Bruce K. Driver

180B Lecture Notes, W2011

January 27, 2011 File:180Lec.tex
## Contents

### 8 First Step Analysis

- 8.1 Finite state space chains .................................................. 43
- 8.2 First Return Times ................................................................. 45
- 8.3 First Step Analysis I ................................................................. 49
- 8.4 First Step Analysis Examples II ............................................... 50
  - 8.4.1 A rat in a maze example Problem 5 on p.131 ................... 54
- 8.5 Computations avoiding the first step analysis ....................... 54
- 8.6 General facts about sub-probability kernels ......................... 56

### 9 Stationary Distributions .................................................... 59

### 10 Invariant distributions and return times .......................... 63

- 10.0.1 Some worked examples .................................................. 64
- 10.1 Random Walk Exercises I ..................................................... 66
- 10.2 Random Walk Stuff II .......................................................... 67

### 11 Long Run Behavior of Discrete Markov Chains .................... 71

- 11.1 The Main Results ............................................................... 71
  - 11.1.1 Finite State Space Remarks ........................................... 76
- 11.2 Examples ........................................................................... 77
- 11.3 The Strong Markov Property .............................................. 80
- 11.4 Irreducible Recurrent Chains ............................................ 83

## References .............................................................................. 87
Basic Probability Facts / Conditional Expectations

0.1 Course Notation

1. \((\Omega, P)\) 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 \(X: \Omega \rightarrow S\) we let
\[
\rho_X(s) := P(X = s).
\]
More generally if \(X_i: \Omega \rightarrow S_i\) for \(1 \leq i \leq n\) we let
\[
\rho_{X_1, \ldots, X_n}(s) := P(X_1 = s_1, \ldots, X_n = s_n)
\]
for all \(s = (s_1, \ldots, s_n) \in S_1 \times \cdots \times S_n\).
3. If \(S = \mathbb{R}\) or \(\mathbb{R}^n\) and \(X: \Omega \rightarrow S\) 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) \(\text{Var}(X) := \mathbb{E}[(X - \mu_X)^2] = \mathbb{E}X^2 - \mu_X^2\) be the variance of \(X\).
   c) \(\sigma_X := \sqrt{\text{Var}(X)}\) be the standard deviation of \(X\).
   d) \(\text{Cov}(X, Y) := \mathbb{E}[(X - \mu_X)(Y - \mu_Y)] = \mathbb{E}[XY] - \mu_X \mu_Y\) be the covariance of \(X\) and \(Y\).
   e) \(\text{Corr}(X, Y) := \text{Cov}(X, Y) / (\sigma_X \sigma_Y)\) be the correlation of \(X\) and \(Y\).

0.2 Some Discrete Distributions

Definition 0.1 (Generating Function). Suppose that \(N: \Omega \rightarrow \mathbb{N}_0\) is an integer valued random variable on a probability space, \((\Omega, \mathcal{B}, P)\). The generating function associated to \(N\) is defined by
\[
G_N(z) := \mathbb{E}[z^N] = \sum_{n=0}^{\infty} P(N = n) z^n \text{ for } |z| \leq 1.
\]  
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,
\[
G_N^{(k)}(z) = \mathbb{E}[N(N - 1) \cdots (N - k + 1) z^{N-k}] \text{ for } |z| < 1
\]
and
\[
G^{(k)}(1) = \lim_{z \uparrow 1} G^{(k)}(z) = \mathbb{E}[N(N - 1) \cdots (N - k + 1)],
\]
where it is possible that one and hence both sides of this equation are infinite. In particular, \(G'(1):= \lim_{z \uparrow 1} \frac{G^{(1)}(z)}{z} = \mathbb{E}N\) and if \(\mathbb{E}N^2 < \infty\),
\[
\text{Var}(N) = G''(1) + G'(1) - [G'(1)]^2.
\]

Proof. By Corollary ?? for \(|z| < 1\),
\[
G_N^{(k)}(z) = \sum_{n=0}^{\infty} P(N = n) \cdot n(n - 1) \cdots (n - k + 1) z^{n-k}
= \mathbb{E}[N(N - 1) \cdots (N - k + 1) z^{N-k}].
\]

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 \(\text{Var}(N)\).

1. Bernoulli\((p) : P(N = 1) = p \) and \(P(N = 0) = 1 - p\). You should find \(\mathbb{E}N = p\) and \(\text{Var}(N) = p - p^2\).
2. Binomial($n, p$) : $P(N = k) = \binom{n}{k}p^k(1-p)^{n-k}$ for $k = 0, 1, \ldots, n$. ($P(N = k)$ 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 $\text{Var}(N) = n(p - p^2)$.

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

4. Poisson($\lambda$) : $P(N = k) = \frac{\lambda^k e^{-\lambda}}{k!}$ for all $k \in \mathbb{N}_0$. You should find $\mathbb{E}N = \lambda = \text{Var}(N)$.

**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 $\text{Var}(N) = 0 + p - p^2$.

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

   \[ \mathbb{E}N = np \quad \text{and} \quad \text{Var}(N) = n(n-1)p^2 + np - (np)^2 = n(p - p^2). \]

3. For the geometric distribution,

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

   Differentiating this equation in $z$ implies,

   \[ \mathbb{E}[Nz^{N-1}] = G'_N(z) = \frac{p[1-z(1-p)] + (1-p)pz}{(1-z(1-p))^2}, \]

   \[ = \frac{p}{(1-z(1-p))^2} \quad \text{and} \]

   \[ \mathbb{E}[N(N-1)z^{N-2}] = G''_N(z) = \frac{2(1-p)p}{(1-z(1-p))^3}. \]

   Therefore,

   \[ \mathbb{E}N = G'_N(1) = 1/p, \]

   \[ \mathbb{E}[N(N-1)] = \frac{2(1-p)p}{p^3} = \frac{2(1-p)p}{p^2}. \]

4. In the Poisson case,

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

   and $G''_N(z) = \lambda^ke^{\lambda(z-1)}$. Therefore, $\mathbb{E}N = \lambda$ and $\mathbb{E}[N \cdot (N - 1)] = \lambda^2$ so that $\text{Var}(N) = \lambda^2 + \lambda - \lambda^2 = \lambda$.

**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 $P(N = k) = P(X_1 = 0, \ldots, X_{k-1} = 0, X_k = 1) = (1-p)^{k-1}p$, so that $N$ is geometric with parameter $p$. Using this representation we easily and intuitively see that

\[ P(N = n + k | N > n) = \frac{P(X_1 = 0, \ldots, X_{n+k-1} = 0, X_{n+k} = 1)}{P(X_1 = 0, \ldots, X_n = 0)} = P(X_{n+1} = 0, \ldots, X_{n+k-1} = 0, X_{n+k} = 1) = P(X_1 = 0, \ldots, X_{k-1} = 0, X_k = 1) = P(N = k). \]
This can be verified by first principles as well:

\[ P (N = n + k | N > n) = \frac{P(N = n + k)}{P(N > n)} = \frac{p (1 - p)^{n+k-1}}{\sum_{k>n} p (1 - p)^{k-1}} \]

\[ = \frac{p (1 - p)^{n+k-1}}{\sum_{j=0}^{\infty} p (1 - p)^{n+j} = (1 - p)^n \sum_{j=0}^{\infty} (1 - p)^j} = \frac{(1 - p)^{k-1}}{1 - (1 - p)} = p (1 - p)^{k-1} = P(N = k). \]

**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 \( n \),

\[ P(\text{Binomial}(n,p) = k) \approx P(\text{Poisson}(pn) = k) = \left(\frac{pn}{k!}\right)^k e^{-pn}. \]

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

**Solution to Exercise (0.2).** We have,

\[ P(S_{n,p_n} = k) = \binom{n}{k} \left(\frac{\lambda_n}{n}\right)^k (1 - \frac{\lambda_n}{n})^{n-k} \]

\[ = \frac{\lambda^k}{k!} \frac{n(n-1)\ldots(n-k+1)}{n^k} \left(1 - \frac{\lambda_n}{n}\right)^{n-k}. \]

The result now follows since,

\[ \lim_{n \to \infty} \frac{n(n-1)\ldots(n-k+1)}{n^k} = 1 \]

and

\[ \lim_{n \to \infty} \ln (1 - \frac{\lambda_n}{n})^{n-k} = \lim_{n \to \infty} \ln (n-k) \ln (1 - \frac{\lambda_n}{n}) \]

\[ = -\lim_{n \to \infty} [(n-k) \lambda_n/n] = -\lambda. \]
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 στόχος 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 and E. Nelson, 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 \( X_t : \Omega \rightarrow S \) (\( S \) is the state space) on some probability space, \( (\Omega, P) \). 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. \( X_n \in S = Perm\{\{1,\ldots,52\}\} \) 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.

---


1.1 180B Course Topics:

2. Use the least squares theory to interpret covariance and correlations.
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 \( \mathbb{E}[f(X)]^2 < \infty \), that is \( h(X) \) is a least squares approximation to \( Y \) among random variables of the form \( f(X) \), i.e.

\[
\mathbb{E}(Y - h(X))^2 = \min_{f} \mathbb{E}(Y - f(X))^2.
\]  

**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**
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

$$E(Y - a_0 X + b_0)^2 = \min_{a,b \in \mathbb{R}} E(Y - a X + b)^2.$$  \hspace{1cm} (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 $E[XY]$. 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 $\min_{c \in \mathbb{R}} E(Y - c)^2 = \text{Var}(Y)$ which occurs for $c = EY$. Thus $EY$ is the least squares approximation to $Y$ by a constant function and $\text{Var}(Y)$ is the least square error associated with this problem.
Covariance and Correlation

Suppose that $\Omega$ 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 $L^1(\mathbb{P})$ and the square integrable random variables by $L^2(\mathbb{P})$. When $X$ is integrable we let $\mu_X := \mathbb{E} X$ be the mean of $X$. If $\Omega$ is a finite set, then

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

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 $\Omega$ is infinite life can become a little more complicated.

Example 2.1. Suppose that $\Omega$ is a geometric with parameter $p$ so that $P(N = k) = p (1 - p)^{k-1}$ for $k \in \mathbb{N} = \{1, 2, 3, \ldots\}$. 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 $XN = \lambda^N$ for some $\lambda > 0$ we have

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

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

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

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

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

it follows that

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

Moreover,

$$(X + Y)^2 = X^2 + Y^2 + 2XY \leq X^2 + Y^2 + 2|XY| \leq 2 (X^2 + Y^2)$$

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

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

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

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

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

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

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

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

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

$$0 \leq \text{Var} (X) \leq \mathbb{E} [X^2] \text{ for all } X \in L^2(\mathbb{P}).$$

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

1. Show that $X$ and $Y$ are independent iff $\text{Cov} (f(X), g(Y)) = 0$ (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, $\mathcal{S}$.)
2. If $X,Y \in L^2(\mathbb{P})$ and $X$ and $Y$ are independent, then $\text{Cov} (X,Y) = 0$.

Note well: we will see in examples below that $\text{Cov} (X,Y) = 0$ 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

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

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

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

from which item 1. easily follows.

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

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

Since

\[ |f_M(X) g_M(Y)| \leq |XY| \leq \frac{1}{2} (X^2 + Y^2) \leq L^1(P), \]

\[ |f_M(X)| \leq |X| \leq \frac{1}{2} (1 + X^2) \leq L^1(P), \]

\[ |g_M(Y)| \leq |Y| \leq \frac{1}{2} (1 + Y^2) \leq L^1(P), \]

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. \( \text{Cov}(X,Y) = 0 \). (These technical details were omitted in class.)


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

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

Therefore,

\[ \mathbb{E}X = \int \int ye^{-y}1_{0<y} dx dy = \int_0^\infty x^2 e^{-y} dy = 2, \]

\[ \mathbb{E}X^2 = \int \int x^2 e^{-y}1_{0<y} dx dy = \int_0^\infty y^3 e^{-y} dy = 3! = 6 \]

\[ \mathbb{E}X = \int \int ye^{-y}1_{0<y} dx dy = \frac{1}{2} \int_0^\infty y^3 e^{-y} dy = 1, \]

\[ \mathbb{E}X^2 = \int \int x^2 e^{-y}1_{0<y} dx dy = \frac{1}{3} \int_0^\infty y^3 e^{-y} dy = \frac{1}{3} \cdot 3! = 2 \]

and

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

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

\[ \text{Corr}(X,Y) = \frac{1}{\sqrt{2}}. \]

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

\[ \text{Var}(S_n) = \sum_{k=1}^n \text{Var}(X_k). \]

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

\[ \text{Var}(X+k) = \text{Cov}(X+k, X+k) = \text{Cov}(X+k, X) + \text{Cov}(X+k, k) \]

\[ = \text{Cov}(X+k, X) = \text{Cov}(X, X) + \text{Cov}(k, X) \]

\[ = \text{Cov}(X, X) = \text{Var}(X), \]

wherein we have used the bilinearity of \( \text{Cov} \) and the property that \( \text{Cov}(Y,k) = 0 \) whenever \( k \) is a constant. \( \blacksquare \)
Example 2.7. Suppose that $X$ and $Y$ are distributed as follows:

\[
\begin{array}{cccc}
\rho_Y & 1/4 & 1/2 & 1/4 \\
\rho_X & 1/4 & 0 & 1/4 \\
1/4 & 1 & 0 & 1/4 \\
3/4 & 0 & 1/4 & 1/4 \\
\end{array}
\]

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 $E[XY] = 0$ while

$$
E[X] = 1 \cdot \frac{1}{4} + 0 \cdot \frac{3}{4} = \frac{1}{4}, \quad \text{and}
$$

$$
EY = (-1) \frac{1}{4} + 0 \frac{1}{4} + 1 \frac{1}{4} = 0
$$

so that 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

$$
\text{Cov}(X,Y) = E[X^3] - E[X^2] \cdot EX = 0
$$

since,

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

On the other hand Cov $(Y,X^2) = \text{Cov}(Y,Y) = \text{Var}(Y) \neq 0$ 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 $EZ = 0$ and

$$
$$

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

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

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 $S_n := \sum_{k=1}^{n} X_k$, $\mu := E X_k$ and $\sigma^2 := \text{Var}(X_k)$ (these are independent of $k$). Show;

$$
E\left[\frac{S_n}{n}\right] = \mu,
$$

$$
E\left(\frac{S_n}{n} - \mu\right)^2 = \text{Var}\left(\frac{S_n}{n}\right) = \frac{\sigma^2}{n}, \text{ and}
$$

$$
P\left(\left|\frac{S_n}{n} - \mu\right| > \varepsilon\right) \leq \frac{\sigma^2}{n \varepsilon^2}
$$

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

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 $\|X\| := \sqrt{\mathbb{E}[X^2]}$.

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

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

where

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

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. $\|X\|^2 := (X, X) \geq 0$ 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,

$$\|X + Y\|^2 = (X + Y, X + Y) = \|X\|^2 + \|Y\|^2 + (X, Y) + (Y, X) = \|X\|^2 + \|Y\|^2 + 2(X, Y). \quad (3.1)$$

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

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

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

$$E[XY] \leq \|X\| \|Y\|.$$

**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)

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

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

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

$$0 \leq \|Z\|^2 = \|X - cY\|^2 = \|X\|^2 - 2c(X, Y) + c^2 \|Y\|^2.$$
Proposition 3.7. 

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

\[
0 \leq \|X - cY\|^2 = \|X\|^2 - (X, Y)^2 / \|Y\|^2
\]

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,

\[
\|X + Y\|^2 = \|X\|^2 + \|Y\|^2 + 2(X, Y)
\]

\[
\leq \|X\|^2 + \|Y\|^2 + 2\|X\|\|Y\| = (\|X\| + \|Y\|)^2.
\]

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

\[
|(X + \Delta X, Y + \Delta Y) - (X, Y)| = |(X, \Delta Y) + (\Delta X, Y) + (\Delta X, \Delta Y)|
\]

\[
\leq \|X\|\|\Delta Y\| + \|Y\|\|\Delta X\| + \|\Delta X\|\|\Delta Y\|
\]

\[
\to 0 \quad \text{as} \quad \Delta X, \Delta Y \to 0,
\]

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 \( S \subset H \) is a finite orthogonal set, then

\[
\left\| \sum_{X \in S} X \right\|^2 = \sum_{X \in S} \|X\|^2.
\]  

(3.2)

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

\[
\|X + Y\|^2 + \|X - Y\|^2 = 2\|X\|^2 + 2\|Y\|^2.
\]  

(3.3)

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

\[
\left\| \sum_{X \in S} X \right\|^2 = \left( \sum_{X \in S} X, \sum_{Y \in S} Y \right) = \sum_{X, Y \in S} (X, Y)
\]

\[
= \sum_{X \in S} (X, X) = \sum_{X \in S} \|X\|^2
\]

and

\[
\|X + Y\|^2 + \|X - Y\|^2
\]

\[
= \|X\|^2 + \|Y\|^2 + 2(X, Y) + \|X\|^2 + \|Y\|^2 - 2(X, Y)
\]

\[
= 2\|X\|^2 + 2\|Y\|^2.
\]

**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. \( \|Y - X\| \geq \|Y - Z\| \) 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 \( Q_Y \) 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 \( f(0) = 0 \). Since

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

we may conclude that

\[
0 = f(0) = -2(Y - X, Z).
\]

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

\[
\|Y - Z\|^2 = \|Y - X + X - Z\|^2 = \|Y - X\|^2 + \|X - Z\|^2 \geq \|Y - X\|^2.
\]

(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

\[
\|X - Z\|^2 = E(X - Z)^2 = 0,
\]

i.e. iff \( X = Z \) a.s.
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

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

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

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

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 $Y$,

$$|\text{Cov}(X,Y)| \leq \sigma(X) \cdot \sigma(Y)$$

or equivalently,

$$|\text{Corr}(X,Y)| \leq 1.$$

**Proof.** This is a simply application of Schwarz’s inequality (Theorem 3.4):

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

Since $\text{Corr}(X,Y) > 0$ iff $\text{Cov}(X,Y) > 0$ 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 $\text{Corr}(X,Y) = 1$ or $\text{Corr}(X,Y) = -1$.)

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

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

where

$$W = \mu_Y + \frac{\text{Cov}(X,Y)}{\text{Var}(X)} (X - \mu_X) = \mu_X + \frac{\text{Cov}(X,Y)}{\text{Var}(X)} (Y - \mu_Y).$$

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

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

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

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

so that

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

**Example 4.3.** Suppose that $P(X \in dx, Y \in dy) = e^{-y} 1_{0<x<y} dx dy$. Recall from Example 2.3 that

$$\begin{align*}
\mathbb{E}X &= 1, \quad \mathbb{E}Y = 2, \\
\mathbb{E}X^2 &= 2, \quad \mathbb{E}Y^2 = 6, \\
\sigma(X) &= 1, \quad \sigma(Y) = \sqrt{2}, \\
\text{Cov}(X,Y) &= 1, \quad \text{Corr}(X,Y) = \frac{1}{\sqrt{2}}.
\end{align*}$$

So in this case

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

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

$$Y = \mu_Y + \frac{\text{Cov} \ (X,Y)}{\text{Var} \ (X)} \ (X - \mu_X)$$

$$= \mu_Y \pm \frac{\sigma_X \sigma_Y}{\sigma_X} \ (X - \mu_X) = \mu_Y \pm \frac{\sigma_Y}{\sigma_X} \ (X - \mu_X),$$

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

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

$$X = \mu_X + \sigma \ (X) Z_1, \text{ and}$$

$$Y = \mu_Y + \sigma \ (Y) \ [\cos \theta \cdot Z_1 + \sin \theta \cdot Z_2],$$

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

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

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

To get $Z_2$ we need to normalize $\bar{Z}_2$ using:

$$\mathbb{E} \bar{Z}_2^2 = \sigma \ (Y)^2 - 2 \frac{(\bar{Y}, \bar{X})}{\sigma \ (X)^2} (\bar{X}, \bar{Y}) + \frac{(\bar{Y}, \bar{X})^2}{\sigma \ (X)^2} \sigma \ (X)^2$$

$$= \sigma \ (Y)^2 - \frac{(\bar{X}, \bar{Y})^2}{\sigma \ (X)^2} = \sigma \ (Y)^2 \left(1 - \text{Corr}^2 \ (X,Y)\right)$$

$$= \sigma \ (Y)^2 \sin^2 \theta.$$

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

$$Z_2 := \frac{\bar{Z}_2}{\sqrt{\bar{Z}_2^2}} = \frac{\bar{Y} - \frac{(\bar{Y}, \bar{X})}{\sigma \ (X)^2} \bar{X}}{\sigma \ (Y) \sin \theta} = \frac{\bar{Y} - \sigma \ (X) \sigma \ (Y) \text{Corr} \ (X,Y)}{\sigma \ (Y) \sin \theta} \frac{\bar{X}}{\sigma \ (X)^2}$$

$$= \frac{\bar{Y} - \frac{\sigma \ (Y)}{\sigma \ (X)} \cos \theta \cdot \bar{X}}{\sigma \ (Y) \sin \theta} = \frac{\bar{Y} - \sigma \ (Y) \cos \theta \cdot Z_1}{\sigma \ (Y) \sin \theta}.$$

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 $\text{Corr} \ (X,Y) = 1$, then $\theta = 0$ and $\bar{Y} = \sigma \ (Y) Z_1 = \frac{\sigma \ (Y)}{\sigma \ (X)} \bar{X}$ while if $\text{Corr} \ (X,Y) = -1$, 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 $f,g : \mathbb{R} \rightarrow \mathbb{R}$ 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 $\text{Cov} \ (f \ (X), g \ (X)) \geq 0$. Hint: let $Y$ be another random variable which has the same law as $X$ and is independent of $X$. Then consider

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

Notation 5.1 (Conditional Expectation 1) Given $Y \in L^1(P)$ and $A \subset \Omega$ let
\[ E[Y : A] := E[1_A Y] \]
and
\[ E[Y | A] = \begin{cases} \frac{E[Y : A]}{P(A)} & \text{if } P(A) > 0 \\ 0 & \text{if } P(A) = 0. \end{cases} \tag{5.1} \]
(In point of fact, when $P(A) = 0$ we could set $E[Y | A]$ to be any real number. We choose 0 for definiteness and so that $Y \rightarrow E[Y | A]$ is always linear.)

Example 5.2 (Conditioning for the uniform distribution). Suppose that $\Omega$ 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 $Y : \Omega \rightarrow \mathbb{R}$ we have $E[Y | A]$ is the expectation of $Y$ restricted to $A$ under the uniform distribution on $A$. Indeed,
\[ E[Y | A] = \frac{1}{P(A)} E[Y : A] = \frac{1}{P(A)} \sum_{\omega \in A} Y(\omega) P(\{\omega\}) = \frac{1}{\#(A)/\#(\Omega)} \sum_{\omega \in A} Y(\omega) = \frac{1}{\#(A)} \sum_{\omega \in A} Y(\omega). \]

Lemma 5.3. If $P(A) > 0$ then $E[Y | A] = E_{P(\cdot | A)} Y$ 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 $Y : \Omega \rightarrow S$ 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
\[ E_{P(\cdot | A)} Y = E_{P(\cdot | A)} \sum_{y \in S} y 1_{y = y} = \sum_{y \in S} y E_{P(\cdot | A)} 1_{y = y} = \sum_{y \in S} y P(Y = y | A) = \sum_{y \in S} y P(Y = y | A) \cdot \frac{P(A)}{P(A)} = \frac{1}{P(A)} \sum_{y \in S} y E[1_A Y | A] = \frac{1}{P(A)} E[1_A Y] \Rightarrow E[Y | A]. \]

Lemma 5.4. No matter whether $P(A) > 0$ or $P(A) = 0$ we always have,
\[ |E[Y | A]| \leq E[|Y| | A] \leq \sqrt{E[|Y|^2 | A]}. \tag{5.2} \]

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
\[ |E[Y | A]| = |E_{P(\cdot | A)} Y| \leq E_{P(\cdot | A)} |Y| \leq \sqrt{E_{P(\cdot | A)} |Y|^2} \cdot E_{P(\cdot | A)} 1 = \sqrt{E_{P(\cdot | A)} |Y|^2}. \]
This completes that proof as $E_{P(\cdot | A)} |Y| = E[|Y| | A]$ and $E_{P(\cdot | A)} |Y|^2 = E[|Y|^2 | A]$.

Notation 5.5 Let $S$ be a set (often $S = \mathbb{R}$ or $S = \mathbb{R}^N$) and suppose that $X : \Omega \rightarrow S$ is a function. (So $S$ 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 $f : S \rightarrow \mathbb{R}$. (This is a subspace of $L^2(P)$ and we let $F_X := \{ f : S \rightarrow \mathbb{R} : f(X) \in L^2(P) \}$.)

Definition 5.6 (Conditional Expectation 2). Given a function $X : \Omega \rightarrow S$ and $Y \in L^2(P)$, we define $E[Y | X] := Q_{V_X} Y$ 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, $E[Y | X] = h(X)$ where $h \in F_X$ is chosen so that \[ [Y - h(X)] \perp V_X, \] i.e. $E[Y | X] = h(X)$ iff $(Y - h(X), f(X)) = 0$ for all $f \in F_X$. So in summary, $E[Y | X] = h(X)$ iff \[ E[Y f(X)] = E[h(X) f(X)] \] for all $f \in F_X$. \tag{5.3}

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

Proof. Take $f = 1$ in Eq. (5.3) 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 \( \Omega \) is a finite set with \( N \) elements, \( P \) is the uniform distribution on \( \Omega \), and \( A \) is a non-empty subset of \( \Omega \), 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

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

**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

\[
\mathbb{E}[Y\,f(X)] = \mathbb{E}[h(X)\,f(X)]
\]

for all bounded \( f \).

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

\[
\mathbb{E}[Y\,f(X)] = \sum_{s \in S} \mathbb{E}[Y\,f(X) : X = s] = \sum_{s \in S'} \mathbb{E}[Y\,f(X) : X = s]
\]

\[
= \sum_{s \in S'} f(s) \mathbb{E}[Y|X = s] \cdot P(X = s)
\]

\[
= \sum_{s \in S'} f(s) h(s) \cdot P(X = s)
\]

\[
= \sum_{s \in S'} f(s) h(s) \cdot P(X = s) = \mathbb{E}[h(X)\,f(X)]
\]

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 \( V_X = \text{span}\{1_{X=s} : s \in S\} \). As \( \{1_{X=s} : s \in S\} \) is an orthogonal set, we may compute

\[
\mathbb{E}[Y|X] = \sum_{s \in S} (Y, 1_{X=s}) 1_{X=s} = \sum_{s \in S} \mathbb{E}[Y|X = s] 1_{X=s}
\]

\[
= \sum_{s \in S} \mathbb{E}[Y|X = s] \cdot 1_{X=s} = h(X).
\]

**Example 5.11.** Suppose that \( X \) and \( Y \) are discrete random variables with joint distribution given as:

\[
\begin{align*}
\rho_Y &= \frac{1}{4} \; \frac{1}{2} \; \frac{1}{4} \\
\rho_X \, Y|X &= 1 \; 0 \; 1 \\
1/4 & \; 1 \; 0 \; 1/4 \; 0 \\
3/4 & \; 0 \; 1/4 \; 1/4 \; 1/4
\end{align*}
\]

We then have

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

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

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

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

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

\[
\mathbb{E}[X|Y = 1] = \frac{1}{1/4} \left( 1 \cdot 0 + 0 \cdot \frac{1}{4} \right) = 0.
\]

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

\[
\begin{array}{c|ccc}
\rho_X & .3 & .35 & .35 \\
\rho_Y & .3 \ Y|X & 1 & 2 & 3 \\
& .6 & 1 & .1 & .2 & .3 \\
& .3 & 2 & .15 & .15 & 0 \\
& 1 & 3 & .05 & 0 & .05 \\
\end{array}
\]

Then

\[
\rho_{X|Y}(1,3) = P(X = 1|Y = 3) = \frac{.05}{.1} = \frac{1}{2},
\]

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

\[
\rho_{X|Y}(3,3) = P(X = 3|Y = 3) = \frac{.05}{.1} = \frac{1}{2}.
\]

Therefore,
We compute, for Example 5.14. Suppose that metric distribution, is a girl. We are given \( P \) and so
\[
\begin{align*}
E[X|Y = 3] &= \frac{1}{2} + 2 \cdot 0 + 3 \cdot \frac{1}{2} = 2 \\
h(3) &= E[X|Y = 3] = \frac{1}{2} \cdot (1 - 0.05 + 2 \cdot 0 + 3 \cdot 0.05) = 2 \\
\end{align*}
\]
or
\[
\begin{align*}
h(3) &= E[X|Y = 3] = \frac{1}{2} \cdot (1 - 0.05 + 2 \cdot 0 + 3 \cdot 0.05) = 2 \\
\end{align*}
\]
Similarly,
\[
\begin{align*}
h(1) &= E[X|Y = 1] = \frac{1}{2} \cdot (1 + 2 \cdot 2 + 3 \cdot 2) = \frac{10}{3}, \\
h(2) &= E[X|Y = 2] = \frac{1}{2} \cdot (1 \cdot 15 + 2 \cdot 15 + 3 \cdot 0) = 1.5 \\
\end{align*}
\]
and so
\[
E[X|Y] = h(Y) = \frac{2}{3} \cdot 1_{Y=1} + 1.5 \cdot 1_{Y=2} + 2 \cdot 1_{Y=3}.
\]

Example 5.13 (Number of girls in a family). Suppose the number of children in a family is a random variable \( X \) with mean \( \mu \), 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,
\[
E[G|X = n] = p \cdot n
\]
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 \( E[1_A_i|X = n] = p \) and so
\[
E[G|X = n] = p \cdot n.
\]
Therefore, \( E[G|X] = p \cdot X \) and
\[
\]

Example 5.14. Suppose that \( X \) and \( Y \) are i.i.d. random variables with the geometric distribution,
\[
P(X = k) = P(Y = k) = (1 - p)^{k-1} p \quad \text{for} \quad k \in \mathbb{N}.
\]
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
\[
P(X = m, Y = n - m) = p^2 (1 - p)^{m-1} (1 - p)^{n-m-1}
\]
and
\[
\begin{align*}
\sum_{k+l=n} P(X = k, Y = l) &= \sum_{k+l=n} (1 - p)^{k-1} p (1 - p)^{l-1} p \\
&= \sum_{k+l=n} p^2 (1 - p)^{n-2} = p^2 (1 - p)^{n-2} \sum_{k=1}^{n-1} 1.
\end{align*}
\]
Thus we have shown,
\[
P(X = m|X + Y = n) = \frac{1}{n-1} \quad \text{for} \quad 1 \leq m < n.
\]
From this it follows that
\[
E[f(X)|X + Y = n] = \frac{1}{n-1} \sum_{m=1}^{n-1} f(m)
\]
and so
\[
E[f(X)|X + Y] = \frac{1}{X+Y-1} \sum_{m=1}^{X+Y-1} f(m).
\]
As a check if \( f(m) = m \) we have
\[
\begin{align*}
E[X|X + Y] &= \frac{1}{X+Y-1} \sum_{m=1}^{X+Y-1} m \\
&= \frac{1}{X+Y-1} \left[ \frac{1}{2} (X+Y-1) (X+Y-1+1) \right] \\
&= \frac{1}{2} (X+Y)
\end{align*}
\]
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 $E[Y]$ as $E[E[Y|X]]$.

Clearly if $x \in L \cup W$ then $E[Y|X = x] = 1$ while if $x \in P$, then $E[Y|X = x] = 1 + E_N$, where $N$ 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 $E_N = \frac{1}{p_x}$. For example if $x = 4$, then $p_x = \frac{3+6}{36} = \frac{9}{36}$ (is the number of ways to roll a 4 and 6 is the number of ways to roll as 7) and hence $1 + E_N = 1 + 4 = 5$. Similar calculations gives us the following table:

$$E[Y|X = x] = \begin{cases} 
2, 3, 7, 11, 12 & \{4, 10\} \{5, 9\} \{6, 8\} \\
\frac{12}{36}, \frac{45}{36}, \frac{46}{36}, \frac{47}{36}, \frac{48}{36} 
\end{cases}$$

(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, $E[Y|X = 6] = 47/11$. Similarly for $x = 8$ and $P(X = 6) = 5/36$.) Putting the pieces together and using the law of total expectation gives,

$$E[Y] = E[E[Y|X]] = 1 \cdot \frac{12}{36} + 45 \cdot \frac{1}{36} + 10 \cdot \frac{4}{36} + 47 \cdot \frac{1}{36} \approx 3.376$$

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

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

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

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

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

$$E[u(X_1, \ldots, X_n)|X_1 \in T_1, \ldots, X_n \in T_n] = E[u(Y_1, \ldots, Y_n)]$$

where $Y_i : \Omega \rightarrow T_i$ for $1 \leq i \leq n$ 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,

$$E[u(X_1, \ldots, X_n)|X_1 \in T_1, \ldots, X_n \in T_n] = \frac{\mathbb{E}[u(X_1, \ldots, X_n); X_1 \in T_1, \ldots, X_n \in T_n]}{P(X_1 \in T_1, \ldots, X_n \in T_n)}$$

$$= \frac{1}{P(X_1 \in T_1, \ldots, X_n \in T_n)} \sum_{t_i \in T_i} u(t_1, \ldots, t_n) \prod_i P(X_i = t_i)$$

$$= \sum_{(t_1, \ldots, t_n) \in T_1 \times \cdots \times T_n} u(t_1, \ldots, t_n) \prod_i \frac{P(X_i = t_i)}{P(X_i \in T_i)}$$

$$= \sum_{(t_1, \ldots, t_n) \in T_1 \times \cdots \times T_n} u(t_1, \ldots, t_n) \prod_i P(Y_i = t_i)$$

$$= E[u(Y_1, \ldots, Y_n)]$$

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

$$Y = \sum_{i=1}^n 1_{\{1, 3, 5\}}(X_i) = \text{number of odd rolls}$$

$$Z = \sum_{i=1}^n 1_{\{3, 4, 6\}}(X_i) = \text{number of times 3, 4, or 6 are rolled}$$

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

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

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

The next lemma generalizes this result.

**Lemma 5.19.** Suppose that \( X_i : \Omega \to S \) for \( 1 \leq i \leq n \) 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
\[ \mathbb{E}[Z_A|Z_B] = Z_B \cdot P(X_1 \in A|X_1 \in B) + (n - Z_B) \cdot P(X_1 \in A|X_1 \notin B). \tag{5.4} \]

**Proof.** Intuitively, for a typical trial there are \( Z_B \) of the \( X_i \) in \( B \) and for these \( i \) we have \( \mathbb{P}[1_A(X_i) \mid X_i \in B] = P(X_1 \in A|X_1 \in B) \). Likewise there are \( n - Z_B \) of the \( X_i \) in \( S \setminus B \) and for these \( i \) we have \( \mathbb{P}[1_A(X_i) \mid X_i \notin B] = P(X_1 \in A|X_1 \notin B) \). 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 \( \cup Q_A \) where \( A \) runs through subsets of \( k \) elements of \( S \) and
\[ Q_A = (\cap_{i \in A} \{X_i \in B\}) \cap (\cap_{i \notin A^c} \{X_i \notin B\}). \]

Then according to Proposition 5.17,
\[ \mathbb{E}[Z_A|Q_A] = \mathbb{E} \left[ \sum_{i=1}^{n} 1_A(Y_i) \right] \]
where \( \{Y_i\} \) are independent and
\[ P(Y_i = s) = P(X_i = s|X_i \in B) = P(X_1 = s|X_1 \in B) \]
for \( i \in A \) and
\[ P(Y_i = s) = P(X_i = s|X_i \notin B) = P(X_1 = s|X_1 \notin B) \]
for \( i \notin A \).

Therefore,
\[ \mathbb{E}[Z_A|Q_A] = \mathbb{E} \left[ \sum_{i=1}^{n} 1_A(Y_i) \right] = \sum_{i \in A} P(A \mid X_1 \in B) + \sum_{i \notin A} P(A \mid X_1 \notin B) = m \cdot P(X_1 \in A|X_1 \in B) + (n - m) \cdot P(X_1 \in A|X_1 \notin B). \]

As the result is independent of the choice of \( A \) with \( \#(A) = m \) we may use Proposition 5.10 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 \( 0 \leq m \leq n \) is arbitrary Eq. (5.4) follows.

As a check notice that \( \mathbb{E}[Z_A = n \cdot P(X_1 \in A) \mid A \) while
\[ \mathbb{E}[Z_A|Z_B] = \mathbb{E}[Z_B \cdot P(X_1 \in A|X_1 \in B) + (n - Z_B) \cdot P(X_1 \in A|X_1 \notin B)] = \mathbb{E}[Z_B \cdot P(X_1 \in A)] + (n - \mathbb{E}[Z_B]) \cdot \mathbb{E}[P(X_1 \in A|X_1 \notin B)] = n \cdot P(X_1 \in A) \cdot \mathbb{E}[Z_B] + (n - \mathbb{E}[Z_B]) \cdot \mathbb{E}[P(X_1 \in A|X_1 \notin B)] = n \cdot \mathbb{E}[Z_B] = \mathbb{E}[Z_A]. \]

\[ \square \]

### 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 \( \mathbb{E}[\mathbb{E}[Y|X]] \leq \mathbb{E}[Y] \). Moreover if \( Y \geq 0 \) then \( \mathbb{E}[Y|X] \geq 0 \) (a.s.).

**Proof.** Let \( \mathbb{E}[Y|X] = h(X) \) (with \( h : S \to \mathbb{R} \)) and then define
\[ f(x) = \begin{cases} 1 & \text{if } h(x) \geq 0 \\ -1 & \text{if } h(x) < 0 \end{cases}. \]

Since \( h(x) f(x) = |h(x)| \), it follows from Eq. (5.3) that
for all bounded functions \( f \). Therefore we may conclude that
\[
\mathbb{E}[Yg(X)|X] = (h \cdot g)(X) = h(X)g(X) = g(X)\mathbb{E}(Y|X).
\]

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

5. If \( X \) and \( Y \) are independent and \( \mu := \mathbb{E}[Y] \), then
\[
\mathbb{E}[Yf(X)] = \mathbb{E}[Y]\mathbb{E}[f(X)] = \mu \mathbb{E}[f(X)] = \mathbb{E}[\mu f(X)]
\]
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 \((\tilde{X}, \tilde{Y})\) are random vectors such that \((X, Y) \overset{d}{=} (\tilde{X}, \tilde{Y})\), i.e. \( \mathbb{E}[f(X, Y)] = \mathbb{E}[f(\tilde{X}, \tilde{Y})] \) for all bounded (or non-negative) functions \( f \). If \( h(X) = \mathbb{E}[u(X, Y)|X] \), then \( \mathbb{E}[u(\tilde{X}, \tilde{Y})|\tilde{X}] = h(\tilde{X}) \).

**Proof.** By assumption we know that
\[
\mathbb{E}[u(X, Y)f(X)] = \mathbb{E}[h(X)f(X)] \quad \text{for all bounded } f.
\]
Since \((X, Y) \overset{d}{=} (\tilde{X}, \tilde{Y})\), this is equivalent to
\[
\mathbb{E}[u(\tilde{X}, \tilde{Y})f(\tilde{X})] = \mathbb{E}[h(\tilde{X})f(\tilde{X})] \quad \text{for all bounded } f
\]
which is equivalent to \( \mathbb{E}[u(\tilde{X}, \tilde{Y})|\tilde{X}] = h(\tilde{X}) \). ■

**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,
\[
\mathbb{E}[S_m|S_n] = \frac{m}{n} S_n \quad \text{for all } m \leq n.
\]
for all \( m \leq n \). To prove this first observe by symmetry\footnote{Apply Theorem 5.23 using \((X_1, S_n) \overset{d}{=} (X_i, S_n)\) for \( 1 \leq i \leq n \).} that
\[
\mathbb{E}(X_i|S_n) = h(S_n) \quad \text{independent of } i.
\]

Therefore
\[
S_n = \mathbb{E}(S_n|S_n) = \sum_{i=1}^n \mathbb{E}(X_i|S_n) = \sum_{i=1}^n h(S_n) = n \cdot h(S_n).
\]
Thus we see that

\[ \mathbb{E}(X_i|S_n) = \frac{1}{n} S_n \]

and therefore

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

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

\[ \mathbb{E}(S_m|S_n) = \mathbb{E}(S_n + X_{n+1} + \cdots + X_m|S_n) = \mathbb{E}(S_n|S_n) + \mathbb{E}(X_{n+1}|S_n) + \cdots + \mathbb{E}(X_m|S_n) = S_n + (m - n) \mu \] if \( m \geq n \)

where \( \mu = \mathbb{E}X_1 \).

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

\[ \mathbb{E}[X|Y] = 18 - \frac{3}{5} Y \] and \( \mathbb{E}[Y|X] = 10 - \frac{1}{3} X \).

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

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

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, \( \rho_{(Y,X)}(y,x) \). Then by definition of \( \rho_{(Y,X)} \), we have, for all bounded or non-negative, \( U \), that

\[ \mathbb{E}[U(Y,X)] = \int \int U(y,x) \rho_{(Y,X)}(y,x) \, dy \, dx. \] (5.5)

The marginal density associated to \( X \) is then given by

\[ \rho_X(x) := \int \rho_{(Y,X)}(y,x) \, dy \] (5.6)

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

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

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

\[ \rho_{(Y,X)}(y,x) = \rho_{(Y|X)}(y,x) \rho_X(x) \text{ for all } (x,y). \] (5.8)

Indeed, if \( \rho_X(x) = 0 \), then

\[ 0 = \rho_X(x) = \int \rho_{(Y,X)}(y,x) \, dy \]

from which it follows that \( \rho_{(Y,X)}(y,x) = 0 \) 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,

\[ \rho(x,y) = \rho_{(Y|X)}(y,x) \rho_X(x) \text{ for a.e. } (x,y). \] (5.9)

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

\[ \int \int \rho_{(Y|X)}(y,x) \rho_X(x) \, dxdy = \int \int \rho_{(Y|X)}(y,x) \rho_X(x) 1_{\rho_X(x) > 0} \, dxdy \]

\[ = \int \int \rho(x,y) 1_{\rho_X(x) > 0} \, dxdy \]

\[ = \int \rho_X(x) 1_{\rho_X(x) > 0} \, dx = \int \rho_X(x) \, dx \]

\[ = 1 = \int \int \rho(x,y) \, dxdy, \]

or equivalently,

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

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

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

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

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

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

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

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

Using Lemma 5.26, we find

\[
\mathbb{E}[U(Y) f(X)] = \iint U(y, x) f(x) \rho_{Y|X}(y, x) \, dy \, dx
\]

\[
= \int \int U(y, x) f(x) \rho_{Y|X}(y, x) \rho_Y(x) \, dx \, dy
\]

\[
= \int \left[ \int U(y, x) \rho_{Y|X}(y, x) \, dy \right] f(x) \rho_Y(x) \, dx
\]

\[
= \int h(x) f(x) \rho_Y(x) \, dx
\]

\[
= \mathbb{E}[h(X) f(X)]
\]

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

\[
\rho_Y(y) = \int_0^1 8xy \cdot 1_{0<y<x<1} \, dx = 8y \int_{x=y}^{x=1} x \cdot dx = 8y \cdot \frac{1}{2} y^2 = 4y \cdot (1 - y^2).
\]

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

\[
\mathbb{E}[u(X, Y) | Y = y] = \int_0^1 \frac{2x \cdot 1_{0<y<x<1}}{(1 - y^2)} u(x, y) \, dx = 2 \int_0^1 \frac{1_{0<y<x<1}}{(1 - y^2)} \int_y^1 u(x, y) \, x \, dx.
\]

and so

\[
\mathbb{E}[u(X, Y) | Y] = 2 \int_0^1 \frac{1}{1 - Y^2} \int_Y^1 u(x, Y) \, x \, dx.
\]

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

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

where

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

**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 \( \mathbb{E}[U(Y) | X] = h_0(X) \) where

\[
h_0(x) = \int U(y, x) \rho_{Y|X}(y, x) \, dy
\]

\[
= \int_{\rho_X(x) > 0} U(y, x) \rho_Y(y) \, dy = \int_{\rho_X(x) > 0} \mathbb{E}[U(Y) | X] = \int_{\rho_X(x) > 0} h(x).
\]

If \( f \) is a bounded function of \( x \), then

\[
\mathbb{E}[h_0(X) f(X)] = \int h_0(x) f(x) \rho_X(x) \, dx = \int_{\rho_X(x) > 0} h_0(x) f(x) \rho_X(x) \, dx
\]

\[
= \int_{\rho_X(x) > 0} h(x) \rho_X(x) \, dx = \int h(x) f(x) \rho_X(x) \, dx
\]

\[
= \mathbb{E}[h(X) f(X)].
\]

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

---

\textbf{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

\[
\text{Var}(Y|X) = E \left[ (Y - E[Y|X])^2 | X \right]
\]

and

\[
\text{Var}(Y) = E \left[ \text{Var}(Y|X) \right] + \text{Var}(E[Y|X]).
\]

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

\[
\text{Var}(Y) = E \left[ \text{Var}(Y|X) \right] + \text{Var}(E[Y|X]).
\]

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

\[
E \left[ \text{Var}(Y|X) \right] = E \left[ Y^2 | X \right] - (E \left[ Y | X \right])^2
\]

\[
= EY^2 - E \left[ (E \left[ Y | X \right])^2 \right] = \text{Var}(Y) + (EY)^2 - E \left[ (E \left[ Y | X \right])^2 \right].
\]

The result follows from this identity and the fact that

\[
\text{Var}(E[Y|X]) = E \left[ \text{Var}(E[Y|X]) \right] - (E \left[ E[Y|X] \right])^2 = E \left[ (E \left[ Y | X \right])^2 \right] - (EY)^2.
\]

5.5 Summary on Conditional Expectation Properties

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

1. \( h(X) = E(Y|X) \), i.e. \( h(X) \) is the conditional expectation of \( Y \) given \( X \).
2. \( E(Y - h(X))^2 \leq E(Y - g(X))^2 \) for all functions \( g \), i.e. \( h(X) \) is the best approximation to \( Y \) among functions of \( X \).
3. \( E(Y \cdot g(X)) = E(h(X) \cdot g(X)) \) 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 \( E(Y|X) \) 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 \( E(Y|X) = h(X) \), where

\[
h(x) = E(Y|X = x) = \frac{E(Y \mid X = x)}{P(X = x)} = \frac{1}{P(X = x)} \sum_y y p(y, x)
\]

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 \( p(y, x) \) (i.e. \( P(Y \in dy, X \in dx) = \rho(y, x)dydx \)). Then \( E(Y|X) = h(X) \), where

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

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

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

1. \( E(Y_1 + Y_2|X) = E(Y_1|X) + E(Y_2|X) \).
2. \( E(aY|X) = aE(Y|X) \) for all constants \( a \).
3. \( E(f(Y)Y|X) = f(X)E(Y|X) \) for all functions \( f \).
4. \( E(E(Y|X)) = E(Y) \).
5. If \( Y \) and \( X \) are independent then \( E(Y|X) = E(Y) \).
6. If \( X \geq 0 \) then \( E(Y|X) \geq 0 \).

**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:

\[
E[Y] = \sum_x E(Y|X = x)p_X(x) \quad \text{(Discrete Case)}
\]

where

\[
E(Y|X = x) = \begin{cases} \frac{E(Y \mid X = x)}{P(X = x)} & \text{if } P(X = x) > 0 \\ 0 & \text{otherwise} \end{cases}
\]

\[
E[U(Y, X)] = \int E(U(Y, X) | X = x) \rho_X(x) dx, \quad \text{(Continuous Case)}
\]

where

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

and

\[
\rho_{Y|X}(y, x) = \begin{cases} \frac{\rho(y, x)}{\rho_X(x)} & \text{if } \rho_X(x) > 0 \\ 0 & \text{if } \rho_X(x) = 0 \end{cases}
\]
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
\[
\mathbb{E} [|f (S_N)|] = \mathbb{E} [|Tf (N)|] < \infty.
\]

**Proof.** If \( f \geq 0 \) we have,
\[
\mathbb{E} [f (S_N)] = \sum_{n=0}^{\infty} \mathbb{E} [f (S_N) : S_N = n] = \sum_{n=0}^{\infty} \mathbb{E} [f (S_n) : S_N = n]
\]
\[
= \sum_{n=0}^{\infty} \mathbb{E} [f (S_n)] P (S_N = n) = \sum_{n=0}^{\infty} (Tf) (n) P (S_N = n)
\]
\[
= \mathbb{E} [Tf (N)].
\]

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 \( \sigma^2 = \text{Var} (X_i) \) independent of \( i \). Assuming that \( N \in L^2 (P) \) is independent of the \( \{X_i\} \), then
\[
\mathbb{E} [S_N] = \mu \cdot \mathbb{E} N
\]
and
\[
\text{Var} (S_N) = \sigma^2 \mathbb{E} [N] + \mu^2 \text{Var} (N) \tag{6.1}
\]
\[
\text{Var} (N) = \frac{91}{6} - \left( \frac{7}{2} \right)^2 = \frac{35}{12}.
\]

**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
\[
Tf (n) = \mathbb{E} [S_n^2] = \text{Var} (S_n) + (\mathbb{E} S_n)^2 = \sigma^2 n + (n \cdot \mu)^2,
\]
we find that
\[
\mathbb{E} [S_N^2] = \mathbb{E} [\sigma^2 N + \mu^2 N^2]
\]
\[
= \sigma^2 \mathbb{E} [N] + \mu^2 \mathbb{E} [N^2].
\]

Combining these results shows,
\[
\text{Var} (S_N) = \sigma^2 \mathbb{E} [N] + \mu^2 \mathbb{E} [N^2] - \mu^2 (\mathbb{E} N)^2
\]
\[
= \sigma^2 \mathbb{E} [N] + \mu^2 \text{Var} (N) \tag{6.2}
\]

**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 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 \( i \)th toss and zero otherwise. In this case
\[
\mathbb{E} X_1 = \frac{1}{2},
\]
\[
\text{Var} (X_1) = \frac{1}{2} - \left( \frac{1}{2} \right)^2 = \frac{1}{4},
\]
\[
\mathbb{E} N = \frac{1}{6} (1 + \cdots + 6) = \frac{1}{6} \cdot \frac{1 \cdot 6}{2} = \frac{7}{2},
\]
\[
\mathbb{E} N^2 = \frac{1}{6} (1^2 + 2^2 + 3^2 + 4^2 + 5^2 + 6^2) = \frac{91}{6}
\]
\[
\text{Var} (N) = \frac{91}{6} - \left( \frac{7}{2} \right)^2 = \frac{35}{12}.
\]
Therefore,

\[ E[Z] = E[X_1] \cdot EN = \frac{1}{2} \cdot \frac{7}{2} = \frac{7}{4} \]

\[ \text{Var}(Z) = \frac{1}{4} \cdot \frac{7}{2} + \left( \frac{1}{2} \right)^2 \cdot \frac{35}{12} = \frac{77}{48} = 1.6042. \]

Alternatively, we have

\[ P(Z = k) = \sum_{n=1}^{6} P(Z = k \mid N = n) P(N = n) \]

\[ = \frac{1}{6} \sum_{n=1}^{6} P(Z = k \mid N = n) \]

\[ = \frac{1}{6} \sum_{n=1}^{6} \left( \frac{n}{k} \right) \left( \frac{1}{2} \right)^n. \]

where

\[ E[Z] = \sum_{k=0}^{6} k P(Z = k) = \sum_{k=1}^{6} k P(Z = k) \]

\[ = \sum_{k=1}^{6} \frac{1}{6} \sum_{n=k}^{6} \left( \frac{n}{k} \right) \left( \frac{1}{2} \right)^n \]

\[ = \frac{7}{4} \]

and

\[ E[Z^2] = \sum_{k=0}^{6} k^2 P(Z = k) = \sum_{k=1}^{6} k^2 \sum_{n=k}^{6} \left( \frac{n}{k} \right) \left( \frac{1}{2} \right)^n \]

\[ = \frac{14}{3} \]

so that

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

We have,

\[ P(Z = 0) = \frac{1}{6} \sum_{n=1}^{6} \binom{n}{0} \left( \frac{1}{2} \right)^n = \frac{21}{128} \]

\[ P(Z = 1) = \frac{1}{6} \sum_{n=1}^{6} \binom{n}{1} \left( \frac{1}{2} \right)^n = \frac{5}{16} \]

\[ P(Z = 2) = \frac{1}{6} \sum_{n=2}^{6} \binom{n}{2} \left( \frac{1}{2} \right)^n = \frac{33}{128} \]

\[ P(Z = 3) = \frac{1}{6} \sum_{n=3}^{6} \binom{n}{3} \left( \frac{1}{2} \right)^n = \frac{1}{6} \]

\[ P(Z = 4) = \frac{1}{6} \sum_{n=4}^{6} \binom{n}{4} \left( \frac{1}{2} \right)^n = \frac{29}{384} \]

\[ P(Z = 5) = \frac{1}{6} \sum_{n=5}^{6} \binom{n}{5} \left( \frac{1}{2} \right)^n = \frac{1}{48} \]

\[ P(Z = 6) = \frac{1}{6} \sum_{n=6}^{6} \binom{n}{6} \left( \frac{1}{2} \right)^n = \frac{1}{384}. \]

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

\[ \mathbb{E}[e^{tS_N} \mid N = n] = \mathbb{E}[e^{tX_n} \mid N = n] = \mathbb{E}[e^{tX_n}] = [mgf_{X_n}(t)]^n \]

so that

\[ \mathbb{E}[e^{tS_N} \mid N] = [mgf_{X_n}(t)]^N = e^{N \ln(mgf_{X_n}(t))}. \]

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

\[ mgf_{SN}(t) = mgf_N(\ln(mgf_{X_n}(t))). \]

**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 \( Z \sim X_1 + \cdots + X_N \). We now compute
\[ P(Z = k) = \sum_{n=k}^{M} P(Z = k | N = n) P(N = n) \]
\[ = \sum_{l=0}^{M-k} P(Z = k | N = k + l) P(N = k + l) \]
\[ = \sum_{l=0}^{M-k} P(Z = k | N = k + l) P(N = k + l) \]
\[ = \sum_{l=0}^{M-k} p^k (1 - p)^{k+l} \binom{k+l}{k} \cdot \binom{M}{k+l} (1 - q)^{M-(k+l)} \]
\[ = (pq)^k \sum_{l=0}^{M-k} \frac{1}{l!(M-k-l)!} q^l (1 - q)^{M-k-l} \]
\[ = \binom{M}{k} (pq)^k \sum_{l=0}^{M-k} \frac{(M-k)!}{l!(M-k-l)!} (1 - p)^l q^l (1 - q)^{M-k-l} \]
\[ = \binom{M}{k} (pq)^k \sum_{l=0}^{M-k} (1 - p)^l q^l (1 - q)^{M-k-l} \]
\[ = \binom{M}{k} (pq)^k [(1 - p) q + (1 - q)]^{M-k} \]
\[ = \binom{M}{k} (pq)^k [1 - pq]^{M-k} \]
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 \( \{\xi_1 \eta_i\}_{i=1}^{M} \) 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
\[ \{N = n\} = \bigcup_{i_1 < \cdots < i_n} B(i_1, \ldots, i_n) \]
so that
\[ P(Z = k | N = n) = \frac{\sum_{i_1 < \cdots < i_n} P((Z = k) \cap B(i_1, \ldots, i_n))}{\sum_{i_1 < \cdots < i_n} P(B(i_1, \ldots, i_n))} \]
\[ = \frac{\sum_{i_1 < \cdots < i_n} P(Z = k | B(i_1, \ldots, i_n)) P(B(i_1, \ldots, i_n))}{\sum_{i_1 < \cdots < i_n} P(B(i_1, \ldots, i_n))} \]
\[ = \frac{\sum_{i_1 < \cdots < i_n} \binom{n}{k} q^k (1 - q)^{n-k} P(B(i_1, \ldots, i_n))}{\sum_{i_1 < \cdots < i_n} P(B(i_1, \ldots, i_n))} \]
\[ = \binom{n}{k} q^k (1 - q)^{n-k} \]
and this gives another more intuitive proof of the result.
Part I

Markov Chains
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.

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

A probability on $S$ is a function, $\pi : S \to [0, 1]$ such that $\sum_{x \in S} \pi(x) = 1$. Further, let $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$, $\Omega := S^{\mathbb{N}_0} = \{\omega = (s_0, s_1, \ldots) : s_j \in S\}$, and for each $n \in \mathbb{N}_0$, let $X_n : \Omega \to S$ be given by

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

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

**Definition 7.2 (Markov probabilities).** A (time homogeneous) Markov probability $P$, on $\Omega$ with transition kernel, $p$, is probability on $\Omega$ such that

$$P(X_{n+1} = x_{n+1} | X_0 = x_0, X_1 = x_1, \ldots, X_n = x_n) = P(X_{n+1} = x_{n+1} | X_n = x_n) = p(x_n, x_{n+1}) \quad (7.2)$$

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 $P(X_0 = x_0, X_1 = x_1, \ldots, X_n = x_n) > 0$. (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 $\Omega$ is sufficiently big that it is no longer so easy to give a rigorous definition of a probability on $\Omega$. For the purposes of this class, a probability on $\Omega$ should be taken to mean an assignment, $P(A) \in [0, 1]$ for all subsets, $A \subseteq \Omega$, such that $P(\emptyset) = 0$, $P(\Omega) = 1$, and

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

whenever $A = \bigcup_{n=1}^\infty A_n$ 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,

$$E[f(X_{n+1}) | X_0, X_1, \ldots, X_n] = E[f(X_{n+1}) | X_n] = \sum_{y \in S} p(X_n, y) f(y) \quad (7.3)$$

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 $\{x_j\} \subseteq S$,

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

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 $\{x_j\} \subseteq S$, then $(\{X_n\}, P, p)$ satisfies Definition 7.2.

**Proof.** ($\Rightarrow$) 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

$$P(X_0 = x_0, X_1 = x_1) = P(X_1 = x_1 | X_0 = x_0) P(X_0 = x_0) = \pi(x_0) p(x_0, x_1).$$

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

$$P(X_0 = x_0, X_1 = x_1, X_2 = x_2) = P(X_2 = x_2 | X_0 = x_0, X_1 = x_1) \cdot P(X_0 = x_0, X_1 = x_1) = P(X_2 = x_2 | X_1 = x_1) \cdot P(X_0 = x_0, X_1 = x_1) = p(x_1, x_2) \cdot \pi(x_0) p(x_0, x_1) = \pi(x_0) p(x_0, x_1) p(x_1, x_2).$$

The formal induction argument is now left to the reader.

($\Leftarrow$) If

$$\pi(x_0) p(x_0, x_1) \cdots p(x_{n-1}, x_n) = P(X_0 = x_0, \ldots, X_n = x_n) > 0,$$
then by Eq. (7.4) and the definition of conditional probabilities we find,

\[
P(X_{n+1} = x_{n+1} | X_0 = x_0, X_1 = x_1, \ldots, X_n = x_n) = \frac{P(X_0 = x_0, X_1 = x_1, \ldots, X_n = x_n | X_{n+1} = x_{n+1})}{P(X_0 = x_0, \ldots, X_n = x_n)}
\]

\[
= \frac{\pi(x_0) p(x_0, x_1) \cdots p(x_{n-1}, x_n) p(x_n, x_{n+1})}{\pi(x_0) p(x_0, x_1) \cdots p(x_{n-1}, x_n)} = p(x_n, x_{n+1})
\]

as desired. □

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

**Notation 7.5** We will abbreviate the expectation \((E_{P_{\pi}})\) with respect to \(P_{\pi}\) by \(E_{\pi}\). Moreover if

\[
\pi(y) = \delta_x(y) := \begin{cases} 1 & \text{if } x = y \\ 0 & \text{if } x \neq y \end{cases}
\]

(7.5)

we will write \(P_x\) for \(P_{\delta_x}\) and \(E_x\) for \(E_{\delta_x}\).

For a general probability, \(\pi\), on \(S\), it follows from Proposition 7.3 and Corollary 7.6 that

\[
P_{\pi} = \sum_{x \in S} \pi(x) P_x \quad \text{and} \quad E_{\pi} = \sum_{x \in S} \pi(x) E_x.
\]

(7.6)

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

\[
E_{\pi} [u(X_0, \ldots, X_n)] = \sum_{x_0, \ldots, x_n \in S} u(x_0, \ldots, x_n) \pi(x_0) p(x_0, x_1) \cdots p(x_{n-1}, x_n).
\]

**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

\[
(p \cdot q)(x,y) := \sum_{z \in S} p(x,z) q(z,y).
\]

(7.7)

We also let

\[
p^n := \underbrace{p \cdot p \cdots \cdot p}_{n \times \text{times}}.
\]

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 \(\pi\) 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,

\[
P_x(X_n = y) = \sum_{x \in S} \pi(x) p^n(x, y) = (\pi \cdot p^n)(y).
\]

**Proof.** We have from Eq. (7.4) that

\[
P_x(X_n = y) = \sum_{x_0, \ldots, x_n-1 \in S} P_x(X_0 = x_0, X_1 = x_1, \ldots, X_{n-1} = x_{n-1}, X_n = y)
\]

\[
= \sum_{x_0, \ldots, x_n-1 \in S} \delta_x(x_0) p(x_0, x_1) \cdots p(x_{n-2}, x_{n-1}) p(x_{n-1}, y)
\]

\[
= \sum_{x_1, \ldots, x_{n-1} \in S} p(x, x_1) \cdots p(x_{n-2}, x_{n-1}) p(x_{n-1}, y) = p^n(x, y).
\]

The formula for \(P_x(X_n = y)\) 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](http://www.zweigmedia.com/RealWorld/markov/markov.html) – see Figure 7.1 below.

![Figure 7.1](http://www.zweigmedia.com/RealWorld/markov/markov.html)
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,

\[
P = \begin{bmatrix}
1 & 2 & 3 \\
1/4 & 1/2 & 1/4 \\
1/2 & 0 & 1/2 \\
1/3 & 1/3 & 1/3
\end{bmatrix}
\]

is represented by the jump diagram in Figure 7.2.

![Jump Diagram](image)

**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

\[
P = \begin{bmatrix}
1 & 2 & 3 \\
1/4 & 1/2 & 1/4 \\
1/2 & 0 & 1/2 \\
1/3 & 1/3 & 1/3
\end{bmatrix}
\]

is shown in Figure 7.3.

![Jump Diagram](image)

**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.

*Example 7.12.* Suppose that \( S = \{1, 2, 3\} \), then

\[
P = \begin{bmatrix}
1/2 & 0 & 1/2 \\
1/3 & 1/3 & 1/3
\end{bmatrix}
\]

has the jump graph given by 7.2.

![Jump Diagram](image)

**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,

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

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.

\[
\begin{pmatrix}
1 & 2 & 3 \\
4 & 5 & 6 \\
7
\end{pmatrix}
\]

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 \( 7 \times 7 \) transition matrix, \( q \), with entries given by \( p_{ij} := P(X_{n+1} = j | X_{n} = i) \), 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,

\[
P = \begin{bmatrix}
1 & 2 & 3 & 4 & 5 & 6 & 7 \\
0 & 1/2 & 0 & 1/2 & 0 & 0 & 0 \\
1/3 & 0 & 1/3 & 0 & 1/3 & 0 & 0 \\
0 & 1/2 & 0 & 0 & 0 & 1/2 & 0 \\
1/3 & 0 & 0 & 0 & 1/3 & 0 & 1/3 \\
0 & 1/3 & 0 & 1/3 & 0 & 1/3 & 0 \\
0 & 0 & 1/2 & 0 & 1/2 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0
\end{bmatrix}
\]

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 7.6. 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 \( \pi_i := P(X_0 = i) \).

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

\[
P[X_{n+1} = x_{n+1} | X_0 = x_0, \ldots, X_n = x_n] = P[\alpha(X_n, \xi_{n+1}) = x_{n+1} | X_0 = x_0, \ldots, X_n = x_n]
\]

\[
= P[\alpha(x_n, \xi_{n+1}) = x_{n+1} | X_0 = x_0, \ldots, X_n = x_n]
\]

\[
= P[\alpha(x_n, \xi_{n+1}) = x_{n+1} = p(x_n, x_{n+1}).
\]

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 \( p = 1 - p \). In this case we have

\[
1-a
\]
The jump diagram for such a walk is given in Figure 7.7. This fits into Exam-
ple 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 $P(\xi = -1) = q = 1 - p$.

**Example 7.16 (See III.3.1 of Karlin and Taylor).** Let $\xi_n$ 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 $P(\xi = k) = a_k$ 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 $X_n \leq s$ 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,

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

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

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

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

**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,

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

It is easy to see the following replacement policy.

$$X_0 > S \implies \text{replace with } S \text{ on hand at the beginning of period 1}$$
This can be verified by first principles as well:

\[
P(N = n + k | N > n) = \frac{P(N = n + k)}{P(N > n)} = \frac{p (1 - p)^{n+k-1}}{\sum_{k>n} p (1 - p)^{k-1}}
\]

\[
= \frac{p (1 - p)^{n+k-1}}{\sum_{j=0}^{\infty} p (1 - p)^{n+j}} = \frac{(1 - p)^n \sum_{j=0}^{\infty} (1 - p)^j}{(1 - p)^n + (1 - p)^{n+1}}
\]

\[
= \frac{(1 - p)^{k-1}}{1 - (1 - p)} = (1 - p)^{k-1} = P(N = k).
\]

**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

\[P(Z = k) = \alpha (1 - \alpha)^{k-1}\]

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 \(\beta\) 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 \(\beta \cdot \alpha + (1 - \beta)(1 - \alpha) = 1 - \alpha - \beta\). (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 \((1 - \alpha)\beta\). (If \(k = 0\).) Finally, for \(k \geq 1\), the probability of a \(k \to k - 1\) transition is \(\alpha (1 - \beta)\).

### 7.2 Hitting Times

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

**Definition 7.19 (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

\[\{T_B = n\} = \{ X_0 \notin B, \ldots, X_{n-1} \notin B, X_n \in B \} = \{ X_0 \in A, \ldots, X_{n-1} \in A, X_n \in B \}\]

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 7.20.** 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;

\[\mathbb{E}T = \sum_{m>0} m P(T = m) = \sum_{0<k\leq m<\infty} P(T = m)\]

\[= \sum_{k=1}^{\infty} P(T \geq k) = \sum_{k=0}^{\infty} P(T > k).
\]

**Notation 7.21.** Let \(Q\) be \(P\) restricted to \(A\), i.e. \(Q_{xy} = P_{xy}\) for all \(x, y \in A\). In particular we have

\[Q^N_{x,y} := \sum_{x_1, \ldots, x_{N-1} \in A} Q_{x,x_1}Q_{x_1,x_2} \cdots Q_{x_{N-1},y} \text{ for all } x, y \in A.
\]

**Corollary 7.22.** Continuing the notation introduced above, for any \(x \in A\) we have

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

and
Proposition 7.23. 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 $P_x(T_B = \infty) \leq \alpha$
for all $x \in A$, then $P_x(T_B = \infty) = 0$ for all $x \in A$.

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

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

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

$$P_x(T_B = \infty) = \sum_{y \in A} Q^m_{x,y} P_y(T_B = \infty)$$

$$\leq \alpha \sum_{y \in A} Q^m_{x,y} = \alpha P_x(T_B > n).$$

Finally letting $n \to \infty$ shows $P_x(T_B = \infty) \leq \alpha P_x(T_B = \infty)$, 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
$E_x T_B = \infty$. The next theorem gives a criteria which avoids this scenario.

Theorem 7.24. 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 $P_x(T_B > n) \leq \alpha$
for all $x \in A$, then

$$E_x(T_B = \infty) \leq \frac{n}{1-\alpha} < \infty$$

for all $x \in A$.

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

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

One easily uses this relationship to show inductively that

$$P_x(T_B > kn) \leq \alpha^k \text{ for all } k = 0,1,2,\ldots.$$
First Step Analysis

We assume the \( \{X_n\}_{n=0}^\infty \) is a Markov chain with values in \( S \) and transition kernel \( P \). I will often write \( p(x,y) \) for \( P_{xy} \).

**Theorem 8.1 (Markov Conditioning).** Let \( \pi \) be a probability on \( S \), \( F(X) = F(X_0, X_1, \ldots) \) be a random variable depending on \( X \), and \( x_0, x_1, \ldots, x_n \in S \) be given. Then

\[
\mathbb{E}_\pi \left[ F(X_0, X_1, \ldots) \mid X_0 = x_0, \ldots, X_n = x_n \right] = \mathbb{E}_x \left[ F(x_0, x_1, \ldots, x_{n-1}, X_0, X_1, \ldots) \right].
\]  
(8.1)

whenever \( P(X_0 = x_0, \ldots, X_n = x_n) > 0 \).

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

\[
\mathbb{E}_x \left[ F(X_0, X_1, \ldots, X_N) : X_0 = x_0, \ldots, X_n = x_n \right] = \sum_{x_{n+1}, \ldots, x_N \in S} F(x_0, x_1, \ldots, x_{n+1}, x_n) \frac{\pi(x_0) p(x_0, x_1) \cdots p(x_{n-1}, x_n)}{p(x_n, x_{n+1}) \cdots p(x_N-1, x_N)}
\]

\[
= P(X_0 = x_0, \ldots, X_n = x_n) \sum_{x_{n+1}, \ldots, x_N \in S} F(x_0, x_1, \ldots, x_{n+1}, x_n) \pi(x_0) p(x_0, x_1) \cdots p(x_{n-1}, x_n) \cdot \frac{1}{p(x_n, x_{n+1}) \cdots p(x_N-1, x_N)}
\]

\[
= P(X_0 = x_0, \ldots, X_n = x_n) \mathbb{E}_{x_n} \left[ F(x_0, x_1, \ldots, x_{n-1}, X_0, X_1, \ldots, X_N-n) \right]
\]

from which Eq. (8.1) follows. \( \blacksquare \)

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

\[
\mathbb{E}_x \left[ F(X_0, X_1, \ldots) \mid X_1 = y \right] = \mathbb{E}_y \left[ F(x, X_0, X_1, \ldots) \right]
\]  
(8.2)

and

\[\text{Remark 8.4. To construct the matrix } Q \text{ and } R \text{ from } P, \text{ let } P' \text{ be } P \text{ with the rows corresponding to } B \text{ omitted. To form } Q \text{ from } P', \text{ remove the columns of } P' \text{ corresponding to } B \text{ and to form } R \text{ from } P', \text{ remove the columns of } P' \text{ corresponding to } A.\]

**Example 8.5.** If \( S = \{1, 2, 3, 4, 5, 6, 7\} \), \( A = \{1, 2, 4, 5, 6\} \), \( B = \{3, 7\} \), and

\[
P = \begin{bmatrix}
1 & 2 & 3 & 4 & 5 & 6 & 7 \\
0 & 1/2 & 0 & 1/2 & 0 & 0 & 0 \\
1/3 & 0 & 1/3 & 0 & 1/3 & 0 & 0 \\
0 & 1/2 & 0 & 0 & 0 & 1/2 & 0 \\
1/3 & 0 & 0 & 0 & 1/3 & 0 & 1/3 \\
0 & 1/3 & 0 & 1/3 & 0 & 1/3 & 0 \\
0 & 0 & 1/2 & 0 & 1/2 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1
\end{bmatrix} 
\]

\[
\mathbb{E}_x \left[ F(X_0, X_1, \ldots) \mid X_1 = y \right] = \mathbb{E}_y \left[ F(x, X_0, X_1, \ldots) \right]
\]  
(8.3)
Corollary 8.6. Given a bounded or non-negative functions \( h : B \to \mathbb{R} \), let 

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

Then \( u : A \to \mathbb{R} \) satisfies Eq. (8.26),

\[
    u(x) = \sum_{y \in A} p(x, y) u(y) + \sum_{y \in B} p(x, y) h(y) \quad \text{for all } x \in A
\]

which we abbreviate by

\[
    u = Qu + Rh \implies u = (I - Q)^{-1} Rh.
\]

In particular, when \( h \equiv 1 \), \( u(x) = P_x (T_B < \infty) \) is a solution to the equation,

\[
    u = Qu + R1_B \quad \text{on } A.
\]

Or if \( b \in B \) and \( h = \delta_b \), then \( u(x) = P_x (X_{T_B} = b : T_B < \infty) \) is the probability that the chain first hits \( B \) at site \( b \) and this \( u \) satisfies,

\[
    u(x) = \sum_{y \in A} p(x, y) u(y) + p(x, b) \quad \text{for all } x \in A.
\]

More generally if \( B_0 \) is a subset of \( B \) and \( h = 1_{B_0} \) then \( u(x) = P_x (X_{T_B} \in B_0 : T_B < \infty) \) is the probability that the chain first enters \( B \) inside of \( B_0 \) and this function \( u(x) \) must satisfy,

\[
    u(x) = \sum_{y \in A} p(x, y) u(y) + \sum_{b \in B} p(x, b),
\]

i.e.

\[
    u = Qu + R1_{B_0} \quad \text{on } A.
\]

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

\[
    F(x_0, X_1, \ldots) = h(X_{T_B}(X)) = h(X_{T_B}(X)) 1_{T_B(X) < \infty},
\]

then for \( x \in A \) we have \( F(x, X_0, X_1, \ldots) = F(X_0, X_1, \ldots) \). Therefore by the first step analysis (Theorem 8.2) we learn

\[
    u(x) = \mathbb{E}_x h(X_{T_B}(X)) = \mathbb{E}_x F(x, X_1, \ldots) = \sum_{y \in S} p(x, y) \mathbb{E}_y F(x, X_0, X_1, \ldots) = \sum_{y \in S} p(x, y) \mathbb{E}_y h(X_{T_B}(X)) = \sum_{y \in A} p(x, y) \mathbb{E}_y h(X_{T_B}(X)) + \sum_{y \in B} p(x, y) h(y)\]

\[
= \sum_{y \in A} p(x, y) u(y) + \sum_{y \in B} p(x, y) h(y) = (P_A u)(x) + (P_{1_B})(x).
\]

\[
\]

**Corollary 8.7.** Given \( g : A \to [0, \infty] \), let \( u(x) := \mathbb{E}_x \left[ \sum_{n<T_B} g(X_n) \right] \). Then \( u(x) \) satisfies

\[
    u(x) = \sum_{y \in A} p(x, y) u(y) + g(x) \quad \text{for all } x \in A
\]

which we abbreviate as

\[
    u = Qu + g \implies u = (I - Q)^{-1} g.
\]

In particular if we take \( g \equiv 1 \) in this equation we learn that
\[
\mathbb{E}_x T_B = \sum_{y \in A} p(x,y)\mathbb{E}_y T_B + 1 \text{ for all } x \in A,
\]
that is
\[
\mathbb{E}_x T_B = \left[ (I - Q)^{-1} 1_A \right]_x = \sum_{y \in A} (I - Q)_{x,y}^{-1}.
\]

**Proof.** Let \( F(x, x_0, x_1, \ldots) = \sum_{n<T_B(x_0, x_1, \ldots)} g(X_n) \) 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
\[
F(x, x_0, x_1, \ldots) = \left\{ \begin{array}{ll} g(x) & \text{if } x_0 \in B \\
g(x) + F(x, x_0, x_1, \ldots) & \text{if } x_0 \in A 
\end{array} \right.
\]
\[
= g(x) + \sum_{x_0 \in A} (1 - \mathbb{P}(x_0 \notin A)) F(x, x_0, x_1, \ldots).
\]
Therefore by the first step analysis (Theorem 8.2) it follows that
\[
u(x) = \mathbb{E}_x F(x, x_0, x_1, \ldots) = \sum_{y \in S} p(x,y)\mathbb{E}_y F(x, x_0, x_1, \ldots)
\]
\[
= \sum_{y \in S} p(x,y)\mathbb{E}_y [g(x) + 1_{x_0 \in A} \cdot F(x, x_0, x_1, \ldots)]
\]
\[
= g(x) + \sum_{y \in A} p(x,y)\mathbb{E}_y [F(x, x_0, x_1, \ldots)]
\]
\[
= g(x) + \sum_{y \in A} p(x,y)\mathbb{E}_y [F(x_0, x_1, \ldots)]
\]
\[
= g(x) + \sum_{y \in A} p(x,y)\mathbb{E}_y [F(x_0, x_1, \ldots)].
\]

\[ \quad \]

### 8.1 Finite state space chains

In this subsection I would like to write out the above theorems in the special case where \( S \) is a finite set. Let \( Q \) and \( R \) be as described in Notation 8.3.

**Theorem 8.8.** Let us continue to use the notation and assumptions as described above. If \( h : B \to \mathbb{R} \) and \( g : A \to \mathbb{R} \) are given functions, then for all \( x \in A \) we have:
\[
\mathbb{E}_x [h(X_{T_B})] = \left[ (I - Q)^{-1} R h \right](x) \quad \text{and} \quad \mathbb{E}_x \left[ \sum_{n<T_B} g(X_n) \right] = \left[ (I - Q)^{-1} g \right](x).
\]

**Remark 8.9.** Here is a story to go along with the above scenario. Suppose that \( g(x) \) is the toll you have to pay for visiting a site \( x \in A \) while \( h(y) \) is the amount of prize money you get when landing on a point in \( B \). Then \( \mathbb{E}_x \left[ \sum_{0 \leq n < T} g(X_n) \right] \) is the expected toll you have to pay before your first exit from \( A \) while \( \mathbb{E}_x [h(X_T)] \) is your expected winnings upon exiting \( B \).

Here are some typical choices for \( h \) and \( g \).

1. If \( y \in B \) and \( h = \delta_y \), then
\[
P_x (X_{T_B} = y) = \left[ (I - Q)^{-1} R \delta_y \right](x) = \left[ (I - Q)^{-1} R \right]_{x,y}.
\]
2. If \( y \in A \) and \( g = \delta_y \), then
\[
\sum_{n<T_B} g(X_n) = \sum_{n<T_B} \delta_y(X_n) = \# \text{ visits to before hitting } B
\]
and hence
\[
\mathbb{E}_x \left( \# \text{ visits to } y \text{ before hitting } B \right) = \left[ (I - Q)^{-1} \delta_y \right](x) = (I - Q)_{x,y}^{-1}.
\]
3. If \( g = 1 \), i.e. \( g(y) = 1 \) for all \( y \in A \), then \( \sum_{n<T_B} g(X_n) = T_B \) and we find,
\[
\mathbb{E}_x T_B = \left[ (I - Q)^{-1} 1 \right]_x = \sum_{y \in A} (I - Q)_{x,y}^{-1},
\]
where \( \mathbb{E}_x T_B \) is the expected hitting time of \( B \) when starting from \( x \). (Notice that by item 2., \( \sum_{y \in A} (I - Q)_{x,y}^{-1} \) is the expected number of visits to sites in \( A \) before hitting \( B \) which is precisely \( \mathbb{E}_x T_B \) as we have already seen.)

**Example 8.10.** Consider the Markov chain determined by
\[
P = \begin{bmatrix}
1 & 2 & 3 & 4 \\
0 & 1/3 & 1/3 & 1/3 \\
3/4 & 1/8 & 1/8 & 0 \\
0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1
\end{bmatrix}
\]

Notice that 3 and 4 are absorbing states. Let \( h_i = P_i (X_n \text{ hits } 3) \) for \( i = 1, 2, 3, 4 \). Clearly \( h_3 = 1 \) while \( h_1 = 0 \) and by the first step analysis we have
\[
h_1 = 1/3 h_2 + 1/3 h_3 + 1/3 h_4 = 1/3 h_2 + 1/3
\]
\[
h_2 = 3/4 h_1 + 1/8 h_2 + 1/8 h_3 = 3/4 h_1 + 1/8 h_2 + 1/8
\]
so that we have

\[ h_1 = \frac{1}{3} h_2 + \frac{1}{3} \]
\[ h_2 = \frac{3}{4} h_1 + \frac{1}{8} h_2 + \frac{1}{8} \]

which have solutions,

\[ P_1(X_n \text{ hits 3}) = h_1 = \frac{8}{15} \approx 0.5333 \]
\[ P_2(X_n \text{ hits 3}) = h_2 = \frac{3}{5} \]

Similarly if we let \( h_i = P_i(X_n \text{ hits 4}) \) instead, from the above equations with \( h_3 = 0 \) and \( h_4 = 1 \), we find

\[ h_1 = \frac{1}{3} h_2 + \frac{1}{3} \]
\[ h_2 = \frac{3}{4} h_1 + \frac{1}{8} h_2 \]

which has solutions,

\[ P_1(X_n \text{ hits 4}) = h_1 = \frac{7}{15} \text{ and } P_2(X_n \text{ hits 4}) = h_2 = \frac{2}{5}. \]

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

\[ P_1(X_n \text{ hits 3}) + P_1(X_n \text{ hits 4}) = 1 \text{ and } P_2(X_n \text{ hits 3}) + P_2(X_n \text{ hits 4}) = 1. \]

We can do these computations using the matrix formalism as well. For this we have

\[
P' = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 0 & 1/3 & 1/3 & 1/3 \\ 3/4 & 1/8 & 1/8 & 0 \end{bmatrix}, \quad Q = \begin{bmatrix} 0 & 1/3 \\ 3/4 & 1/8 \end{bmatrix}, \quad \text{and } R = \begin{bmatrix} 1/3 & 1/3 \\ 1/8 & 0 \end{bmatrix},
\]

so that

\[
(I - Q)^{-1} = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1/3 & 1 \end{bmatrix}, \\
(1 - Q)^{-1} P = \begin{bmatrix} 1/3 & 1 \\ 1/8 & 0 \end{bmatrix}.
\]

so that

\[
E_i(\# \text{ visits to } j \text{ before hitting } B) = \frac{\mathbb{I}_{ij}}{2} \begin{bmatrix} 2 & 5 \\ 3 & 8 \end{bmatrix},
\]

\[
P_i(X_{T(3,4)} = j) = \frac{\mathbb{I}_{ij}}{2} \begin{bmatrix} 7 & 3 \\ 8 & 2 \end{bmatrix} \approx \begin{bmatrix} 0.5333 & 0.4667 \\ 0.6 & 0.4 \end{bmatrix},
\]

\[
E_i T_{(3,4)} = \frac{i}{2} \begin{bmatrix} 7 & 8 \\ 5 & 5 \end{bmatrix} = \begin{bmatrix} 7 & 8 \\ 5 & 5 \end{bmatrix}.
\]

The output of one simulation from [www.zweigmedia.com/RealWorld/markov/markov.html](http://www.zweigmedia.com/RealWorld/markov/markov.html) is in Figure 8.1 below.

**Example 8.11.** 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.

\[
\begin{bmatrix} 1 & 2 & 3 \text{ (food)} \\ 4 & 5 & 6 \end{bmatrix}.
\]

Recall that the transition matrix for this chain with sites 3 and 7 absorbing is given by,

\[
P = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 0 & 1/2 & 0 & 1/2 & 0 & 0 & 0 \\ 1/3 & 0 & 1/3 & 0 & 1/3 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1/3 & 0 & 0 & 0 & 1/3 & 0 & 1/3 \\ 0 & 1/3 & 0 & 1/3 & 0 & 1/3 & 0 \\ 0 & 0 & 1/2 & 0 & 1/2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}.
\]
8.1 Finite state space chains

Fig. 8.1. 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 \( \frac{171}{171 + 154} = 0.52615 \) which is a little lower than the predicted value of 0.533. You can try your own runs of this simulator.

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 \( P \) leaves the matrix,

\[
P' = \begin{bmatrix}
0 & 1/2 & 0 & 1/2 & 0 & 0 & 0 & 1 & 2 & 4 & 5 & 6 \\
1/3 & 0 & 0 & 0 & 1/3 & 0 & 2 \\
0 & 1/3 & 0 & 1/3 & 0 & 1/3 & 4 \\
0 & 0 & 1/2 & 0 & 1/2 & 0 & 6
\end{bmatrix}
\]

Deleting the 3 and 7 columns of \( P' \) gives

\[
Q = P_{A,A} = \begin{bmatrix}
1 & 2 & 4 & 5 & 6 \\
0 & 1/2 & 1/2 & 0 & 0 & 0 \\
1/3 & 0 & 0 & 1/3 & 0 & 2 \\
1/3 & 0 & 0 & 1/3 & 0 & 4 \\
0 & 1/3 & 1/3 & 0 & 1/3 & 5 \\
0 & 0 & 0 & 1/2 & 0 & 6
\end{bmatrix}
\]

and deleting the 1, 2, 4, 5, and 6 columns of \( P' \) gives

\[
R = P_{A,B} = \begin{bmatrix}
3 & 7 \\
0 & 0 & 1 & 3 & 4 \\
0 & 0 & 5 \\
1/2 & 0 & 6
\end{bmatrix}
\]

Therefore,

\[
I - Q = \begin{bmatrix}
1 & -1/2 & 1/2 & 0 & 0 \\
-1/3 & 0 & 1 & -1/2 & 0 \\
0 & -1/3 & 1/3 & 1 & -1/3 \\
0 & 0 & 0 & -1/2 & 1
\end{bmatrix}
\]

and using a computer algebra package we find

\[
(I - Q)^{-1} = \begin{bmatrix}
1 & 2 & 4 & 5 & 6 \\
0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 1 & 1/3 \\
0 & 0 & 1 & 2 & 1/3 \\
0 & 0 & 0 & 0 & 1/3
\end{bmatrix}
\]

In particular we may conclude,

\[
\begin{bmatrix}
E_1T \\
E_2T \\
E_3T \\
E_4T \\
E_5T \\
E_6T
\end{bmatrix} = (I - Q)^{-1} 1 = \begin{bmatrix}
17 \\
14 \\
14 \\
15 \\
15 \\
13
\end{bmatrix}
\]

and see Figure S.2 for the corresponding jump diagram for this chain.
where

\[
\begin{bmatrix}
P_1 (X_T = 3) \\
P_2 (X_T = 3) \\
P_3 (X_T = 3) \\
P_4 (X_T = 3) \\
P_5 (X_T = 3) \\
P_6 (X_T = 3)
\end{bmatrix} = (I - Q)^{-1} \mathbf{R} =
\begin{bmatrix}
7/12 \\
5/12 \\
1/6 \\
1/6 \\
1/6 \\
1/6
\end{bmatrix}.
\]

Since the event of hitting 3 before 7 is the same as the event \( \{ X_T = 3 \} \), the desired hitting probabilities are

\[
\begin{bmatrix}
P_1 (X_T = 3) \\
P_2 (X_T = 3) \\
P_3 (X_T = 3) \\
P_4 (X_T = 3) \\
P_5 (X_T = 3) \\
P_6 (X_T = 3)
\end{bmatrix} = \begin{bmatrix}
7/12 \\
5/12 \\
1/6 \\
1/6 \\
1/6 \\
1/6
\end{bmatrix}.
\]

We can also derive these hitting probabilities from scratch using the first step analysis. In order to do this let

\[ h_i = P_i (X_T = 3) = P_i (X_n \text{ hits } 3 \text{ (food) before } 7 \text{ (trapped)}) . \]

By the first step analysis we will have,

\[
h_i = \sum_j P_1 (X_T = 3 | X_1 = j) P_i (X_1 = j) = \sum_j p(i, j) P_1 (X_T = 3 | X_1 = j) = \sum_j p(i, j) P_j (X_T = 3) = \sum_j p(i, j) h_j
\]

where \( h_3 = 1 \) and \( h_7 = 0 \). Looking at the jump diagram (Figure 8.2) 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 then solve to find,

\[
\begin{bmatrix}
k_1 \\
k_2 \\
k_4 \\
k_5 \\
k_6
\end{bmatrix} = \begin{bmatrix}
5/12 \\
1/4 \\
7/12 \\
1/3 \\
1/6
\end{bmatrix}.
\]

Notice that the sum of the hitting probabilities in Eqs. (8.7) and (8.8) add up to 1 as they should.
Example 8.12 (A modified rat maze). Here is the modified maze,

\[
\begin{bmatrix}
1 & 2 & 3 & 4 & 5 & 6 \\
0 & 1/2 & 0 & 1/2 & 0 & 0 \\
1/3 & 0 & 1/3 & 0 & 1/3 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 \\
1/3 & 0 & 0 & 0 & 1/3 & 1/3 \\
0 & 1/2 & 0 & 1/2 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1
\end{bmatrix}.
\]

The transition matrix with 3 and 6 made into absorbing states\(^2\) is:

\[
P =
\begin{bmatrix}
1 & 2 & 3 & 4 & 5 & 6 \\
0 & 1/2 & 0 & 1/2 & 0 & 0 \\
1/3 & 0 & 1/3 & 0 & 1/3 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 \\
1/3 & 0 & 0 & 0 & 1/3 & 1/3 \\
0 & 1/2 & 0 & 1/2 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1
\end{bmatrix},
\]

\[
Q =
\begin{bmatrix}
1 & 2 & 4 & 5 \\
0 & 1/2 & 1/2 & 0 \\
1/3 & 0 & 0 & 1/3 \\
0 & 1/2 & 1/2 & 0 \\
1/3 & 0 & 0 & 1/3 \\
0 & 1/2 & 1/2 & 0
\end{bmatrix},
\]

\[
R =
\begin{bmatrix}
3 & 6 \\
0 & 0 & 1 \\
1/3 & 0 & 2 \\
1/3 & 0 & 4 \\
0 & 1/3 & 4 \\
0 & 0 & 5
\end{bmatrix}.
\]

\[
(I_4 - Q)^{-1} =
\begin{bmatrix}
1 & 2 & 4 & 5 \\
2 & 3 & 3 & 1 \\
1 & 2 & 1 & 1 \\
1 & 1 & 2 & 1 \\
3 & 3 & 3 & 2 \\
5 & 5 & 5 & 2
\end{bmatrix},
\]

\[
(I_4 - Q)^{-1} R =
\begin{bmatrix}
1 & 2 & 4 & 5 \\
1/2 & 3/4 & 1/2 & 1/4 \\
1/2 & 3/4 & 1/2 & 1/4 \\
1/2 & 3/4 & 1/2 & 1/4 \\
1/2 & 3/4 & 1/2 & 1/4 \\
1/2 & 3/4 & 1/2 & 1/4
\end{bmatrix},
\]

\[
(1/4 - Q)^{-1} =
\begin{bmatrix}
1 & 1 & 1 \\
1 & 1 & 1 \\
1 & 1 & 1 \\
1 & 1 & 1 \\
1 & 1 & 1 \\
1 & 1 & 1
\end{bmatrix}.
\]

\[
(1/4 - Q)^{-1} =
\begin{bmatrix}
6 & 1 \\
5 & 2 \\
5 & 4 \\
6 & 5
\end{bmatrix}.
\]

So for example, 

\[
P(4(X_T = 3(\text{food}))) = 1/3, \quad E_4(\text{Number of visits to 1}) = 1,
\]

\[
E_5(\text{Number of visits to 2}) = 3/2 \text{ and } E_1T = E_5T = 6 \text{ and } E_2T = E_4T = 5.
\]

Therefore,

\[
h_6 = 1/2 (1 + h_5)
\]

\[
h_5 = 1/3 (h_2 + h_4 + h_6)
\]

\[
h_4 = 1/2 h_1
\]

\[
h_2 = 1/3 (1 + h_1 + h_5)
\]

\[
h_1 = 1/2 (h_2 + h_4).
\]

The solutions to these equations are,

\[
h_1 = 4/9, \quad h_2 = 2/3, \quad h_4 = 2/9, \quad h_5 = 5/9, \quad h_6 = 7/9.
\]  \hspace{2em} (8.9)

Similarly if \(h_i = P_i (X_n \text{ hits 7 before 3})\) we have \(h_7 = 1, \quad h_3 = 0\) and

\[
h_6 = 1/2 h_5
\]

\[
h_5 = 1/3 (h_2 + h_4 + h_6)
\]

\[
h_4 = 1/2 (h_1 + 1)
\]

\[
h_2 = 1/3 (h_1 + h_3)
\]

\[
h_1 = 1/2 (h_2 + h_4)
\]

whose solutions are

\[
h_1 = 5/9, \quad h_2 = 1/3, \quad h_4 = 7/9, \quad h_5 = 4/9, \quad h_6 = 2/9.
\]  \hspace{2em} (8.10)

Notice that the sum of the hitting probabilities in Eqs. (8.9) and (8.10) add up to 1 as they should.

8.2 First Return Times

Definition 8.13 (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 \(\infty\).
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 \( x \).

**Exercise 8.1.** Let \( x \in X \). Show;

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

\[ P_x (R_x > n + 1) \leq \sum_{y \neq x} p(x, y) P_y (T_x > n). \] \hspace{1cm} (8.11)

b Use Eq. (8.11) 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. (8.11) on \( n \in \mathbb{N}_0 \) to show

\[ \mathbb{E}_x [R_x] \leq P_x (R_x > 0) + \sum_{y \neq x} p(x, y) \mathbb{E}_y [T_x]. \] \hspace{1cm} (8.12)

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. (8.12) combined with Corollary 7.25 shows that \( \mathbb{E}_x [R_x] < \infty \).
References