.10 we have seen that if $\mathbb{E}_x T_0 < \infty$ for some and hence all $x < 0$ then

$$\mathbb{E}_x T_0 = (q-p)^{-1} x + a\left[1 - (q/p)^x\right]$$

for some $a \leq 0$. Since the right side of this equation is minimized by taking $a = 0$ we again have by Corollary 8.29 that

$$\mathbb{E}_x T_0 = (q-p)^{-1} x \text{ for all } x < 0.$$

Corollary 8.30. *If $P_x(T_B = \infty) = 0$ for all $x \in A$ and $h : B \to \mathbb{R}$ is a bounded measurable function, then $u(x) := \mathbb{E}_x[h(X_{T_B})]$ is the **unique** solution to Eq. (8.38).*

Corollary 8.31. *Suppose now that $A = B^c$ is a finite subset of S such that $P_x(T_B = \infty) < 1$ for all $x \in A$. Then there exists $C < \infty$ and $\beta \in (0,1)$ such that $P_x(T_B > n) \leq C\beta^n$ and in particular $\mathbb{E}_x T_B < \infty$ for all $x \in A$.*

Proof. Let $\alpha_0 = \max_{x \in A} P_x(T_B = \infty) < 1$. We know that

$$\lim_{n\to\infty} P_x(T_B > n) = P_x(T_B = \infty) \leq \alpha_0 \text{ for all } x \in A.$$

Therefore if $\alpha \in (\alpha_0, 1)$, using the fact that A is a finite set, there exists an n sufficiently large such that $P_x(T_B > n) \leq \alpha$ for all $x \in A$. The result now follows from item 8. of Theorem 8.28. ∎

9

Markov Chains in the Long Run (Results)

Through out this chapter $\{X_n\}_{n=0}^{\infty}$ will be a Markov chain on a discrete state space S with Markov kernel $p : S \times S \to [0,1]$ along with the following notation.

Notation 9.1 *For $i, j \in S$ we define the following quantities;*

$$T_i := \min\{n \geq 1 : X_n = i\} = \textbf{\textit{first hitting time of}}\ i \tag{9.1}$$

$$R_i := \min\{n \geq 1 : X_n = i\} = \textbf{\textit{first passage time of i,}} \tag{9.2}$$

$$M_i := \sum_{n \geq 1} 1_{X_n = i} = \textbf{\textit{the number of visits to}}\ i\ \textit{after}\ n = 0, \tag{9.3}$$

$$f_{i\,i}^{(n)} = P_i(R_i = n), \tag{9.4}$$

$$f_{i,i} = \sum_{n=0}^{\infty} f_{i,i}^{(n)} = \sum_{n=0}^{\infty} P_i(R_i = n) = P_i(R_i < \infty), \tag{9.5}$$

$$m_i = \mathbb{E}_i[R_i : R_i < \infty] = \sum_{n=0}^{\infty} n f_{i,i}^{(n)}\ \textit{and} \tag{9.6}$$

$$m_{i,j} = \mathbb{E}_i R_j. \tag{9.7}$$

9.1 A Touch of Class

Definition 9.2. *A state j is **accessible** from i (written $i \to j$) iff $P_i(T_j < \infty) > 0$ and $i \longleftrightarrow j$ (i **communicates** with j) iff $i \to j$ and $j \to i$. Notice that $i \to j$ iff there is a path, $i = x_0, x_1, \ldots, x_n = j \in S$ such that $p(x_0, x_1)\, p(x_1, x_2) \ldots p(x_{n-1}, x_n) > 0$.*

Definition 9.3. *For each $i \in S$, let $C_i := \{j \in S : i \longleftrightarrow j\}$ be the **communicating class of** i. The state space, S, is partitioned into a disjoint union of its communicating classes. If there is only one communication class we say that the chain is **irreducible.***

Definition 9.4. *A communicating class $C \subset S$ is **closed** provided the probability that X_n leaves C given that it started in C is zero. In other words $P_{ij} = 0$ for all $i \in C$ and $j \notin C$.*

The notion of being closed just introduced follows the usual mathematical conventions in that; C is closed for a chain X iff the X can not leave C if it starts in C. In particular it makes sense to restrict a Markov chain to a closed communication class. Mathematically this means that $\{p(x,y)\}_{x,y \in C}$ form a Markov transition matrix, i.e.

$$\sum_{y \in C} p(x,y) = 1 \text{ for all } x \in C.$$

Example 9.5. Consider the Markov chain with jump diagram given in Figure 9.1. In this example the communicating classes are $\{1,2\}$, $\{3,4\}$, and $\{5\}$ with

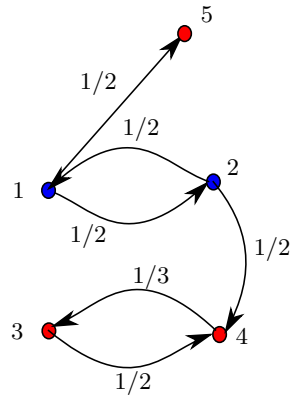


Fig. 9.1. A 5 state Markov chain with 3 communicating classes.

the latter classes being closed. The class $\{1,2\}$ is not closed.

Example 9.6. Let $\{X_n\}_{n=0}^{\infty}$ denote the fair random walk on $S = \mathbb{Z}$, then this chain is irreducible. On the other hand if $\{X_n\}_{n=0}^{\infty}$ is the fair random walk on $\{0, 1, 2, \ldots\}$ with 0 being an absorbing state, then the communication classes are $\{0\}$ (closed) and $\{1, 2, \ldots\}$ (not closed).

Definition 9.7. *For each $i \in S$, let $d(i)$ be the greatest common divisor of $\{n \geq 1 : \mathbf{P}_{i\,i}^n > 0\}$. We refer to $d(i)$ as the **period** of i. We say a site i is **aperiodic** if $d(i) = 1$.*

Example 9.8. Each site of the fair random walk on $S = \mathbb{Z}$ has period 2. While for the fair random walk on $\{0, 1, 2, \dots\}$ with 0 being an absorbing state, each $i \geq 1$ has period 2 while 0 has period 1, i.e. 0 is aperiodic.

Theorem 9.9. *The period function is constant on each communication class of a Markov chain.*

Proof. Let $x, y \in C$ and $a = d(x)$ and $b = d(y)$. Now suppose that $\mathbf{P}_{xy}^m > 0$ and $\mathbf{P}_{yx}^n > 0$, then $\mathbf{P}_{x,x}^{m+n} \geq \mathbf{P}_{xy}^m \mathbf{P}_{yx}^n > 0$ and so $a | (m+n)$. Further suppose that $\mathbf{P}_{y,y}^l > 0$ for come $l \in \mathbb{N}$, then

$$\mathbf{P}_{x,x}^{m+n+l} \geq \mathbf{P}_{xy}^m \mathbf{P}_{y,y}^l \mathbf{P}_{yx}^n > 0$$

and therefore $a | (m+n+l)$ which coupled with $a | (m+n)$ implies $a | l$. We may therefore conclude that $a \leq b$ (in fact $a | b$) as $b = \gcd\left(\left\{l \in \mathbb{N} : \mathbf{P}_{j,j}^l > 0\right\}\right)$. Similarly we show that $b \leq a$ and therefore $b = a$. ∎

Lemma 9.10. *If $d(i)$ is the period of site i, then*

1. *if $m \in \mathbb{N}$ and $\mathbf{P}_{i,i}^m > 0$ then $d(i)$ divides m and*
2. *$\mathbf{P}_{i,i}^{nd(i)} > 0$ for all $n \in \mathbb{N}$ sufficiently large.*
3. *If i is aperiodic iff $\mathbf{P}_{i,i}^n > 0$ for all $n \in \mathbb{N}$ sufficiently large.*

In summary, $A_i := \left\{m \in \mathbb{N} : \mathbf{P}_{i,i}^m > 0\right\} \subset d(i)\mathbb{N}$ and $d(i)\, n \in A_i$ for all $n \in \mathbb{N}$ sufficiently large.

Proof. Choose $n_1, \dots, n_k \in \{n \geq 1 : \mathbf{P}_{i\,i}^n > 0\}$ such that $d(i) = \gcd(n_1, \dots, n_k)$. For part 1. we also know that $d(i) = \gcd(n_1, \dots, n_k, m)$ and therefore $d(i)$ divides m. For part 2., if $m_i \in \mathbb{N}$ we have,

$$\left(\mathbf{P}^{\sum_{l=1}^k m_l n_l}\right)_{i,i} \geq \prod_{l=1}^k \left[\mathbf{P}_{i,i}^{n_l}\right]^{m_l} > 0.$$

This observation along with the number theoretic Lemma 9.15 below is enough to show $\mathbf{P}_{i,i}^{nd(i)} > 0$ for all $n \in \mathbb{N}$ sufficiently large. The third item is a special case of item 2. ∎

Example 9.11. Suppose that $\mathbf{P} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$, then $\mathbf{P}^m = \mathbf{P}$ if m is odd and $\mathbf{P}^m = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ if m is even. Therefore $d(i) = 2$ for $i = 1, 2$ and in this case $\mathbf{P}_{i,i}^{2n} = 1 > 0$ for all $n \in \mathbb{N}$. However observe that $\mathbf{P}^2$ is no longer irreducible – there are now two communication classes.

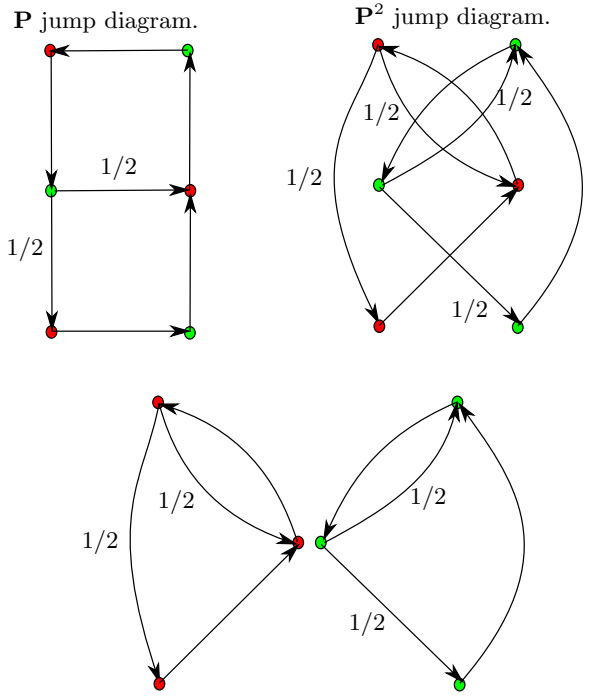
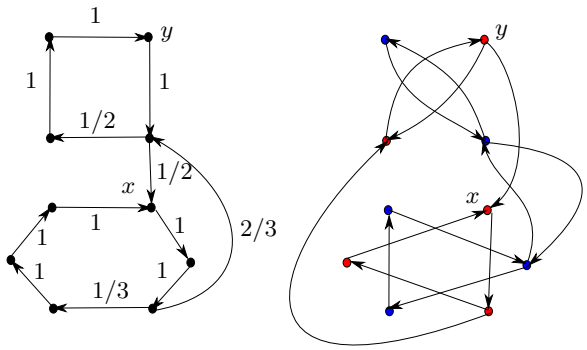


Fig. 9.2. All arrows are assumed to have weight 1 unless otherwise specified. Notice that each state has period $d = 2$ and that $\mathbf{P}^2$ is the transition matrix having two aperiodic communication classes.

Example 9.12. Consider the Markov chain with jump diagram given in Figure 9.2.In this example, $d(i) = 2$ for all i and all states for $\mathbf{P}^2$ are aperiodic. However $\mathbf{P}^2$ is no longer irreducible. This is an indication of the what happens in general. In terms of matrices,

$$\mathbf{P} = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ \hline 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{matrix} \text{ and } \mathbf{P}^2 = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ \hline 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{matrix}.$$

Example 9.13. Consider the Markov chain with jump diagram given in Figure 9.3.Assume there are no implied jumps from a site back to itself, i.e. $\mathbf{P}_{i,i} = 0$ for all i. This chain is then irreducible and has period 2. This chain is irreducible and has period 2. To calculate the period notice that starting at y there is an



P jump diagram. P^2 jump diagram.

Fig. 9.3. Assume there are no implied jumps from a site back to itself, i.e. $\mathbf{P}_{i,i} = 0$ for all i. This chain is then irreducible and has period 2.

obvious loop of length 4 and starting at x there is one of length 6. Therefore the period must divide both 4 and 6 and so must be either 2 or 1. The period is not 1 as one can only return to a site with an even number of jumps in this picture. If on the other hand there was any one vertex, i, such that $\mathbf{P}_{i,i} = 1$, then the period of the chain would have been one, i.e. the chain would have been aperiodic. Further notice that the jump diagram for $\mathbf{P}^2$ is no longer irreducible. The red vertices and the blue vertices split apart. This has to happen as is a consequence of Proposition 9.14 below.

Proposition 9.14. *If $\mathbf{P}$ is the Markov matrix for a finite state irreducible aperiodic chain, then there exists $n_0 \in \mathbb{N}$ such that $\mathbf{P}^n_{ij} > 0$ for all $i, j \in S$ and $n \geq n_0$.*

Proof. Let $i, j \in S$. By Lemma 9.10 with $d(i) = 1$ we know that $\mathbf{P}^m_{i,i} > 0$ for all m large. As $\mathbf{P}$ is irreducible there exists $a \in \mathbb{N}$ such that $\mathbf{P}^a_{ij} > 0$ and therefore $\mathbf{P}^{a+m}_{i,j} \geq \mathbf{P}^m_{i,i}\mathbf{P}^a_{i,j} > 0$ for all m sufficiently large. This shows for all $i, j \in S$ there exists $n_{i,j} \in \mathbb{N}$ such that $\mathbf{P}^n_{ij} > 0$ for all $n \geq n_{i,j}$. Since there are only finitely many steps we may now take $n_0 := \max\{n_{i,j} : i, j \in S\} < \infty$. ∎

9.1.1 A number theoretic lemma

Lemma 9.15 (A number theory lemma). *Suppose that 1 is the greatest common denominator of a set of positive integers, $\Gamma := \{n_1, \dots, n_k\}$. Then there exists $N \in \mathbb{N}$ such that the set,*

$$A = \{m_1 n_1 + \cdots + m_k n_k : m_i \geq 0 \text{ for all } i\},$$

contains all $n \in \mathbb{N}$ with $n \geq N$. More generally if $q = \gcd(\Gamma)$ (perhaps not 1), then $A \subset q\mathbb{N}$ and contains all points qn for n sufficiently large.

Proof. First proof. The set $I := \{m_1 n_1 + \cdots + m_k n_k : m_i \in \mathbb{N} \text{ for all } i\}$ is an ideal in $\mathbb{Z}$ and as $\mathbb{Z}$ is a principle ideal domain there is a $q \in I$ with $q > 0$ such that $I = q\mathbb{Z} = \{qm : m \in \mathbb{Z}\}$. In fact $q = \min(I \cap \mathbb{N})$. Since $q \in I$ we know that $q = m_1 n_1 + \cdots + m_k n_k$ for some $m_i \in \mathbb{N}$ and so if l is a common divisor of $n_1, \dots, n_k$ then l divides q. Moreover as $I = q\mathbb{Z}$ and $n_i \in I$ for all i, we know that $q | n_i$ as well. This shows that $q = \gcd(n_1, n_2, \dots, n_k)$.

Now suppose that $n \gg n_1 + \cdots + n_k$ is given and large (to be explained shortly). Then write $n = l(n_1 + \cdots + n_k) + r$ with $l \in \mathbb{N}$ and $0 \leq r < n_1 + \cdots + n_k$ and therefore,

$$\begin{aligned} nq &= ql(n_1 + \cdots + n_k) + rq \\ &= ql(n_1 + \cdots + n_k) + r(m_1 n_1 + \cdots + m_k n_k) \\ &= (ql + rm_1) n_1 + \cdots + (ql + rm_k) n_k \end{aligned}$$

where

$$ql + rm_i \geq ql - (n_1 + \cdots + n_k) m_i$$

which is greater than 0 for l and hence n sufficiently large.

Second proof. (The following proof is from Durrett [1].) We first will show that A contains two consecutive positive integers, a and $a + 1$. To prove this let,

$$k := \min\{|b - a| : a, b \in A \text{ with } a \neq b\}$$

and choose $a, b \in A$ with $b = a + k$. If $k > 1$, there exists $n \in \Gamma \subset A$ such that k does not divide n. Let us write $n = mk + r$ with $m \geq 0$ and $1 \leq r < k$. It then follows that $(m + 1) b$ and $(m + 1) a + n$ are in A,

$$(m + 1) b = (m + 1)(a + k) > (m + 1) a + mk + r = (m + 1) a + n,$$

and

$$(m + 1) b - (m + 1) a + n = k - r < k.$$

This contradicts the definition of k and therefore, $k = 1$.

Let $N = a^2$. If $n \geq N$, then $n - a^2 = ma + r$ for some $m \geq 0$ and $0 \leq r < a$. Therefore,

$$n = a^2 + ma + r = (a + m) a + r = (a + m - r) a + r(a + 1) \in A.$$

∎

9.2 Transience and Recurrence Classes

Definition 9.16 (First return time). *For any $x \in S$, let $R_x := \min\{n \geq 1 : X_n = x\}$ where the minimum of the empty set is defined to be ∞.*

On the event $\{X_0 \neq x\}$ we have $R_x = T_x := \min\{n \geq 0 : X_n = x\}$ – the first hitting time of x. So R_x is really manufactured for the case where $X_0 = x$ in which case $T_x = 0$ while R_x is the *first return time* to x.

Definition 9.17. *A state $i \in S$ is:*

1. ***transient*** *if $P_i(R_i < \infty) < 1$ ($\iff$ $P_i(R_i = \infty) > 0$),*
2. ***recurrent*** *if $P_i(R_i < \infty) = 1$ ($\iff$ $P_i(R_i = \infty) = 0$),*
   a) ***positive recurrent*** *if $1/(\mathbb{E}_i R_i) > 0$, i.e. $\mathbb{E}_i R_i < \infty$,*
   b) ***null recurrent*** *if it is recurrent ($P_i(R_i < \infty) = 1$) and $1/(\mathbb{E}_i R_i) = 0$, i.e. $\mathbb{E}_i R_i = \infty$.*

We let S_t, S_r, S_{pr}, and S_{nr} be the transient, recurrent, positive recurrent, and null recurrent states respectively.

Theorem 9.18 (Class properties). *Each of the conditions on a state $i \in S$ in Definition 9.17 is a class property. More explicitly if $i, j \in S$ communicate ($i \longleftrightarrow j$), then i is transient or positive recurrent or null recurrent iff j has the same property.*

Lemma 9.19 (Recurrent classes are closed). *Let $C \subset S$ be a communicating class. Then*

$$C \text{ not closed} \implies C \text{ is transient}$$

or equivalently put,

$$C \text{ is recurrent} \implies C \text{ is closed.}$$

Proof. If C is not closed and $i \in C$, there is a $j \notin C$ such that $i \to j$, i.e. there is a path $i = x_0, x_1, \ldots, x_n = j$ with all of the $\{x_l\}_{l=0}^n$ being distinct such that

$$P_i(X_0 = i, X_1 = x_1, \ldots, X_{n-1} = x_{n-1}, X_n = x_n = j) > 0.$$

Since $j \notin C$ we must have $j \nrightarrow C$ and therefore on the event,

$$A := \{X_0 = i, X_1 = x_1, \ldots, X_{n-1} = x_{n-1}, X_n = x_n = j\},$$

$X_m \notin C$ a.s. for all $m \geq n$ and therefore $R_i = \infty$ a.s. on the event A which has positive probability, i.e. $P_i(R_i = \infty) \geq P_i(A) > 0$. ∎

Proposition 9.20 (Return time estimates). *Let $x \in S$ and $\pi : S \to [0,1]$ be a probability on S.*

1. *If there exists $\alpha < 1$ such that $P_y(T_x = \infty) \leq \alpha$ for all $y \in S$ then $P_x(R_x = \infty) = 0$.*
2. *If there exists $\alpha < 1$ and $n \in \mathbb{N}$ such that $P_y(T_x > n) \leq \alpha$ for all $y \in S$, then*

$$\mathbb{E}_\pi[R_x] \leq 1 + \frac{n}{1-\alpha} < \infty.$$

Proof. 1. By Corollary 8.5 our hypothesis guarantees $P_y(T_x = \infty) = 0$ for all $y \in S$. Hence using the first step analysis we find,

$$\begin{aligned} P_\pi(R_x = \infty) &= \sum_{y \in S \setminus \{x\}} p(x,y) P_\pi(R_x = \infty | X_1 = y) \\ &= \sum_{y \in S \setminus \{x\}} p(x,y) P_y(T_y = \infty) = 0 \end{aligned}$$

wherein we have use $R_x = 1$ if $X_1 = x$.

2. By Corollary 8.6, our hypothesis guarantees $\mathbb{E}_y(T_x = \infty) = n/(1-\alpha)$ for all $y \in S$. Hence using the first step analysis we find,

$$\begin{aligned} \mathbb{E}_\pi(R_x) &= p(x,x) + \sum_{y \in S \setminus \{x\}} p(x,y) \mathbb{E}_\pi(R_x | X_1 = y) \\ &= p(x,x) + \sum_{y \in S \setminus \{x\}} p(x,y)[1 + \mathbb{E}_y T_x] = 1 + \frac{n}{1-\alpha} < \infty. \end{aligned}$$

The last item now follows using the exact same techniques in the proof of Corollary 8.7. ∎

Corollary 9.21. *If $C \subset S$ is a finite closed communication class, then C is positively recurrent and in fact $\mathbb{E}_y R_x < \infty$ for any $x, y \in C$.*

Proof. Since C is closed we may restrict our Markov chain to C and since C is a communication class we know that $P_y(T_x = \infty) < 1$ for all $x, y \in C$. Because C is finite it follows that $\max_{x,y \in C} P_y(T_x > n) = \alpha < 1$ for some n sufficiently large – see the proof of Corollary 8.7. Therefore Proposition 9.20 applies to show $\mathbb{E}_y R_x < \infty$ for all $x, y \in C$. ∎

Corollary 9.22. *If $\#(S) < \infty$ and C is a communication class in S. If C is closed then every $x \in C$ is positively recurrent and if C is not closed then ever $x \in C$ is transient. We will refer to the class C as being (positively) recurrent or transient respectively. We also have the equivalence of the following statements:*

Proposition 9.23. *1. C is closed.*
2. C is positive recurrent.
3. C is recurrent.

In particular if $\#(S) < \infty$, then the recurrent (= positively recurrent) states are precisely the union of the closed communication classes and the transient states are what is left over.

Proof. This is a simple combination of the results in Lemma 9.19 and Corollary 9.21. See Corollary 9.40 for another proof. ∎

Example 9.24. Let $\mathbf{P}$ be the Markov matrix with jump diagram given in Figure 9.1 above and repeated below in Figure 9.4. As we saw in Example 9.5 above, the communication classes are $\{\{1,2\},\{3,4\},\{5\}\}$. The latter two are closed and hence positively recurrent while $\{1,2\}$ is transient. Each of the classes is aperiodic since $\mathbf{P}_{i,i} > 0$ for all $i = 1,2,3,4,5$ in this example.

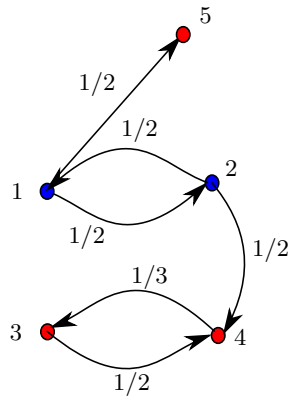


Fig. 9.4. A 5 state Markov chain with 3 communicating classes.

Proposition 9.25. *Suppose that $C \subset S$ is a finite communicating class and $T = \inf\{n \geq 0 : X_n \notin C\}$ be the first exit time from C. If C is not closed, then not only is C transient but $\mathbb{E}_i T < \infty$ for all $i \in C$ and in particular*

$$\mathbb{E}_j[M_i] \leq \mathbb{E}_j T < \infty \text{ for all } i,j \in C.$$

Proof. These results follow from Corollary 8.6 and the fact that

$$T = 1 + \sum_{i \in C} M_i.$$

∎

Warning: when $\#(S) = \infty$ or more importantly $\#(C) = \infty$, life is not so simple.

Remark 9.26. Let $\{X_n\}_{n=0}^{\infty}$ denote the fair random walk on $\{0,1,2,\dots\}$ with 0 being an absorbing state. The communication classes are now $\{0\}$ and $\{1,2,\dots\}$ with the latter class not being closed and hence transient. Using Exercise 8.6 or Exercise 8.7, it follows that $\mathbb{E}_i T = \infty$ for all $i > 0$ which shows we can not drop the assumption that $\#(C) < \infty$ in the first statement in Proposition 9.25. Similarly, using the fair random walk example, we see that it is not possible to drop the condition that $\#(C) < \infty$ for the equivalence statements as well.

The next examples show that if $C \subset S$ is closed and $\#(C) = \infty$, then C could be recurrent or it could be transient. Transient in this case means the chain goes off to "infinity," i.e. eventually leaves in finite subset of C never to return again.

Example 9.27. Let $S = \mathbb{Z}$ and $X = \{X_n\}$ be the standard fair random walk on $\mathbb{Z}$, i.e. $P(X_{n+1} = x \pm 1 | X_n = x) = \frac{1}{2}$. Then S itself is a closed class and every element of S is (null) recurrent. Indeed, using Exercise 8.4 or Exercise 8.5 and the first step analysis we know that

$$\begin{aligned} P_0[R_0 = \infty] &= \frac{1}{2}\left(P_0[R_0 = \infty | X_1 = 1] + P_0[R_0 = \infty | X_1 = -1]\right) \\ &= \frac{1}{2}\left(P_1[T_0 = \infty] + P_{-1}[T_0 = \infty]\right) = \frac{1}{2}(0+0) = 0. \end{aligned}$$

This shows 0 is recurrent. Similarly using Exercise 8.6 or Exercise 8.7 and the first step analysis we find,

$$\begin{aligned} \mathbb{E}_0[R_0] &= \frac{1}{2}\left(\mathbb{E}_0[R_0 X_1 = 1] + \mathbb{E}_0[R_0 | X_1 = -1]\right) \\ &= \frac{1}{2}\left(1 + \mathbb{E}_1[T_0 X] + 1 + \mathbb{E}_{-1}[T_0]\right) = \frac{1}{2}(\infty + \infty) = \infty \end{aligned}$$

and so 0 is null recurrent. As this chain is invariant under translation it follows that every $x \in \mathbb{Z}$ is a null recurrent site.

Example 9.28. Let $S = \mathbb{Z}$ and $X = \{X_n\}$ be a biassed random walk on $\mathbb{Z}$, i.e. $P(X_{n+1} = x+1 | X_n = x) = p$ and $P(X_{n+1} = x-1 | X_n = x) = q := 1-p$ with $p > \frac{1}{2}$. Then every site of is now transient. Recall from Exercises 8.8 and 8.9 (see Eq. (8.31)) that

$$P_x(T_0 < \infty) = \begin{cases} (q/p)^x & \text{if } x \geq 0 \\ 1 & \text{if } x < 0 \end{cases}. \tag{9.8}$$

Using these result and the first step analysis implies,

$$\begin{aligned} P_0[R_0 = \infty] &= pP_0[R_0 = \infty | X_1 = 1] + qP_0[R_0 = \infty | X_1 = -1] \\ &= pP_1[T_0 = \infty] + qP_{-1}[T_0 = \infty] \\ &= p\left[1 - (q/p)^1\right] + q(1-1) \\ &= p - q = 2p - 1 > 0. \end{aligned}$$

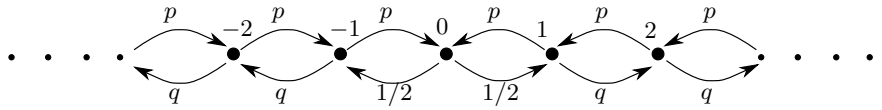


Fig. 9.5. A positively recurrent Markov chain.

Example 9.29. Again let $S = \mathbb{Z}$ and $p \in \left(\frac{1}{2}, 1\right)$ and suppose that $\{X_n\}$ is the random walk on $\mathbb{Z}$ described the jump diagram in Figure 9.5. In this case using the results of Exercise 8.10 we learn that

$$\begin{aligned}\mathbb{E}_0[R_0] &= \frac{1}{2}\left(\mathbb{E}_0[R_0|X_1 = 1] + \mathbb{E}_0[R_0|X_1 = -1]\right)\\ &= \frac{1}{2}\left(1 + \mathbb{E}_1[T_0 X] + 1 + \mathbb{E}_{-1}[T_0]\right)\\ &= 1 + \frac{1}{2}\left(\frac{1}{p-q} + \frac{1}{p-q}\right) = 1 + \frac{1}{p-q} = \frac{2p}{2p-1} < \infty.\end{aligned}$$

This shows the site 0 is positively recurrent. Thus according to Theorem 9.18, every site in $\mathbb{Z}$ is positively recurrent. (Notice that $\mathbb{E}_0[R_0] \to \infty$ as $p \downarrow \frac{1}{2}$, i.e. as the chain becomes closer to the unbiased random walk of Example 9.27.

Theorem 9.30 (Recurrence Conditions). *Let $j \in S$. Then the following are equivalent;*

1. *j is recurrent, i.e. $P_j(R_j < \infty) = 1$,*
2. *$P_j(X_n = j \text{ i.o. } n) = 1$,*
3. *$\mathbb{E}_j M_j = \sum_{n=1}^{\infty} \mathbf{P}_{jj}^n = \infty$.*

Moreover if $C \subset S$ is a recurrent communication class, then $P_i(\cap_{j\in C}\{X_n = j \text{ i.o. } n\}) = 1$ for all $i \in C$.In words, if we start in C then every state in C is visited an infinite number of times.

Theorem 9.31 (Transient States). *Let $j \in S$. Then the following are equivalent;*

1. *j is transient, i.e. $P_j(R_j < \infty) < 1$,*
2. *$P_j(X_n = j \text{ i.o. } n) = 0$, and*
3. *$\mathbb{E}_j M_j = \sum_{n=1}^{\infty} \mathbf{P}_{jj}^n < \infty$.*

More generally if $\nu : S \to [0,1]$ is any probability and $j \in S$ is transient, then

$$\mathbb{E}_\nu M_j = \sum_{n=1}^{\infty} P_\nu(X_n = j) = \mathbb{E}_\nu M_j < \infty \implies \begin{cases} \lim_{n\to\infty} P_\nu(X_n = j) = 0 \\ P_\nu(X_n = j \text{ i.o. } n) = 0. \end{cases} \tag{9.9}$$

Example 9.32. Let us revisit the fair random walk on $\mathbb{Z}$ describe before Exercise 8.4. In this case $P_0(X_n = 0) = 0$ if n is odd and

$$P_0(X_{2n} = 0) = \binom{2n}{n}\left(\frac{1}{2}\right)^{2n} = \frac{(2n)!}{(n!)^2}\left(\frac{1}{2}\right)^{2n}.$$

Making use of Stirling's formula, $n! \sim \sqrt{2\pi} n^{n+\frac{1}{2}} e^{-n}$, we find,

$$\left(\frac{1}{2}\right)^{2n}\frac{(2n)!}{(n!)^2} \sim \left(\frac{1}{2}\right)^{2n}\frac{\sqrt{2\pi}(2n)^{2n+\frac{1}{2}}e^{-2n}}{2\pi n^{2n+1}e^{-2n}} = \sqrt{\frac{1}{\pi}}\frac{1}{\sqrt{n}}$$

and therefore,

$$\sum_{n=0}^{\infty} P_0(X_n = 0) = \sum_{n=0}^{\infty} P_0(X_{2n} = 0) \sim 1 + \sum_{n=1}^{\infty}\sqrt{\frac{1}{\pi}}\frac{1}{\sqrt{n}} = \infty$$

which shows again that this walk is recurrent.

Example 9.33. The above method may easily be modified to show that the biased random walk on $\mathbb{Z}$ (see Exercise 8.8) is transient. In this case $\frac{1}{2} < p < 1$ and

$$P_0(X_{2n} = 0) = \binom{2n}{n} p^n (1-p)^n = \binom{2n}{n}\left[p(1-p)\right]^n.$$

Since $p(1-p)$ has a maximum at $p = \frac{1}{2}$ of $\frac{1}{4}$ we have $\rho_p := 4p(1-p) < 1$ for $\frac{1}{2} < p < 1$. Therefore,

$$P_0(X_{2n} = 0) = \binom{2n}{n}\left[\rho_p \frac{1}{4}\right]^n = \binom{2n}{n}\left(\frac{1}{2}\right)^{2n}\rho_p^n \sim \sqrt{\frac{1}{\pi}}\frac{1}{\sqrt{n}}\rho_p^n.$$

Hence

$$\sum_{n=0}^{\infty} P_0(X_n = 0) = \sum_{n=0}^{\infty} P_0(X_{2n} = 0) \sim 1 + \sum_{n=1}^{\infty}\frac{1}{\sqrt{n}}\rho_p^n \le 1 + \frac{1}{1-\rho_p} < \infty$$

which again shows the biased random walk is transient.

9.3 Invariant / Stationary (sub) distributions

As a warm-up let us again consider the Markov chain whose jump diagram is given in Figure 9.1. Let us further suppose that we start the chain at 1. We would like to compute $\lim_{n\to\infty} P_1(X_n = j)$ for $j = 1, 2, \ldots, 5$. Let $B = \{3, 4, 5\}$. Since there is a positive chance of hitting B from either 1 or 2 we know that

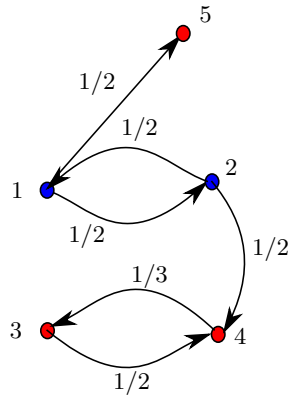


Fig. 9.6. A 5 state Markov chain – sites 1 and 2 are transient and 5 is absorbing.

$\mathbb{E}_i T_B < \infty$ for $i = 1, 2$. Let $h_i = P_i(X_{T_B} = 5)$ for $i = 1, 2$. Then $h_5 = 1$, $h_3 = h_4 = 0$ and the first step analysis shows,

$$h_1 = \frac{1}{2}h_5 + \frac{1}{2}h_2 = \frac{1}{2} + \frac{1}{2}h_2$$
$$h_2 = \frac{1}{2}h_1 + \frac{1}{2}h_4 = \frac{1}{2}h_1$$

and therefore, $h_1 = \frac{1}{2} + \frac{1}{4}h_1$ or $P_1(X_{T_B} = 5) = h_1 = \frac{2}{3}$. With this information in hand we may now conclude

$$\lim_{n\to\infty} P_1(X_n = 1) = 0 = \lim_{n\to\infty} P_1(X_n = 2),$$
$$\lim_{n\to\infty} P_1(X_n = 5) = \frac{2}{3}, \text{ and } \lim_{n\to\infty} P_1(X_n \in \{2,4\}) = \frac{1}{3}.$$

The question now becomes how the chain distributes itself within the closed class $\{2,4\}$. We are now going to address this issue.

Through out this chapter $\{X_n\}_{n=0}^{\infty}$ will be a Markov chain on a discrete state space S with Markov kernel $p : S \times S \to [0,1]$ and corresponding matrix $\mathbf{P}$. If $\pi_j := \lim_{n\to\infty} P_\nu(X_n = j)$ exists then

$$\begin{aligned}\pi_j &= \lim_{n\to\infty} P_\nu(X_{n+1} = j) = \lim_{n\to\infty} \sum_{k\in S} P_\nu(X_{n+1} = j | X_n = k) P_\nu(X_n = k) \\ &= \lim_{n\to\infty} \sum_{k\in S} P_\nu(X_n = k)\mathbf{P}_{kj} \overset{?}{=} \sum_{k\in S} \lim_{n\to\infty} P_\nu(X_n = k)\mathbf{P}_{kj} \\ &= \sum_{k\in S} \pi_k \mathbf{P}_{kj}.\end{aligned}$$

Thus we expect that any "limiting distributions" should also be "stationary" or "invariant" distributions.

Definition 9.34. *A function, $\pi : S \to [0,1]$ is a **sub-probability** if $\sum_{j\in S} \pi(j) \le 1$. We call $\pi(S) := \sum_{j\in S} \pi(j)$ the **mass** of π. So a probability is a sub-probability with mass one.*

Definition 9.35. *We say a sub-probability, $\pi : S \to [0,1]$, is **invariant or stationary** relative to $\mathbf{P}$ if $\pi\mathbf{P} = \pi$, i.e.*

$$\sum_{i\in S} \pi(i)\, p_{ij} = \pi(j) \text{ for all } j \in S. \tag{9.10}$$

*An invariant probability, $\pi : S \to [0,1]$, is called an **invariant distribution.***

Lemma 9.36. *A probability $\pi : S \to [0,1]$ is an invariant distribution for $\mathbf{P}$ iff*

$$P_\pi(X_n = j) = \pi(j) \text{ for all } j \in S \text{ and } n \in \mathbb{N}.$$

Proof. A simple induction argument shows that $\pi = \pi\mathbf{P}$ implies that $\pi = \pi\mathbf{P}^n$ for all $n \in \mathbb{N}$. This remark along with the following identity completes the proof;

$$P_\pi(X_n = j) = \sum_{i\in S} \pi(i) P_i(X_n = j) = \sum_{i\in S} \pi(i)\mathbf{P}^n_{ij} = (\pi\mathbf{P}^n)_j.$$

■

Example 9.37. Suppose that $S = \{1,2,3\}$, and

$$\mathbf{P} = \begin{array}{c} \begin{matrix} 1 & 2 & 3 \end{matrix} \\ \begin{bmatrix} 0 & 1 & 0 \\ 1/2 & 0 & 1/2 \\ 1 & 0 & 0 \end{bmatrix} \end{array} \begin{matrix} 1 \\ 2 \\ 3 \end{matrix}$$

has the jump graph given by 9.7. Notice that $\mathbf{P}^2_{11} > 0$ and $\mathbf{P}^3_{11} > 0$ that $\mathbf{P}$ is

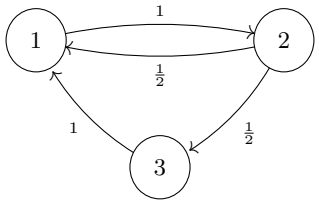


Fig. 9.7. A simple 3 state jump diagram.

"aperiodic." We now find the invariant distribution,

$$\operatorname{Nul}(\mathbf{P}-I)^{\mathrm{tr}}=\operatorname{Nul}\begin{bmatrix}-1&\frac{1}{2}&1\\1&-1&0\\0&\frac{1}{2}&-1\end{bmatrix}=\mathbb{R}\begin{bmatrix}2\\2\\1\end{bmatrix}.$$

Therefore the invariant distribution is given by

$$\pi=\frac{1}{5}\begin{bmatrix}2&2&1\end{bmatrix}=\begin{bmatrix}0.4&0.4&0.2\end{bmatrix}.$$

Let us now observe that

$$\mathbf{P}^2=\begin{bmatrix}\frac{1}{2}&0&\frac{1}{2}\\\frac{1}{2}&\frac{1}{2}&0\\0&1&0\end{bmatrix}$$

$$\mathbf{P}^3=\begin{bmatrix}0&1&0\\1/2&0&1/2\\1&0&0\end{bmatrix}^3=\begin{bmatrix}\frac{1}{2}&\frac{1}{2}&0\\\frac{1}{4}&\frac{1}{2}&\frac{1}{4}\\\frac{1}{2}&0&\frac{1}{2}\end{bmatrix}$$

$$\mathbf{P}^{20}=\begin{bmatrix}\frac{409}{1024}&\frac{205}{512}&\frac{205}{1024}\\\frac{205}{512}&\frac{409}{1024}&\frac{205}{1024}\\\frac{205}{512}&\frac{205}{512}&\frac{51}{256}\end{bmatrix}=\begin{bmatrix}0.399\,41&0.400\,39&0.200\,20\\0.400\,39&0.399\,41&0.200\,20\\0.400\,39&0.400\,39&0.199\,22\end{bmatrix}$$

and so it certainly appears that $\lim_{n\to\infty}\mathbf{P}^n_{ij}=\pi_j$ independent of i.

Example 9.38. Suppose that $\{X_n\}_{n=0}^{\infty}$ is the fair random walk on $S=\mathbb{Z}$ so that $P(X_{n+1}=x\pm1|X_n=x)=\frac{1}{2}$ for all $x\in S$ and $n\in\mathbb{N}_0$. This chain has no stationary distribution. To see this suppose $\pi:S\to[0,1]$ were to exist, then by definition,

$$\pi(y)=\sum_{x\in S}\pi(x)\,p(x,y)=\frac{1}{2}\left[\pi(y-1)+\pi(y+1)\right].$$

From Exercise 8.3 we know that the general solution to this equation is of the form,

$$\pi(x)=a+bx \text{ for some } a,b\in\mathbb{R}.$$

In order for $\pi(x)\geq0$ for all x we must have $b=0$ and $a\geq0$. However this is no choice for a such that $\sum_{x\in S}\pi(x)=1$. We will see explicitly in the next chapter that when $\#(S)<\infty$, every chain will have at least one stationary distribution.

Exercise 9.1 (2 - step M, see 7.2). Consider the following simple (i.e. no-brainer) two state "game" consisting of moving between two sites labeled 1 and 2. At each site you find a coin with sides labeled 1 and 2. The probability of flipping a 2 at site 1 is $a\in[0,1]$ and a 1 at site 2 is $b\in[0,1]$. We assume that $0<a+b<2$, i.e. neither both of a and b are zero or 1. If you are at site i at time n, then you flip the coin at this site and move or stay at the current site as indicated by coin toss. We summarize this scheme by the "jump diagram" of Figure 9.8. It is reasonable to suppose that your location, X_n, at time n is modeled by a Markov process with state space, $S=\{1,2\}$. Explain (briefly) why this is a time homogeneous chain and find the one step transition probabilities,

$$p(i,j)=P(X_{n+1}=j|X_n=i) \text{ for } i,j\in S.$$

Use your result and basic linear (matrix) algebra to compute, $\lim_{n\to\infty}P(X_n=1)$. Your answer should be independent of the possible starting distributions, $\pi=(\pi_1,\pi_2)$ for X_0 where $\pi_i:=P(X_0=i)$.

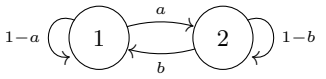


Fig. 9.8. The generic jump diagram for a two state Markov chain.

Solution to Exercise (7.2). The Markov matrix for this chian is

$$\mathbf{P}=\begin{bmatrix}1-a&a\\b&1-b\end{bmatrix}.$$

If $P(X_0=i)=\nu_i$ for $i=1,2$ then

$$P(X_n=j)=\sum_{k=1}^{2}\nu_k\mathbf{P}^n_{k,j}=\left[\nu\mathbf{P}^n\right]_j$$

where we now write $\nu=(\nu_1,\nu_2)$ as a row vector. A simple computation shows that

$$\begin{aligned}\det\left(\mathbf{P}^{\mathrm{tr}}-\lambda I\right)&=\det(\mathbf{P}-\lambda I)\\&=\lambda^2+(a+b-2)\lambda+(1-b-a)\\&=(\lambda-1)(\lambda-(1-a-b)).\end{aligned}$$

For any Markov matrix, $\sum_j\mathbf{P}_{ij}=1$ and therefore $\mathbf{P1}=\mathbf{1}$ where $\mathbf{1}$ is the column vector with all entries being 1. Thus we always know that $\lambda_1=1$ is an eigenvalue of $\mathbf{P}$. The second eigenvalue is $\lambda_2=1-a-b$. We now find the eigenvectors of $\mathbf{P}^{\mathrm{tr}}$;

$$\operatorname{Nul}\left(\mathbf{P}^{\mathrm{tr}}-\lambda_1 I\right)=\operatorname{Nul}\left(\begin{bmatrix}-a&b\\a&-b\end{bmatrix}\right)=\mathbb{R}\cdot\begin{bmatrix}b\\a\end{bmatrix}$$

while

$$\text{Nul}\left(\mathbf{P}^{\text{tr}}-\lambda I_2\right)=\text{Nul}\left(\begin{bmatrix} b & b \\ a & a \end{bmatrix}\right)=\mathbb{R}\cdot\begin{bmatrix} 1 \\ -1 \end{bmatrix}.$$

Writing

$$\nu=\alpha\,(b,a)+\beta\,(1,-1)\,,$$

we find

$$\begin{aligned} 1&=\nu\cdot(1,1)=\alpha\,(b,a)\cdot(1,1)=\alpha\,(a+b)\,,\\ \alpha&=\frac{1}{a+b}\text{ and}\\ \beta&=v_1-\alpha b=\alpha a-v_2. \end{aligned}$$

At any rate we have

$$\nu=\frac{1}{a+b}\,(b,a)+\beta\,(1,-1)\,.$$

and therefore,

$$\nu\mathbf{P}^n=\frac{1}{a+b}\,(b,a)\,\mathbf{P}^n+\beta\,(1,-1)\,\mathbf{P}^n=\frac{1}{a+b}\,(b,a)+\beta\,(1,-1)\,\lambda_2^n.$$

By our assumptions that $a+b\neq 2$ we have $|\lambda_2|<1$ and therefore

$$\lim_{n\to\infty}\nu\mathbf{P}^n=\frac{1}{a+b}\,(b,a)$$

and we have shown

$$\lim_{n\to\infty}P\,(X_n=1)=\frac{b}{a+b}\text{ and }\lim_{n\to\infty}P\,(X_n=2)=\frac{a}{a+b}$$

independent of the starting distribution ν. Also observe that the convergence is exponentially fast. For the two degenerate cases not considered here see Examples 10.6 and 10.7 below.

9.4 The basic limit theorems

Theorem 9.39. *Suppose that $\mathbf{P}=(p_{ij})$ is an irreducible Markov kernel and $\pi_j:=\frac{1}{\mathbb{E}_jR_j}$ for all $j\in S$. Then:*

1. *For all $i,j\in S$, we have*

$$\lim_{N\to\infty}\frac{1}{N}\sum_{n=0}^{N}1_{X_n=j}=\pi_j\quad P_i-a.s. \tag{9.11}$$

and

$$\lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N}P_i\,(X_n=j)=\lim_{N\to\infty}\frac{1}{N}\sum_{n=0}^{N}\mathbf{P}_{ij}^n=\pi_j. \tag{9.12}$$

2. *If $\mu:S\to[0,1]$ is an invariant sub-probability, then either $\mu(i)>0$ for all i or $\mu(i)=0$ for all i.*
3. *$\mathbf{P}$ has at most one invariant distribution.*
4. *$\mathbf{P}$ has a (necessarily unique) invariant distribution, $\mu:S\to[0,1]$, iff $\mathbf{P}$ is positive recurrent in which case $\mu(i)=\pi(i)=\frac{1}{\mathbb{E}_iR_i}>0$ for all $i\in S$.*

(These results may of course be applied to the restriction of a general non-irreducible Markov chain to any one of its communication classes.)

Proof. These results are the contents of Theorem **??** and Propositions **??** and **??** below. ∎

Using this result we can give another proof of Proposition 9.25.

Corollary 9.40. *If C is a closed finite communicating class then C is positive recurrent. (Recall that we already know that C is recurrent by Corollary **??**.)*

Proof. For $i,j\in C$, let

$$\pi_j:=\lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N}P_i\,(X_n=j)=\frac{1}{\mathbb{E}_jR_j}$$

as in Theorem **??**. Since C is closed,

$$\sum_{j\in C}P_i\,(X_n=j)=1$$

and therefore,

$$\sum_{j\in C}\pi_j=\lim_{N\to\infty}\frac{1}{N}\sum_{j\in C}\sum_{n=1}^{N}P_i\,(X_n=j)=\lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N}\sum_{j\in C}P_i\,(X_n=j)=1.$$

Therefore $\pi_j>0$ for some $j\in C$ and hence all $j\in C$ by Theorem 9.39 with S replaced by C. Hence we have $\mathbb{E}_jR_j<\infty$, i.e. every $j\in C$ is a positive recurrent state. ∎

Theorem 9.41. *Let $\mathbf{P}$ is be an irreducible Markov chain and ν is a probability on S.*

1. *If $\mathbf{P}$ is null-recurrent then $\lim_{n\to\infty}\mathbf{P}_{ij}^n=0$ for all $i,j\in S$ and more generally*

$$\lim_{n\to\infty}P_\nu\,(X_n=i)=0$$

2. *If $\mathbf{P}$ is a positive-recurrent and aperiodic Markov transition kernel, then*

$$\lim_{n\to\infty}\mathbf{P}_{ij}^n=\frac{1}{\mathbb{E}_iR_j}=:\pi_j$$

and more generally,

$$\lim_{n\to\infty}P_\nu\,(X_n=i)=\pi_j.$$

More generally, if C is an aperiodic communication class and ν is any probability on S, then

$$\lim_{n\to\infty} P_\nu(X_n = i) := \lim_{n\to\infty} \sum_{j\in S} \nu(j)\, \mathbf{P}^n_{ji} = P_\nu(R_i < \infty) \frac{1}{\mathbb{E}_j(R_j)} \text{ for all } i \in C.$$

If C is transient or null-recurrent then no matter whether C is aperiodic or not we will have $\lim_{n\to\infty} P_\nu(X_n = i) = 0$ for all $i \in C$.

Theorem 9.42 (General Convergence Theorem). *Let $\nu : S \to [0,1]$ be any probability, $i \in S$, C be the communicating class containing i,*

$$\{X_n \text{ hits } C\} := \{X_n \in C \text{ for some } n\},$$

and

$$\pi_i := \pi_i(\nu) = \frac{P_\nu(X_n \text{ hits } C)}{\mathbb{E}_i R_i},$$

where $1/\infty := 0$. Then:

1. *P_ν – a.s.,*
$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} 1_{X_n = i} = \frac{1}{\mathbb{E}_i R_i} 1_{\{X_n \text{ hits } C\}},$$
2.
$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} P_\nu(X_n = i) = \pi_i,$$
3. *π is an invariant sub-probability for $\mathbf{P}$, and*
4. *the mass of π is*
$$\sum_{i\in S} \pi_i = \sum_{C:\ pos.\ recurrent} P_\nu(X_n \text{ hits } C) \le 1.$$

The loss of mass can only occur when $\#(S) = \infty$ and this loss of mass happens throught loss of mass (sand) to infinity in the transient and null - recurrent classes with inifinitely many points. For example in the fair random walk a unit lump of sand starting at zero spreads out with $1/2$ going to $+\infty$ and the other half going to $-\infty$. While for the biased random walk with $p > \frac{1}{2}$, the sand all gets shoveled to $+\infty$.

10

Finite State Space Results and Examples

For this subsection suppose that $S = \{1, 2, \dots, n\}$ and $\mathbf{P}_{ij}$ is a Markov matrix.

Proposition 10.1. *The Markov matrix $\mathbf{P}$ on a finite state space has at least one invariant distribution.*

Proof. If $\mathbf{1} := \begin{bmatrix} 1 & 1 & \dots & 1 \end{bmatrix}^{\mathrm{tr}}$, then $\mathbf{P1} = \mathbf{1}$ from which it follows that

$$0 = \det(\mathbf{P} - I) = \det\left(\mathbf{P}^{\mathrm{tr}} - I\right).$$

Therefore there exists a non-zero row vector ν such that $\mathbf{P}^{\mathrm{tr}}\nu^{\mathrm{tr}} = \nu^{\mathrm{tr}}$ or equivalently that $\nu\mathbf{P} = \nu$. At this point we would be done if we knew that $\nu_i \geq 0$ for all i – but we don't. So let $\pi_i := |\nu_i|$ and observe that

$$\pi_i = |\nu_i| = \left|\sum_{k=1}^n \nu_k \mathbf{P}_{ki}\right| \leq \sum_{k=1}^n |\nu_k| \mathbf{P}_{ki} \leq \sum_{k=1}^n \pi_k \mathbf{P}_{ki}.$$

We now claim that in fact $\pi = \pi\mathbf{P}$. If this were not the case we would have $\pi_i < \sum_{k=1}^n \pi_k \mathbf{P}_{ki}$ for some i and therefore

$$0 < \sum_{i=1}^n \pi_i < \sum_{i=1}^n \sum_{k=1}^n \pi_k \mathbf{P}_{ki} = \sum_{k=1}^n \sum_{i=1}^n \pi_k \mathbf{P}_{ki} = \sum_{k=1}^n \pi_k$$

which is a contradiction. So all that is left to do is normalize π_i so $\sum_{i=1}^n \pi_i = 1$ and we are done. ∎

Proposition 10.2. *Suppose that $\mathbf{P}$ is an irreducible Markov matrix on ($\#(S) = \infty$ ok here) and suppose $\pi : S \to [0,1]$ is not identically zero function such that $\pi = \pi\mathbf{P}$, then $\pi(j) > 0$ for all $j \in S$. In particular, if $\pi : S \to [0,1]$ is an invariant distribution of $\mathbf{P}$ then $\pi(j) > 0$ for all $j \in S$.*

Proof. Let $i \in S$ such that $\pi(i) > 0$ and let $j \in S$ be arbitrary. Since $\mathbf{P}$ is irreducible, there exists $n \in \mathbb{N}$ such that $\mathbf{P}_{ij}^n > 0$. The result now follows because;

$$\pi(j) = \sum_k \pi(k) \mathbf{P}_{kj}^n \geq \pi(i) \mathbf{P}_{ij}^n > 0.$$

∎

Corollary 10.3. *If $\mathbf{P}$ is an irreducible Markov matrix on a finite state space S, then $\mathbf{P}$ has precisely one invariant distribution π. (We will give a second proof of this in Proposition 10.4 below as well.)*

Proof. Suppose that λ and π are two invariant distributions which are necessarily positive by Proposition 10.2. Let $c = \min_i \frac{\pi(i)}{\lambda(i)} > 0$ and set $\alpha(i) = \pi(i) - c\lambda(i)$ for all $i \in S$. Notice that $\alpha(i) \geq 0$ and is equal to zero at some $i \in S$. Moreover $\alpha = \alpha\mathbf{P}$ and therefore according to Proposition 10.2 we must have $\alpha = 0$, i.e. $\pi(i) = c\lambda(i)$ for all $i \in S$. Summing this equation on i shows that $1 = c \cdot 1$, i.e. $c = 1$ and so $\pi = \lambda$. ∎

Recall that $\mathbf{P}$ is **irreducible** means that for all $i \neq j$ there exists $n \in \mathbb{N}$ such that $\mathbf{P}_{ij}^n > 0$. Alternatively put this implies that $P_i(T_j < \infty) = P_i(R_j < \infty) > 0$ for all $i \neq j$. By Corollary 8.31 we know that $\mathbb{E}_i[R_j] = \mathbb{E}_i T_j < \infty$ for all $i \neq j$ and from Exercise 7.4 that $\mathbb{E}_i R_i < \infty$ also holds. The fact that $\mathbb{E}_i R_i < \infty$ for all $i \in S$ will come out of the proof of the next proposition as well.

Proposition 10.4. *If $\mathbf{P}$ is irreducible, then there is precisely one invariant distribution, π, which is given by $\pi_i = 1/(\mathbb{E}_i R_i) > 0$ for all $i \in S$.*

Proof. First observe that

$$R_j(i, X) = \begin{cases} R_j(i, j, X_2, \dots) = 1 & \text{if } X_0 = j \\ 1 + R_j(X) & \text{if } X_0 \neq j \end{cases} = 1 + 1_{X_0 \neq j} R_j(X).$$

Therefore by the first step analysis,

$$\begin{aligned} \mathbb{E}_i[R_j] &= \mathbb{E}_i[R_j(X)] = \mathbb{E}_{p(i,\cdot)}[R_j(i, X)] \\ &= \mathbb{E}_{p(i,\cdot)}[1 + 1_{X_0 \neq j} R_j(X)] \\ &= 1 + \sum_{k \neq j} \mathbf{P}_{ik} \mathbb{E}_k[R_j]. \end{aligned} \tag{10.1}$$

Here is a slight recasting of this same argument;

$$\begin{aligned} \mathbb{E}_i[R_j] &= \sum_{k=1}^n \mathbb{E}_i[R_j | X_1 = k] \mathbf{P}_{ik} = \sum_{k \neq j} \mathbb{E}_i[R_j | X_1 = k] \mathbf{P}_{ik} + \mathbf{P}_{ij} 1 \\ &= \sum_{k \neq j} (\mathbb{E}_k[R_j] + 1) \mathbf{P}_{ik} + \mathbf{P}_{ij} 1 = \sum_{k \neq j} \mathbb{E}_k[R_j] \mathbf{P}_{ik} + 1. \end{aligned}$$

which is again Eq. (10.1).

Now suppose that π is any invariant distribution for $\mathbf{P}$, then multiplying Eq. (10.1) by π_i and summing on i shows

$$\sum_{i=1}^{n}\pi_i\mathbb{E}_i[R_j]=\sum_{i=1}^{n}\pi_i\sum_{k\neq j}\mathbf{P}_{ik}\mathbb{E}_k[R_j]+\sum_{i=1}^{n}\pi_i 1$$
$$=\sum_{k\neq j}\pi_k\mathbb{E}_k[R_j]+1.$$

Since $\sum_{k\neq j}\pi_k\mathbb{E}_k[R_j]<\infty$ we may cancel it from both sides of this equation in order to learn $\pi_j\mathbb{E}_j[R_j]=1$. This shows that $\pi_j>0$, $\mathbb{E}_j[R_j]<\infty$, and $\pi_j=1/(\mathbb{E}_jR_j)$ for all $j\in S$. ∎

We may use Eq. (10.1) to compute $\mathbb{E}_i[R_j]$ in examples. To do this, fix j and set $v_i:=\mathbb{E}_iR_j$. Then Eq. (10.1) states that $v=\mathbf{P}^{(j)}v+\mathbf{1}$ where $\mathbf{P}^{(j)}$ denotes $\mathbf{P}$ with the j^{th} – column replaced by all zeros. Thus we have

$$\mathbb{E}_iR_j=\left[\left(I-\mathbf{P}^{(j)}\right)^{-1}\mathbf{1}\right]_i, \tag{10.2}$$

i.e.

$$\begin{bmatrix}\mathbb{E}_1R_j\\ \vdots\\ \mathbb{E}_nR_j\end{bmatrix}=\left(I-\mathbf{P}^{(j)}\right)^{-1}\begin{bmatrix}1\\ \vdots\\ 1\end{bmatrix}. \tag{10.3}$$

Remark 10.5. We can also derive Eq. (10.2) by first principles as well;

$$\begin{aligned}\mathbb{E}_iR_j&=\sum_{n=0}^{\infty}P_i(R_j>n)=1+\sum_{n=1}^{\infty}P_i(R_j>n)\\ &=1+\sum_{n=1}^{\infty}P_i(X_1\neq j,\dots,X_n\neq j) \qquad (10.4)\\ &=1+\sum_{n=1}^{\infty}\sum_{x_1,\dots,x_n\in S\setminus\{j\}}p(i,x_1)p(x_1,x_2)\dots p(x_{n-1},x_n)\\ &=1+\sum_{n=1}^{\infty}\left(\left[P^{(j)}\right]^n\mathbf{1}\right)_i=\sum_{n=0}^{\infty}\left(\left[P^{(j)}\right]^n\mathbf{1}\right)_i=\left[\left(I-P^{(j)}\right)^{-1}\mathbf{1}\right]_i.\end{aligned}$$

Multiplying Eq. (10.4) by $\pi(i)$ and summing on i implies,

$$\mathbb{E}_\pi R_j=1+\sum_{n=1}^{\infty}P_\pi(X_1\neq j,\dots,X_n\neq j).$$

Assuming that π is an invariant distribution of the chain this leads to

$$\begin{aligned}\mathbb{E}_\pi R_j&=1+\sum_{n=1}^{\infty}P_\pi(X_1\neq j,\dots,X_n\neq j)\\ &=1+\sum_{n=1}^{\infty}P_\pi(X_0\neq j,\dots,X_{n-1}\neq j)\\ &=1+\sum_{n=0}^{\infty}P_\pi(X_0\neq j,\dots,X_n\neq j)\\ &=1+(1-\pi(j))+\sum_{n=1}^{\infty}P_\pi(X_0\neq j,X_1\neq j,\dots,X_n\neq j)\\ &=1+(1-\pi(j))+\sum_{n=1}^{\infty}\begin{bmatrix}P_\pi(X_1\neq j,\dots,X_n\neq j)\\ -P_\pi(X_0=j,X_1\neq j,\dots,X_n\neq j)\end{bmatrix}\\ &=1+\sum_{n=1}^{\infty}P_\pi(X_1\neq j,\dots,X_n\neq j)+(1-\pi(j))\\ &\qquad-\sum_{n=1}^{\infty}P_\pi(X_0=j,X_1\neq j,\dots,X_n\neq j)\\ &=\mathbb{E}_\pi R_j+(1-\pi(j))-\pi(j)\sum_{n=1}^{\infty}P_j(X_1\neq j,\dots,X_n\neq j)\\ &=\mathbb{E}_\pi R_j+1-\pi(j)\left[1+\sum_{n=1}^{\infty}P_j(X_1\neq j,\dots,X_n\neq j)\right]\\ &=\mathbb{E}_\pi R_j+1-\pi(j)\,\mathbb{E}_jR_j.\end{aligned}$$

10.1 Some worked examples

Example 10.6. Let $S=\{1,2\}$ and $\mathbf{P}=\begin{bmatrix}0&1\\1&0\end{bmatrix}$ with jump diagram in Figure 10.1. In this case $\mathbf{P}^{2n}=I$ while $\mathbf{P}^{2n+1}=\mathbf{P}$ and therefore $\lim_{n\to\infty}\mathbf{P}^n$ does not

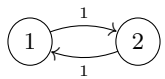


Fig. 10.1. A non-random chain.

exist. On the other hand it is easy to see that the invariant distribution, π, for $\mathbf{P}$ is $\pi=\begin{bmatrix}1/2&1/2\end{bmatrix}$ and, moreover,

$$\frac{\mathbf{P}+\mathbf{P}^2+\cdots+\mathbf{P}^N}{N} \to \frac{1}{2}\begin{bmatrix}1&1\\1&1\end{bmatrix} = \begin{bmatrix}\pi\\\pi\end{bmatrix}.$$

Let us compute

$$\begin{bmatrix}\mathbb{E}_1R_1\\\mathbb{E}_2R_1\end{bmatrix} = \left(\begin{bmatrix}1&0\\0&1\end{bmatrix} - \begin{bmatrix}0&1\\0&0\end{bmatrix}\right)^{-1}\begin{bmatrix}1\\1\end{bmatrix} = \begin{bmatrix}2\\1\end{bmatrix}$$

and

$$\begin{bmatrix}\mathbb{E}_1R_2\\\mathbb{E}_2R_2\end{bmatrix} = \left(\begin{bmatrix}1&0\\0&1\end{bmatrix} - \begin{bmatrix}0&0\\1&0\end{bmatrix}\right)^{-1}\begin{bmatrix}1\\1\end{bmatrix} = \begin{bmatrix}1\\2\end{bmatrix}$$

so that indeed, $\pi_1 = 1/\mathbb{E}_1R_1$ and $\pi_2 = 1/\mathbb{E}_2R_2$. Of course $R_1 = 2$ (P_1 -a.s.) and $R_2 = 2$ (P_2 -a.s.) so that it is obvious that $\mathbb{E}_1R_1 = \mathbb{E}_2R_2 = 2$.

Example 10.7. Again let $S = \{1,2\}$ and $\mathbf{P} = \begin{bmatrix}1&0\\0&1\end{bmatrix}$ with jump diagram in Figure 10.2. In this case the chain is not irreducible and every $\pi = [a\ b]$ with

Fig. 10.2. A simple non-irreducible chain.

$a+b=1$ and $a,b \geq 0$ is an invariant distribution.

Example 10.8. Suppose that $S = \{1,2,3\}$, and

$$\mathbf{P} = \left[\begin{array}{ccc}1&2&3\\\hline 0&1&0\\1/2&0&1/2\\1&0&0\end{array}\right]\begin{matrix}1\\2\\3\end{matrix}$$

has the jump graph given by 10.3. Notice that $\mathbf{P}^2_{11} > 0$ and $\mathbf{P}^3_{11} > 0$ that $\mathbf{P}$ is "aperiodic." We now find the invariant distribution,

$$\operatorname{Nul}(\mathbf{P}-I)^{\mathrm{tr}} = \operatorname{Nul}\begin{bmatrix}-1&\frac{1}{2}&1\\1&-1&0\\0&\frac{1}{2}&-1\end{bmatrix} = \mathbb{R}\begin{bmatrix}2\\2\\1\end{bmatrix}.$$

Therefore the invariant distribution is given by

$$\pi = \frac{1}{5}\begin{bmatrix}2&2&1\end{bmatrix}.$$

Let us now observe that

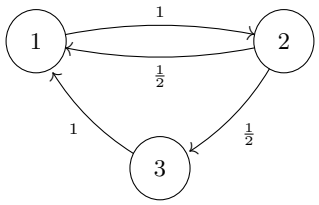


Fig. 10.3. A simple 3 state jump diagram.

$$\mathbf{P}^2 = \begin{bmatrix}\frac{1}{2}&0&\frac{1}{2}\\\frac{1}{2}&\frac{1}{2}&0\\0&1&0\end{bmatrix}$$

$$\mathbf{P}^3 = \begin{bmatrix}0&1&0\\1/2&0&1/2\\1&0&0\end{bmatrix}^3 = \begin{bmatrix}\frac{1}{2}&\frac{1}{2}&0\\\frac{1}{4}&\frac{1}{2}&\frac{1}{4}\\\frac{1}{2}&0&\frac{1}{2}\end{bmatrix}$$

$$\mathbf{P}^{20} = \begin{bmatrix}\frac{409}{1024}&\frac{205}{512}&\frac{205}{1024}\\\frac{205}{512}&\frac{409}{1024}&\frac{205}{1024}\\\frac{205}{512}&\frac{205}{512}&\frac{51}{256}\end{bmatrix} = \begin{bmatrix}0.399\,41&0.400\,39&0.200\,20\\0.400\,39&0.399\,41&0.200\,20\\0.400\,39&0.400\,39&0.199\,22\end{bmatrix}.$$

Let us also compute $\mathbb{E}_2R_3$ via,

$$\begin{bmatrix}\mathbb{E}_1R_3\\\mathbb{E}_2R_3\\\mathbb{E}_3R_3\end{bmatrix} = \left(\begin{bmatrix}1&0&0\\0&1&0\\0&0&1\end{bmatrix} - \begin{bmatrix}0&1&0\\1/2&0&0\\1&0&0\end{bmatrix}\right)^{-1}\begin{bmatrix}1\\1\\1\end{bmatrix} = \begin{bmatrix}4\\3\\5\end{bmatrix}$$

so that

$$\frac{1}{\mathbb{E}_3R_3} = \frac{1}{5} = \pi_3.$$

Example 10.9. The transition matrix,

$$\mathbf{P} = \begin{array}{c}\begin{matrix}1&2&3\end{matrix}\\\begin{bmatrix}1/4&1/2&1/4\\1/2&0&1/2\\1/3&1/3&1/3\end{bmatrix}\end{array}\begin{matrix}1\\2\\3\end{matrix}$$

is represented by the jump diagram in Figure 10.4. This chain is aperiodic. We find the invariant distribution as,

$$\begin{aligned}\operatorname{Nul}(\mathbf{P}-I)^{\mathrm{tr}} &= \operatorname{Nul}\left(\left(\begin{bmatrix}1/4&1/2&1/4\\1/2&0&1/2\\1/3&1/3&1/3\end{bmatrix} - \begin{bmatrix}1&0&0\\0&1&0\\0&0&1\end{bmatrix}\right)^{\mathrm{tr}}\right)\\ &= \operatorname{Nul}\left(\left(\begin{bmatrix}-\frac{3}{4}&\frac{1}{2}&\frac{1}{3}\\\frac{1}{2}&-1&\frac{1}{3}\\\frac{1}{4}&\frac{1}{2}&-\frac{2}{3}\end{bmatrix}\right)\right) = \mathbb{R}\begin{bmatrix}1\\\frac{5}{6}\\1\end{bmatrix} = \mathbb{R}\begin{bmatrix}6\\5\\6\end{bmatrix}\end{aligned}$$

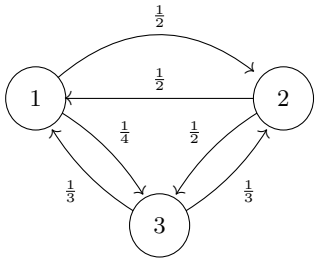


Fig. 10.4. In the above diagram there are jumps from 1 to 1 with probability 1/4 and jumps from 3 to 3 with probability 1/3 which are not explicitly shown but must be inferred by conservation of probability.

$$\pi = \frac{1}{17}\begin{bmatrix} 6 & 5 & 6 \end{bmatrix} = \begin{bmatrix} 0.352\,94 & 0.294\,12 & 0.352\,94 \end{bmatrix}.$$

In this case

$$\mathbf{P}^{10} = \begin{bmatrix} 1/4 & 1/2 & 1/4 \\ 1/2 & 0 & 1/2 \\ 1/3 & 1/3 & 1/3 \end{bmatrix}^{10} = \begin{bmatrix} 0.352\,98 & 0.294\,04 & 0.352\,98 \\ 0.352\,89 & 0.294\,23 & 0.352\,89 \\ 0.352\,95 & 0.294\,1 & 0.352\,95 \end{bmatrix}.$$

Let us also compute

$$\begin{bmatrix} \mathbb{E}_1 R_2 \\ \mathbb{E}_2 R_2 \\ \mathbb{E}_3 R_2 \end{bmatrix} = \left(\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} - \begin{bmatrix} 1/4 & 0 & 1/4 \\ 1/2 & 0 & 1/2 \\ 1/3 & 0 & 1/3 \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{11}{5} \\ \frac{17}{5} \\ \frac{13}{5} \end{bmatrix}$$

so that

$$1/\mathbb{E}_2 R_2 = 5/17 = \pi_2.$$

Example 10.10. Consider the following Markov matrix,

$$\mathbf{P} = \begin{bmatrix} 1 & 2 & 3 & 4 & \\ \hline 1/4 & 1/4 & 1/4 & 1/4 & 1 \\ 1/4 & 0 & 0 & 3/4 & 2 \\ 1/2 & 1/2 & 0 & 0 & 3 \\ 0 & 1/4 & 3/4 & 0 & 4 \end{bmatrix}$$

with jump diagram in Figure 10.5. Since this matrix is doubly stochastic (i.e $\sum_{i=1}^4 \mathbf{P}_{ij} = 1$ for all j as well as $\sum_{j=1}^4 \mathbf{P}_{ij} = 1$ for all i), it is easy to check that $\pi = \frac{1}{4}\begin{bmatrix} 1 & 1 & 1 & 1 \end{bmatrix}$. Let us compute $\mathbb{E}_3 R_3$ as follows

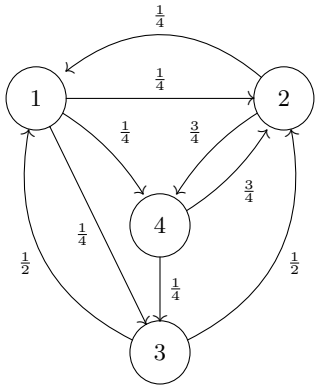


Fig. 10.5. The jump diagram for Q.

$$\begin{bmatrix} \mathbb{E}_1 R_3 \\ \mathbb{E}_2 R_3 \\ \mathbb{E}_3 R_3 \\ \mathbb{E}_4 R_3 \end{bmatrix} = \left(\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} - \begin{bmatrix} 1/4 & 1/4 & 0 & 1/4 \\ 1/4 & 0 & 0 & 3/4 \\ 1/2 & 1/2 & 0 & 0 \\ 0 & 1/4 & 0 & 0 \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$
$$= \begin{bmatrix} \frac{50}{17} \\ \frac{52}{17} \\ 4 \\ \frac{30}{17} \end{bmatrix}$$

so that $\mathbb{E}_3 R_3 = 4 = 1/\pi_4$ as it should be. Similarly,

$$\begin{bmatrix} \mathbb{E}_1 R_2 \\ \mathbb{E}_2 R_2 \\ \mathbb{E}_3 R_2 \\ \mathbb{E}_4 R_2 \end{bmatrix} = \left(\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} - \begin{bmatrix} 1/4 & 0 & 1/4 & 1/4 \\ 1/4 & 0 & 0 & 3/4 \\ 1/2 & 0 & 0 & 0 \\ 0 & 0 & 3/4 & 0 \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$
$$= \begin{bmatrix} \frac{54}{17} \\ 4 \\ \frac{44}{17} \\ \frac{50}{17} \end{bmatrix}$$

and again $\mathbb{E}_2 R_2 = 4 = 1/\pi_2$.

Example 10.11. Consider the following example,

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

with jump diagram given in Figure 10.6.We have

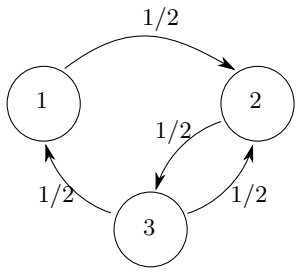


Fig. 10.6. The jump diagram associated to P.

$$\mathbf{P}^2 = \begin{bmatrix} 1/2 & 1/2 & 0 \\ 0 & 1/2 & 1/2 \\ 1/2 & 1/2 & 0 \end{bmatrix}^2 = \begin{bmatrix} \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \end{bmatrix}$$

and

$$\mathbf{P}^3 = \begin{bmatrix} 1/2 & 1/2 & 0 \\ 0 & 1/2 & 1/2 \\ 1/2 & 1/2 & 0 \end{bmatrix}^3 = \begin{bmatrix} \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \end{bmatrix}.$$

To have a picture what is going on here, imaging that $\pi = (\pi_1, \pi_2, \pi_3)$ represents the amount of sand at the sites, 1, 2, and 3 respectively. During each time step we move the sand on the sites around according to the following rule. The sand at site j after one step is $\sum_i \pi_i p_{ij}$, namely site i contributes p_{ij} fraction its sand, π_i, to site j. Everyone does this to arrive at a new distribution. Hence π is an invariant distribution if each π_i remains unchanged, i.e. $\pi = \pi\mathbf{P}$. (Keep in mind the sand is still moving around it is just that the size of the piles remains unchanged.)

As a specific example, suppose $\pi = (1, 0, 0)$ so that all of the sand starts at 1. After the first step, the pile at 1 is split into two and 1/2 is sent to 2 to get $\pi_1 = (1/2, 1/2, 0)$ which is the first row of $\mathbf{P}$. At the next step the site 1 keeps 1/2 of its sand ($= 1/4$) and still receives nothing, while site 2 again receives the other 1/2 and keeps half of what it had ($= 1/4 + 1/4$) and site 3 then gets ($1/2 \cdot 1/2 = 1/4$) so that $\pi_2 = \begin{bmatrix} \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \end{bmatrix}$ which is the first row of $\mathbf{P}^2$. It turns out in this case that this is the invariant distribution. Formally,

$$\begin{bmatrix} \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \end{bmatrix} \begin{bmatrix} 1/2 & 1/2 & 0 \\ 0 & 1/2 & 1/2 \\ 1/2 & 1/2 & 0 \end{bmatrix} = \begin{bmatrix} \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \end{bmatrix}.$$

In general we expect to reach the invariant distribution only in the limit as $n \to \infty$.

Notice that if π is any stationary distribution, then $\pi\mathbf{P}^n = \pi$ for all n and in particular,

$$\pi = \pi\mathbf{P}^2 = \begin{bmatrix} \pi_1 & \pi_2 & \pi_3 \end{bmatrix} \begin{bmatrix} \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \end{bmatrix} = \begin{bmatrix} \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \end{bmatrix}.$$

Hence $\begin{bmatrix} \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \end{bmatrix}$ is the unique stationary distribution for $\mathbf{P}$ in this case.

Example 10.12 (§3.2. p108 Ehrenfest Urn Model). Let a beaker filled with a particle fluid mixture be divided into two parts A and B by a semipermeable membrane. Let $X_n = (\#$ of particles in $A)$ which we assume evolves by choosing a particle at random from $A \cup B$ and then replacing this particle in the opposite bin from which it was found. Suppose there are N total number of particles in the flask, then the transition probabilities are given by,

$$p_{ij} = P(X_{n+1} = j \mid X_n = i) = \begin{cases} 0 & \text{if } j \notin \{i-1, i+1\} \\ \frac{i}{N} & \text{if } \quad j = i-1 \\ \frac{N-i}{N} & \text{if } \quad j = i+1. \end{cases}$$

For example, if $N = 2$ we have

$$(p_{ij}) = \left[\begin{array}{ccc} 0 & 1 & 2 \\ \hline 0 & 1 & 0 \\ 1/2 & 0 & 1/2 \\ 0 & 1 & 0 \end{array}\right] \begin{matrix} 0 \\ 1 \\ 2 \end{matrix}$$

and if $N = 3$, then we have in matrix form,

$$(p_{ij}) = \left[\begin{array}{cccc} 0 & 1 & 2 & 3 \\ \hline 0 & 1 & 0 & 0 \\ 1/3 & 0 & 2/3 & 0 \\ 0 & 2/3 & 0 & 1/3 \\ 0 & 0 & 1 & 0 \end{array}\right] \begin{matrix} 0 \\ 1 \\ 2 \\ 3 \end{matrix}.$$

In the case $N = 2$,

$$\begin{bmatrix} 0 & 1 & 0 \\ 1/2 & 0 & 1/2 \\ 0 & 1 & 0 \end{bmatrix}^2 = \begin{bmatrix} \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 1 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \end{bmatrix}$$

$$\begin{bmatrix} 0 & 1 & 0 \\ 1/2 & 0 & 1/2 \\ 0 & 1 & 0 \end{bmatrix}^3 = \begin{bmatrix} 0 & 1 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 1 & 0 \end{bmatrix}$$

and when $N = 3$,

$$\begin{bmatrix} 0 & 1 & 0 & 0 \\ 1/3 & 0 & 2/3 & 0 \\ 0 & 2/3 & 0 & 1/3 \\ 0 & 0 & 1 & 0 \end{bmatrix}^2 = \begin{bmatrix} \frac{1}{3} & 0 & \frac{2}{3} & 0 \\ 0 & \frac{7}{9} & 0 & \frac{2}{9} \\ \frac{2}{9} & 0 & \frac{7}{9} & 0 \\ 0 & \frac{2}{3} & 0 & \frac{1}{3} \end{bmatrix}$$

$$\begin{bmatrix} 0 & 1 & 0 & 0 \\ 1/3 & 0 & 2/3 & 0 \\ 0 & 2/3 & 0 & 1/3 \\ 0 & 0 & 1 & 0 \end{bmatrix}^3 = \begin{bmatrix} 0 & \frac{7}{9} & 0 & \frac{2}{9} \\ \frac{7}{27} & 0 & \frac{20}{27} & 0 \\ 0 & \frac{20}{27} & 0 & \frac{7}{27} \\ \frac{2}{9} & 0 & \frac{7}{9} & 0 \end{bmatrix}$$

$$\begin{bmatrix} 0 & 1 & 0 & 0 \\ 1/3 & 0 & 2/3 & 0 \\ 0 & 2/3 & 0 & 1/3 \\ 0 & 0 & 1 & 0 \end{bmatrix}^{25} \cong \begin{bmatrix} 0.0 & 0.75 & 0.0 & 0.25 \\ 0.25 & 0.0 & 0.75 & 0.0 \\ 0.0 & 0.75 & 0.0 & 0.25 \\ 0.25 & 0.0 & 0.75 & 0.0 \end{bmatrix}$$

$$\begin{bmatrix} 0 & 1 & 0 & 0 \\ 1/3 & 0 & 2/3 & 0 \\ 0 & 2/3 & 0 & 1/3 \\ 0 & 0 & 1 & 0 \end{bmatrix}^{26} \cong \begin{bmatrix} 0.25 & 0.0 & 0.75 & 0.0 \\ 0.0 & 0.75 & 0.0 & 0.25 \\ 0.25 & 0.0 & 0.75 & 0.0 \\ 0.0 & 0.75 & 0.0 & 0.25 \end{bmatrix}$$

:

$$\begin{bmatrix} 0 & 1 & 0 & 0 \\ 1/3 & 0 & 2/3 & 0 \\ 0 & 2/3 & 0 & 1/3 \\ 0 & 0 & 1 & 0 \end{bmatrix}^{100} \cong \begin{bmatrix} 0.25 & 0.0 & 0.75 & 0.0 \\ 0.0 & 0.75 & 0.0 & 0.25 \\ 0.25 & 0.0 & 0.75 & 0.0 \\ 0.0 & 0.75 & 0.0 & 0.25 \end{bmatrix}$$

We also have

$$(\mathbf{P} - I)^{\mathrm{tr}} = \begin{bmatrix} -1 & 1 & 0 & 0 \\ \frac{1}{3} & -1 & \frac{2}{3} & 0 \\ 0 & \frac{2}{3} & -1 & \frac{1}{3} \\ 0 & 0 & 1 & -1 \end{bmatrix}^{\mathrm{tr}} = \begin{bmatrix} -1 & \frac{1}{3} & 0 & 0 \\ 1 & -1 & \frac{2}{3} & 0 \\ 0 & \frac{2}{3} & -1 & 1 \\ 0 & 0 & \frac{1}{3} & -1 \end{bmatrix}$$

and

$$\mathrm{Nul}\left((\mathbf{P} - I)^{\mathrm{tr}}\right) = \begin{bmatrix} 1 \\ 3 \\ 3 \\ 1 \end{bmatrix}.$$

Hence if we take, $\pi = \frac{1}{8}\begin{bmatrix} 1 & 3 & 3 & 1 \end{bmatrix}$ then

$$\pi\mathbf{P} = \frac{1}{8}\begin{bmatrix} 1 & 3 & 3 & 1 \end{bmatrix}\begin{bmatrix} 0 & 1 & 0 & 0 \\ 1/3 & 0 & 2/3 & 0 \\ 0 & 2/3 & 0 & 1/3 \\ 0 & 0 & 1 & 0 \end{bmatrix} = \frac{1}{8}\begin{bmatrix} 1 & 3 & 3 & 1 \end{bmatrix} = \pi$$

is the stationary distribution. Notice that

$$\frac{1}{2}\left(\mathbf{P}^{25} + \mathbf{P}^{26}\right) \cong \frac{1}{2}\begin{bmatrix} 0.0 & 0.75 & 0.0 & 0.25 \\ 0.25 & 0.0 & 0.75 & 0.0 \\ 0.0 & 0.75 & 0.0 & 0.25 \\ 0.25 & 0.0 & 0.75 & 0.0 \end{bmatrix} + \frac{1}{2}\begin{bmatrix} 0.25 & 0.0 & 0.75 & 0.0 \\ 0.0 & 0.75 & 0.0 & 0.25 \\ 0.25 & 0.0 & 0.75 & 0.0 \\ 0.0 & 0.75 & 0.0 & 0.25 \end{bmatrix}$$

$$= \begin{bmatrix} 0.125 & 0.375 & 0.375 & 0.125 \\ 0.125 & 0.375 & 0.375 & 0.125 \\ 0.125 & 0.375 & 0.375 & 0.125 \\ 0.125 & 0.375 & 0.375 & 0.125 \end{bmatrix} = \begin{bmatrix} \pi \\ \pi \\ \pi \\ \pi \end{bmatrix}.$$

Example 10.13. Let us consider the Markov matrix,

$$\mathbf{P} = \begin{bmatrix} 1 & 2 & 3 \\ \hline 0 & 1 & 0 \\ 1/2 & 0 & 1/2 \\ 1 & 0 & 0 \end{bmatrix} \begin{matrix} 1 \\ 2 \\ 3 \end{matrix}.$$

In this case we have

$$\mathbf{P}^{25} = \begin{bmatrix} 0 & 1 & 0 \\ 1/2 & 0 & 1/2 \\ 1 & 0 & 0 \end{bmatrix}^{25} \cong \begin{bmatrix} 0.399\,9 & 0.400\,15 & 0.199\,95 \\ 0.400\,02 & 0.399\,9 & 0.200\,07 \\ 0.400\,15 & 0.399\,9 & 0.199\,95 \end{bmatrix}$$

$$\mathbf{P}^{26} = \begin{bmatrix} 0 & 1 & 0 \\ 1/2 & 0 & 1/2 \\ 1 & 0 & 0 \end{bmatrix}^{26} \cong \begin{bmatrix} 0.400\,02 & 0.399\,9 & 0.200\,07 \\ 0.400\,02 & 0.400\,02 & 0.199\,95 \\ 0.399\,9 & 0.400\,15 & 0.199\,95 \end{bmatrix}$$

$$\mathbf{P}^{100} = \begin{bmatrix} 0 & 1 & 0 \\ 1/2 & 0 & 1/2 \\ 1 & 0 & 0 \end{bmatrix}^{100} \cong \begin{bmatrix} 0.4 & 0.4 & 0.2 \\ 0.4 & 0.4 & 0.2 \\ 0.4 & 0.4 & 0.2 \end{bmatrix}$$

and observe that

$$\begin{bmatrix} 0.4 & 0.4 & 0.2 \end{bmatrix}\begin{bmatrix} 0 & 1 & 0 \\ 1/2 & 0 & 1/2 \\ 1 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0.4 & 0.4 & 0.2 \end{bmatrix}.$$

so that $\pi = \begin{bmatrix} 0.4 & 0.4 & 0.2 \end{bmatrix}$ is a stationary distribution for $\mathbf{P}$.

10.2 Extra Homework Problems

Exercises 10.1 – 10.4 refer to the following Markov matrix:

$$\mathbf{P} = \left[\begin{array}{cccccc|c} 1 & 2 & 3 & 4 & 5 & 6 & \\ \hline 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 1/2 & 1/2 & 0 & 0 & 0 & 0 & 2 \\ 0 & 0 & 1/2 & 1/2 & 0 & 0 & 3 \\ 0 & 0 & 1 & 0 & 0 & 0 & 4 \\ 0 & 1/2 & 0 & 0 & 0 & 1/2 & 5 \\ 0 & 0 & 0 & 1/4 & 3/4 & 0 & 6 \end{array}\right] \tag{10.5}$$

We will let $\{X_n\}_{n=0}^{\infty}$ denote the Markov chain associated to $\mathbf{P}$.

Exercise 10.1. Make a jump diagram for this matrix and identify the recurrent and transient classes. Also find the invariant destitutions for the chain restricted to each of the recurrent classes.

Exercise 10.2. Find all of the invariant distributions for $\mathbf{P}$.

Exercise 10.3. Compute the hitting probabilities, $h_5 = P_5\left(X_n \text{ hits } \{3,4\}\right)$ and $h_6 = P_6\left(X_n \text{ hits } \{3,4\}\right)$.

Exercise 10.4. Find $\lim_{n\to\infty} P_6\left(X_n = j\right)$ for $j = 1,2,3,4,5,6$.

References

1. Richard Durrett, *Probability: theory and examples*, second ed., Duxbury Press, Belmont, CA, 1996. MR MR1609153 (98m:60001)
2. William Feller, *An Introduction to Probability Theory and Its Applications. Vol. I*, John Wiley & Sons Inc., New York, N.Y., 1950. MR MR0038583 (12,424a)
3. Olav Kallenberg, *Foundations of modern probability*, second ed., Probability and its Applications (New York), Springer-Verlag, New York, 2002. MR MR1876169 (2002m:60002)
4. J. R. Norris, *Markov chains*, Cambridge Series in Statistical and Probabilistic Mathematics, vol. 2, Cambridge University Press, Cambridge, 1998, Reprint of 1997 original. MR MR1600720 (99c:60144)
5. Sheldon M. Ross, *Stochastic processes*, Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics, John Wiley & Sons Inc., New York, 1983, Lectures in Mathematics, 14. MR MR683455 (84m:60001)