A Mathematical Introduction to Markov Chains[1]

Martin V. Day[2]

May 13, 2018

[2]day@math.vt.edu

Contents

Preface

I have taught various undergraduate courses on Markov chains and stochastic processes at Virginia Tech over the years. In every instance I used a different published text and was always disappointed in their emphases and coverage. So I began to write short sections of notes to bring out the ideas and organize the material the way I thought appropriate. After my retirement in 2016 I decided to put those notes together, fill in some of the gaps, and try to turn them into a coherent treatment. This document is the current status of that effort.

So far I have organized the material in an order that makes logical sense to me and developed proofs based on that organization. Here are some of the ideas that have guided my selection and organization of the material.

- This is a topic in *mathematics*. Although Markov chains are used in many applications, and specific applications help to illustrate the ideas, I want the mathematics of Markov chains to be the focus. Students should see topics from their previous mathematics courses at work here: linear algebra, infinite series, elementary real analysis and differential equations.

- There are some ideas that unify the different Markov processes considered here. One is the central role of the *generator*. For discrete time chains this is the matrix $\mathbf{A} = \mathbf{P} - \mathbf{I}$. For a countably infinite state space this is an infinite matrix (viewed an operator on sequences). For Markov jump processes this is the operator $\mathcal{A}$ of (11.18). And for Brownian motion it is $\mathcal{A} = \frac{1}{2}\frac{\partial^2}{\partial x^2}$. One of the common roles it plays is to identify families of martingales:

$$M_n = f(X_n) - \sum_0^{n-1} \mathbf{A} f(X_k), \quad \text{or } M_t = f(X_t) - \int_0^t \mathcal{A} f(X_s)\, ds$$

  as the case may be. Although I won't try to develop semigroup theory per se, or the full martingale problem equivalence, I do hope students will recognize that there are some central ideas that are common to the different types of Markov processes discussed and which hold the whole business together conceptually.

- Many problems about Markov processes can be reduced to solving a system of equations for functions of the state variable which involve $\mathbf{A}$ or $\mathcal{A}$. Calculation of hitting probabilities, mean hitting times, determining recurrence vs. transience, and explosion vs. non-explosion, are all considered in this way.

- One of my disappointments with some published texts is the reliance on a merely intuitive understanding of the Markov property and conditional expectations. Although the measure-theoretic foundations of such things cannot be developed at the undergraduate level, I want students to recognize that there is some underlying mathematical structure which makes those parts of our reasoning rigorous, even if we don't always draw it out fully. So I try to explain in Chapter 3 how the whole conception of probability theory is founded on the Kolmogorov model of an underlying probability space with random variables as (measurable) functions. I present in that chapter some important working tools (dominated and monotone convergence theorems, the strong law, ...) but with no attempt to prove them. In the chapters which follow I will use those tools but will sweep some measure-theoretic technicalities under the rug (such qualifications that functions be measurable or "except on a set of probability 0"). I readily acknowledge that as a consequence my treatment falls short of the level of rigor expected in

advanced texts. Even so, I have tried to make the treatment mathematically honest in that although a few deeper results are presented without proof, their statements are correct. With regard to conditional probabilities, instead of introducing sigma-algebras, as a rigorous discussion would require, I use the idea of functional dependence. Where a rigorous treatment would say that "Y is measurable with respect to the sigma-algebra $\mathcal{F}_n$ generated by $X_0, \ldots X_n$", I say that "Y is $X_{0:n}$-determined", meaning that $Y = \phi(X_{0:n})$ for some function $\phi(\cdot)$. This formulation allows the essential features of conditionals to be presented in a way comprehensible to undergraduates but without all the measure-theoretic machinery necessary for complete rigor.

- Contemporary software makes it possible to carry out calculations and perform simulations to exhibit various properties. Using MATLAB for instance students can easily perform Markov chain calculations which are larger than they could attempt by hand. Facility with such calculations should be part of a contemporary mathematics education, so I want this text to both presume and cultivate those skills. The m-files referred in the text to are available from a web page accompanying this text.

But now some candid acknowledgment of what this document is *not*. I hesitate to call it a "book" for several reasons. There are additional topics which ought to be included, some of which appear as brown-colored text and some chapters which need to be expanded and supplied with more problems. Many ideas could use better introduction. The use of MATLAB is rather thin in the latter chapters. The difficulty level is uneven and probably strays beyond what most undergraduates can handle in places. Surely there are inconsistencies in my notation. There are also certain to be many typos, misspellings, even mathematically incorrect statements (though I hope not many) which further editing would improve[1]. Whether I will eventually improve all these things only time will tell. I hope that what I have written out so far at least provides an organization and development of the material that others may find useful in some way, even if it is not quite suitable for a course text (yet).

Prerequisites, in addition to the standard freshman-sophomore calculus and differential equations courses, would be a real analysis or advanced calculus course which covers the connections between continuity and sequential convergence, the properties of infinite series and power series, a linear algebra course which includes the study of diagonalization of matricies and eigenvalues, and (for Chapter 11) a course on differential equations which includes the matrix exponential. Although not strictly necessary it would also be helpful if the student has encountered some basic ideas about random variables previously. An appendix provides a brief synopsis of some supporting mathematical topics that may have escaped students' backgrounds.

Martin Day; Lynchburg, VA, Dec. 2017

[1] Please feel free to point out any failings or suggestions by e-mail if you wish.

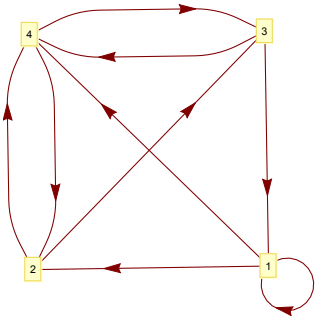
Chapter 1

Introduction and Preliminaries

A *stochastic process* is a mathematical quantity which is both time-dependent and random. We typically denote it as something like X_t where the time variable t is written as a subscript. We will call the value of X_t its *state* at time t. The possible states might be integers, real numbers, vectors, or possibly other types of quantities. There are different types of stochastic processes, depending on what is assumed about the relation between X_s and X_t for different times s and t. Our main interest will be in *Markov processes*. The essential feature of these is that if we know the current state, X_s, then the probabilistic description of future states, X_t for $t > s$, does not involve any additional information about the past: X_u for $u < s$. In brief, future behavior depends on the present but not past history. This property allows Markov processes to be studied in terms of their time evolution and allows a rather rich mathematical analysis. Central to that analysis is a matrix $\mathbf{A}$ (or operator $\mathcal{A}$ in Chapters 11 and 12) which shows up over and over in the equations describing the process. This is called the *generator*, or infinitesimal generator, or characteristic operator in different settings. Our presentation is organized to highlight the several roles that $\mathbf{A}$ (or $\mathcal{A}$) plays in the analysis of the Markov processes we consider.

1.1 Overview and Examples

We begin here with a preview of some of the types of Markov processes we will be considering. The simplest are finite state Markov chains. Time is limited to integer values: $t = 0, 1, 2, \ldots, n, \ldots$. There is a finite set $\mathcal{S}$ of possible states. At time $t = n$ the chain is located at one these states $X_n = i \in \mathcal{S}$ and then jumps to another state for the next time, say $X_{n+1} = j$. A collection of transition probabilities $p_{i,j}$ describe how likely each of these $X_n = i \to X_{n+1} = j$ transitions is. A simple example is given in Example 2.1 on page 8, illustrated in the graph below. The verticies are the possible states $\mathcal{S} = \{1, 2, 3, 4\}$ and values for all the $p_{i,j}$ are given in the matrix $\mathbf{P}$ on page 8. Each arrow $i \to j$ in the graph indicates a possible transition ("possible" meaning positive probability: $p_{i,j} > 0$).



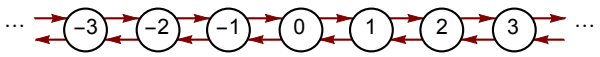
The chain X_n itself moves from one state to another (according to the prescribed $p_{i,j}$) as time progresses.

Chapter 2 considers finite state Markov chains in general, and how we can determine the probabilities of various behaviors for them. We will find that these problems reduce to matrix equations all involving the matrix (generator)

$$\mathbf{A} = \mathbf{P} - \mathbf{I},$$

where $\mathbf{P} = [p_{i,j}]$ is the matrix of transition probabiities.

We can also have Markov chains with an infinite set of states. Although much of what we will learn in Chapter 2 carries over to the infinite state setting, the analysis becomes more difficult. In preparation for that Chapter 3 summarizes some fundamental features of probabilities, including conditional probabilities, that are vital to understanding stochastic processes in all but the simplest settings. Chapter 4 then resumes our exploration of Markov chains in the case of infinitely many states. For an example consider the symmetric random walk in one dimension. The states are the integers. At each step the chain moves either up or down by one $X_{n+1} = X_n \pm 1$, each with probability $\frac{1}{2}$. Thus X_n "walks" back and forth through the integers.



This seems simple enough. But some questions about it are not so easy to answer. If we assume the chain starts at $X_0 = 0$ can we determine whether or not it will return to 0 sometime in the future? Is it possible that the chain can wander away never to return to 0 again? It turns out that this particular example does eventually return to 0 (see Example 4.6) but the average amount of time it takes to do so is infinite! Issues of what happens in the long run will be a main concern in that chapter.

Chapters 5–8 consider several other issues associated with Markov chains: estimating transition probabilities from observations, the concept of entropy for a Markov chain and it's importance in information and coding theory, and optimization problems associated with Markov chains (such as betting strategies). All of these are important topics for more complicated types of Markov processes, but they are easier to encounter for the first time for Markov chains where the technicalities are much tamer.

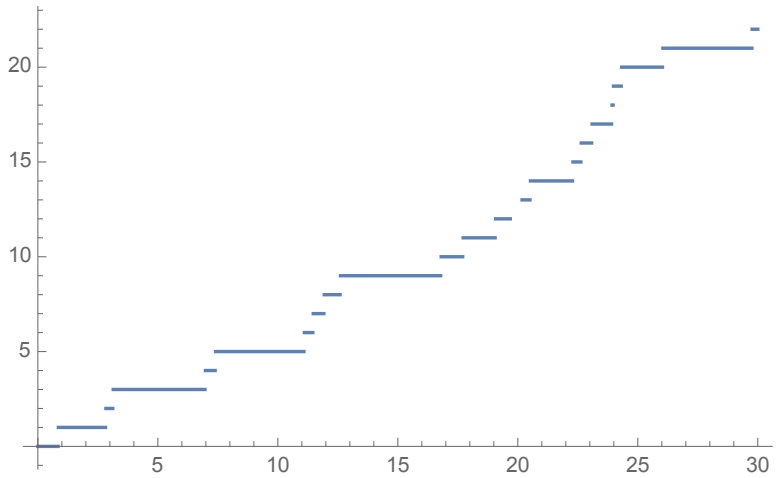
Chapter 9 introduces another class of stochastic processes called *martingales*. We will see in Section 9.2 for instance that there is a fundamental connection between Markov processes and martingales involving the generator $\mathbf{A}$ or $\mathcal{A}$. We include them here both because they are important for more advanced study of stochastic processes and because they are needed for Chapter 10.

Chapter 10 is the one chapter devoted to a particular application: mathematical finance. Since the 1990s applications of stochastic processes to the analysis of financial markets has enjoyed tremendous growth, both as an academic discipline as well as a career specialty in the financial industry. It now probably attracts more student interest in stochastic processes than any other application area. Although most research in this area involves more complicated continuous time models (we will touch on that in Chapter 12) Chapter 10 will introduce some of the fundamental ideas in the context of Markov chains.

The final two chapters concern Markov processes for which the time variable is allowed to vary continuously over the nonnegative real numbers. We still write X_t but now all $0 \le t$ are considered, rather than just $t = 0, 1, 2, \ldots$. The simplest type of continuous time Markov processes are jump processes, the topic of Chapter 11. The most basic example is the Poisson process. Its states are the nonnegative integers. If $X_t = k$ the process stays at k a random amount of time T and then makes an instantaneous jump to $X_{t+T} = k + 1$. By giving T an exponential distribution,

$$P(T \le s) = 1 - e^{-\lambda s},$$

this is a Markov process, as we will see. Here is a graph of a typical sample of X_t (using $\lambda = 1$).

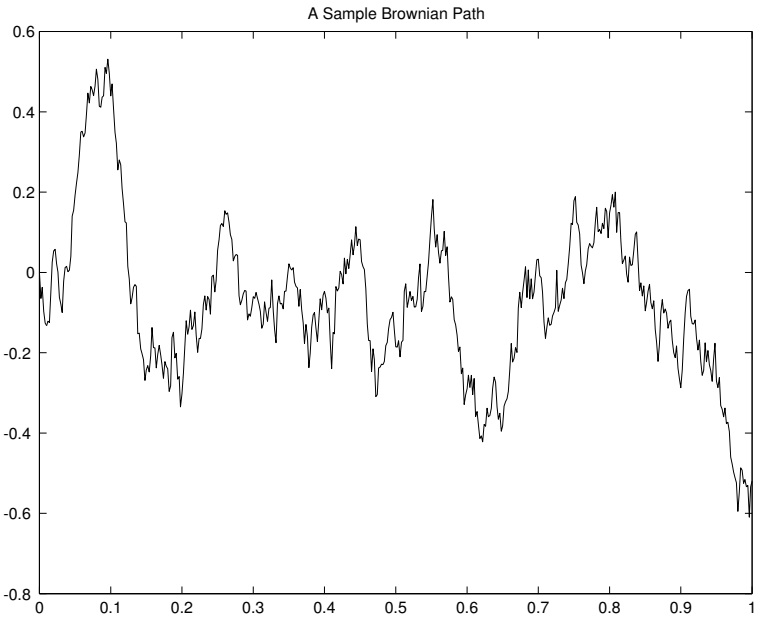


It waits then moves up, waits then moves up, moving step by step up through the integers. All the randomness comes from the waiting times at each state. This can be generalized by allowing the waiting time parameter λ to depend on the current state, λ_k, and by letting the state it jumps to be random as well, determined by some collection $q_{k,j}$ of transition probabilities. The generator takes the form of a difference operator,

$$\mathcal{A}f(k) = \sum_j \lambda_k q_{k,j}[f(j) - f(k)],$$

and the equations describing probabilities become ordinary differential equations (sometimes infinitely many such equations). There are many applications of this general type of process. Some will be indicated in Chapter 11.

There are also Markov processes for which X_t is continuous. These are generally called diffusions. The premier example is Brownian Motion, the topic of Chapter 12. Here is the graph of a typical sample of Brownian Motion.



It is continuous, but very rough, not smooth at all. Brownian motion has many remarkable properties. The generator now becomes a differential operator

$$\mathcal{A}f(x) = \frac{1}{2}f''(x)$$

and the equations describing probabilities become partial differential equations. A careful treatment of Brownian Motion requires graduate-level mathematical analysis, which is beyond the level of our treatment here. Our goal will only be to give an overview of its features, and techniques for working with it. This includes a heuristic introduction to stochastic calculus, and a brief return to mathematical finance in continuous time.

1.2 About Matlab

We will often use simulations to illustrate or examine the behavior of a particular process. This involves the use of computer-generated pseudo-random numbers to construct sample behaviors of a particular Markov process. Section A.3 has some discussion of pseudo-random numbers and their use in MATLAB. Some of our examples and homework problems involve matrix calculations, usually larger than anyone would want to do by hand. We will freely use MATLAB to perform those calculations. You should likewise feel free to rely on MATLAB (or an alternative if your prefer) to carry out matrix calculations required in homework problems. Appendix A includes a brief list of MATLAB commands which are relevant to our use. In general however it is assumed that students have some facility with MATLAB. We will not try to present a self-contained introduction.

1.3 About Notation

Our notation for standard number systems is as follows.

- $\mathbb{N}$ is the natural numbers $\{1, 2, 3, \ldots\}$.
- $\mathbb{Z}$ if the full set of integers $\{\ldots, -3, -2, -1, 0, 1, 2, 3, \ldots\}$.
- $\mathbb{Z}^+$ is the nonnegative integers $\{0, 1, 2, 3, \ldots\}$.
- $\mathbb{R}$ is the set of all real numbers.
- $\mathbb{R}^+$ is the nonnegative real numbers $[0, \infty)$.
- $\mathbb{Z}^d$ is the d-dimensional integer lattice. A typical $z \in \mathbb{R}^d$ is $z = (z_1, \ldots z_d)$, each $z_i \in \mathbb{Z}$.
- $\mathbb{R}^d$ is the d-dimensional Euclidean space. A typical $x \in \mathbb{R}^d$ is $x = (x_1, \ldots x_d)$, each $x_i \in \mathbb{R}$.

For a sequence x_i, $i = 0, 1, \ldots$ we will use $x_{m:n}$ as a shorthand notation for the finite segment where i runs from m to n: $(x_m, \ldots x_n)$.

The number of elements in a set A will be denoted $\#A$.

When referring to a function as a whole, rather than one of its values, we will typically write $f(\cdot)$. The usual $f(x)$ refers to the value of $f(\cdot)$ when evaluated at x. If $f(\cdot, \cdot)$ is a function of two variables, $f(\cdot, y)$ refers to the function of one variable obtained by fixing the value y for the second variable.

Matricies will be in boldface, usually uppercase with their entries in lowercase: $\mathbf{P} = [p_{i,j}]$. The column vector of all 0s or all 1s will be denoted [0] or [1] respectively, the size as determined from the context. When possible we will write the vector comprised of the collection of all values by using boldface: $\mathbf{f} = [f(i)]$ or $\mathbf{u} = [u_i]$. Row vectors will be indicated with parentheses: $(\pi_1, \ldots, \pi_m)$ but again we will sometimes use boldface to indicate the row vector of all possible values: $\mathbf{v} = (v_1, v_2, \ldots) = (v_i)$.

We will continue to use matrix notation even when the set $\mathcal{S}$ of indicies is a countably infinite set. If $B, C \subseteq \mathcal{S}$ are subsets of the index set then the submatrix consisting of those $a_{i,j}$ with $i \in B$, $j \in C$ will be denoted $\mathbf{A}_{BC}$. For instance if $C = \mathcal{S} \setminus B$ then we can think of $\mathbf{A}$ in blocks:

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{BB} & \mathbf{A}_{BC} \\ \mathbf{A}_{CB} & \mathbf{A}_{CC}. \end{bmatrix}.$$

When f is a function of the indices of the matrix we will write $\mathbf{P}f$ for the function obtained by matrix multiplication, with the values of f arranged as a column:

$$\mathbf{P}f(i) = \sum_j p_{i,j} f(j).$$

Thus $\mathbf{P}f(i)$ is the i^{th} entry of the matrix product, the same thing as $(\mathbf{Pf})_i$. When $f(\cdot,\cdot)$ is a function of two variables, $\mathbf{P}f$ will operate on just the first variable:

$$\mathbf{P}f(i,s) = \sum_k p_{i,k} f(k,s)$$

When the set of indicies $\mathcal{S}$ is countably infinite then these expressions refer to infinite series. For instance if the indicies for $\mathbf{P}$ range over all integers ($\mathcal{S} = \mathbb{Z}$) then

$$\mathbf{P}f(i) = \sum_{j\in\mathbb{Z}} a_{i,j} f(j) = \sum_{j=-\infty}^{\infty} a_{i,j} f(j).$$

Random variables will be denoted with uppercase letters, e.g. X or $\mathcal{T}$, and their possible values by lowercase: $P(X = x) = \cdots$. This applies to sequences of their values as well: $P(X_{1:100} = x_{1:100}) = \cdots$. (But not everything in uppercase is random, for instance sets, matrices, Var[·], $P(\cdot)$, $E[\cdot]$.) Stochastic processes consist of random variables which depend on time. We will always put the time variable in the subscript position: X_t.

Chapter 2

Markov Chains: Finite States and Discrete Time

We begin with the simplest type of stochastic process: a Markov chain. This is a process which moves around randomly within a set $\mathcal{S}$ of *states* in accord with some prescribed *transition probabilities* $p_{i,j}$ $(i, j \in \mathcal{S})$. This chapter introduces the basic features of a Markov chains **assuming that the state space $\mathcal{S}$ is finite**. We will see how the Markov property allows us to reduce many problems concerning a Markov chain to matrix equations, which can then be solved with the techniques of linear algebra. In Chapter 4 we will extend these considerations to an infinite state space.

Board games are good examples. Consider the game of Monopoly for instance. The Monopoly playing board consists of 40 positions arranged around a square. Let's label the positions 1 through 40. These comprise the states; the state space is the set

$$\mathcal{S} = \{1, 2, \ldots, 40\}.$$

If your token is at position i then on your next move you roll a pair of dice to determine a random number $2 \le D \le 12$ and then move your token from position i to position $j = i + D$, reduced mod 40 to remain in $\mathcal{S}$ if you "Pass Go". (We are ignoring all the rules about going to jail, or getting sent different places because of a "Chance" card you drew, financial transactions, rolling doubles and so forth. We are just rolling the dice and moving.) Different outcomes of the dice roll occur with different probabilities:

$$P(D = 2) = \frac{1}{36},\ P(D = 3) = \frac{2}{36}\ \cdots\ P(D = 7) = \frac{6}{36}\ \cdots\ P(D = 11) = \frac{2}{36},\ P(D = 12) = \frac{1}{36}.$$

(A concise formula is $P(D = k) = \frac{6-|7-k|}{36}$ for $k = 2, \ldots, 12$ and 0 otherwise.) The positions j you can move to from position i are different for different i; some are not possible and some are more likely than others. The transition probability $p_{i,j}$ is the probability of moving from i to j.

$$p_{i,j} = P(D = k) \text{ if } j = i + k \pmod{40}.$$

For instance,

$$p_{3,9} = \frac{5}{36},\ p_{7,25} = 0,\ p_{32,4} = \frac{1}{36}.$$

All together the $p_{i,j}$ form a 40×40 matrix $\mathbf{P} = [p_{i,j}]$, which is called the *transition matrix*.

Your position/state at time n is denoted X_n. The Monopoly Markov chain produces a sequence of positions: X_0, X_1, X_2, ... Knowing that $X_0 = 1$ and what $\mathbf{P}$ is allows us to calculate the probability that the first several moves work out a particular way. For instance

$$P(X_0 = 1, X_1 = 5, X_2 = 13, X_3 = 20) = p_{1,5}p_{5,13}p_{13,20} = \frac{3}{36}\frac{5}{36}\frac{6}{36} = \frac{5}{2592} = .00192901. \tag{2.1}$$

We will do some more complicated calculations with this chain in Examples 2.7 and 2.5 below.

2.1 Definition and Transition Probabilities

In general a Markov chain X_n is like a board game. The set $\mathcal{S}$ of states is the set of positions on the board. If there are m of them we might as well label them with numbers:

$$\mathcal{S} = \{1, 2, \ldots, m\}.$$

There is an $m \times m$ transition matrix $\mathbf{P} = [p_{i,j}]$ with the properties that $p_{i,j} \geq 0$ and

$$\text{for each } i,\ p_{i,1} + \cdots p_{i,m} = 1; \quad \text{i.e. } \sum_{j \in \mathcal{S}} p_{i,j} = 1.$$

For Monopoly we always start at $X_0 = 1$ but in general we allow X_0 to be chosen randomly from $\mathcal{S}$ in accord with some specified *initial distribution* $\mu = (\mu_1, \ldots, \mu_m)$:

$$P(X_0 = i) = \mu_i.$$

(We require $0 \leq \mu_i$ and $\sum_1^m \mu_i = 1$.) Given μ and $\mathbf{P}$ we generate the Markov chain X_n by starting with a random choice of X_0 in accord with the specified probabilities μ. Then we take the resulting $i = X_0$ and randomly pick X_1 in accord with the probabilities $p_{i,1}, \ldots, p_{i,m}$. Then we take $j = X_1$ and randomly pick X_2 in accord with the probabilities $p_{j,1}, \ldots, p_{j,m}$. Continue to choose X_3, X_4, ... in the same fashion: if $k = X_n$ then we pick X_{n+1} in accord with the probabilities $p_{k,1}, \ldots, p_{k,m}$. The particular value of X_n determines which set of probabilities we use to choose X_{n+1}. Row k of $\mathbf{P}$ contains the probabilities for transitions out of state k.

There are some important points to make before we go further. First, the X_n are random variables, not values that are fixed or predetermined. If you move your token back to $X_0 = 1$ and roll the dice anew you will generate a different sequence $X_0, X_1, X_2, \ldots$ than before, even though $\mathbf{P}$ has not changed. It is the *probabilities* associated with the outcomes X_n which we can determine through careful mathematical analysis, not the actual outcomes X_n.

Second there is a subtlety in the above description of a Markov chain that you may not have noticed. When we identify $k = X_n$ and then use $p_{k,1}, \ldots, p_{k,m}$ to randomly choose X_{n+1}, the idea is that the earlier outcomes $X_0, \ldots, X_{n-1}$ have no influence on the probabilities for choosing X_{n+1}. *It is only the immediately preceding state $k = X_n$ which determines the probabilities of X_{n+1}.* This is the feature that distinguishes a Markov chain from other non-Markov sequences of random variables. To illustrate suppose that X_0 and X_1 are the results of two independent (single) dice throws, but for higher n we just repeat X_0 (if n is even) or X_1 (if n is odd). We don't re-roll the dice after $n = 1, 2$ we just re-use the results of those initial two dice rolls. For instance if the first dice roll is 3 and the second is 5 then

$$3 = X_0 = X_2 = X_4 = \cdots \text{ and } 5 = X_1 = X_3 = X_5 = \cdots .$$

This is a stochastic process X_n on the set of states $\mathcal{S} = \{1, 2, 3, 4, 5, 6\}$ but it is *not* a Markov chain! If it were then there would be no certainty that $X_0 = X_3$. That would only happen only with probability $1/6$.

Some notation will help us describe Markov chains more efficiently. We will use $s_{1:n}$ to refer to the finite sequence of s_i for $i = 1, 2, \ldots n$:

$$s_{1:n} = (s_1, s_2, \ldots s_n).$$

If we specify $s_0 = 0$, $s_1 = 5$, $s_2 = 13$ and $s_3 = 20$ then instead of writing

$$X_0 = 0, X_1 = 5, X_2 = 15, X_3 = 19$$

we can just write

$$X_{0:3} = s_{0:3}.$$

This notation is particularly convenient for talking about long finite sequences. We can describe the probabilities associated with a Markov process with initial distribution μ and transition matrix $\mathbf{P}$ using the formula

$$P(X_{0:n} = s_{0:n}) = \mu_{s_0} \prod_{i=1}^{n} p_{s_{i-1}, s_i} \tag{2.2}$$

for all finite sequences $s_{0:n}$ of states.

Equation (2.2) describes how to determine the probably that $X_{0:n}$ follows a specified sequence $s_{0:n}$ of states. If we want the probability that $X_{0:n}$ does something that can occur in multiple ways then we just add up $P(X_{0:n} = s_{0:n})$ for all the appropriate sequences $s_{0:n}$. For instance suppose we want the probability that $X_3 = 5$. We want to consider all $s_{0:3}$ with $s_3 = 5$.

$$P(X_3 = 5) = \sum_{s_{0:3} \text{ with } s_3=5} P(X_{0:3} = s_{0:3}) = \sum_{s_0} \sum_{s_1} \sum_{s_2} \mu_{s_0} p_{s_0,s_1} p_{s_1,s_2} p_{s_2,5}. \tag{2.3}$$

In this last expression we recognize the occurrence of a matrix product: $\sum_{s_2} p_{s_1,s_2} p_{s_2,5}$ is the $s_1, 5$ entry of the matrix product $\mathbf{PP} = \mathbf{P}^2$. In fact the expression above is just the 5^{th} entry of the (row) matrix $\mu\mathbf{P}^3$.

$$P(X_3 = 5) = (\mu\mathbf{P}^3)_5.$$

Said another way, as a row vector

$$\mu\mathbf{P}^3 = (P(X_3 = 1), \ldots, P(X_3 = 5), \ldots, P(X_3 = m)).$$

Starting with μ as the vector of $P(X_0 = i)$ probabilities, $\mu\mathbf{P}$ is the vector of $P(X_1 = i)$ probabilities, $\mu\mathbf{P}^2$ is the vector of $P(X_2 = i)$ probabilities, $\mu\mathbf{P}^3$ is the vector of $P(X_3 = i)$ probabilities, and so on. Multiplication (on the right) by $\mathbf{P}$ converts the (row) vector of $P(X_n = i)$ probabilities into the (row) vector of $P(X_{n+1} = i)$ probabilities. We could say $\mathbf{P}$ propagates the distribution of X_n one step forward in time to the distribution of X_{n+1}. This connection between the probabilities of X_n and multiplication by the matrix $\mathbf{P}$ is what makes Markov chains amenable to mathematical analysis.

We will use $p_{i,j}(n)$ to denote the i, j entry of $\mathbf{P}^n$, so $\mathbf{P}^n = [p_{i,j}(n)]$. The formula

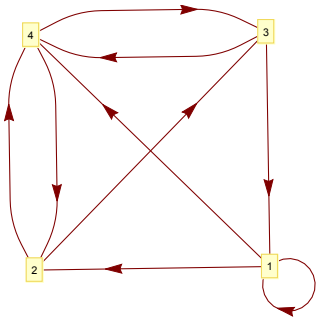
$$p_{i,j}(m + n) = \sum_{k \in \mathcal{S}} p_{i,k}(m) p_{k,j}(n) \tag{2.4}$$

is just the definition of matrix product $\mathbf{P}^{m+n} = \mathbf{P}^m \mathbf{P}^n$. In the terminology of Markov processes this is called the *Chapman-Kolmogorov equation*.

2.1.1 Simulation and Examples

Example 2.1. As a simple example consider the chain on $\mathcal{S} = \{1, 2, 3, 4\}$ with transition matrix

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



We have illustrated the chain as a directed graph with arrows for those state transitions which are possible, i.e. $p_{i,j} > 0$. For instance there is no $2 \to 1$ arrow because $p_{2,1} = 0$. If we start the chain at $X_0 = 1$ a typical sample run is

$$X_{0:20} = (1, 4, 2, 3, 4, 2, 3, 1, 4, 3, 1, 1, 4, 2, 4, 2, 3, 1, 2, 3, 4).$$

To find the probabilities $P(X_{100} = i)$ we can calculate the 100^{th} power of $\mathbf{P}$:

$$\mathbf{P}^{100} = \begin{bmatrix} 0.11605 & 0.22244 & 0.30948 & 0.35203 \\ 0.11605 & 0.22244 & 0.30948 & 0.35203 \\ 0.11605 & 0.22244 & 0.30948 & 0.35203 \\ 0.11605 & 0.22244 & 0.30948 & 0.35203 \end{bmatrix}.$$

Observe that all rows seem to be the same. This particular chain has the property that after many steps the initial state has little influence:

$$P(X_{100} = 1) \approx 0.11605,\ P(X_{100} = 2) \approx 0.22244,\ P(X_{100} = 3) \approx 0.30948,\ P(X_{100} = 3) \approx 0.35203$$

regardless of X_0. $\mathbf{P}^n$ for other large n produces the same result. This is the phenomenon of convergence to an equilibrium distribution, which we will consider more carefully in Section 2.4 below.

How to Simulate

Here is a simple MATLAB m-file to produce sample runs of a Markov chain, as we did in the above example.

mc.m

```
function x=mc(x0,P,k)
%mc(x0,P,k) simulates the evolution of a Markov chain.  x0 is either an
%initial state in {1, ..., n} or an an initial distribution.  P is the nxn
%transition matrix.  k (optional) specifies the run length: 0 ... k.
if nargin==2
   k=1;
end
crs=cumsum(P,2);                    % Cumulative row sums of P
if length(x0)>1
   y=find(rand(1)<cumsum(x0),1);    % Generate initial state
   x=[y,zeros(1,k)];                % Preallocate x
else
   x=[x0,zeros(1,k)];
end
u=rand(1,k);                        % uniform random values
for i=1:k
   x(i+1)=find(u(i)<crs(x(i),:),1);  % i to i+1 chain transition
end
% M. Day, August 13, 2014.
```

Example 2.2. This example is called Feller's Breeding Problem; see Example XV(*2.l*) and Section V.6 of [22]. The idea behind it is this. Most mammals carry two copies of each chromosome, one coming from each parent. If we focus on a particular gene for which two different versions (alleles) are possible, A and a, the genotype of an individual can be AA, Aa, or aa. Suppose that a pair of individuals (the parents) reproduce. There are six different parental genotype pairs:

1:AA&AA, 2:AA&Aa, 3:Aa&Aa, 4:Aa&aa, 5:aa&aa, 6:AA&aa

These will be the states of the Markov chain. A single step of the chain is the result of the following reproductive mechanism (which we will illustrate starting from state 3: Aa&Aa). First the two copies of the chromosome in each parent separate. Then one of the two chromosomes from each parent is randomly chosen and the selected chromosomes from each parent are combined. That produces an offspring of one of the the three genotypes AA, Aa, or aa with certain probabilities ($\frac{1}{4}, \frac{1}{2}, \frac{1}{4}$ in our example). A large population of such children are produced from the original parental pair. The relative frequencies of the three genotypes in this population should be the probabilities just calculated. From this population two individuals are selected at random to be the next parental pair. The following attempts to illustrate the process.

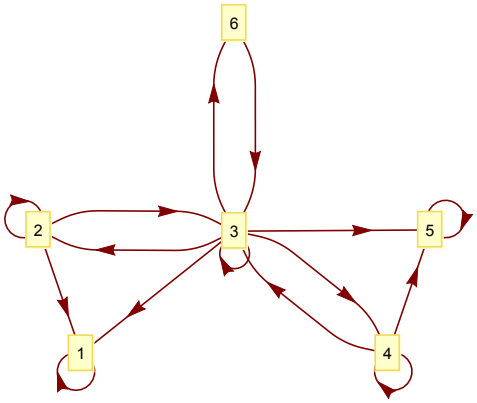
$$\begin{array}{c} \text{Aa \& Aa} \\ \downarrow \\ \text{(A or a) \& (A or a)} \\ \downarrow \\ \text{AA } (\frac{1}{4}) \text{ or Aa } (\frac{1}{2}) \text{ or aa } (\frac{1}{4}) \end{array}$$

$$\downarrow$$

AA&AA $(\frac{1}{4}\cdot\frac{1}{4})$ or AA&Aa $(2\cdot\frac{1}{4}\cdot\frac{1}{2})$ or Aa&Aa $(\frac{1}{2}\cdot\frac{1}{2})$ or Aa&aa $(2\cdot\frac{1}{2}\cdot\frac{1}{4})$ or aa&aa $(\frac{1}{4}\cdot\frac{1}{4})$ or AA&aa $(2\cdot\frac{1}{4}\cdot\frac{1}{4})$

For our example of state 3, these are the probabilities $p_{3,i}$ in the third row of the transition matrix. The other rows are worked out similarly. Here is the result, and a graph indicating the possible transitions among parental pair states.

$$\mathbf{P} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1/4 & 1/2 & 1/4 & 0 & 0 & 0 \\ 1/16 & 1/4 & 1/4 & 1/4 & 1/16 & 1/8 \\ 0 & 0 & 1/4 & 1/2 & 1/4 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$



Observe that it is not possible to leave states 1 or 5; once there the chain is stuck there forever. We call these *absorbing* states. Now look at the 100-step transition probabilities.

$$\mathbf{P}^{100} = \begin{bmatrix} 1. & 0. & 0. & 0. & 0. & 0. \\ 0.75 & 0. & 0. & 0. & 0.25 & 0. \\ 0.5 & 0. & 0. & 0. & 0.5 & 0. \\ 0.25 & 0. & 0. & 0. & 0.75 & 0. \\ 0. & 0. & 0. & 0. & 1. & 0. \\ 0.5 & 0. & 0. & 0. & 0.5 & 0. \end{bmatrix}$$

This indicates that in the long run the chain is certain to end up in either state 1: AA&AA or state 5: aa&aa. The only issue is the probability of landing in one of these as opposed to the other. We can read that off from the row of $\mathbf{P}^{100}$ corresponding to the initial state.

2.2 Hitting Probabilities and Means

Example 2.2 suggests another category of probabilities, those without a time specification. Suppose C and D are two disjoint subsets of $\mathcal{S}$. How might we calculate the probabilities that the following occur?

- X_n eventually reaches C.
- X_n never reaches D.
- X_n reaches C before D.

For the first of these we mean the probability that

$$X_n \in C \text{ for some } n \geq 0.$$

We are asking for the probability that the chain is in C at *some* time but we are not specifying when that has to happen. Probabilities like these cannot be calculated just by finding a certain power of $\mathbf{P}$. A different approach is needed. The key is to explore how these probabilities depend on the initial state $X_0 = i$.

It will be helpful to indicate the initial state $X_0 = i$ in our notation. This is usually done with a subscript on $P(\cdot)$: writing $P_i(\cdot)$ means that probabilities are computed assuming $X_0 = i$. For instance

$$P_3(X_{10} = 23) \text{ means } P(X_{10} = 23) \text{ assuming that } X_0 = 3.$$

We can think of $P_i(X_{10} = 23)$ as a function $u(i)$ of the initial state $i \in \mathcal{S}$. For $P_i(X_{10} = 23)$ we have a formula for this function

$$u(i) = p_{i,23}(10).$$

But for probabilities like the bullets above the answer is not so immediate.

Consider the first bullet and let $u(i)$ be its probability as a function of the initial state i:

$$u(i) = P_i(X_n \in C \text{ for some } 0 \le n < \infty).$$

What can say about this function? For $i \in C$ this is trivial.

$$u(i) = 1 \text{ for } i \in C.$$

For $i \notin C$ whether the chain will ever reach C depends on how the chain evolves in the future. Let the chain move forward one step to $X_1 = j$. If $j \in C$ then we have succeeded in reaching C. Starting from $X_0 = i$ this happens with probability $\sum_{j \in C} p_{i,j}$. But if $X_1 = j \notin C$ the probability of reaching C at some point in the future is $u(j)$. Starting from $X_0 = i$ the probability that $X_1 =$ some $j \notin C$ but then $X_n \in C$ at some $n > 1$ is $\sum_{j \notin C} p_{i,j} u(j)$. So putting these together, for $i \notin C$ we have

$$\begin{aligned} u(i) &= \sum_{j \in C} p_{i,j} + \sum_{j \notin C} p_{i,j} u(j) \\ &= \sum_{j \in \mathcal{S}} p_{i,j} u(j) \\ &= \mathbf{P} u(i). \end{aligned} \tag{2.5}$$

This is a system of equations which when solved yields the values of $u(i)$. We have derived it heuristically, but will give a more formal derivation using conditional expectations in Section 3.7. If we let

$$\mathbf{A} = \mathbf{P} - \mathbf{I}$$

then we can express the equations as

$$\mathbf{A} u(i) = 0 \text{ for } i \notin C, \quad u(i) = 1 \text{ for } i \in C.$$

Example 2.3. Consider Example 2.2. Let's determine the probability that the chain eventually reaches $C = \{1, 2\}$. Note that it is possible for the chain to reach 5 first and be trapped there forever, never reaching C. We want to determine $u(i) = P_i(\text{the chain eventually reaches states 1 or 2})$. We know $u(1) = u(2) = 1$. Here is the system of equations we need to solve; $u(i) = \mathbf{P} u(i)$ for $i = 3, 4, 5, 6$:

$$\begin{aligned} \begin{bmatrix} u(3) \\ u(4) \\ u(5) \\ u(6) \end{bmatrix} &= \begin{bmatrix} \frac{1}{16} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{16} & \frac{1}{8} \\ 0 & 0 & \frac{1}{4} & \frac{1}{2} & \frac{1}{4} & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ u(3) \\ u(4) \\ u(5) \\ u(6) \end{bmatrix} \\ &= \begin{bmatrix} \frac{5}{16} \\ 0 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{16} & \frac{1}{8} \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} u(3) \\ u(4) \\ u(5) \\ u(6) \end{bmatrix} \end{aligned}$$

$$\begin{bmatrix} \frac{3}{4} & -\frac{1}{4} & -\frac{1}{16} & -\frac{1}{8} \\ -\frac{1}{4} & \frac{1}{2} & -\frac{1}{4} & 0 \\ 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} u(3) \\ u(4) \\ u(5) \\ u(6) \end{bmatrix} = \begin{bmatrix} \frac{5}{16} \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

The matrix on the left is singular. In particular the value of $u(5)$ can be treated as a free variable. But because 5 is an absorbing state we know $u(5) = 0$. Using this leads to the solution

$$u(1) = 1, u(2) = 1, u(3) = \frac{5}{8}, u(4) = \frac{5}{16}, u(5) = 0, u(6) = \frac{5}{8}.$$

So in particular $P_4($ the chain eventually reaches states 1 or 2$) = \frac{5}{16}$.

Observe that this approach involves finding all values of $u(i) = P_i(X_n \in C \text{ for some } 0 \leq n < \infty)$ simultaneously, not one i at a time. This example also illustrates that the equations for $u(i)$ need not have a unique solution. (But if we had used $u(1) = u(2) = 1$ *as well as* $u(5) = 0$ from the outset the resulting system *would* have had a unique solution.)

Example 2.4. Let the state space be $\mathcal{S} = \{0, \ldots, k\}$. If $0 < X_n < k$ the chain moves one step to the right with probability p or one step to the left with probability $q = 1 - p$. If $X_n = 0$ or k it just stays where it is. This is what we would call a *random walk* on $\{0, \ldots, k\}$ with *absorbing* endpoints. Let $u(i)$ be the probability of eventually reaching $C = \{0\}$. We know $u(0) = 1$, $u(k) = 0$ (because k is absorbing), and for $0 < i < k$ we have $u(i) = \mathbf{P}u(i)$, which reduces to

$$u(i) = pu(i + 1) + qu(i - 1).$$

We can solve this explicitly if we rearrange it as

$$[u(i + 1) - u(i)] = \frac{q}{p}[u(i) - u(i - 1)].$$

From here we see that

$$u(i + 1) - u(i) = (q/p)^i[u(1) - 1].$$

If $p \neq q$ we can continue to find that for $j \geq 1$

$$u(j) = u(0) + \sum_{i=0}^{j-1}[u(i + 1) - u(i)] = 1 + \sum_{i=0}^{j-1}(q/p)^i[u(1) - 1] = 1 + \frac{(q/p)^j - 1}{(q/p) - 1}[u(1) - 1]. \tag{2.6}$$

In order for $u(k) = 0$ we must have

$$[u(1) - 1] = -\frac{(q/p) - 1}{(q/p)^k - 1},$$

which gives us the solution

$$u(i) = \frac{(q/p)^k - (q/p)^i}{(q/p)^k - 1}.$$

In the case of $p = q = 1/2$ we need to calculate equation (2.6) differently:

$$u(j) = 1 + (j - 1)[u(1) - 1].$$

Now $u(k) = 0$ implies $[u(1) - 1] = -1/(k - 1)$ and so

$$u(i) = \frac{k - i}{k}.$$

Matrix Equations and Solvability

In equation (2.5) we know the values of $u(i)$ for $i \in C$. Let $B = \mathcal{S} \setminus C$. The $u(i), i \in B$ are the terms we need to solve for. We can rearrange equation (2.5) by separating the known and unknown terms into two (column) vectors:

$$\mathbf{u}_B = [u(i)]_{i \in B}, \quad \mathbf{u}_C = [u(i)]_{i \in C},$$

and break up the $i \in B$ rows of $\mathbf{P}$ into two submatricies:

$$\mathbf{P}_{BB} = [p_{i,j}]_{i \in B, j \in B}, \quad \mathbf{P}_{BC} = [p_{i,j}]_{i \in B, j \in C}.$$

With this notation we can express our system of equations (2.5) as

$$\mathbf{u}_B = \mathbf{P}_{BB}\mathbf{u}_B + \mathbf{P}_{BC}\mathbf{u}_C. \tag{2.7}$$

The task is to solve for $\mathbf{u}_B$ with given values for $\mathbf{u}_C$. For this purpose we would like to know when

$$\mathbf{P}_{BB} - \mathbf{I} = \mathbf{A}_{BB}$$

is invertible, and if it is not invertible how we can identify which of the many possible solutions is the correct one. We will answer the first of these questions in Theorem 2.3 below and the second in Theorem 4.1.

In the next example we illustrate how these calculations can be carried out in MATLAB.

Example 2.5. Consider again the Monopoly playing board with squares $\mathcal{S} = \{1, \ldots, 40\}$. The transition probabilities are the result of adding two dice rolls to the current position, with "wrap around", i.e. reduction mod 40. The following script generates the transition matrix for a *single* dice roll.

gP40.m

```
%Script to generate single dice roll transition matrix on 40x40 game board
%with "wrap around".
P40=spdiags(ones(40,6)/6,1:6,40,46);
P40(:,1:6)=P40(:,1:6)+P40(:,41:46);
P40=full(P40(:,1:40));
```

So the transition matrix for a double dice roll is given by

```
P=P40^2;
```

To illustrate hitting probability calculations let's find the probability of reaching "Go to Jail" (31) *without* first landing on "Chance" $(8, 23, 37)$. (This is an example of the third bullet above, with $C = \{31\}$, $D = \{8, 23, 37\}$.) For purposes of our calculation let's combine the target and "avoid" sets as $C = \{8, 23, 31, 37\}$ and take $\mathbf{u}_C$ defined by $u(8) = u(23) = u(37) = 0$ and $u(31) = 1$. We need to solve $\mathbf{u}_B = \mathbf{P}_{BB}\mathbf{u}_B + \mathbf{P}_{BC}\mathbf{u}_C$ for $\mathbf{u}_C$. Here is how to do the calculations in MATLAB.

```
C=[8,23,31,37]; B=setdiff(1:40,C);
PBB=P(B,B); PBC=P(B,C); uC=[0;0;1;0];
cond(eye(36)-PBB)                    %To check invertibility
uB=(eye(36)-PBB)\(PBC*uC);
uB(1)
```

The resulting value $u(1) = .2135$ is the probability of landing on "Go to Jail" without first landing on "Chance" starting from $X_0 = 1$. For initial state $X_0 = 28$ we find $u(28) =$`uB(26)`$= .2343$ because state 28 is entry number 26 in B. (Be careful about indicies when looking up specific results!)

2.2.1 Hitting Times

Suppose $C \subseteq \mathcal{S}$ is a subset of states. As the chain X_n proceeds we can watch for the first time the chain lands in C or the first *positive* time this happens. This determines two time-valued random variables.

$$\mathcal{T}_C = \begin{cases} \min\{n \geq 0 : X_n \in C\} & \text{if } X_n \in C \text{ for some } n \geq 0, \\ \infty & \text{otherwise.} \end{cases}$$

$$\mathcal{T}_C^+ = \begin{cases} \min\{n > 0 : X_n \in C\} & \text{if } X_n \in C \text{ for some } n > 0, \\ \infty & \text{otherwise.} \end{cases}$$

These only differ if $X_0 \in C$; in that situation $\mathcal{T}_C = 0$ but $\mathcal{T}_C^+$ is the first time X_n *returns* to C; $\mathcal{T}_C^+ \geq 1$. We will call $\mathcal{T}_C$ the *hitting time* of C and $\mathcal{T}_C^+$ the *first return time* to C. We will be concerned with $\mathcal{T}_C$ for the moment, but $\mathcal{T}_C^+$ will be important when we talk about recurrence in Section 2.3. (If $C = \{a\}$ contains a single state we will write $\mathcal{T}_C = \mathcal{T}_a$ and $\mathcal{T}_C^+ = \mathcal{T}_a^+$.)

These definitions allow more concise expressions for the three bullets on page 10:

- $\mathcal{T}_C < \infty$
- $\mathcal{T}_D = \infty$
- $\mathcal{T}_C < \mathcal{T}_D$.

For instance $u(i) = P_i(\mathcal{T}_C < \infty)$ is the $u(i)$ of (2.5).

About $P(\mathcal{T}_C = \infty)$

The equations for $v(i) = P_i(\mathcal{T}_C = \infty)$ are easily obtained from the $u(i)$ equations (2.5) because

$$v(i) = P_i(\mathcal{T}_C = \infty) = 1 - P_i(\mathcal{T}_C < \infty) = 1 - u(i),$$

or $\mathbf{v} = [1] - \mathbf{u}$. Since $[1] = \mathbf{P}[1]$ it follows that $\mathbf{v} = \mathbf{P}\mathbf{v}$ on $B = \mathcal{S} \setminus C$, or

$$\mathbf{v}_B = \mathbf{P}_{BB}\mathbf{v}_B + \mathbf{P}_{BC}\mathbf{v}_C.$$

This is the same as equation (2.7) except that now $\mathbf{v}_C = \mathbf{0}$, because for $i \in C$ we have $P_i(\mathcal{T}_C = \infty) = 0$. Thus the equation is simply

$$\mathbf{v}_B = \mathbf{P}_{BB}\mathbf{v}_B. \tag{2.8}$$

Only if $\mathbf{A}_{BB} = \mathbf{P}_{BB} - \mathbf{I}$ is singular can there be a nonzero solution, and in that case there are infinitely many solutions. We will need to understand which of them actually gives us the correct $v(i)$. Theorem 2.3 will tell us that singularity means there is a closed communication class in B, which means some $v(i) = 1$.

The Distribution of $\mathcal{T}_C$

The distribution of $\mathcal{T}_C$ is described by the probabilities

$$w(i, n) = P_i(\mathcal{T}_C = n).$$

Clearly

$$w(i, 0) = \begin{cases} 1 & \text{if } i \in C \\ 0 & \text{if } i \notin C. \end{cases}$$

For $n \geq 1$ we have $w(i, n) = 0$ if $i \in C$. If $i \notin C$ then $\mathcal{T}_C = n$ will require the chain to go from $X_0 = i$ to some $X_1 = j \notin C$ and then $n - 2$ additional steps staying out of C, and then finally landing in C for the first time on the last one, X_n. This has probability $p_{i,j}w(j, n-1)$. Adding this up over the possible $j \in \mathcal{S}$ we find that for $i \notin C$

$$\begin{aligned} w(i, n) &= \sum_{j \in \mathcal{S}} p_{i,j} w(j, n-1) \\ &= \mathbf{P}w(i, n-1), \end{aligned}$$

using the notation introduced in Section 1.3. With $B = \mathcal{S} \setminus C$ we can express all this as

$$\begin{aligned} \mathbf{w}_C(0) &= [1], \\ \mathbf{w}_B(0) &= [0], \\ \mathbf{w}_C(n) &= [0] \text{ for } n \geq 1, \\ \mathbf{w}_B(1) &= \mathbf{P}_{BC}[1], \\ \mathbf{w}_B(n+1) &= \mathbf{P}_{BB}\mathbf{w}_B(n) \text{ for } n > 1. \end{aligned} \tag{2.9}$$

Iterating the last line makes computation straightforward.

Next, assuming $P_i(\mathcal{T}_C < \infty) = 1$, consider the mean hitting times

$$v(i) = E_i[\mathcal{T}_C] = \sum_{n}^{\infty} n P_i(\mathcal{T}_C = n).$$

Clearly $v(i) = 0$ if $i \in C$. For $i \notin C$ we have to go at least one step to $X_1 = j$ and then some number of additional steps whose mean is $v(j)$. So for $i \notin C$

$$\begin{aligned} v(i) &= 1 + \sum_{j \in \mathcal{S}} p_{i,j} v(j) \\ &= 1 + \mathbf{P} v(i). \end{aligned} \tag{2.10}$$

Or simply

$$\mathbf{v}_B = [1] + \mathbf{P}_{BB} \mathbf{v}_B.$$

As noted above, for $P_i(\mathcal{T}_C < \infty) = 1$ for all $i \in B$ it must be that $(\mathbf{I} - \mathbf{P}_{BB})$ is nonsingular, in which case the unique solutions is $(\mathbf{I} - \mathbf{P}_{BB})^{-1}[1]$.

Example 2.6. Consider again the absorbing random walk of Example 2.4. We will calculate the mean time to absorbtion at either of the endpoints, $v(i) = E_i[\mathcal{T}_{\{0,k\}}]$. We know $v(0) = 0 = v(k)$ and for $0 < i < k$

$$v(i) = 1 + pv(i+1) + qv(i-1).$$

For $p = q = 1/2$ the solution is simply $v(i) = i(k-i)$. For $p \neq q$ it works out to be

$$v(i) = \frac{k}{q-p} \left(\frac{i}{k} - \frac{(q/p)^i - 1}{(q/p)^k - 1} \right).$$

Problem 2.10 asks you to confirm these formulas.

Example 2.7. Here is another MATLAB calculation based on the Monopoly board. We will compute the mean number of steps until landing on "Go to Jail" (31). First construct the transition matrix as in Example 2.5. Then calculate as follows.

```
C=31;
B=setdiff(1:40,C);
PBB=P(B,B);
det(eye(39)-PBB)              %to check invertibility
vB=(eye(39)-PBB)\ones(39,1);
vB(1)
```

This works out to $v(1) = 38.5841$.

2.3 State Classification

Next we consider properties of a Markov chain which depend simply on the network of possible transitions, i.e. features visible in our graphical representation of a chain. We will need some terminology.

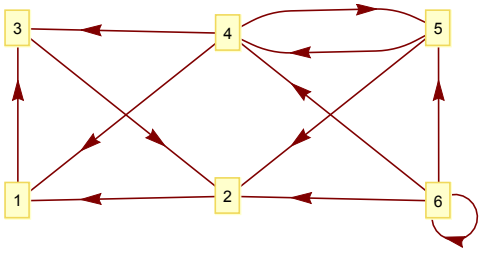
Definition. *Let* $\mathbf{P}$ *be the transition martix for a Markov chain with state space* $\mathcal{S}$.

- *For* $i, j \in \mathcal{S}$ *if* $p_{i,j}(n) > 0$ *for some* $n \geq 0$ *we say* j is reachable from i *and write* $i \rightsquigarrow j$.
- *When both* $i \rightsquigarrow j$ *and* $j \rightsquigarrow i$ *we say* i *and* j communicate *and write* $i \leftrightsquigarrow j$.
- *The chain is* irreducible *if* $i \leftrightsquigarrow j$ *for every pair* $i, j \in \mathcal{S}$.
- *We say* $C \subseteq \mathcal{S}$ *is* closed *if* $i \in C$ *and* $i \rightsquigarrow j$ *implies* $j \in C$.
- *The greatest common divisor of the set of* $n > 0$ *for which* $p_{i,i}(n) > 0$ *is called the* period *of state* i. *When the period is* 1 *we call* i *an* aperiodic *state.*

Intuitively $i \rightsquigarrow j$ simply means that starting at $X_0 = i$ there is a chance of finding $X_n = j$ at *some* time n. Since $p_{i,i}(0) = 1$ the definition says that $i \rightsquigarrow i$ and $i \leftrightsquigarrow i$ are always true. Moreover if $i \rightsquigarrow j$ and $j \rightsquigarrow k$ then from (2.4) we see that $i \rightsquigarrow k$. In fact $\leftrightsquigarrow$ is an equivilence relation and so partitions $\mathcal{S}$ into equivalence classes. I.e. every state belongs to exactly one communication class. An equivalence class is called a *communication class* of states. To say the chain is irreducible means it consists of a single closed communication class; i.e. $i \leftrightsquigarrow j$ for every pair of states.

Example 2.8. With $\mathcal{S} = \{1, \ldots, 6\}$ consider the following Markov chain.

$$\mathbf{P} = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0.3 & 0 & 0.3 & 0 & 0.4 & 0 \\ 0 & 0.5 & 0 & 0.5 & 0 & 0 \\ 0 & 0.25 & 0 & 0.25 & 0.25 & 0.25 \end{bmatrix}$$



We can see three communication classes:

- $\{1, 2, 3\}$ is a closed class of states with period 3.
- $\{4, 5\}$ is a non-closed class of states with period 2.
- $\{6\}$ is a non-closed class with period 1.

This only depends on which $p_{i,j}$ are positive and which are 0, not on any other information about their values. In other words it only depends on what we can see in the graph.

The following lemma provides a useful technical fact about closed sets of states.

Lemma 2.1 (Maximal Lemma)**.** *Let* $\mathbf{P}$ *be the transition matrix for a Markov chain. Suppose* $B \subseteq \mathcal{S}$*,* $\phi : B \to \mathbb{R}$ *is bounded above by a constant* β *(i.e.* $\phi(a) \leq \beta$ *for all* $a \in B$*) and satisfies*

$$\phi(i) \leq \sum_{j \in B} p_{i,j}\phi(j) \text{ for all } i \in B.$$

If either $\beta > 0$ *or* $B = \mathcal{S}$ *then*

$$M = \{i \in B : \phi(i) = \beta\}$$

is a closed set of states.

When $\mathcal{S}$ is finite every function $\phi : B \to \mathbb{R}$ is bounded above. The lemma remains true for infinite state spaces so we have stated it without making $\mathcal{S}$ finite a hypothesis. When $B = \mathcal{S}$ the lemma says that if $\phi \leq \mathbf{P}\phi$ then the states where ϕ takes its maximum value form a closed class.

Proof. Suppose that either $\beta > 0$ or $B = \mathcal{S}$ and consider any $i \in M$. We want to show that if $p_{i,j} > 0$ then $j \in M$. We know

$$\beta = \phi(i) \leq \sum_{k \in B} p_{i,k}\phi(k) \leq \sum_{k \in B} p_{i,k}\beta \leq \sum_{j \in \mathcal{S}} p_{i,j}\beta = \beta.$$

($\beta > 0$ or $B = \mathcal{S}$ is needed for the last inequality.) Since the outside values are equal, all the inequalities must in fact be equalities. Equality of the last two summations means that $p_{i,j}\beta = 0$ for any $j \in \mathcal{S} \setminus B$. So if $\beta > 0$ then any j with $p_{i,j} > 0$ must also belong to B. (This is trivial if $B = \mathcal{S}$.) Thus $p_{i,j} > 0$ implies $j \in B$. Next observe that equality of the first two summations implies that any $k \in B$ with $p_{i,k} > 0$ must have $\phi(k) = \beta$, else we would have a strict inequality $p_{i,k}\phi(j) < p_{i,k}\beta$ between the first and second summations. Thus M contains all $j \in \mathcal{S}$ with $p_{i,j} > 0$. This proves the lemma. □

Closed communication classes are particularly important. For finite chains there is always at least one closed class.

Lemma 2.2. *If* $B \subseteq \mathcal{S}$ *is a finite closed set of states then it contains a closed communication class.*

Proof. For each $i \in B$ define

$$D_i = \{s \in \mathcal{S} : i \rightsquigarrow s\}.$$

Clearly D_i is closed, and since B is closed $D_i \subseteq B$. Choose $i^* \in B$ for which $\#D_{i^*}$ is as small as possible. We claim that D_{i^*} is a communication class. Consider any $j \in D_{i^*}$. We know $i^* \rightsquigarrow j$; we just need to show that $j \rightsquigarrow i^*$. Obviously $D_j \subseteq D_{i^*}$. But since the size of D_{i^*} is minimal it must be that $D_j = D_{i^*}$. Since $i^* \in D_{i^*}$ it follows that $i^* \in D_j$. Therefore $j \rightsquigarrow i^*$, completing the proof. □

We can now answer one of the questions we asked in Section 2.2.

Theorem 2.3. *Suppose* $\mathbf{P}$ *is a transition matrix on a state space* $\mathcal{S}$ *and* $B \subseteq \mathcal{S}$ *is a finite subset. The matrix* $\mathbf{P}_{BB} - \mathbf{I}$ *is singular if and only if there is a closed communication class entirely contained in* B*.*

Proof. Suppose $D \subseteq B$ is a closed communication class and let $C = B^c$. Let $v(i) = P_i(\mathcal{T}_C = \infty)$. We know from equation (2.8) that this provides a solution of the equation

$$(\mathbf{P}_{BB} - \mathbf{I})v(i) = 0 \text{ for all } i \in B.$$

But since $D \subseteq B$ is closed $\mathcal{T}_C = \infty$ for $X_0 = i \in D$. Thus $v(i) = 1$ for some $i \in D$. Thus there exists a nontrivial solution to the above equation, proving that $\mathbf{P}_{BB} - \mathbf{I}$ is singular.

Conversely assume that $\mathbf{P}_{BB} - \mathbf{I}$ is singular. Then there exists a nontrivial solution to $(\mathbf{P}_{BB} - \mathbf{I})v(i) = 0$ for all $i \in B$. After replacing v with $-v$ if necessary we can assume that some of the $v(i)$ are positive. Let $\beta = \max_{k \in B} v(k)$. It follows from the Maximal Lemma 2.1 that $\{i \in B : v(i) = \beta\}$ is a closed set of states in B which by Lemma 2.2 contains a closed communication class. □

Here are some basic properties related to the period of a state.

Lemma 2.4. *Suppose* X_n *is a Markov chain with state space* $\mathcal{S}$*.*

a) If $i \in \mathcal{S}$ *has period* d *then* $p_{i,i}(n) = 0$ *if* n *is not a multiple of* d*.*

b) If i *has period* d *there exists* K *so that* $p_{i,i}(kd) > 0$ *for all* $k \geq K$*.*

c) If $i \leftrightsquigarrow j$ *then* i *and* j *have the same period.*

The lemma is true even for infinite state spaces. However the following corollary depends on $\mathcal{S}$ being finite. Note that by c) of the lemma all states in a communication class C have the same period. That is why in the corollary we can just say that "C is aperiodic".

Corollary 2.5. *Suppose* X_n *is a Markov chain with finite state space* $\mathcal{S}$*. If* C *is a closed aperiodic class there exists* N *and a positive constant* $\alpha > 0$ *so that* $p_{i,j}(n) \geq \alpha$ *for all* $i, j \in C$ *and all* $n \geq N$*.*

Proof of the Lemma. By definition the period d of a state i is the greatest common divisor of $W = \{n > 0 : p_{i,i}(n) > 0\}$. In particular if $p_{i,i}(n) > 0$ then d divides n. This proves a) of the lemma. It is a basic property of greatest common divisors that all multiples kd beyond some Kd can be expressed as a finite sum

$$kd = \sum_1^\ell n_i.$$

for some integers $n_i \in W$. (See Lemma A.9 in the Appendix and note that $\alpha_i n_i = n_i + n_i + \cdots + n_i$ using α_i repeated terms.) Based on that we can say that

$$p_{i,i}(kd) \geq p_{i,i}(n_1) \cdots p_{i,i}(n_\ell) > 0.$$

This proves b).

For c), by hypothesis there exist k and ℓ for which $p_{i,j}(k) > 0$ and $p_{j,i}(\ell) > 0$. Suppose i has period d and j has period g. It follows that $p_{j,j}(k + \ell) > 0$ and therefore g divides $k + \ell$. If $p_{i,i}(n) > 0$ then

$$p_{j,j}(\ell + n + k) \geq p_{j,i}(\ell) p_{i,i}(n) p_{i,j}(k) > 0$$

and so g divides $k + n + \ell$, and therefore must divide n itself. We conclude that g is a common divisor of W and so divides d. The same argument with i and j reversed shows that d divides g. Therefore $d = g$. □

Proof of the Corollary. Suppose $i, j \in C$. Since $i \rightsquigarrow j$ there exists m with $p_{i,j}(m) > 0$. By b) of the lemma (with period $d = 1$) there exists K with $p_{j,j}(n) > 0$ for all $n \geq K$. It follows that

$$p_{i,j}(m+n) \geq p_{i,j}(m)p_{j,j}(n) > 0$$

for all $m + n \geq m + K$. Thus for every pair $i, j \in C$ there exists $N_{i,j}$ such that $p_{i,j}(n) > 0$ for all $n \geq N_{i,j}$. Now just take N to be the maximum of $N_{i,j}$ over all pairs in C. Let $\alpha = \min_{i,j \in C} p_{i,j}(N)$. Then for any $n > N$ and $i, j \in C$ write $n = N + m$. We have (since $p_{i,k}(m) = 0$ for $k \notin C$)

$$p_{i,j}(n) = \sum_{k \in C} p_{i,k}(m)p_{k,j}(N) \geq \sum_{k \in C} p_{i,k}(m)\alpha = \alpha.$$

□

Definition. *A state i is called* recurrent *if*

$$P_i(X_n = i \text{ for some } n \geq 1) = 1.$$

If this probability is < 1 state i is called transient.

Recurrence means that starting from $X_0 = i$ the chain is certain to return to i eventually. The definition of can be expressed as

$$P_i(\mathcal{T}_i^+ < \infty) = 1.$$

It is important to use $\mathcal{T}_i^+$ rather than $\mathcal{T}_i$ here because $P_i(\mathcal{T}_i < \infty) = 1$ is always true. In fact

$$P_i(\mathcal{T}_i^+ < \infty) = \sum_j p_{i,j} P_j(\mathcal{T}_i < \infty). \tag{2.11}$$

Part a) of the next theorem gives a simple characterization of the recurrent states. In Chapter 4 we will see that *both parts of this theorem can fail for infinite state spaces!*

Theorem 2.6. *Suppose $\mathcal{S}$ is finite.*

a) A state i is recurrent if and only if it belongs to a closed communication class.

b) If i is a recurrent state, then $E_i[\mathcal{T}_i^+] < \infty$.

Proof. Suppose i is recurrent and let

$$u(j) = P_j(\mathcal{T}_i < \infty).$$

For $j \neq i$ we know from (2.5) that $u(j) = \mathbf{P}u(j)$. By recurrence and (2.11) we know that this holds for i as well:

$$u(i) = 1 = P_i(X_n = i \text{ for some } n \geq 1) = \sum_j p_{i,j} u(j).$$

The Maximal Lemma 2.1 tells us that $\{j \in \mathcal{S} : u(j) = 1\}$ is a closed set of states. Since this set contains i it follows that $i \rightsquigarrow j$ implies $u(j) = 1$. But $u(j) = 1$ implies that $j \rightsquigarrow i$ so j is in the same communication class as i. Thus the communication class of i consists of all j with $i \rightsquigarrow j$, and is therefore closed. This proves the "only if" part of a).

Now let C be the communication class of i and assume it is closed. Let $B = C \setminus \{i\}$. Every $j \in B$ has $j \rightsquigarrow i$ since C is the communication class of i. So C contains no closed classes. By Theorem 2.3 this means $\mathbf{P}_{BB} - \mathbf{I}$ is invertible. So there is a unique solution to

$$(\mathbf{P}_{BB} - \mathbf{I})u = \mathbf{P}_{BB^c}[1],$$

which by (2.5) must be $u(j) = P_j(\mathcal{T}_{B^c} < \infty)$. Because C is closed no states outside C can be reached from $j \in B$ so this is the same as $u(j) = P_j(\mathcal{T}_{\{i\}} < \infty)$. Now observe that $u = [1]$ also satisfies this equation. Therefore by uniqueness of the solution we see that

$$P_j(\mathcal{T}_{\{i\}} < \infty) = 1 \text{ for all } j \in B.$$

By (2.11) it follows that $P_i(\mathcal{T}^+_{\{i\}} < \infty) = 1$, completing the proof of a).

For b), assume i is a recurrent state. Observe that the mean of $\mathcal{T}_i^+$ can be written

$$\begin{aligned} E_i[\mathcal{T}_i^+] &= \sum_{k=1}^{\infty} k P_i(\mathcal{T}_i^+ = k) \\ &= \sum_{k=1}^{\infty} \sum_{n=1}^{k} 1 \cdot P_i(\mathcal{T}_i^+ = k) \\ &= \sum_{n=1}^{\infty} \sum_{k=n}^{\infty} P_i(\mathcal{T}_i^+ = k) \\ &= \sum_{n=1}^{\infty} P_i(\mathcal{T}_i^+ \geq n) \\ &= \sum_{n=0}^{\infty} P_i(\mathcal{T}_i^+ > n). \end{aligned}$$

We can produce a concise matrix expression for $P_i(\mathcal{T}_i^+ > n)$. Of course

$$P_i(\mathcal{T}_i^+ > 0) = 1.$$

Let C be the communication class if i, which by a) must be closed, and let $B = C \setminus \{i\}$. Then $P_i(\mathcal{T}_i^+ > 1)$ is just $P_i(X_1 \in B)$. Let $\mathbf{P}_{BB}$ be our usual submatrix of the $p_{j,k}$ for $j, k \in B$. Then

$$P_i(X_1 \in B) = \left(\mathbf{P}_{BB}[1]\right)_i,$$

the i-entry of $\mathbf{P}_{BB}[1]$. In general

$$\begin{aligned} P_i(\mathcal{T}_i^+ > n) &= P_i(X_k \in B \text{ for each } k = 1, \ldots n) \\ &= \left(\mathbf{P}_{BB}^n[1]\right)_i. \end{aligned}$$

Therefore

$$E_i[\mathcal{T}_i^+] = 1 + \sum_{n=1}^{\infty} \left(\mathbf{P}_{BB}^n[1]\right)_i.$$

Next we will show that this series is convergent, more specifically that $\mathbf{P}_{BB}^n[1] \to [0]$ geometrically. First, $\mathbf{P}_{BB}[1] \leq [1]$ and so $\mathbf{P}_{BB}^n[1]$ is monotone decreasing. Therefore $L = \lim \mathbf{P}_{BB}^n[1]$ exists. Now $L = \mathbf{P}_{BB}L$ so $(\mathbf{I} - \mathbf{P}_{BB})L = [0]$. Since $\mathbf{I} - \mathbf{P}_{BB}$ is invertible (B contains no closed classes) we deduce that $L = [0]$. So there must exist N with $\mathbf{P}_{BB}^N[1] \leq \frac{1}{2}[1]$. It follows that $\mathbf{P}_{BB}^n[1] \leq 2(\frac{1}{2})^{\frac{n}{N}}[1]$. Finally,

$$E_i[\mathcal{T}_i] = 1 + \sum_{1}^{\infty} p_{i,B} \mathbf{P}_{BB}^n[1] \leq 1 + 2(\#B) \sum_{1}^{\infty} \frac{1}{2^{\frac{n}{N}}} < \infty.$$

□

2.4 Equilibrium

Examples 2.1 and 2.2 made observations about the distribution of X_n for large n. The behavior of X_n in the long run, as $n \to \infty$, is the subject of this section. Except in trivial cases, X_n keeps moving from one state to another. So we can't expect X_n to actually converge as $n \to \infty$. It can't have an equilibrium in the same sense as a differential equation $\dot{x}(t) = F(x(t))$. In that setting an equilibrium is a point x^* so that if $x(0) = x^*$ then $x(t) = x^*$ for all $t \geq 0$, i.e. a rest point for $x(t)$. But a Markov chain can converge to an equilibrium in a different sense, namely that $P(X_n = i) \to \pi_i$ as $n \to \infty$. Such a $\pi = (\pi_1, \ldots, \pi_n)$ is

a statistical equilibrium or *equilibrium distribution*. If the initial distribution is μ then we should find π as $\lim_{n\to\infty} \mu\mathbf{P}^n$. If such a limiting distribution does exist then it follows that

$$\pi = \lim \mu\mathbf{P}^n = \lim \mu\mathbf{P}^{n+1} = (\lim \mu\mathbf{P}^n)\,\mathbf{P} = \pi\mathbf{P}.$$

In other words if we start the chain with π as its initial distribution then $P(X_n = i) = \pi_i$ for all $n \geq 0$. Thus π is a rest point for the distribution of X_n, although X_n itself never comes to rest. An equlibrium distribution is often called a *stationary* or *invariant* distribution.

Finding an equilibrium distribution means solving the matrix equation

$$\pi = \pi\mathbf{P}. \tag{2.12}$$

You might recognize this as saying that π is a (left) eigenvector of $\mathbf{P}$ with eigenvalue $\lambda = 1$. But not just any left eigenvector will be an equilibrium distribution; it has also to be a legitimate probability distribution: $\pi_i \geq 0$ and $\sum_i \pi_i = 1$. Not every eigenvector has that property.

Given a (finite state) Markov chain, the questions we want to address are the following.

- Do equilibrium distributions exist? Can there be more than one?
- Given a non-equilibrium initial distribution μ what can we say about $\lim_{n\to\infty} \mu\mathbf{P}^n$? Does this necessarily converge?
- What about the long run average state of the chain, $\frac{1}{n}\sum_1^n X_i$? Does this converge as $n \to \infty$?

Example 2.9. Let's consider again Example 2.1. Taking the transpose of (2.12) to get $(\mathbf{I} - \mathbf{P})^T\pi^T = [0]$ and solving we find a one-parameter family of solutions

$$c\mu \text{ where } \mu = [30/91, 115/182, 80/91, 1]^T.$$

choosing $c = 1/\sum \mu_i = 182/517$ we find that the only equilibrium distribution is

$$\pi = (60/517, 115/517, 160/517, 182/517) \approx (0.116054, 0.222437, 0.309478, 0.352031).$$

This is what the rows of $\mathbf{P}^n$ converged to in Example 2.1.

Example 2.10. Consider again Example 2.2. To find equilibrium distributions we again solve $(\mathbf{I} - \mathbf{P})^T\pi^T = [0]$ to find a 2-parameter family of solutions:

$$\pi = (\beta, 0, 0, 0, \alpha, 0).$$

Not all of these are probability distributions. To be a probability distribution we need $\alpha \geq 0$, $\beta \geq 0$ and $\alpha + \beta = 1$. So the equilibrium distributions are

$$\pi = (1 - \alpha, 0, 0, 0, \alpha, 0) \text{ for } 0 \leq \alpha \leq 1.$$

In Example 2.2 we exhibited $\mathbf{P}^{100}$. Observe that every row was of this form for some α.

Example 2.11. Consider Example 2.8 again. The solutions of $(\mathbf{I} - \mathbf{P})^T\pi^T = [0]$ are $\pi = (c, c, c, 0, 0, 0)$ for any constant c. There is only one equlibrium distribution:

$$\pi = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3}, 0, 0, 0).$$

If you explore $\mathbf{P}^n$ however you will find that it does *not* converge as $n \to \infty$; see Problem 2.2.

2.4.1 Existence and Uniqueness of Equilibrium Distributions

We first want to prove that equilibrium distributions always exist (assuming a finite state space). This follows from a simple real analysis argument. Let K be the set of all probability distributions on $\mathcal{S} = \{1, \ldots, m\}$:

$$K = \{\mu = (\mu_i) \in \mathbb{R}^m : \mu_i \geq 0 \text{ and } \sum_1^m \mu_i = 1\}.$$

It should be clear that K is closed and bounded and therefore a compact subset of $\mathbb{R}^m$. Start with any $\mu \in K$. Define a sequence in K by

$$\pi^{(k)} = \frac{1}{k} \sum_{n=1}^{k} \mu \mathbf{P}^n.$$

You can check that each $\pi^{(k)}$ is again in K: $\pi^{(k)}[1] = [1]$ and $\pi^{(k)} \geq 0$. Since K is compact there must be a convergent subsequence:

$$\pi^{(k')} \to \pi \in K.$$

Now observe that after cancellation of common terms we have

$$\pi^{(k)} - \pi^{(k)}\mathbf{P} = \frac{1}{k}(\mu\mathbf{P} - \mu\mathbf{P}^{k+1}).$$

All terms of the right are bounded by $\frac{1}{k}$. Therefore $\pi^{(k)} - \pi^{(k)}\mathbf{P} \to [0]$ as $k \to \infty$. It follows that

$$\pi = \lim \pi^{(k')} = \lim \pi^{(k')}\mathbf{P} = \pi\mathbf{P},$$

proving that π is an equilibrium distribution. This proves the first part of our next theorem.

Theorem 2.7. *Suppose* $\mathbf{P}$ *is the transition matrix for a Markov chain on a finite state space* $\mathcal{S}$*. There exists an equilibrium distribution* π*. Every equilibrium distribution vanishes on all non-closed communication classes. The equilibrium distribution is unique if and only if there is only one closed communication class.*

In preparation for the rest of the proof we establish a couple lemmas. We separate the states into two disjoint subsets: $\mathcal{S} = R \cup T$ where

$$R = \{i \in \mathcal{S} : \text{ the communication class of } i \text{ is closed}\}$$
$$T = \{i \in \mathcal{S} : \text{ the communication class of } i \text{ is not closed}\}.$$

R is the set of recurrent states and T is the set of transient states. An important observation is that

$$p_{i,j}(n) = 0 \text{ for all } n \geq 0,\ i \in R,\ j \in T. \tag{2.13}$$

This is because if $i \rightsquigarrow j$ and the class of i is closed then j would be in the same class as i. But the class of j is not closed since $j \in T$.

Lemma 2.8. *Suppose* X_n *is a Markov chain on a finite state space* $\mathcal{S}$*. For all* $i, j \in T$

$$p_{i,j}(n) \to 0 \text{ as } n \to \infty.$$

Proof. We will show that $P_i(X_n \in T) \to 0$. This will imply the lemma since $p_{i,j}(n) \leq P_i(X_n \in T)$ for every

$j \in T$. The first step is to observe that $P_i(X_n \in T)$ is decreasing in n.

$$\begin{aligned} P_i(X_{n+m} \in T) &= \sum_{j \in T} p_{i,j}(n+m) \\ &= \sum_{j \in T} \sum_{k \in \mathcal{S}} p_{i,k}(n) p_{k,j}(m) \\ &= \sum_{j \in T} \sum_{k \in T} p_{i,k}(n) p_{k,j}(m), \text{ by (2.13)} \\ &= \sum_{k \in T} p_{i,k}(n) \sum_{j \in T} p_{k,j}(m) \\ &\leq \sum_{k \in T} p_{i,k}(n) \\ &= P_i(X_n \in T). \end{aligned}$$

Next, if $i \in T$ there must be $j \in R$ for which $i \rightsquigarrow j$. This is because $\{j : i \rightsquigarrow j\}$ is a closed set of states, and so by Lemma 2.2 must contain a closed communication class, which must be contained in R. For this $j \in R$ there exists n_i with $p_{i,j}(n_i) > 0$ and so $P_i(X_{n_i} \in T) < 1$. By the monotonicity we just proved we know $P_i(X_n \in T) \leq P_i(X_{n_i} \in T) < 1$ for all $n \geq n_i$. Doing this for each $i \in T$ and taking $m = \max_{i \in T}(n_i)$ we see that there is $\epsilon > 0$ so that

$$\sum_{j \in T} p_{i,j}(m) < 1 - \epsilon \text{ for all } i \in T.$$

Therefore

$$P_i(X_{n+m} \in T) = \sum_{j \in T} p_{i,j}(n) P_j(X_m \in T) \leq (1 - \epsilon) P_i(X_n \in T).$$

We can now conclude that $\lim_n P_i(X_n \in T) = 0$ as desired. □

The next lemma is the key to proving uniqueness in Theroem 2.7.

Lemma 2.9. *Suppose* $\mathbf{P}$ *is the transition matrix for an irreducible Markov chain and* $\nu \in \mathbb{R}^n$ *is nonzero with* $\nu = \nu\mathbf{P}$*. Then either* $\nu_i > 0$ *for all* i *or* $\nu_i < 0$ *for all* i*.*

Proof. Let

$$A = \{i : \nu_i > 0\} \text{ and } B = \{j : \nu_j < 0\}.$$

We suppose that both A and B are nonempty, and will see that that leads to a contradiction. Let $f(i) = 1$ if $i \in A$ and $f(j) = 0$ otherwise. Since $\nu = \nu\mathbf{P}^m$ for any m we have $\nu f = \nu\mathbf{P}^m f$. Pick m so that for some $j \in B$, $k \in A$ we have $p_{j,k}(m) > 0$. This is possible since $\mathbf{P}$ is irreducible.

For $i \in A$ we have

$$\sum_k p_{i,k}(m) f(k) \leq \sum_k p_{i,k}(m) 1 = 1 = f(i).$$

For $j \in B$ we have

$$\sum_k p_{j,k}(m) f(k) \geq 0,$$

and by our choice of m this is strictly positive for at least one such j. Using these inequalities we have

$$\begin{aligned}\sum_{i\in A}\nu_i &= \nu f\\ &= \nu\mathbf{P}^m f\\ &= \sum_{i\in A}\nu_i\left[\sum_k p_{i,k}^{(m)} f(k)\right] + \sum_{j\in B}\nu_j\left[\sum_k p_{j,k}^{(m)} f(k)\right]\\ &\le \sum_{i\in A}\nu_i + \sum_{j\in B}\nu_j\left[\sum_k p_{j,k}^{(m)} f(k)\right]\\ &< \sum_{i\in A}\nu_i,\end{aligned}$$

this last inequality because all the terms in the $j\in B$ sum are nonpositive and at least one of them is strictly negative. This contradiction proves that either A or B is empty.

Assume A is nonempty, with some $i'\in A$. By the above B is empty, so $\nu_j\ge 0$ for all k. Consider any k. Because $\mathbf{P}$ is irreducible there is an m with $p_{i',k}^{(m)}>0$. So

$$\nu_k = \sum_i \nu_i p_{i,k}^{(m)} \ge \nu_{i'} p_{i',k}^{(m)} > 0.$$

Thus all k belong to A. The case of B nonempty is analogous. □

Proof of Theorem 2.7. We have already proven that an equilibrium distribution exists. Suppose π is any equilibrium distribution. Since $\pi=\pi\mathbf{P}^n$ we have

$$\pi_j = \sum_{i\in\mathcal{S}} \pi_i p_{i,j}(n).$$

As before, let T be the set of transient states, i.e. those whose communication classes are not closed. If $j\in T$ we can limit the sum to $i\in T$ because $p_{i,j}(n)=0$ for $i\in R$. Therefore for $j\in T$

$$\pi_j = \sum_{i\in T} \pi_i p_{i,j}(n).$$

By Lemma 2.8 $p_{i,j}(n)\to 0$ for $i,j\in T$. Therefore $\pi_j=0$ for all $j\in T$.

We turn now to the uniqueness assertion. Suppose C is a closed communication class. For $i\in C$ we know that $p_{i,j}=0$ for all $j\notin C$. This means we can consider X_n as a Markov chain on C, i.e. with C as the entire state space. So there exists an equilibrium distribution π^C consisting of π_i^C for $i\in C$. Extend this to a distribution on the full state space by

$$\pi_j = \begin{cases}\pi_j^C & \text{if } j\in C\\ 0 & \text{if } j\notin C.\end{cases}$$

We claim that this is a stationary distribution on the full $\mathcal{S}$. For $j\in C$

$$\sum_{i\in\mathcal{S}}\pi_i p_{i,j} = \sum_{i\in C}\pi_i^C p_{i,j} = \pi_j^C = \pi_j.$$

For $j\notin C$ we know so that

$$\sum_{i\in\mathcal{S}}\pi_i p_{i,j} = \sum_{i\in C}\pi_i^C p_{i,j} = 0 = \pi_j,$$

because $p_{i,j}=0$ if $i\in C$ and $j\notin C$. This shows that π is indeed an equilibrium distribution which is nonzero only for the states in C. If there is a second (disjoint) closed communication class $\tilde{C}$ then there is another equilibrium distribution $\tilde{\pi}$ that arises in the same way but using $\tilde{C}$. Clearly $\pi\ne\tilde{\pi}$ since one vanishes

on C and the other does not. So if there is more than one closed communication class then equilibrium distributions are not unique.

Finally suppose C is the only closed communication class and that both π and $\tilde{\pi}$ are equilibrium distributions. Then as shown above they both vanish outside C; they both can be considered as equilibrium distributions for the Markov chain on the reduced state space C. On C the chain is irreducible so Lemma 2.9 applies. Let $\nu = \pi - \tilde{\pi}$. Since $\nu\mathbf{P} = \nu$ we conclude that either $\pi_i \leq \tilde{\pi}_i$ for all $i \in C$ or $\tilde{\pi}_i \leq \pi_i$ for all $i \in C$. But since $\sum_{i\in C} \pi_i = 1 = \sum_{i\in C} \tilde{\pi}_i$ it follows that $\pi = \tilde{\pi}$, proving uniqueness. □

2.4.2 Convergence of $\mathbf{P}^n$

We showed on page 19 that *if* $\mu\mathbf{P}^n \to \pi$ then π is an equilibrium distribution. Now we want to consider whether or not $\mu\mathbf{P}^n$ really does converge. Problem 2.2 is one example in which it can fail to converge (depending on μ). If there are multiple closed communication classes Theorem 2.7 implies that this limit can have different values depending on μ. We will assume that the chain is irreducible, eliminating that complication. The feature that can prevent convergence is periodicity. We are going to show that for irreducible *aperiodic* chains, $\mu\mathbf{P}^n \to \pi$, no matter what μ is.

Theorem 2.10. *Suppose* $\mathbf{P}$ *is the transition matrix for an irreducible, aperiodic Markov chain on a finite state space* $\mathcal{S}$*. For any initial distribution* μ*,* $\mu\mathbf{P}^n \to \pi$ *as* $n \to \infty$*, where* π *is the unique equilibrium distribution for* $\mathbf{P}$*. In particular*

$$\mathbf{P}^n \to \Pi = \begin{bmatrix} \cdots \pi \cdots \\ \vdots \\ \cdots \pi \cdots \end{bmatrix},$$

the matrix Π *which has* π *as each of its rows.*

This theorem is why we can get a good approximation to π by computing $\mathbf{P}^n$ for a large n, like we did in Example 2.1.

Proof. Let π be the equilibrium distribution (which exists by Theorem 2.7) and μ any initial distribution. Let $\mu^{(n)} = \mu\mathbf{P}^n$. We know

$$\mu_j^{(n+1)} - \pi_j = \sum_i (\mu_i^{(n)} - \pi_i) p_{i,j}$$

and therefore

$$\begin{aligned} |\mu_j^{(n+1)} - \pi_j| &\leq \sum_i |\mu_i^{(n)} - \pi_i| p_{i,j} \\ \sum_j |\mu_j^{(n+1)} - \pi_j| &\leq \sum_j \sum_i |\mu_i^{(n)} - \pi_i| p_{i,j} \\ &= \sum_i \sum_j |\mu_i^{(n)} - \pi_i| p_{i,j} \\ &= \sum_i |\mu_i^{(n)} - \pi_i|. \end{aligned}$$

This shows that $\sum_i |\mu_i^{(n)} - \pi_i|$ is monotone nonincreasing as $n \to \infty$ and therefore has a nonnegative limit $L = \lim_n \sum_i |\mu_i^{(n)} - \pi_i|$.

The above applies for any Markov chain and any equilibrium distribution π. We want to show that for an irreducible aperiodic chain the limit L is actually $L = 0$. From Corollary 2.5 we know that for such a chain there is a positive $\alpha > 0$ and integer N so that

$$p_{i,j}(N) \geq \alpha \text{ for all } i, j.$$

This implies $0 \le \sum_j [p_{i,j}(N) - \alpha] = 1 - m\alpha < 1$, were $m = \#\mathcal{S}$. Observe that

$$\begin{aligned}\mu_j^{(n+N)} - \pi_j &= \sum_i (\mu_i^{(n)} - \pi_i) p_{i,j}(N) \\ &= \sum_i (\mu_i^{(n)} - \pi_i)[p_{i,j}(N) - \alpha + \alpha] \\ &= \sum_i (\mu_i^{(n)} - \pi_i)[p_{i,j}(N) - \alpha],\end{aligned}$$

because $\alpha \sum_i (\mu_i^{(n)} - \pi_i) = 0$. Now proceed as before.

$$\begin{aligned}|\mu_j^{(n+N)} - \pi_j| &\le \sum_i |\mu_i^{(n)} - \pi_i| [p_{i,j}(N) - \alpha] \\ \sum_j |\mu_j^{(n+N)} - \pi_j| &\le \sum_j \sum_i |\mu_i^{(n)} - \pi_i| [p_{i,j}(N) - \alpha] \\ \sum_j |\mu_j^{(n+N)} - \pi_j| &\le \sum_j \sum_i |\mu_i^{(n)} - \pi_i| [p_{i,j}(N) - \alpha] \\ &= \sum_i \sum_j |\mu_i^{(n)} - \pi_i| [p_{i,j}(N) - \alpha] \\ &= (1 - m\alpha) \sum_i |\mu_i^{(n)} - \pi_i|.\end{aligned}$$

It follows from this that $L \le (1 - m\alpha)L$ and therefore $L = 0$. This in turn implies that $\mu_i^{(n)} \to \pi_i$ for each i. □

2.4.3 Eigenvalues of P

This section examines the eigenvectors of $\mathbf{P}$ in more detail. It is elementary that $\lambda = 1$ is an eigenvalue because we can exhibit a (right) eigenvector:

$$\mathbf{P}[1] = 1[1].$$

We call [1] a *right* eigenvector because we are multiplying it on the right side of $\mathbf{P}$, as is customary in discussions of eigenvectors. An equlibirium distribution is a *left* eigenvector for the eigenvalue $\lambda = 1$ because we multiply it on the left of $\mathbf{P}$:

$$\pi\mathbf{P} = 1\pi.$$

This is the same as saying π is a right eigenvector of $\mathbf{P}^T$: $\mathbf{P}^T\pi^T = 1\pi^T$. The eigenvalues and characteristic polynomial are the same for $\mathbf{P}$ and $\mathbf{P}^T$ but the eigenvectors are not.

If λ is any eigenvalue (possibly complex) and $\mathbf{v} \ne [0]$ is a corresponding (right) eigenvector then for any positive power n we have

$$\mathbf{P}^n\mathbf{v} = \lambda^n\mathbf{v}.$$

All entries of $\mathbf{P}^n$ are between 0 and 1 so the left side of this is bounded. So the right side is also bounded. That means $|\lambda| \le 1$ since otherwise $|\lambda^n| \to \infty$ as $n \to \infty$ and the right side would be unbounded. Thus all eigenvalues are bounded (in complex modulus) by 1:

$$|\lambda| \le 1.$$

What about the multiplicity of the eigenvalue $\lambda = 1$? By that we mean the power m to which $(\lambda - 1)$ appears in the factorization of the characteristic polynomial:

$$p(\lambda) \doteq \det(\lambda\mathbf{I} - \mathbf{P}) = (\lambda - 1)^m \cdots .$$

First let's consider the irreducible case. If $\mathbf{v}$ is an eigenvector with eigenvalue 1 then $\mathbf{v} = \mathbf{P}\mathbf{v}$. Let $\beta = \max v_i$. It follows from Lemma 2.1 that $\{i : v_i = \beta\}$ is a closed set of states. If the chain is irreducible this must be

all of $\mathcal{S}$ since $\mathcal{S}$ is itself the only closed set. Therefore $\mathbf{v} = \beta[1]$. In other words there is only one linearly independent eigenvector for $\lambda = 1$. Could it be that $m > 1$ even though there is only one independent eigenvector? For matricies in general yes, that is possible. But if so then it turns out that there must be a second *generalized* vector $\mathbf{u} \neq [0]$ for which $(\mathbf{I} - \mathbf{P})\mathbf{u} = \mathbf{v}$. (This is part of the Jordan basis for $\mathbf{P}$. For a discussion look for a treatment of the Jordan canonical form in an intermediate level linear algebra text such as Lang [38].) Rearranging, $\mathbf{P}\mathbf{u} = \mathbf{u} - \mathbf{v}$. Iterating this we find that

$$\mathbf{P}^n\mathbf{u} = \mathbf{u} - n\mathbf{v}.$$

But this is not possible for a transition matrix because the left side is bounded in n but the right side is not. This shows that (in the irreducible case) the multiplicity can *not* be $m > 1$. I.e. the factorization of the characteristic polynomial $p(\lambda)$ has only a single factor of $(\lambda - 1)$.

In the general (not irreducible) case if there are m closed communication classes then there will be exactly m factors of $(\lambda - 1)$ in $p(\lambda)$, and m linearly independent eigenvectors. We saw this emerging in the uniqueness part of the proof of Theorem 2.7, where there was a different equilibrium distribution π^C for each closed class C. It turns out that those form a basis for the vector space of all solutions to $\nu(\mathbf{P} - \mathbf{I}) = (0)$, although we will skip the details of a complete argument.

Example 2.12. Consider again Example 2.10. The characteristic polynomial has exactly two factors of $(\lambda - 1)$:

$$\begin{aligned} p(\lambda) &= \lambda^6 - \frac{13\lambda^5}{4} + \frac{15\lambda^4}{4} - \frac{13\lambda^3}{8} - \frac{\lambda^2}{32} + \frac{3\lambda}{16} - \frac{1}{32} \\ &= (\lambda - 1)^2 q(\lambda) \end{aligned}$$

where

$$q(\lambda) = \lambda^4 - \frac{5\lambda^3}{4} + \frac{\lambda^2}{4} + \frac{\lambda}{8} - \frac{1}{32}.$$

(Since $q(1) = 3/32$ there are no additional factors of $(\lambda - 1)$.) This is consistent with our observations of this section since there are two closed communication classes, and a two-dimensional eigenspace.

What about other eigenvalues (complex) with $|\lambda| = 1$? This depends on the existence of periodic closed classes. For simplicity assume the chain is irreducible, so that there is only one closed class to consider. If the chain is aperiodic then we know from Theorem 2.10 that $\mathbf{P}^n \to \Pi$. Suppose $|\lambda| = 1$ and $\mathbf{v}$ is a (right) eigenvector. Then

$$\lambda^n\mathbf{v} = \mathbf{P}^n\mathbf{v} \to \Pi\mathbf{v} = c[1]$$

where $c = \sum_i \pi_i v_i$. Then λ^n itself must be convergent, which is only possible if $\lambda = 1$. If however the chain has positive period $k > 1$ then the k^{th} roots of unity will be eigenvectors. This holds in the general case, closed class by closed class, but we again omit the details and just look at an example.

Example 2.13. Consider Example 2.8 above. This has one closed class of period 3. The characteristic polynomial works out to be

$$\begin{aligned} p(\lambda) &= \lambda^6 - 0.25\lambda^5 - 0.2\lambda^4 - 0.95\lambda^3 + 0.25\lambda^2 + 0.2\lambda - 0.05 \\ &= (\lambda^3 - 1)q(\lambda) \\ &= (\lambda - 1)\left(\lambda - \frac{i\sqrt{3}}{2} + \frac{1}{2}\right)\left(\lambda + \frac{i\sqrt{3}}{2} + \frac{1}{2}\right)q(\lambda), \end{aligned}$$

where

$$q(\lambda) = \lambda^3 - 0.25\lambda^2 - 0.2\lambda + 0.05.$$

We see that the three cube roots of unity, $1, \frac{1 \pm i\sqrt{3}}{2}$ are the eigenvalues with $|\lambda| = 1$.

2.5 An Example: Google Page Rank

Many details of how Google's search engine works are closely guarded company secrets. However the founders, S. Brin and L. Page, have published a general description (see the citations in [39]) so the basic method

is known. When you do a search Google uses the extensive catalog of web pages produced by its crawlers to produce a list of web pages relevant to your search criteria. Then it sorts that list using an importance score that it has calculated for each web page. That importance score is called its *PageRank*. The PageRank scores of web pages have nothing to do with your particular search. They are pre-computed (and updated regularly) so that they are available when needed for a search. We want to talk about how PageRank works because it has a natural interpretation in terms of Markov chains.

Think of the internet as a huge directed graph. Each node in the graph is a web page and each edge $i \to j$ represents a link from page i to page j. PageRank tries to use *just this link information* and nothing more to assign an importance score $r(i)$ to each web page i. The basic idea is that the ranking $r(m)$ of a web page m should be determined by

- the ranking $r(i)$ of those sites which link to m: $i \to m$
- but the contribution of $r(i)$ to the ranking of m is diluted if i also links to lots of other sites j.

A quantitative expression of this is

$$r(m) = \sum_{i:i\to m} \frac{r(i)}{\mathcal{O}(i)}$$

where $\mathcal{O}(i) = \#\{j : i \to j\}$ is the number of out-links from i. Observe that the above is an expression of the form

$$r(m) = \sum_i r(i)p_{i,m} \text{ where } p_{i,m} = \begin{cases} \frac{1}{\mathcal{O}(i)} & \text{if } i \to m \\ 0 & \text{if } i \nrightarrow m. \end{cases}$$

Also note that $0 \le p_{i,m}$ and $\sum_m p_{i,m} = 1$ (unless there are no outlinks from i). Thus $\mathbf{P} = [p_{i,m}]$ is the transition matrix of a Markov chain (provided each page has at least one out-link) and the rankings $r(\cdot)$ are a (left) einvector. They would form an equilibrium distribution if normalized so that $\sum_m r(m) = 1$. We will call $\mathbf{P}$ the *raw Google matrix*. A Markov chain with transition matrix $\mathbf{P}$ moves from web page $X_n = i$ by randomly picking one of the outgoing links $i \to j$ (all being equally likely) and following it to $X_{n+1} = j$.

For several reasons a modified transition matrix $\tilde{\mathbf{P}}$ is used in place of $\mathbf{P}$ for calculating the PageRank stationary distribution $r = r\tilde{\mathbf{P}}$. The size of $\tilde{\mathbf{P}}$ is massive (over 8 billion $\times$ 8 billion in 2005) so the calculation of r is a significant challenge. The basic algorithm is to use an iteration: $r^{(k+1)} = r^{(k)}\tilde{\mathbf{P}}$. Getting this to converge quickly is of paramount importance, as are methods to parallelize or otherwise speed up the calculation. This is one purpose for modifying $\mathbf{P}$.

Another issue is that some pages have no out-links: $\mathcal{O}(i) = 0$. For such i we have $p_{i,m} = 0$ for all m so $\mathbf{P}$ is not a proper transition matrix. To fix that we can simply use $\frac{1}{n}(1, \ldots, 1)$ for row i in a preliminary modification $\bar{\mathbf{P}}$, n being the size of $\mathbf{P}$.

$$\bar{p}_{i,m} = \begin{cases} \frac{1}{\mathcal{O}(i)} & \text{if } i \to m \\ 0 & \text{if } i \nrightarrow m \text{ and } \mathcal{O}(i) > 0 \\ \frac{1}{n} & \text{if } \mathcal{O}(i) = 0. \end{cases}$$

This $\bar{\mathbf{P}}$ is well-defined and a valid transition matrix. But there is still a problem because it may not produce an irreducible chain. There can be many closed communication classes. As a consequence,

- $\bar{\mathbf{P}}$ does not have a unique equilibrium distribution,
- the limit of $r^{(0)}\bar{\mathbf{P}}^k$ depends on $r^{(0)}$,
- any equilibrium distribution r will have $r(i) = 0$ for i not in closed classes, so these web pages would get a ranking of 0.

Google's solution to this is to use a *personalization vector* v:

$$0 < v_i, \ \sum_i v_i = 1.$$

Form the matrix $\mathbf{V}$ which uses v for each of its rows, and then take

$$\tilde{\mathbf{P}} = \alpha\bar{\mathbf{P}} + (1-\alpha)\mathbf{V},$$

using a parameter $0 < \alpha < 1$. This $\tilde{\mathbf{P}}$ is still a valid transition matrix, called the *Google matrix*. Using the Google matrix instead of $\bar{\mathbf{P}}$ has several advantages.

- $\tilde{p}_{i,j} > 0$ for all i, j so the matrix is irreducible, aperiodic, so there is a unique equilibrium distribution r (but it depends on v and α). Moreover $r^{(0)}\tilde{\mathbf{P}}^k$ is theoretically guaranteed to converge as $k \to \infty$ by Theorem 2.10.
- The associated Markov chain, if at state i, makes a $\bar{\mathbf{P}}$ transition with probability $1-\alpha$ but a transition using the probabilities from v with probability α. This is going to increase the ranking of those pages with larger v_i values and decrese those with smaller v_i values, giving Google a way to modify the rankings in accord with "commercial considerations" (i.e. paid advertising). The strength of this modification is controlled by the value of α.
- The effect of α on the eigenvalues of $\tilde{\mathbf{P}}$ is predictable. If the eigenvalues of $\bar{\mathbf{P}}$ are $1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ then the eigenvalues of $\tilde{\mathbf{P}}$ will be $1 > \alpha\lambda_2 \geq \cdots \geq \alpha\lambda_n$. This is significant because the rate of convergence of $r^{(0)}\tilde{\mathbf{P}}^k$ depends on the second largest eigenvalue $\alpha\lambda_2$. The smaller this is the faster the iteration will converge. So there is a tradeoff in determining the value to use for α. Smaller α will allow the stationary distribution r to be calculated faster, but will also make r more influenced by v and less by the actual structure of the internet. (It is reported that at least initially Google used something like $\alpha = .85$.)
- The matrix $\tilde{\mathbf{P}}$ does not actually need to be stored in computer memory. The $\tilde{\mathbf{P}}$ iteration can be done in terms of $\bar{\mathbf{P}}$.

$$\begin{aligned} r^{(k+1)} &= r^{(k)}\tilde{\mathbf{P}} \\ &= r^{(k)}(\alpha\bar{\mathbf{P}} + (1-\alpha)V) \\ &= \alpha r^{(k)}\bar{\mathbf{P}} + (1-\alpha)v, \end{aligned}$$

  because $\sum_i r^{(k)}(i) = 1$ implies that $r^{(k)}\mathbf{V} = v$.

Problems

Problem 2.1
Let X_n be the Markov Chain on $\mathcal{S} = \{1, 2, 3\}$ with

$$\mathbf{P} = \begin{bmatrix} \frac{1}{2} & \frac{1}{3} & \frac{1}{6} \\ 0 & \frac{1}{3} & \frac{2}{3} \\ \frac{1}{2} & 0 & \frac{1}{2} \end{bmatrix}.$$

Let the initial distribution be

$$P(X_0 = 1) = \frac{1}{4},\ P(X_0 = 2) = \frac{1}{2},\ P(X_0 = 2) = \frac{1}{4}.$$

a) Calculate $P(X_3 = 2)$.

b) Calculate $P(X_n \neq 3 \text{ for all } n \leq 5)$, i.e. the probability that X_n avoids 3 from $n = 0$ to $n = 5$.

..... Avoid

Problem 2.2
Using the transition matrix from Example 2.8 and $\mu = (\frac{1}{3}, \frac{1}{6}, \frac{1}{6}, 0, \frac{1}{3}, 0)$ calculate $\mu\mathbf{P}^n$ for several successive

large values of n. (On the order of a hundred is large enough.) Does it appear that $\mu\mathbf{P}^n$ converges as $n \to \infty$? If so what does it converge to? If not is there any discernable pattern as $n \to \infty$? Is there a different initial distribution μ for which $\mu\mathbf{P}^n$ does converge?

3class

Problem 2.3
In Example 2.1 calculate mean time to reach state 2 for each initial state.

MeanHit2

Problem 2.4
In Feller's Breeding Problem, Example 2.2, calculate the mean number of steps $E_i[\mathcal{T}_{\{1,5\}}]$ to reach one of the two absorbing states for each initial state i.

FellerT

Problem 2.5
A gambler is betting on the outcomes of succesive (fair) coin tosses. He starts with \$2 and wants to reach \$10. On each coin toss he wagers the smaller of what he currently has and the amount he needs to reach \$10. He cannot wager more than he currently has, but also does not wager more than necessary to reach \$10. (For instance if he has \$2 he wagers \$2 but if he has \$6 he wagers \$4.) If he wins the coin toss he receives back his wager plus an equal amount but if he looses the toss he looses his wager. This can be considered as a Markov Chain on $\mathcal{S} = \{0, 2, 4, 6, 8, 10\}$. Work out the transition probabilities and then answer these questions. What is the probability that the gambler will reach his goal of \$10 before going broke? What is the mean number of coin tosses until he either succeeds or goes broke? (Taken from Norris [45].)

Gamble

Problem 2.6
Let X_n be the symmetric random walk on $\{0, \dots m\}$ with absorption at 0 and m. Show that for any initial position $0 < X_0 < m$ the probability that X_n is eventually absorbed at either 0 or m is 1.

NoAbsorb

Problem 2.7
You are playing a simple board game with N spaces arranged in a circle: $1 \to 2 \to \cdots \to N \to 1$. You start at space 1, roll a (single) dice and move ahead that many spaces. For $N = 10$ calculate the mean number of rolls until you return to 1 for the first time.

Loop10

Problem 2.8
Derive equation (2.10) from $v(i) = \sum_n nw(i, n)$ and the equations for $w(i, n)$.

MeanHit

Problem 2.9
Equations (2.9) tell us how to calculate $w(i, n) = P_i(\mathcal{T}_C = n)$. In this problem we want to consider the distribution of $\mathcal{T}_C^+$:

$$w^+(j, n) = P_i(\mathcal{T}_C^+ = n).$$

For $X_0 = i \in B$ we know that $\mathcal{T}_C = \mathcal{T}_C^+$, so $w(i, n) = w^+(i, n)$. Suppose $j \in C$. Explain why $w^+(j, 0) = 0$,

$$w^+(j, 1) = \sum_{k \in C} p_{i,k}, \text{ and}$$
$$w^+(j, n + 1) = \sum_{i \in B} p_{j,i} w(i, n) \text{ for } n \geq 1.$$

In other words, $\mathbf{w}_C^+(1) = \mathbf{P}_{CC}[1]$ and $\mathbf{w}_C^+(n+1) = \mathbf{P}_{CB}\mathbf{w}_B(n)$ for $n \geq 1$. Also explain why

$$E_j[\mathcal{T}_C^+] = 1 + \sum_{i \in \mathcal{S}} p_{j,i} v(i),$$

where $v(i) = E_i[\mathcal{T}_C]$.

. . . . . MeanRetPlus

Problem 2.10
Confirm that the expressions for $v(i)$ given in Example 2.6 really do satisfy the equations as claimed.

. . . . . RWmeanA

Problem 2.11
Chutes & Ladders is a well-known children's board game[1]. At each turn you roll a dice (actually the game comes with a 6-position spinner, not a dice) which determines how many spaces you move. If you land at the bottom of a ladder you immediately move your position to the ladder's top, and if you land at the top of a chute you immediately move your piece to the chute's bottom. Your first task is to produce (in MATLAB) the transition matrix for this as a Markov chain. The file `gp6.m` is a script which will generate the 100x100 transition matrix if there were no chutes or ladders. You need to decide how to modify that to get the correct transition matrix. (Some columns will be all 0s.)

Use that to compute the average number of turns it takes to reach the finish position at #100. What is the mean time to finish? Calculate the probability of reaching the Finish position (for the first time) on the n^{th} play for enough values of n to determine which time n is the most likely time to finish. Compute the probability of finishing without ever using the big chute (87 to 24).

. CL

Problem 2.12
Consider the Markov chain on $S = \{1, 2, 3, 4\}$ with transition matrix

$$\mathbf{P} = \begin{bmatrix} \frac{1}{3} & \frac{2}{3} & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{4} & 0 & \frac{1}{4} & \frac{1}{2} \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

a) Find the transient and recurrent states and identify the irreducible classes.

b) For each $i \in S$ determine the value of $\lim_{n\to\infty} p_{i1}(n)$, and explain.

. . . . . P2

Problem 2.13
Let X_n be an irreducible aperiodic Markov Chain on $\{1, 2, \ldots m\}$ with transition probabilities $p_{i,j}$ and stationary distribution π. Let $\mathcal{D} = \{(i,j) : p_{i,j} > 0\}$ and let $Y_n = (X_{n-1}, X_n)$ $(n = 1, 2, \ldots)$. In other words Y_n records the most recent transition of X_n.

a) Explain why Y_n is also a Markov chain and determine its transition probabilities $p_{(i,j),(k,\ell)}$.

b) Show that Y_n is also irreducible and aperiodic.

c) If π_i are the equilibrium probabilities for X_n, what are the equilibrium probabilities $\pi_{(i,j)}$ for Y_n? Verify that your formula for $\pi_{(i,j)}$ satisfies the equilibrium equation

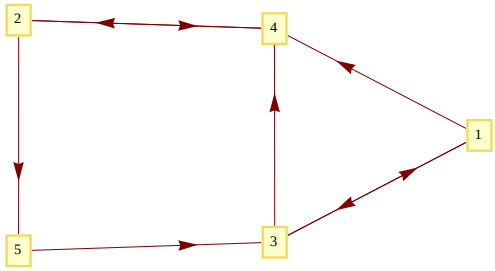
$$\pi_{(k,\ell)} = \sum_{(i,j)} \pi_{(i,j)} p_{(i,j),(k,\ell)}.$$

[1] You can see a picture of the playing board at

`http://myprettypennies.com/wp-content/uploads/2011/06/chutesladders.gif`

Pairs

Problem 2.14
Consider the "toy" internet consisting of five web pages with links as illustrated.



a) Before you do any calculation, just based on your visual inspection speculate which pages you think should have the highest and lowest rankings. (There is no right answer here.)

b) With no modification by means of a personalization vector ($\alpha = 1$) find the page ranks of each of the pages.

c) The owners of page #1 are not happy that page #2 has a higher ranking. They negotiate to have the page rankings recalculated using a personalization vector $\mathbf{v} = (1/3, 1/6, 1/6, 1/6, 1/6)$. How small would α need to be in order for page #1 to have a higher ranking than page #2.

HW6A

For Further Study

There are *many* books on the material of this chapter. Feller [22] is a classic and Karlin [33] is also a well-known standard. Some written at an introductory level are Norris [45], Ross [52], Hoel, Port, & Stone [28], and Lawler [40]. Most books treat the finite and infinite state cases at the same time and so also cover the material of our Chapter 4. One text that is exclusively about the finite state case is Kemeny & Snell [34]. The properties of the eigenvalues of $\mathbf{P}$ are part of the Perron-Frobenius theory of positive matricies, which you can find in Berman and Plemmons [4] and the appendix to Karlin [33].

Section 2.5 is based on the 2005 paper by A.Langville an C. Meyer [39], which would be a good place to start if you want to lean more about that.

Chapter 3

Basics of Probability Theory

The problems of the previous chapter all boiled down to calculations with matricies and systems of linear equations. In Chapter 4 will consider Markov chains with (countably) *infinite* state spaces. Sums over all possible states with now be infinite series. In some cases we will need to work with limits of probabilities and random variables. We will need to have a clearer idea of what random variables actually are and how to work with their expected values. This chapter provides an introduction to the systematic mathematical framework used to study probability theory and the properties we will need as we continue. We will describe the idea of independence and some of the most important theorems about sequences of independent random variables. We will introduce conditional probabilities and conditional expectations, essential tools for working with stochastic processes. Many of these are more difficult mathematical issues than you might expect. A full treatment of these things requires a graduate level course in measure theory. (See "For Further Study" at the end of the chapter for a couple references.) Our discussion here is *only* intended to give us the working knowledge that we will need in the rest of this book, not a fully justified rigorous development. There are a few technical details which we will simply overlook because they do not affect our use of these things and would only be an encumbrance to our discussion.

3.1 Infinite Sequences and the Kolmogorov Model

Our basic description of a Markov chain in Section 2.1 prescribed how the transition matrix $\mathbf{P}$ and initial distribution μ determine the probability that over a prescribed time period ($t = 0, 1, \ldots, n$) the Markov chain $X_{0:n}$ produces one particular finite sequence of states $s_{0:n}$; see (2.2). For the probability of something that can happen in multiple ways we added up the individual probabilities of all the specific ways it can be achieved. That is how we determined $P(X_n = a)$ for a given state $a \in \mathcal{S}$ and time n; see (2.3). This is based on a presumption that one probability can be obtained as a certain sum of other probabilities. In Section 2.2 we considered probabilities such as $P(X_n = a \text{ for some } n)$ which are even more complex because the event in question may happen at many different times n, not just a single n. We did develop a way to calculate some probabilities like this by solving an certain matrix equation. But how does all this relate back to (2.2)? We presumed some properties which we didn't really think about then, but will now.

The full outcome of a Markov chain is an *infinite* sequence of states: $s_0 = X_0$, $s_1 = X_1$, ..., $s_k = X_k$, ... with each $s_k \in \mathcal{S}$. Let Ω be the set of all infinite sequences $\omega = (s_0, s_1, \ldots)$ of states $s_k \in \mathcal{S}$. Thus each $\omega \in \Omega$ is a particular sequence, $\omega = (s_0, s_1, \ldots)$, or $\omega = s_{0:\infty}$ for short. When we contemplate the probability that $X_0 = 0, X_1 = 5, X_2 = 15, X_3 = 19$ we are considering the set $C \subseteq \Omega$ of those sequences with the prescribed values in the first four positions:

$$C = \{\omega \in \Omega : \omega = (0, 5, 15, 19, \cdots), \text{ the terms coming after 19 allowed to be anything}\}.$$

When we say $P(X_0 = 0, X_1 = 5, X_2 = 15, X_3 = 19) = .000771605$ as we did in (2.1) we are associating the numerical value .000771605 with this particular subset $C \subseteq \Omega$. To say $X_0 = 0, X_1 = 5, X_2 = 15, X_3 = 19$ is the same as saying $X_{0:\infty} \in C$. We will write $P(C)$ in place of $P(X_{0:\infty} \in C)$ to emphasize this view that we are assigning a probability to a set $C \subseteq \Omega$.

Every probability we consider consists of identifying a particular subset of sequences in Ω and assigning a numerical value to that subset. For sets of the same general type as C above (2.2) prescribes the values that we want to associate with them. In the case of (2.3) the probability that $X_3 = 5$ corresponds to the set $B \subseteq \Omega$ described by

$$B = \{\omega = s_{0:\infty} : s_3 = 5\}.$$

Strictly speaking this is not a set of the type C as above, but it *is* a union of such.

$$B = \cup C_{s_0,s_1,s_2} \text{ where } C_{s_0,s_1,s_2} = \{\omega \in \Omega : \omega = (s_0, s_1, s_2, 5, \ldots)\}.$$

(The union is over all $s_0, s_1, s_2 \in \mathcal{S}$.) Each C_{s_0,s_1,s_2} *is* of the type associated with (2.2). What (2.3) presumes is that the numerical value $P(B)$ assigned to B is the sum of the numerical values assigned to the (disjoint) pieces of the union:

$$P(B) = P(\cup C_{s_0,s_1,s_2}) = \sum P(C_{s_0,s_1,s_2}).$$

Assuming that $\mathcal{S}$ is finite, this sum has finitely many terms.

Next consider the probability that $X_n = 2$ for *some* n. This is the same as saying $X_{0:\infty} \in A$ where $A \subseteq \Omega$ is

$$A = \{\omega = s_{0:\infty} : \text{ there exists } n \text{ for which } s_n = 2\}. \tag{3.1}$$

Again, this is not a set for which (2.2) prescribes a value. But again we can break it down into a union:

$$A = \cup_{n=0}^{\infty} B_n$$

where

$$B_n = \{\omega = s_{0:\infty} \ \in \Omega : s_k \neq a \text{ for } k < n \text{ and } s_n = a\}.$$

Each B_n in turn can be written as a finite union of sets like C. So by means of unions of subsets of Ω we can connect $P(X_n = a \text{ for some } n)$ back to sets for which (2.2) applies. But now we are dealing with an infinite series, not a finite sum.

By viewing every probability as a numerical value $P(A)$ assigned to a subset $A \subseteq \Omega$ we can see how different probabilities are related to each other in terms of writing one subset as a union of others. The prescription (2.2) does not cover all the subsets of Ω whose probabilities we are interested in, but our natural presumption that the probabiilty of a union of disjoint pieces is the sum of the probabilities of the individual pieces allows us to determine the other probabilities we were interested in. It is this presumed additive property of probabilities that allows us to connect basic sets like (2.2) to more complicated sets like (3.1).

3.1.1 The Fundamental Properties of Probability

Here are the basic properties of probabilities assigned to subsets of Ω which we naturally presume, essentially the *axioms* of probability. The second and third bullets are the additive properties which we presumed in the previous chapter. Suppose A, B, A_n, B_n are subsets of Ω.

- $P(\emptyset) = 0$, $P(\Omega) = 1$, and $0 \leq P(A) \leq 1$ for any $A \subseteq \Omega$.
- If A and B are disjoint (meaning $A \cap B = \emptyset$) then $P(A \cup B) = P(A) + P(B)$.
- The preceding extends to countable unions: if $A = \cup_1^\infty B_k$ and every pair B_i, B_j $(i \neq j)$ is disjoint, then
$$P(A) = \sum_1^\infty P(B_k).$$
- We always have
$$P(A^c) = 1 - P(A).$$
- If $A_1 \subseteq A_2 \subseteq \cdots$ and $A = \cup_1^\infty A_n$, then $P(A) = \lim_n P(A_n)$. The same is true if $A_1 \supseteq A_2 \supseteq \cdots$ and $A = \cap_1^\infty A_n$. In other words we can pass to the limit when taking the probabilities of an increasing or decreasing sequence of sets.

These properties describe a sort of additive consistency among the $P(A)$ for various $A \subseteq \Omega$ (with one caveat to be explained below). In brief, probabilities work like areas of geometric figures: you can calculate the probability of something by decomposing it into disjoint pieces (sets) and adding up the probabilities of the pieces. The last bullet says that you can also get probabilities through limit operations in terms of monotone sequences of sets. (Actually these properties are redundant. The second, fourth and fifth can be proven from the first and third.)

There is a general theorem (the Kolmogorov Consistency Theorem) which says that starting with (2.2) as the prescription of $P(C)$ for those specific types of $C \subseteq \Omega$ the values of $P(A)$ for more complicated subsets of Ω are uniquely determined by the above consistency rules. There is one technical caveat however: it turns out that it is *not* possible to assign probabilities to *all* subsets $A \subseteq \Omega$ in a way consistent with the above rules. There are certain "bad" sets A (called *non-measurable* sets) for which no $P(A)$ can be assigned. These A can't be written down explicitly but are proven to exist by abstract arguments. However they never occur in probabilistic calculations of the type we are considering, so we will simply ignore their possibility and proceed as if all $A \subset \Omega$ have a $P(A)$ associated with them consistent with (2.2) and the above rules.

It would be discouraging if all our probability calculations had to work with this cumbersome description of subsets of the set Ω of all sequences and the consistency rules. Take heart, because we almost never work directly with Ω and it subsets. We work mostly with the intuitively natural properties of probabilities, like the additivity properties we presumed in Chapter 1, and some others regarding expected values which we will come to soon. What we have been describing above is the underlying mathematical formulation of probability theory from which the properties we actually work with are derived. It's similar to calculus: we use things like the product rule for derivatives or the method of substitution of integrals on a regular basis, but not the definitions of derivative and integral in terms limit sfrom which those working properties are derived.

3.1.2 Random Variables

If we are interested in $P(X_3 = 7)$ then we use $C = \{\omega = s_{0:\infty} : s_3 = 7\}$. To build the random variable X_3 itself into the formulation above we take the view that X_3 is actually a function, $X_3 : \Omega \to \mathcal{S}$, defined by $X_3(\omega) = s_3$, where $\omega = s_{0:\infty}$. I.e. X_3 is the third-term-reader function of a sequence ω. Then $C = \{\omega : X_3(\omega) = 7\}$.

For mathematical purposes every random variable Y is viewed this way. It is a function $Y(\omega)$ with domain Ω. A random variable is an ω-reader which only reveals some aspect of ω to us. Then something like $P(Y < 12)$ is taken to mean $P(C)$ where $C = \{\omega \in \Omega : Y(\omega) < 12\}$. Every random variable is some particular ω-reader and all probabilities involving Y are $P(C)$ for the appropriate set of C determined by testing ω using the ω-reader Y. (We are sweeping some technicalities under the rug again here. What if C were one of those bad sets for which no value $P(C)$ is assigned? The full definition of a random variable includes a requirement that $Y(\cdot)$ is a well-enough-behaved function of $\omega \in \Omega$ that this never happens. But enough; put that back under the rug and don't worry about it.) We think of the whole scheme this way: the inscrutable random influences of the universe choose one full outcome sequence $\omega = s_{0:\infty}$ and then what we see when we look at the random variables of interest are their values $X_3(\omega)$, $Y(\omega)$ and any others all evaluated at the same selected ω.

If there are two random variables X and Y and we want the probability that X and Y obey some relation, say $X \leq Y$, then we determine the corresponding set

$$B = \{\omega \in \Omega \,:\, X(\omega) \leq Y(\omega)\},$$

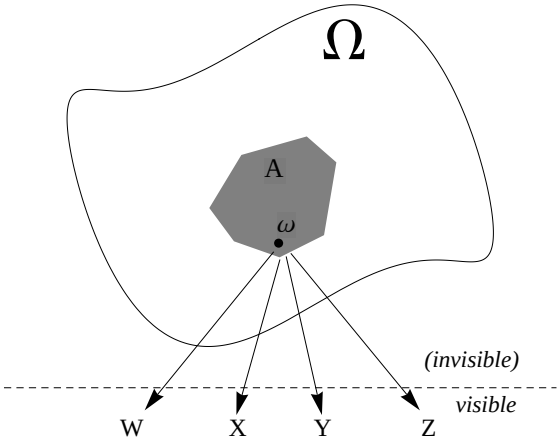
and then calculate $P(X \leq Y)$ by finding $P(B)$. In this way every outcome that we can describe in terms of the random variables corresponds to some particular subset of Ω, to which a probability is assigned. This is true for things like (3.1) as well. The set A is

$$A = \{\omega : X_n(\omega) = a \text{ for some } n\}$$

and then $P(A)$ gives the value of $P(X_n = a \text{ for some } n)$.

What we have described here is called the Kolmogorov model of probability theory. It consists of a set Ω called the *sample space*. Its subsets $A \subseteq \Omega$ are called *events* and have probabilities $P(A)$ assigned to them

in way that satisfies the properties of Section 3.1.1. Random variables are functions defined on Ω. We have considered Ω to be specifically the set of all infinite sequences of states but theoretical probabilists don't usually worry about what Ω actually consists of. We just assume that Ω and the assignment of probabilities to its subsets exist somewhere in the unseen background. We typically work only with events that can be described directly in terms of the random variables, without any direct reference to Ω. For that reason we usually leave out the "ω" and just write $A = \{X = 2\}$ or $B = \{X \leq Y\}$. However this use of set notation for events is a reminder of this underlying mathematical structure.



The simplest type of random variable is one which takes only the values 0 or 1. This is sometimes called a *Bernoulli* random variable. Suppose $A \subseteq \Omega$. Using A we can define the random variable

$$1_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{if } \omega \notin A. \end{cases}$$

This is a Bernoulli random variable. Its probabilities are

$$P(1_A = 1) = P(A)$$
$$P(1_A = 0) = P(A^c) = 1 - P(A).$$

Every Bernoulli random variable X is of this form for some choice of A: $X = 1_A$ where $A = \{\omega : X(\omega) = 1\}$.

3.2 Expectations

Now that we have said precisely what a random variable X is we want to discuss its expectation $E[X]$. This is intended to be the average of the possible values of X with the different values, each weighted according to its probability. For instance if X only takes integer values (i.e. $X(\omega) \in \mathbb{Z}$ for all ω) then

$$E[X] = \sum_{k \in \mathbb{Z}} kP(X = k), \tag{3.2}$$

provided the series is absolutely convergent. A random variable whose possible values are limited to a countable set is called a *discrete* random variable. Most of the random variables associated with Markov chains are discrete, taking only integer or positive integer values. Here is an example.

Example 3.1. A *Poisson* random variable with parameter $\lambda > 0$ has probabilities

$$P(X = n) = \frac{\lambda^n}{n!} e^{-\lambda}, \ n = 0, 1, \ldots.$$

Its expected value is

$$\begin{aligned} E[X] &= \sum_{n=0}^{\infty} n \frac{\lambda^n}{n!} e^{-\lambda} \\ &= \sum_{n=1}^{\infty} \lambda \frac{\lambda^{n-1}}{(n-1)!} e^{-\lambda} \\ &= \lambda \end{aligned}$$

A random variable X is called *continuous* if its probabilities are described by a density function ($f(x) \geq 0$ with $\int_{-\infty}^{\infty} f(x)\,dx = 1$) by means of the formula

$$P(a \leq X \leq b) = \int_a^b f(x)\,dx.$$

In that case we calculate its expectation with an integral:

$$E[X] = \int_{-\infty}^{\infty} x f(x)\,dx. \tag{3.3}$$

Here are two examples.

Example 3.2. A *uniform* random variable on $[a, b]$ has density

$$f(x) = \begin{cases} 0 & \text{if } x < a \\ \frac{1}{b-a} & \text{if } a \leq x \leq b \\ 0 & \text{if } b < x. \end{cases}$$

Its expected value is

$$E[X] = \int_{-\infty}^{\infty} x f(x)\,dx = \int_a^b \frac{x}{b-a}\,dx = \frac{1}{b-a}(\frac{1}{2}b^2 - \frac{1}{2}a^2) = \frac{a+b}{2}.$$

Example 3.3. A *normal* random variable with with parameters (μ, σ^2) has density $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$. Its expected value is

$$\begin{aligned} E[X] &= \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,dx \\ &= \mu \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,dx + \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} (x-\mu) e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,dx \\ &= \mu + 0 = \mu. \end{aligned}$$

The first integral is because $\int f = 1$. The second is $= 0$ by symmetry about μ.

Some other common types of discrete and continuous random variables are identified in Example 3.5 and the Appendix.

If $\phi : \mathbb{R} \to \mathbb{R}$ is a function then $\phi(X)$ is another random variable, so we can talk about its expectation, $E[\phi(X)]$. For instance we can consider things like $E[X^2]$, $E[e^X]$, $E[|X|]$. If X is discrete then to calculate $E[\phi(X)]$ as we described it above would require working out all the probabilities $P(\phi(X) = y)$ in order to determine

$$E[\phi(X)] = \sum_y y P(\phi(X) = y).$$

But by breaking each $P(\phi(X) = y)$ up into a sum of pieces according to the values $X = k$ for which $\phi(k) = y$ and rearranging we find that

$$\begin{aligned} E[\phi(X)] &= \sum_y yP(\phi(X) = y) \\ &= \sum_y y \sum_{k:\phi(k)=y} P(X = k) \\ &= \sum_y \sum_{k:\phi(k)=y} \phi(k)P(X = k) \\ &= \sum_k \phi(k)P(X = k). \end{aligned} \tag{3.4}$$

The first line is a sum over values of ϕ; the last line is a sum over values of X. For a continuous random variable to follow (3.3) strictly would require determining the density of $\phi(X)$, but similar to our calculation above it turns out that

$$E[\phi(X)] = \int_{-\infty}^{\infty} \phi(x)f(x)\,dx \tag{3.5}$$

produces the same result. We can take these two formulas (3.4) and (3.5), as our working definitions of $E[\phi(X)]$ in the discrete and continuous cases.

Integrability

There is one technical issue we need to be aware of when talking about expectations. This is the possibility that an ambiguous $\infty - \infty$ somehow arises in the calculation of $E[X]$.

Example 3.4. Suppose that X takes only integer values (both positive and negative), with probabilities $P(X = 0) = 0$ and $P(X = n) = \frac{3}{\pi^2 n^2}$ for $n \neq 0$. (The $\frac{3}{\pi^2}$ is so that $\sum_{-\infty}^{\infty} P(X = n) = 1$. Recall that $\sum_1^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}$.) If we try to calculate $E[X]$ we are faced with

$$\sum_{n=-1}^{-\infty} \frac{-3}{\pi^2 n} + \sum_{n=1}^{\infty} \frac{+3}{\pi^2 n},$$

which is of the form $-\infty + \infty$ since both series are divergent.

We can deal with just a $+\infty$ or just a $-\infty$ in an expectation. In particular we can always talk about $E[\,|X|\,]$ if we allow $+\infty$ as a possible value. In the example we do have $E[\,|X|\,] = \infty$. But in cases like the example where the positive and negative contributions to $E[X]$ are *both* infinitely large there is no satisfactory way to resolve the cancellation in $\infty - \infty$; we have to consider $E[X]$ undefined in such cases.

Definition. *We say that a random variable X is* integrable *or* has finite mean *when*

$$E[\,|X|\,] < \infty.$$

To say that X is integrable is to say that both the positive and negative contributions to $E[X]$ are finite, so $\infty - \infty$ does not occur and $E[X]$ will have a finite value. Look back at equation (3.2) and notice that we said the series should converge *absolutely*. That is what integrability requires for discrete random variables. For a continuous random variable to be integrable means that the improper integrals

$$\int_{-\infty}^{0} xf(x)\,dx \text{ and } \int_{0}^{\infty} xf(x)\,dx \text{ are } \textit{both} \text{ convergent.}$$

This agrees with the usual convention in integral calculus for $\int_{-\infty}^{\infty} xf(x)\,dx$ to be considered convergent. Integrabiity will be a technical hypothesis for many of the results we describe below.

Properties

We have given formulas for the expected values of random variables that are either discrete or continuous. There do exist random variables which are *neither* discrete nor continuous. They won't arise in any of our considerations, but the fact that they exist means that a full account of expectations requires a more general definition for $E[X]$ which accommodates all types of random variables. That general definition takes the Kolmogorov model point of view that a random variable X is a function defined on an underlying set Ω and forms a type of integral, sometimes denoted

$$E[X] = \int_\Omega X(\omega)\, P(d\omega). \tag{3.6}$$

Although we won't attempt to explain how that works, just thinking of $E[X]$ as an integral of X as a function helps several of the properties of expectations listed in the proposition below seem natural.

One more thing before we list properties of expectations. Sometimes we want to consider just the part of an expectation corresponding to a particular subset $A \subseteq \Omega$. We use the notation $E[X; A]$ for this. The "; A" in $E[X; A]$ is like the interval of integration $[a, b]$ in $\int_a^b g(x)\, dx$; it specifies how much of the expectation or integral we want. In fact using an integral notation of (3.6) we could write

$$E[X; A] = \int_A X(\omega)\, P(d\omega).$$

The full expected value is $E[X] = E[X; \Omega]$. A formula for it is

$$E[X; A] = E[X \cdot 1_A]. \tag{3.7}$$

The effect of multiplying by 1_A is to create a new random variable $X \cdot 1_A$ which agrees with X for $\omega \in A$ but is just 0 otherwise:

$$X \cdot 1_A(\omega) = \begin{cases} X & \text{if } \omega \in A \\ 0 & \text{if } \omega \in A^c. \end{cases}$$

Here then are some of the most elementary properties of expectations. These hold for all types of random variables (discrete, continuous or otherwise).

Proposition 3.1. *If X, Y are both integrable random variables and $A, B \subset \Omega$ are events, then*

a) $E[1_A] = P(A)$.

b) $E[aX + bY] = aE[X] + bE[Y]$ *for any constants* a, b.

c) $X(\omega) \le Y(\omega)$ *for all* ω *implies* $E[X] \le E[Y]$.

d) $|E[X]| \le E[\,|X|\,]$.

e) $E[X] = E[X; A] + E[X; A^c]$.

f) $E[X; A \cup B] = E[X; A] + E[X; B]$, *provided* $A \cap B = \emptyset$.

Without a full definition of (3.6) we can't write actual proofs of all of these. But we can offer convincing arguments for several of them.

Part a) is elementary from (3.2): since 1_A only takes two possible values we calculate

$$E[1_A] = 1P(1_A = 1) + 0P(1_A = 0) = P(A).$$

If we remember that $E[X]$ is really a type of integral many of these properties are familiar by analogy with the integral $\int_a^b f(x)\, dx$ of calculus. For instance when written this way b) says

$$\int [aX(\omega) + bY(\omega)]\; P(d\omega) = a \int X(\omega)\, P(d\omega) + b \int Y(\omega)\, P(d\omega).$$

Part f) is says

$$\int_{A\cup B} X(\omega)\,dP = \int_A X(\omega)\,dP + \int_B X(\omega)\,dP,$$

which is a lot like the familiar $\int_a^c = \int_a^b + \int_b^c$.

The *variance* of a random variable is

$$\mathrm{Var}[X] = E[(X-m)^2], \text{ where } m = E[X].$$

We can use the properties above to derive a common formula for it.

$$\begin{aligned}
\mathrm{Var}[X] &= E[(X-m)^2] \\
&= E[X^2 - 2mX + m^2] \\
&= E[X^2] - 2mE[X] + m^2E[1], \text{ using b)} \\
&= E[X^2] - 2m^2 + m^2, \text{ since } E[X] = m \\
&= E[X^2] - m^2 \\
&= E[X^2] - E[X]^2.
\end{aligned}$$

For a nonnegative integer valued random variable $X \in \mathbb{Z}^+$ there is a useful alternate formula for the expected value:

$$E[X] = \sum_{n=0}^{\infty} P(X > n). \tag{3.8}$$

This can be derived by interchanging orders in a double summation:

$$\begin{aligned}
E[X] &= \sum_{n=1}^{\infty} nP(X=n) \\
&= \sum_{n=1}^{\infty}\sum_{m=1}^{n} P(X=n) \\
&= \sum_{m=1}^{\infty}\sum_{n=m}^{\infty} P(X=n) \\
&= \sum_{m=1}^{\infty} P(X \ge m) \\
&= \sum_{n=0}^{\infty} P(X > n).
\end{aligned}$$

(In fact we have used this already, in the proof of Theorem 2.6.) There is a version of this for continuous random variables too; see Problem 3.4. Here is an application of formula (3.8) to the calculation of an expected value.

Example 3.5. A *geometric* random variable with parameter $0 < p < 1$ has probabilities

$$P(X=n) = p(1-p)^n, \ n = 0, 1, \ldots.$$

Its expected value is

$$E[X] = \sum_{n=0}^{\infty} np(1-p)^n.$$

This can be worked out directly (by differentiating a power series). However it is easier to work it out using

(3.8). First observe that

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

Therefore by (3.8) we have

$$\begin{aligned} E[X] &= \sum_{n=0}^{\infty} (1-p)^{n+1} \\ &= (1-p) \sum_{n=0}^{\infty} (1-p)^n \\ &= (1-p)\frac{1}{p} \\ &= \frac{1-p}{p}. \end{aligned}$$

3.2.1 Limits in Expectations

We are going to encounter situations in which we need the expectation of a limit. To be more explicit, if we have an infinite sequence $X_1, X_2, \ldots$ of random variables we will sometimes want to say that

$$\lim_{n\to\infty} E[X_n] = E[\lim_n X_n],$$

assuming that both limits exist. This is often correct but not always. If the limit inside the righthand expectation does exist it defines a new random variable

$$\lim_{n\to\infty} X_n(\omega) = Y(\omega) \tag{3.9}$$

for all $\omega \in \Omega$. This is called "pointwise convergence" of X_n to Y. Sometimes the closest we can get to (3.9) is that it holds with probability 1, not that it holds for every single ω. So for practical purposes we want to allow (3.9) to fail for some ω, provided that set of ω for which it fails has probability 0. This is called *almost sure convergence*. The Strong Law of Large Numbers below will give one nontrivial situation in which this kind of convergence holds.

Definition. *Let X_n be a sequence of random variables. We say that X_n converges to a random variable Y* almost surely*, written "$X_n \to Y$ a.s.", if*

$$P(\lim_{n\to\infty} X_n = Y) = 1;$$

Example 3.6. Suppose X_n takes only two possible values, 0 or 2^n, with probabilities

$$P(X_n = 0) = 1 - 2^{-n}, P(X_n = 2^n) = 2^{-n}.$$

We claim that $X_n \to 0$ almost surely. This is because

$$\begin{aligned} P(X_n > 0 \text{ for some } n \geq m) &= \sum_{k=m}^{\infty} P(X_i = 0 \text{ for all } m \leq i < k \text{ and } X_k > 0) \\ &\leq \sum_{k=m}^{\infty} P(X_k > 0) \\ &= \sum_{k=m}^{\infty} 2^{-k} \\ &= 2^{1-m}. \end{aligned}$$

Let

$$A_m = \{X_n > 0 \text{ for some } n \geq m\}.$$

These are a decreasing sequence of sets, $A_1 \supseteq A_2 \supseteq \cdots$ and their intersection is

$$\cap_{m=1}^{\infty} A_m = \{X_n > 0 \text{ for infinitely many } n\}.$$

Therefore the last bullet on page 33 implies that

$$P(X_n > 0 \text{ for infinitely many } n) = \lim P(A_m) = \lim 2^{1-m} = 0.$$

So with probability 1, $X_n > 0$ happens only a finite number of times. This implies that $X_n \to 0$ with probability 1.

Observe that this proof of $P(X_n \to 0) = 1$ used features of the Kolmorogov model but did *not* depend on knowing what the underlying Ω actually is or how the $X_n(\omega)$ are defined as functions of $\omega \in \Omega$. This illustrates our comments at the end of Section 3.1.2 about not needing to work directly with Ω.

This example also shows that it is possible for

$$\lim E[X_n] \neq E[\lim X_n].$$

It is easy to check that $E[X_n] = 1$ for all n so that $1 = \lim E[X_n]$ while $E[\lim X_n] = E[0] = 0$. Limits and expectations do *not* always commute!

We now state three important results describing when it *is* correct to say $E[\lim X_n] = \lim E[X_n]$.

Theorem 3.2 (Monotone Convergence)**.** *Suppose that $X_n \to Y$ almost surely and that*

$$0 \leq X_1 \leq X_2 \leq \cdots \leq X_n \leq \cdots$$

Then $E[Y] = \lim E[X_n]$, provided the limit on the right exists.

Theorem 3.3 (Dominated Convergence)**.** *Suppose $X_n \to Y$ almost surely and that there exists an* integrable *random variable W so that $|X_n| \leq W$ (with probability 1) for all n. Then $E[Y] = \lim E[X_n]$.*

Theorem 3.4 (Fatou's Lemma)**.** *Suppose $X_n \to Y$ almost surely and that $X_n \geq 0$ for each n. Then $E[Y] \leq \lim E[X_n]$, provided the limit on the right exists.*

These are general properties which hold for most notions of integration, including (3.6) as well as the Riemann integral $\int_a^b$ of calculus, subject to some technical limitations on the integrands. Although we will not prove them in general they are powerful tools which we will use as needed in various places below. When Ω is countable they reduce to the results on infinite series proven in the Appendix as Theorems A.7 and A.8.

3.3 Independence and Dependence

Suppose I roll two conventional dice and denote their outcomes as D_1 and D_2. These are two random variables, each taking one of the values $1, \ldots, 6$ with equal probability. If I tell you that $D_1 = 4$ and then ask you if that information influences the probabilities for D_2 you would say "no." This is what we mean by saying that D_1 and D_2 are independent random variables.

Suppose that instead of telling you D_1 and D_2 directly I first tell you wheter or not their sum is less than 5. I.e. I tell you the value of the random variable

$$X = \begin{cases} 1 & \text{if } D_1 + D_2 < 5 \\ 0 & \text{if } D_1 + D_2 \geq 5. \end{cases}$$

To be specific, suppose I tell you that $X = 1$. That *would* influence the probabilities of D_2. It would be impossible for D_2 to be 5, or 6. Moreover knowing that $X = 1$ would mean that $D_1 = 1$ and $D_1 = 2$ are *not* equally likely! Here we have two random variables, X and D_2, which are *not* independent. There is some sort of partial dependence between them, but not enough for you to be able to deduce the specific value of D_2 from knowledge of X.

As another possibility consider

$$Y = 2^{D_1} 3^{D_2}.$$

If you are told that $Y = 72$ then you *can* deduce that $D_2 = 2$ (from the prime factorization $72 = 2^3 3^2$). Here D_2 is completely dependent on Y. We will say that D_2 is Y*-determined*.

These simple examples illustrate that there is a range of degrees of dependency between a pair of random variables. Independence (as for D_1 and D_2) and complete dependence (as for Y and D_2) are at opposite ends of the scale. The partial dependence (as for X and D_2) is somewhere in the middle. To describe partial dependence more carefully involves joint distributions and conditional probabilities, which we will come to in Section 3.5 below.

Complete dependence means that there is a functional relation between the random variables, $D_2 = f(Y)$ in our example above, where $f(k)$ is the function which gives the power of 3 in the prime factorization of k. In general we will say that X is $Y_{1:m}$*-determined* if there is a function $f(y_1, \ldots y_m)$ so that

$$X = f(Y_1, \ldots Y_m)$$

always holds.

Here is the definition of independence.

Definition. *Two random variables X and Y are called* independent *when for any two sets A and B*

$$P(X \in A \text{ and } Y \in B) = P(X \in A)P(Y \in B).$$

When we have several random variables $X_1, X_2, \ldots, X_n$ they are independent *when*

$$P(X_1 \in A_1,\ X_2 \in A_2, \ldots, X_n \in A_n) = P(X_1 \in A_1)P(X_2 \in A_2) \cdots P(X_n \in A_n).$$

for every choice of sets $A_1, \ldots A_n$. An infinite sequence X_1, X_2, ... is independent *when every finite subsequence $X_1, \ldots, X_n$ is.*

Technically the sets A, B, A_i should be assumed to be measurable, but that is getting into the finer mathemaitcal details which we are choosing to ignore. Assuming that the random variables are real-valued it is sufficient to just use intervals $A_i = (-\infty, a_i]$:

$$P(X_1 \leq a_1,\ X_2 \leq a_2, \ldots, X_n \leq a_n) = P(X_1 \leq a_1)P(X_2 \leq a_2) \cdots P(X_n \leq a_n).$$

for any choice of $a_1, \ldots a_n$. For integer-valued random variables it is enough to just check individual values:

$$P(X = i \text{ and } Y = j) = P(X = i)P(Y = j)$$

for all i, j.

It is important to note that for X, Y and Z to be independent requires more than just X and Y independent, Y and Z independent, and X and Z independent: pairwise independence is not the same as independence of the whole collection! See Problem 3.6 below.

One consequence of independence is that the expected value of the product is the product of expected values:

$$E[XY] = E[X]\,E[Y], \tag{3.10}$$

assuming X and Y are independent and both integrable. The verification of this for integer-valued random variables is as follows.

$$\begin{aligned} E[XY] &= \sum_i \sum_j ijP(X = i \text{ and } Y = j) \\ &= \sum_i \sum_j ijP(X = i)P(Y = j), \text{ using independence} \\ &= \left(\sum_i iP(X = i)\right)\left(\sum_j jP(Y = j)\right) \\ &= E[X]E[Y]. \end{aligned}$$

(The middle step requires both series to be absolutely convergent, which is equivalent to integrability of X and Y.) It is easy to see that (3.10) generalizes to the formula

$$E[g(X)h(Y)] = E[g(X)]\,E[h(Y)],$$

for any two functions $g(x)$, $h(y)$, provided that $g(X)$ and $h(Y)$ are integrable. This can be checked directly, as above. Alternately, first recognize that if $U = g(X)$ is X-dependent and $V = h(Y)$ is Y-dependent, then independence of U and V is a consequence of independence of X and Y. So provided U and V are integrable we find that $E[UV] = E[U]E[V]$ as an application of (3.10).

Example 3.7. Suppose we take a fair dice and roll it repeatedly, generating a sequence of independent dice-roll random variables $D_1, D_2, D_3 \ldots$ Use the outcomes define a new random variable N to be the number of rolls we made *before* the first instance of a 1 or 2. In other words N is how many times in a row we got a 3 or larger. We can calculate the distribution of N using the independence of the D_i.

$$\begin{aligned} P(N = k) &= P(D_1 \geq 3, D_2 \geq 3, \ldots, D_k \geq 3, D_{k+1} \leq 2) \\ &= P(D_1 \geq 3)P(D_2 \geq 3)P(D_k \geq 3) \cdots P(D_k \geq 3)P(D_{k+1} \leq 2) \\ &= (4/6)^k 2/6 = (2/3)^k 1/3. \end{aligned}$$

This is a geometric distribution with parameter 1/3, as defined in Example 3.5. To demonstrate with MATLAB here is a function m-file to produce one sample of N.

Nb.m

```
function n=Nb()
%Nb produces a single simulation of the random variable N
%
n=0;                    %Initialize
d=randi(6,[1,1]);
while not(d==1 || d==2) %Repeatedly sample until finding a 1 or 2
   n=n+1;               %n is #rolls when 1 or 2 was found
   d=randi(6,[1,1]);
end
```

The following code with generate 10000 samples of N and then look at the frequencies at which the different values appeared, and then compare those to the theoretical probabilities above.

```
data=zeros(1,10000);
for i=1:10000
    data(i)=Nb;
end
m=max(data)
hist(data,0:m)
counts=histc(data,0:m)
counts/10000
(1/3)*(2/3).^(0:m)
```

3.3.1 Sums of Independent Random Variables

Suppose X and Y are independent random variables and we define their sum to be the random variable

$$S = X + Y.$$

It is always possible, in principle, to work out the distribution of S from the joint distribution of (X, Y). But when X and Y are independent the distribution of S has a nice relationship to the distributions of X and Y.

Suppose X and Y are independent, both discrete taking integer values: $P(X = i) = p_i$ and $P(Y = j) = q_j$. There are several ways that $S = k$ can occur; any combination of $X = i$ and $Y = j$ with $i + j = k$ will do it. So we find

$$\begin{aligned} \{S = k\} &= \cup_i \{X = i \text{ and } Y = k - i\} \\ P(S = k) &= \sum_i P(X = i \text{ and } Y = k - i) \\ &= \sum_i p_i q_{k-i}, \text{ using independence.} \end{aligned}$$

We can limit the sum to those i with $p_i > 0$. If we know that $0 \le X$ and $0 \le Y$ then

$$P(S = k) = \sum_{i=0}^{k} p_i q_{k-i} = p_0 q_k + p_1 q_{k-1} + \cdots + p_k q_0. \tag{3.11}$$

Example 3.8. Suppose D_1 and D_2 are independent, discrete random variables, uniform on $\{1, 2, \ldots 6\}$. In other words they are a conventional pair of fair dice. Let $X = D_1 + D_2$. The distribution of X works out to be.

$$P(X = k) = \begin{cases} \frac{k-1}{36} & \text{for } k = 2, \ldots, 7 \\ \frac{13-k}{36} & \text{for } k = 8, \ldots 12. \end{cases}$$

For instance, since $p_i = q_i = \frac{1}{6}$ for $i = 1, \ldots 6$, we have

$$\begin{aligned} P(X = 6) &= p_1 q_5 + p_2 q_4 + p_3 q_3 + p_4 q_2 + p_5 q_1 \\ &= 5(1/6)^2 = \frac{5}{36}. \end{aligned}$$

Example 3.9. Suppose $(Z_1, Z_2, \ldots)$ is a sequence of independent Bernoulli random variables with the same distribution: $p = P(Z_i = 1)$ for all i. Consider

$$X = \sum_{i=1}^{n} Z_i.$$

If we specify some sequence d_i of 0s and 1s, and $k = \sum_1^n d_i$, then by independence

$$P(Z_1 = d_1,\ Z_2 = d_2, \ldots \text{ and } Z_n = d_n) = p^k (1 - p)^{n-k}.$$

If we want $P(X = k)$ then we have to count how many different ways we can choose $(d_1, \ldots, d_n)$ with $k = \sum_1^n d_i$. This is given by the binomial coefficient

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}.$$

So we find that

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

In other words X is binomial with parameters (n, p). We could calculate the mean of X by working out the sum $\sum_{j=0}^n j\binom{n}{j}p^j(1-p)^{n-j}$ but it is far easier to use the independence:

$$E[X] = E\left[\sum_{i=1}^n Z_i\right] = \sum_{i=1}^n E[Z_i] = np.$$

For the second moment we have

$$E[X^2] = E\left[\left(\sum_{i=1}^n Z_i\right)^2\right] = E\left[\sum_1^n Z_i^2 + \sum_{i\neq j} Z_i Z_j\right] = \sum_1^n E[Z_i^2] + \sum_{i\neq j} E[Z_i Z_j] = np + n(n-1)p^2.$$

So the variance is

$$\mathrm{Var}(X) = E[X^2] - E[X]^2 = np + n(n-1)p^2 - (np)^2 = np(1-p).$$

Suppose Y is a second binomial random variable, independent of X and with parameters (m, p). We can produce such a Y by using more of the Z_i:

$$Y = \sum_{i=n+1}^{n+m} Z_i,$$

and therefore

$$X + Y = \sum_{i=1}^{n+m} Z_i$$

must also be binomial, but with parameters $(n+m, p)$. This fact could be worked out as the convolution of the two binomial distributions of X and Y directly, but that is a more tedious calculation.

3.4 Famous Theorems for I.I.D. Sequences

Suppose that $X_1, X_2, \ldots$ is a sequence of independent random variables and they all have the same distribution, i.e. $P(X_n \leq t)$ is the same for all n. We call this an *independent identically distributed (i.i.d.)* sequence. An infinite sequence of coin flips or an infinite sequence of dice rolls would be an example.

Suppose $X_1, X_2, \ldots$ is an i.i.d. sequence. We present below three famous theorems concerning the sequence S_n of partial sums,

$$S_n = \sum_{i=1}^n X_i.$$

These results are centerpieces of probability theory. We will use them in various ways below. At the end of the chapter we will cite references where proofs can be found. We will focus our efforts on understanding and illustrating these theorems.

Theorem 3.5 (The Renewal Theorem)**.** *Let X_i be an i.i.d. sequence of random variables with finite mean $m = E[X_i]$, which take only positive integer values, and such that $P(X_i$ is divisible by $k) < 1$ for every integer $k > 1$. Let $S_n = \sum_{i=1}^n X_i$ and $g_k = P(S_n = k$ for some $n)$. Then*

$$\lim_{k\to\infty} g_k = \frac{1}{m}.$$

The hypothesis that $P(X_i \text{ is divisible by } k) < 1$ for every integer $k > 1$ means that X is not concentrated on the multiples of any positive integer $k > 1$; this is usually described by saying that X is *non-arithmetic* or *aperiodic*.

Example 3.10. Suppose we are playing a game in which we start at 0 and at each turn roll a pair of dice to determine how many steps we move to our next position. Our successive positions are

$$S_0 = 0, S_1 = X_1, S_2 = X_1 + X_2, \dots$$

where the X_i are i.i.d. with the dice pair distribution; see Problem 3.5. (We move on a straight line, not on a game board that wraps around.) If you pick a position k, the value g_k defined in the theorem is the probability that you will land at position k at some time. The probability of skipping over k would be $1 - g_k$. The average gap between the positions we land at is $E[X] = 7$. The Renewal Theorem says that for large k

$$g_k \approx 1/E[X] = 1/7 = .1428571.$$

In other words the probability that k does not fall in a gap is $\approx \frac{1}{\text{mean gap size}}$ for large k. Thinking about it that way makes the theorem seem reasonable.

We can compute the sequence $\{g_k\}$ numerically to see if this appears to be true or not. First notice that $g_0 = 1$ because we start with $S_0 = 0$. Now consider $k > 0$; we are interested in the event

$$B = \{k = \sum_{i=1}^{n} X_i \text{ for some } n\}.$$

We can break this up as a disjoint union based on the value $j = X_1$: $B = \cup_1^k B_j$ where

$$\begin{aligned} B_j &= \{X_1 = j\} \cap B \\ &= \{X_1 = j\} \cap \{k - j = \sum_{i=1}^{n} X_{i+1} \text{ for some } n\} \end{aligned}$$

Because the X_i are independent we can write

$$P(B_j) = P(X_1 = j) P(k - j = \sum_{i=1}^{n} X_{i+1} \text{ for some } n).$$

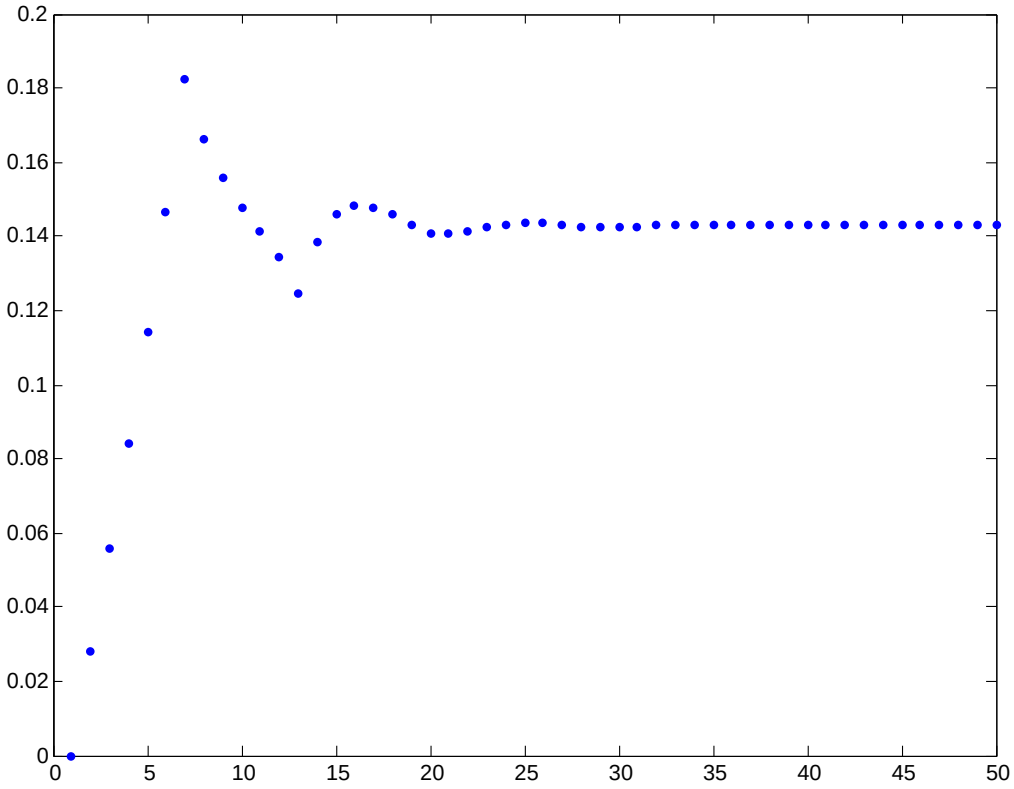
The first term is just $p_j = P(X_1 = j)$. The second term is in fact g_{k-j}, because the X_i are identically distributed. So $P(B_j) = p_j g_{k-j}$ for $j = 0, \dots, k$. Therefore

$$g_k = \sum_{j=0}^{k} P(B_j) = \sum_{j=0}^{k} p_j g_{k-j}.$$

For our dice pair distribution, $p_0 = 0$, so the right side only involves $k - j = 0, \cdots, k - 1$. Thus we can calculate the g_k recursively from this, starting with $g_0 = 1$.

$$\begin{aligned} g_1 &= p_1 g_0 \\ g_2 &= p_1 g_1 + p_2 g_0 \\ &\vdots \\ g_k &= p_1 g_{k-1} + \cdots + p_k g_0 \\ &\vdots \end{aligned}$$

Of course $p_j = 0$ for $j > 12$. Here is a graph of the resulting values.



Notice that $g_1 = 0$ because $p_1 = 0$. It is amusing to observe that the largest value is $g_7 = 0.182227$ ("lucky 7") and the smallest value after that is $g_{13} = 0.124704$ ("unlucky 13"). Looking at the calculated values we see that g_k does indeed seem to converge to a value $\approx .143$, as Theorem 3.5 said it would.

Theorem 3.6 (Strong Law of Large Numbers)**.** *Suppose $X_1, X_2, \ldots$ is a sequence of independent identicaly distributed random variables with finite mean $m = E[X_n]$. Then*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} X_i = m$$

with probability 1. *If $X_i \geq 0$ and $E[X_i] = \infty$ then $\frac{1}{n}\sum_{i=1}^{n} X_i \to \infty$ as $n \to \infty$ with probability* 1.

This is what people sometimes refer to as the "law of averages". In terms of the Kolmogorov model, what it says is that the event

$$C = \left\{ \omega \in \Omega \;\middle|\; \lim_{n} \frac{1}{n} \sum_{i=1}^{n} X_i = m \right\}$$

has $P(C) = 1$. We will illustrate this using a MATLAB experiment in Example 3.12 below.

Theorem 3.6 sys that for large n the random variable $\frac{1}{n}S_n - m$ is nearly 0 with very high probability. But it is not exactly 0; it still has some randomness. The next famous result says that when rescaled appropriately,

$$\frac{\sqrt{n}}{\sigma}\left(\frac{1}{n}S_n - m\right) = \frac{S_n - nm}{\sqrt{n}\sigma}$$

where $\sigma^2 = \mathrm{Var}(X_i)$, the distribution of $\frac{1}{n}S_n - m$ converges to a particular shape.

Theorem 3.7 (The Central Limit Theorem)**.** *Suppose X_i is an i.i.d. sequence of random variables with finite mean m and finite variance σ^2. Then*

$$\lim_{n\to\infty} P\left(a < \frac{S_n - nm}{\sqrt{n}\sigma} < b\right) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}\, dy.$$

In other words the distribution of $\frac{S_n - nm}{\sqrt{n}\sigma}$ is very nearly that of a standard normal random variable. We might say it this way:

$$S_n \approx nm + \sqrt{n}\sigma Y,$$

where Y is a standard normal random variable.

The Central Limit Theorem is one reason the normal distribution is so ubiquitous; it arises naturally when the randomness is the cumulative effects of many small independent random influences. We can illustrate with some computer calculations.

Example 3.11. Suppose the X_i are i.i.d. uniform random variables on $[0, 1]$. That means they have density

$$f(x) = \begin{cases} 1 & \text{for } 0 \le x \le 1 \\ 0 & \text{otherwise.} \end{cases}$$

The mean and variance of X_i are

$$m = \int_0^1 x \cdot 1\, dx = 1/2, \quad \sigma^2 = \int_0^1 (x - 1/2)^2\, dx = 1/12.$$

For small values of n we can work out the density $f_n(s)$ of $S_n = \sum_{i=1}^n X_i$ explicitly. For instance the density of S_2 is

$$\begin{aligned} f_2(s) &= \int_{-\infty}^{\infty} f(t) f(s-t)\, dt \\ &= \begin{cases} s & \text{for } 0 \le s \le 1 \\ 2 - s & \text{for } 1 < s \le 2 \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

The densities of S_3 are calculated recursively from the convolution formula

$$f_{n+1}(s) = \int_{-\infty}^{\infty} f_n(t) f(s-t)\, dt.$$

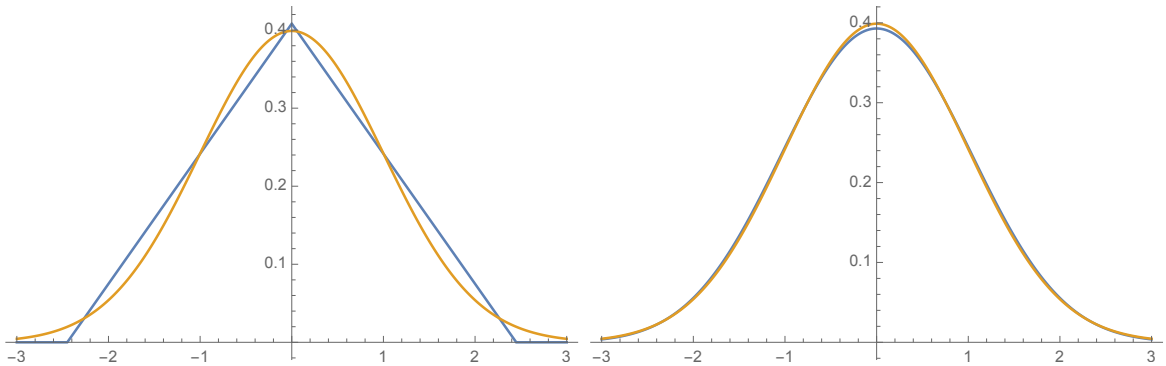
These calculations get increasingly tedious as n gets larger due to the large number of pieces in the formula. But computer algebra software can do the work (for modest values of n). We would like to compare the standard normal density to the density of

$$\frac{S_n - nm}{\sqrt{n\sigma^2}},$$

which is

$$f_n(nm + \sqrt{n\sigma^2}y)\sqrt{n\sigma^2}.$$

Here are the comparative plots for $n = 2$ and $n = 10$. (The orange curve is the standard normal density.)



You can see how close the two densities are for $n = 10$. This is essentially what Theorem 3.7 predicted.

Example 3.12. We can examine Theorem 3.6 with a simple MATLAB experiment.

```
X=rand(1,10000); % Uniform-[0,1] r.v.s; mean=.5, var=1/12
S=cumsum(X);
N=1:10000;
L=S./N;
plot(L) % We can see the SLLN convergence to the mean = .5
```

A subtle but important point is that Theorem 3.7 does *not* say that the random variables $\frac{S_n - nm}{\sqrt{n}\sigma}$ converge as $n \to \infty$, only that the *probabilities* associated with these random variables converge, i.e. their distributions converge.

```
% Continuing with the same data generated above ...
C=(S-.5*N).*sqrt(12./N);
plot(C) % There is apparently no convergence!
```

In fact it can be proven that $\lim_n \frac{S_n - nm}{\sqrt{n}\sigma}$ *diverges* as $n \to \infty$ with probability 1. Convergence of random variables and convergence of their distributions are *not* the same thing! This brings out an important distinction between the Renewal Theorem and the Central Limit Theorem on one hand, which are theorems about limits of probabilities, as opposed to the Strong Law of Large Numbers on the other hand, which is about the limit of actual values of the random variables $\frac{1}{n}S_n$.

3.5 Elementary Conditional Probabilities

The full description of the dependency between two or more random variables is described by their *joint* distribution. For a pair X and Y that means describing the probabilities that the pair (X, Y) falls in different subsets of the plane. A Markov chain consists of a sequence of random variables, one for each moment of time: X_0, X_1, X_2, ... Because the random variables are associated with successive moments in time, it is natural to try to describe their joint distribution in terms of their successive dependence on each other. We want to describe how X_1 depends on X_0, and how X_2 depends on (X_0, X_1), and so forth. From that we hope to work out ways to calculate the probabilities of the complicated events of interest. Conditional probabilities and expectations are what we use to describe how one random variable depends on another. These are the topics of this section. Let's start with an example.

Example 3.13. Let D_1 and D_2 be an independent pair of fair dice. Both dice are rolled by someone who only tells us the sum $X = D_1 + D_2$, not what D_1 and D_2 are separately. Suppose we are told that $X = 5$. Now, without knowing any more about the outcomes of the individual dice rolls, we are asked what we think the probability of $D_2 = 3$ is in light of our our knowledge that $X = 5$. We know $P(D_2 = 3) = \frac{1}{6}$, but that does not take into account the extra information we have in knowing that $X = 5$. The possible dice pair outcomes that consistent with $X = 5$ are $(D_1, D_2) = (1, 4), (2, 3), (3, 2), (4, 1)$, each of which has probability $\frac{1}{36}$. Although these account for only $\frac{1}{9}$ of the overall probability, we know that one of these four *did* happen, because we were told that $X = 5$. So based on our knowledge that $X = 5$ we know that these four possibilities account for *all* the possible outcomes consistent with $X = 5$. They are all equally likely, but only one of them has $D_2 = 3$. So we would say the that probability of $D_2 = 3$ *given* that $X = 5$ is $\frac{1}{4}$, the *fraction* of $P(X = 5)$ which corresponds to $D_2 = 3$:

$$\frac{P(X = 5 \text{ and } D_2 = 3)}{P(X = 5)} = \frac{1/36}{1/9} = \frac{1}{4}.$$

This is what we mean by the conditional probability $P(D_2 = 3 \,|\, X = 5) = \frac{1}{4}$.

In general the conditional probability of an event B given an event A is defined to be the fraction of $P(A)$ which corresponds to B:

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

But notice that we need $P(A) > 0$ for this is to be defined.

Definition. *For events A and B, the* conditional probability *of B given A is a value $P(B \mid A)$ which satisfies*

$$P(A \cap B) = P(B \mid A)P(A).$$

Stating it this way avoids the problem of dividing by $P(A) = 0$. If $P(A) > 0$ then $P(B \mid A)$ has to be the value we gave above, but if $P(A) = 0$ then *any value* will work for $P(B \mid A)$.

Suppose that X and Y are discrete random variables with joint distribution $P(X = i \text{ and } Y = j) = p_{ij}$. We work out $P(Y = m \mid X = n)$ by taking $A = \{X = n\}$ and $B = \{Y = m\}$:

$$P(Y = m \mid X = n) = \frac{P(X = n \text{ and } Y = m)}{P(X = n)} = \frac{p_{nm}}{\sum_j p_{nj}}.$$

This is what we did in the dice example above. The distribution of X is

$$p_n = \sum_j p_{nj}.$$

Let's use the notation

$$p_{m|n} = \frac{p_{nm}}{p_n}. \tag{3.12}$$

(If $p_n = 0$ it won't matter what we take $p_{m|n}$ to be. The essential relationship is $p_{nm} = p_{m|n}p_n$, which holds in any case.) The collection of these $p_{m|n}$ values describe the *conditional distribution of* Y given X. Observe that $\sum_m p_{m|n} = 1$ for each individual n; this describes a distribution with respect to the index m which depends on n as a parameter.

If we are given the distribution of X (i.e. the values of $p_n = P(X = n)$) and the conditional distribution of Y given X (i.e. the values of $p_{m|n} = P(Y = m \mid X = n)$) then we can reconstruct the joint distribution:

$$P(X = n \text{ and } Y = m) = p_{nm} = p_{m|n}p_n = P(Y = m \mid X = n)P(X = n).$$

Now suppose there is a third random variable Z. We can form the conditional distribution of Z given X and Y in the same way as above. Writing "$X = n$ and $Y = m$" as "$(X, Y) = (n, m)$" will shorten the notation a bit.

$$P(Z = k \mid X = n \text{ and } Y = m) = \frac{P((X, Y, Z) = (n, m, k))}{P((X, Y) = (n, m))},$$

The the conditional distribution of Z given both X and Y is just the collection of these values for all possible n, m, k. (We won't try to give it a "$p_{\cdots}$" notation.) If these are known along with the distribution of X and the conditional distribution of Y give X, then we can build the "triple" joint distribution

$$\begin{aligned} P(X = n,\ Y = m, Z = k) &= P(Z = k \mid X = n \text{ and } Y = m)P(X = n \text{ and } Y = m) \\ &= P(Z = k \mid X = n \text{ and } Y = m)P(Y = m \mid X = n)P(X = n). \end{aligned}$$

This is what we meant by saying that we can describe the joint distribution in a step-by-step way.

3.5.1 Basic Properties

Here are some basic properties of conditional probabilities which help us work with them. (All of these are subsumed by Propostion 3.8 below.)

a) $P(A) = P(A|B)P(B) + P(A|B^c)P(B^c)$

b) If B_i is a partition (a collection of disjoint events with $P(\cup B_i) = 1$) then

$$P(A) = \sum_i P(A|B_i)P(B_i).$$

c) *Bayes Formula* (reversal of conditioning order): assuming $P(A) > 0$,

$$P(B|A) = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|B^c)P(B^c)}.$$

More generally, if B_i is a partition as in b),

$$P(B_j|A) = \frac{P(A|B_j)P(B_j)}{\sum_i P(A|B_i)P(B_i)}.$$

Observe that a) is just b) with $B_1 = B$, $B_2 = B^c$.

3.5.2 Examples

Here are some examples illustrating the use of conditional probabilities.

Example 3.14. **The Monte Hall Problem.** You are the contestant on a game show. There are three closed doors #1, #2, and #3. The host tells you that there is a new car behind one of them, but there is a goat behind each of the other two. You will get to pick a door and then keep whatever is found behind it. You make an initial choice of one of the doors; lets say you pick #1. But before opening your selected door #1 the host opens a different door, let's say #2, and you see that there is a goat behind it. Now the host asks if you would like to change your choice to the other unopened door, #3, or if you want to stay with your original pick of door #1. What should you do? (We presume you prefer a new car to a new goat.)

This problem became notorious when a reader wrote to Marilyn vos Savant's "Ask Marilyn" column in Parade magazine in 1990 asking her what the answer was. In her column vos Savant said that it was best to switch to the other door. Thousands of people wrote in to disagree with her, claiming that switching made no difference to your chances of winning the car. These dissenters included several with math Ph.D.s disagreed. Even the famous Paul Erdős was convinced vos Savant was wrong, until a computer simulation changed his mind; see [29]. The analysis of this problem is a great example of the use of conditional probabilities

Let C be a random variable which indicates which door hides the new car. We will assume that $P(C = 1) = P(C = 2) = P(C = 3) = \frac{1}{3}$. Those who said switching made no difference reasoned that by showing us the goat behind door #2 the host has told us that $C \neq 2$. Therefore

$$P(C = 1 \,|\, C \neq 2) = \frac{P(C = 1 \text{ and } C \neq 2)}{P(C \neq 2)} = \frac{P(C = 1)}{P(C = 2) + P(C = 3)} = \frac{1/3}{2/3} = \frac{1}{2},$$

and similarly, $P(C = 3 \,|\, C \neq 2) = \frac{1}{2}$. Thus conditional on $C \neq 2$ the other two possibilities, $C = 1$ or 3, are equally likely so there is no benefit in switching.

But now think about another way. If you were given the choice of door #1 or *the better* of doors #2 or 3, certainly you would choose the latter, because

$$P(C \in \{2, 3\}) = \frac{2}{3} > P(C = 1) = \frac{1}{3}.$$

By opening door #2 to reveal the goat the host has shown you which of doors #2 and 3 actually is the better one. So by switching to door #3 after being shown the goat behind door #2 you are actually choosing the better of doors #2 and 3, and thus will get the car with probability $\frac{2}{3}$. This reasoning says you double your chances of winning the car by switching to door #3.

Is one of these arguments wrong? How can we explain their different conclusions? The resolution depends on clarifying exactly what we learned when the host opened door #2. Do we learn something from the fact that it was door #2 that he opened and not door #3? To make this explicit, lets introduce another random variable H which is the number of the door the host opens before giving us the chance to switch. *Everything depends on what we believe about H.* Is H always 2? In other words would the host have opened door #2 even if it held the car? Is H chosen from $\{2, 3\}$ independently of C; might the host might flip a coin to decide whether to open door #2 or #3 and then open that door regardless of its contents? Or is H selected to insure that $H \neq C$?

The show's producers would probably be unhappy if the host opened the door with the car, because then we would know that neither of the unopened doors has the car and there would be no interest left in the game. So lets assume $P(H = C) = 0$ and $P(H = 1) = 0$. Then H must be 3 if $C = 2$, and H must be 2 if $C = 3$. Only when $C = 1$ does the host have any choice of door to open. We don't know how he will decide in that case, but lets assume his selection is governed by some conditional probabilities

$$P(H = 2 \,|\, C = 1) = \gamma, \quad P(H = 3 \,|\, C = 1) = 1 - \gamma,$$

and $P(H = 1 \,|\, C = 1) = 0$. Knowing these we can calculate the distribution of H:

$$P(H = 2) = \sum_1^3 P(H = 2 \,|\, C = i)P(C = i) = \frac{1}{3}(\gamma + 0 + 1) = \frac{\gamma + 1}{3},$$
$$P(H = 3) = 1 - P(H = 2) = \frac{2 - \gamma}{3}.$$

We want $P(C = 3 \,|\, H = 2)$, which we can calclulate using Bayes formula c) above.

$$\begin{aligned} P(C = 3 \,|\, H = 2) &= \frac{P(H = 2 \,|\, C = 3)P(C = 3)}{\sum_{i=1}^3 P(H = 2 \,|\, C = i)P(C = i)} \\ &= \frac{P(H = 2 \,|\, C = 3)}{P(H = 2 \,|\, C = 1) + P(H = 2 \,|\, C = 2) + P(H = 2 \,|\, C = 3)} \\ &= \frac{1}{\gamma + 0 + 1} = \frac{1}{\gamma + 1}. \end{aligned}$$

A similar calculation yields

$$P(C = 2 \,|\, H = 3) = \frac{1}{1 - \gamma + 1 + 0} = \frac{1}{2 - \gamma}.$$

Let S be the remaining door other than #1 and H, the door we will have the option to switch to after learning what H is. What we are interested in is the following.

$$\begin{aligned} P(C = S) &= P(C = 3 \,|\, H = 2)P(H = 2) + P(C = 2 \,|\, H = 3)P(H = 3) \\ &= \frac{1}{\gamma + 1}\frac{\gamma + 1}{3} + \frac{1}{2 - \gamma}\frac{2 - \gamma}{3} = \frac{2}{3}, \text{ and} \\ P(C = 1) &= 1 - P(C = S) = \frac{1}{3}. \end{aligned}$$

This justifies the conclusion that (always) switching doors doubles your probability of winning the car, under the hypothesis that $H \neq 1$ and $H \neq C$. By "always" we mean regardless of whether $H = 2$ or $H = 3$. The conditional probabilities $P(C = S \,|\, H = j)$ are different for $j = 2$ and $j = 3$, but it turns out that in both cases $P(C = S \,|\, H = j) > P(C = 1 \,|\, H = j)$ as long as $0 < \gamma < 1$. So you can't improve the overall strategy by deciding whether or not to switch based on the value of H.

If you rework these calculations assuming that $H = 2$ always, or that H is either 2 or 3 chosen independently of C, then it turns out that the first line of reasoning is correct (see Problem 3.24):

$$P(C = 1 \,|\, H \neq C) = P(C = S \,|\, H \neq C) = \frac{1}{2}.$$

So we see that different assumptions about H lead to different conclusions. But you will probably agree that allowing $H = C$ or $H = 1$ is unreasonable for the game show. On this basis we find that vos Savant was right; you *are* twice as likely to win the car if you switch doors, and a careful conditional probability discussion explains why[1].

[1]This affair embarrassed the many math Ph.D.s who had written vos Savant to disagree and in some cases scold her. Are you wondering how many apologies she got? Read this:

`http://query.nytimes.com/gst/fullpage.html?res=9D0CEFDD1E3FF932A15754C0A967958260`.

Example 3.15. Suppose we have a coin that produces heads with probability p and tails with probability $q = 1 - p$. We flip the coin (independently) n times and let X be the number of heads and Y the number of tails. Since Y is X-determined ($Y = n - X$) we don't expect X and Y to be independent. But suppose we "randomize" n; to be specific, suppose that N is a Poisson random variable with parameter λ: $P(N = n) = \frac{\lambda^n}{n!}e^{-\lambda}$ for $n \geq 0$. We observe N first and then flip our coin N times to determine X and Y. The remarkable fact is that this results in X and Y which *are* independent!

The essential fact is that

$$P(X = k \,|\, N = n) = \binom{n}{k} p^k q^{n-k}, \quad \text{where } q = 1 - p.$$

Therefore we have

$$\begin{aligned} P(X = k \text{ and } Y = \ell) &= P(X = k \text{ and } N = k + \ell) \\ &= P(X = k \,|\, N = k + \ell) P(N = k + \ell) \\ &= \binom{k+\ell}{k} p^k q^\ell \cdot \frac{\lambda^{k+\ell}}{(k+\ell)!} e^{-\lambda} \\ &= \frac{p^k}{k!} \frac{q^\ell}{\ell!} \lambda^{k+\ell} e^{-\lambda}. \end{aligned}$$

On the other hand,

$$\begin{aligned} P(X = k) &= \sum_{n=0}^{\infty} P(X = k \,|\, N = n) P(N = n) \\ &= \sum_{n=k}^{\infty} \binom{n}{k} p^k q^{n-k} \frac{\lambda^n}{n!} e^{-\lambda} \\ &= \sum_{n=k}^{\infty} \frac{p^k}{k!} \lambda^k \frac{q^{n-k}}{(n-k)!} \lambda^{n-k} e^{-\lambda} \\ &= \frac{p^k}{k!} \lambda^k \left(\sum_{n=k}^{\infty} \frac{q^{n-k}}{(n-k)!} \lambda^{n-k} \right) e^{-\lambda} \\ &= \frac{p^k}{k!} \lambda^k e^{\lambda q} e^{-\lambda} \\ &= \frac{p^k}{k!} \lambda^k e^{-p\lambda} \end{aligned}$$

This is a Poisson distribution with parameter $p\lambda$. With a similar calculation we find that

$$P(Y = \ell) = \frac{q^\ell}{\ell!} \lambda^\ell e^{-q\lambda}.$$

So

$$P(X = k) P(Y = \ell) = \frac{p^k}{k!} \lambda^k e^{-p\lambda} \frac{q^\ell}{\ell!} \lambda^\ell e^{-q\lambda} = \frac{p^k}{k!} \lambda^k \frac{q^\ell}{\ell!} \lambda^\ell e^{-\lambda}.$$

3.5.3 Elementary Conditional Expectation

Suppose X is a discrete random variable and A is an event with $P(A) > 0$. We can calculate the conditional expectation $E[X \,|\, A]$ just as we would $E[X]$ but using the conditional distribution of X given A in place of the distribution of X:

$$E[X|A] = \sum_x x P(X = x \,|\, A). \tag{3.13}$$

Notice that this can be written using a restricted expectation:

$$E[X|A] = \sum_x x P(X = x \text{ and } A)/P(A) = \frac{E[X; A]}{P(A)}.$$

As for conditional probabilities we rearrange this into a definition which makes sense even if $P(A) = 0$.

Definition. *For an integrable random variable and an event A the* conditional expectation of X given A *is* $E[X|A]$ *defined by*

$$E[X|A]P(A) = E[X; A].$$

If $P(A) = 0$ then *any* value qualifies as $E[X|A]$.

Example 3.16. Consider again the random variables of Example 3.13.

$$\begin{aligned}
P(D_2 = 1|X = 5) &= 1/4 \\
P(D_2 = 2|X = 5) &= 1/4 \\
P(D_2 = 3|X = 5) &= 1/4 \\
P(D_2 = 4|X = 5) &= 1/4 \\
P(D_2 = 5|X = 5) &= 0 \\
P(D_2 = 6|X = 5) &= 0.
\end{aligned}$$

Therefore

$$E[D_2|X = 5] = (1 + 2 + 3 + 4)/4 = 2.5.$$

Here are some basic properties of conditional expectations. X and Y are assumed to have finite mean and $P(A) > 0$.

a) $E[1_B|A] = P(B|A)$.

b) $E[aX + bY|A] = aE[X|A] + bE[Y|A]$ for any constants a, b.

c) $X(\omega) \le Y(\omega)$ for all ω implies $E[X|A] \le E[Y|A]$.

d) $|E[X|A]| \le E[\,|X|\,|A]$.

e) If B_i is a partition of Ω then

$$E[X] = \sum_i E[X|B_i]P(B_i).$$

3.6 Generalized Conditional Expectation: $E[Y|X]$

Consider Example 3.16 once again. We have been thinking of $E[D_2|X = n]$ one value of n at a time. The value of $E[D_2|X = n]$ will be different for different choices of n. In Example 3.16 we conditioned on $X = 5$. If we condition on $X = 4$ instead then

$$\begin{aligned}
P(D_2 = 1|X = 4) &= 1/3 \\
P(D_2 = 2|X = 4) &= 1/3 \\
P(D_2 = 3|X = 4) &= 1/3 \\
P(D_2 = 4|X = 4) &= 0 \\
P(D_2 = 5|X = 4) &= 0 \\
P(D_2 = 6|X = 4) &= 0
\end{aligned}$$

and so

$$E[D_2|X = 4] = (1 + 2 + 3)/3 = 2.$$

Working out the rest of the cases we find

$$E[D_2|X=n] = \begin{cases} 1 & n=2 \\ 1.5 & n=3 \\ 2 & n=4 \\ & \vdots \\ 5.5 & n=11 \\ 6 & n=12 \end{cases}$$
$$= n/2.$$

We now want to introduce the idea of the *generalized* conditional $E[D_2|X]$. Notice that no value n for $X=n$ is specified in this notation. This is not a typo; we are trying to indicate a way of combining all the different values of $E[D_2|X=n]$ into a single more comprehensive object: $E[D_2|X]$. The connection between the two is that $E[D_2|X]$ is the value that results from using the actual value of X in place of n in $E[D_2|X=n]$. For instance let's say that $X=5$ then the value of $E[D_2|X]$ is $E[D_2|X=5]=2.5$. But if $X=4$ then the value of $E[D_2|X]$ is $E[D_2|X=4]=2$. The value of $E[D_2|X]$ depends on the value that X actually takes, not some specific n which we might select in advance. In our particular example this works out to be

$$E[D_2|X] = X/2.$$

The upshot is that $E[D_2|X]$ is not a mere number but *a new random variable*, an X-determined random variable.

Provided X is a discrete random variable and Y is integrable we can describe the generalized conditional expectation of Y given X this way:

$$E[Y|X] = \Phi(X) \text{ where } \Phi(\cdot) \text{ is the function } \Phi(x) = E[Y|X=x]. \tag{3.14}$$

The difference in notation is subtle. When you see the "$=x$" in "$E[Y|X=x]$" that tells you that we mean the elementary conditional expectation (3.13) using the particular value x for the outcome of X. (This is a number, not a random variable.) When you see "$E[Y|X]$" with no "$=x$" we mean the random variable $E[Y|X]$ which has different values depending on what X is. (*Our notation here is not standard.* What we denote by "$E[Y|X]$" is usually denoted "$E[Y|\sigma(X)]$" or "$E[Y|\mathcal{F}_X]$" in more advanced treatments. We use a simplified notation in order to avoid discussion of σ-algebras, which would take us into measure theory.)

There is a formula which characterizes the generalized conditional expectation. Suppose C is a set of possible X values and consider the expected value of Y restricted to the event $X \in C$.

$$\begin{aligned} E[Y; X\in C] &= \sum_y \sum_{x\in C} yP(Y=y \text{ and } X=x) \\ &= \sum_y \sum_{x\in C} yP(Y=y|X=x)P(X=x) \\ &= \sum_{x\in C} \left[\sum_y yP(Y=y|X=x)\right] P(X=x) \\ &= \sum_{x\in C} E[Y|X=x]P(X=x) \\ &= E[\,E[Y|X]\,; X\in C]. \end{aligned} \tag{3.15}$$

This formula leads to the second part of the following definition of the generalized conditional expectation.

Definition. *If X and Y are random variables and Y is integrable then $E[Y|X]$ is an integrable random variable with the following properties.*

1. *$E[Y|X]$ is X-determined, i.e. $E[Y|X]=\Phi(X)$ for some function $\Phi(\cdot)$.*

2. For any set C of possible values for X,

$$E[Y; X \in C] = E[\,E[Y|X]; X \in C].$$

The generalized conditional probability $P(B|X)$ is just the special case of $Y = 1_B$:

$$P(B|X) = E[1_B|X].$$

Although this definition may seem perplexing in comparison to the simplicity of (3.13) it is actually quite important. The working properties of conditional expectations are expressed in terms of generalized conditionals; see Proposition 3.8 below. Moreover all the properties of *elementary* conditionals (Section 3.5.1 and equation (3.13)) follow from it.

To condition on a particular event A use the generalized conditional on $X = 1_A$. Since $E[Y|X]$ is X-determined we must have $E[Y|X] = \Phi(X)$ where $\Phi(\cdot)$ has only two values, $\Phi(0)$ and $\Phi(1)$. Using part 2 of the definition,

$$E[Y; A] = E[Y|X = 1] = E[\Phi(X); X = 1] = \Phi(1)P(A), \tag{3.16}$$

so $\Phi(1) = \frac{E[Y;A]}{P(A)} = E[Y|A]$. Repeating this for A^c leads to $\Phi(0) = E[Y|A^c]$. In other words we have deduced (3.14) from the definition in the case that X is a Bernouli random variable. In particular,

$$E[Y|1_A] = \Phi(1_A) = E[Y|A]1_A + E[Y|A^c]1_{A^c}.$$

If X is a discrete random variable, with

$$\{X = i\} = B_i,\ P(B_i) > 0,$$

the same argument implies that $\Phi(i) = E[Y|B_i]$, so that

$$E[Y|X] = \sum_i E[Y|B_i]1_{B_i}.$$

The point is that (3.14) follows from the definition, and $\Phi(x) = E[Y|X = x]$ is the *only* function $\Phi(x)$ which works in part 2 of the definition.

Now let's think about the case of continuous random variables. Suppose X and Y have a joint density

$$P(Y \le b \text{ and } X \le a) = \int_{-\infty}^{a}\int_{-\infty}^{b} f(x,y)\,dy\,dx.$$

In this situation we have trouble defining the elementary conditional $E[Y|X = c]$ as in (3.13) because $P(X = c) = 0$. We can guess our way to the correct formula as follows. Let $f_X(x)$ be the marginal density of X and assume $f_X > 0$. Then we can define the *conditional density* $f(y|x)$ by optimistically just following the pattern of our basic definition but with densities in place of probabilities:

$$f(y|x) = \frac{f(x,y)}{f_X(x)}. \tag{3.17}$$

This *conditional density* has properties similar to $p_{m|n}$ in (3.12). For instance it is a density in y for each value of x individually,

$$\int_{-\infty}^{\infty} f(y|x)\,dy = 1 \quad \text{for each } x.$$

We speculate that conditional expectations with respect to X should computed by

$$E[Y|\ X = x] = \int y f(y|x)\,dy. \tag{3.18}$$

This seems reasonable just by analogy with the discrete case. But the true reason for (3.18) is that this formula satisfies our definition of the generalized conditional defined above. Take the right side of (3.18) as the definition of a function

$$\Phi(x) = \int y f(y|x)\,dy.$$

Our claim is that this works in the definition of $E[Y|X] = \Phi(X)$ above. Let's check, using an interval $C = [a, b]$.

$$\begin{aligned} E[Y; X \in C] &= \int_a^b \int y f(x,y)\, dy\, dx \\ &= \int_a^b \left[\int y f(y|x) f_X(x)\, dy \right] dx \\ &= \int_a^b \Phi(x) f_X(x)\, dx \\ &= E[\Phi(X); X \in C]. \end{aligned}$$

This confirms that (3.18) does indeed satisfy property 2 of the definition. *That* is why (3.18) is correct!

Example 3.17. Consider the following joint density.

$$f(x,y) = \begin{cases} x + \frac{3}{2}y^2 & \text{if } 0 \le x \le 1 \text{ and } 0 \le y \le 1 \\ 0 & \text{otherwise.} \end{cases}$$

(You can check that $\iint f(x,y) = 1$.) The marginal density of X is

$$f_X(x) = \int f(x,y)\, dy = \begin{cases} x + \frac{1}{2} & \text{if } 0 \le x \le 1 \\ 0 & \text{otherwise.} \end{cases}$$

So for $0 \le x, y \le 1$ we have

$$f(y|x) = \frac{x + \frac{3}{2}y^2}{x + \frac{1}{2}} = \frac{2x + 3y^2}{2x + 1}.$$

Therefore for $0 \le x \le 1$ we have

$$\Phi(x) = \int y f(y|x)\, dy = \int_0^1 y \frac{2x + 3y^2}{2x + 1}\, dy = \frac{4x + 3}{8x + 4}.$$

(It doesn't matter what $\Phi(x)$ is for other values of x since those never occur as values of X.) So we find

$$E[Y|X] = \frac{4X + 3}{8X + 4}.$$

Our calculations for discrete and jointly continuous random variables show how to find $E[Y|X]$ in those particular cases. In fact $E[Y|X]$ always exists, provided Y is integrable. The conditioning random variable X can be a vector of several different random variables: for instance $X = (Z, W, \Theta, \ldots)$, in which case our existence claim is that there does exist a function $\Phi(z, w, \theta, \ldots)$ for which $\Phi(Z, W, \Theta, \ldots)$ has property 2 of the definition, and so can rightfully be called $E[Y|(Z, W, \Theta, \ldots)]$. The proof of this grand existence claim is part of the general theory of the Kolmogorov model and beyond our purposes here. We will take it for granted.

The really important properties for working with conditionals are expressed most concisely in the generalized conditional point of view. The following proposition collects several such properties. When we write "$X \equiv c$" we mean that X is the random variable that always takes the value c, $P(X = c) = 1$. We can call this a *constant* random variable.

Proposition 3.8. *Suppose X, Y, Z are random variables, with Y, Z (and YZ for 10) integrable.*

1. *If $Y \equiv c$ is a constant random variable then $E[Y|X] \equiv c$.*
2. *For any two constants α, β,*

$$E[\alpha Y + \beta Z|X] = \alpha E[Y|X] + \beta E[Z|X]$$

3. *If $Z \le Y$, then $E[Z|X] \le E[Y|X]$.*

4. *$|E[Y|X]| \le E[|Y||X]$.*

5. *If $X \equiv c$ is a constant random variable then $E[Y|X] \equiv E[Y]$ (a constant random variable).*

6. *Independence of X and Y implies $E[Y|X] \equiv E[Y]$.*

7. *For any X-determined event A,*
$$E[Y; A] = E[\, E[Y|X]\,; A].$$
When $P(A) = 1$ this says that $E[Y] = E\,[E[Y|X]]$. For conditional probabilities this says that $P(B \cap A) = E[P(B|X); A]$ and $P(B) = E[P(B|X)]$.

8. *$E[Y|X] = E[E[Y|(X, Z)]|X]$.*

9. *If X is Z-determined, then $E[Y|(X, Z)] = E[Y|Z]$.*

10. *If Y is X-determined, then*
$$E[Y|X] = Y$$
and more generally
$$E[YZ|X] = YE[Z|X].$$

We will explain some of these and leave others as problems.

Consider item 3 for discrete random variables. $E[Z|X] = \Psi(X)$ and $E[Y|X] = \Phi(X)$, where

$$\Psi(x) = E[Z; X = x]/P(X = x), \quad \Phi(x) = E[Y; X = x]/P(X = x).$$

Since $Z \le Y$ implies $E[Z; X = x] \le E[Y; X = x]$, we have $\Psi(x) \le \Phi(x)$. This implies $E[Z|X] \le E[Y|X]$ as claimed. A similar argument can be given in the jointly continuous case.

For item 6 we will show that the constant random variable $\Gamma \equiv E[Y]$ satisfies the definition of $E[Y|X]$. As a constant Γ is X-dependent. For any set C the random variable $1_C(X)$ is independent of Y, since X is. Using that independence in (3.10) we have

$$E[Y; X \in C] = E[Y 1_C(X)] = E[Y]E[1_C(X)] = E[Y]P(X \in C) = E[\Gamma; X \in C].$$

This implies item 6. Item 5 is a special case because the constant random variable $Y \equiv c$ is independent of X.

Item 7 is just a restatement of part 2 of the definition, since an X-determined event must be of the form $A = \{X \in C\}$ for some set of values C. We get conditional probabilities by taking $Y = 1_B$:

$$P(B|X) = E[1_B|X].$$

In particular

$$\begin{aligned} P(B \cap A) &= E[1_B; A] \\ &= E[E[1_B|X]; A] \\ &= E[P(B|X); A]. \end{aligned}$$

Item 8 is called the *Tower Law*. It says we can find generalized conditionals in stages by conditioning on more "informative" random variables first and less informative ones after. To verify it, let $\Gamma = E[E[Y|(X, Z)]|X]$. We want to see that Γ satisfies the defininiton of $E[Y|X]$. Γ is X-determined because it is $E[\cdot|X]$ of something. For the second part, if C is a set of X values, then

$$\begin{aligned} E[Y; X \in C] &= E[Y; (X, Z) \in C \times \mathbb{R}] \\ &= E[E[Y|(X, Z)]; (X, Z) \in C \times \mathbb{R}] \\ &= E[\Gamma; (X, Z) \in C \times \mathbb{R}] \\ &= E[\Gamma; X \in C]. \end{aligned}$$

Item 9 is because if X is Z-determined, then being (X, Z)-determined is equivalent to being Z-determined. Therefore the definitions of $E[Y|(X, Z)]$ and $E[Y|Z]$ coincide.

To explain item 10 let's start with a Bernoulli random variable for Y: $Y = 1_B$ where B is an X-determined event. The claim is that $E[1_B Z|X] = 1_B E[Z|X]$. We check this by verifying that the right side satisfies the definition of the left side. The right side is X-determined because both factors are. We know

$$E[|1_B E[Z|X]|] \leq E[|E[Z|X]] < \infty$$

since $E[Z|X]$ must be integrable by its own definition. Next consider any set C of possible X values.

$$\begin{aligned} E[1_B E[Z|X]; X \in C] &= E[E[Z|X]; X \in C \cap B] \\ &= E[Z; X \in C \cap B], \text{ because } C \cap B \text{ is } X\text{-determined} \\ &= E[1_B Z; X \in C]. \end{aligned}$$

This completes the verification that $E[1_B Z|X] = 1_B E[Z|X]$, which is what we wanted to show in the special case of $Y = 1_B$. Now suppose Y is a discrete X-determined random variable taking values y_i. The events

$$B_i = \{Y = y_i\}$$

are all X-determined, and we can write

$$Y = \sum_i y_i 1_{B_i}.$$

Using what we just showed for each of the 1_{B_i} together with part 1) of the proposition, we have

$$E[YZ|X] = E\left[\sum_i y_i 1_{B_i} Z \middle| X\right] = \sum_i y_i E[1_{B_i} Z|X] = \sum_i y_i 1_{B_i} E[Z|X] = Y E[Z|X].$$

If the set of y_i is infinite these are infinite series and more technical justification is needed. The details of the continuous case are more involved as well. But this calculation gives you the general idea.

Here are a couple examples which use properties of the generalized conditional.

Example 3.18. Consider again Example 3.15 above, in which we first observe a Poisson random variable N (parameter λ) and then toss a coin with $P(\text{heads}) = p$ the number of times given by N and let X be the number of heads observed. One way to describe the joint distribution is to say that N is Poisson, and that the *conditional distribution* of X given N is binomial with parameters (N, p). I.e.

$$P(X = k|N = n) = p^k (1-p)^{n-k} \binom{n}{k} 1_{0 \leq k \leq n}.$$

We can use 7) of the proposition above to calculate $E[X]$. First, using the mean of binomial distributions,

$$E[X|N = n] = pn,$$

or as a generalized conditional,

$$E[X|N] = pN.$$

Therefore

$$E[X] = E[E[X|N]] = E[pN] = pE[N] = p\lambda.$$

Don't make the mistake of thinking that X is a binomial random variable. It's not; we worked out its distribution in Example 3.15.

Example 3.19. (From [52]) You are lost in a system of underground tunnels. You find yourself in a chamber with three tunnels leading out of it. Let's name them #1, #2, #3 but the chamber is pitch black so you can't tell one from another. If you take tunnel #1 you will reach safety in 2 minutes. If you take tunnel #2 in 3 minutes you find yourself back in the same chamber. If you take tunnel #3 you again find yourself back in the same chamber but after 5 minutes. Each time you reach this chamber you grope around until randomly finding one of the tunnels to try (equal probabilities of finding each). The problem is to determine

the expected time until you reach safety, starting from the chamber. Let Y indicate the first tunnel you choose to follow and X your total time to reach safety; we want $E[X]$. Here is what we know.

$$P(Y = i) = 1/3 \text{ for each of } i = 1, 2, 3$$

and

$$\begin{aligned} E[X|Y = 1] &= 2 \\ E[X|Y = 2] &= 3 + E[X] \\ E[X|Y = 3] &= 5 + E[X]. \end{aligned}$$

Using these, we can calculate

$$E[X] = E[E[X|Y]] = \frac{1}{3} \cdot 2 + \frac{1}{3} \cdot (3 + E[X]) + \frac{1}{3} \cdot (5 + E[X]) = \frac{1}{3} \cdot (10 + 2E[X]).$$

This is an equation that $E[X]$ must satisfy; solving we find that

$$E[X] = 10.$$

3.7 The Markov Property

We now turn our attention back to Markov chains to see how they are related to some of the things we have discussed in this chapter, especially conditional expectations.

The first thing to note is that transition probabilities are really conditional probabilities. Equation (2.2) says that

$$\begin{aligned} P(X_{0:k+1} = s_{0:k+1}) &= \mu_{s_0} \prod_{i=1}^{k+1} p_{s_{i-1},s_i} \\ &= \left[\mu_{s_0} \prod_{i=1}^{k} p_{s_{i-1},s_i} \right] p_{s_k,s_{k+1}} \\ &= P(X_{0:k} = s_{0:k}) p_{s_k,s_{k+1}}. \end{aligned}$$

In other words

$$P(X_{k+1} = s_{k+1}|X_{0:k} = s_{0:k}) = p_{s_k,s_{k+1}}. \tag{3.19}$$

Forming a sum over s_{k+1} leads to

$$\begin{aligned} E[f(X_{k+1})|X_{0:k} = s_{0:k}] &= \sum_{s_{k+1} \in \mathcal{S}} p_{s_k,s_{k+1}} f(s_{k+1}) \\ &= \mathbf{P} f(s_k). \end{aligned}$$

As a generalized conditional expectation,

$$E[f(X_{k+1})|X_{0:k}] = \mathbf{P} f(X_k). \tag{3.20}$$

In general a conditional on $X_{0:k}$ would be an $X_{0:k}$-determined random variable, something of the form $\Phi(X_{0:k})$. But in (3.20) the right side depends *only* on X_k itself, not the earlier values $X_0, \ldots, X_{k-1}$. This is the Markov property and equation (3.20) is its most succinct expression.

There are some other expressions of the Markov property. For instance applying the Tower Law gives us the conditional properties of multiple steps.

$$E[f(X_{k+2})|X_{0:k}] = E[E[f(X_{k+2})|X_{0:k+1}]|X_{0:k}] = E[\mathbf{P} f(X_{k+1})|X_{0:k}] = \mathbf{P}\mathbf{P} f(X_k) = \mathbf{P}^2 f(X_k).$$

Of course this pattern continues to X_{k+m} with $\mathbf{P}^m$. In fact let's redo the calculation leading to (3.20) by going m steps beyond k rather than just one. And to make it easier to follow lets use s_i for the states up to k and a_j for those after k: $X_{k+j} = a_j$.

$$\begin{aligned} P(X_{0:k} = s_{0:k} \text{ and } X_{k+1:k+m} = a_{1:m}) &= \left[\mu_{s_0} \prod_{i=1}^{k} p_{s_{i-1},s_i}\right] \left[p_{s_k,a_1} \prod_{j=2}^{m} p_{a_{j-1},a_j}\right] \\ &= P(X_{0:k} = s_{0:k}) P_{s_k}(X_{1:m} = a_{1:m}). \end{aligned}$$

Notice the second factor on the right especially. What was $X_{k+j} = a_j$ on the left has become $X_j = a_j$ on the right. Consequently

$$P(X_{k+1:k+m} = a_{1:m} | X_{0:k} = s_{0:k}) = P_{s_k}(X_{1:m} = a_{1:m}). \tag{3.21}$$

For a function f of m variables this becomes

$$E[f(X_{k+1,m+k+1})|X_{0:k} = s_{0:k}] = E_{s_k}[f(X_{1:m})].$$

A further extension to functions $\Phi(a_{1:\infty})$ of infinite sequences of states says that

$$E[\Phi(X_{k+1:\infty})|X_{0:k} = s_{0:k}] = E_{s_k}[\Phi(X_{1:\infty})],$$

or

$$E[\Phi(X_{k+1:\infty})|X_{0:k} = s_{0:k}] = w(X_k) \text{ where } w(x) = E_x[\Phi(X_{1:\infty})]. \tag{3.22}$$

This is the same as (3.20) except that we are taking the conditional expectation of a function of the entire future of the chain, not just the next state. (The proof of this extension from functions of m variables to infinite sequences requires the more advanced techniques of measure theory, which we are deliberately not pursuing.)

3.7.1 Hitting Probability Equations

The equations for hitting probabilities that we considered in Section 2.2 were obtained simply from heuristic considerations. In fact they follow from the expression (3.22) of the Markov property. Consider the first contact time $\mathcal{T}_C$ of some set of states $C \subseteq \mathcal{S}$. We will focus on the event $\mathcal{T}_C < \infty$, i.e. that the state of the chain is in C either initially or at some time in the future. The indicator random variable for this event is a function of the full outcome of the chain:

$$1_{\mathcal{T}_C < \infty} = \Phi(X_{0:\infty})$$

where

$$\Phi(s_{0:\infty}) = \begin{cases} 1 & \text{if } s_n \in C \text{ for some } n \geq 0 \\ 0 & \text{if no } s_n \text{ belongs to } C. \end{cases}$$

We are interested in

$$u(i) = P_i(\mathcal{T}_C < \infty) = E_i[\Phi(X_{0:\infty})].$$

For initial states $i \in C$ we know that $u(i) = P_i(\mathcal{T}_C < \infty) = 1$, since if $X_0 = i \in C$ then $\mathcal{T}_C = 0$.

Now consider initial states $X_0 = i \notin C$. Then $\mathcal{T}_C$ is determined by $X_{1:\infty}$; to be precise, $\Phi(X_{0:\infty}) = \Phi(X_{1:\infty})$. The Markov property in the form (3.22) says that

$$E_i[\Phi(X_{1:\infty})|X_1] = u(X_1).$$

Combining these things with the Tower Law we find that

$$\begin{aligned} u(i) &= E_i[\Phi(X_{0:\infty})] \\ &= E_i[\Phi(X_{1:\infty})] \\ &= E_i[E_i[\Phi(X_{1:\infty})|X_1]] \\ &= E_i[u(X_1)] \\ &= \sum_{j \in \mathcal{S}} p_{i,j} u(j). \end{aligned}$$

This is equation (2.5). For a similar derivation of the mean hitting time equations (2.10) see Problem 3.26.

3.7.2 Stopping Times and the Strong Markov Property

There is yet another generalization of the Markov Property which allows the k in (3.21) to be replaced by a time-valued random variable $\mathcal{K}$ of a special type called a *stopping time*. A stopping time is a time-valued random variable $\mathcal{K}$ with the special property that at any moment of time the question "has $\mathcal{K}$ happend yet?" can be be answered based on the history of the chain up to that moment. In our present context to be time-valued means the possible values of $\mathcal{K}$ are $0, 1, 2, \ldots$ and possibly ∞. Here is the formal definition.

Definition. *A random variable $\mathcal{K}$ taking values in $\{0, 1, 2, \ldots, \infty\}$ with the property that for each n*

$$\{\mathcal{K} \le n\} \text{ is } X_{0:n}\text{-determined} \tag{3.23}$$

is called a stopping time.

The time-valued random variables $\mathcal{T}_C$ and $\mathcal{T}_a^+$ are examples of stopping times. For instance whether or not $\mathcal{T}_C \le 5$ is true can be determined by examining $X_0, \ldots X_5$ to see if any of those states are in C or not. An example of a time-valued random variable with is *not* a stopping time is the *last* time $X_n \in C$ (or ∞ if there is no last time).

$$\mathcal{L}_C = \max\{n : X_n \in C\};$$

To know that $\mathcal{L}_C \le 5$ requires examining all the *future* states $X_6, X_7, \ldots$ to be sure that none of them belong to C; it can't be determined just from $X_{1:5}$. That's why $\mathcal{L}_C$ is not a stopping time.

The *strong* Markov property says that the k in the Markov property (3.21) can be replaced by a stopping time $\mathcal{K}$. However (3.21) does not make much sense if $\mathcal{K} = \infty$ because there would be nothing that comes after $\mathcal{K}$. So we will phrase it this way, limiting the statement to the event $\mathcal{K} < \infty$:

$$P(\mathcal{K} < \infty \text{ and } X_{\mathcal{K}+1:\mathcal{K}+m} = a_{1:m} | X_{0:\mathcal{K}}) = P_{X_{\mathcal{K}}}(X_{1:m} = a_{1:m}) 1_{\mathcal{K}<\infty}. \tag{3.24}$$

(A strong Markov version of (3.22) involving the conditional expectation of a function $\Phi(X_{\mathcal{K}+1,\infty})$ of the entire future after $\mathcal{K}$ is also true. But the above version will be enough for our purposes.)

We want to justify the strong Markov property. Let G refer to the event

$$G = \{\mathcal{K} < \infty \text{ and } X_{\mathcal{K}+1:\mathcal{K}+m} = a_{1:m}\}.$$

(It is important to keep in mind that the states of $a_{1:m}$ are part of the specification of G. The a_i are not to be viewed as variables in the following but as fixed values.) To justify (3.24) we need to show that the right side of (3.24) has the two properties required by the definition of the left side. The first property is that $P(G|X_{0:\mathcal{K}})$ is $X_{0:\mathcal{K}}$-determined. In other words the proposed expression for $P(G|X_{0:\mathcal{K}})$ should be something that can be evaluated knowing $s_{0:k} = X_{0:\mathcal{K}}$ but nothing more about the trajectory of the chain; it should be obtained by plugging $s_{0:k} = X_{0:\mathcal{K}}$ into some kind of function $\Gamma(s_{0:k})$ of a finite/infinite sequence $s_{0:k}$ of states. Now suppose we are told what $s_{0:k} = X_{0:\mathcal{K}}$ is. Then we *can* determine the exact value of the right side of (3.24). If $s_{0:k}$ is an infinite sequence then $\mathcal{K} = \infty$ so the right side is 0. If $s_{0:k}$ is a finite sequence then the number of terms tells us the value of $k = \mathcal{K}$ and the last term tells us the sstate $X_{\mathcal{K}} = X_k = s_k$. The right side of (3.24) becomes

$$P_{X_{\mathcal{K}}}(X_{1:m} = a_{1:m}) 1_{\mathcal{K}<\infty} = P_{s_k}(X_{1:m} = a_{1:m}) = p_{s_k,a_1} p_{a_1,a_2} \cdots p_{a_{m-1},a_m}.$$

So the function $\Gamma(\cdot)$ of finite/infinite sequences defined by

$$\Gamma(s_{0:k}) = \begin{cases} p_{s_k,a_1} p_{a_1,a_2} \cdots p_{a_{m-1},a_m} & \text{if } k < \infty \\ 0 & \text{if } k = \infty \end{cases}$$

allows us to express the right side of (3.24) as $\Gamma(X_{0:\mathcal{K}})$, as desired.

The second part of the definition of $P(G|X_{0:\mathcal{K}})$ is that

$$P(G \text{ and } X_{0:\mathcal{K}} \in C) = E[\Gamma(X_{0:\mathcal{K}}); X_{0:\mathcal{K}} \in C]$$

should be correct for any set C of finite/infinite sequences of states. Suppose we can verify that for any particular string of states $s_{0:k}$ the formula

$$P(G \text{ and } X_{0:\mathcal{K}} = s_{0:k}) = P(X_{0:\mathcal{K}} = s_{0:k})\Gamma(s_{0:k}) \tag{3.25}$$

is correct. By adding this up for the different choices of $s_{0:k} \in C$ the second part of the definition will follow. So to check (3.25) we need to *carefully* work out both sides to confirm that they agree. Notice that if $k = \infty$ then both sides are 0, so we can assume k is finite. Consider any finite sequence of states $s_{0:k}$.

$$P(G \text{ and } X_{0:\mathcal{K}} = s_{0:k}) = P(X_{0:k} = s_{0:k},\ X_{k+1,k+m} = a_{1:m},\ \text{and } \mathcal{K} = k).$$

Now *because $\mathcal{K}$ is a stopping time* whether or not $\mathcal{K} = k$ is determined by the specific states $X_{0:k} = s_{0:k}$. So if $s_{0:k}$ is a finite sequence of states for which $\mathcal{K} = k$ then the "$\mathcal{K} = k$" in the above probability is redundant; it is simply

$$\begin{aligned} P(G \text{ and } X_{0:\mathcal{K}} = s_{0:k}) &= P(X_{0:k} = s_{0:k},\ X_{k+1,k+m} = a_{1:m}) \\ &= P(X_{0:k} = s_{0:k})P(X_{k+1,k+m} = a_{1:m} | X_{0:k} = s_{0:k}) \\ &= P(X_{0:k} = s_{0:k})P_{s_k}(X_{1,m} = a_{1:m}) \\ &= P(X_{0:\mathcal{K}} = s_{0:k})\Gamma(s_{0:k}). \end{aligned}$$

And if $s_{0:k}$ is a finite sequence of states for which $\mathcal{K} \neq k$ then $X_{0:\mathcal{K}} \neq s_{0:k}$ so both sides of (3.25) are 0. Either way, (3.25) is correct, completing our verification of (3.24).

3.7.3 Long Run Results for Chains

To finish this chapter we will put the strong Markov property to work. Suppose $a \in \mathcal{S}$ is a recurrent state a. (See Theorems 2.6 b) and 4.2.) We want to consider the sequence of times that the chain is at state a. Starting from $X_0 = a$ let $\mathcal{T}_a^{(1)} = \mathcal{T}_a^+$, the first return time as defined on page 13. Then recursively define the subsequent return times by

$$\mathcal{T}_a^{(k+1)} = \min\{n > \mathcal{T}_a^{(k)} : X_n = a\}, \text{ or } +\infty \text{ if this set is empty.}$$

Each $\mathcal{T}_a^{(k)}$ is a stopping time, because the event $\{\mathcal{T}_a^{(k)} \leq n\}$ is something we can determine by examining the states $X_{0:n}$ to see if the state a occurs at least k times. In brief, $\{\mathcal{T}_a^{(k)} \leq n\}$ is $X_{0:n}$-determined. Our first task is to show that, assuming a is recurrent, all $\mathcal{T}_a^{(k)}$ are finite (with probability 1) and that the waiting times between them

$$\mathcal{W}_k = \mathcal{T}_a^{(k)} - \mathcal{T}_a^{(k-1)}$$

are independent, identically distributed random variables. (Take $\mathcal{T}_a^{(0)} = 0$, so that $\mathcal{W}_1 = \mathcal{T}_a^{(1)} - 0 = \mathcal{T}_a^+$ is well-defined.) Once we prove the lemma we will know that all the $\mathcal{W}_k$ have the same distribution as $\mathcal{T}_a^+$.

Lemma 3.9. *Suppose a is a recurrent state for a Markov chain. Then $P_a(\mathcal{T}_a^{(k)} < \infty) = 1$ for every $k \geq 1$ and the $\mathcal{W}_k$ form an i.i.d. sequence.*

Proof. By the recurrence hypothesis, $\mathcal{T}_a^{(1)} = \mathcal{T}_a^+$ is finite with probability 1. We proceed by induction: assume $\mathcal{T}_a^{(k)}$ is finite with probability 1. Let Φ be the function of infinite sequences $s_{1:\infty}$ which identifies the first m for which $s_m = a$. In other words

$$\mathcal{T}_a^+ = \Phi(X_{1:\infty}).$$

Then

$$\begin{aligned} \mathcal{W}_{k+1} &= \Phi(X_{\mathcal{T}_a^{(k)}+1:\infty}) \\ \mathcal{T}_a^{(k+1)} &= \mathcal{T}_a^{(k)} + \Phi(X_{\mathcal{T}_a^{(k)}+1:\infty}). \end{aligned}$$

The strong Markov property tells us that

$$\begin{aligned} P_a(\mathcal{W}_{k+1} = \ell | X_{0:\mathcal{T}_a^{(k)}}) &= P_a(\Phi(X_{\mathcal{T}_a^{(k)}+1:\infty}) = \ell | X_{0:\mathcal{T}_a^{(k)}}) \\ &= P_{X_{\mathcal{T}_a^{(k)}}}(\Phi(X_{1:\infty}) = \ell) \\ &= P_a(\mathcal{T}_a^+ = \ell). \end{aligned} \tag{3.26}$$

Summing over ℓ we see that $P_a(\mathcal{T}_a^+ < \infty) = 1$ implies that

$$P_a(\mathcal{W}_{k+1} < \infty | X_{0:\mathcal{T}_a^{(k)}}) = 1$$

and therefore

$$P_a(\mathcal{W}_{k+1} < \infty) = E_a[P_a(\mathcal{W}_{k+1} < \infty | X_{0:\mathcal{T}_a^{(k)}})] = 1.$$

By induction this implies the finiteness assertion for $\mathcal{T}_a^{(k+1)} = \mathcal{T}_a^{(k)} + \mathcal{W}_{k+1}$. Moreover because the right side of (3.26) does not depend on $X_{0:\mathcal{T}_a^{(k)}}$ this means that $\mathcal{W}_{k+1}$ is independent of $X_{0:\mathcal{T}_a^{(k)}}$ and has the same distribution as $\mathcal{T}_a^+$. Since all of $\mathcal{W}_1, \ldots, \mathcal{W}_k$ are $X_{0:\mathcal{T}_a^{(k)}}$-determined it follows that $\mathcal{W}_{k+1}$ is independent of $\mathcal{W}_1, \ldots, \mathcal{W}_k$. □

A consequence of this lemma is that the $\mathcal{T}_a^{(k)}$ are the partial sums of an i.i.d. sequence:

$$\mathcal{T}_a^{(k)} = \sum_{i=1}^{k} \mathcal{W}_i.$$

The following theorem now harvests the application of the Strong Law (Theorem 3.6) and Renewal Theorem (Theorem 3.5) to this observation.

Theorem 3.10. *Suppose X_n is a Markov chain with $X_0 = a$ where a is a recurrent state. Then*

$$\frac{1}{n}\sum_{k=1}^{n} 1_a(X_k) \to 1/r_a \text{ as } n \to \infty \text{ with probability } 1,$$

where $r_a = E_a[\mathcal{T}_a^+]$ (the case of $r_a = \infty$ included). If in addition a has period 1 then

$$p_{a,a}(n) \to 1/r_a \text{ as } n \to \infty.$$

We will see later that this generalizes considerably.

Proof. We can apply the standard Law of Large Numbers for i.i.d. random variables as follows. Let $\mathcal{T}_a^{(1)} < \mathcal{T}_a^{(2)} < \cdots < \mathcal{T}_a^{(k)} < \cdots$ be the sequence of return times to a and $\mathcal{W}_k = \mathcal{T}_a^{(k)} - \mathcal{T}_a^{(k-1)}$ the waiting times between visits, as defined above. The mean of the $\mathcal{W}_k$ is

$$E_a[\mathcal{W}_k] = E_a[\mathcal{T}_a^+] = r_a.$$

By the Strong Law (Theorem 3.6),

$$\frac{1}{\ell}\sum_{j=1}^{\ell} \mathcal{W}_j \to r_a \text{ almost surely as } \ell \to \infty.$$

Now $\sum_{j=1}^{\ell} \mathcal{W}_j = \mathcal{T}_a^{(\ell)}$. So the above says that

$$\frac{\mathcal{T}_a^{(\ell)}}{\ell} \to r_a.$$

Taking reciprocals,

$$\frac{\ell}{\mathcal{T}_a^{(\ell)}} \to \frac{1}{r_a} \text{ almost surely as } \ell \to \infty.$$

Each $n \geq 0$ falls between two successive $\mathcal{T}^k$: $\mathcal{T}^\ell \leq n < \mathcal{T}^{\ell+1}$. With n and ℓ related this way $\ell \to \infty$ as $n \to \infty$ because all $\mathcal{T}^\ell$ are finite. Also notice that $\sum_{k=1}^n 1_a(X_k) = \ell$. Therefore

$$\frac{\ell}{\mathcal{T}^\ell} \geq \frac{\ell}{n} = \frac{\sum_{k=1}^n 1_a(X_k)}{n} > \frac{\ell}{\mathcal{T}^{\ell+1}} = \frac{\ell}{\ell+1}\frac{\ell+1}{\mathcal{T}^{\ell+1}}.$$

Both sides of this converge to $1/r_a$ as $\ell \to \infty$. This shows that

$$\frac{\sum_{k=1}^n 1_a(X_k)}{n} \to \frac{1}{r_a} \text{ almost surely as } n \to \infty.$$

The second part of the theorem follows by applying the Renewal Theorem to the $\mathcal{W}_k$. We have that $p_{a,a}(n) = P_a(X_n = a) = P(\mathcal{T}_a^{(k)} = n \text{ for some } k)$. Since $\mathcal{T}_a^{(k)} = \sum_{i=1}^k \mathcal{W}_i$ Theorem 3.5 says that

$$\lim_{n\to\infty} p_{a,a}(n) = \frac{1}{E[\mathcal{W}_i]} = \frac{1}{r_a}.$$

□

Problems

Problem 3.1
Suppose X is geometric random variable with parameter $p = 2/3$. Find $P(X \text{ is an odd number})$. Suppose Y is a Poisson random variable with parameter $\lambda = 2$. What is $P(Y > 4)$? (See Examples 3.1, 3.5 and page 234 for the definitions of these distributions.)

. . . . . SimpleCalc

Problem 3.2
Calculate the mean and variance of a Poisson random variable with parameter λ. (This will be similar to the calculation for binomial random variables, except that we use the Taylor series $e^x = \sum_{n=0}^\infty \frac{1}{n!}x^n$ instead of the binomial formula.)

. . . . . Poisson

Problem 3.3
A *Cauchy* random variable has density

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

Explain why a Cauchy random variable is *not* integrable.

. . . . . Cauchy

Problem 3.4
Suppose X is a continuous random variable with $P(X \geq 0) = 1$. (That means $f(x) = 0$ for all $x < 0$.) Assume also that $E[X] < \infty$. Let $F(x)$ be the distribution function $F(x) = P(X \leq x)$. (There is a little more on this in Section A.2.) Derive the formula

$$E[X] = \int_0^\infty 1 - F(x)\,dx.$$

by starting with the formula we gave to the define $E[X]$, writing $x = \int_0^x 1\,dy$, and then changing the order of integration in the double integral. This is the analogue of (3.8) for continuous random variables.

. . . . . DistnMean

Problem 3.5
Suppose D_1 and D_2 are the results of two independent dice rolls. Use the independence of D_1 and D_2 to

work out the probabilities $P(D_1 + D_2 = k)$. Your results should agree with the formula given on page 6. If there are *three* independent dice, D_1, D_2 and D_3 what is $P(D_1 + D_2 + D_3 = 7)$?

. . . . . Dice

Problem 3.6
Suppose D_1 and D_2 are the results of two independent dice rolls. Let Z be the sum of D_1 and D_2 reduced modulo 6:

$$Z = \begin{cases} D_1 + D_2 & \text{if } D_1 + D_2 \leq 6 \\ D_1 + D_2 - 6 & \text{if } D_1 + D_2 > 6. \end{cases}$$

Show that D_1 and Z are independent, that D_2 and Z are independent, but that D_1, D_2 and Z taken together are *not* independent!

. . . . . 3Dep

Problem 3.7
In Example 3.11 we calculated the density of $S = X + Y$, where X and Y are independent uniform random variables on $[0, 1]$. Illustrate this with a simulation. Produce two lists X and Y of uniform samples, add them to get a list of samples of S, and produce a histogram to view the results. In your simulation, what fraction of the samples fell in the interval $[.5, 1.5]$? Compare that to the theoretical value of $P(.5 \leq Z \leq 1.5)$.

. . . . . SumU

Problem 3.8
If X and Y are independent standard normal random variables, show that $X^2 + Y^2$ is exponential with $\lambda = 1/2$.

. . . . . NormRad

Problem 3.9
Suppose X and Y are independent, both with finite variance. Show that

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y).$$

. . . . . VarSum

Problem 3.10
Work out $E[X^2]$ and $\text{Var}(X)$ where X is as in Example 3.15. You can do this in the same way as Example 3.18

. . . . . Xvar

Problem 3.11
Suppose X and Y are independent Poisson random variables with parameters λ_X and λ_Y. Show that $X + Y$ is also a Poisson random variable and determine it's parameter.

. . . . . PP

Problem 3.12
Explain why the hypotheses of Theorems 3.2 and 3.3 fail in Example 3.6. What about Theorem 3.4 – is it applicable to the example? Is what it claims consistent with what we found in the example?

. . . . . NConv

Problem 3.13
Suppose we toss a (fair) coin independently and repeatedly. Explain why the probability of eventually seeing a head must be 1. What properties of the Kolmogorov model does your reasoning depend on?

. . . . . AllHeads

Problem 3.14
Let X_i be an i.i.d. sequence of exponential random variables with parameter 1, and $S_n = X_1 + \cdots X_n$ the partial sums.

a) Show that S_n has density

$$f_n(x) = \frac{x^{n-1}}{(n-1)!}e^{-x} \text{ for } x > 0,$$

and $f_n(x) = 0$ for $x \leq 0$.

b) Explain why $\{S_{n+1} \leq t\} \subseteq \{S_n \leq t\}$ and therefore

$$P(S_n \leq t < S_{n+1}) = P(S_n \leq t) - P(S_{n+1} \leq t).$$

c) Calculate the value of $P(S_n \leq t < S_{n+1})$ (Hint: $\int_0^t f_n(x) - f_{n+1}(x)\,dx = ?$)

d) Define the random variable N to be the largest $n \geq 0$ for which $S_n \leq \lambda$. (Take $S_0 = 0$.) What kind of random variable is N?

. . . . . SumExpon

Problem 3.15
Suppose that D_i is an i.i.d. sequence of random variables with the uniform distribution on $\{0, \ldots, 9\}$. Show that

$$U = \sum_{i=1}^{\infty} D_i 10^{-i}$$

is a uniform random variable on $[0, 1]$.

. . . . . DigUnif

Problem 3.16
The file `P1S.mat` contains 10000 sampled values (each) of two random variables, X and Y. By examining the data (as we have illustrated with several examples in class) decide what you think the distributions of these random variables are, including the values of any parameters. Download the file and then read the data into Matlab with the command `load P1S` . The file will need to be in your default directory for Matlab to find it.

. . . . . Data

Problem 3.17
If U is uniform on $[0, 1]$ and $0 < p < 1$ then

$$Z = \begin{cases} 1 & \text{if } U < p \\ 0 & \text{otherwise} \end{cases}$$

is a Bernouli random variable with parameter p. Explain why this is so. In MATLAB this could be implemented by `Z=rand()<p`. Using this and the observation of Example 3.9 write an m-file for a command `binomsample(n,p,size)` to produce a pseudo-random number from the binomial distribution with parameters (n, p).

. . . . . BinomSim

Problem 3.18
Write an m-file for a command `randgeo(p,size)` to produce random values from a geometric distribution with parameter p, using the F^* method. (See Section A.3.2.)

. . . . . RandGeo

Problem 3.19
Suppose X is a nonnegative random variable with the property that

$$P(X > t+1 \,|\, X > t) < 1 - \epsilon$$

for some $0 < \epsilon < 1$ and all $t \geq 0$. Show that

$$P(X > n) \leq (1-\epsilon)^n \text{ for all integers } n \geq 1.$$

Use formula of Problem 3.4 to prove that

$$E[X] \leq \frac{1}{\epsilon}.$$

(Note that there is no assumption here that X is a discrete random variable. You may want to bound $P(X > t)$ above by $P(X > n)$ for some integer n related to t in some way.)

..... MLbound

Problem 3.20
(Taken from [25].) There are five coins, indistinguishable by touch. Two are conventional, with a head on one side and a tail on the other. Two of them have a head on both sides. One has a tail on both sides. You choose one of the coins without looking as you do (so each of them is equally likely). Then still without looking at the coin you flip it and finally look to find that the side facing up has a head on it. What is the probability that the side facing down is also a head?

You can set this up by letting C be a random variable giving the number of heads on the coin you draw ($C = 0$, 1, or 2). Let $U = 1$ if the flipped coin lands with a head facing up (and $U = 0$ if a tail is up), and similarly let $D = 1$ or 0 for the side facing down. If you start by writing down $P(C = k)$ and $P(U = i,\ D = j \,|\, C = k)$ you should be able to work out everything you need.

..... HT5

Problem 3.21
Suppose the lifetime of a machine is a random variable T with p.m.f. $P(T = k) = \frac{1}{N+1}$ for $k = 0, 1, \ldots, N$. Find the conditional mean of T given that $T > n$: $E[T \,|\, T > n]$.

..... CMT

Problem 3.22
Suppose X_i are independent (fair) dice rolls, what is $P(\lim_{n\to\infty} X_n \text{ exists }) =$? Justify your answer.

..... DLim

Problem 3.23
Suppose X_i is an i.i.d. sequence with $P(X_i = j) = p_j$; $j = 0, 1, \ldots, n$. Let Y be the value of X_i for the smallest i with $X_i \neq 0$. What is the distribution of Y?

..... Xnot0

Problem 3.24
In the Monte Hall problem, Example 3.14, assume that H is independent of C with

$$P(H = 2) = p, \quad P(H = 3) = 1 - p,$$

where $0 < p \leq 1$. Verify that $P(C = 1 \,|\, H \neq C) = P(C = S \,|\, H \neq C) = \frac{1}{2}$.

..... MHall

Problem 3.25
The Two Envelope Problem. You are given two sealed envelopes. One of them contains $\$X$ and the other contains $\$2X$, but you don't know which is which. You get to pick one and keep its contents. We will assume X is a nonnegative discrete random variable, with finite mean.

The problem is often presented as a paradox, based on the following reasoning. Suppose you pick one of the two envelopes; denote its contents by Y. The other envelope must contain either $2Y$ or $Y/2$. Presumably, since you have no information to help you know which envelope contains the larger amount, the probabilities for the two possibilities for the other envelope are both $1/2$. So the expected value of the contents of the other envelope is

$$\frac{1}{2}(2Y) + \frac{1}{2}(Y/2) = \frac{5}{4}Y > Y. \tag{3.27}$$

Thus it seems that you will increase your expected reward by always picking one envelope and then switching to the other before opening it. But this is ludicrous; you could repeat the argument after switching to the other envelope and argue that it is better to switch back, and then continue switching back and forth ad infinitum. Something is obviously wrong with this reasoning. In this problem you will analyze this more carefully.

In addition to the notation above we will denote the p.m.f. of X by

$$p(x) = P(X = x).$$

We consider this to be defined for all $x > 0$, even though it is positive only for the countable number of values that X can actually take. Our basic assumption is that

$$P(Y = X \,|\, X) = \frac{1}{2}.$$

Let $\tilde{Y}$ be the contents of the envelope that you did *not* pick.

a) Find $q(y) = P(Y = y)$ and $\tilde{q}(y) = P(\tilde{Y} = y)$ in terms of $p(\cdot)$. Use this to show $E[Y] = E[\tilde{Y}]$. (I.e. switching envelopes does not effect the mean.)

b) Find $E[Y|X]$ and then use item 7) from Proposition 3.8 to find the relation between $E[Y]$ and $E[X]$.

c) A second way to come to the same conclusion as a) is to observe that

$$\tilde{Y} = 2\frac{X^2}{Y}.$$

Explain why this is true, and use it to find $E[\tilde{Y}|X]$ and then $E[\tilde{Y}]$.

d) Define

$$s(y) = P(Y = X \,|\, Y = y)$$

and work out a formula for it in terms of $p(\cdot)$.

e) Equation (3.27) seems to say that $E[\tilde{Y}|Y] = \frac{5}{4}Y$. Find a correct formula for $E[\tilde{Y}|Y]$ in terms of $s(Y)$.

f) What would have to be true about $s(\cdot)$ for (3.27) to be true? Show that there are no discrete random variables $X \geq 0$ for which (3.27) holds.

g) Suppose we change the rules so that you are allowed to look at the contents Y of the envelope you picked and then decide whether you want to keep it or switch to the other envelope $\tilde{Y}$ (but without peeking at $\tilde{Y}$). What strategy should you follow to maximize the expected value of the envelope you keep? (You can base your decision on the observed value of Y.)

h) An example which has appeared in the literature on this problem is $X = 2^N$ where N is a geometric random variable with $p = \frac{2}{3}$. Calculate $s(y)$ for this example, and observe that $s(Y) > \frac{1}{3}$ with probability 1. What does this mean in light of your answer to g)? Doesn't this contradict a)?

i) Again under the revised rules of g), what strategy should you follow to maximize the probability that you end up with the larger of the two envelopes? (Your answer here will be different than in g)!)

TwoEnvelopes

Problem 3.26
Write out the derivation of equation (2.10) in a way similar to what we did on page 61 for equation (2.5). To do this write $\mathcal{T}_C = \Psi(X_{0:\infty})$ and note that for $s_0 = i \notin C$ we have

$$\Psi(s_{0:\infty}) = 1 + \Psi(s_{1:\infty}).$$

If we write $\mathcal{T}_C^+ = \Psi^+(X_{0:\infty})$ then observe that in all cases

$$\Psi^+(s_{0:\infty}) = 1 + \Psi(s_{1:\infty}),$$

which leads to

$$E[\mathcal{T}_C^+] = 1 + E[v(X_1)],$$

where $v(s)$ is still $E_s[\mathcal{T}_C]$.

MHeqn

Problem 3.27
Consider again the pairs chain of Problem 2.13 Suppose $f(i,j)$ is a function of two variables. Using Theorem 3.10 what can we say about

$$\lim_{n\to\infty} \frac{1}{n} \sum_{a=1}^{n} f(X_{n-1}, X_n)?$$

ErgPairs

For Further Study

An expanded introduction to the Kolmogorov model is the book by Pfiffer [47]. The first chapter of Grimmett & Stirzaker [25] also provides an introduction to much of this chapter's material.

We acknowledged in Section 3.1.1 that some important aspects of the Kolmogorov model have been ignored in our discussion. Specifically when Ω is an uncountable it is generally impossible to assign a probability $P(A)$ to *every* subset $A \subseteq \Omega$. The resolution is to only define $P(A)$ for certain subsets of Ω, but not others. The mathematics of all the technicalities involved is the subject of measure theory, a graduate level topic. A good reference for that written from a probabilistic perspective is Billingsley [6]. Volume 1 of Rogers and Williams [51] also covers this material in the first two chapters.

Proofs of the Strong Law and Central Limit Theorem (Theorems 3.6 and 3.7) as well as the convergence theorems of Section 3.2.1 can be found in Billingsley [6]. A proof of The Renewal Theorem 3.5 can be found in Feller [22]; see his Theorem 3, Chapter XIII. Grimett & Stirzaker [25] §6.4 also offers a proof. There is a short proof due to S. Port [48] which relies on **T1** from Spitzer [60], page 276. In fact our Renewal Theorem is a special case of Spitzer's **P2**, page 278. It's also proven in Norris [45] §1.8 using a coupling argument.

Chapter 4

Infinite State Markov Chains

This chapter considers Markov chains for which the state space $\mathcal{S}$ is *countably infinite*, such as $\mathbb{N}$, $\mathbb{Z}$ or $\mathbb{Z}^d$. Many of the results from Chapter 2 remain true in this setting. But there are new phenomena which are only possible for an infinite state space: transience and null recurrence.

4.1 Introduction

The transition matrix $\mathbf{P}$ is now an infinite matrix: there are entries $p_{i,j}$ for all $i, j \in \mathcal{S}$. (Mathematicians would call it a *linear operator* rather than a matrix in this setting.) We can still write matrix-vector products as before, provided we remember that these are now infinite series. For instance consider

$$\mathbf{P}u(i) = \sum_{j \in \mathcal{S}} p_{i,j} u(j).$$

If $\mathcal{S} = \mathbb{N}$ we would write $\sum_{j \in \mathcal{S}}$ as $\sum_{j=1}^{\infty}$. If $\mathcal{S} = \mathbb{Z}$ it would be $\sum_{j=-\infty}^{\infty}$. Since $\sum_{j \in \mathcal{S}} p_{i,j} = 1$ the series above will converge if $u(\cdot)$ is bounded. If $u(\cdot) \geq 0$ the series might diverge, in which case $\mathbf{P}u(i) = \infty$. More generally if $\sum_{j \in \mathcal{S}} p_{i,j}|u(j)| < \infty$ then the above series will converge to a finite value. (See the Appendix for more on these issues.) An initial distribution $\mu = [\mu_i]$ now has infinitely many entries ($0 \leq \mu_i$, $\sum_{i \in \mathcal{S}} \mu_i = 1$). The calculation of $\mu\mathbf{P}$ and $\mathbf{P}^n$ all involve (convergent) infinite series. The distribution of X_n is as before given by

$$P_\mu(X_n = s) = (\mu \mathbf{P}^n)_s.$$

Theoretically this is fine, but for purposes of calculation we can't usually work out $\mathbf{P}^n$ explicitly in examples.

We took advantage of the properties of finite matricies in Chapter 2, but now that our matricies are infinite we can't presume that all the usual properties of matricies carry over to the infinite setting. We need to be careful about which results for the finite state space case do or do not carry over to infinite state spaces. For this reason we were careful to include "for finite state space" in the statements of those results in Chapter 2 which only hold for finite state spaces. The definitions of reachable, communicate, irreducible, closed, period are all as on page 15, and Lemma 2.4 still holds. Most of our attention in this chapter will be on recurrence and its alternatives in the more complicated infinite state space setting. We will give new definitions of recurrent and transient below but they will be equivalent to those we gave on page 18. However the simple characterization of Theorem 2.6 is *not* correct for infinite state spaces.

We have selected and organized the results to try to present a reasonably organized collection of ideas. Although we don't want the technical details to become overwhelming, it is inevitable that the level of difficulty is higher in this chapter than previously. It may be wise to focus on the results themselves and their use in examples first, and save careful reading of their proofs for later. Also, to keep the complexity from getting out of hand **we will assume throughout this chapter that the chain is irreducible**.

Some Simulations

To begin our discussion let's look at an example.

Example 4.1. The symmetric random walk is the Markov chain on $\mathcal{S} = \mathbb{Z}$ for which $X_{n+1} = X_n \pm 1$ each with probability 1/2. In other words

$$p_{i,j} = \frac{1}{2} \text{ if } j = i \pm 1 \text{ and 0 otherwise.}$$

The following MATLAB code will simulate and plot a 10000-step sample, starting from $X_0 = 0$.

```
Y=2*randi([0,1],[10000,1])-1;
X=cumsum(Y);
D1=sqrt(sum(X.*X,2));
plot(D1,'.')
```

We should observe that X_n does seem to always return to 0 if we wait long enough. Since this is a 2-periodic chain $p_{0,0}(n) = 0$ if n is odd. If n is even ($n = 2k$) then for $X_0 = X_n$ requires exactly half of the transitions to be to the right and half to the left. This implies that

$$p_{0,0}(2k) = \binom{2k}{k} \frac{1}{2^{2k}} = \frac{(2k)!}{(k!)^2 2^{2k}}.$$

It is not obvious from this formula, but it turns out that $\lim_n p_{0,0}(n) = 0$. (In fact on page 89 we will see that $p_{0,0}(2k) \sim \frac{1}{\sqrt{\pi k}}$.) This seems to be contrary to Corollary 2.5. Although X_n is not aperiodic, observed at just the even times $Y_k = X_{2k}$ *is* aperiodic, and irreducible if we take the even integers as the state space. So this really *is* contrary to how finite state chains behave.

Example 4.2. The symmetric random walk in 3 dimensions is a Markov chain on $\mathcal{S} = \mathbb{Z}^3$. At each stage one of the 6 possible increments $(\pm 1, 0, 0), (0, \pm 1, 0), (0, 0, \pm 1)$ is choosen (with equal probabilities of 1/6) and added to the current state X_n to get the next state X_{n+1}. This is easy to simulate starting from $X_0 = (0, 0, 0)$, but a plot similar to that of the previous example the result would take a 4-dimensional graph. Instead we can plot $|X_n|$ to look for returns to $0 = |(0, 0, 0)|$. The following code will do it.

```
J=2*randi([0,1],[10000,1])-1;
C=randi([1,3],[10000,1]);
Y=zeros([10000,3]);
for i=1:10000
    Y(i,C(i))=J(i);
end
X=cumsum(Y);
D3=sqrt(sum(X.*X,2));
plot(D3,'.')
```

We should observe that X_n does *not* always return to $(0, 0, 0)$. Instead it appears to eventually wander away and never come back. Finite irreducible chains do not do that.

4.2 Hitting Time Equations

The different types of infinite sate chains are characterized in terms of the values of $u(i) = P_i(\mathcal{T}_i^+ < \infty)$ and $v(i) = E_i[\mathcal{T}_i^+]$. In order to study these let's look again at the equations for $u(i) = P_i(\mathcal{T}_C < \infty)$ and $v(i) = E_i[\mathcal{T}_C]$ and update what we said about these previously (Section 2.2) to infinite state spaces. The new feature of the following theorem is that it explains the relation of $u(\cdot)$ and $v(\cdot)$ to other solutions of the same systems of equations.

Theorem 4.1. *Consider an irreducible Markov chain. Let $C \subsetneq \mathcal{S}$ and $B = \mathcal{S} \setminus C$. The hitting probabilities $u(i) = P_i(\mathcal{T}_C < \infty)$ solve the following:*

$$u(i) = \begin{cases} 1 & \text{for } i \in C \\ \sum_{j \in \mathcal{S}} p_{i,j} u(j) & \text{for } i \in B. \end{cases}$$

If $\phi(\cdot)$ is a nonnegative function satisfying

$$\phi(i) \geq \begin{cases} 1 & \text{for } i \in C \\ \sum_{j \in \mathcal{S}} p_{i,j}\phi(j) & \text{for } i \in B \end{cases}$$

then $u(i) \leq \phi(i)$ for all $i \in \mathcal{S}$.

The mean hitting times $v(i) = E_i[\mathcal{T}_C]$ satisfy

$$v(i) = \begin{cases} 0 & \text{for } i \in C \\ 1 + \sum_{j \in B} p_{i,j}v(j) & \text{for } i \in B. \end{cases}$$

(This holds even if some $v(i) = \infty$.) If $\psi(j), j \in \mathcal{S}$ is a nonnegative (finite-valued) function satisfying $\psi(i) \geq 1 + \sum_{j \in B} p_{i,j}\psi(j)$ for $i \in B$ then $v(i) \leq \psi(i)$.

Proof. The equations for $u(i)$ and $v(i)$ are the same as in Chapter 2, and were derived in the preceding chapter using the strong Markov property. For our purposes here we will rederive them using an iterative approach.

Let's start with $u(i)$, breaking it down further by defining

$$u(i,n) = P_i(\mathcal{T}_C \leq n).$$

We know that

$$u(i,n) = 1 \text{ for all } n \text{ if } i \in C$$

and

$$u(i,0) = 0 \text{ if } i \in B.$$

For $i \in B$ the key observation is that

$$\mathcal{T}_C(X_{0:\infty}) = 1 + \mathcal{T}_C(X_{1:\infty}).$$

Therefore using properties of conditional expectations and the Markov property we can say

$$\begin{aligned} u(i,n+1) &= P_i(\mathcal{T}_C(X_{0:\infty}) \leq n+1) \\ &= P_i(1 + \mathcal{T}_C(X_{1:\infty}) \leq n+1) \\ &= P_i(\mathcal{T}_C(X_{1:\infty}) \leq n) \\ &= E_i[P_i(\mathcal{T}_C(X_{1:\infty}) \leq n | X_1)] \\ &= E_i[P_{X_1}(\mathcal{T}_C(X_{0:\infty}) \leq n)] \\ &= E_i[u(X_1,n)] \\ &= \sum_{j \in \mathcal{S}} p_{i,j}u(j,n). \end{aligned}$$

In matrix form this looks like

$$\begin{aligned} \mathbf{u}_C(n) &= [1] \\ \mathbf{u}_B(0) &= [0] \\ \mathbf{u}_B(n+1) &= \mathbf{P}_{BB}\mathbf{u}_B(n) + \mathbf{P}_{BC}[1]. \end{aligned}$$

The definition of $u(\cdot,\cdot)$ implies that $u(i,n) \leq u(i,n+1)$. Using the last bullet on page 33 we can say

$$u(i) = P_i(\mathcal{T}_C < \infty) = \lim_n P_i(\mathcal{T}_C \leq n) = \lim_n u(i,n).$$

Next we can use the Monotone Convergence Theorem for infinite series, Theorem A.8, to let $n \to \infty$ and conclude that

$$\begin{aligned} \mathbf{u}_C &= [1] \\ \mathbf{u}_B &= \mathbf{P}_{BB}\mathbf{u}_B + \mathbf{P}_{BC}[1], \end{aligned}$$

which are the equations for $u(i)$.

Next, suppose there exists a function $\phi \geq 0$ as described:

$$\begin{aligned}\phi_C &\geq [1]\\ \phi_B &\geq \mathbf{P}_{BB}\phi_B + \mathbf{P}_{BC}[1],\end{aligned}$$

Then $\phi(\cdot) \geq u(\cdot, 0)$ and by induction $\mathbf{u}_B(n) \leq \phi_B$:

$$\mathbf{u}_B(n+1) = \mathbf{P}_{BB}\mathbf{u}_B(n) + \mathbf{P}_{BC}[1] \leq \mathbf{P}_{BB}\phi_B + \mathbf{P}_{BC}[1] \leq \phi_B.$$

Therefore

$$\mathbf{u}_B = \lim_n \mathbf{u}_B(n) \leq \phi_B,$$

and

$$\mathbf{u}_C = [1] \leq \phi_C.$$

This proves all the theorem's claims about $u(i)$.

The argument for $v(i)$ is similar. Define

$$v(i, n) = E_i[\min(n, \mathcal{T}_C)].$$

We know that

$$v(i, n) = 0 \text{ for all } n \text{ if } i \in C$$

and

$$v(i, 0) = 0 \text{ if } i \in B.$$

For $i \in B$ conditional calculation gives

$$\begin{aligned}v(i, n+1) &= E_i[\min(n+1, \mathcal{T}_C(X_{0:\infty})]\\ &= E_i[\min(n+1, 1 + \mathcal{T}_C(X_{1:\infty})]\\ &= E_i[1 + \min(n, \mathcal{T}_C(X_{1:\infty})]\\ &= 1 + E_i[\min(n, \mathcal{T}_C(X_{1:\infty})]]\\ &= 1 + E_i[E_i[\min(n, \mathcal{T}_C(X_{1:\infty}))|X_1]]\\ &= 1 + E_i[E_{X_1}[\min(n, \mathcal{T}_C(X_{0:\infty}))]]\\ &= 1 + E_i[v(X_1, n)]\\ &= 1 + \sum_{j \in B} p_{i,j} v(j, n),\end{aligned}$$

since $v(j, n) = 0$ for $j \notin B$. In matrix form

$$\begin{aligned}\mathbf{v}_C(n) &= [0]\\ \mathbf{v}_B(0) &= [0]\\ \mathbf{v}_B(n+1) &= [1] + \mathbf{P}_{BB}\mathbf{v}_B(n).\end{aligned}$$

Again $v(i, n) \leq v(i, n+1)$ and

$$v(i) = E_i[\mathcal{T}_C] = \lim_n E_i[\min(n, \mathcal{T}_C)] = \lim_n v(i, n).$$

It follows as above that

$$\begin{aligned}\mathbf{v}_C &= [0]\\ \mathbf{v}_B &= [1] + \mathbf{P}_{BB}\mathbf{v}_B.\end{aligned}$$

Suppose there exists a function $\psi \geq 0$ as described:

$$\psi_C \geq [0]$$
$$\psi_B \geq [1] + \mathbf{P}_{BB}\psi_B.$$

Then $\psi(\cdot) \geq v(\cdot, 0)$ and by induction $\mathbf{v}_B(n) \leq \psi_B$. Therefore

$$\mathbf{v}_B = \lim_n \mathbf{v}_B(n) \leq \psi_B,$$

and

$$\mathbf{v}_C = [0] \leq \psi_C.$$

□

One-Dimensional Reflecting Random Walk

We want to illustrate the solution of the equations above using a random walk. To simplify things we will make it a *reflecting* random walk on $\mathbb{Z}^+$ so that all states s are nonnegative: $s \geq 0$. When $X_n > 0$ it moves one step to the right with probability p and one step to the left with probability $q = 1 - p$. But if $X_n = 0$ then $X_{n+1} = 1$ with probability 1. (This is what makes it reflecting.) This chain is irreducible, but like the unreflected random walk on $\mathbb{Z}$ has period 2. We will consider the hitting time $\mathcal{T}_0$ of $C = \{0\}$. Following our usual notation let

$$u(i) = P_i(\mathcal{T}_0 < \infty) \text{ and } v(i) = E_i[\mathcal{T}_0].$$

We know $u(0) = 1$ and $v(0) = 0$. For $i > 0$

$$u(i) = pu(i+1) + qu(i-1) \text{ and } v(i) = 1 + pv(i+1) + qv(i-1).$$

If we write $\Delta u(i) = u(i) - u(i-1)$ we can rearrange the $u(i)$-equation as

$$\Delta u(i+1) = \frac{p}{q}\Delta u(i) \text{ for } i \geq 1.$$

The Symmetric Case: $p = q = \frac{1}{2}$.

The equation simplifies to $\Delta u(i+1) = \Delta u(i)$ so for all $i \geq 1$ we have $\Delta u(i) = \Delta u(1)$, and consequently for $i > 0$ we have

$$u(i) = u(0) + \sum_{k=1}^{i} \Delta u(i) = 1 + i\alpha$$

where $\alpha = \Delta u(1)$. This is the general solution of the u-equations, involving a single undetermined parameter α. There are many nonnegative solutions: any $\alpha \geq 0$ will produce one. According to the result above we need the smallest nonnegative solution, which is clearly for $\alpha = 0$. Thus

$$u(i) = P_i(\mathcal{T}_0 < \infty) = 1 \text{ for all } i.$$

Since

$$P_0(\mathcal{T}_0^+ < \infty) = p_{0,1}P_1(\mathcal{T}_0 < \infty) = 1$$

we see that 0 is a recurrent state. As we will see below that means that all states are recurrent.

But let's also consider the mean return time,

$$v(i) = E_i[\mathcal{T}_0].$$

We have $v(0) = 0$ and for $i > 0$

$$v(i) = 1 + \frac{1}{2}v(i+1) + \frac{1}{2}v(i-1),$$

which we can rearrange as

$$\Delta v(i+1) = -2 + \Delta v(i) \text{ for } i > 0.$$

(Notice that if $v(1), \ldots v(i)$ are all finite then $v(i+1)$ would have to be finite too. So there cannot be a solution which is finite for some $i > 0$ but infinite for others. Either $v(i) < \infty$ for all $i > 0$ or $v(i) = \infty$ for all $i > 0$.) Considering the possibility of a finite solution we find for $i > 0$ that

$$v(i) = -2(i-1) + \beta$$

where $\beta = \Delta v(1)$. Using this we have

$$\begin{aligned} v(i) &= v(0) + \sum_{k=1}^{i} \Delta v(k) \\ &= 0 + \sum_{k=1}^{i} \Delta v(k) \\ &= -2 \sum_{k=1}^{i} (k-1) + (i-1)\Delta v(1) \\ &= -i(i-1) + (i-1)\beta \\ &\to -\infty \text{ as } i \to \infty, \end{aligned}$$

regardless of the value of β. So *there is no finite nonnegative solution*: $E_i[\mathcal{T}_0] = \infty$ for all $i > 0$, even though $P_i(\mathcal{T}_0 < \infty) = 1$. Since $E_0[\mathcal{T}_0^+] = p_{0,1} E_1[\mathcal{T}_0] = \infty$.

The Asymmetric Case $p > q$.

Now we find that for $i \geq 1$

$$\Delta u(i+1) = (q/p)^i \Delta u(1).$$

Using this in $u(i) = u(0) + \sum_{k=1}^{i} \Delta u(k)$ leads to

$$u(i) = 1 + \frac{(q/p)^i - 1}{q/p - 1} \alpha, \tag{4.1}$$

where $\alpha = \Delta u(1)$. Because $p > q$ we have $q/p < 1$ and $\frac{(q/p)^i - 1}{q/p - 1} > 0$. If $\alpha \geq 0$ then the resulting $u(i)$ are all nonnegative. But in fact some negative values of α lead to nonnegatvie soutions as well: any $\alpha \geq \frac{q}{p} - 1$. Using the smallest of these gives the correct solution:

$$u(i) = 1 + \frac{(q/p)^i - 1}{q/p - 1} \left(\frac{q}{p} - 1 \right) = (q/p)^i \text{ for } i > 0.$$

Observe that $P_i(\mathcal{T}_0 < \infty) = (q/p)^i < 1$. Thus there is a positive probability of never returning to 0. In fact for large i the probability of ever reaching 0 is very small. This is a consequence of the chain's preference to move to the right.

Since there is positive probability that $\mathcal{T}_0 = \infty$ there is no point in calculating $E_i[\mathcal{T}_0]$; we know it is infinite.

The Asymmetric Case $p < q$.

If $p < q$, so that the chain wants to move to the left, the situation is different. The calculations leading to (4.1) are still applicable, but now $\frac{(q/p)^i - 1}{q/p - 1} \to +\infty$ as $i \to \infty$. We get a nonnegative solution only for $\alpha \geq 0$. Taking the smallest possibility, $\alpha = 0$, we find that

$$u(i) = 1 \text{ for all } i \geq 1.$$

Hence return to 0 is certain. Solving the equations for $v(i)$ leads to

$$v(i) = \frac{i}{q-p} + \beta \frac{(q/p)^i - 1}{q/p - 1}$$

for an arbitrary constant β. The smallest nonnegative solution occurs for $\beta = 0$. So we find that

$$E_i[\mathcal{T}_0] = \frac{i}{q-p} < \infty \text{ for } i > 0.$$

This formula makes intuitive sense; the chain must make i steps down to reach 0 and each step takes an average of $\frac{1}{q-p}$ transitions. In particular the mean return times are finite. Starting from $X_0 = 0$ we have

$$E_0[\mathcal{T}_0^+] = 1 + p_{0,1}E_1[\mathcal{T}_0] = 1 + 1 \cdot \frac{1}{q-p} = \frac{2q}{q-p}.$$

(See Problem 3.26.)

Observe that even though $B = \{1, 2, 3, \ldots\}$ contained no closed communication classes the equations to be solved did *not* have unique solutions, unlike the finite case of Theorem 2.3.

4.3 Transience and Recurrence

An irreducible chain on a *finite* state space has the property that all states are visited infinitely many times with probability 1; see Theorem 2.6 and Lemma 3.9. But the examples above show that this need *not* be so for infinite state spaces. It is possible for $P_i(\mathcal{T}_i^+ < \infty)$ to be < 1 or $= 1$, and if this probability is $= 1$ it is possible for $E_i[\mathcal{T}_i^+]$ to be either $< \infty$ or $= \infty$. This leads to a three-way classification of the recurrence type of a state.

Definition. *Let X_n be a Markov chain. A state $a \in \mathcal{S}$ is called* transient *if*

$$P_a(\mathcal{T}_a^+ < \infty) < 1.$$

and recurrent *if*

$$P_a(\mathcal{T}_a^+ < \infty) = 1.$$

A recurrent state a is called positive *recurrent if*

$$E_a[\mathcal{T}_a^+] < \infty$$

and null *recurrent if*

$$E_a[\mathcal{T}_a^+] = \infty.$$

To be a transient state means that $P_a(\mathcal{T}_a^+ = \infty) > 0$, i.e. there is positive probability of never returning to a.

Our next goal is to understand the implications of these different recurrence types. We will see that for an irreducible chain the recurrence type is common to all states. We will also look at some necessary and sufficient conditions for a chain to be one of the particular types. The two theorems of the present section gather a number of equivalent characterizations. The first concerns recurrence vs. transience. Bear in mind that if a state is not recurrent then it is transient, so the failure of any part of this theorem is equivalent to transience.

Theorem 4.2. *Suppose X_n is an irreducible Markov chain on state space $\mathcal{S}$ and $a \in \mathcal{S}$. The following are equivalent.*

1. *a is recurrent.*
2. *$P_b(\mathcal{T}_a^+ < \infty) = 1$ for all $b \in \mathcal{S}$.*
3. *All $b \in \mathcal{S}$ are recurrent.*
4. *$\sum_0^\infty p_{a,a}(n) = \infty$.*
5. *$P_a(X_n = a \text{ for infinitely many } n) = 1$.*

Because of part 3, instead of refering to an individual state as recurrent we will call the entire chain recurrent if its states are.

Corollary 4.3. *Suppose X_n is a transient irreducible Markov chain on $\mathcal{S}$. Then*

$$\sum_0^\infty p_{i,j}(n) < \infty \text{ for all } i, j \in \mathcal{S}.$$

Proof of the Corollary. Since $j \rightsquigarrow i$ there exists k so that $p_{j,i}(k) > 0$. We know that

$$p_{i,i}(n+k) \geq p_{i,j}(n)p_{j,i}(k).$$

By hypothesis i is transient and so by part 4 of the theorem $\sum_{n=0}^\infty p_{i,i}(n) < \infty$. Therefore

$$p_{j,i}(k)\left(\sum_{n=0}^\infty p_{i,j}(n)\right) \leq \sum_{n=0}^\infty p_{i,i}(k+n) \leq \sum_{m=0}^\infty p_{i,i}(m) < \infty.$$

Since $p_{j,i}(k) > 0$ it follows that $\sum_{n=0}^\infty p_{i,j}(n) < \infty$. □

We turn to the proof of the theorem itself. We will use the following notation for the distribution of $\mathcal{T}_a^+$ assuming $X_0 = i$:

$$f_{i,a}(n) = P_i(X_k \neq a \text{ for all } 0 < k < n \text{ and } X_n = a) = P_i(\mathcal{T}_a^+ = n).$$

In particular $f_{i,a}(0) = 0$ for all $i \in \mathcal{S}$. This is consistent with our definition of $\mathcal{T}_a^+$ because $\mathcal{T}_a^+ \neq 0$. For a to be recurrent [transient] means that $\sum_1^\infty f_{a,a}(n) = 1$ [< 1]. The proof of the Theorem 4.2 depends on the following equations. (See Problem 4.5 for verification.)

$$\begin{aligned} f_{i,a}(0) &= 0 \\ p_{i,a}(0) &= \begin{cases} 1 & \text{if } i = a \\ 0 & \text{if } i \neq a \end{cases} \\ p_{i,a}(n) &= \sum_{k=1}^n f_{i,a}(k)p_{a,a}(n-k) \text{ for } n \geq 1. \end{aligned} \tag{4.2}$$

Proof of Theorem 4.2. It is elementary to see that 1 follows from any of 2, 3, or 5. That $1 \Rightarrow 5$ was proven as Lemma 3.9. So our proof needs to show that $1 \Rightarrow 2$, that $1 \Leftrightarrow 4$ and that $4 \Rightarrow 3$.

$1 \Rightarrow 2$: Suppose 1. By hypothesis 2 is true for $b = a$, so suppose $b \neq a$. Since the chain is irreducible there exists a sequence $s_{0:m}$ of states with $s_0 = a$ and $s_m = b$ for which $P_a(X_{0:m} = s_{0:m}) > 0$. By chosing the shortest such sequence we can assume $s_1, \ldots, s_{m-1}$ are all distinct from a. But then by the Markov property

$$P_a(\mathcal{T}_a^+ = \infty | X_{0:m} = s_{0:m}) = P_b(\mathcal{T}_a^+ = \infty).$$

So we have

$$\begin{aligned} P_a(\mathcal{T}_a^+ = \infty) &\geq P_a(X_{0:m} = s_{0:m} \text{ and } \mathcal{T}_a^+ = \infty) \\ &= P_a(X_{0:m} = s_{0:m})P_b(\mathcal{T}_a^+ = \infty). \end{aligned}$$

Since the left side of this inequality is 0 and $P_a(X_{0:m} = s_{0:m}) > 0$ it must be that $P_b(\mathcal{T}_a^+ = \infty) = 0$.

$1 \Leftrightarrow 4$: Let $F_{i,i} = \sum_{n=1}^\infty f_{i,i}(n)$ and $M_{i,i} = \sum_{n=0}^\infty p_{i,i}(n)$. Summing both sides of the third formula in (4.2) we find

$$\begin{aligned} \sum_{n=0}^\infty p_{i,i}(n) &= 1 + \sum_{n=0}^\infty \sum_{k=1}^n f_{i,i}(k)p_{i,i}(n-k) \\ &= 1 + \sum_{m=0}^\infty \sum_{k=1}^\infty f_{i,i}(k)p_{i,i}(m) \\ M_{i,i} &= 1 + F_{i,i}M_{i,i}. \end{aligned}$$

If $F_{i,i} = 1$ this is only possible if $M_{i,i} = \infty$. Suppose that $F_{i,i} < 1$. Summing the recurrence relation up to $n = m$ and then including extra terms on the right we find that

$$\begin{aligned}\sum_{n=0}^{m} p_{a,a}(n) &= 1 + \sum_{n=0}^{m}\sum_{k=1}^{n} f_{a,a}(k)p_{a,a}(n-k) \\ &\leq 1 + \sum_{\ell=0}^{m}\sum_{k=1}^{m} f_{a,a}(k)p_{a,a}(\ell) \\ &\leq 1 + F_{a,a}\sum_{n=0}^{m} p_{a,a}(n)\end{aligned}$$

and therefore

$$\sum_{n=0}^{m} p_{a,a}(n) \leq \frac{1}{1-F_{a,a}}.$$

Letting $m \to \infty$ we find that

$$M_{a,a} \leq \frac{1}{1-F_{a,a}} < \infty.$$

Thus $F_{a,a} < 1$ iff $M_{a,a} < \infty$. Since the definition of recurrence for a is that $F_{a,a} = 1$ this proves $1 \Leftrightarrow 4$. *Note that this equivalence $1 \Leftrightarrow 4$ does not depend on the hypothesis of irreducibility.*

$4 \Rightarrow 3$: Suppose 4 holds and consider any $i \in \mathcal{S}$. Since the chain is irreducible there exist k and ℓ so that $p_{a,i}(k) > 0$ and $p_{i,a}(\ell) > 0$. It follows that

$$p_{i,i}(\ell + n + k) \geq p_{i,a}(\ell)p_{a,a}(n)p_{a,i}(k).$$

Therefore

$$\begin{aligned}\sum_{m=0}^{\infty} p_{i,i}(m) &\geq \sum_{n=0}^{\infty} p_{i,i}(\ell + n + k) \\ &\geq p_{i,a}(\ell)\left(\sum_{n=0}^{\infty} p_{a,a}(n)\right)p_{a,i}(k) \\ &= \infty.\end{aligned}$$

This establishes 4 with i in place of a and so by $4 \Rightarrow 1$ it follows that i is recurrent. Thus $4 \Rightarrow 3$ holds. □

The next theorem concerns positive recurrence. A recurrent chain that fails any of the equivalent conditions must be null recurrent.

Theorem 4.4. *Suppose X_n is a recurrent irreducible Markov chain on state space $\mathcal{S}$. Suppose $a \in \mathcal{S}$ and $C \subsetneq \mathcal{S}$ is a nonempty, finite set. The following are equivalent.*

1. *a is positive recurrent.*
2. *$E_b[\mathcal{T}_a^+] < \infty$ for all $b \in \mathcal{S}$.*
3. *All $b \in \mathcal{S}$ are positive recurrent.*
4. *$E_b[\mathcal{T}_C^+] < \infty$ for all $b \in \mathcal{S}$.*

Positive recurrence of one state implies positive recurrence of all states. So just as for recurrence, we will say that the chain itself is positive recurrent rather than a specific state is. Also observe that since C is arbitrary (proper, finite, nonempty) the theorem implies that if 4 holds for one such C then it holds for every such C.

In preparation for the proof we establish the following result. Although the number of steps to reach a specific state a may have infinite mean, the number of visits to a *particular* state b before reaching a always has finite mean.

Lemma 4.5. *Suppose X_n is an irreducible Markov chain. Given $a, b \in \mathcal{S}$ let $N_{b\neg a}$ be the number of visits to b prior to the first return to a:*

$$N_{b\neg a} = \sum_{n=0}^{\mathcal{T}_a^+ - 1} 1_{\{b\}}(X_n).$$

Then $E_i[N_{b\neg a}] < \infty$ for all $i \in \mathcal{S}$. Moreover $w(i) = E_i[N_{b\neg a}]$ satisfies

$$w(b) = 1 + \sum_{j \neq a} p_{b,j} w(j)$$
$$w(i) = \sum_{j \neq a} p_{i,j} w(j) \text{ for } i \neq b.$$

We will solve these equations in an example after the proof.

Notice that $N_{b\neg a}$ counts X_0 but not $X_{\mathcal{T}_a^+}$. Let's consider what would be different if instead we used

$$N_{b\neg a}^+ = \sum_{n=1}^{\mathcal{T}_a^+} 1_{\{b\}}(X_n).$$

If $b = a$ then $N_{b\neg a} = N_{b\neg a}^+ = 1_{\{b\}}(X_0)$. If $b \neq a$ and $X_0 \neq b$ then again $N_{b\neg a} = N_{b\neg a}^+$. Only when $X_0 = b \neq a$ are they different. In that case $N_{b\neg a} = 1 + N_{b\neg a}^+$.

Proof. First notice that if $b = a$ then $N_{b\neg a} = 1_{\{b\}}(X_0)$ and $w(i) = 1_{\{b\}}(i)$, which does satisfy the equations. So we can assume $b \neq a$ for the rest of the proof.

In order for $N_{b\neg a} > 0$ the chain has to reach b before returning to a: $\mathcal{T}_b < \mathcal{T}_a^+$. Using this with the strong Markov property

$$E_i[N_{b\neg a}] = E_i[N_{b\neg a}; \mathcal{T}_b < \mathcal{T}_a^+] = P_i(\mathcal{T}_b < \mathcal{T}_a^+) E_b[N_{b\neg a}] \leq E_b[N_{b\neg a}].$$

So it is enough to prove that $E_b[N_{b\neg a}] < \infty$. This we can do with a simple calculation. Let $p = P_b(\mathcal{T}_b^+ < \mathcal{T}_a^+)$. From irreducibility it follows that $p < 1$. Now observe that

$$\begin{aligned} P_b(N_{b\neg a} \geq 1) &= 1 \\ P_b(N_{b\neg a} \geq 2) &= p \\ P_b(N_{b\neg a} \geq 3) &= p^2 \\ &\vdots \\ P_b(N_{b\neg a} \geq k) &= p^{k-1} \end{aligned}$$

Subtracting successive lines above we find that for $k \geq 1$

$$P_b(N_{b\neg a} = k) = P_b(N_{b\neg a} \geq k) - P_b(N_{b\neg a} \geq k+1) = p^{k-1} - p^k = (1-p)p^{k-1}.$$

In other words for $X_0 = b$ the random variable $N_{b\neg a}$ has a geometric distribution with parameter $1 - p$. Therefore

$$E_b[N_{b\neg a}] = \sum_{k=1}^{\infty} k(1-p)p^{k-1} = \frac{1-p}{(1-p)^2} = \frac{1}{1-p} < \infty.$$

As for the equations for $w(i)$, observe that

$$N_{b\neg a}(X_{0:\infty}) = 1_b(X_0) + 1_{\{a\}^c}(X_1) N_{b\neg a}(X_{1:\infty}).$$

Using this with the Markov property

$$\begin{aligned} E[N_{b\neg a}(X_{0:\infty})|X_1] &= 1_b(X_0) + 1_{\{a\}^c}(X_1) E[N_{b\neg a}(X_{1:\infty})|X_1] \\ &= 1_b(X_0) + 1_{\{a\}^c}(X_1) E_{X_1}[N_{b\neg a}(X_{0:\infty})] \\ &= 1_b(X_0) + 1_{\{a\}^c}(X_1) w(X_1). \end{aligned}$$

Therefore

$$\begin{aligned} w(i) &= E_i[N_{b\neg a}(X_{0:\infty})] \\ &= E_i[E[N_{b\neg a}(X_{0:\infty})|X_1]] \\ &= 1_b(i) + E_i[1_{\{a\}^c}(X_1)w(X_1)] \\ &= 1_b(i) + \sum_{j\neq a} p_{i,j}w(j), \end{aligned}$$

which are the desired equations. □

The equations for $w(\cdot)$ can be organized as follows. Let

$$C = \mathcal{S} \setminus \{a\}$$

and $\mathbf{d} = [1_b(\cdot)]$ the vector of all 0s except for a 1 in the b^{th} position. The equations for $i \neq a$ (i.e. $i \in C$) can be expressed using our submatrix notation as

$$\mathbf{w}_C = \mathbf{d}_C + \mathbf{P}_{CC}\mathbf{w}_C.$$

(Theorem 2.3 guarantees a unique solution.) Finally

$$w(a) = 1_b(a) + \mathbf{P}_{aC}\mathbf{w}_C.$$

Example 4.3. Consider Example 2.1 again. Let's take $a = 2$ and $b = 4$. The equations for $i \neq 2$ are

$$\begin{bmatrix} w(1) \\ w(3) \\ w(4) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} + \begin{bmatrix} 0.2 & 0 & 0.4 \\ 0.3 & 0 & 0.7 \\ 0 & 0.5 & 0 \end{bmatrix} \begin{bmatrix} w(1) \\ w(3) \\ w(4) \end{bmatrix}$$

Solving these we find $[w(1), w(3), w(4)]^T = (0.869565, 1.47826, 1.73913)$. Then

$$w(2) = 0 \ + \ 0 * w(1) + .6 * w(3) + .4 * w(4) = 1.58261.$$

In particular,

$$E_2[N_{4\neg 2}] = w(4) = 1.73913.$$

The next lemma is $4 \Rightarrow 3$ of the theorem. We are presenting it separately because it is the most difficult part of the theorem, and once we establish it the rest of the proof of the theorem will be relatively simple.

Lemma 4.6. *Suppose X_n is an irreducible recurrent Markov chain and $B \subseteq \mathcal{S}$ is a nonempty finite proper subset with the property that $E_i[\mathcal{T}_B^+] < \infty$ for all $i \in \mathcal{S}$. Every $a \in \mathcal{S}$ is positive recurrent.*

Recall that $\mathcal{T}_B^+$ is the first time *after* $n = 0$ that $X_n \in B$. For $X_0 \notin B$ we know $\mathcal{T}_B^+ = \mathcal{T}_B$. But for $X_0 \in B$ we have $\mathcal{T}_B = 0 < \mathcal{T}_B^+$. In particular for all $i \in \mathcal{S}$

$$E_i[\mathcal{T}_B^+] = 1 + \sum_{j\in\mathcal{S}} p_{i,j}E_j[\mathcal{T}_B]. \tag{4.3}$$

(See Problem 3.26.)

Proof. Consider any $a \in \mathcal{S}$. Our task is to show that $E_a[\mathcal{T}_a^+] < \infty$. Let $v(i) = E_i[\mathcal{T}_B]$. By hypothesis this is finite. We have $v(b) = 0$ for $b \in B$ and for $i \notin B$

$$v(i) = 1 + \sum_j p_{i,j}v(j) \geq 1 + \sum_{j\neq a} p_{i,j}v(j).$$

Let

$$w(i) = \sum_{b\in B} E_i[N_{b\neg a}].$$

This is the mean number of times $X_n \in B$ prior to its first return to a. By summing the equations from Lemma 4.5 over $b \in B$ (a *finite* sum) it follows that $w(i) < \infty$,

$$w(b) = 1 + \sum_{j \neq a} p_{b,j} w(j) \text{ for } b \in B$$
$$w(i) = \sum_{j \neq a} p_{i,j} w(j) \text{ for } i \notin B.$$

Consider

$$\psi(i) = v(i) + Kw(i).$$

We will see that for a careful choice of constant $K \geq 0$ this will work as the ψ as in the second part of Theorem 4.1 with $C = \{a\}$. Clearly $\psi \geq 0$. Let's check the inequality we need for ψ. Consider $i \in B$.

$$\begin{aligned}
\psi(i) &= v(i) + Kw(i) \\
&= 0 + K(1 + \sum_{j \neq a} p_{i,j} w(j)) \\
&= K + \sum_{j \neq a} p_{i,j} \psi(j) - \sum_{j \neq a} p_{i,j} v(j) \\
&\geq K + 1 - E_i[\mathcal{T}_B^+] + \sum_{j \neq a} p_{a,j} \psi(j).
\end{aligned}$$

If we choose $K = \max_{i \in B} E_i[\mathcal{T}_B^+]$ then we have the desired inequality:

$$\psi(i) \geq 1 + \sum_{j \neq a} p_{i,j} \psi(j),$$

It now follows from Lemma 4.5 that $E_i[\mathcal{T}_a] \leq \psi(i)$ for all $i \neq a$.

Finally, from (4.3) we have

$$\begin{aligned}
E_a[\mathcal{T}_a^+] &= 1 + \sum_j p_{a,j} E_j[\mathcal{T}_a] \\
&\leq 1 + \sum_{j \neq a} p_{a,j} \psi(j) \\
&= 1 + \sum_{j \neq a} p_{a,j} v(j) + K \sum_{j \neq a} p_{a,j} w(j) \\
&\leq E_a[\mathcal{T}_B^+] + K \sum_{b \in B} E_b[N_{b\neg a}] < \infty,
\end{aligned}$$

using the bound on w from the proof of Lemma 4.5. □

We can now prove the theorem.

Proof of Theorem 4.4. Assume 1, namely that $E_a[\mathcal{T}_a^+] < \infty$. Let $v(i) = E_i[\mathcal{T}_a]$. Then $E_a[\mathcal{T}_a^+] = 1 + \sum_i p_{a,i} v(i)$, which is finite by hypothesis. Therefore $v(i) < \infty$ whenever $p_{a,i} > 0$. For any $i \neq a$ with $v(i) < \infty$ since $v(i) = \sum_j p_{i,j} v(j)$ it follows that $v(j) < \infty$ whenever $p_{i,j} > 0$, and consequently for all j with $a \rightsquigarrow j$. Since the chain is irreducible we must have $v(i) = E_i[\mathcal{T}_a] < \infty$ for all $i \in \mathcal{S}$. Thus $1 \Rightarrow 2$.

Now observe that 2 means that Lemma 4.6 applies with $C = \{a\}$. Consequently $2 \Rightarrow 3$.

Next suppose 3 and consider any finite, nonempty subset $C \subseteq \mathcal{S}$. Consider any $c \in C$. By hypothesis c is positive recurrent, so because $1 \Rightarrow 2$ we know that $E_b[\mathcal{T}_c^+] < \infty$ for every $b \in \mathcal{S}$. Since $\mathcal{T}_C^+ \leq \mathcal{T}_c^+$ it follows that for every $b \in \mathcal{S}$

$$E_b[\mathcal{T}_C^+] \leq E_b[\mathcal{T}_c^+] < \infty.$$

This shows that $3 \Rightarrow 4$.

Lemma 4.6 says that $4 \Rightarrow 3$. Clearly $3 \Rightarrow 1$. This completes the proof. □

4.3.1 Generating Functions

The method of generating functions is an elegant approach to some of the issues of this chapter. This section introduces their use. Given a sequence $a(0), a(1), \ldots$ its generating function is

$$\hat{a}(s) = \sum_0^\infty a(n)s^n,$$

in other words it's just the power series using the sequence as coefficients. The domain of this function is the interval of convergence of the infinite series. It always includes $s = 0$. If the $a(n)$ are bounded it includes all $|s| < 1$ and is differentiable for those s. If the $a(n) \geq 0$, as they will be for us, $\hat{a}(s)$ is an increasing function, and

$$\sum_0^\infty a(n) = \hat{a}(1-) \quad \left(= \lim_{s \to 1^-} \hat{a}(s) \right),$$

even in the case $\sum a(n) = \infty$.

Generating functions are particularly useful with convolutions. If $a(n)$ and $b(n)$ are two sequences, their *convolution* is the sequence $c(n)$ given by

$$c(n) = \sum_{i=0}^n a(i)b(n-i).$$

This is often written

$$c = a * b.$$

The generating function of the convolution is

$$\begin{aligned}
\hat{c}(s) &= \sum_{n=0}^\infty s^n \sum_{i=0}^n a(i)b(n-i) \\
&= \sum_{n=0}^\infty \sum_{i=0}^n s^i a(i)\, s^{n-i} b(n-i) \\
&= \sum_{i=0}^\infty \sum_{n=i}^\infty s^i a(i)\, s^{n-i} b(n-i) \text{ after interchanging order of summation,} \\
&= \sum_{i=0}^\infty \sum_{j=0}^\infty s^i a(i)\, s^j b(j) \text{ after changing to } j = n-i, \\
&= \left(\sum_{i=0}^\infty s^i a(i) \right) \left(\sum_{j=0}^\infty s^j b(j) \right) \\
&= \hat{a}(s)\hat{b}(s),
\end{aligned}$$

the product of the generating functions, provided both are defined. This comes up many places in probability.

Suppose X is a nonnegative integer valued random variable and $p_X(n) = P(X = n)$ is its distribution. The *moment generating function of* X is generating function of the sequence $p_X(n)$:

$$\hat{p}_X(s) = \sum_0^\infty s^n p_X(n) = E[s^X].$$

Observe that

$$E[X] = \sum_0^\infty n p_X(n) = \left[\sum_1^\infty n s^{n-1} p_X(n) \right]_{s=1} = \hat{p}'_X(1-).$$

This can be a convenient way to calculate moments. Suppose we have two nonnegative integer-valued random variables, X and Y, which are independent. Their sum $Z = X + Y$ has moment generating function

$$\hat{p}_Z(s) = E[s^{X+Y}] = E[s^X s^Y] = E[s^X]E[s^Y] = \hat{p}_X(s)\hat{p}_Y(s).$$

This is a manifestation of the generating function of a convolution since

$$p_Z(n) = \sum_{i=0}^{n} p_X(i) p_Y(n-i) \quad \text{i.e. } p_Z = p_X * p_Y.$$

Example 4.4. If X is λ-Poisson,

$$\hat{p}_X(s) = E[s^X] = \sum_0^\infty s^n \frac{\lambda^n}{n!} e^{-\lambda} = e^{\lambda(s-1)},$$

converging for all s. The mean of X is given by

$$p'_X(1) = \lambda e^{\lambda(1-1)} = \lambda.$$

Suppose X and Y are independent, both Poisson but with parameters λ and μ, then $Z = X + Y$ has generating function

$$\hat{p}_Z(s) = \hat{p}_X(s)\hat{p}_Y(s) = e^{\lambda(s-1)} e^{\mu(s-1)} = e^{(\lambda+\mu)(s-1)}$$

which implies that Z is itself Poisson but with parameter $(\lambda + \mu)$.

Observe that the formulas (4.2) can be expressed as convolutions:

$$\begin{aligned} p_{a,a} &= (1, 0, 0, \ldots) + f_{a,a} * p_{a,a} \\ p_{i,a} &= f_{i,a} * p_{a,a} \text{ for } i \neq a. \end{aligned}$$

So the generating functions are related by

$$\begin{aligned} \hat{p}_{a,a}(s) &= 1 + \hat{f}_{a,a}(s)\hat{p}_{a,a}(s) \\ \hat{p}_{i,a}(s) &= \hat{f}_{i,a}(s)\hat{p}_{a,a}(s) \text{ for } i \neq a. \end{aligned}$$

This means that if we can determine $\hat{p}_{a,a}(s)$ then we can easily obtain $\hat{f}_{a,a}(s)$ by simple algebraic manipulation:

$$1 - \hat{f}_{a,a}(s) = \frac{1}{\hat{p}_{a,a}(s)}. \tag{4.4}$$

Recurrence means that

$$1 = \sum_0^\infty f_{a,a}(n) = \hat{f}_{a,a}(1-).$$

But from (4.4) we see that this is true if and only iff $\hat{p}_{a,a}(1-) = \infty$. Since $\hat{p}_{a,a}(1-) = \sum_0^\infty p_{a,a}(n)$ we have rediscovered part 4 of Theorem 4.2. Moreover postive recurrence is equivalent to

$$\hat{f}'_{a,a}(1-) = \sum_1^\infty n f_{a,a}(n) < \infty.$$

This is something we can calculate if we know $\hat{f}_{a,a}(s)$ explicitly.

There is more. If we assemble the $\hat{p}_{i,j}(s)$ into a matrix $\hat{\mathbf{P}}(s) = [\hat{p}_{i,j}(s)]$ we can write

$$\begin{aligned} \hat{\mathbf{P}}(s) &= \sum_{n=0}^\infty \mathbf{P}^n s^n \\ &= \mathbf{I} + \sum_{n=1}^\infty \mathbf{P}^n s^n \\ &= \mathbf{I} + s\mathbf{P}\left(\sum_{n=0}^\infty \mathbf{P}^n s^n\right) \\ &= \mathbf{I} + s\mathbf{P}\hat{\mathbf{P}}(s), \end{aligned}$$

and therefore

$$(\mathbf{I} - s\mathbf{P})\hat{\mathbf{P}}(s) = \mathbf{I}.$$

In the finite state space case (where we understand matrix inverses) we can take this one more step to obtain the formula

$$\hat{\mathbf{P}}(s) = (\mathbf{I} - s\mathbf{P})^{-1}. \tag{4.5}$$

This provides a symbolic/algebraic approach to determining recurrence, transience, mean return times for finite state chains.

Example 4.5. Let's apply equation (4.5) to Example 2.1. Calculating $(\mathbf{I} - s\mathbf{P})^{-1}$ we obtain

$$\hat{\mathbf{P}}(s) = \begin{bmatrix} -\frac{5}{s-5} & -\frac{10s\left(s^2-s-2\right)}{3s^4-8s^3-35s^2-10s+50} & \frac{2s^2(5s+11)}{3s^4-8s^3-35s^2-10s+50} & \frac{4s\left(3s^2+2s+5\right)}{3s^4-8s^3-35s^2-10s+50} \\ 0 & \frac{5\left(s^2-2\right)}{3s^3+7s^2-10} & -\frac{2s(s+3)}{3s^3+7s^2-10} & -\frac{2s(3s+2)}{3s^3+7s^2-10} \\ 0 & \frac{5s^2}{-3s^3-7s^2+10} & \frac{2\left(s^2-5\right)}{3s^3+7s^2-10} & \frac{10s}{-3s^3-7s^2+10} \\ 0 & \frac{5s}{-3s^3-7s^2+10} & -\frac{s(3s+5)}{3s^3+7s^2-10} & \frac{10}{-3s^3-7s^2+10} \end{bmatrix}.$$

(That's a *not* a calculation we would want to do by hand; it's a job for a symbolic software package.) Extracting the diagonals $\hat{p}_{a,a}(s)$ and using them in formula (4.4) above we find

$$\begin{aligned} \hat{f}_{1,1}(s) &= \frac{s}{5} \\ \hat{f}_{2,2}(s) &= -\frac{s^2(3s+2)}{5\left(s^2-2\right)} \\ \hat{f}_{3,3}(s) &= -\frac{s^2(3s+5)}{2\left(s^2-5\right)} \\ \hat{f}_{4,4}(s) &= \frac{1}{10}s^2(3s+7) \end{aligned}$$

We see that $\hat{f}_{1,1}(1) < 1$ so state 1 is transient, but $\hat{f}_{2,2}(1) = 1$ so state 2 is recurrent. Likewise states 3 and 4 are recurrent either by similar calculations or because they communicate with 2. We can calculate $E_2[\mathcal{T}_2^+]$ from

$$\begin{aligned} E_2[\mathcal{T}_2^+] &= \hat{f}'_{2,2}(1-) \\ &= \left.\frac{s\left(-3s^3+18s+8\right)}{5\left(s^2-2\right)^2}\right|_{s=1} \\ &= 23/5. \end{aligned}$$

Since we found a finite value the closed class $\{2,3,4\}$ is positive recurrent. Of course we already knew that since the state space is finite, but the point here is how we reached that conclusion from generating function calculations.

Example 4.6. We can apply the generating function approach to the one-dimensional random walk on $\mathbb{Z}$: $p_{i,i+1} = p$, $p_{i,i=1} = q$ where $p+q = 1$. It is possible to write down $p_{0,0}(n)$ and $\hat{p}_{0,0}(s)$ explicitly, and therefore $\hat{f}_{0,0}(s)$ as well. First, $p_{0,0}(n) = 0$ if n is odd. I.e. this chain has period 2. For $n = 2k$ we have

$$p_{0,0}(2k) = \binom{2k}{k} p^k q^k.$$

That's because there must be exactly k up-transitions and k down-transitions. So the generating function is

$$\hat{p}_{0,0}(s) = \sum_{k=0}^{\infty} \binom{2k}{k} p^k q^k s^{2k} = \frac{1}{\sqrt{1-4pqs^2}}.$$

(It takes some effort to work out this series. One way is to start from the Taylor or generalized binomial series

$$\frac{1}{\sqrt{1-x}} = \sum_{k=0}^{\infty} \frac{1}{2^{2k}} \binom{2k}{k} x^k \text{ for } |x| < 1$$

and then substitute $x = 4pqs^2$.)

Now use formula (4.4) to obtain

$$\hat{f}_{0,0}(s) = 1 - \sqrt{1 - 4pqs^2}.$$

It follows (after simplification) that

$$P_0(\mathcal{T}_0 < \infty) = \hat{f}_{0,0}(1) = 1 - |p - q|.$$

So for $p \neq q$ the random walk is transient because $\hat{f}_{0,0}(1) < 1$. For $p = q = 1/2$ it is recurrent, and using $\hat{f}_{0,0}(s) = 1 - \sqrt{1 - s^2}$ we find that

$$E_0[\mathcal{T}_0^+] = \hat{f}'_{0,0}(1-) = \infty$$

so the symmetric random walk in one dimension is null recurrent.

In general a recurrent state is null-recurrent if and only if $p_{i,i}(n) \to 0$. That will follow from Theorems 4.13 and 4.14 below. We can almost prove it using generating functions. Rearrange equation (4.4) as

$$\frac{1 - \hat{f}_{a,a}(s)}{1 - s} = \frac{1}{(1-s)\hat{p}_{a,a}(s)}.$$

By L'Hopital's Rule

$$\lim_{s \to 1^-} \frac{1 - \hat{f}_{a,a}(s)}{1 - s} = \lim_{s \to 1^-} \hat{f}'_{a,a}(s) = E_a[\mathcal{T}_a^+].$$

If $\lim_{n\to\infty} p_{a,a}(n) = L$ exists then it is not hard to show that $\lim_{s\to 1^-}(1-s)\hat{p}_{a,a}(s) = L$. So if $L > 0$ then $E_a[\mathcal{T}_a^+] < \infty$ and a is positive recurrent, but if $L = 0$ then $E_a[\mathcal{T}_a^+] = \infty$ and a is null recurrent. This is not a proof however because it presumes that $\lim_{n\to\infty} p_{a,a}(n)$ exists. Section 4.4 we will see that at least in the aperiodic case $\lim_{n\to\infty} p_{a,a}(n) = L$ does always exist, but when the period is > 1 it need not.

4.3.2 Sufficient Conditions for Transience/Recurrence

Suppose we have an irreducible Markov chain X_n with transition matrix $\mathbf{P}$. How can we tell if it is transient, null or positive recurrent? If we can somehow work out the generating $\hat{p}_{a,a}(s)$ then the above tells us how to answer our question. But it is rare that an explicit formula for $\hat{p}_{a,a}(s)$ is possible. Theorem 4.1 identifies $P_i(\mathcal{T}_a^+)$ and $E_i[\mathcal{T}_a^+]$ in terms of systems of equations involving $\mathbf{A} = \mathbf{I} - \mathbf{P}$, but again only in special cases can we solve those explicitly.

There are a variety of results which imply transience, null or positive recurrence. We want to present a trio of results which allow us to conclude transience, recurrence, or positive recurrence based on the existence of solutions to $\mathbf{A}\phi \leq 0$ or $\mathbf{A}\psi \leq -1$ with various other properties. These can be convenient because we only have to solve inequalities, for which we can sometimes guess solutions, and in addition the results below allow the inequalities to *fail* for a finite set of states which again makes guessing a bit easier. (Direct application of Theorem 4.1 with $B = \{b\}$ allows failure at only one state b.) We will state the results first then turn to their proofs. This will be followed by application to branching processes and random walks in higher dimensions. Readers may prefer to look at those applications before reading the proofs.

Theorem 4.7. *Suppose a function $\phi : S \to \mathbb{R}$ exists for which $\phi(s) \to 0$ as $|s| \to \infty$, $\mathbf{A}\phi(s) \leq 0$ for all s except those in a finite set B, and $\phi(b) > 0$ for $b \in B$. Then the Markov chain is transient.*

Theorem 4.8. *Suppose a function $\phi : S \to \mathbb{R}$ exists satisfying $\phi(s) \to +\infty$ as $|s| \to \infty$ and $\mathbf{A}\phi(s) \leq 0$ for all but finitely many s. Then the Markov chain is recurrent.*

Both these theorems seem to presume that $|s|$ is defined for $s \in \mathcal{S}$ in order for $\lim_{|s|\to\infty} \phi(s)$ to make sense. If you like you can just assume just assume that $\mathcal{S} \subseteq \mathbb{Z}^d$. But actually all that is needed is the appropriate definition of $\lim_{|s|\to\infty} h(s) = \ell$: for any $\epsilon > 0$ there are at most finitely many $s \in \mathcal{S}$ for which $|h(s) - \ell| < \epsilon$ fails.

Theorem 4.9. *Suppose a function $\psi : S \to \mathbb{R}$ exists which is bounded below and $\mathbf{A}\psi(s) \leq -1$ for all but finitely many s. Then the Markov chain is positive recurrent.*

Note that $\mathbf{A}\psi(s) < 0$ is not sufficient for positive recurrence; see Problem 4.18.

4.3.3 The Proofs

In preparation for the proofs of the theorems we need to establish some lemmas. We will need something like Lemma 4.6 for transience and recurrence.

Lemma 4.10. *Suppose X_n is an irreducible Markov chain and $C \subseteq \mathcal{S}$ is a nonempty subset and there exists $a \in \mathcal{S} \setminus C$ for which $P_a(\mathcal{T}_C^+ < \infty) < 1$. Then the chain is transient.*

Proof. Consider any $c \in C$. Clearly $\mathcal{T}_C^+ \leq \mathcal{T}_c^+$. According Theorem 4.2, if the chain was recurrent then $P_a(\mathcal{T}_c^+ < \infty) = 1$ which would imply that $P_a(\mathcal{T}_C^+ < \infty) = 1$. Since this is contrary to the hypothesis the chain must be transient. □

Lemma 4.11. *Suppose X_n is an irreducible Markov chain and there is a finite proper subset $B \subsetneq \mathcal{S}$ for which $P_i(\mathcal{T}_B < \infty) = 1$ for all $i \in \mathcal{S}$. Then the chain is recurrent.*

Proof. Let $u(i) = P_i(\mathcal{T}_B < \infty)$. Pick any $a \in B$ and let $v(i) = P_i(\mathcal{T}_a < \infty)$. We know that $v(a) = 1$ and $v(j) = \sum p_{j,k} v(k)$ for $j \neq a$. By irreducibility we know that $v(i) > 0$ for every i.

We claim that the smallest value of v occurs in B. Let $\alpha = \min_B v$. Consider $\phi = v(\cdot)/\alpha$. Then $\phi \geq 1$ on B and $\mathbf{A}\phi \leq 0$ on B^c. By Theorem 4.1 it follows that $u \leq \phi$ on B^c. But since $u \equiv 1$ this means that $1 \leq \phi$ and therefore $\alpha \leq v$ on B^c. As a consequence $\alpha \leq v(i)$ for all i.

We claim that $\alpha = 1$. Since B is finite there exists $i \in B$ with $v(i) = \alpha$. If $i = a$ then $\alpha = v(a) = 1$ by definition of v. Suppose $i \neq a$. Then $\alpha = v(i) = \sum_j p_{i,j} v(j)$ and $v(j) \geq \alpha$ implies that $v(j) = \alpha$ for any j with $p_{i,j} > 0$. It follows that $v(j) = \alpha$ for all $i \rightsquigarrow j$. That incudes $j = a$. Having shown that $\alpha = 1$ it follows that all $v(j) \geq 1$. But $v(j) \leq 1$ since by definition $v(j)$ is a probability. Therefore all $v(j) = 1$. I.e. $P_j(\mathcal{T}_a < \infty) = 1$ for all j. This establishes recurrence of the chain, by Theorem 4.2. □

In the theory of partial differential equations there is a standard result called called the maximum principle. We have named the following lemma "The Maximum Principle" because its analogue for Brownian Motion (with $\mathbf{A}$ replaced by $\mathcal{A} = \frac{1}{2}\frac{\partial^2}{\partial x^2}$) *is* the standard PDE version.

Lemma 4.12 (Maximum Principle)**.** *Suppose X_n is an irreducible Markov chain and $B \subsetneq \mathcal{S}$ is a finite proper subset of the state space. If $\phi : \mathcal{S} \to \mathbb{R}$ is bounded above, $\mathbf{A}\phi(i) \geq 0$ for all $i \in B$, and $\phi(j) \leq M$ for all $j \notin B$ then*

$$\phi(i) \leq M \text{ for all } i \in B.$$

Proof. Let $L = \max_B \phi$. We will show that if $L > M$ then the chain is not irreducible. Let $i \in B$ with $\phi(i) = L$. We know from $\mathbf{A}\phi(i) \geq 0$ that

$$\phi(i) \leq \sum_j p_{i,j} \phi(j).$$

If $L > M$ then all $\phi(j) \leq L$. It follows that any j with $p_{i,j} > 0$ must have $\phi(j) = L$, which means $j \in B$. Continuing in this way we find that $\{i \in B : \phi(i) = L\}$ is a closed set of states contained in B. But since the chain is irreducible and there exists a state outside B this is not possible. Thus we must have $L \leq M$, which proves the lemma. □

We are now ready to prove our three theorems.

Proof of Theorem 4.7. We want to apply the first part of Theorem 4.1, but that requires $\phi \geq 0$ and $\phi \geq 1$ on B. So we need to modify ϕ first. We are going to use $\tilde{\phi} = (1 - \alpha) + \beta\phi$ for carefully choosen constants $\alpha, \beta > 0$.

Since $\phi(s) \to 0$ as $|s| \to \infty$ it must be that ϕ is bounded: $|\phi| \leq c$ for some $c > 0$. Since $\phi > 0$ on the finite set B there is $0 < \epsilon < 1$ with $\phi(b) \geq \epsilon$ for $b \in B$. First choose $\beta > 0$ with $c\beta < 1/2$ and then choose $\alpha > 0$ with $\alpha < \min(1/2, \beta\epsilon)$. Now consider $\tilde{\phi} = (1-\alpha) + \beta\phi$. For $b \in B$ we have

$$\tilde{\phi}(b) = (1-\alpha) + \beta\phi(b) \geq (1-\alpha) + \beta\epsilon > 1,$$

since $\alpha < \beta\epsilon$. Outside of B since $\beta \geq 0$ it follows that $\mathbf{A}\tilde{\phi} \leq 0$, and

$$\tilde{\phi}(i) = (1-\alpha) + \beta\phi(i) \geq (1-\alpha) - \beta c > 0,$$

since $\alpha + \beta c < \frac{1}{2} + \frac{1}{2} = 1$. We are now justified in applying the first part of Theorem 4.1 to $\tilde{\phi}$ to conclude that

$$P_i(\mathcal{T}_B < \infty) \leq \tilde{\phi}(i) \text{ for } i \notin B.$$

Since $\phi \to 0$ we have $\tilde{\phi}(i) \to 1 - \alpha$ as $|i| \to \infty$. So there exists $i \notin B$ with $\tilde{\phi}(i) < 1$. By Lemma 4.10 the chain must be transient. □

Proof of Theorem 4.8. Let B be a nonempty finite set containing the exceptions to $\mathbf{A}\phi \leq 0$. We want to show that $P_i(\mathcal{T}_B = \infty) = 0$ for all $i \notin B$. This is equivalent to $P_i(\mathcal{T}_B < \infty) = 1$ which implies recurrence by Lemma 4.11. By adding a constant to ϕ we can assume that $\phi \geq 0$. Let $v(i) = P_i(\mathcal{T}_B = \infty)$. We know that $v(b) = 0$ for $b \in B$ and $\mathbf{A}v(i) = 0$ for $i \notin B$. Consider any $\epsilon > 0$ and let $u = v - \epsilon\phi$. We know $u(b) = 0 - \epsilon\phi(b) \leq 0$ for $b \in B$. As $|i| \to \infty$ we know that $u(i) \to -\infty$. That means that the set $F = \{i \in \mathcal{S} : u(i) > 0\}$ is finite. Now if $i \in F$ then $i \notin B$ because $u < 0$ on B. So for all $i \in F$ we have

$$\mathbf{A}u(i) = \mathbf{A}v(i) - \epsilon\mathbf{A}\phi(i) \geq 0.$$

This means the maximum principle of Lemma 4.12 applies. Since $u(j) \leq 0$ for all $j \notin F$ we conclude that $u(i) \leq 0$ for all $i \in F$ as well. Therefore for *every* $i \in \mathcal{S}$ we have

$$v(i) \leq \epsilon\phi(i).$$

Since this is true for every $\epsilon > 0$ we can let $\epsilon \downarrow 0$ and conclude that $v(i) = 0$, which is what we wanted to show. □

Proof of Theorem 4.9. By adding a constant we can assume $\psi \geq 0$. Let B be the finite set where $\mathbf{A}\psi \leq -1$ fails. It follows from the second part of Theorem 4.1 that $E_i[\mathcal{T}_B] \leq \psi(i) < \infty$ for $i \notin B$. Now Theorem 4.4 implies positive recurrence. □

4.3.4 Branching Processes

A branching process Z_n is a nonnegative integer-valued Markov process which evolves as follows. If $Z_n = m$ then we take m independent random variables Y_i with a prescribed common distribution and let

$$Z_{n+1} = Y_1 + \cdots Y_m.$$

The Y_i should be nonnegative integer-valued. Think of Z_n as the size of a population of some organism. Each individual lives one unit of time and then dies while giving birth to a random number of offspring, distributed as the Y_i. We will assume $Z_0 = 1$. If $P(Y_i = 0) = 0$ then $Z_{n+1} \geq Z_n$. But if $Y_i = 0$ is possible, then it is possible for the population to reach $Z_n = 0$ at some time n. Then $Z_{n+k} = 0$ for all $k \geq 0$. I.e. the population dies out and become extinct: 0 is an absorbing state for the chain Z_n.

We want to focus on whether or not extinction is certain: $P_1(\mathcal{T}_0 < \infty) = 1$ or < 1. We can address this using our results about recurrence or transience. Note however that Z_n is not an irreducible chain, because of the absorbing state at 0. But we are only concerned with what happens up to the first visit to 0. If we modify what the chain does from 0 we can make it irreducible without changing the probability of reaching 0: just let $p_{0,1} = 1$. In other words when the population becomes extinct then one individual is miraculously born at the next time. We might call this a *branching process with spontaneous regeneration.* Provided only that $P(Y_i = 0)$ and $P(Y_i > 1)$ are both positive this makes the chain irreducible, so the results of this chapter all apply. We will proceed with this regenerating version of the branching process.

Our question is whether this chain is recurrent or transient. If it is recurrent then it is certain to eventually reach 0, i.e. go extinct, but if it is transient then there is a positive probability of surviving forever. The answer depends simply on the mean number of offspring for a single parent,

$$\mu = E[Y_i].$$

It is natural to speculate that larger μ means higher probability of avoiding extinction, so we expect the chain to be transient for larger μ and recurrent for smaller μ. Consider $\phi(i) = i$ as a candidate for our Therorems 4.8 or 4.7.

$$\sum_j p_{k,j}\phi(j) = E\left[\sum_1^k Y_i\right] = k\mu.$$

So we can apply one of our theorems if $k\mu \leq k$, i.e. if $\mu \leq 1$. Since $\phi(i) = i \to \infty$ as $i \to \infty$ we can apply Therorem 4.8 if $\mu \leq 1$. Note that our calculation of $\sum_j p_{k,j}\phi(j)$ above is actually incorrect if $k = 0$, because of our modification of the process. But the theorem allows a finite number of exceptions to $\mathbf{P}\phi(k) \leq \phi(k)$ so this is not an obstacle. We conclude that the chain is recurrent if $\mu \leq 1$, meaning that the original branching process is certain to eventually go extinct.

If $\mu > 1$ it certainly seems reasonable to expect $Z_n \to \infty$ which would mean the chain is transient. We can establish this using Therorem 4.7 but it takes a little more care to find an appropriate ϕ. To use the theorem we need $\phi(i) \to 0$ as $i \to \infty$. We will see that $\phi(i) = \gamma^i$ will work if we choose $0 < \gamma < 1$ carefully. Using independence of the Y_i,

$$\sum_j p_{k,j}\phi(j) = E[\gamma^{\sum_1^k Y_i}] = E[\gamma^{Y_i}]^k.$$

We need to choose γ so that $E[\gamma^{Y_i}]^k \leq \gamma^k$, i.e. so that

$$E[\gamma^{Y_i}] \leq \gamma$$

Now $E[s^{Y_i}] = \hat{g}(s)$ is the generating function for Y_i. So we are hoping to find $0 < \gamma < 1$ for which $\hat{g}(\gamma) \leq \gamma$. The two sides of this inequality are equal for $\gamma = 1$. Now $\hat{g}'(1) = E[Y_i] = \mu$, so that under our assumption that $\mu > 1$ we have $\hat{g}(\gamma) < \gamma$ for γ close to but sightly less that 1. So a γ as desired does indeed exist if $\mu > 1$. Since $\phi(i) = \gamma^i \to 0$ as $i \to \infty$ Therorem 4.7 does apply. (Again, that our calculation for $\mathbf{P}\phi(k) \leq \phi(k)$ may be wrong for $k = 0$ makes no difference because the theorem allows a finite number of exceptions.) Thus the chain is transient if $\mu > 1$, meaning that the original branching process has positive probability of surviving forever.

A more precise question is what the actual probability of extinction actually is: $\eta = P_1(\mathcal{T}_0 < \infty) =?$ This has a very nice answer in terms of the generating function $\hat{g}(s)$ for Y. It turns out that η is the smallest nonnegative fixed point of $\hat{g}$: $\eta = \hat{g}(\eta)$. Perhaps this is not too surprizing given that $\hat{g}(\gamma) \leq \gamma$ already emerged in our discussion above. See Section 5.4 of [25] for a discussion of this description of η.

4.3.5 Random Walks in Higher Dimensions

Example 4.6 considered the random walk in one dimension, both the symmetric and asymmetric cases. Consider now the symmetric random walk in $\mathbb{Z}^d$. This moves up or down *one coordinate at a time* with equal probabilities of $\frac{1}{2d}$, but no diagonal moves. Our goal is to determine its transience or recurrence.

There are a couple ways to resolve this, but they all involve some careful calculation. One way is to determine the convergence or divergence of $\sum_{n=0}^{\infty} p_{0,0}(n)$ in some explicit or approximatecalculation. This is possible using Stirling's asymptotic formula for $n!$. For $d = 1$ we are concerned with the convergence of

$$\sum_{n=0}^{\infty} p_{0,0}(n) = \sum_{k=0}^{\infty} \binom{2k}{k} \frac{1}{2^{2k}}.$$

Stirling's asymptotic formula for $n!$ says that

$$n! \sim \sqrt{2\pi n}(n/e)^n \text{ as } n \to \infty.$$

Using this

$$\begin{aligned}\binom{2k}{k}\frac{1}{2^{2k}} &= \frac{(2k)!}{(k!)^2}\frac{1}{2^{2k}} \\ &\sim \frac{\sqrt{2\pi}\sqrt{2k}(2k/e)^{2k}}{2\pi k(k/e)^{2k}}\frac{1}{2^{2k}} \\ &= \frac{1}{\sqrt{\pi}}\frac{1}{\sqrt{k}}.\end{aligned}$$

Since $\sum 1/\sqrt{k}$ is divergent it follows that $\sum p_{0,0}(n)$ is divergent, implying recurrence of the symmetric random walk in one dimension. We already knew this from Example 4.6. The point here is to indicate how this can be done with an accurate enough analysis of $p_{0,0}(n)$.

In $d = 2$ dimensions the chain again has period 2, so $p_{0,0}(\text{odd}) = 0$. The explicit formula for the even transitions is

$$p_{0,0}(2k) = \left[\binom{2k}{k}\frac{1}{2^{2k}}\right]^2.$$

This deceptively simple result is tricky to justify; see Norris [45]. But given this formula the above tells us that

$$p_{0,0}(2k) \sim \left[\frac{1}{\sqrt{\pi}}\frac{1}{\sqrt{k}}\right]^2 = \frac{1}{\pi k}.$$

Since $\sum 1/k$ diverges so does $\sum p_{0,0}(n)$. Thus the 2-dimensional symmetric random walk is also recurrent, in fact null recurrent by the remarks at the top of page 86 since $p_{0,0}(n) \to 0$.

In $d = 3$ the calculations are harder yet. After a more intricate analysis (which we omit) it turns out that

$$p_{0,0}(2k) \le Ck^{-3/2} \text{ for some constant } C.$$

This is enough to say that $\sum p_{0,0}(n)$ is *con*vergent so the 3-dimensional symmetric random walk is transient! In dimensions higher than 3 the random walk is also transient.

Varadhan [63] observes out that these conclusions can also be reached using our Theorems 4.7 and 4.8. For the symmetric random walk in $d = 2$ dimensions it turns out that

$$v(x) = \log(|x|) - 1/|x|$$

has $\mathbf{P}v(x) \le v(x)$ for $|x| > 1$ and $v(x) \to \infty$ as $|x| \to \infty$, so that recurrence follows from Theorem 4.8. It takes a two-dimensional walk at least as long as a one-dimensional walk to return to 0 so null recurrence for $d = 2$ follows from null recurrence for $d = 1$.

For $d \ge 3$ the function

$$v(x) = 1/\sqrt{|x|}$$

satisfies $\mathbf{P}v(x) \le v(x)$ if $|x| > 1$ and has $v(x) \to 0$ as $|x| \to \infty$ so Theorem 4.7 implies transience in all dimensions $d \ge 3$. Problem 4.19 asks you to supply more of the details.

4.4 Equilibrium Distributions and Ergodicity

This final section is concerned with the existence of equlibrium distributions and convergence of $\mathbf{P}^n$ for infinite state spaces. In Section 2.4.1 we proved that on a finite state space a equilibrium distribution always exists. In this chapter we have limited ourselves to irreducible chains. An irreducible chain on a finite state space is always positive recurrent, and we will see that on infinite state spaces positive recurrent chains again always have equilibrium distributions. But we will also see that equilibrium distributions do *not* exist for null recurrent or transient chains.

Example 4.7. Let X_n be the reflecting random walk on $\{0, 1, 2, \ldots\}$ as discussed on page 75. We know from those calculations that the chain is positive recurrent if and only if $p < q$. If you consider solutions of $\pi = \pi\mathbf{P}$ you will find that there is a nonnegative solution with $\sum \pi_n = 1$ when $p/q < 1$, but not when $p/q \ge 1$; see Problem 4.14. Based on the results we are about to prove this gives another demonstration that the chain is positive recurrent if and only if $p < q$.

4.4.1 The Transient and Null-Recurrent Cases

Theorem 2.6 tells us that an irreducible chain on a finite state space is always positive recurrent. Only for infinite state spaces can an irreducible chain be either transient or null-recurrent. The following theorem says that in those cases there are no equilibrium distributions and that $\mathbf{P}^n \to [0]$.

Theorem 4.13. *For an irreducible Markov chain which is either transient or null-recurrent the following hold.*

a) $\mathbf{P}^n \to [0]$ *(entry-wise).*

b) No equilibrium distribution exists.

c) For each $a \in \mathcal{S}$ *and any initial distribution we have* $\frac{1}{N}\sum_1^N \delta_a(X_n) \to 0$ *with probability* 1.

Proof. In the transient case Corollary 4.3 tells us that $p_{i,j}(n) \to 0$. In the null-recurrent case we know from Theorem 3.10 that $p_{i,i}(n) \to 0$ for each $i \in \mathcal{S}$. To extend this to $p_{i,j}(n)$ recall from (4.2) that

$$p_{i,j}(n) = \sum_{k=0}^{n} f_{i,j}(k) p_{j,j}(n-k).$$

Because $\sum_{k=0}^{\infty} f_{i,j}(k) \le 1$ we can use the Dominated Convergence Theorem to take $\lim_n \sum_k = \sum_k \lim_n$ of the right side, to obtain

$$\lim_{n\to\infty} p_{i,j}(n) = \sum_{k=0}^{\infty} f_{i,j}(k) \lim_{n\to\infty} p_{j,j}(n-k) = \sum_{k=0}^{\infty} f_{i,j}(k) 0 = 0.$$

Therefore $\mathbf{P}^n \to [0]$ entry-wise, which is what a) claimed.

To prove b) suppose there existed a equilibrium distribution μ. Then for any $a \in \mathcal{S}$ and every $n \ge 1$ we have

$$\mu_a = \sum_{j\in\mathcal{S}} \mu_j p_{j,a}(n).$$

Every term of the series $\to 0$ and since $\sum \mu_j = 1$ the Dominated Convergence Theorem applies and we conclude that

$$\mu_a = \sum_{j\in\mathcal{S}} \lim_n \mu_j p_{j,a}(n) = 0.$$

This is not possible since $\sum \mu_a = 1$, so no equilibrium distribution can exist.

Now we turn to c). In the transient case X_n visits a only a finite number of times. (See Problem 4.4.) Therefore

$$\frac{1}{N}\sum_1^N \delta_a(X_n) \to 0.$$

In the null-recurrent case we know from Theorem 3.10 that c) holds if $X_0 = a$. (This is the case of $r_a = \infty$ in that theorem.) For an arbitrary initial distribution we know from Theorem 4.4 that $\mathcal{T}_a < \infty$ with probability 1 and therefore $\frac{N}{N+\mathcal{T}_a} \to 1$. It follows that

$$\lim_N \frac{1}{N}\sum_1^N \delta_a(X_n) = \lim_N \frac{1}{N}\sum_{\mathcal{T}_a+1}^{\mathcal{T}_a+N} \delta_a(X_n).$$

Therefore, using the strong Markov Property and the fact that $X_{\mathcal{T}_a} = a$,

$$\begin{aligned} P\left(\lim_N \frac{1}{N}\sum_1^N \delta_a(X_n) = 0\right) &= P\left(\lim_N \frac{1}{N}\sum_{\mathcal{T}_a+1}^{\mathcal{T}_a+N} \delta_a(X_n) = 0\right) \\ &= E\left[P\left(\lim_N \frac{1}{N}\sum_{\mathcal{T}_a+1}^{\mathcal{T}_a+N} \delta_a(X_n) = 0 \middle| X_{0:\mathcal{T}_a}\right)\right] \\ &= E\left[P_a\left(\lim_N \frac{1}{N}\sum_1^N \delta_a(X_n) = 0\right)\right] \\ &= E[1] = 1. \end{aligned}$$

□

4.4.2 The Positive Recurrent Case

This is the infinite state space case most like Chapter 2. We will prove that there does exist an invariant distribution. The argument we gave on page 21 does not work in an infinite state space. (Although a convergent subsequence $\pi^{(m')} \to \pi$ does exist the limit need not be a probability distribution: without additional hypotheses we cannot justify $\pi\mathbf{P} = \lim_{m'}(\pi^{(m')}\mathbf{P}) = \mathbf{1}$ because this is now an interchange of limit with infinite series.) Our proof will present a different construction based on positive recurrence.

Theorem 4.14. *Suppose X_n is an irreducible, positive recurrent Markov chain.*

a) There is a unique equilibrium distribution given by

$$\pi_a = 1/E_a[\mathcal{T}_a^+] \text{ for each } a \in \mathcal{S}.$$

b) If the chain is aperiodic then $\lim_n p_{i,a}(n) = \pi_j$ for each $i, a \in \mathcal{S}$.

c) For each $a \in \mathcal{S}$ and any initial distribution we have $\frac{1}{N}\sum_{n=1}^N 1_a(X_n) \to \pi_a$ with probability 1.

Proof. Pick any state initial $s \in \mathcal{S}$ and let $\mathcal{T}_s^+$ be our usual first return time. Let $m_s = E_s[\mathcal{T}_s^+]$ which is finite by hypothesis. The idea is to start the chain at $X_0 = s$ and follow it from X_1 to $X_{\mathcal{T}_s^+}$ counting how many times it visits each state i prior to its first return to s:

$$N_{i\neg s} = \sum_{n=0}^{\mathcal{T}_s^+-1} 1_i(X_n),$$

using the notation of Lemma 4.5. Let π_i be the mean of these, normalized:

$$\pi_i = E_s\left[N_{i\neg s}\right]/E_s[\mathcal{T}_s^+]. \tag{4.6}$$

These are a probability distribution, because

$$\sum_i E_s\left[\sum_{n=0}^{\mathcal{T}_s^+-1} 1_i(X_n)\right] = E_s\left[\sum_{n=0}^{\mathcal{T}_s^+-1}\sum_i 1_i(X_n)\right] = E_s\left[\mathcal{T}_s^+\right].$$

We claim that the π_i in fact comprise an equilibrium distribution for the chain: $\pi\mathbf{P} = \pi$. To see this consider any state k. Observe that if $X_0 = s$ then $1_k(X_0) = 1_k(X_{\mathcal{T}_s^+}) = 1_k(s)$. This justifies the first line of the

following.

$$E_s\left[\sum_{n=0}^{\mathcal{T}_s^+-1} 1_k(X_n)\right] = E_s\left[\sum_{n=0}^{\mathcal{T}_s^+-1} 1_k(X_{n+1})\right]$$
$$= E_s\left[\sum_{n=0}^{\infty} 1_k(X_{n+1})1_{\mathcal{T}_s^+>n}\right]$$
$$= \sum_{n=0}^{\infty} E_s\left[1_k(X_{n+1})1_{\mathcal{T}_s^+>n}\right].$$

Now lets work on an individual term from this summation.

$$E_s\left[1_k(X_{n+1})1_{\mathcal{T}_s^+>n}\right] = E_s\left[E_s\left[1_k(X_{n+1})1_{\mathcal{T}_s^+>n}|X_{0:n}\right]\right]$$
$$= E_s\left[E_s\left[1_k(X_{n+1})|X_{0:n}\right]1_{\mathcal{T}_s^+>n}\right]$$
$$= E_s\left[p_{X_n,k}1_{\mathcal{T}_s^+>n}\right]$$
$$= E_s\left[\sum_j 1_j(X_n)p_{j,k}1_{\mathcal{T}_s^+>n}\right]$$
$$= \sum_j E_s\left[1_j(X_n)1_{\mathcal{T}_s^+>n}\right]p_{j,k}$$

Taking $\sum_{n=0}^{\infty}$ of this we find that

$$E_s\left[\sum_{n=1}^{\mathcal{T}_s^+} 1_k(X_n)\right] = \sum_{n=0}^{\infty}\sum_j E_s\left[1_j(X_n)1_{\mathcal{T}_s^+>n}\right]p_{j,k}$$
$$= \sum_j E_s\left[\sum_{n=0}^{\infty} 1_j(X_n)1_{\mathcal{T}_s^+>n}\right]p_{j,k}$$
$$= \sum_j E_s\left[\sum_{n=0}^{\mathcal{T}_s^+-1} 1_j(X_n)1_{\mathcal{T}_s^+>n}\right]p_{j,k}$$

Dividing both sides by $E_s[\mathcal{T}_s^+]$ this becomes

$$\pi_k = \sum_j \pi_j p_{j,k},$$

proving that π is a equilibrium distribution. Observe that $\pi_s = 1/r_s$ for the particular state s used in the construction. But this construction of π_i appears to depend on the choice of s. And we have *not* established the uniqueness of π yet. Once we do establish uniqueness *then* we will be able to say that $\pi_i = 1/r_i$ for all $i \in \mathcal{S}$.

Consider any $i, j \in \mathcal{S}$. We know

$$p_{i,j}(n) = \sum_{k=0}^{n} f_{i,j}(k)p_{j,j}(n-k).$$

By Theorem 3.10 we know that $p_{j,j}(n-k) \to 1/r_j$. By Theorem 4.2 part 5) we know that $\sum_{k=0}^{\infty} f_{i,j}(k) = 1$. The Dominated Convergence Theorem allows us to let $n \to \infty$ in the above to obtain

$$\lim_n p_{i,j}(n) = \sum_{k=0}^{\infty} \lim_n [f_{i,j}(k)p_{j,j}(n-k)] = \sum_{k=0}^{\infty} f_{i,j}(k)1/r_j = 1/r_j.$$

Now consider *any* invariant distribution π. Again applying the Dominated Convergence Theorem we find

$$\pi_k = \lim_n \sum_j \pi_j p_{j,k}(n) = \sum_j \pi_j \lim_n p_{j,k}(n) = \sum_j \pi_j 1/r_k = 1/r_k.$$

This shows that π is unique, completing the proof of a) and b).

Part c) is proven in the same way as for Theorem 4.13: apply Theorem 3.10, this time with $r_a < \infty$. □

Example 4.8. Let's come back to Example 2.1 once again. We have encountered the equilibrium distribution several times above. In Example 2.1 we looked at $\mathbf{P}^n$ for large n and saw that all rows seemed to converge to

$$\pi \approx (0.11605, 0.22244, 0.30947, 0.35203),$$

which we later realized (page 19) had to be the equilibrium distribution. Part b) of the theorem above said this had to happen. Later we calculated π exactly as a (left) eigenvector in Example 2.9:

$$\pi = (60/517, 115/517, 160/517, 182/517).$$

Now let's consider it again from the perspective of Theorem 4.14. Let's pick a state, say $s = 2$, and calculate π using the formula (4.6) of the proof. For that we will need to find the mean

$$E_2\left[\sum_{n=0}^{\mathcal{T}_2^+ - 1} 1_i(X_n)\right]$$

for each state i. Find these by solving the equations of Lemma 4.5. For $i = 4$ we did the calculation in Example 4.3. We know that for $i = 2$ the result is 1 (the case of $a = b$ in Lemma 4.5). We have to repeat that calculation for $i = 1$ and $i = 3$. The results (from Problem 4.6) are

$$E_2\left[\sum_{n=0}^{\mathcal{T}_2^+ - 1} 1_1(X_n)\right] = 0.521739$$

$$E_2\left[\sum_{n=0}^{\mathcal{T}_2^+ - 1} 1_2(X_n)\right] = 1$$

$$E_2\left[\sum_{n=0}^{\mathcal{T}_2^+ - 1} 1_3(X_n)\right] = .3913$$

$$E_2\left[\sum_{n=0}^{\mathcal{T}_2^+ - 1} 1_4(X_n)\right] = 1.58261.$$

The sum of these gives

$$E_2[\mathcal{T}_2^+] = \sum_i E_2\left[\sum_{n=0}^{\mathcal{T}_2^+ - 1} 1_i(X_n)\right] = 4.49565.$$

So from equation (4.6) we find the equlibrium distribution to be

$$\pi = (0.116054, 0.222437, 0.309478, 0.352031),$$

agreeing with what we calculated previously.

Finally we can examine part c) of the theorem with a simulation. If we produce a sample run of $10,000$ steps (using the code on page 9 for instance) and count how many visits the chain makes to each state we obtain $(1129, 2175, 3125, 3571)$. Dividing by $N = 10,000$ we obtain the following approximation to π:

$$(0.1129, 0.2175, 0.3125, 0.3571)$$

Problems

Problem 4.1
Suppose we roll a conventional fair dice repeatedly. At each time n let Y_n be the number of rolls since we last observed a 6. Explain why Y_n is a Markov chain and find its transition probabilities.

. . . . . Sixes

Problem 4.2
Suppose X_n and Y_n are independent Markov chains both with transition matrix

$$\mathbf{P} = \begin{bmatrix} 0 & 1/2 & 1/2 & 0 \\ 0 & 0 & 1/2 & 1/2 \\ 1/2 & 0 & 0 & 1/2 \\ 1/2 & 1/2 & 0 & 0 \end{bmatrix}$$

Let $\mathcal{J}$ denote the first time that the two chains are in the same state. In other words $\mathcal{J}$ is the first n for which $X_n = Y_n$. Do a calculation using the 16-state Markov chain (X_n, Y_n) to compute $E[\mathcal{J}]$ assuming $X_0 = 1$ and $Y_0 = 3$. (Turn in your MATLAB code as well as the results.)

. . . . . FE2

Problem 4.3
Let X_n be a (generalized) random walk on $\mathbb{Z}^d$. We can express it as

$$X_n = X_0 + \sum_{i=1}^{n} Y_i,$$

where the Y_i are an i.i.d. $\mathbb{Z}^d$-valued random variables. Let $\mu = E[Y_i]$ (this is a vector). Use the Strong Law of Large Numbers to show that if $\mu \neq [0]$ then the chain is transient.

. . . . . RWT

Problem 4.4
Theorem 4.2 says that $\sum_{n=0}^{\infty} p_{i,i}(n) = \infty$ is equivalent to $\sum_{n=0}^{\infty} f_{i,i}(n) = 1$. (The notation $f_{i,j}(n)$ was introduced on page 78.) One way to understand the connection is to recognize that $\sum_{n=0}^{\infty} p_{i,i}(n)$ gives the mean number of returns to i, which should be infinite if i is recurrent. To be precise let N_i be the total number of visits X_n makes to i:

$$N_i = \sum_{n=0}^{\infty} 1_{\{i\}}(X_n).$$

Prove that $E_i[N_i] = \sum_{n=0}^{\infty} p_{i,j}(n)$.

Thus $E_i[N_i]$ is infinite or finite when i is recurrent or transient respectively. Observe that this would be infinite if $P_i(N_i = \infty) > 0$. So if i is transient then starting from i the chain is certain to visit i only a finite number of times. This gives another argument for part 2 of Theorem 4.2, but explain why it also shows that $P_i(N_i = \infty)$ can only be $= 0$ or $= 1$.

. . . . . MeanRet

Problem 4.5
Prove the equations (4.2).

. . . . . f-convl

Problem 4.6
Repeat the calculation of Example 4.3 using $b = 1$ and $b = 3$ (keeping $a = 2$).

. . . . . MSH

Problem 4.7
We know from Theorem 2.6 that an irreducible Markov chain on a finite state space is positive recurrent. For another proof observe that

$$\mathcal{T}_a^+ = \sum_{b \neq a} N_{b\neg a}.$$

Now use Lemma 4.5 to prove $E_a[\mathcal{T}_a^+] < \infty$ if $\mathcal{S}$ is finite.

. . . . . FinRecur

Problem 4.8
Suppose X_n is an irreducible chain and there exists $b \neq a$ for which

$$P_b(\mathcal{T}_a^+ < \infty) = 1.$$

It is not necessarily true that a is recurrent. (We really need the "for all b" in Theorem 4.2 part 2.) Find one of the examples we have discussed which illustrates this.

. . . . . NoRec

Problem 4.9
Referring to (4.5), prove that for a finite state space the matrix $\mathbf{I} - s\mathbf{P}$ is invertible for $|s| < 1$. (Hint: Suppose $(\mathbf{I} - s\mathbf{P})\mathbf{u} = [0]$. Show that $(\mathbf{I} - s^2\mathbf{P}^2)\mathbf{u} = [0]$ and in fact $(\mathbf{I} - s^n\mathbf{P}^n)\mathbf{u} = [0]$ for all positive integers n. Conclude that $\mathbf{u} = [0]$.)

. . . . . IsPinv

Problem 4.10
Find a stochastic Lyapunov function (ϕ as in Theorem 4.8) which implies that the symmetric random walk on $\mathbb{Z}$ is recurrent. (Drawing a graph and guessing should work.)

. . . . . Lya1

Problem 4.11
Consider the reflecting random walk on $\mathbb{Z}^+$: $i \to i \pm 1$ with probabilities $p, 1-p$ respectively when $i \geq 1$; and $0 \to 0, 1$ with probabilities $p, 1-p$. Consider $\phi(i) = \gamma^i$ where $p = \frac{1}{\gamma+1} < \frac{1}{2}$ and explain why recurrence follows from Theorem 4.8.

. . . . . RRWphi

Problem 4.12
The formula for the equilibrium distribution of Theorem 4.14 part a) can be confirmed another way. Assume that $p_{i,i}(n) \to \pi_i$. You may take for granted that this implies that $(1-s)\hat{p}_{i,i}(s) \to \pi_i$ as $s \to 1^-$. Now $E_i[\mathcal{T}_i^+] = \hat{f}_{i,i}(1-)$. Use the generating function relationship (4.4) to establish the relation between $E_i[\mathcal{T}_i^+]$ and π_i.

Secondly we would like see that $\pi_i = 1/E_i[\mathcal{T}_i^+]$ is a probability distribution: $\sum_i \pi_i = 1$. To prove this directly explain why

$$\hat{p}_{i,a}(s) = \frac{\hat{f}_{i,a}(s)}{1 - \hat{f}_{a,a}(s)}, \quad \sum_a \hat{p}_{i,a}(s) = \frac{1}{1-s}, \quad \hat{f}_{i,a}(1-) = 1$$

and then put these pieces together to show that $\sum_a 1/E_a[\mathcal{T}_a^+] = 1$. (You will probably encounter an interchange of limit and summation. You do not need to provide the technical justification for that.)

. . . . . EqFor

Problem 4.13
For the asymmetirc random walk with $p < q$ (page 76) find a formula for the stationary distribution π_i and verify that it satisfies the equation $\pi = \pi\mathbf{P}$.

. . . . . Exit

Problem 4.14
Let X_n be the reflecting random walk on $\{0, 1, 2, \ldots\}$ as discussed on page 75 and let $\rho = p/q$. Verify the assertion of Example 4.7 that there is a nonnegative solution of $\pi = \pi\mathbf{P}$ with $\sum \pi_n = 1$ when $\rho < 1$, but not when $\rho \geq 1$.

. . . . . RfRW

Problem 4.15
In this problem you are to simulate the "reflecting random walk" of Problem 4.13, using $p = 1/3$. Starting at $X_0 = 0$ produce a sample path up to $n = 10000$. Compare the frequency of visits to 0, 1, 2, 3, 4, 5 to their theoretical limits as calculated in Problem 4.13.

. . . . . SimExit

Problem 4.16
Let X_n be the chain on $\mathbb{Z}^+$ which jumps $i \to i + 1$ with probability q_i and $i \to 0$ with probability $1 - q_i$. (Assume $0 < q_i < 1$.) Under what conditions on q_i is the chain transient? Work out and attempt to solve the equations that an equlibrium distribution π_i must satisfy. Under what conditions on q_i is the null-recurrent, and under what conditions is it positive recurrent? In the positive recurrent case find π_i.

. . . . . JO

Problem 4.17
Consider the Markov chain X on $S = \{0, 1, 2, 3, \ldots\}$ with transition probabilities

$$p_{01} = 1, \quad \text{for } i \geq 1 \; p_{ij} = \begin{cases} \frac{1}{i+1} & \text{if } j = 2i \\ \frac{i}{i+1} & \text{if } j = i - 1, \end{cases} \quad \text{and 0 for all other cases.}$$

a) Explain why this chain is irreducible.

b) Show that $E[X_{n+1} \,|\, X_n = i] = i$ for all $i \geq 1$.

c) Show that the chain is recurrent. [Hint: Part b) should tell you something to use for $\mathbf{A}\phi \leq 0$.]

. . . . . LyapEx

Problem 4.18
This problem explains the comment following Theorem 4.9. First, suppose there exists a function $\psi : S \to \mathbb{R}$ exists which is bounded below and $\mathbf{A}\psi(s) \leq -c$ for all but finitely many s, where $c > 0$ is some positive constant. Then observe that the theorem applies to ψ/c. In other words if there is a negative upperbound for $\mathbf{A}\psi(s)$, allowing a finite number of exceptions, the chain is positive recurrent. The comment following the theorem says that $\mathbf{A}\psi(s) < 0$ for all but finitely many s is *not* enough however. As an example consider the symmetric case of the reflecting random walk of page 75. Our calculations there showed that this is null-recurrent. But show that $\psi(i) = \sqrt{i}$ does have $\mathbf{A}\psi(s) < 0$ for all but finitely many s. Why is there no $c > 0$ so that $\mathbf{A}\psi(s) \leq -c$ for all but finitely many s?

. . . . . NotPosRec

Problem 4.19
In this problem your are to confirm the applicability of Theorems 4.7 and 4.8 for random walks using

$$\phi(x) = \log(|x|) - 1/|x| \text{ for } d = 2$$
$$\phi(x) = 1/\sqrt{|x|} \text{ for } d = 3$$

as claimed on page 90. First notice that these are undefined at $x = 0$ so simply take $v(0) = 0$ to complete the definitions. If you are unable to verify $\mathrm{A}\phi(s) \leq 0$ by working with the formulas then see if you can produce good experimental evidence: write a program to compute $\mathrm{A}\phi(s)$ for all $|s| \leq N$ for some large N and look for instances of $\mathrm{A}\phi(s) > 0$.

. . . . . VarLyaVer

For Further Study

In addition to the books already cited at the end of Chapter 2 some more advanced treatments are those of Grimmett and Stirzaker [25], Bremaud [10], Stroock [57], and Chung [15].

Chapter 5

Hidden Markov Chains and Elementary Filtering

For Further Study

Until such time as I write this chapter [54] provides an introduction and some references.

Chapter 6

Statistics of Markov Chains

Suppose we can observe a sequence $X_0, X_1, \ldots$ of random variables which we believe or assume come from a Markov chain, but we don't know the transition probabilities $\mathbf{P} = [p_{i,j}]$. How can we estimate the $p_{i,j}$ from our observations? The natural thing to do is to observe $X_{0:N}$ for some large N and count how many transitions of each type occurred:

$$c_{i,j} = \sum_{n=1}^{N} 1_{(X_{n-1},X_n)=(i,j)}.$$

Then the number of transitions out of state i is

$$C_i = \sum_j c_{i,j}.$$

A natural estimate of the transition probabilities is

$$\hat{p}_{i,j} = \frac{c_{i,j}}{C_i}.$$

This $\hat{p}_{i,j}$ is what we call a *statistical estimator*, a function of $X_{0:N}$ which we expect or hope will give a good approximation to the true value of the transition probability:

$$\hat{p}_{i,j} \approx p_{i,j}.$$

We want to discuss how $\hat{p}_{i,j}$ arises as the maximum likelihood estimator for $p_{i,j}$ and then apply the idea to the English language considered as a Markov chain.

6.1 The Maximum Likelihood Estimate of Transition Probabilities

Let's first talk about the basic problem of identifying the distribution of a random variable Y based on observing a value of its outcome, $y = Y$. Suppose that there is a family of possible distributions for Y determined by a parameter θ. For each θ there is a probability $P_\theta(Y = y)$ that the outcome of Y was the observed value y. We want to decide which value of θ is the the most likely based on our observation. Since we *did* observe $Y = y$ it seems reasonable to believe that the values of θ which give larger values for $P_\theta(Y = y)$ are the better guesses for the true value. The *maximum likelihood estimator* of θ is

$$\hat{\theta}(y) = \text{ the value of } \theta \text{ which maximizes } P_\theta(Y = y).$$

This is a general strategy for finding a statistical estimator in any situation which fits the above description. In many situations a formula for $\hat{\theta}(y)$ can be worked out and then studied in detail to determine its properties.

We want to apply this general maximum likelihood strategy to our problem of estimating the transition probabilities of a Markov chain. Let $s_{0:N}$ be the sequence of states that we see in our observation of $X_{0:N}$. By examining $s_{0:N}$ we see what states comprise $\mathcal{S}$. (Maybe there are some rare states that did not actually

occur in our observation, so that $\mathcal{S}$ should be a bit bigger. But there is no way we can know that from this observation alone. Hopefully we observed a long enough run of the chain so that every state did actually occur at least once in our observation.) Having identified $\mathcal{S}$ we know the number m of states and the size $m \times m$ that the transition matrix needs to be. To make the connection with the general maximum likelihood idea $X_{0:N}$ is playing the role of Y; $s_{0:N}$ is playing the role of y and θ is the transition matrix $\mathbf{P}$. The chain also has an initial distribution μ but we only have one observation of a random variable under that distribution: $s_0 = X_0$. We can't hope to estimate μ based on a single observation,so we won't even try. We have

$$\begin{aligned} P_\theta(Y = y) &= P_\mu(X_{0:n} = s_{0:N}) \\ &= \mu_{s_0} p_{s_0,s_1} p_{s_1:s_2} \cdots p_{s_{n-1},s_n}) \\ &= \mu_{s_0} \prod_{(i,j)} p_{i,j}^{c_{i,j}}, \end{aligned} \tag{6.1}$$

where $c_{i,j}$ are the observed transition counts as above. We view the right side as a function of $\theta = \mathbf{P} = [p_{i,j}]$.

Our task is the find the $m \times m$ transition matrix $\mathbf{P}$ which maximizes (6.1). Now $\mathbf{P} = [p_{i,j}]$ consists of m^2 scalar values, each between 0 and 1. But they are constrained by the requirement that each row must sum to 1: $\sum_j p_{i,j} = 1$ for each j. So there are m constraints. Thus we are faced with a constrained optimization problem. To maximize (6.1) is equivalent to maximizing its logarithm. So we want to maximize

$$\begin{aligned} f(\mathbf{P}) &= \log \left[\mu_{s_0} \prod_{(i,j)} p_{i,j}^{c_{i,j}} \right] \\ &= \log(\mu_{s_0}) + \sum_{(i,j)} c_{i,j} \log(p_{i,j}) \end{aligned}$$

over $0 \le p_{i,j}$ subject to the constraints $g_i(\mathbf{P}) = 0$ for each $i = 1, \ldots, m$ where

$$g_i(\mathbf{P}) = -1 + \sum_j p_{i,j}.$$

Lagrange multipliers is a standard technique for locating candidates for the maximizing $\hat{\mathbf{P}}$. (See [41] for a discussion of the method.) To carry this out we introduce a multiplier λ_i for each constraint and seek values for them and $\hat{\mathbf{P}} = [\hat{p}_{i,j}]$ so that

$$\nabla f(\hat{\mathbf{P}}) = \sum_i \lambda_i \nabla g_i(\hat{\mathbf{P}}).$$

We have

$$\frac{\partial}{\partial p_{i,j}} f(\mathbf{P}) = \frac{c_{i,j}}{p_{i,j}}, \quad \frac{\partial}{\partial p_{i,j}} g_i(\mathbf{P}) = 1, \text{ and } \frac{\partial}{\partial p_{i,j}} g_k(\mathbf{P}) = 0 \text{ if } i \neq k.$$

So we seek values for which $\frac{c_{i,j}}{\hat{p}_{i,j}} = \lambda_i$ and therefore

$$\hat{p}_{i,j} = \frac{c_{i,j}}{\lambda_i}.$$

To satisfy the constraints we need $1 = \sum_j \hat{p}_{i,j}$ which implies

$$\lambda_i = \sum_j c_{i,j} = C_i.$$

In this way we are led to the formula

$$\hat{p}_{i,j}(X_{0:N}) = \frac{c_{i,j}}{C_i}$$

as the maximum likelihood estimator. If some $C_i = 0$ then $c_{i,j} = 0$ for all j. In that case we can choose the i^{th} row of $\mathbf{P}$ arbitrarily, subject to $\sum_j p_{i,j} = 1$. The choice will not affect the value of $f(\mathbf{P})$ since those

terms all have coefficient 0 in f. In general the method of Lagrange multipliers only constitutes necessary conditions for a maximizer, so we should ask for some kind of argument to guarantee that our $\hat{\mathbf{P}}$ does truly maximize the constrained optimization problem. See Problem 6.1 for that.

Theorem 4.14 allows us to prove that (for irreducible recurrent chains) our estimator is *consistent*; this means that it converges to the true value in the limit.

$$\lim_{N\to\infty} \hat{p}_{i,j}(X_{0:N}) = p_{i,j}$$

Bear in mind that $\hat{p}_{i,j}(X_{0:N})$ is a random variable; if we recalculate it with a new observation of $X_{0:N}$ we will get a slightly different value. Hopefully its value is very close to $p_{i,j}$. The above limit says that this is true in the same probabilistic sense as the law of large numbers. If π is the stationary distribution of the chain, part c) of Theorem 4.14 says that

$$\frac{1}{N}C_i = \frac{1}{N}\sum_1^N 1_{\{i\}}(X_n) \to \pi_i \text{ with probability 1.}$$

Problem 2.13 says that the chain of pairs $Y_n = (X_{n-1}, X_n)$ is also recurrent (on the state space $\mathcal{D}$) with stationary distribution $\pi_{(i,j)} = \pi_i p_{i,j}$. Applying Theorem 4.14 to the chain of pairs we see that with probabiity 1 we also have

$$\frac{1}{N}c_{i,j} = \frac{1}{N}\sum_1^N 1_{\{(i,j)\}}(X_{n-1}, X_n) \to \pi_{(i,j)} = \pi_i p_{i,j}.$$

(See Problem 3.27.) Putting these together we find that with probability 1, as $N \to \infty$ we have

$$\hat{p}_{i,j}(X_{0:N}) = \frac{c_{i,j}}{C_i} = \frac{\frac{1}{N}c_{i,j}}{\frac{1}{N}C_i} \to \frac{\pi_i p_{i,j}}{\pi_i} = p_{i,j}.$$

6.2 English Language as a Markov Chain

The statistical properties of the English language are important for many purposes. Although it is a rather crude approximation, we will consider modeling English as a Markov chain. To keep this relatively simple we will consider English text as a Markov chain with state space

$$\mathcal{E} = \{\text{A,B,C, } \ldots, \text{ X,Y,Z,}\textvisiblespace\}.$$

Thus we will make no distinction between upper and lower cases letters, ignore all punctuation, numerals and non-alphabetic characters, but we do include the space character $\textvisiblespace$. To estimate the transition probabilities we want to take a large sample of English text and count the frequencies of the 27^2 different transitions, then estimate the transition probabilities as described above.

This requires a relatively large sample of English text. One well-known collection of English texts, assembled for this kind of purpose, is the Brown Corpus [24]. This is a collection of excerpts from a wide range of published sources, selected by a conference of linguists and language scholars. It contains about a million words in total. After eliminating things like hyphenated words, abbreviations, words involving numbers (e.g. "I.B.M.", "24-HOUR") the Brown Corpus data leads to the estimated transition probabilities in the file `trprob.dat`. You will explore some simple implications of this in Problem 6.2. The equilibrium distribution of this transition matrix gives the approximate frequencies of the characters in $\mathcal{E}$ in English text. Here are the results, sorted from most to least likely.

$\textvisiblespace$	E	T	A	O	I	N	S	R
.17613	.10344	.07651	.06656	.06270	.06021	.05832	.05329	.05048
H	L	D	C	U	M	F	P	G
.04516	.03371	.03242	.02549	.02234	.02087	.01920	.01666	.01602
W	Y	B	V	K	X	J	Q	Z
.01550	.01410	.01266	.00818	.00534	.00161	.00130	.00088	.00077

Given the inclusion of the space character it is probably no surprise that it occurs more frequently than any of the letters. It is common knowledge that E is the most frequent letter (by a significant margin). The most likely letter to start a word is the state i which maximizes

$$P(X_{n+1} = i | X_n = 27) = p_{27,i}.$$

We find that $i = 20$ (T).

Problem 6.1
In this problem you will confirm that the Lagrange mulitplier calculation of Section 6.1 did in fact produce the true maximizer for the constrained optimization problem. First observe that for any $\hat{x} > 0$

$$\log(x) \leq \log(\hat{x}) + \frac{1}{\hat{x}}(x - \hat{x}).$$

Therefore

$$\begin{aligned} f(\mathbf{P}) &\leq f(\hat{\mathbf{P}}) + \sum_{(i,j)} c_{i,j} \frac{1}{\hat{p}_{i,j}} (p_{i,j} - \hat{p}_{i,j}) \\ &= f(\hat{\mathbf{P}}) + \sum_{(i,j)} C_i (p_{i,j} - \hat{p}_{i,j}) \end{aligned}$$

Explain why the latter inequality is true even if some $c_{i,j} = 0$. Finally, use this to show that $f(\mathbf{P}) \leq f(\hat{\mathbf{P}})$ if $\hat{\mathbf{P}}$ satisfies the constraints.

. . . . . JustMax

Problem 6.2
Using the estimated transition probabilities in the file `trprob.mat`, what letter is most likely to appear at the beginning of a word? What letter is most likely to appear as the second letter of a word? What is the most likely two-letter combination? Can you calculate the mean word length?

. . . . . English

Problem 6.3
Using the estimated transition probabilities in the file `trprob.mat`, produce two sample sentences of 30 characters, each generated by simulating the Markov chain (starting with the stationary distribution). Do any actual English words appear in your sample sentences?

. . . . . EngSim

Problem 6.4
For an irreducible positive recurrent Markov chain, assuming X_0 has the invariant distribution π, find a formula (in terms of π and $\mathbf{P}$) for

$$P[X_n = j | X_{n+1} = i].$$

Using your answer to the above and the English language transition probabilities in the file `trprob.mat`, what letter is most likely to be the last in a word?

. . . . . HW6B

For Further Study

If you want to explore the data from the Brown corpus on your own, you can obtain it at `http://ota.ahds.ac.uk/desc/0668` . The Brown corpus is not the only option; for some others see `http://clu.uni.no/icame/manuals/`. Google has been generating massive amounts of data of this type in conjunction with its Google Books project; see `https://books.google.com/ngrams/` – you can download their data from the link near the bottom of that page. Our equilibrium distribution provides *first order*

statistics for the English language, and the transition probabilities provide *second order* statistics. Some third order statistics are available at `http://www.data-compression.com/english.shtml`. There is also some interesting information at `http://norvig.com/ngrams/`.

Chapter 7

Entropy and Information

The variance of a random variable X is an indicator of how spread out the distribution of X is. A small variance means X is concentrated near its mean with high probability. A large variance means that the likely values of X are spread over a large range. So we might say that the variance measures how random X is, in some sense.

This chapter introduces the concept of entropy, which provides a different way to measure the "amount of randomness" in X, one that has no regard for the actual values of X. Suppose for instance that X is either 0 or 1, each with probability 1/2 (i.e. a Bernoulli random variable). The variances of X and $10X$ are different (1/4 and 25). But in the sense we consider here they have the same amount of randomness, because they each have two equally likely outcomes. Think of X as revealing to us some information about the the underlying state of the world, those otherwise unseen things which determine the value of X. Before we observe X we are uncertain about what its outcome will be. When we observe X that uncertainty is resolved and we know a little more about the state of the world.

For another example, suppose X is the sum of an independent pair of fair dice, a random variable with possible values in $\{2, 3, \ldots, 12\}$. Let Y be a random variable which takes each of the values in $\{1, 2, \ldots, 10\}$ with equal probability. X has more possible outcomes than Y but some are more likely than others. Which of these two random variables conveys more information when observed? In other words which of them has greater entropy? After we define entropy in the next section we will be able to answer this question with a simple calculation.

After proving the basic theoretical properties of entropy the rest of the chapter develops some simple versions of results from information theory to illustrate how entropy plays an important role in the analysis of coding and transmission of textual information.

7.1 Definition and Properties of Entropy

Definition. *If X is a random variable with only a finite number of different possible outcomes, $p_i = P(X = x_i) > 0$ for $i = 1, \ldots n$, we define its* entropy *to be*

$$H(X) = \sum_{i=1}^{n} -p_i \log_2(p_i). \tag{7.1}$$

By "*different* possible outcomes" we mean that the x_i are distinct. Notice that (unlike $\mathrm{Var}[X]$) this definition does not care what the values x_i of X actually are, only how many different outcomes are possible and what the set of probabilities p_i is. Thus X and $10X$ will always have the same entropy. Since $\lim_{x\to 0} x \log_2(x) = 0$, we will consider 0 to be the value of $x \log_2(x)$ at 0. This makes $x \log_2(x)$ a continuous function on $[0, \infty)$ and allows us to include $P(X = x_i) = p_i = 0$ in the sum without changing its value.

The change of base formula connects the logarithm base b to the natural logarithm: $\log_b(x) = \ln(x)/\ln(b)$. Changing the base of the logarithm in the definition would simply multiply the entropy by a constant, i.e. making a different choice of units. The choice of base 2 means that a Bernoulli random variable X taking the values 0 or 1 with equal probability has $H(X) = 1$. Think of X as a revealing a single binary digit or

"bit" of information. Thus our choice of base 2 means that X has entropy 1 (in bits). A constant random variable Y, with $P(Y = c) = 1$ for some c, has $H(Y) = 0$. Since there is nothing to be learned by observing Y (we know its value without looking) $H(Y) = 0$ is consistent with the interpretation that Y conveys no information.

Now we can compare the entropies of the X and Y that we contemplated above: X is the sum of an independent pair of (fair) dice and Y takes the values $\{1, \ldots, 10\}$ with equal probability. It is a simple calculation to find that

$$H(X) = 3.2744, \quad H(Y) = 3.3219.$$

Thus Y is (slightly) more random than X, in the sense of entropy.

The usefulness of entropy as a measure of randomness depends on its mathematical properties, which are established in the following theorem.

Theorem 7.1. *Suppose X and Y are both random variables, each taking only a finite number of different values.*

a) $0 \le H(X)$ *with equality if and only if X is constant.*

b) If X takes n distinct values, $H(X) \le \log_2(n)$ with equality if and only if the possible values of X are equally likely: $P(X = x_i) = 1/n$ for all n distinct values x_i.

c) $H((X, Y)) \le H(X) + H(Y)$, with equality if and only if X and Y are independent.

d) $H((X, Y)) = H(X) + H(Y \,\|\, X)$, where

$$H(Y \,\|\, X) = \sum_i p_i \sum_j -p_{j|i} \log_2(p_{j|i}).$$

(Here $p_i = P(X = x_i)$, $p_{j|i} = P(Y = y_j \,|\, X = x_i)$.)

e) $H(Y \,\|\, X) \le H(Y)$, with equality if and only if X and Y are independent.

Part b) says that the highest entropy occurs when the possible outcomes are equally likely. It also suggests that if X has infinitely many outcomes the entropy (if we extend the idea using infinite series) could be infinite. In d) and c) (X, Y) refers to the random pair. Part c) says that if X and Y are independent then the entropy associated with observing both of them is just the sum of their individual entropies. That makes sense because observing X first gives us no help in knowing what Y will be. But if they are not independent, then observing X does tell us something about Y so that we learn less from observing Y after X than we would by observing Y by itself first. The quantity $H(Y \,\|\, X)$ in d) and e) is called the *conditional entropy* of Y given X. (The usual notation is "$H(Y \,|\, X)$". But we are using the double vertical bar instead to avoid confusion with conditional expectation.) It measures the amount of *additional* information conveyed by observing Y after observing X first, *averaged* over the possible X values.

Proof. The fact that $-x \log_2(x) \ge 0$ for $0 \le x \le 1$ explains why $0 \le H(X)$. Suppose $H(X) = 0$ then $-p_i \log_2(p_i) = 0$ for all i. But this only happens if each p_i is either 0 or 1. Since $\sum p_i = 1$ it must be that exactly one $p_i = 1$ and all the others are 0. That means there is one value x for which $P(X = x) = 1$. This proves a).

For b), observe that $\phi(x) = x \log_2(x)$ is strictly convex, because $\frac{d^2}{dx^2}\phi(x) = (x \ln(2))^{-1} > 0$ for $x > 0$. We can apply Jensen's Inequality (see A.1 in the Appendix) as follows

$$\sum_1^n \frac{1}{n}[p_i \log_2(p_i)] = \sum_1^n \frac{1}{n}\phi(p_i) \ge \phi(\sum_1^n \frac{1}{n}p_i) = (\sum_1^n \frac{1}{n}p_i) \log_2(\sum_1^n \frac{1}{n}p_i) = \frac{1}{n}\log_2(\frac{1}{n}).$$

This is equivalent to $H(X) \le \log_2(n)$. Since $\phi(x)$ is strictly convex we can only have equality if all p_i are equal, which means that $p_i = 1/n$ for each i.

Part c) follows from d) and e).

For d), $P((X,Y)=(x_i,y_j)) = p_i p_{j|i}$. So

$$\begin{aligned} H((X,Y)) &= \sum_i \sum_j -p_i p_{j|i} \log_2(p_i p_{j|i}) \\ &= \sum_i \sum_j -p_i p_{j|i} \log_2(p_i) + \sum_i \sum_j -p_i p_{j|i} \log_2(p_{j|i}) \\ &= \sum_i -p_i \log_2(p_i) + \sum_i p_i \sum_j -p_{j|i} \log_2(p_{j|i}) \\ &= H(X) + H(Y \,\|\, X). \end{aligned}$$

Part e) is Jensen's Inequality for $\phi(x) = x\log_2(x)$ again: for each j we have

$$(\sum_i p_i p_{j|i}) \log_2(\sum_i p_i p_{j|i}) = \phi(\sum_i p_i p_{j|i}) \le \sum p_i \phi(p_{j|i}) = \sum_i p_i [p_{j|i} \log_2(p_{j|i})].$$

Using $\sum_i p_i p_{j|i} = q_j = P(Y = y_j)$ and summing the above over j we get

$$\sum_j q_j \log_2(q_j) \le \sum_i p_i \sum_j p_{j|i} \log_2(p_{j|i}),$$

which is equivalent to $H(Y \,\|\, X) \le H(Y)$. The two sides are equal if and only if we have equality in each application of Jensen's Inequality, which means that for each j, $p_{j|i}$ is the same for all values of i: $c_j = p_{j|i}$ for some set of values c_j. But multiplying by p_i and summing over i we see this is only possible for $c_j = q_j$. Thus equality holds in e) if and only if for all i, j we have

$$P(X = x_i, Y = y_j) = p_i p_{j|i} = p_i q_j,$$

which means that X and Y are independent. □

There is a converse to this theorem, which says that our definition (7.1) is the *only* way to define H with these properties, and so that $H(X) = 1$ for a Bernoulli random variable. See Khinchin [35] for the proof.

7.2 Entropy of a Markov Source

Information theory studies systems which transmit information from one person or place to another. Suppose that someone provides a source text, which is a sequence of letters or symbols from a source alphabet $\mathcal{A}$. (For English we will take the source alphabet to be our $\mathcal{E}$={A, . . . , Z, ␣}.) Then by means of some translation and transmission mechanism that source text is sent to someone else. How we do the translation and transmission should probably depend on properties of the source texts we expect to encounter, particularly if we hope to do it efficiently. To explore the significance of entropy we will assume that the source text is produced by a stationary Markov chain X_n with a state space denoted $\mathcal{A}$ (for "alphabet"). This is a rather crude description of a natural language, as we saw in Chapter 6. (See Problem 6.3.) The advanced mathematical study of information theory considers more general types of stochastic processes as the source. But a Markov source will give us enough mathematical structure to develop the ideas that we want to exhibit. (Note that independent and identically distributed X_n is a special case of a Markov chain.)

Notation

Some terminology and notation will help us talk about sequences and parts of sequences using symbols from an alphabet $\mathcal{A}$.

- An individual element of an alphabet $a \in \mathcal{A}$ will be called a *character*.
- The set of all infinite sequences in $\mathcal{A}$ is denoted $\mathcal{A}^{\mathbb{N}}$, and we will denote such a sequence by $a_{1:\infty}$. We will try to reserve the term "sequence" for infinite sequences.

- We will use *segment* to refer to a finite sequence. So $\mathcal{A}^n$ is the set of all segments with exactly n characters, i.e. n-segments (often called "n-grams" in the literature). We will denote n-segments by $a_{1:n}$ or sometimes by single Greek letters α, β, ...
- The collection of all segments (any finite length) will be denoted $\mathcal{A}^*$. So

$$\mathcal{A}^* = \cup_{n=1}^{\infty} \mathcal{A}^n.$$

  Such a segment will be denoted $a_{1:*}$, which means that it is an $a_{1:n}$ for some n. By writing $a_{1:*}$ we are leaving the length unspecified.
- Sequences or segments can be formed by concatenating shorter segments. For instance if $\alpha = (A, B, C)$, $\beta = (D, E, F)$, $\gamma = (G, H, I)$ are each 3-segments then by $\alpha\beta\gamma$ we simply mean the 9-segment

$$\alpha\beta\gamma = (A, B, C, D, E, F, G, H, I).$$

We view the Markov chain as our source text generator. In our notation, each realization of the Markov chain produces one sequence $a_{1:\infty} = X_{1:\infty}$. As each successive character is produced by the chain we learn a little more about the text as a whole. We can't calculate the entropy of the full random sequence $X_{1:\infty}$ because there are infinitely many possible outcomes. But we *can* calculate the entropy of the initial n-segment $X_{1:n}$. The increase in entropy as we go from n-segments to $(n+1)$-segments will be given by a conditional entropy. It follows from Theorem 7.1 and the Markov property that

$$\begin{aligned} H(X_{1:n}) &= H(X_{1:n-1}) + H(X_n \,\|\, X_{n-1}) \\ &\vdots \\ &= H(X_1) + (n-1)H(X_n \,\|\, X_{n-1}), \end{aligned}$$

since $H(X_n \,\|\, X_{n-1})$ is the same for all values of n. (That's because we are assuming the chain is stationary.) In fact it gives the *entropy per term* of the chain:

$$H_\Delta = \lim_{n\to\infty} \frac{1}{n} H(X_{1:n}) = H(X_1 \,\|\, X_0) = \sum_i \pi_i \sum_j -p_{i,j} \log_2(p_{i,j}). \tag{7.2}$$

We interpret H_Δ as the (average) amount of information *per term* produced by the chain. We will see in the next section that this plays an essential role in characterizing the performance requirements of transmission mechanisms.

Shannon-Breiman-McMillan Theorem

The entropy per term of a Markov chain arises in another somewhat unexpected (but important) way. Let $X_{0:n-1}$ be the initial n-segment produced by the chain. If $a_{0:n-1}$ is the initial n-segment of a sequence $a_{0:\infty}$, then the probability that $X_{0:n-1} = a_{0:n-1}$ is

$$p^{(n)}(a_{0:\infty}) = \pi_{a_0} p_{a_0,a_1} p_{a_1,a_2} \cdots p_{a_{n-2},a_{n-1}}.$$

We will consider this as a function $p^{(n)} : \mathcal{A}^{\mathbb{N}} \to [0, 1]$. In other words $p^{(n)}(a_{0:\infty})$ gives the probability that $X_i = a_i$ for $i = 0, \ldots n-1$., i.e. the probability that the chain matches $a_{0:\infty}$ through the first n outcomes starting with X_0.

Theorem 7.2 (Shannon-Breiman-McMillan)**.** *Under the assumptions described above,*

$$\frac{1}{n} \log_2 p^{(n)}(X_{0:\infty}) \to -H_\Delta \text{ almost surely.}$$

For any infinite sequence $a_{0:\infty}$ we expect $p^{(n)}(a_{0:\infty}) \to 0$ as $n \to \infty$. What the above says is that the sequences $a_{0:\infty} = X_{0:\infty}$ that are actually produced by the Markov chain are virtually certain to be ones for which (in a rough sense)

$$p^{(n)}(X_{0:\infty}) \sim 2^{-nH_\Delta} \text{ for large } n.$$

Think of $p^{(n)}(X_{0:\infty})$ this way: first we observe $X_{0:n-1}$ and then we ask "what was the probability of seeing what we just saw?" The theorem says that for asymptotically large n the answer is approximately 2^{-nH_Δ}, with vanishing relative error. The proof below shows that this is in fact a consequence of the strong law for the Markov chain.

Proof. We can write

$$p^{(n)}(X_{0:\infty}) = \pi_{X_0} \prod_{(i,j)} p_{(i,j)}^{N_{(i,j)}(n-1)},$$

where $N_{(i,j)}(n-1)$ counts the number of $i \to j$ transitions that the chain has made by time $n-1$. Let

$$F_{(i,j)}(n-1) = \frac{1}{n-1} N_{(i,j)}(n-1),$$

what we would call the *empirical frequency* of (i,j) transitions. Then we have

$$\frac{1}{n}\log_2(p^{(n)}(X_{0:\infty})) = \frac{1}{n}\log_2(\pi_{X_0}) + \frac{n-1}{n}\sum_{(i,j)} F_{(i,j)}(n-1)\log_2(p_{(i,j)}).$$

Problem 2.13 showed that as a consequence of the strong law $F_{(i,j)}(n-1) \to \pi_i p_{i,j}$ almost surely, as $n \to \infty$. So almost surely we have

$$\frac{1}{n}\log_2(p^{(n)}(X_{0:\infty})) \to \sum_{(i,j)} \pi_i p_{i,j} \log_2(p_{(i,j)}) = -H_\Delta.$$

□

Theorem 7.2 only considers random text produced by a stationary Markov chain. In general the entropy of a stochastic source (not necessarily Markov) is defined as

$$H_\Delta = \lim_{n\to\infty} \frac{1}{n} H(X_{0:n-1}).$$

Our theorem establishes the existence of this limit for stationary Markov chains, but it can be shown to exist much more generally. (For English language estimates indicate that $H_\Delta \le 1.75$, which is significantly lower than the 3.36477 which results from our Markov chain model; see Problem 7.3.) The proof of the Shannon-Breimann-McMillan Theorem in that greater generality was a major achievement back in 1953; see [35].

7.3 Coding

Suppose that our source text is written in conventional English prose. English is often predictable enough that we don't need to transmit much to communicate efficiently. For instance if I sent my wife the text "i lv u" I think she would get the message, and I would have sent only 6 characters instead of 10 as the correct spelling would require. The same possibility arises for our Markov source X_n using a source alphabet with $\#\mathcal{A} = A$ characters. The largest possible entropy per term H_Δ is $\log_2 A$. But for all but the simplest chains $H_\Delta < \log_2 A$. The fact that $H_\Delta < \log_2 A$ means that there is typically less information in a source segment than a segment of the same length in $\mathcal{A}^*$ can convey. We ought be able to express the source text in some more efficient form than its original sequence in $\mathcal{A}$, so that we don't waste resources when we transmit it. That's what we do naturally when we resort to abbreviations like "i lv u". It may also be that the transmission mechanism sends signals of a very different form than our source text: modulated radio frequency waves, or perhaps just a binary sequence. So our source text must be translated or "coded" into the form that can be physically transmitted, but in a way that allows it to be decoded after reception.

In this section we want to see how entropy provides a theoretical limit on how efficiently we can abbreviate or code the source text so that it can be decoded after transmission. We will assume that the transmission mechanism works by sending a sequence of symbols taken from a *code alphabet* $\mathcal{B}$ which may be different

from the source alphabet $\mathcal{A}$. For instance $\mathcal{B} = \{0, 1\}$ is a natural choice for digital communication. Our task is to translate the source text (a_i) into a sequence (b_j) in the code alphabet for transmission, and to do so in an efficient way. Throughout we will use $A = \#\mathcal{A}$ and $B = \#\mathcal{B}$ for the numbers of characters in each of these alphabets.

7.3.1 Examples

First we consider two examples for coding the English alphabet $\mathcal{A} = \mathcal{E}$ into the binary alphabet $\mathcal{B} = \{0, 1\}$.

A Fixed Length Code for $\mathcal{E}$

Suppose we number the characters in $\mathcal{E}$ from 1 for A to 27 for ␣, and convert the number of each character to binary form using 5 binary digits.

$$\chi(A) = 00001,\ \chi(B) = 00010,\ \chi(C) = 00011,\ \chi(D) = 00100,\ \chi(E) = 00101, \ldots,\ \chi(\text{␣}) = 11011.$$

This is what we will call a $(1, 5)$ code; every individual English letter gets converted to 5 binary digits. For example the source text "YES SIR" would get coded as the binary segment

$$11001\,00101\,10011\,11011\,10011\,01001\,10010$$

(We have typed this with spaces to help you see the 5-bit blocks, but the spaces are not part of the coded text.)

Morse Code for $\mathcal{E}$

Morse code was developed starting in the 1840s as a system for communicating text through a medium that can be switched back and forth between just two states, "on" or "off". Think of a single telegraph wire that can be connected or disconnected from a electrical voltage, or a signal light that can be either on or off. By turning the medium on and off according to prescribed patterns, messages can be sent "through the wire". Each letter is represented using some combination of "dots" $(\cdot)$ and "dashes" $(-)$; see the table below. (There are also codes for numerals, various punctuation marks and other symbols, but we will limit our discussion to the uppercase letters and the space-between-words character ␣ of our alphabet $\mathcal{E}$.) A dot is transmitted as "on" for one unit of time; a dash is transmitted as "on" for three units of time. The dots and dashes comprising a single letter are separated by "off" for one of time. The letters of the same word are separated by "off" for three units of time. The space-between-words is "off" for seven units of time. We can view this as sending a sequence of 0's and 1's, 0 being "off" and 1 being"on", each for a single unit of time. For instance our sample message "YES SIR" would be translated to

$$-\cdot--|\cdot|\cdots\text{␣}\cdots|\cdot\cdot|\cdot-\cdot$$

where | is the space-between-letters and ␣ is the space between words. This is not quite a code in the binary alphabet $\mathcal{B} = \{0, 1\}$, but there is a natural way to view it in that from. Let each dot becomes 1 followed by a 0 (within-the-letter-space); each dash becomes 111 followed by a 0 (within-the-letter-space). By *always* using the within-the-letter-space after 1 or 111, then the space-between-letters | becomes only 00. And likewise with the last letter of a word including 000 (within-the-letter-space and between-the-letter-space) the space-between-words ␣ becomes only 0000. Following this scheme Y=1110101110111000, E=1000, S=10101000, ␣=0000, and so forth. So our "YES SIR" becomes

$$1110101110111000\ 1000\ 10101000\ 0000\ 10101000\ 101000\ 1011101000.$$

(Again we have printed it with some spaces to help you see the breaks between individual character codes.)

In this form the same message results in a longer code segment than using the preceeding code. But bear in mind that Morse code was designed to be used with human beings as coders/decoders; there were no digital electronic systems then. An actual telegraph operator transmits 111 $(-)$ by just holding the key down for 3 time units, and 1 $(\cdot)$ by holding the key down for one time unit. Strings of 0s are just leaving the key up for various amounts of time. A trained telegraph operator does not think in terms of 0s and 1s.

He has learned to recognize the audible patterns for each letter without needing to think about it[1]. But by writing it this way each 0 and 1 uses exactly the same amount of time, and is convenient for our analysis below. Each letter requires a certain number of units of time to transmit. For instance Y requires 16 time units; S requires 8. These time requirements are included in the table below.

A	B	C	D	E	F	G	H	I
·−	−···	−·−·	−··	·	··−·	−−·	····	··
8	12	14	10	4	12	12	10	6
J	K	L	M	N	O	P	Q	R
·−−−	−·−	·−··	−−	−·	−−−	·−−·	−−·−	·−·
16	12	12	10	8	14	14	16	10
S	T	U	V	W	X	Y	Z	␣
···	−	··−	···−	·−−	−··−	−·−−	−−··	
8	6	10	12	12	14	16	14	4

This is what we call a variable length code, specifically a $(1, *)$ code. Each individual English letter is converted to some segment of 0s and 1s, but some letters get longer code segments than others. In fact if you compare the code lengths (rows 3, 6, 9 of the table) to the probabilities from Chapter 6 you will see that the order of the letters from shortest to longest Morse codes is quite close to the order from highest to lowest probability. The mean codeword length for Morse code is $\bar{\ell} = 8.16735$.

In the above examples each source character $a \in \mathcal{A}$ is coded as a segment $(b_1, \ldots b_\ell) \in \mathcal{B}^*$. For the fixed length example $\ell = 5$ for all letters but for Morse code the length of the segment depends on the source letter: $\ell = \ell(a)$. More generally, instead of coding each source character separately we can break our source into *blocks* or source segments of some length n (n-segments) and code each such block into a specific code segment. If we code blocks of length n into code segments of a fixed length m we will say we have a (n, m) code. If the code segments are of variable length we will say we have a $(n, *)$ code. So Morse code is a $(1, *)$ code and the 5-bit example is a $(1, 5)$ code. (One could also consider using source blocks of variable size, or even more complicated schemes that don't decompose into segments that are coded separately at all, but we won't pursue those possibilities.)

We will use χ refer to a code. Ulitmately the code is a function $\chi : \mathcal{A}^{\mathbb{N}} \to \mathcal{B}^{\mathbb{N}}$, but it works by mapping $\mathcal{A}^n \to \mathcal{B}^m$ for an (n, m) code, or $\mathcal{A}^n \to \mathcal{B}^*$ for a $(n, *)$ code, and concatenating the results:

$$\chi(a_1, a_2, \ldots) = \chi(a_{1\cdots n})\chi(a_{n+1\cdots 2n}) \cdots .$$

It may be that certain source segments $a_{1:n}$ never actually occur, so that no code $\chi(a_{1:n})$ needs to be assigned to them. So in general we will view our code as $\chi : \mathcal{S} \to \mathcal{B}^*$ for some $\mathcal{S} \subseteq \mathcal{A}^n$, as long as $\mathcal{S}$ includes all source n-segments which occur with positive probability.

7.3.2 Theoretical Bounds

We now want to look at simple versions of Shannon's coding theorems, which are centerpieces of information theory. These results give some theoretical bounds which relate the entropy H_Δ of our Markov source to features of any code χ which can be successfully decoded to recover the original source text.

For a fixed length (n, m) to be decodable simply means that χ maps each n-segment in $\mathcal{S}$ to a distinct code segment in $\mathcal{B}^m$. I.e. $\chi : \mathcal{S} \to \mathcal{B}^m$ should be one-to-one. This requires $B^m \geq \#\mathcal{S}$. If all source n-segments are possible ($\mathcal{S} = \mathcal{A}^n$), then we must have $B^m \geq A^n$, which we can write as $\frac{m}{n} \log_2 B \geq \log_2 A$. We also know from Theorem 7.1 that $\log_2 A \geq H_\Delta$. So we have

$$\frac{m}{n} \log_2 B \geq H_\Delta. \tag{7.3}$$

This inequality depends on the assumptions that all n-segments in $\mathcal{A}^n$ are coded (i.e. are in $\mathcal{S}$), and that χ is 1-to-1.

[1]Listen to it at `http://morsecode.scphillips.com/translator.html`.

Variable length codes achieve efficiency by using shorter code segments for more likely source segments. But it is not enough for χ to be one-to-one on $\mathcal{S}$ if we want to be able to decode messages that consist of more than one $\mathcal{S}$-segment put together. The following example illustrates.

Example 7.1. Suppose we number the characters in $\mathcal{E}$ from 1 for A to 27 for $\_$, and convert the number of each character to binary form with*out* leading 0s.

$$\chi(A) = 1,\ \chi(B) = 10,\ \chi(C) = 11,\ \chi(D) = 100,\ \chi(E) = 101, \ldots,\ \chi(\_) = 11011.$$

This is one-to-one on individual letters. But observe that all of the segments BC, EA, BAA code to 1011. So it is not one-to-one on segments. The problem is that in the coded text we can't tell where the code for one $\mathcal{S}$-segment ends and the next one begins.

Let $\mathcal{S} \subseteq \mathcal{A}^n$. We will say that an $(n, *)$ code $\chi : \mathcal{S} \to \mathcal{B}^*$ is *decodable* if $P(X_{1:n} \in \mathcal{S}) = 1$ and χ is 1-to-1 on $\mathcal{S}^* = \cup_{k=1}^{\infty} \mathcal{S}^k$, the collection of all finite messaages we can make up from multiple $\mathcal{S}$-segments. This means that if two segments (of possibly different lengths but both formed from $\mathcal{S}$) have $\chi(a^*) = \chi(b^*)$, then $a^* = b^*$. Theorem 7.5 is going to give us a generalization of (7.3) for decodable variable length codes. In preparation for it we need a couple lemmas.

Lemma 7.3. *Suppose a code $\chi : \mathcal{S} \to \mathcal{B}^*$ maps each $\alpha \in \mathcal{S}$ into a code segment of length $\ell(\alpha)$. If χ is decodable then*

$$\sum_{\alpha \in \mathcal{S}} B^{-\ell(\alpha)} \le 1.$$

Note that there are no probabilities involved here.

Proof. Let L be the largest value of $\ell(\alpha)$. Start by observing that

$$\left(\sum_{\alpha \in \mathcal{S}} B^{-\ell(\alpha)} \right)^k = \sum_{j=1}^{Lk} c_j B^{-j},$$

where c_j counts the number of nk-segments in $\mathcal{S}^k$ with coded length $= j$. The decodable property requires that $c_j \le B^j$, and therefore

$$\left(\sum_{\alpha \in \mathcal{S}} B^{-\ell(\alpha)} \right)^k \le Lk.$$

Taking the kth root,

$$\sum_{\alpha \in \mathcal{S}} B^{-\ell(\alpha)} \le (Lk)^{1/k}.$$

Taking the limit as $k \to \infty$ we find that $(Lk)^{1/k} \to 1$, and therefore

$$\sum_{\alpha \in \mathcal{S}} B^{-\ell(\alpha)} \le 1.$$

□

The inequality of this lemma is called the *Kraft inequality* in the literature. The lemma has a converse.

Lemma 7.4. *If $\ell : \mathcal{S} \to \mathbb{N}$ with $\sum_{\alpha \in \mathcal{S}} B^{-\ell(\alpha)} \le 1$ then there exists a decodable code $\chi : \mathcal{S} \to \mathcal{B}^*$ with code segment lengths $\ell(\alpha)$ for $\alpha \in \mathcal{S}$.*

Proof. We are given code lengths $\ell(\alpha)$ satisfying the Kraft inequality and need to show how a decodable code χ can be built with the prescribed code lengths.

Let c_j be the number of $\sigma \in \mathcal{S}$ with $\ell(\sigma) = j$. We have to have at least c_1 distinct characters in $\mathcal{B}$ to accommodate the assignment of those σ with $\ell(\sigma) = 1$. So we need to have

$$c_1 \le B.$$

If that is true then we take a subset $F_1 \subseteq \mathcal{B}$ with exactly c_1 elements and assign one of them as $\chi(\alpha)$ for each α with $\ell(\alpha) = 1$.

That leaves $B - c_1$ characters in $\mathcal{B}$ which will be used as the first character of $\chi(\alpha)$ when $\ell(\alpha) \geq 2$. In particular there are $(B - c_1)B = B^2 - c_1 B$ length 2 codes available for the c_2 segments of $\mathcal{S}$ requiring length 2 codes. We can make c_2 such assignments provided

$$c_2 \leq B^2 - c_1 B.$$

Making these assignments determines a set $F_2 \subseteq \mathcal{B}^2$ of 2-character codes

$$F_2 = \{\chi(\alpha) : \alpha \in \mathcal{S} \text{ and } \ell(\alpha) = 2\},$$

none of which use a first character from F_1.

That leaves $B^2 - c_1 B - c_2$ unassigned length 2 codes with first character not in F_1 and first two characters not in F_2. These we can use as the first two characters for codes of length 3 or more. There are a total of $(B^2 - c_1 B - c_2)B = B^3 - c_1 B^2 - c_2 B$ length 3 codes of this type which are available to use as $\chi(\alpha)$ for those α with $\ell(\alpha) = 3$. We need assign c_3 of these so we require

$$c_3 \leq B^3 - c_1 B^2 - c_2 B.$$

Making these assignments determines a set $F_3 \subseteq \mathcal{B}^3$ of 3-character codes

$$F_3 = \{\chi(\alpha) : \alpha \in \mathcal{S} \text{ and } \ell(\alpha) = 3\},$$

none of which use a first character from F_1 or an initial pair from F_2.

Provided that is satisfied we assign a set F_3 of exactly c_3 of these available length 3 codes as the codes $\chi(\alpha)$ for those α with $\ell(\alpha) = 3$.

Continuing in this way we see that we can make all the desired assignments provided we can show that the equalities

$$\begin{aligned} c_1 &\leq B \\ c_2 &\leq B^2 - c_1 B \\ &\vdots \\ c_L &\leq B^L - c_1 B^{L-1} - \cdots - c_{L-1} B, \end{aligned}$$

all hold. These all follow from the Kraft inequality $\sum_{\alpha \in \mathcal{S}} B^{-\ell(\alpha)} \leq 1$.

The resulting code χ is decodable using the following algorithm. Given a coded a coded segment $b_1 b_2 b_3 \ldots$ first check whether $b_1 \in F_1$. If so the first segment of the source text is α with $\chi(\alpha) = b_1$. But if not then check whether $b_1 b_2 \in F_2$. If so the first segment of the source text is α with $\chi(\alpha) = b_1 b_2$. But if not then check whether $b_1 b_2 b_3 \in F_3$ ... In this way we will eventually identify the first segment α of the source text and which portion of the coded text is $\chi(\alpha) = b_1 \cdots b_{\ell(\alpha)}$. We now remove that from the coded text and repeat the process to find the next segment of the source text, and so on. □

These two lemmas tell tell us when a given assignment of code lengths $\ell(\alpha)$ for $\alpha \in \mathcal{S}$ can be achieved with a decodable $(n, *)$ code but have nothing to do with the entropy of the Markov source, other than that $P(X_{0:n} \in \mathcal{S}) = 1$. To achieve an efficient code we would want to choose the code lengths so that more likely source texts have shorter code lengths. The next result gives a lower bound on the mean code length per source n-segment that a decodable code may have in terms of the entropy per term.

Theorem 7.5. *For any $(n, *)$ code $\chi : \mathcal{S} \to \mathcal{B}^*$ which is decodable we must have*

$$\frac{1}{n} \bar{\ell} \log_2 B \geq H_\Delta, \tag{7.4}$$

where $\bar{\ell} = E[\ell(X_{0:n-1})]$ is the mean code length per source n-segment.

Observe that this generalizes (7.3) because for a fixed length (n, m) code $\bar{\ell} = m$.

Proof. From the lemma above, $\sum_{\alpha\in\mathcal{S}} B^{-\ell(\alpha)} \le 1$. Therefore

$$\log_2\left(\sum_{\alpha\in\mathcal{S}} B^{-\ell(\alpha)}\right) \le 0.$$

We now apply Jensen's inequality using the convex function $-\log_2(x)$ in the second line:

$$\begin{aligned}
0 &\le -\log_2\left(\sum_{\alpha\in\mathcal{S}} B^{-\ell(\alpha)}\right)\\
&= -\log_2\left(\sum_{\alpha\in\mathcal{S}} p(\alpha)\frac{1}{p(\alpha)}B^{-\ell(\alpha)}\right) \le \sum_{\alpha\in\mathcal{S}} -p(\alpha)\log_2\left(\frac{1}{p(\alpha)}B^{-\ell(\alpha)}\right)\\
&= \sum_{\alpha\in\mathcal{S}} -p(\alpha)\log_2(B^{-\ell(\alpha)}) + \sum_{\alpha\in\mathcal{S}} -p(\alpha)\log_2\left(\frac{1}{p(\alpha)}\right)\\
&= \bar{\ell}\log_2 B - H(X_{0:n-1}).
\end{aligned}$$

Therefore, using Theorem 7.1 e)

$$\bar{\ell}\log_2 B \ge H(X_{0:n-1}) = H(X_0) + (n-1)H_\Delta \ge nH_\Delta.$$

Dividing by n gives the inequality (7.4). □

7.3.3 Nearly Optimal Codes

Theorem 7.5 gives us a theoretical lower bound on $\bar{\ell}/n$ for any code which successfully codes a Markov source with entropy H_Δ. Now we will to see that the lower bound is "sharp", meaning that exist codes for which $\bar{\ell}/n$ is arbitrarily close to the lower bound in the theorem.

Fixed Length Codes

Suppose χ is a 1-1 (n,m) code. Then $\bar{\ell} = m$. If every n-segment in $\mathcal{A}$ has positive probability and χ is decodable we know that

$$\frac{m}{n}\log_2 B \ge \log_2 A.$$

We also know that $\log_2 A \ge H_\Delta$, with equality if and only if the X_n of the chain are i.i.d. with uniform distribution on $\mathcal{A}$. Otherwise $\log_2 A > H_\Delta$ so no decodable (n,m) code can have a value of $\frac{m}{n}\log_2 B$ which approaches the lower bound H_Δ in (7.4). We can't find fixed length codes which are nearly optimal in the sense of Theorem 7.9.

But we can pursue a different idea: allow the code to "fail" on a set of low-probability segments. We will choose a set $\mathcal{T} \subset \mathcal{A}^n$ of at most $B^m - 1$ source segments which will be coded "faithfully," i.e. in a 1-to-1 manner. There remains at least one segment in $\mathcal{B}^m$ which we will designate as the error code $\oslash$. All $\alpha \notin \mathcal{T}$ will be coded with the error code, $\chi(\alpha) = \oslash$. This means that some source sequences will produce code sequences which are not fully decodable. If that happens a coding error has occurred: $\chi(X_{0:n-1}) = \oslash$. We might consider $\frac{1}{n}\log_2(\#\mathcal{T})$ as a measure of the inefficiency of the code, the average amount of information that is *faithfully* transmitted per source character, and $P(\chi(X_{0:n-1}) = \oslash)$ a measure of the inaccuracy. For an efficient and accurate code we would like both of these to be small.

The next two results address the idea that that H_Δ is the lower limit of $\frac{1}{n}\log_2(\#\mathcal{T})$ for accurate codes. The first of them says that for any sufficiently large n we can design a code with $\frac{1}{n}\log_2(\#\mathcal{T})$ as close to H_Δ as we like and probability of error as small as we like.

Theorem 7.6. *Given any $\epsilon, \delta > 0$ there exists N so that for every $n \ge N$ there exists an (n,m) code χ as described above with $\frac{1}{n}\log_2(\#\mathcal{T}) < H_\Delta + \delta$ and $P(\chi(X_{0:n-1}) = \oslash) < \epsilon$.*

Proof. Let $\epsilon, \delta > 0$ be given. The Shannon-Breiman-McMillan Theorem says that there exists N so that for any $n \geq N$ we have

$$P\left(\frac{1}{n}\log_2 p^{(n)}(X_{0:\infty}) > -H_\Delta - \delta\right) > 1 - \epsilon.$$

In other words if $\mathcal{T}$ is the set of n-segments $a_{0:n-1}$ for which

$$p^{(n)}(a_{0:n-1}) > 2^{-n(H_\Delta+\delta)}$$

then

$$P(X_{0:n-1} \in \mathcal{T}) > 1 - \epsilon.$$

It follows that

$$2^{-n(H_\Delta+\delta)}\#\mathcal{T} \leq \sum_{\alpha\in\mathcal{T}} p(\alpha) \leq 1,$$

and so $\frac{1}{n}\log_2 \#\mathcal{T} \leq H_\Delta + \delta$. Now just choose m so that $\#\mathcal{T} < B^m - 1$. Then we can map the segments from $\mathcal{T} \subset \mathcal{A}^n$ faithfully into $\mathcal{B}^m$. And the probability of encountering a segment not in $\mathcal{T}$ is less than $1 - P(\mathcal{T}) < \epsilon$. Thus we have an (n, m) code with the desired properties. □

From the proof it appears that we may need to use a very large block size n to get $\frac{1}{n}\log_2 \#\mathcal{T}$ close to H_Δ. The next theorem says that if $\frac{1}{n}\log_2 \#\mathcal{T} \leq H_\Delta - \delta$ then for long source texts we are very likely to encounter coding errors. Even if the block size n is small, once the overall length of the source text nk is big enough the probability of encountering a coding error will be large. Using single blocks of very large size won't overcome this. In brief, H_Δ is the lowest that $\frac{1}{n}\log_2(\#\mathcal{T})$ can go for codes with low probability of coding errors on large blocks.

Theorem 7.7. *Given the Markov chain and values $\epsilon, \delta > 0$ there exists N with the following property. Suppose χ is any (n, m) code with*

$$\frac{1}{n}\log_2(\#\mathcal{T}) < H_\Delta - \delta$$

and $\mathcal{T}$ is the set of n-segments which are coded uniquely by χ. If $nk > N$ then

$$P(X_{0:nk-1} \notin \mathcal{T}^k) > 1 - \epsilon.$$

Proof. First observe that an (n, m) code can be considered as a (nk, mk) code by concatenating k blocks at a time. The set of nk-segments on which the concatenated code is one-to-one is simply $\mathcal{T}^k$, and

$$\frac{1}{nk}\log_2(\#\mathcal{T}^k) = \frac{1}{nk}\log_2((\#\mathcal{T})^k) = \frac{1}{n}\log_2(\#\mathcal{T}).$$

This means it is enough to prove the theorem just for $k = 1$.

The hypothesis that $\frac{1}{n}\log_2(\#\mathcal{T}) < H_\Delta - \delta$ implies

$$\#\mathcal{T} = 2^{\log_2(\#\mathcal{T})} < 2^{n(H_\Delta-\delta)}.$$

Let $\mathcal{V}_n$ be the set of those n-segments with $p^{(n)}(\alpha) < 2^{-n(H_\Delta-\delta/2)}$. Then

$$P(X_{0:n-1} \in \mathcal{T} \cap \mathcal{V}_n) \leq 2^{n(H_\Delta-\delta)}2^{-n(H_\Delta-\delta/2)} = 2^{-n\delta/2}.$$

Now we can write

$$\begin{aligned} P(X_{0:n-1} \in \mathcal{T}) &\leq P(X_{0:n-1} \notin \mathcal{V}_n) + P(X_{0:n-1} \in \mathcal{T} \cap \mathcal{V}_n) \\ &\leq 1 - P(X_{0:n-1} \in \mathcal{V}_n) + 2^{-n\delta/2}. \end{aligned}$$

The right side does not depend on the code χ, only on the chain and the values of n and δ. We know that $2^{-n\delta/2} \to 0$ as $n \to \infty$, and the Shannon-Breiman-McMillan implies that $P(X^{(n)} \in \mathcal{V}_n) \to 1$. So there is an N for which the right side is $< \epsilon$ for all $n > N$. The inequality applies to all (n, m) codes with $\frac{1}{n}\log_2(\#\mathcal{T}) < H_\Delta - \delta$. Thus if $n > N$ and $\frac{1}{n}\log_2(\#\mathcal{T}) < H_\Delta - \delta$ then we have a failure probability of at least $1 - \epsilon$. □

Variable Length Codes

Next we will see that variable length codes can get arbitrarily close to the theoretical limit of (7.4) without incurring errors. Consider first a finite set $\mathcal{S}$ of segments with probabilities $p(\sigma)$. We want to find a code $\chi : \mathcal{S} \to \mathcal{B}^*$ with for which the mean codeword length $\bar{\ell}$

$$\bar{\ell}(\chi) = \sum_{\sigma \in \mathcal{S}} p(\sigma)\ell(\sigma)$$

is close to the smallest possible. For the resulting code on $\mathcal{A}^*$ to be decodable requires that χ to be 1-to-1 on $\mathcal{S}$, but that alone is not sufficient, as Example 7.1 showed. A sufficient condition is that no codeword $\chi(\sigma)$ occurs as an initial segment of some other codeword: $\chi(\rho) \neq \chi(\sigma)01\cdots$. (Example 7.1 does not satisfy that. For instance $\chi(D) = 100 = \chi(B)0$.) Then as we scan the coded text from left to right, as soon as we recognize a codeword we can be sure that there is no other possible decoding to worry about; we can record the source segment for the codeword we just found and then resume scanning where we left off. A code with this property is called a *prefix-free code*. (Often this is misleadingly shortened to "prefix code".) The code constructed in the proof of Lemma 7.4 *was* prefix-free, which was why it worked.

We can think of a prefix-free code as a (directed) tree with at most B branches (one for each $b \in \mathcal{B}$) descending from each non end node. The ends of the tree (nodes with no branches emanating from them) are the elements of $\mathcal{S}$ corresponding to the coded sequence that leads from the top of the tree to that end node. For example see the picture at the end of this section for a binary ($B = 2$) code for $\mathcal{E}$. A code associated with such a tree is always a prefix-free code because if you reach an end node there are no more branches to follow, so no other source segment could lead you to that same end node and then past it to a different end node.

Given a finite set $\mathcal{S}$ with probabilities $p(\sigma),\ \alpha \in \mathcal{S}$ the next theorem guarantees the existence of a prefix-free code with a particular bound on its mean codeword length $\bar{\ell}$.

Lemma 7.8. *Let Y be a random variable taking values in a finite set F, and $\mathcal{B}$ any code alphabet. Let $B = \#\mathcal{B} > 1$. There exists a prefix-free code $\chi : F \to \mathcal{B}^*$ with*

$$\bar{\ell} \leq H(Y)/\log_2 B + 1,$$

where $\bar{\ell} = E[\ell(Y)]$.

Proof. Let

$$p(\sigma) = P(Y = \sigma) \text{ for } \sigma \in F.$$

Consider the values

$$\ell(\sigma) = \lceil -\log_B(p(\sigma)) \rceil,\ \sigma \in F.$$

(By $\lceil x \rceil$ we mean the ceiling function, i.e. the result of rounding x up to the next highest integer.) We are going to show that there does exist a prefix-free code using codewords with lengths $\ell(\sigma)$. Since $\ell(\sigma) \leq -\log_B(p(\sigma)) + 1$ such a code will have

$$\bar{\ell} = \sum_{\mathcal{S}} p(\sigma)\ell(\sigma) \leq \sum_{\mathcal{S}} p(\sigma)(-\log_B(p(\sigma)) + 1) = H(Y)/\log_2 B + 1,$$

which is the inequality claimed by the lemma.

To see that there is a prefix-free code with with codeword lengths $\ell(\sigma)$ we just need to check that the inequality of Lemma 7.4 holds. (Inspection of the proof of that lemma shows that it produces a prefix-free code.) Our definition of the $\ell(\sigma)$ values implies $B^{-\ell(\sigma)} \leq p(\sigma)$. Summing this, we know that

$$\sum_{\mathcal{S}} B^{-\ell(\sigma)} \leq \sum_{\mathcal{S}} p(\sigma) = 1.$$

□

If we use $F = \mathcal{A}^n$ and $Y = X_{0:n-1}$ in this lemma then $H(Y) = H(X_1) + (n-1)H_\Delta$. We know from Theorem 7.5 (actually the last line of the proof), that for any decodable code

$$H(X_{0:n-1}) \le \bar{\ell} \log_2 B$$

and the above result shows that there always exists a code for which

$$\bar{\ell} \log_2 B \le H(X_{0:n-1}) + \log_2 B.$$

Dividing by n, there exists a prefix-free $(n, *)$ code with

$$\frac{1}{n} H(X_{0:n-1}) \le \frac{1}{n} \bar{\ell} \log_2 B \le \frac{1}{n} H(X_{0:n-1}) + \frac{1}{n} \log_2 B.$$

Since $\frac{1}{n} H(X_{0:n-1}) \to H_\Delta$ and $\frac{1}{n} \log_2 B \to 0$ as $n \to \infty$, this says that by choosing a large n there exists an $(n, *)$ prefix-free code with $\frac{1}{n} \bar{\ell} \log_2 B$ arbitrarily close to the optimum in Theorem 7.5, proving the following.

Theorem 7.9. *Given $\epsilon > 0$ there exists a decodable $(n, *)$ code for which all source texts are decodable and*

$$\frac{1}{n} \bar{\ell} \log_2 B < H_\Delta + \epsilon.$$

Taken together, Theorems 7.5 and 7.9 show that the entropy per term H_Δ is the infimum of $\frac{1}{n} \bar{\ell} \log_2 B$ over all decodable $(n, *)$ codes. In this sense it delineates exactly how much efficiency is or is not possible for a given Markov source.

When $B = 2$ there is a nice iterative construction which will always produce a prefix-free code with the smallest possible $\bar{\ell}$ in Lemma 7.8, called the Huffman code. Rather than prove these various assertions about the Huffman code, we will simply illustrate the procedure with our English alphabet $F = \mathcal{E}$, and the probabilities we indicated in Chapter 6.

First we sort the alphabet according to the probabilities (smallest to largest):

$$\mathcal{E} = \{\text{Z, Q, J, X, K, V, B, Y, W, G, P, F, M, U, C, D, L, H, R, S, N, I, O, A, T, E, } \text{␣}\}.$$

Next pick the two characters with the smallest probabilities, Z and Q in our case. Now form a new alphabet $\mathcal{S}'$ in which Z and Q are replaced by a single new symbol with probability equal to to the sum of those we just combined: if we denote the new symbol by [ZQ] its probability will be .00165. This gives the new alphabet, which we again sort by probabilities:

$$\mathcal{E}' = \{\text{J, X, [ZQ], K, V, B, Y, W, G, P, F, M, U, C, D, L, H, R, S, N, I, O, A, T, E, } \text{␣}\}.$$

Now when we find an optimal code χ' for $\mathcal{S}'$ and use it to construct a a code χ for $\mathcal{S}$ by letting χ and χ' be the same for all but the new combined symbol, and appending a 0 or 1 to $\chi'([\text{ZQ}])$ in order to "split" the character [ZQ]: $\chi(\text{Q}) = \chi'([\text{ZQ}])1$, $\chi(\text{Z}) = \chi'([\text{ZQ}])0$. Now we repeat the procedure on $\mathcal{S}'$: combine J and X to get [JX] with probability .002924 and a new alphabet

$$\mathcal{E}'' = \{\text{[ZQ], [JX], K, V, B, Y, W, G, P, F, M, U, C, D, L, H, R, S, N, I, O, A, T, E, } \text{␣}\}.$$

We continue like this

$$\mathcal{E}''' = \{\text{[[ZQ][JX]], K, V, B, Y, W, G, P, F, M, U, C, D, L, H, R, S, N, I, O, A, T, E, } \text{␣}\}$$

$$\vdots$$

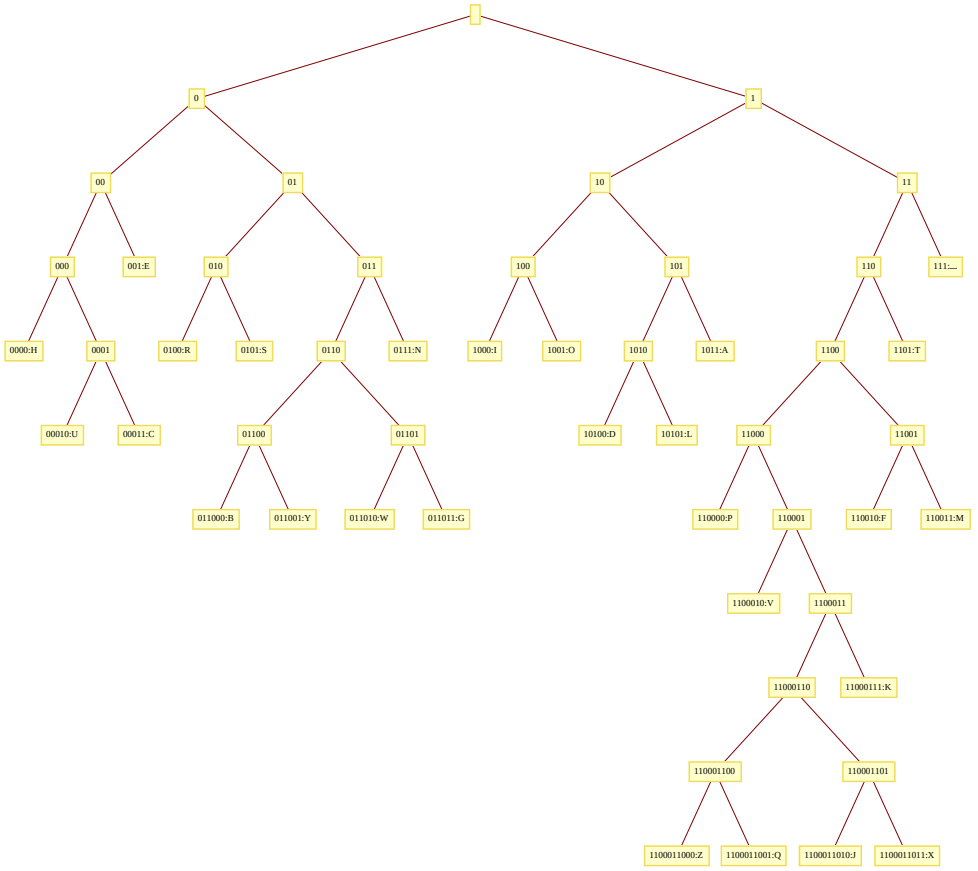
until we get down to a reduced alphabet of just two characters, for which an optimal code is obvious. Then we work backward from that, building the code back up until we reach $\mathcal{S}$. This procedure can be shown to always produce an optimal prefix-free code; see [8] for a proof. Below is the resulting Huffman code for our English alphabet $\mathcal{E}$, presented as a binary tree. As an example, our "YES SIR" codes as

011001001010111101011000010 0

The entropy for this alphabet and probability assignments is $H(\mathcal{E}) = 4.10644$, the entropy per term for the Markov chain is $H_\Delta = 3.36477$. The mean codeword length for the Huffman code is $\bar{\ell} = 4.13793$. Since $n = 1$ and $\log_2(B) = 1$ we knew that the best we could do was

$$H(\mathcal{E}) = 4.10644 \leq \bar{\ell} = 4.13793.$$

That's pretty good; compare this to the mean codeword length for Morse code, which works out to be 8.16735. To get closer to H_Δ we would need to use an $(n, *)$ code with a larger n.



Summary

In conclusion, we can view both $\frac{1}{n}\bar{\ell}(\chi)\log_2(B)$ and $\frac{1}{n}\log_2(\#\mathcal{T})$ as measures of the average amount of information (in bits per source term) that the code χ can transmit *in a decodable form* (or nearly decodable in the case of fixed length codes). For both the variable and fixed length codes we have found that the entropy per term of the source provides the theoretical lower bound on these measures for effective codes. The theory of information begun by C. Shannon in 1948, and continued by many others since, develops these ideas much more extensively.

Problem 7.1
Let X be a random variable and $Y = f(X)$. I.e. the value of Y is the result of applying a prescribed function f to the result X. Prove that $H(Y) \leq H(X)$. (f need not be one-to-one.)

. . . . . function

Problem 7.2
In the Monte Hall problem of Example 3.14, compare the entropy of $1_{C\neq 2}$ with H (assuming H is as described using the conditional probabilities γ and $1-\gamma$). Which has greater entropy?

. . . . . MHent

Problem 7.3
Using the Markov transition probabilities for English from Chapter 6 calculate the entropy per term for English text.

. . . . . MCEnglish

Problem 7.4
Let X_n be the Markov chain on $\mathcal{S} = \{a, b, c, d\}$ with transition matrix

$$P = \begin{pmatrix} \frac{13}{16} & \frac{1}{16} & \frac{1}{16} & \frac{1}{16} \\ 0 & 0 & 0 & 1 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ 1 & 0 & 0 & 0 \end{pmatrix},$$

started with the corresponding equilibrium distribution. Calculate the entropy H_Δ for this chain. With $B = \{0, 1\}$ find a Huffman $(1, *)$ code and its mean word length $\bar{\ell}$. Find a Huffman $(2, *)$ code (i.e. a Huffman code for $X_{0:1}$ on $\mathcal{S}^2$) and it's mean word length. Compare the value of H_Δ and the two values of $\bar{\ell}$ and discuss in light of the results of this chapter.

. . . . . MC4

Problem 7.5
What does Lemma 7.3 say if all $\ell(a) = 1$? What if all $\ell(a) = m$?

. . . . . One

For Further Study

Most of this chapter is based on Blahut [8] and Khinchin [35]. If you want to read more about coding and cryptography you might also consider [37] and [62].

Chapter 8

Optimization of Markov Chains

In this chapter we consider some optimization problems involving Markov chains.

8.1 Optimal Stopping

Suppose there is a hike you want to take, but you are waiting for a day with good weather. Each day when you get up you check the weather and decide whether to go or wait for a better day. Imagine that the weather from day to day is described by a Markov chain X_n with a state space $\mathcal{S}$ consisting of some set of possible weather descriptions: snowy, heavy rain, light rain, overcast, clear but cold, clear and moderate, hot, ... Your personal hiking preferences are described by a function ϕ which assigns a numerical value to each weather description: perhaps $\phi(\text{snowy}) = 0$ and $\phi(\text{clear and moderate}) = 10$ for instance. Given the transition probabilities for the chain and your weather preference function ϕ how should you pick the day on which you take your hike? To be specific, if $\mathcal{T}$ denotes the day on which you eventually take your hike, how can you choose $\mathcal{T}$ to make $E[\phi(X_{\mathcal{T}})]$ as large as possible? If there is no cost for waiting you could just wait for the perfect hiking conditions. But suppose you have already rented some equipment for the trip and it costs you additional rental fees for each extra day you rent it while waiting. Now it is not so simple.

For another example suppose the chain X_n describes the market price on day n of a share of stock in the XYZ Corporation. You own an (american) call option for one share of this stock with an exercise price of \$20 and an expiration date of $T = 365$. This option entitles you to buy one share of the stock for \$20, regardless of the market price, at any time $\mathcal{T} \leq T$. (It's like a sale coupon for one share of the stock, good until T.) If $X_n = 36$ and you decide to exercise your option at time $\mathcal{T} = n$ you will pay \$20 and receive the share of stock worth \$36, for a net gain of $36 - 20 = \$16$. But on a day when $X_n = 12$, why would you pay \$20 for something you could buy directly for \$12? Your option would be worthless to you in those circumstances. In general if you exercise your option at time n the benefit to you will be

$$\phi(X_n) = \max(0, X_n - 20).$$

Assuming you know the transition probabilities of the Markov chain, the problem is to choose $\mathcal{T} \leq T$ so as to maximize $E(\phi(X_{\mathcal{T}}))$.

Each of the above is an *optimal stopping problem*. The general form of this problem will involve a Markov chain X_n with state space $\mathcal{S}$ and transition matrix $\mathbf{P}$. There will be a specified *reward function* $\phi : \mathcal{S} \to \mathbb{R}$ and a *continuation cost function* $c : \mathcal{S} \to \mathbb{R}$. We observe X_n as time proceeds, deciding at each time n whether to stop now and receive reward $\phi(X_n)$ or pay $c(X_n)$ and continue, stopping at some later time. At time n our decision whether to stop or not can depend *only* on the values $X_0, \ldots, X_n$ observed so far, *not* on any future values X_k, $k > n$. The time at which we eventually stop will be denoted $\mathcal{T}$. Its value depends both on the strategy we are using to make our stop-or-continue decisions and the evolution of the chain. Thus $\mathcal{T}$ is a time-valued random variable with a special dependency property which we will describe below, what we call a stopping time. We may require that $\mathcal{T}$ be obey an upper bound, $\mathcal{T} \leq T$; these are called *finite time horizon* problems. If any $\mathcal{T} \leq \infty$ is allowed we call it an *infinite time horizon problem*. ($\mathcal{T} = \infty$ corresponds to forever putting off the decision to stop; you pay continuation costs indefinitely and never

receive the reward.) The goal is to formulate a strategy for selecting $\mathcal{T}$ to maximize the expected reward:

$$E\left[\phi(X_{\mathcal{T}})1_{\mathcal{T}<\infty} - \sum_{n=0}^{\mathcal{T}-1} c(X_n)\right]. \tag{8.1}$$

The $1_{\mathcal{T}<\infty}$ means that we only get the reward $\phi(X_{\mathcal{T}})$ if we do stop at a finite time; if $\mathcal{T} = \infty$ there is no reward, but we do pay the continuation costs. We will make the following assumptions.

- $0 \le \phi(\cdot) \le b_\phi$ for a constant b_ϕ,
- $0 \le c(\cdot)$.

There are many possible generalizations of this problem. For instance we might want to allow $\phi(\cdot)$ to be unbounded. If we were considering when to sell a revenue generating asset we might want to allow $c(x) < 0$ in order to describe the benefit of continuing to hold the asset. In financial situations it is appropriate to use a discount rate $\lambda > 0$ to compute present values of future rewards (and costs), seeking a strategy to maximize

$$E\left[e^{-\lambda\mathcal{T}}\phi(X_{\mathcal{T}})1_{\mathcal{T}<\infty} - \sum_{n=0}^{\mathcal{T}-1} e^{-\lambda n}c(X_n)\right]. \tag{8.2}$$

The theory is cleaner for a positive discount rate, but weakening our assumptions on $\phi(\cdot)$ and $c(\cdot)$ introduce more technical issues. We will limit our discussion the undiscounted problem (8.1) under the above hypotheses on $\phi(\cdot)$ and $c(\cdot)$ because the main ideas are present without as many technicalities.

Stopping Times

We will need a strategy to tell us what to look for as we observe $X_0, X_1, \ldots$ and decide when to stop. We could choose $\mathcal{T}$ in some mindless way, for instance just decide to stop at $\mathcal{T} = 5$ regardless of what the chain does and take $\phi(X_5)$ as our reward, for better or worse. But this makes no use of our observations of X_n; surely we can do better by using those observations together with knowledge of the transition probabilities. If there are no continuation costs ($c(\cdot) \equiv 0$) we might look for a state $s^* \in \mathcal{S}$ which maximizes $\phi(\cdot)$ and wait until the first time $\mathcal{T} = n$ that which $X_n = s^*$:

$$\mathcal{T} = \min\{n : X_n = s^*\}.$$

If X_n is recurrent then $P(\mathcal{T} < \infty) = 1$ and this clearly gives us the largest possible reward $\phi(X_{\mathcal{T}})$. But if X_n is transient , or $\phi(s)$ has no maximum, or $c \neq 0$ then the above strategy is usually a poor choice. In general we need to design a strategy more intelligently.

The time-valued random variable $\mathcal{T}$ is the result of following some stopping strategy as the chain evolves. Each different strategy corresponds to a different random variable $\mathcal{T}$. But $\mathcal{T}$ cannot be just any time-valued random variable. The event $\{\mathcal{T} = 5\}$ can depend on $X_0, \ldots, X_5$ but *not* on X_6 or any of the later states of the chain. Similarly $\{\mathcal{T} = 6\}$ is allowed to depend on $X_0, \ldots, X_6$, but not X_7. In other words $\mathcal{T}$ must be a stopping time, as we defined on page 62.

For example suppose our strategy was to take $\mathcal{T}$ to be the first time n for which $\phi(X_n)$ exceeds some threshold, say $\phi(X_n) \ge 25$. Then to decide whether $\mathcal{T} \le 12$ or not we only need to examine $X_0, \ldots X_{12}$. So this *is* a stopping time. (It could be however that $\phi(X_n) \ge 25$ never happens, in which event we would say $\mathcal{T} = \infty$.) Generally the first time something happens is a stopping time. In contrast suppose $\mathcal{M}$ is the time at which $\phi(X_n)$ takes its largest value over all $n = 0, 1, 2, \ldots$. In general to decide whether $\mathcal{M} \le 12$ requires us to look ahead at *all* the X_n to see if any of the future positions of the chain give larger ϕ-values than $X_0, \ldots X_{12}$ did. This $\mathcal{M}$ may be a legitimate random variable, but in general it is not a stopping time.

Our problem is to maximize (8.1) over all stopping times $\mathcal{T}$ in the infinite horizon case, or all stopping times with $\mathcal{T} \le T$ in the finite horizon case.

There are various ways we can modify stoping times to obtain new stopping times. For instance if $\mathcal{T}$ is a stopping time and N is an integer (not random), then

$$\mathcal{T} \wedge N = \begin{cases} \mathcal{T} & \text{if } \mathcal{T} \le N \\ N & \text{if } N < \mathcal{T} \end{cases}$$

is also a stopping time. That is because the event $\{\mathcal{T} \wedge N \leq n\}$ is the same as $\{\mathcal{T} \leq n\}$ for $n \leq N$, and for $n > N$ is the certain event Ω, which always depends on $X_{0:n}$. We can also form the maximum:

$$\mathcal{T} \vee N = \begin{cases} \mathcal{T} & \text{if } \mathcal{T} \geq N \\ N & \text{if } \mathcal{T} < N \end{cases}$$

The following is the key technical result which is needed for the results of this chapter. Please recall our "operator notation": for a function $f(s,n)$ of state and time we write $\mathbf{P}f(i,n)$ for

$$\mathbf{P}f(i,n) = \sum_{j \in \mathcal{S}} p_{i,j} f(j,n).$$

Thus the Markov property says that

$$E[f(X_{n+1}, n+1) | X_{0:n} = i_{0:n}] = \mathbf{P}f(i_n, n+1),$$

or as a generalized conditional,

$$E[f(X_{n+1}, n+1) | X_{0:n}] = \mathbf{P}f(X_n, n+1).$$

Lemma 8.1. *Suppose $\mathcal{T}$ is a bounded stopping time, and $f : \mathcal{S} \times \mathbb{Z}^+ \to \mathbb{R}$ is a bounded function. Then for any $k = 0, 1, 2, \ldots$ we have*

$$E[f(X_{\mathcal{T}}, \mathcal{T}) \,|\, X_{0:k}] 1_{\mathcal{T}>k} = f(X_k, k) 1_{\mathcal{T}>k} + E\left[\sum_{n=k}^{\mathcal{T}-1} [\mathbf{P}f(X_n, n+1) - f(X_n, n)] \,\middle|\, X_{0:k} \right] 1_{\mathcal{T}>k}.$$

Proof. The key calculation is this:

$$\begin{aligned} E[f(X_{n+1}, n+1) - f(X_n, n) \,|\, X_{0:n}] &= E[f(X_{n+1}, n) \,|\, X_{0:n}] - f(X_n, n) \\ &= \mathbf{P}f(X_n, n+1) - f(X_n, n). \end{aligned}$$

This is just the basic Markov property and the fact that $f(X_n, n)$ is $X_{0:n}$-determined. Since $\{\mathcal{T} > n\}$ is $X_{0:n}$-determined, Proposition 3.8 implies that

$$\begin{aligned} E\left[[f(X_{n+1}, n+1) - f(X_n, n)] 1_{\mathcal{T}>n} \,|\, X_{0:n} \right] &= E[f(X_{n+1}, n+1) - f(X_n, n) \,|\, X_{0:n}] 1_{\mathcal{T}>n} \\ &= [\mathbf{P}f(X_n, n+1) - f(X_n, n)] \, 1_{\mathcal{T}>n}. \end{aligned}$$

Now by the Tower Law, for $k \leq n$ we can say

$$E\left[[f(X_{n+1}, n+1) - f(X_n, n)] 1_{\mathcal{T}>n} \,|\, X_{0:k} \right] = E\left[[\mathbf{P}f(X_n, n+1) - f(X_n, n)] \, 1_{\mathcal{T}>n} \,|\, X_{0:k} \right]. \tag{8.3}$$

Next write $f(X_{\mathcal{T}}, \mathcal{T}) 1_{\mathcal{T}>k}$ using a telescoping sum:

$$f(X_{\mathcal{T}}, \mathcal{T}) 1_{\mathcal{T}>k} = f(X_k, k) 1_{\mathcal{T}>k} + \sum_{n=k}^{\mathcal{T}-1} [f(X_{n+1}, n+1) - f(X_n, n)] 1_{\mathcal{T}>k}.$$

Suppose N is an (integer) upper bound on $\mathcal{T}$. Then we can write

$$\sum_{n=k}^{\mathcal{T}-1} [\cdots] 1_{\mathcal{T}>k} = \sum_{n=k}^{N} [\cdots] 1_{\mathcal{T}>n},$$

because on both sides the sum is over $n = k, \ldots, \mathcal{T} - 1$. Therefore

$$f(X_{\mathcal{T}}, \mathcal{T}) 1_{\mathcal{T}>k} = f(X_k, k) 1_{\mathcal{T}>k} + \sum_{n=k}^{N} [f(X_{n+1}, n+1) - f(X_n, n)] 1_{\mathcal{T}>n}.$$

Take conditional expectations of both sides and use (8.3) to complete the proof:

$$\begin{aligned}
E[f(X_\mathcal{T},\mathcal{T}) \,|\, X_{0:k}]1_{\mathcal{T}>k} &= E[f(X_\mathcal{T},\mathcal{T})1_{\mathcal{T}>k} \,|\, X_{0:k}] \\
&= E\left[f(X_k,k)1_{\mathcal{T}>k} + \sum_{n=k}^{N}[f(X_{n+1},n+1) - f(X_n,n)]1_{\mathcal{T}>n} \,|\, X_{0:k}\right] \\
&= E\left[f(X_k,k)1_{\mathcal{T}>k} + \sum_{n=k}^{N}\left[\mathbf{P}f(X_n,n+1) - f(X_n,n)\right]1_{\mathcal{T}>n} \,\middle|\, X_{0:k}\right] \\
&= E\left[f(X_k,k)1_{\mathcal{T}>k} + \sum_{n=k}^{\mathcal{T}-1}\left[\mathbf{P}f(X_n,n+1) - f(X_n,n)\right]1_{\mathcal{T}>k} \,\middle|\, X_{0:k}\right] \\
&= f(X_k,k)1_{\mathcal{T}>n} + E\left[\sum_{n=k}^{\mathcal{T}-1}\left[\mathbf{P}f(X_n,n+1) - f(X_n,n)\right] \,\middle|\, X_{0:k}\right]1_{\mathcal{T}>k}.
\end{aligned}$$

□

Problems with Finite Time Horizon

Let's first consider problems with a specified time constraint

$$\mathcal{T} \le T,$$

where T is a prescribed constant integer. At each time $n \le T$ our decision whether or not to stop (if we haven't already) may involve the value of n. Like many of the problems in preceding chapters, we can solve this by working with an appropriate function of position and time. We define $V(s,n)$ to denote the maximum expected *future* reward less costs assuming we start a time n from $X_n = s$:

$$V(s,n) = \sup_{n\le\mathcal{T}\le T} E_{s,n}\left[\phi(X_\mathcal{T}) - \sum_{t=n}^{\mathcal{T}-1} c(X_t)\right]. \tag{8.4}$$

(The notation $E_{s,n}[\cdot]$ means that we start the chain at time n (not time 0 as before) and with initial state $X_n = s$.) Of course only stoping times $n \le \mathcal{T} \le T$ are included in the supremum. This V is usually called the *optimal value function*.

Suppose for the moment that we have somehow determined the values of $V(s,n)$ for all $s \in \mathcal{S}$ and $n = 0,\dots,T$. Imagine that we start at time 0 and reach $t = n$ and $X_n = s$ without having stopped yet. We have to decide whether to stop now or continue at least one step longer. If $V(s,n) = \phi(s)$ then we should stop now, because no strategy for stopping in the future will produce a better expected future reward (less costs). But if $\phi(s) < V(s,n)$ then there is a better policy than stopping now, so we should continue. So it seems clear that best strategy is to stop the first time that $X_n = s$ has $V(s,n) = \phi(s)$. Of course if we reach time T then we have to stop.

To actually carry out this strategy we need work out $V(s,n)$ in advance. We can (in principle) solve for $V(s,n)$ by working backward from $n = T$ to $n = 0$. If we are at the final time $n = T$ there is no choice left; we most stop now and take our reward. So we have

$$V(s,T) = \phi(s).$$

Next suppose we are at time $n < T$ and the current state is $X_n = s$. Suppose also that we know the values of $V(n+1,\cdot)$. We can either stop now or continue at least one more step. Suppose we continue at least one more step. We pay $c(s)$ to continue and find ourselves at some $X_{n+1} = x$. Then we should follow an optimal strategy for starting at $X_{n+1} = x$. Now observe that if $\mathcal{T} > n$ then

$$\phi(X_\mathcal{T}) - \sum_{t=n}^{\mathcal{T}-1} c(X_t) = -c(X_n) + \left[\phi(X_\mathcal{T}) - \sum_{t=n+1}^{\mathcal{T}-1} c(X_t)\right]$$

The quantity in square brackets is what an optimal strategy starting from X_{n+1} would minimize (in mean). Its optimized mean, conditional on X_{n+1}, is $V(X_{n+1}, n+1)$. So it seems that the best we could do with a stopping time $\mathcal{T} > n$ from $X_n = s$ would be

$$\begin{aligned} E_{s,n}[-c(X_n) + V(X_{n+1}, n+1)] &= -c(s) + E_{s,n}[V(X_{n+1}, n+1)] \\ &= -c(s) + \mathbf{P}V(s, n+1). \end{aligned} \tag{8.5}$$

On the other hand if we stop at time n our reward is simply $\phi(s)$. So $V(s,n)$ must be the larger of $\phi(s)$ and the expression in (8.5). So we calculate $V(\cdot, n)$ from $V(\cdot, n+1)$ as

$$V(s,n) = \max(\phi(s), -c(s) + \mathbf{P}V(s, n+1)), \tag{8.6}$$

for each state $s \in \mathcal{S}$.

Thus starting with $V(\cdot, T) = \phi(\cdot)$ and iterating (8.6) backwards ($n = T-1, T-2, \ldots, 3, 2, 1, 0$) we will be able to construct the optimal value function $V(\cdot, \cdot)$, and with that in hand we know what an optimal stopping strategy should be.

Example 8.1. To illustrate the above procedure let's consider a primitive example of an American call option. X_n will be the Markov chain on $\mathcal{S} = \{5, 10, 15, 20, 25, 30\}$ with transitions $X_n \to X_n - 5, X_n + 0, X_n - 5$ with equal probabilities of $1/3$ except at the end points, where the two possible transitions $X_n \to X_n + 0, X_n \pm 5$ are equally likely. We take

$$\phi(s) = \max(s - 15, 0).$$

There is no continuation cost: $c \equiv 0$. We take $T = 10$ The calculation is rather simple in MATLAB. We use an 6×11 array `v` in which `v(i,n)` will hold $V(5i, n-1)$. The first index `i` coresponds to state $s = 5i$; since MATLAB indexes arrays starting at 1, not 0,the second index `n` corresponds to $t = n - 1$. In brief, we calculate the values of $\phi(s)$ and put them in a (column) vector:

```
phi=[0,0,0,5,10,15]'
```

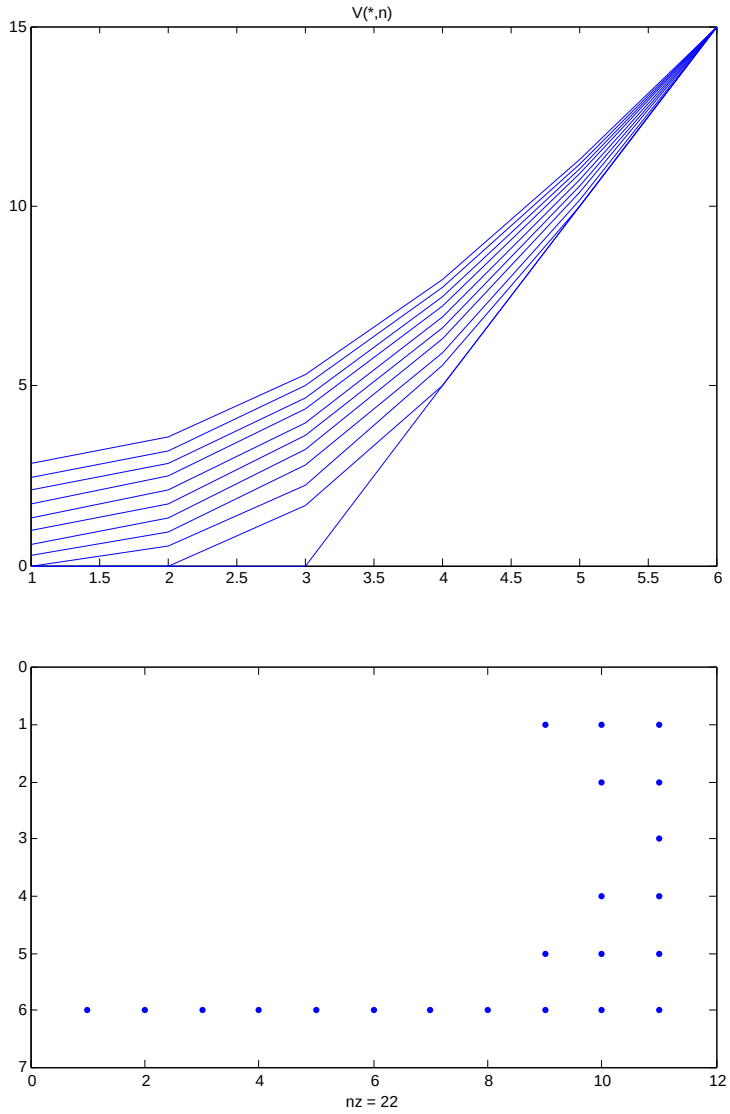
We assign the values $V(\cdot, T) = \phi(\cdot)$ with the statement

```
v(:,11)=phi;
```

and then carry out (8.6) by iterating

```
v(:,i)=max(phi,P*v(:,i+1));
```

as `i` goes from 10 to 1. The results are displayed in the following graphic; the higher values of n are the lower curves, and the horizontal axis uses the array index `i` instead of the corresponding value $s = 5i$. The second graph (produced by `spy(v==phi*ones(1,11))`) shows the `(i,n)` pairs at which it is optimal to stop. The optimal strategy (produced by `spy(v==phi*ones(1,11))`) is to stop ($\mathcal{T}^* = t$) when $(t+1, X_t)$ first hits a dot in the figure.



We now want to *prove* that our prescription above does produce an optimal strategy. We could do this directly by by writing out proofs for all the steps in our reasoning above but that is long and tedious. For instance we would have to be explicit about how we take optimal stopping times, one for each X_{n+1} starting value, and assemble them into a single stopping time for starting at $X_n = s$. Instead we start with the function defined by (8.6) and then prove that it *is* in fact the optimal value function.

Theorem 8.2. *The function $V(s,n)$ for $s \in \mathcal{S}$ and $n = 0, 1, \ldots, T$ defined by $V(\cdot, T) = \phi(\cdot)$ and the iteration (8.6) is the optimal value function for the finite time horizon optimal stopping problem with $\mathcal{T} \leq T$. Given a starting time n and position $X_n = s$, define the stopping time $\mathcal{T}^*$ to be the smallest t with $n \leq t \leq T$ for which $V(X_t, t) = \phi(X_t)$. Then $\mathcal{T}^*$ is optimal.*

Proof. We assume that $V(\cdot,\cdot)$ is the function produced by the iteration (8.6) starting from $V(\cdot, T) = \phi(\cdot)$. Consider any stopping time $n \leq \mathcal{T} \leq T$ and any starting point $X_n = s$. Applying Lemma 8.1 we have

$$V(s,n) = E_{s,n}\left[V(X_{\mathcal{T}}, \mathcal{T}) - \sum_{t=n}^{\mathcal{T}-1}[\mathbf{P}V(X_t, t+1) - V(X_t, t)]\right].$$

But we know from (8.6) that

$$V(X_t, t) \geq -c(X_t) + \mathbf{P}V(X_t, t+1),$$

which rearranged says that

$$-[\mathbf{P}V(X_t, t+1) - V(X_t, t)] \geq -c(X_t).$$

We also know that $V(X_{\mathcal{T}}, \mathcal{T}) \geq \phi(X_{\mathcal{T}})$. Making these substitutions we find that

$$V(s,n) \geq E_{s,n}\left[\phi(X_{\mathcal{T}}) - \sum_{t=n}^{\mathcal{T}-1} c(X_t)\right].$$

Now consider $\mathcal{T}^*$. As the *first* time (n or larger) that $V(X_t, t) = \phi(X_t)$ we know that $\mathcal{T}^*$ is a stopping time. Since $V(X_T, T) = \phi(X_T)$ is always true, we know that $\mathcal{T}^* \leq T$. The t's in the sum $\sum_{t=n}^{\mathcal{T}^*-1}$ are all $t < \mathcal{T}^*$, so it must be that $V(X_t, t) \neq \phi(X_t)$. But then (8.6) says that

$$V(X_t, t) = -c(X_t) + \mathbf{P}V(X_t, n+1),$$

so that

$$-c(X_t) = [\mathbf{P}V(X_t, t+1) - V(X_t, t)].$$

And since $V(X_{\mathcal{T}^*}, \mathcal{T}^*) = \phi(X_{\mathcal{T}^*})$ the substitutions we made above are all equalities in this case, so we find that

$$V(s,n) = E_{s,n}\left[\phi(X_{\mathcal{T}^*}) - \sum_{t=n}^{\mathcal{T}^*-1} c(X_t)\right].$$

It now follows that V defined by (8.6) is the same as (8.4). □

Problems with Infinite Time Horizon

Next we remove the restriction $\mathcal{T} \leq T$ and allow any $\mathcal{T} \leq \infty$. The situation is a little simpler now, because the optimal value function will be a function of the state variable alone, $V(x)$. It will no longer depend on the time variable t. But by including the possibility that $\mathcal{T} = \infty$ we are allowing strategies that continue waiting for some combination of circumstances that might never actually occur. If that happens there is no payoff, because

$$\phi(X_{\mathcal{T}})1_{\mathcal{T}<\infty} = 0 \text{ when } \mathcal{T} = \infty.$$

This complicates things because waiting for the first time $V(X_n) = \phi(X_n)$ may produce $\mathcal{T} = \infty$ and may be non-optimal!

Dealing with unbounded and possibly infinite $\mathcal{T}$ is the principal technical issue we face in the proofs below. The next lemma provides the technical tool that will allow us to use $\mathcal{T} = \lim_{N\to\infty} \mathcal{T} \wedge N$ in certain expectations.

Lemma 8.3. *Suppose that $\mathcal{T}$ is a stopping time for which*

$$-\infty < E_s\left[\phi(X_{\mathcal{T}})1_{\mathcal{T}<\infty} - \sum_{n=0}^{\mathcal{T}-1} c(X_n)\right].$$

Then

$$\lim_{N\to\infty} E_s\left[\phi(X_{\mathcal{T}\wedge N})1_{\mathcal{T}<\infty} - \sum_{n=0}^{(\mathcal{T}\wedge N)-1} c(X_n)\right] = E_s\left[\phi(X_{\mathcal{T}})1_{\mathcal{T}<\infty} - \sum_{n=0}^{\mathcal{T}-1} c(X_n)\right].$$

Proof. Let

$$B = E_s\left[\phi(X_{\mathcal{T}})1_{\mathcal{T}<\infty} - \sum_{n=0}^{\mathcal{T}-1} c(X_n)\right].$$

Then

$$E_s\left[\sum_{n=0}^{\mathcal{T}-1} c(X_n)\right] \leq E_s\left[\phi(X_{\mathcal{T}})1_{\mathcal{T}<\infty}\right] - B \leq b_\phi - B < \infty.$$

Since

$$\sum_{n=0}^{(\mathcal{T}\wedge N)-1} c(X_n) \to \sum_{n=0}^{\mathcal{T}-1} c(X_n)$$

it follows from the Dominated (or Monotone) Convergence Theorem that

$$E_s\left[\sum_{n=0}^{\mathcal{T}\wedge N-1} c(X_n)\right] \to E_s\left[\sum_{n=0}^{\mathcal{T}-1} c(X_n)\right].$$

Since

$$\phi(X_{\mathcal{T}\wedge N})1_{\mathcal{T}<\infty} \to \phi(X_{\mathcal{T}})1_{\mathcal{T}<\infty}$$

and $|\phi(X_{\mathcal{T}\wedge N})1_{\mathcal{T}<\infty}| \le b_\phi$ the Dominated Convergence Theorem applies again to tell us that

$$E_s\left[\phi(X_{\mathcal{T}\wedge N})1_{\mathcal{T}<\infty}\right] \to E_s\left[\phi(X_{\mathcal{T}})1_{\mathcal{T}<\infty}\right].$$

Subtracting these two limits gives the assertion of the lemma. □

We start by characterizing the optimal value function,

$$V(s) = \sup_{\mathcal{T}} E_s\left[\phi(X_{\mathcal{T}})1_{\mathcal{T}<\infty} - \sum_{n=0}^{\mathcal{T}-1} c(X_n)\right],$$

where $\mathcal{T}$ ranges over all stopping times. The same reasoning which led to (8.6) now suggests that $V(\cdot)$ should solve equation (8.7) of the following theorem. But as before, we will start by defining $V(\cdot)$ in terms of this equation and then proving that $V(\cdot)$ really is the optimal value function.

Lemma 8.4. *There exists a bounded function $V : \mathcal{S} \to \mathbb{R}$ which solves*

$$V(s) = \max(\phi(s), -c(s) + \mathbf{P}V(s)),\ s \in \mathcal{S}, \tag{8.7}$$

and which is less than or equal to every other function $u(s)$ which solves

$$u(s) \ge \max(\phi(s), -c(s) + \mathbf{P}u(s)) \text{ for all } s \in \mathcal{S}.$$

Notice that because $\phi(s)$ is nonnegative any solution V must be nonnegative. Likewise any function $u(\cdot)$ satisfying the inequality of the theorem must be nonnegative.

Proof. Starting with $v_0(s) = \phi(s)$, define the sequence of functions

$$v_{n+1}(s) = \max(\phi(s), -c(s) + \mathbf{P}v_n(s)). \tag{8.8}$$

The bound $\phi(s) \le b_\phi$ for all s implies that $\mathbf{P}\phi(s) \le b_\phi$, and therefore $-c(s) + \mathbf{P}\phi(s) \le b_\phi$ as well (because $c(s) \ge 0$). Therefore $v_1(s) \le b_\phi$. It follows by induction that $v_n(s) \le b_\phi$ for all n and s.

Observe also that $v_1(s) \ge \phi(s) = v_0(s)$ for all s. This implies that

$$\begin{aligned}
\mathbf{P}v_1(s) &\ge \mathbf{P}v_0(s)\\
-c(s) + \mathbf{P}v_1(s) &\ge -c(s) + \mathbf{P}v_0(s)\\
\max(\phi(s), -c(s) + \mathbf{P}v_1(s)) &\ge \max(\phi(s), -c(s) + \mathbf{P}v_0(s))\\
v_2(s) &\ge v_1(s).
\end{aligned}$$

We can repeat this inductively to see that $(v_n(s))$ is an increasing sequence, for each s. Since it is a *bounded* increasing sequence we know that its limit exists. So we can define

$$V(s) = \lim_{n\to\infty} v_n(s).$$

This function obeys the same bounds $0 \le V(s) \le b_\phi$ as all the $v_n(s)$. By taking $\lim_{n\to\infty}$ on both sides of (8.8) we see that $V(\cdot)$ solves (8.7). (Note that if S is infinite then $\mathbf{P}v_n(s)$ is an infinite series, but since all

terms are nonnegative and the $v_n(s)$ are increasing, $\lim_n \mathbf{P}v_n(s) = \mathbf{P}(\lim_n v_n)(s)$ is valid by the monotone convergence theorem for infinite series.)

Now suppose $u(\cdot)$ solves the u-inequality of the theorem. Then

$$v_0(s) = \phi(s) \le u(s).$$

It follows that

$$v_1(s) = \max(\phi(s), -c(s) + \mathbf{P}v_0(s)) \le \max(\phi(s), -c(s) + \mathbf{P}u(s)) \le u(s).$$

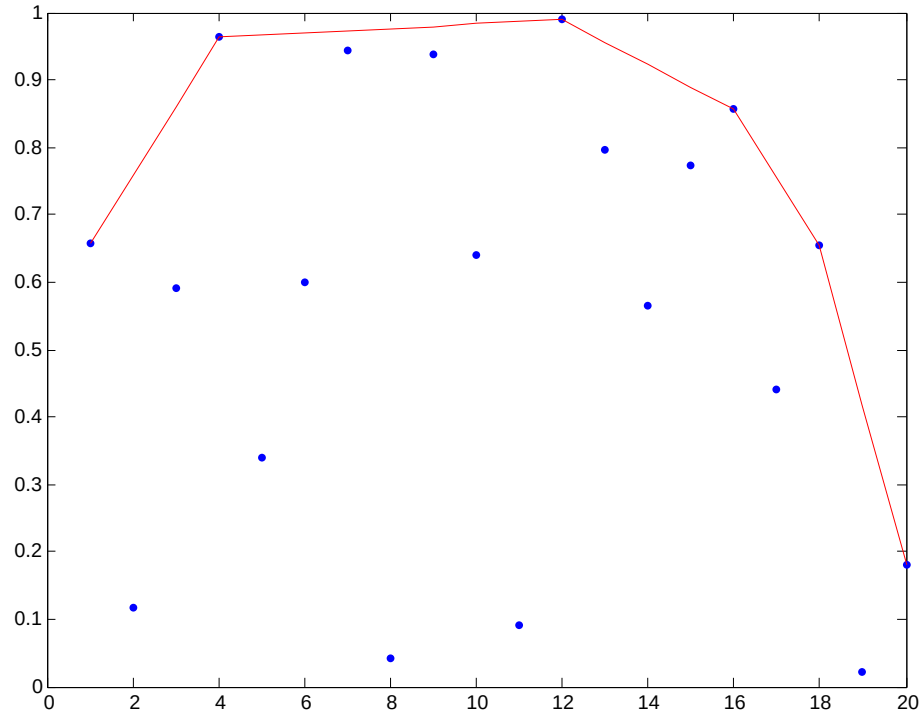
Repeating this inductively we find that $v_n(s) \le u(s)$ for all n, and consequently $V(s) \le u(s)$ as well. □

In the case of $c(s) \equiv 0$ the u-equation of the theorem says that $\phi(s) \le u(s)$ and $\mathbf{P}u(s) \le u(s)$. The latter is often called a *superharmonic* or *excessive* function for the chain. Thus for $0 \equiv c(\cdot)$, the function $V(\cdot)$ is referred to as the *minimal superharmonic majorant* or *minimal excessive majorant* of $\phi(\cdot)$. The next example illustrates this graphically.

Example 8.2. Let X_n be the random walk on $\{1, \dots, 20\}$ with absorption at the endpoints, 1 and 20. Take $0 \equiv c(\cdot)$ and the reward function $\phi(i)$ as illustrated by the dots in the figure below. For $u(\cdot)$ to be superharmonic in this example means that for each $1 < i < 20$,

$$\frac{u(i-1) + u(i+1)}{2} \le u(i).$$

Graphically this says that $u(i)$ must lie on or above the straight line connecting its two neighbors on the graph. The solid line is the graph of $V(\cdot)$, the minimal superharmonic majorant. This is what you get if you draw a string tightly over the points on the graph of $\phi(\cdot)$ and tie the ends at $\phi(1)$ and $\phi(20)$.



The set $H = \{1, 4, 12, 16, 18, 20\}$ of points where $V(i) = \phi(i)$ are the states where an optimal policy should stop, as we will see.

Theorem 8.5. *The function $V(s)$ of Lemma 8.4 is the optimal value function of the optimal stopping problem.*

Proof. Since $V(s)$ refers to the function constructed in the proof of Theorem 8.4, we will want a different notation to refer to the optimal value function. We will use

$$W(s) = \sup_{\mathcal{T}} E_s \left[\phi(X_{\mathcal{T}}) 1_{\mathcal{T}<\infty} - \sum_{n=0}^{\mathcal{T}-1} c(X_n) \right].$$

Our goal is to prove that $V(s) = W(s)$. (Once the proof is over we will no longer need the separate notation W.)

Begin by observing that the iteration (8.8) is the same as (8.6), just using different notation. They both start with $\phi(\cdot)$. What we called $v_n(s)$ in (8.8) is identical with what we called $v(s, T-n)$ in (8.6). So our $v_n(s)$ is the optimal value function for the stopping problem with stopping times contained by $\mathcal{T} \leq n$. Such a $\mathcal{T}$ is still allowed in the unconstrained problem, so we know that for $\mathcal{T} \leq n$ we must have

$$E_s \left[\phi(X_{\mathcal{T}}) 1_{\mathcal{T}<\infty} - \sum_{n=0}^{\mathcal{T}-1} c(X_n) \right] \leq W(s).$$

Moreover the supremum on the left side over $\mathcal{T} \leq n$ gives us $v_n(s)$. This shows that $v_n(s) \leq W(s)$ for each n. Taking the limit implies that

$$V(s) \leq W(s).$$

To establish inequality in the other direction we want to show that for any stopping time,

$$E_s \left[\phi(X_{\mathcal{T}}) 1_{\mathcal{T}<\infty} - \sum_{n=0}^{\mathcal{T}-1} c(X_n) \right] \leq V(s). \tag{8.9}$$

This will imply that $W(s) \leq V(s)$ and complete the proof. So consider any stopping time $\mathcal{T}$. If it turns out that

$$E_s \left[\phi(X_{\mathcal{T}}) 1_{\mathcal{T}<\infty} - \sum_{n=0}^{\mathcal{T}-1} c(X_n) \right] < 0$$

then we don't have to do anything, because we know $0 \leq V(s)$. So suppose that

$$0 \leq E_s \left[\phi(X_{\mathcal{T}}) 1_{\mathcal{T}<\infty} - \sum_{n=0}^{\mathcal{T}-1} c(X_n) \right]. \tag{8.10}$$

Pick any integer N and consider the bounded stopping time $\mathcal{T} \wedge N$. By Theorem 8.2 we can say that

$$E_s \left[\phi(X_{\mathcal{T}\wedge N}) 1_{(\mathcal{T}\wedge N)<\infty} - \sum_{n=0}^{(\mathcal{T}\wedge N)-1} c(X_n) \right] \leq v_N(s) \leq V(s).$$

Notice that $1_{\mathcal{T}<\infty} \leq 1 = 1_{(\mathcal{T}\wedge N)<\infty}$, so we make the left side even smaller if we make that change.

$$E_s \left[\phi(X_{\mathcal{T}\wedge N}) 1_{\mathcal{T}<\infty} - \sum_{n=0}^{(\mathcal{T}\wedge N)-1} c(X_n) \right] \leq v_N(s) \leq V(s).$$

By virtue of (8.10) we can apply Lemma 8.3. Letting $N \to \infty$ implies (8.9). Taking the supremum over all stopping times $\mathcal{T}$ we find that

$$W(s) \leq V(s),$$

completing the proof. □

Example 8.3. The Best Offer Problem.

Imagine you are trying to sell your house. Offers come in one at a time given by an i.i.d. sequence X_n. We will suppose the $X_n > 0$ are discrete with $p_i = P(X_n = x_i)$ and that $E[|X_n|] < \infty$. After receiving

an offer you must decide whether to accept it or reject it. If you accept the offer X_n your receive X_n in exchange for your house. If you reject it you pay $c > 0$ in taxes, maintenance and upkeep costs before the next offer comes in. The problem is to determine the strategy for deciding when to accept or reject an offer so as to maximize your expected net gain.

We can view X_n as a Markov chain with $p_{i,j} = p_j$ for the transition $x_i \to x_j$. The reward function is $\phi(x) = x$ and the continuation cost is a constant $c > 0$. We seek the optimal value function $V(x)$ (defined for the set of possible offer values $x = x_i$), which must satisfy

$$V(x) = \max(x, -c + \mathbf{P}V(x)).$$

Now $-c + \mathbf{P}V(x) = -c + E[V(X)]$ does not depend on x, so is a constant which we will call α. Thus $V(x)$ is of the form

$$V(x) = \max(x, \alpha).$$

We need to determine the value of α. It is determined by the equation

$$\begin{aligned} \alpha &= -c + E[V(X)] \\ &= -c + E[\max(X, \alpha)] \\ c &= E[\max(X, \alpha) - \alpha] \\ &= E[\max(X - \alpha, 0)]. \end{aligned}$$

So if we define the function $\psi(a)$ by

$$\psi(a) = E[\max(X - a, 0)] = \sum_i (x_i - a)^+ p_i$$

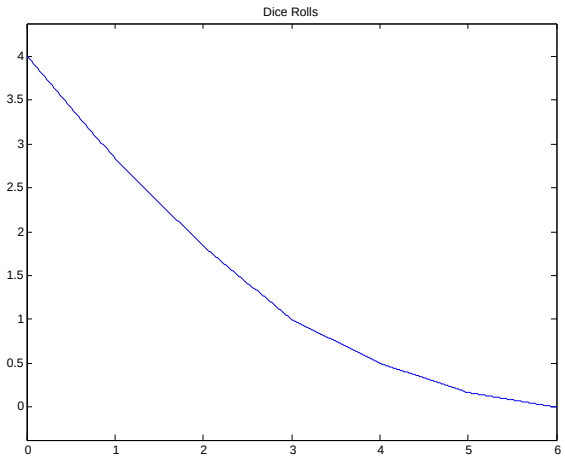
the α will be the value of a which solves $\psi(a) = c$.

Some elementary properties of $\psi(a)$ are as follows.

- $\psi(a) \geq 0$;
- $\psi(a)$ is a continuous function of a;
- $\psi(a)$ is strictly decreasing for $a < \max_i x_i$, and $\psi(a) = 0$ for $\max x_i \leq a$;
- $\lim_{a \to -\infty} \psi(a) = \infty$, $\psi(0) = E[X]$, and $\lim_{a \to +\infty} \psi(a) = 0$.

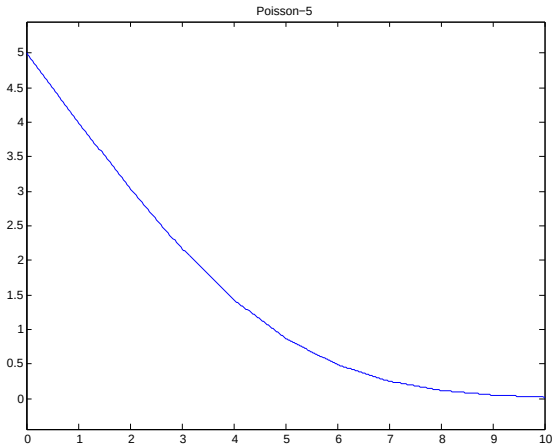
It follows that for any $0 < c$ there will be a unique solution α. For $c < E[X]$ the solution will be $\alpha > 0$ but for $E[X] \leq c < \max x_i$ the solution will be $\alpha \leq 0$. What this means for the optimal strategy is that if you get an offer $X_i < \alpha$ then $X_n < V(X_n)$ so it is best to reject this offer and wait. But when you get an offer $X_n \geq \alpha$ then $X_n = V(X_n)$ so you can't do any better than stopping now. Notice that if $E[X] \leq c$ then $\alpha \leq 0$ so *every* offer will satisfy $\alpha \leq X_n$; you should simply take the first offer regardless!

Let's consider some particular cases. First suppose the X_n are the outcomes of a fair dice roll, and $c = 1$. Here is the graph of $\psi(a)$.



We find that $\psi(3) = 1 = c$ so $\alpha = 3$. We accept the first offer $X_n \geq 3$.

For a different example suppose the X_n has a Poisson distribution with $\lambda = 5$. The graph of $\psi(a)$ is as displayed here.



If we again take $c = 1$ we solve $\psi(a) = 1$ numerically to find that $\alpha = 4.780\ldots$. Since the X_n are always integers, the optimal policy is to accept the first offer $X_n \geq 5$.

Next we want to explore the connection between the optimal value function V and optimal stoping times. By analogy with the finite time horizon case, we would expect to be able to form an optimal stopping time by finding the set

$$H = \{s \in \mathcal{S} : V(s) = \phi(s)\},$$

and taking $\mathcal{T}_H$ to be the first time that $X_n \in H$. Unfortunately this can fail, as the next example shows.

Example 8.4. Consider the chain with state space $\mathbb{N}$ and the following transition probabilities.

$$p_{n,1} = 1/n^2,\ p_{n,n+1} = 1 - 1/n^2,\ p_{n,k} = 0 \text{ otherwise.}$$

In other words the only possible transitions are $n \to n+1$ and $n \to 1$, with probabilities $1/n^2$ and $1 - 1/n^2$ respectively. We want to consider the optimal stopping problem with reward function $\phi(n) = 1 - 1/n$ for $1 < n$ and $\phi(1) = 1$ and $c \equiv 0$. For an initial position $n > 1$ pick any $m > n$ and let $\mathcal{T}^m$ be the first time the chain reaches either 1 or m. Clearly

$$1 \geq E_n[\phi(X_{\mathcal{T}^m})] \geq 1 - 1/m.$$

It follows that $V(n) = 1$ for all n. So $H = \{1\}$ and $\mathcal{T}_H$ is the first time that $X_n = 1$. But the chain is transient, in other words $P(\mathcal{T}_H = \infty) > 0$. To see this, consider any $X_0 = n > 1$.

$$\begin{aligned} P_n(\mathcal{T}_H > k) &= P_n(X_1 = n+1, X_2 = n+2, \ldots, X_k = n+k) \\ &= p_{n,n+1} p_{n+1,n+2} \cdots p_{n+k-1,n+k}. \end{aligned}$$

Now

$$p_{j,j+1} = 1 - 1/j^2 = \frac{j^2 - 1}{j^2} = \frac{(j-1)(j+1)}{j^2}.$$

So we find that

$$\begin{aligned} P_n(\mathcal{T}_H > k) &= \frac{(n-1)(n+1)}{n^2} \frac{(n)(n+2)}{(n+1)^2} \frac{(n+1)(n+3)}{(n+2)^2} \cdots \frac{(n+k-2)(n+k)}{(n+k-1)^2} \\ &= \frac{n-1}{n} \frac{n+k}{n+k-1}. \end{aligned}$$

Therefore

$$P_n(\mathcal{T}_h = \infty) = \lim_{k \to \infty} P_n(\mathcal{T}_H > k) = \frac{n-1}{n} > 0.$$

Thus,

$$E_n[\phi(X_{\mathcal{T}})1_{\mathcal{T}_H<\infty}] = 1 - \frac{n-1}{n} = \frac{1}{n}.$$

We see that $\mathcal{T}_H$ is a rather bad strategy for large $X_0 = n$. In fact no optimal strategy exists for this example, because an optimal strategy would have to produce $X_{\mathcal{T}^*} = 1$ with $P(\mathcal{T}^* < \infty) = 1$, which is impossible because of the transience of the chain!

For infinite horizon problems in general it is possible that no optimal strategy exists. The following gives us some sufficient conditions for $\mathcal{T}_H$ to be optimal.

Theorem 8.6. *Consider the infinite horizion optimal stopping problem and let $H = \{s \in \mathcal{S} : V(s) = \phi(s)\}$ and $\mathcal{T}_H$ be the first time that $X_n \in H$. $\mathcal{T}_H$ will be an optimal stopping time if any of the following is true.*

a) $\mathcal{T}_H < \infty$ with probability 1.

b) there exists some optimal strategy $\mathcal{T}^*$.

c) $\mathcal{S}$ is finite.

Part a) says that the possibility of $\mathcal{T}_H = \infty$ is the only thing that prevents $\mathcal{T}_H$ from being optimal. Part b) says that $\mathcal{T}_H$ will be optimal if anything is. Part c) says that non-optimality of $\mathcal{T}_H$ is never a concern for finite state spaces.

To apply this in the Best Offer Problem, note that for $0 < c$ we have $\alpha < \max x_i$ so that with probability 1 there will eventually be an offer with $\alpha \leq X_n$. This means part a) of the theorem applies; the strategy we described above is indeed optimal!

The proof of the theorem is a bit too technical for us so we will not include it. Versions of the three parts of the theorem, under hypotheses not quite the same as ours, have been proven at various places in the literature. Part a) is Taylor's Corollary 1; [61]. Part b) is his Theorem 3 as well as Theorem 3 in Ferguson [23]. Part c) is explained in Dynkin & Yushkevtich [20].

Finally when there is no optimal strategy, we can always produce a nearly optimal one in a manner similar to $\mathcal{T}_H$. For $\epsilon > 0$ let

$$H_\epsilon = \{s \in \mathcal{S} : \phi(s) \geq V(s) - \epsilon\}.$$

and define

$$\mathcal{T}_{H_\epsilon} = \min\{n : X_n \in H_\epsilon\}.$$

Theorem 8.7. *For $\epsilon > 0$, $\mathcal{T}_{H_\epsilon} < \infty$ with probability 1 and is an ϵ-optimal strategy, meaning that*

$$E_s\left[\phi(X_{\mathcal{T}_{H_\epsilon}}) - \sum_{n=0}^{\mathcal{T}_{H_\epsilon}-1} c(X_n)\right] \geq V(s) - \epsilon.$$

A proof of a version of this can be found in Taylor [61]; see his Theorem 2 (ii). It is also discussed in Dynkin & Yushkevtich [20].

Example 8.5. An example which appears in many treatments of optimal stopping is the *problem of optimal choice*. (See [20] Chapter 3 for instance, although their analysis is different than ours). You are interviewing applicants for a job. There is a pool of $N \geq 2$ applicants, the number N being known to you. Once you have interviewed a set of applicants you will be able to rank order them from most qualified to least qualified. The goal is to offer the job to the most qualified applicant in the pool. (Second best is not acceptable!) This would be simple if you could interview them all before making the job offer, but the rules are that after each interview you must offer the job to the person you just interviewed or dismiss them and go on to interview another applicant. You can't call an applicant back once you have dismissed them. What should your strategy be to maximize the probability that you offer the job to the most qualified applicant from the original pool? If you interview an applicant and find that they are not the best so far, then clearly you don't want to offer the job to them. Only when the most recently interviewed applicant is the best you have seen so far do you need to decide whether or not to offer them the job.

It may not be obvious that there is a Markov chain here, but there is. Let the state space S consist of the ordered pairs (n, i) where $n = 1, \ldots, N$ and $i = 0, 1$ together with a dead state Δ. The interpretation of

$X_n = (n, i)$ is that n applicants have been interviewed and the most recent one is the best so far if $i = 1$, but not the best so far if $i = 0$. The chain starts in state $(1, 1)$. From (n, i) there are two possible transitions:

$$(n, i) \to (n + 1, 0) \text{ with probability } \frac{n}{n+1}$$
$$(n, i) \to (n + 1, 1) \text{ with probability } \frac{1}{n+1}.$$

These transition probabilities may seem intuitively reasonable; we will give a careful justification at the end of our calculations. For now we will take them for granted. To complete the specification of the chain we include the probability 1 transitions to the dead state:

$$(N, i) \to \Delta \text{ and } \Delta \to \Delta$$

You might notice that this is really a finite time horizon problem, but for a Markov chain whose transition probabilities depend on the time of the transition, a *non*homogenous chain. So we have made it homogeneous by making the time part of the state, and adding the dead state to account for "time has run out". In this way we can treat it as an infinite horizon problem. Since there are only a finite number of states, there will be an optimal strategy. Our goal is to describe it explicitly.

The reward for a given state is the probability that the person you just interviewed is actually the best in the original pool. This means $\phi(n, 0) = 0$, and we take $\phi(\Delta) = 0$ because if you didn't offer the job to any of the applicants, then you have certainly failed to offer it to the best. For $n \leq N$ we want $\phi(n, 1)$ to be the probability that the future states of the chain are precisely $(n + 1, 0)$, $(n + 2, 0)$, $\ldots$, $(N, 0)$ and then Δ. Based on the transition probabilities, this probability is

$$\phi(n, 1) = \frac{n}{n+1}\frac{n+1}{n+2}\cdots\frac{N-1}{N} = \frac{n}{N}.$$

(There is no continuation cost: $c \equiv 0$.)

Now let's consider the value function V. For $n < N$ (8.7) becomes the following equations describing V

$$\begin{aligned} V(n, 0) &= \frac{n}{n+1}V(n + 1, 0) + \frac{1}{n+1}V(n + 1, 1) \\ V(n, 1) &= \max\left(\frac{n}{N}, \frac{n}{n+1}V(n + 1, 0) + \frac{1}{n+1}V(n + 1, 1)\right) \\ &= \max\left(\frac{n}{N}, V(n, 0)\right). \end{aligned}$$

For $n = N$ we have

$$V(N, 1) = 1,\ V(N, 0) = 0.$$

Starting with these we can work backwards to determine $V(n, i)$. Observe that starting at $n = N$ and progressing downward the $V(n, 0)$ values start small and the $V(n, 1)$ values start large. In fact as long as $V(n, 0) \leq \frac{n}{N}$ we will have $V(n, 1) = \frac{n}{N}$, $(n, 1) \in H$. So for some segment of the larger n values, $n = k + 1, \ldots, N - 1, N$ we will have $V(n, 1) = \frac{n}{N}$ and

$$V(n, 0) = \frac{n}{n+1}V(n + 1, 0) + \frac{1}{n+1}\frac{n+1}{N} = \frac{n}{n+1}V(n + 1, 0) + \frac{1}{N}$$

We can work this out recursively:

$$\begin{aligned}
V(N-1,0) &= \frac{1}{N} \\
&= \frac{N-1}{N}\left[\frac{1}{N-1}\right] \\
V(N-2,0) &= \frac{N-2}{N-1}\frac{1}{N}+\frac{1}{N} \\
&= \frac{N-2}{N}\left[\frac{1}{N-1}+\frac{1}{N-2}\right] \\
V(N-3,0) &= \frac{N-3}{N-2}\frac{N-2}{N}\left[\frac{1}{N-1}+\frac{1}{N-2}\right]+\frac{1}{N} \\
&= \frac{N-3}{N}\left[\frac{1}{N-1}+\frac{1}{N-2}+\frac{1}{N-3}\right] \\
&\vdots \\
V(n,0) &= \frac{n}{N}\left[\frac{1}{N-1}+\cdots+\frac{1}{n}\right]
\end{aligned}$$

This remains correct as long as $V(n,0) \leq \frac{n}{N}$, i.e. as long as

$$\frac{1}{N-1}+\cdots+\frac{1}{n} \leq 1.$$

But at some k we will find that

$$\frac{1}{N-1}+\cdots+\frac{1}{k} > 1 \geq \frac{1}{N-1}+\cdots+\frac{1}{k+1}.$$

That means that

$$V(k,0) = \frac{k}{N}\left[\frac{1}{N-1}+\cdots+\frac{1}{k}\right]$$

but $(k,1) \notin H$, and

$$V(k,1) = V(k,0).$$

As we continue to $n = k-1$ we see that $V(k-1,0) = V(k,0)$, and since $\frac{k-1}{N} < \frac{k}{N}$ we have $(k-1,1) \notin H$ and $V(k-1,1) = V(k-1,0)$ as well. This will continue for all $n \leq k$:

$$V(n,0) = V(n,1) = V(k,0) \text{ for all } n \leq k.$$

So we have completely solved the problem! Given N we identify k as the largest value for which

$$\frac{1}{N-1}+\cdots+\frac{1}{k} > 1.$$

Then the optimal stopping set is

$$H = \{(n,1) : k < n\}.$$

This means that the optimal strategy is to interview but dismiss the first k applicants regardless of how good they are, and then offer the job to the next applicant who is better qualified than all who were interviewed previously. The optimal value function is

$$V(n,0) = \begin{cases} \frac{n}{N}\left[\frac{1}{N-1}+\cdots+\frac{1}{n}\right] & \text{for } k < n \\ \frac{k}{N}\left[\frac{1}{N-1}+\cdots+\frac{1}{k}\right] & \text{for } n \leq k; \end{cases}$$

$$V(n,1) = \begin{cases} \frac{n}{N} & \text{for } k < n \\ \frac{k}{N}\left[\frac{1}{N-1}+\cdots+\frac{1}{k}\right] & \text{for } n \leq k; \end{cases}$$

The probability of success is $V(1,1) = V(k,0)$.

We can calculate the values k and $V(k,0)$ for various N. For instance when $N = 5$ it turns out that $k = 3$ and the probability of success is $V(1,1) = .4333$. When $N = 20$ then $k = 8$ and $V(1,1) = .3854$.

Now we return to justify the transition probabilities that we specified above. Let

$$A = \{1, 2, \ldots, N\}$$

be the set of applicants, numbered in the order in which they will be interviewed. Their qualifications produce a rank ordering of them. We will represent that by a function ρ so that $\rho(i)$ indicates what position in the ordering applicant i has. For instance $\rho(5) = 2$ means that the fifth applicant has the second worst qualifications. Our goal is to offer the job to that applicant j with $\rho(j) = N$. Now ρ is a permutation on N elements, a one-to-one mapping from A to itself. (A is doing double duty here, both as the set of applicants and as the set of ranks.) There are $N!$ different rankings ρ, and *our basic assumption will be that they are all equally likely*, the actual ρ being chosen from among them using a uniform distribution.

Now suppose you have interviewed applicants 1 through n. The ranks of these applicants form a subset $B \subseteq A$:

$$B = \{\rho(1), \rho(2), \ldots, \rho(n)\}.$$

Let's use b_1, b_2, ..., b_n to indicate the elements of B listed in numerical order.

For example, suppose $N = 10$ and

$$(\rho(1), \rho(2), \ldots, \rho(10)) = (8, 4, 2, 9, 3, 1, 10, 5, 6, 7).$$

With $n = 4$ we have

$$\begin{aligned} B &= \{8, 4, 2, 9\} \\ &= \{2, 4, 8, 9\} \text{ listed in order,} \end{aligned}$$

so that

$$b_1 = 2,\ b_2 = 4,\ b_3 = 8,\ b_4 = 9.$$

After the fourth interview you *don't* know what B is, but you do know the *relative order* within B, i.e. in our example you know that

$$\rho(3) < \rho(2) < \rho(1) < \rho(4).$$

This corresponds to a certain permutation α of $\{1, \ldots n\}$. In our example

$$\alpha(1) = 3,\ \alpha(2) = 2,\ \alpha(3) = 1,\ \alpha(4) = 4.$$

In general,

$$\rho(i) = b_{\alpha(i)},\ \text{for } i = 1, \ldots, n. \tag{8.11}$$

The applicants you have not interviewed yet form the set $C = A^c$. There is a relative ordering γ within C as well. γ is a permutation of $\{n+1, \ldots N\}$ so that if the elements of C listed in order are c_{n+1}, c_{n+2}, $\ldots c_N$ then

$$\rho(j) = c_{\gamma(j)},\ \text{for } j = n+1, \ldots, N. \tag{8.12}$$

In our example,

$$c_5 = 1,\ c_6 = 3,\ c_7 = 5,\ c_8 = 6,\ c_9 = 7,\ c_{10} = 10,$$

and

$$\gamma(5) = 6,\ \gamma(6) = 5,\ \gamma(7) = 10,\ \gamma(8) = 7,\ \gamma(9) = 8,\ \gamma(10) = 9.$$

This gives us a decomposition of ρ into three parts: a subset $B \subseteq A$ of n elements, a permutation α on n elements, and a permutation γ on $N - n$ elements. Given the subset size n, each ρ determines B, α and γ uniquely. Conversely given any B, α and γ we can reconstruct ρ from (8.11) and (8.12). (The complement C and the numberings b_i, c_j are completely determined by the choice of B.) In other words we have a

one-to-one correspondence between ρ and triples (B, α, γ). We can confirm this by counting. There are $\binom{N}{n}$ choices for B, $n!$ choices for α and $(N-n)!$ choices for γ, so the total number of triples is

$$\binom{N}{n} n!(N-n)! = N!,$$

agreeing with the total number of choices for ρ.

The overall ranking ρ is not known to you, but once you have interviews the first n applicants you know their relative ranking α. If $\alpha(n) = n$ you the chain is in state $(n, 1)$ and if $\alpha(n) < n$ the chain is in state $(n, 0)$. For a given α there are a total of

$$\binom{N}{n}(N-n)! = \frac{N!}{n!}$$

possible ρ's consistent with it, each equally likely.

When you interview the $(n+1)$st applicant the overall ranking ρ doesn't change, but you will have a new relative ranking $\tilde{\alpha}$ on the $n+1$ applicants you have interviewed so far. It has to be consistent with α, meaning that the the relative rankings of $1, \ldots, n$ according to $\tilde{\alpha}$ must agree with those according to α. All that changes is where the new interviewee $n+1$ falls in the ranking relative to $1, \ldots, n$. So there are a total of $n+1$ possibilities for $\tilde{\alpha}$ for a given α. Each $\tilde{\alpha}$ is consistent with a total of

$$\binom{N}{n+1}(N-(n+1))! = \frac{N!}{(n+1)!},$$

possible ρ's. So the conditional probability one of the consistent $\tilde{\alpha}$'s given α is

$$\frac{\frac{N!}{(n+1)!}}{\frac{N!}{n!}} = \frac{1}{n+1}.$$

Of the $n+1$ possibilities for $\tilde{\alpha}$ which are consistent with α, exactly one has $\tilde{\alpha}(n+1) = n+1$ so corresponds to a chain state of $(n+1, 1)$.

Problems with Discounting

This yet-to-be-written section is to give a brief summary of how the above results generalize to discounted case: equation (8.2). This will be worthwhile for discussion of American options in Chapter 10.

8.2 Dynamic Programming and Optimal Control

This yet-to-be-written section is to discuss dynamic programming and optimal control of finite horizon and discounted problems. Consideration of gambling strategies is a nice place to start the discussion. Some references are Bertsekas [5] and Puterman [49].

8.3 Optimizing the Mean per Step

This yet-to-be-written section is to discuss optimal control of the long run average cost. See Howard [30] and Bertsekas [5] for instance.

Problems

Problem 8.1
Consider our standard optimal stopping problem except that instead of $\phi(\cdot) \geq 0$ we assume only that $\phi(\cdot)$ is bounded below:

$$\phi(\cdot) \geq -b$$

for some constant b. Show that replacing the reward by $\tilde{\phi}(\cdot) = \phi(\cdot) + b$ gives an equivalent problem for which $\tilde{\phi}(\cdot) \geq 0$ is satisfied. How are the optimal value functions $V(\cdot)$ and $\tilde{V}(\cdot)$ related? How are optimal stopping rules for the two versions related?

. . . . . neg

Problem 8.2
You roll a fair dice until you either choose to stop or else roll a 1 (at which time you are required to stop). Your reward is the result of the final dice roll. Your goal is to design a stopping strategy to produce the largest possible expected reward. Describe this as an optimal stopping problem of the type considered in this chapter, with a Markov chain having an absorbing state at 1. Find the optimal value function V and use that to identify the optimal strategy.

Suppose someone proposes the "beat the mean" strategy: stop the first time that either $\phi(X_n) \geq E[\phi(X_1)]$ or $X_n = 1$. Is that optimal?

Repeat the problem using the final dice roll to the fourth power as the reward. Is the beat the mean strategy optimal here? (Modified from [25].)

. . . . . OSB

Problem 8.3
You are to roll a dice up to N times. After each roll you may stop and keep the value of your most recent roll, or (if you have not used up your N rolls) roll again. If you get as far as N rolls you must stop and accept the value of the N^{th} roll. You want to find the strategy which maximizes the expected value of your final roll. Certainly if you roll a 6 you should stop, and if you roll a 1 you should continue (if you can). For each of the other values $k = 2, 3, 4, 5$ there is a number m_k so that you should stop if $X_n = k$ and $N - n \leq m_k$ but roll again if $N - n > m_k$. Find the values of m_k.

. . . . . HW6D

Problem 8.4
Consider a game of chance which works as follows. If your winnings so far are $X_n (< 1000)$ and you play one more round you will with probability p win an additional $(1000 - X_n)/2$ so that $X_{n+1} = (1000 + X_n)/2$, or else with probability $1 - p$ you will lose everything so that $X_{n+1} = 0$. Thus your winnings will never reach or exceed 1000 but could get close if you have a run of good luck. If you stop playing you get to keep your current winnings. You are forced to stop playing once you loose everything. You want to decide when to stop playing so as to maximize the expected value of your winnings at the time you stop. It turns out that the optimal strategy is to stop playing the first time $X_n \geq \alpha$ but to keep playing as long as $X_n < \alpha$ for some threshold value α. Find the value of α. (You are not being asked to determine the full optimal value function $V(x)$ but what you are told about it above is enough to find α.)

. . . . . FE4

Problem 8.5
You are driving along a one-way street with parallel parking on one side. Suppose that the parking places are occupied with probability p, each parking space being independent of the others. Your destination is at parking space #100 and you want to park as close as possible to that. You start at parking space #1 and as you drive along you can only see the parking space closest to you; you can't see ahead. You can't turn around and go back. So as you encounter each empty parking space you must decide whether to park there or keep going and hope to find a vacant space closer to #100. If $\mathcal{T}$ is the number of the space you end up parking in, what should your strategy be to minimize $E[|\mathcal{T} - 100|]$? (Modified from [49].)

. . . . . Parking

Problem 8.6
Suppose X_n is a Markov chain (on a countable state space S) and $f : S \to \mathbb{R}$ is a bounded function. We know that there exists a smallest nonnegative excessive function v_0 which majorizes f. This means that v_0

is the smallest function so that for each $x \in S$ all the following inequalities hold:

$$v_0(x) \geq 0, \quad v_0(x) \geq f(x), \quad v_0(x) \geq \sum_{y \in S} p_{x,y} v_0(y). \tag{8.13}$$

a) Show that for each $x \in S$, one of the above inequalities must be an equality.

b) Suppose that v is some function so that for each $x \in S$ the inequalities (8.13) hold (for v) and one of them is an equality. (For different x, the equality may be for a different one of the three inequalities.) Give an example to show that it is not necessarily true that $v = v_0$.

c) Suppose in addition to the hypotheses of b) that the Markov chain is irreducible and recurrent, and that there is at least one x at which either $v(x) = 0$ or $v(x) = f(x)$. Prove that $v = v_0$. (Hint: Use Problem 1.)

S9Mid2

Problem 8.7
Suppose X_n is our standard symmetric random walk on $\mathbb{Z}$. Take the reward function $\phi(i) = 10e^{-i^2/100}$ and continuation cost $c(i) = |i|$. Find an optimal policy for the infinite horizon optimal stopping problem. (Observe that if $|X_n|$ is large enough you should stop immediately, because the cost of continuing for a single step is more than you can ever make up for from a future reward $\phi(X_\mathcal{T})$.)

RW

Problem 8.8
Consider the general optimal stopping problem as in Theorem 8.6.

a) Is it possible that $H = \emptyset$?

b) Show that if S is finite then H is not empty.

c) Explain why any finite, communicating, recurrent class must contain a point of H.

phimax

Problem 8.9
This problem explores the asymptotic behavior of the solution to the problem of optimal choice as $N \to \infty$. The special value of k depends on N so we will denote it by k_N here. It is the integer determined by

$$\frac{1}{N-1} + \cdots + \frac{1}{k_N} > 1 \geq \frac{1}{N-1} + \cdots + \frac{1}{k_N - 1}.$$

Use the inequalities

$$\int_m^N \frac{1}{x}\, dx < \frac{1}{N-1} + \cdots + \frac{1}{m-1} < \int_{m-1}^{N-1} \frac{1}{x}\, dx$$

to show that

$$\frac{N}{e} < k_N < \frac{N-1}{e} + 2 - \frac{1}{e}$$

and therefore

$$\frac{k_N}{N} \to \frac{1}{e} \text{ as } N \to \infty.$$

Also show that $\frac{1}{N-1} + \cdots + \frac{1}{k_N - 1}$ is between $\ln(N/k_N)$ and $\ln((N-1)/(k_N - 1))$, and therefore converge to $\ln(e) = 1$. What is $\lim_{N\to\infty} V(1,1)$?

AsymOpt

Problem 8.10
In the problem of optimal choice, Example 8.5, suppose we change the goal to offering the job to the most qualified applicant as possible, so you want to maximize $E[\rho(\mathcal{T})]$, the mean rank of the selected applicant. The interviewing rules stay the same.

a) Formulate the problem: specify an appropriate Markov chain and reward function to use.

b) Can you work out an optimal strategy? Try it for $N = 10$ first (using numerical calculations if necessary), and then see if you can work it out in general.

..... BestChoice

For Further Study

Some references on this material are Ferguson [23], Dynkin & Yushkevitch [20], Bertsekas [5], Puterman [49].

The same problems can be posed for continuous time processes, but there are more technicalities to deal with in that case. See Taylor [61] for a summary treatment of optimal stopping in that setting.

Chapter 9

Martingales

A martingale M_n is another type of stochastic process, characterized by different properties than a Markov process. A stochastic process X_n is Markov if the dependence of its future values X_{n+1} on the past $X_{0:n}$ only involves the most recent past, X_n. We have expressed this in the form

$$E[f(X_{n+1})|X_{0:n}] = \mathbf{P}f(X_n).$$

A martingale is characterized by the property that its mean future change, $M_{n+1} - M_n$, given the past, is 0:

$$E[M_{n+1} - M_n|X_{0:n}] = 0, \text{ or equivalently } E[M_{n+1}|X_{0:n}] = M_n.$$

There is no requirement of limited dependence on the past, as for Markov processes.

Why are martingales important? Here are three general reasons.

- Many important relationships are concisely described by saying the certain expressions are martingales. We will see this in Chapter 10, where all the pricing formulas reduce to the martingale property of the present value of various financial assets. Even the Markov property turns out to be equivalent to certain expressions being martingales, as we will see in Section 9.2 below. Thus martingales are valuable as a unifying concept.
- There is a substantial theory of martingales. This includes theorems about convergence $\lim_{n\to\infty} M_n$ (see Section 9.4) as well as a "calculus of martingales" (see Section 9.3) which describes how new martingales can be constructed from old martingales. In Section 9.6 we will see several applications of the convergence theory. And in Chapter 10 the calculus of martingales will play a vital role.
- Martingales are an essential part of contemporary stochastic process theory, a required topic for anyone who wants to be literate in the mathematics of stochastic processes.

9.1 Defining Martingales

We assume that there is an underlying process X_n. (For us it will be a Markov chain but it doesn't need to be.) Based on X_n many auxiliary processes can be constructed. Suppose for instance that X_n is the familiar symmetric random walk on $\mathbb{Z}$, and we define $M_0 = X_0^3$ and

$$M_n = X_n^3 - 3(X_0 + X_1 + \cdots + X_{n-1}) \text{ for } n \geq 1. \tag{9.1}$$

This is not itself a Markov process. One way to see this is to observe that the two possible values of M_{n+1} are

$$\begin{aligned} M_{n+1} &= (X_n \pm 1)^3 - X_n^3 - 3X_n + M_n \\ &= M_n \pm (1 + 3X_n^2). \end{aligned}$$

So the states that M_{n+1} can move to from M_n depend on more that just M_n itself; they depend on X_n as well. We can determine X_n from the full history $M_{0:n}$ but not from M_n alone. Said differently, $P(M_{n+1} = x|M_{0:n})$ is not M_n-determined; for the same M_n value this conditional probability will be different for different X_n values.

However M_n from (9.1) does have a different expectation property: its future increments have zero conditional mean given the past: $E[M_{n+1} - M_n|X_{0:n}] = 0$, or as it is usually presented,

$$E[M_{n+1}|X_{0:n}] = M_n.$$

For our example this holds because $E[X_{n+1}^3|X_{0:n}] = X_n^3 + 3X_n$. (Check that for yourself.) Using this we find that

$$\begin{aligned} E[M_{n+1}|X_{0:n}] &= E[X_{n+1}^3 - X_n^3 - 3X_n + M_n|X_{0:n}] \\ &= E[X_{n+1}^3|X_{0:n}] - X_n^3 - 3X_n + M_n \\ &= X_n^3 + 3X_n - X_n^3 - 3X_n + M_n \\ &= M_n \end{aligned}$$

Definition. *Given an underlying process X_n, a martingale is a real-valued process M_n which is integrable and satisfies*

$$E[M_{n+1}|X_{0:n}] = M_n$$

for each $n = 0, 1, 2, 3, \ldots$.

Some immediate consequences of the definition are as follows.

- M_n is $X_{0:n}$-determined for each n.
- $E[M_m|X_{0:n}] = M_n$ for all $n < m$.
- $E[X_n] = E[X_0]$ for all n.

Generalizing the definition, an integrable, $X_{0:n}$-determined process M_n is called a *submartingale* if

$$E[M_{n+1}|X_{0:n}] \geq M_n \text{ for all } n,$$

and a *supermartingale* if

$$E[M_{n+1}|X_{0:n}] \leq M_n \text{ for all } n.$$

For M_n to be a martingale is to be both a submartingale and a supermartingale.

Examples

The example in (9.1) above is just one of many martingales associated with the standard symmetric random walk X_n on $\mathbb{Z}$. Here are several others.

- $M_n = X_n$.
- $M_n = X_n^2 - n$.
- $M_n = \theta^{X_n}/\bar{\theta}^n$ where $\bar{\theta} = (\theta + \theta^{-1})/2$ (provided $\bar{\theta} \neq 0$).
- $Y_0 = 0$, $Y_n = 1 - 2^n$ if $X_1 = -1, X_2 = -2, \ldots, X_n = -n$ and 1 otherwise.
- $M_n = E[Z|X_{0:n}]$ for an integrable random variable Z.

9.2 Martingales and Markov Chains

Suppose that the underlying process X_n is a Markov chain with state space $\mathcal{S}$ and transition matrix $\mathbf{P}$. If $f : \mathcal{S} \to \mathbb{R}$ is a (bounded) function and $g : \mathcal{S} \to \mathbb{R}$ is the function determined from it by

$$g(s) = \mathbf{A}f(s)$$

then

$$M_n = f(X_n) - \sum_0^{n-1} g(X_k),$$

is a martingale. This is easy to check.

$$\begin{aligned}
E[M_{n+1}|X_{0:n}] &= E[f(X_{n+1})|X_{0:n}] - \sum_0^n g(X_k) \\
&= \mathbf{P}f(X_n) - \sum_0^n g(X_k) \\
&= f(X_n) + \mathbf{A}f(X_n) - \sum_0^n g(X_k) \\
&= f(X_n) + g(X_n) - \sum_0^n g(X_k) \\
&= f(X_n) - \sum_0^{n-1} g(X_k) \\
&= M_n.
\end{aligned}$$

The example (9.1) is just this using $f(n) = n^3$ for the random walk. The reasoning runs in reverse as well: if M_n as defined above is a martingale then it must be that

$$E[f(X_{n+1})|X_{0:n}] = \mathbf{P}f(X_n).$$

And if that is true for all (bounded) functions $f(\cdot)$ then it follows that for all $x_n, j \in \mathcal{S}$

$$P(X_{n+1} = j|X_{0:n} = x_{0:n}) = p_{x_n,j}.$$

This means that X_n is a Markov chain with transition matrix $\mathbf{P}$. We have proven the following martingale characterization of the Markov chain with transition matrix $\mathbf{P}$.

Theorem 9.1. *Suppose that X_n, $n = 0, 1, 2, \ldots$ is a stochastic process with values in a countable set $\mathcal{S}$ and $\mathbf{P}$ is a transition matrix on $\mathcal{S}$. Then X_n is a Markov chain with transition matrix $\mathbf{P}$ if and only if*

$$M_n = f(X_n) - \sum_0^{n-1} \mathbf{A}f(X_k),$$

is a martingale for every bounded function $f : \mathcal{S} \to \mathbb{R}$.

This characterization of X_n resembles a differential equation. One way to describe a solution $y(t)$ to $y'(t) = G(y(t))$ would be to say that for every differentiable function $f(x)$

$$f(y(t)) - \int_0^t g(y(s))\, ds \text{ is a constant}$$

where $g(y) = \langle \nabla f(y), G(y) \rangle$. In the Markov setting "is a constant" is replaced by "is a martingale" and instead of $g(y) = \langle \nabla f(y), G(y) \rangle$ we use $g(s) = \mathbf{A}f(s)$.

The theorem can be generalized in many ways. For instance we can include time as well as state dependence. If

$$\begin{aligned} g(s,n) &= \mathbf{P}f(s,n+1) - f(s,n) \\ &= \mathbf{A}f(s,n+1) + [f(s,n+1) - f(s,n)] \end{aligned}$$

then the following is a martingale

$$M_n = f(X_n, n) - \sum_0^{n-1} g(X_k, k).$$

This, in conjunction with Theorem 9.6 below, imply Lemma 8.1 of the previous chapter.

A multiplicative version is as follows. Suppose that $c \neq 0$ and

$$\mathbf{P}f(s) = c(s)f(s) \text{ for all } s \in \mathcal{S}.$$

Then

$$M_n = f(X_n) \prod_{k=0}^{n-1} \frac{1}{c(X_k)} \tag{9.2}$$

is a martingale, provided f and $1/c$ are bounded. The essential calculation is that

$$E[f(X_{n+1})/c(X_n)|X_{0:n}] = \mathbf{P}f(X_n)/c(X_n) = f(X_n).$$

(This is a discrete version of what is often called a Feynmann-Kac formula in continuous time settings.) Again the converse is true: if (9.2) is a martingale for all such f and c then X_n is Markov with transition matrix $\mathbf{P}$.

Example 9.1. Let X_n be the standard symmetric random walk on $\mathbb{Z}$ and $f(k) = \theta^k$. Then $\mathbf{P}f(k) = \bar{\theta}f(k)$ where $\bar{\theta} = (\theta + 1/\theta)/2$. This gives the third example in Section 9.1.

9.3 Discrete Stochastic Integrals

Imagine an i.i.d. sequence $X_1, X_2, \ldots$ with $X_n \geq 0$ and $1 = E[X_n]$. Think of these as the outcomes of a sequence of dice rolls or some other game of game of chance. A gambler is going to place wagers W_n on the outcomes of these games, specifically W_{n-1} is the wager placed on the nth game. This means that the gambler pays W_{n-1} before the game is played and then receives $W_{n-1}X_n$ after, for a net gain of $W_{n-1}(X_n - 1)$. The assumption that $E[X_n] = 1$ means that this is a fair game in the sense that constant wagers $W_n = 1$ will produce zero average net gain in the long run:

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} (X_i - 1) = 0.$$

(This is the Strong Law of Large Numbers.) Suppose the gambler starts with Y_0 in cash. His cash holdings after the nth game will be

$$Y_n = Y_0 + \sum_{i=1}^{n} W_{i-1}(X_i - 1).$$

Under the reasonable assumption that W_n depends only on $X_{0:n}$, the gambler's fortune Y_n is a martingale:

$$\begin{aligned} E[Y_{n+1}|X_{0:n}] &= E[Y_n + W_n(X_{n+1} - 1)|X_{0:n}] \\ &= Y_n + W_n E[(X_{n+1} - 1)|X_{0:n}] \\ &= Y_n + W_n E[(X_{n+1} - 1)] \text{ (by independence)} \\ &= Y_n + 0 = Y_n. \end{aligned}$$

An interesting example is the following doubling strategy. Suppose $X_n = 0$ or 2 with equal probabilities (flip a coin, heads=2, tails=0). We start with $Y_0 = 0$. Our goal is to win \$1. We wager $W_0 = 1$ on the first

play. If we win we have $Y_1 = 1$ and now we take all future wagers to be $W_n = 0$. But if we loose we have $Y_1 = -1$. We now double our wager to $W_1 = 2$ for the second play. If we win on the second play we have $Y_2 = 1$ and now make all future $W_n = 0$. If we lose the second play we will have $Y_2 = -3$. We double our bet again to $W_2 = 4$ for the third play. We keep doubling our wagers until the first time we win a play. This produces the fourth in our list of example martingales above. The strategy *will* eventually succeed: $Y_n = 1$ once n is large enough. But it is possible that Y_n will take large negative values before that happens; you may need to be a courageous gambler to ride out a string of bad luck and large gambling debts ($Y_n < 0$) before you eventually win and have $Y_n = 1$.

In general the martingale property says that $E[Y_n] = E[Y_0] = 0$ for even n. That means there is no betting strategy which will achieve the goal of $Y_n = 1$ with certainty in a fixed finite number of steps. For the doubling strategy in particular you have to be willing to play for an arbitrarily long time for the strategy to succeed with probability 1. The managers of a gambling casino would want to impose some kind of rules to prevent people from successfully playing such a doubling strategy. Otherwise they would have lots of people starting with $Y_0 = 0$ and leaving with 1, at the expense of the casino. So that raises the question of what rules could put in place to prevent it. We will come back to that in Section 9.6.1.

Let M_n be the fortune of a gambler who always wagers $W_n = 1$ every time:

$$M_n = M_0 + \sum_{i=1}^{n} (X_i - 1).$$

This is a martingale. The fortune Y_n of a gambler following a more complicated wagering scheme W_n can be expressed in terms of M_n as

$$Y_n = Y_0 + \sum_{k=1}^{n} W_{k-1}(M_k - M_{k-1}) \tag{9.3}$$

In fact starting with any martingale M_n and a sequence W_n of $X_{0:n}$-determined random variables, (9.3) will always produce a new martingale Y_n. (There are some technical conditions that are needed to justify this. Something needs to be assumed to insure that $W_{n-1}\Delta M_n$ is integrable.)

Theorem 9.2. *Assume M_n is a martingale (with $E[M_n^2] < \infty$ for each n) and W_k are $X_{0:n}$-determined random variables (with $E[W_k^2] < \infty$) then Y_n defined by* (9.3) *is also a martingale.*

Proof. Using the notation $\Delta M_n = M_n - M_{n-1}$, the argument is

$$\begin{aligned} E[Y_{n+1}|X_{0:n}] &= E[Y_n + W_n \Delta M_{n+1}|X_{0:n}] \\ &= Y_n + W_n E[\Delta M_{n+1}|X_{0:n}] \\ &= Y_n + W_n \cdot 0 \\ &= Y_n \end{aligned}$$

□

The construction (9.3) is called a *discrete stochastic integral* or sometimes a *martingale transform*. This always produces a new martingale Y_n from the original martingale M_n, for any "integrand" process W_n which for each n is $X_{0:n}$-determined. Such a stochastic process W_n is called a *non-anticipating* stochastic process. I like to think of W_n as associated with the time *interval* between $t = n$ and $t = n+1$; non-anticipating means W_n should be known at the start of that interval. We will see the continuous-time version of stochastic integrals in Chapter 12.

9.4 Martingale Convergence Theorems

Consider the example

$$M_n = E[Z|X_{0:n}]$$

described above (for any integrable random variable). If Z is dependent on the full sequence $X_{0:\infty}$ then $E[Z|X_{0:\infty}] = Z$. We might expect to obtain this in the limit as $n \to \infty$:

$$\lim_n M_n = \lim_n E[Z|X_{0:n}] \stackrel{?}{=} E[Z|X_{0:\infty}] = Z.$$

This is indeed the case. Even if a martingale does not come to us in the form $M_n = E[Z|X_{0:n}]$ it is often the case that $Z = \lim M_n$ does exist. There are certainly martingales which do not converge as $n \to \infty$; an example is the symmetric random walk $M_n = X_n$ itself. But if they are bounded in mean they do converge. Here are some basic results on convergence of martingales, stated without proof. (See "For Further Study" at the end of the chapter for references to proofs.)

Theorem 9.3. *Suppose M_n is a supermartingale with the property that $E[|M_n|]$ is bounded (i.e. there is a value $0 \le B < \infty$ with $E[|M_n|] \le B$ for all n). Then with probability* 1

$$\lim_{n\to\infty} M_n \text{ exists.}$$

Every martingale is a supermatringale. And if M_n is a submartingale then $-M_n$ is a supermatringale. So the theorem applies to martingales, submartingales, and supermartingales.

Observe that for a nonnegative supermartigale,

$$E[|M_n|] = E[M_n] \le E[M_0],$$

so the theorem applies.

Corollary 9.4. *If M_n is a nonnegative supermartingale, then $\lim_{n\to\infty} M_n$ exists with probability* 1.

Given that $M_\infty = \lim_{n\to\infty} M_n$ exists the next question is whether it is legitimate to let $m \to \infty$ in $M_n = E[M_m|X_{0:n}]$ to obtain

$$M_n = \lim_{m\to\infty} E[M_m|X_{0:n}] = E[\lim_{m\to\infty} M_m|X_{0:n}] = E[M_\infty|X_{0:n}].$$

The answer to this is "yes" if M_n satisfies a condition called "uniform integrability", which is a bit too technical for us to describe here. However a simple sufficient condition is the following.

Theorem 9.5. *Suppose M_n is a martingale with the property that $E[M_n^2]$ is bounded. Then $M_\infty = \lim_{n\to\infty} M_n$ exists with probability* 1*, has $E[M_\infty^2] < \infty$, and*

$$M_n = E[M_\infty|X_{0:n}].$$

This says that many martingales are of the form in the last bullet on page 141.

Example 9.2. Suppose X_n is an irreducible Markov chain and ψ is a nonnegative superharmonic function: $\mathbf{A}\psi \le 0$. Then it follows from Theorem 9.1 that $\psi(X_n)$ is a nonnegative supermartingale and so by the corollary must converge almost surely: $\psi(X_n) \to c$. If X_n is recurrent then it visits all states infinitely often, which means that ψ must be constant. In particular for a irreducible, recurrent chain all bounded harmonic functions are constant. If a nonconstant, nonnegative, superharmonic function exists for an irreducible chain then the chain must be transient. This is another result in the same category as Theorem 4.7.

Example 9.3. The doubling strategy above is an example when this theorem does *not* apply!

9.5 Optional Stopping

Martingales also interact nicely with stopping times. Recall that a stopping time is a time-valued random variable $\mathcal{T}$ with the property that

$$\{\mathcal{T} \le n\} \text{ is } X_{0:n}\text{-determined for each } n.$$

If M_n is a (sub- or super-) martingale we can define its stopped version

$$M_{\mathcal{T}\wedge n} = \begin{cases} M_n & \text{if } n < \mathcal{T} \\ M_{\mathcal{T}} & \text{if } \mathcal{T} \le n. \end{cases}$$

This is usually referred to as "optional stopping" (not to be confused with "optimal stopping").

Theorem 9.6. *If M_n is a (sub- or super-) martingale and $\mathcal{T}$ is a stopping time, then $M_{\mathcal{T}\wedge n}$ is also a (sub- or super-) martingale. Moreover, in the case that M_n is a submartingale*

$$E[M_{\mathcal{T}\wedge n}] \leq E[M_n],$$

or $\geq$ in the case of a supermartingale, or $=$ for a martingale.

Proof. The supermartingale and martingale cases follow easily from the submartingale case. Suppose that M_n is a submartingale. We want to show that

$$E[M_{\mathcal{T}\wedge(n+1)} \mid X_{0:n}] \geq M_{\mathcal{T}\wedge n}.$$

Write $M_{\mathcal{T}\wedge(n+1)} = M_{\mathcal{T}\wedge n} + 1_{\{\mathcal{T}>n\}} \cdot (M_{n+1} - M_n)$. As a consequence,

$$E[M_{\mathcal{T}\wedge(n+1)}|X_{0:n}] = M_{\mathcal{T}\wedge n} + 1_{\{\mathcal{T}>n\}}E[(M_{n+1} - M_n)|X_{0:n}] \geq M_{\mathcal{T}\wedge n} + 0.$$

For the other inequality asserted in the theorem write

$$\begin{aligned} M_{\mathcal{T}\wedge n} &= \sum_{k=0}^{n} 1_{\{\mathcal{T}=k\}} M_k \quad + 1_{\{\mathcal{T}>n\}} M_n \\ M_n &= \sum_{k=0}^{n} 1_{\{\mathcal{T}=k\}} M_n \quad + 1_{\{\mathcal{T}>n\}} M_n. \end{aligned}$$

So the key argument is that for each $k = 0, \ldots, n$

$$E[1_{\{\mathcal{T}=k\}} M_k] \leq E[1_{\{\mathcal{T}=k\}} E[M_n|X_{0:k}]] = E[E[1_{\{\mathcal{T}=k\}} M_n|X_{0:k}]] = E[1_{\{\mathcal{T}=k\}} M_n].$$

□

In the case of a martingale (so all the inequalities are equalities) it follows that $E[M_0] = E[M_{\mathcal{T}\wedge n}]$. We can ask what happens in the limit as $n \to \infty$. If $\mathcal{T} < \infty$ with probability 1, or $M_n \to M_\infty$ then

$$M_{\mathcal{T}\wedge n} \to M_{\mathcal{T}}.$$

It turns out that if Theorem 9.5 applies, i.e. if $E[M_n^2]$ is bounded, then the limit *can* be taken under the expectation and we obtain

$$E[M_0] = E[M_{\mathcal{T}}].$$

9.6 Applications

Now that we have laid out some of the basic properties, let's look at a couple applications.

9.6.1 Casino Policies

Consider again the martingale Y_n resulting from a betting strategy W_n applied to a game represented by a martingale M_n, i.e. Y_n is as in (9.3). Instead of making the wagers $W_n = 0$ after the gambler's goal is reached it is a little more natural to let $\mathcal{T}$ be the time the gambler achieves his goal and consider $Y_{\mathcal{T}}$ as the final result of the betting strategy. Let's consider the problem the casino faces with strategies such as the doubling strategy which produces $Y_0 < Y_{\mathcal{T}}$. Certainly they want this to be possible, meaning to have positive probability, because it is the prospect of winning more than you started with that attracts customers. But they don't want $E[Y_0] < E[Y_{\mathcal{T}}]$, because the mean of $E[Y_0 - Y_{\mathcal{T}}] < 0$ over many customers is what they would see in their long term profits (or losses). So what policies would prevent $E[Y_0] < E[Y_{\mathcal{T}}]$?

First of all we can assume that $\mathcal{T} < \infty$ with probability 1. No gambler can play forever. We know that $E[Y_0] = E[Y_{\mathcal{T}\wedge n}]$ and $Y_{\mathcal{T}\wedge n} \to Y_{\mathcal{T}}$. If the casino put a limit on the number of times the gambler can play, $\mathcal{T} \leq k$ for some fixed k, then by Theorem 9.6

$$E[Y_{\mathcal{T}}] = E[Y_{\mathcal{T}\wedge k}] = E[Y_0].$$

That would accomplish the casino's goals, but might be a bit too heavy-handed for customers to be encouraged to come.

One way to think about the problem is this. To avoid $E[Y_0] < E[Y_\mathcal{T}]$ the casino wants to be able to pass $\lim_{n\to\infty}$ through the expectation

$$E[Y_0] = \lim_{n\to\infty} E[Y_{\mathcal{T}\wedge n}]$$

to conclude that $E[Y_0] = E[Y_\mathcal{T}]$. Stated this way, the casino cares about the applicability of one of our convergence results, dominated or monotone convergence. We can't hope for monotone convergence; martingales cannot be monotone unless they are constant, and that wouldn't make for a very interesting gambling opportunity. Dominated convergence would require the existence of some integrable random variable Z with $|Y_{\mathcal{T}\wedge n}| \le Z$. This would mean cutting off the gambler's betting if they have won too much. But all the casino really needs is a lower bound on Y_n, it turns out. This is accomplished by requiring gamblers to play using chips or tokens. The gambler starts by converting his initial $\$Y_0$ to chips. All wagers must be made using the chips which the gamble holds. Wagers are paid up-front before each play, which insures that $W_n \le Y_n$. The gambler can never loose more than his wager, which insures $0 \le Y_n$. If the gambler runs out of chips he can't place any more wagers; he must stop. This system insures that $Y_n \ge 0$; the gambler can't go into debt by producing $Y_n < 0$. This will serve the casino's purpose because of Fatou's Lemma, Theorem 3.4. Consider $X_n = Y_{\mathcal{T}\wedge n}$. Observe that

$$\lim X_n = \lim Y_{\mathcal{T}\wedge n} = Y_\mathcal{T}$$

and

$$\lim E[X_n] = \lim E[Y_{\mathcal{T}\wedge n}] = \lim E[Y_0] = E[Y_0].$$

So Fatou's Lemma guarantees that

$$E[Y_\mathcal{T}] \le E[Y_0],$$

even better than equality from the casino's point of view! In other words there can be no betting strategy which achieves $Y_0 < Y_\mathcal{T}$ with certainty *and* which maintains $Y_n \ge 0$ for all n. Requiring all gamblers to use chips enforces $Y_n \ge 0$ (and probably has other benefits to the casino management) and so protects them from sure-fire winning strategies!

9.6.2 Branching Processes

The classical branching process was first introduced to study the extinction of a reproducing species. The idea is that $X_n \in \mathbb{N}$ represents the number of individuals of a certain species which exist in the nth generation. To make the transition $X_n \to X_{n+1}$ each of the X_n individuals gives birth to a random number of children (and then the parent dies). The numbers of offspring from each parent are independent, but with a common distribution $p_i = P(Y = i)$. So if $X_n = k$ then we take k independent copies of Y: $Y_1, \ldots, Y_k$ and form

$$X_{n+1} = \sum_{i=1}^{k} Y_i.$$

Provided $P(Y = 0) > 0$ it is possible for the population to become extinct, $X_{n+1} = 0$, in which case it remains extinct forever. In other words 0 is an absorbing state. The resulting X_n is a Markov chain, called a branching process or sometimes a Galton-Watson process. The basic theory is focused on the extinction probability assuming that $X_0 = 1$:

$$P(\mathcal{T}_0 < \infty),$$

where $\mathcal{T}_0$ is the first time $X_n = 0$.

We can learn a lot from martingale theory here. Consider

$$M_n = X_n/\mu^n.$$

This is the martingale from (9.2) using $f(x) = x$, since

$$\mathbf{P}f(n) = E[\sum_{1}^{n} Y_i] = n\mu = \mu f(n).$$

A simple consequence of $E[M_n] = E[M_0] = 1$ is that $E[X_n] = \mu^n$.

Consider the case of $\mu < 1$. If $\mathcal{T}_0$ is the first time $X_n = 0$, i.e. the time of extinction , then $P(\mathcal{T}_0 > n) \leq E[X_n] = \mu^n \to 0$ as $n \to \infty$. We conclude that $P(\mathcal{T}_0 < \infty) = 1$. Extinction is certain.

Suppose $\mu = 1$ and $p_1 = P(Y = 1) < 1$. In this case $M_n = X_n$. Theorem 9.3 implies that X_n converges as $n \to \infty$. But X_n is integer valued. The only way it can converge is if it is constant after some time n. In other words X_n must reach an absorbing state with probability 1. But if $p_1 < 1$ the only absorbing state is 0. It follows that X_n reaches 0 with probability 1. Extinction is certain.

Now consider that case of $\mu > 1$. Nonextinction is possible in this case. But Theorem 9.3 still says that

$$M_\infty = \lim_{n\to\infty} X_n/\mu^n$$

exists. In other words there is a limiting *normalized* population size. This means that asymptotically the population grows exponentially, with only the coefficient being random:

$$X_n \sim \mu^n M_\infty \text{ as } n \to \infty.$$

9.6.3 Stochastic Lyapunov Functions

The proofs we gave for Theorems 4.7 and 4.8 do not provide much insight into how properties of a solution to $\mathbf{A}\phi \leq 0$ are related to transience or recurrence. The connection between martingales and Markov chains allows us to present alternate arguments which may be a bit more satisfying intuitively. Suppose X_n is an irreducible Markov chain with an infinite state space. For technical reasons we will also assume that for each i there are only a finite number of j with $p_{i,j} > 0$, i.e. that there are only a finite number of possible transitions from each state. For simplicity lets suppose the state space is the integer lattice $\mathbb{Z}^d$ in d dimensions, so that we can talk about $|X_n|$. Suppose we can find a nonnegative function ϕ which satisfies

$$\mathbf{P}\phi(x) \leq \phi(x) \text{ for all } r < |x|.$$

This means the inequality is allowed to fail for a finite number of states. Then $\phi(X_n)$ will be a supermartingale, as long as $|X_n| > r$. That's because if $|X_n| > r$ then

$$E[\phi(X_{n+1})|X_{0:n}] = \mathbf{P}\phi(X_n) \leq \phi(X_n).$$

This may fail once $|X_n| \leq r$, but if we stop at the first time that happens,

$$\mathcal{T}_r = \text{ the first time that } |X_n| \leq r,$$

then $\phi(X_{\mathcal{T}_r \wedge n})$ will be a nonnegative supermartingale. This must have a limit as $n \to \infty$. If we also know something about $\lim_{|x|\to\infty} \phi(x)$ we will be able to make some deductions about the probability that $\mathcal{T}_r = \infty$, which is intimately related to the issue of recurrence.

Let's define a second stopping time,

$$\mathcal{T}_R = \text{ the first time that } |X_n| \geq R,$$

where $r < R$. We must have

$$\mathcal{T}_R \to \infty \text{ as } R \to \infty. \tag{9.4}$$

This is because the chain cannot jump to ∞ in a finite number of steps; for any n

$$P_x(\max(|X_1|, |X_2|, \ldots, |X_n|) \geq R) \to 0 \text{ as } R \to \infty.$$

Since $P_x(\mathcal{T}_R \leq n) = P_x(\max(|X_1|, |X_2|, \ldots, |X_n|) \geq R)$ we see that $\lim_{R\to\infty} \mathcal{T}_R \leq n$ has probability 0 for each n. Thus (9.4) holds with probability 1.

The set of x with $r < |x| < R$ is a finite set, and assuming the chain is irreducible we know

$$P_x(\mathcal{T}_r \wedge \mathcal{T}_R < \infty) = 1. \tag{9.5}$$

That's because the chain must eventually jump out of the finite set of $r < |x| < R$. Now we have a *bounded* supermartingale $\phi(X_{\mathcal{T}_r \wedge \mathcal{T}_R \wedge n})$. (This where we are using the assumption that for each i the set $\{j : p_{ij} > 0\}$ is bounded.) It follows that

$$\phi(x) \geq \lim_{n\to\infty} E_x[\phi(X_{\mathcal{T}_r \wedge \mathcal{T}_R \wedge n})] = E_x[\phi(X_{\mathcal{T}_r \wedge \mathcal{T}_R})].$$

If we have lower bounds

$$\begin{aligned} 0 < B_r \leq \phi(x) \text{ for all } |x| \leq r \\ B_R \leq \phi(x) \text{ for all } |x| \geq R, \end{aligned}$$

(note that $B_r > 0$ is slightly more than we get from the hypotheses of Theorem 4.7) then we can say

$$B_r P_x(\mathcal{T}_r < \mathcal{T}_R) + B_R P_x(\mathcal{T}_r > \mathcal{T}_R) \leq \phi(x). \tag{9.6}$$

From here the conclusions of Theorems 4.7 and 4.8 are pretty easy.

Suppose $\phi(x) \to 0$ as $|x| \to \infty$. Equation (9.6) implies

$$P_x(\mathcal{T}_r < \mathcal{T}_R) \leq \phi(x)/B_r.$$

Letting $R \to \infty$ in this we find

$$P_x(\mathcal{T}_r < \infty) \leq \phi(x)/B_r.$$

Since the right side $\to 0$ as $|x| \to \infty$ we see that it is not possible that $P_x(\mathcal{T}_r < \infty) = 1$ for all x, as would be the case if the chain were recurrent. We conclude that the chain must be transient.

Suppose instead that $\phi(x) \to \infty$ as $|x| \to \infty$. Using the other half of (9.6) we have that

$$B_R P_x(\mathcal{T}_R < \mathcal{T}_r) \leq \phi(x),$$

and therefore

$$P_x(\mathcal{T}_R < \mathcal{T}_r) \leq \phi(x)/B_R.$$

Since $B_R \to \infty$ as $R \to \infty$ it follows that

$$\lim_{R\to\infty} P_x(\mathcal{T}_R < \mathcal{T}_r) = 0.$$

From here it follows that

$$P_x(\mathcal{T}_r < \infty) = 1.$$

This means that the chain must be recurrent.

9.7 Change of Measure and Martingales

There is one more important role of martingales that we should introduce here. In some circumstances there are reasons to consider more than one assignment of probabilities. In terms of the Kolmogorov model $P(A)$ denotes the probability assigned to events $A \subseteq \Omega$. There may be a second way of assigning them, $Q(A)$, that we want to consider as well. We will encounter this in Chapter 10 and again in Chapter 12, where market prices of contingent claims are associated with a different probability assignment Q than the original P which describes the stock price evolution.

One way an alternate probability assignment Q can be described is this: take a nonnegative random variable $Z \geq 0$ with $E[Z] = 1$. Use it to define $Q(A)$ for $A \subseteq \Omega$ by

$$Q(A) = E[Z; A].$$

We can now check that $Q(A)$ *is* a legitimate assignment of probabilities, in that all the properties listed on page 33 are satisfied. Every random variable X now has *two* means, $E[X]$ calculated using the original P, and $E^Q[X]$ calculated using the new probability Q. But it is easy to convert:

$$E^Q[X] = E[XZ].$$

A process M_n can be a martingale with respect to Q while not a martingale with respect to P. That is because $E[M_{n+1}|X_{0:n}]$ and $E^Q[M_{n+1}|X_{0:n}]$ will be different in general, so they can't both be $= M_n$. The technical issue we will face in Chapter 12 is how Q-martingales are related to P-martingales. For starters we know that

$$\zeta_n = E[Z|X_{0:n}]$$

Is a P-martingale. Now suppose M_n is a Q-martingale. If we go back to the definition of generalized conditionals, the martingale property says that for any $X_{0:n}$-determined event A

$$E^Q[M_{n+1}; A] = E^Q[M_n; A].$$

Now

$$E^Q[M_n; A] = E[M_n Z 1_A] = E[E[M_n Z 1_A|X_{0:n}]] = E[E[Z|X_{0:n}]M_n 1_A] = E[\zeta_n M_n 1_A] = E[\zeta_n M_n; A].$$

The same argument (conditioning on $X_{0:n+1}$ instead) shows that $E^Q[M_{n+1}; A] = E[M_{n+1}\zeta_{n+1}; A]$. So we see that the Q-martingale property of M_n is equivalent to

$$E[M_{n+1}\zeta_{n+1}; A] = E[\zeta_n M_n; A]$$

for all $X_{0:n}$-determined events A. This means that

$$E[M_{n+1}\zeta_{n+1}|W_{0:n}] = M_n\zeta_n,$$

namely that $M_n\zeta_n$ is a P-martingale. What we have found is that for M_n to be a Q-martingale is equivalent to $M_n\zeta_n$ being a P-martingale! We will see this at work in Chapter 12.

Problems

Problem 9.1
Suppose M_n is a martingale (and each M_n is square-integrable) and let $\Delta M_n = M_n - M_{n-1}$. Although

$$\Delta(M_n^2) = M_n^2 - M_{n-1}^2 \quad \text{and} \quad (\Delta M_n)^2 = (M_n - M_{n-1})^2$$

are not the same, show that

$$E[\Delta(M_n^2)|X_{0:n-1}] = E[(\Delta M_n)^2|X_{0:n-1}].$$

In fact the following are both martingales:

$$A_n = M_n^2 - \sum_1^n (\Delta M_k)^2,$$

$$B_n = M_n^2 - \sum_1^n \overline{(\Delta M_k)^2},$$

where

$$\overline{(\Delta M_k)^2} = E[(\Delta M_k)^2|X_{0:n-1}].$$

Show in particular that

$$E[M_n^2] = E[M_0^2] + \sum_1^n E[(\Delta M_k)^2]. \tag{9.7}$$

MQV

Problem 9.2
Let X_n be the asymmetric random walk on $\mathbb{Z}$ for which with $X_{n+1} = X_n + 1$ has probability p and $X_{n+1} = X_n - 1$ has probability $q = 1 - p$. Assume $0 < p < 1$ and $p \neq 1/2$. Find a value θ so that θ^{X_n} is a martingale.

Let N be a positive integer and suppose $0 < X_0 < N$. Let $\mathcal{T}_N$ be the first time that $X_n = 0$ or $X_n = N$. Explain why $M_n = \theta^{X_{n\wedge\mathcal{T}_N}}$ is a bounded martingale. Taking for granted that $\mathcal{T}_N < \infty$ with probability 1, explain why optional stopping implies

$$\theta^{X_0} = \theta^0 P_{X_0}(X_{\mathcal{T}_N} = 0) + \theta^N P_{X_0}(X_{\mathcal{T}_N} = N).$$

Deduce a formula for $P_{X_0}(X_{\mathcal{T}_N} = N)$ in terms of p and q. If $p \neq 1/2$ you can use the same martingale to *prove* that $\mathcal{T}_N < \infty$ with probability 1 – explain how.

ASYMRW

Problem 9.3
Let $Z_n = (X_n, Y_n)$ be the standard two-dimensional random walk on $\mathbb{Z}^2$. Show that

$$M_n = X_n^2 + Y_n^2 - n$$

is a martingale. (This is an instance of the martingale at the bottom of page 142. What can you conclude about $E_{(x_0,y_0)}[|Z_n|^2]$? Let $\mathcal{T}_r$ be the first time that $|Z_n| \geq r$. Use optional stopping to produce an upper bound on $E_{(0,0)}[\mathcal{T}_r]$. (Hint: $|Z_{\mathcal{T}_r}| \leq r + 1$.)

RW2

Problem 9.4
Suppose you are gambling on a sequence of fair coin tosses, represented as in Section 3.3 by an i.i.d. sequence with $P(X_i = 0) = P(X_i = 2) = 1/2$. You start with $Y_0 = 100$ and plan to gamble until the first time $\mathcal{T}$ when Y_n is either 0 or 1000. Until that happens (i.e. while $n < \mathcal{T}$) you follow some wagering strategy for which your wagers W_n satisfy $1 \leq W_n \leq \min(Y_n, 1000 - Y_n)$. This means you don't give up until $\mathcal{T}$ happens, you never wager more than your current cash holding Y_{n-1} or more than necessary to reach your goal of $Y_n = 1000$ on the next game. (And of course W_n should be $X_{0:n}$-determined.) Explain why any such wagering strategy does eventually terminate ($P(\mathcal{T} < \infty) = 1$), and the probability of success is $P(Y_{\mathcal{T}} = 1000) = 1/10$, *regardless of strategy*. (Hint: Use optional stopping again.)

FairHouse

Problem 9.5
Let S_n be our usual simple, symmetric random walk on $\mathbb{Z}$, starting at $S_0 = 0$. Given $K \neq 0$ let

$$\mathcal{T} = \inf\{n \geq 0 \,:\, S_n = K\}$$

be the first time that the random walk reaches K. We know that S_n is a martingale and $\mathcal{T}$ is a stopping time.

a) Explain why the optional stopping result $E[S_0] = E[S_{\mathcal{T}}]$ is *false*. (This is elementary.)

b) Explain why $E[S_n^2]$ is *not* bounded, so that the sufficient condition mentioned at the end of Section 9.5 is not satisfied.

S9G

For Further Study

There are many nice introductory treatments of martingales. We have drawn material from Grimmet & Stirzaker [25], Rogers & Williams [51]. In particular the convergence theorems of Section 9.3 are discussed in Sections II.48, 49 of [51]. See also §5.5 of Durrett [18]. More references are indicated at the end of Chapter 11.

The standard treatment of branching processes is based on generating functions. See for instance Example 5.1.1 in Norris [45], Section 5.4 in Grimmiet and Stirzaker [25], Example 4.4 in Varadhan [63], or Section 6.4 of Whittle [65].

Chapter 10

Mathematical Finance in Discrete Time

A financial market allows people and organizations to buy and sell various types of assets and to form binding agreements (contracts) under which the participants agree to exchange of assets under terms which are contingent on future market conditions. Familiar examples of financial assets are stocks and bonds. Examples of contingent contracts are financial derivatives and options. Virtually every asset has a price, determined by the market to be the price at which buyers and sellers agree to trade. This applies to contracts as well. These prices change over time, often in seemingly random ways. Market participants often have long term goals that they hope to achieve by means of their transactions. They need strategies to guide them in pursuit of their goals. Mathematical finance tries to develop models to help design such strategies. Given the apparent randomness of the evolution of prices it is natural that these models involve stochastic processes.

In this chapter we will consider a Markov chain model which is relatively simple, but which allows us to exhibit some of the ideas which are important in mathematical finance. There will be a single stock, whose share price S_n evolves over time according to a random walk. There will also be a bank account whose value B_n evolves deterministically in accord with a fixed interest rate. In this setting we want to study the price of a financial derivative or *contingent claim*, examples of which are described below. We will revisit some of the same issues in a continuous time setting in Chapter 12.

10.1 Stocks and Bonds

To own a share of stock in a particular company is simply to own part of the company itself. If the company does well your share will become more valuable and others will be more eager to buy it from you and willing to pay a higher price. Thus its market price will go up. If the company does poorly the market price of your share will decrease. Initially a company will sell shares to raise capital to get the business going. After that shares are bought and sold on the stock market and the price is determined solely by that market; the price of a share is the price at which buyers and sellers are currently agreeing to buy and sell. That can vary from day to day for all kinds of reasons, both material as well as psychological. There are other complicating features. There can be different "classes" of shares that entitle their owners to different voting privileges in company decision making. Some companies pay their stockholders dividends, essentially passing some of the company's profits back to the owners. Other companies don't pay dividends. We won't try to handle all those complexities. We will just describe the price per share of a particular stock as a Markov chain S_n, with no dividend payments. The value of S_n is the market price, meaning that at time n shares of the company can be bought or sold for $\$S_n$ apiece. To keep things simple we do not consider any sales charges, transaction costs, or brokerage fees.

We will assume that S_n follows a random walk through a set of possible (positive) prices $\cdots < s_i < s_{i+1} < \cdots$ with transition probabilities

$$P(S_{n+1} = s_{i+1} \,|\, S_n = s_i) = p_i \text{ and } P(S_{n+1} = s_{i-1} \,|\, S_n = s_i) = 1 - p_i, \tag{10.1}$$

for a set of $0 < p_i < 1$. In this finance literature this is sometimes called a "binary tree". Notice that we allow only two possible transitions: $s_i \to s_{i+1}$ and $s_i \to s_{i-1}$. There is a reason for keeping the structure so simple. See Problem 10.4.

A bond is essentially an I.O.U. issued by a company or government. It is a promise to pay a specified amount to the bond holder on a specified date of maturity. The specified payment amount is called the "face value" or "par value". Companies and governments issue bonds as a way of borrowing money. They *must* pay it back on the specified date (else no one will buy their bonds in the future, thus ruining their chances of raising more money, not to mention all the legal consequences). Some bonds make a series of interest payments in addition to the final payment. Like stocks there is a market where bonds can be bought and sold, again at market-determined prices. The stochastic modeling of a bond's market price B_t is more complicated than for stocks, because the random evolution of B_t prior to maturity has to lead to a deterministic final value B_T.

We are not going to consider a stochastic model for bond prices. Instead our "bond" will be completely predictable. Its market value at time n will be

$$B_n = (1+r)^n,$$

where r is a constant called the *interest rate per period*. In particular the price at time $n = 0$ is $B_0 = 1$. If the date of maturity is T, then its face value is $B_T = (1+r)^T$. This is really more like a fixed-rate certificate of deposit than a real-world bond. It functions as the bank account in our considerations below and so we will refer to it as *the bank*. It will be convenient to refer to the value invested in the bank in terms of the number of deposit certificates rather than cash value. One certificate is simply the value at time n of an initial deposit of $B_0 = 1$, so a certificate is worth B_n at time n. To put \$100 in the bank at time n is to buy $y = 100/B_n$ certificates. If I keep those until time $n+6$ and sell them I will recover a cash amount of

$$yB_{n+6} = \frac{100}{B_n} B_{n+6} = 100(1+r)^6.$$

This is the original \$100 plus 6 time periods of compounded interest. We will call this "buying y deposit certificates". Using the terminology of certificates rather than value allows us to describe both investments, stock and bank, in a common format. In our model the bank provides a safe or risk-free asset, in contrast to the stock which is the risky asset. The interest rate r is often called the *risk free* rate for this reason.

10.2 Contingent Claims

In addition to the basic assets S_n and B_n we want to consider financial contracts that are dependent on future asset prices. For example, suppose S_n represents the market price at time n of a bushel of wheat. If I am a baker I may want to arrange for my next year's supply of wheat, and I want to establish *now* what price I will have to pay in the future. To do this I could sign a *forward contract* with a wheat supplier. Under this contract the supplier would promise to deliver a certain quantity of wheat (lets say 1000 bushels) next year (on date T) and I promise to pay him \$$K$ on delivery, where K is a value we write into the contract *today*. I may have to pay the wheat producer something up front to get them to sign that contract, especially if K is low. If after the contract has been signed wheat prices rise, my contract becomes increasingly valuable, so some third party may want to buy it from me. If the price of wheat falls, I may wish I was not bound to pay the price required by the contract and may want to sell it to someone else. The basic problem we are interested in is how to determine the market price for this contract.

If I'm really worried about the price of wheat falling, a different contract arrangement might be more attractive. Imagine a contract that does not *require* me to buy 1000 bushels of wheat for \$$K$ but only guarantees that I can *if I want to*. Then if wheat prices are low on the exercise date T I can simply ignore the contract and buy at the lower market price. This kind of contract is called a *call option*. The value of this contract to me at time T is

$$\begin{cases} 1000S_T - K & \text{if } 1000S_T - K \geq 0 \\ 0 & \text{if } 1000S_T - K < 0, \end{cases}$$

where S_T is the price per bushel of wheat at time T. The value of 0 on the second line is because in those circumstances I am better off buying at the market price and not using the contract.

In general the time T at which a contract like this matures is called the *exercise time*. The value of the contract at the exercise time all be

$$V_T = \phi(S_T)$$

for some function $\phi(s)$ of the final stock price, called the *exercise value function*. (Actually this is called a *European* option. So-called *American* options are a bit different; see Section 10.6.5 below.) Different choices of the function $\phi(s)$ correspond to different types of options.

1. $\phi(s) = s - K$, a *forward contract*;
2. $\phi(s) = \begin{cases} s-K & \text{if } s-K \geq 0 \\ 0 & \text{if } s-K < 0 \end{cases}$, a *call option*;
3. $\phi(s) = \begin{cases} 0 & \text{if } s-K \geq 0 \\ K-s & \text{if } s-K < 0 \end{cases}$, a *put option*;
4. $\phi(s) = s$, the stock itself;
5. $\phi(s) = K$, the bank account itself.

There are other $\phi(s)$ which are of interest in finance. Some of them will be mentioned in Section 10.5 below. Mathematically we can consider any function for $\phi(s)$. We want to consider the pricing problem: assuming that the contract can be bought and sold just like other marketable assets, how can we determine the market price V_n of the contract at times $n < T$ prior to expiry? We expect this to be a $S_{0:n}$-determined stochastic process but we want to understand how it works in more detail.

10.3 No-Arbitrage Pricing

Prices are determined by what rational people will do in the market. Let's start with an extremely simple version of the pricing problem to see how this works out.

10.3.1 A Single Branch

Consider just two times: $n = 0$ and $n = 1(= T)$. The stock price is initially $S_0 = s_0$. At time $n = 1$ there are two possible values, $S_1 = s_{\pm 1}$ with probabilities p (for $S_1 = s_{+1}$) and $1 - p$ (for $S_1 = s_{-1}$). We assume

$$s_{-1} < s_{+1} \text{ and } 0 < p < 1.$$

A certificate of deposit purchased today for $B_0 = 1$ will be worth $B_1 = 1 + r$ tomorrow. The value of the contract tomorrow is $\phi(S_1) = V_1$, which will have two possible values depending on which way the stock price goes. Here is a specific example.

Example 10.1. Take $s_0 = 20$, $s_{-1} = 10$, $s_{+1} = 30$, $p = 1/2$, and $r = .05$. For the exercise value function take $\phi(s) = s - 20$. This is what we called a forward contract above. The final value of our contract is $V_1 = \phi(S_1) = S_1 - 20$, which depends on which value S_1 turns out to have. What is its value V_0 today? A (naive) guess might be that

$$V_0 = \frac{1}{1.05} E[V_1] = \frac{1}{1.05}\left[\frac{1}{2}(30-20) + \frac{1}{2}(10-20)\right] = 0,$$

the mean present value of tomorrow's contact value. If that were right this contract should be free in today's market; people would be willing to enter into this contract with each other without any money changing hands. If so I could take advantage of this market in the following way.

- I get one of these contracts at no cost.

- I sell 1 share of stock in the market, producing $\$s_0 = \$20$.
- I invest my \$20 in the bank, i.e. buy 20 certificates at $B_0 = 1$ each.

The contract requires that tomorrow I must pay \$20 and will receive one share of stock. Its value then will be $S_1 - 20$, which might turn out to be either positive or negative; we won't know until tomorrow. I gained \$20 by selling the share of stock today but I immediately invested it in the bank. So my net spending today is \$0. This has cost me nothing so far.

Tomorrow I use my contract; I pay the \$20 and receive one share of stock (which replaces the one I sold yesterday). My investment in the bank has grown in value because of the interest rate; it's value today is

$$20B_1 = 20(1.05) = 21.$$

This covers the \$20 I just paid you with \$1 to spare. I have made \$1 through this transaction and otherwise have exactly the same assets as I did yesterday before all this started. It doesn't matter whether S_1 is 30 or 10; it comes out the same either way. I made \$1 with no risk at all!

If I were greedy I could have acquired a thousand of those free contracts and carried out the above plan a thousand times over, to make a cool \$1000 overnight. Of course I wouldn't be the only person to see this free money opportunity. People would be clamoring to get these free contracts, and no one would be willing to issue them, at least not without being paid something for it. Obviously this contact is not really worthless and the market would be telling us that. Our guess that $V_0 = E[V_1]/(1+r)$ is wrong!

So what is the correct value V_0 of this contract today? To answer this we can design a strategy using just S_n and B_n which will leave us with exactly $V_1 = S_1 - 20$ tomorrow, the same as the contract would, and then ask how much it would cost today to execute that strategy. The strategy will be to buy one share of stock today and "borrow" some amount of money from the bank, so that tomorrow the amount "owed" to the bank comes to exactly 20. The amount to borrow today is $20/1.05 = 19.0476\cdots$. The value today of this combination is

$$S_0 - 19.0476\cdots = .9524\cdots$$

In other words starting with \$0.9524 today, I can borrow the remaining \$19.0476 I need to buy one share of stock and my strategy will be in place. I spend \$0.9524 today and have exactly $V_1 = S_0 - 20$ tomorrow. Thus

$$V_0 = \$0.9524\cdots.$$

The reason I could make money if the contract were available for free is that I could get for free something that was actually worth \$0.9524. I just capitalized on that, turning \$0.9524 today into $0.9524 \cdot 1.05 = \$1$ tomorrow.

In real-world forward contracts the value of K in our $\phi(s) = s - K$ is adjusted so that the contract really *is* worth $V_0 = 0$ today. In our example here $K = 20$ is not that value. Rethinking the above, we see that for my strategy to produce no net gain or loss, the correct value of K must be $K = S_0(1+r) = 21$.

The key idea from this example is that we can "replicate" tomorrow's $V_T = \phi(S_T)$ with a certain combination of stock and deposit certificates which we purchase *today*. Let's work this idea out more generally in our single branch setting. Consider a "portfolio" consisting of α shares of stock and β deposit certificates. The idea is to choose α and β so that the portfolio's value tomorrow is the same as $\phi(S_1)$ for *both* possible values of S_1. The value tomorrow of this portfolio will be

$$\alpha S_1 + \beta B_1 = \alpha s_{\pm 1} + \beta(1+r),$$

depending on whether $S_1 = s_{\pm 1}$. For the portfolio to match the option value V_1 in both cases α and β must be chosen to solve the following equations:

$$\begin{aligned} \alpha s_{+1} + \beta(1+r) &= \phi(s_{+1}) \\ \alpha s_{-1} + \beta(1+r) &= \phi(s_{-1}). \end{aligned}$$

Solving for α and β results in

$$\alpha = \frac{\phi(s_{+1}) - \phi(s_{-1})}{s_{+1} - s_{-1}}, \quad \beta = \frac{s_{+1}\phi(s_{-1}) - s_{-1}\phi(s_{+1})}{(1+r)(s_{+1} - s_{-1})}.$$

We will call this the *replicating portfolio* for the contingent claim with exercise value $\phi(S_1)$. To buy this portfolio today ($t = 0$) costs

$$c = \alpha S_0 + \beta B_0 = \frac{\phi(s_{+1}) - \phi(s_{-1})}{s_1 - s_{-1}} s_0 + \frac{s_1\phi(s_{-1}) - s_{-1}\phi(s_{+1})}{(1+r)(s_1 - s_{-1})} 1.$$

This value c *must* be the market value V_0 of the contract today. To see why, consider what I can do if V_0 is different from c. Suppose $V_0 > c$. Then I will sell someone a copy of the contract, i.e. they pay me V_0 and I promise to deliver V_1 to them tomorrow. Now I use c of what they just paid me to buy a copy of my replicating portfolio, and put the difference of $V_0 - c > 0$ in the bank. Tomorrow my replicating portfolio is worth exactly enough for me to fulfill my obligation to the person I sold the contract to. But I now have a bank investment worth $(1+r)(V_0 - c)$. That's risk-free profit for my being smart enough to see this strategy.

What if $V_0 < c$? Now I can make risk-free profit another way. This time I *sell* α shares of stock and β deposit certificates. That brings in c in revenue. I take V_0 of this and buy a copy of the contract, leaving $V_0 - c > 0$ in cash which I put in the bank. Tomorrow I cash in my option which pays me V_1, exactly enough for me to buy back at today's prices the α and β shares of stock and deposit certificates, but still leaving me the value $(1 + r)(V_0 - c)$ from the bank shares I bought with the yesterday's leftover cash. I'm right back where I started yesterday in terms of stock and bank investments but plus $(1 + r)(V_0 - c)$ in risk-free profit.

Each of these scenarios for $V_0 \neq c$ constitute an *arbitrage opportunity*, a way to buy and sell assets on the market consisting of S_t, B_t, and V_t to make a guaranteed profit. **Our basic premise is that the market sets the prices so that no arbitrage opportunities exist.** If such an opportunity did exist there would be a huge demand for the undervalued assets, which would bring their prices up until until the arbitrage opportunities are eliminated. If S_t and B_t are the correct market prices then the option price at $t = 0$ must be $V_0 = c$. The market enforces the no-arbitrage principle of prices, which is a sort of balance among all the prices in the market.

We will call (α, β) an *arbitrage portfolio* if its value $V_t = \alpha S_t + \beta B_t$ has these properties.

1. $V_0 = 0$,
2. $P(V_1 \geq 0) = 1$, and
3. $P(V_1 > 0) > 0$.

In other words an arbitrage portfolio is a strategy with which we can start from nothing, have no chance of loosing money and a positive probability of making money. What we considered above was actually a portfolio in the market with *three* assets, S_t, B_t, and V_t. Could there be arbitrage opportunities in our simple single-branch model consisting of just S_t and B_t alone? The answer should be "no" for any reasonable model.

Lemma 10.1. *There are no arbitrage portfolios for the single-branch model S_t, B_t described above if and only if*

$$s_{-1} < (1 + r)s_0 < s_{+1}.$$

Proof. Recall that we are assuming $0 < p < 1$.

Suppose that α, β is an arbitrage portfolio. First observe that either α or β must be nonzero; otherwise $V_1 > 0$ would be impossible. Since $V_0 = 0$ and both S_0 and B_0 are positive, *both* $\alpha \neq 0$ and $\beta \neq 0$ are necessary for an arbitrage portfolio. That $V_0 = 0$ implies

$$\frac{\beta}{\alpha} = -\frac{S_0}{B_0}.$$

Therefore

$$V_1 = \alpha S_1 + \beta B_1 = \alpha \left(S_1 - \frac{B_1}{B_0} S_0 \right) = \alpha \left(S_1 - (1 + r)S_0 \right)$$

Now consider the possibility of $\alpha > 0$. To be an arbitrage portfolio we must have *both* $s_{\pm 1} - (1+r)s_0 \geq 0$. Looking just at the smaller of these, if there is an arbitrage portfolio with $\alpha > 0$ then

$$s_{-1} - (1 + r)s_0 \geq 0. \tag{10.2}$$

Conversely suppose (10.2) holds. Consider the portfolio with $\alpha = 1$ and $\beta = -s_0$. We see that (10.2) implies that $V_1 \geq 0$ in both cases. With probability $p > 0$ have

$$V_1 = S_1 - (1+r)s_0 = s_{+1} - (1+r)s_0 > s_{-1} - (1+r)s_0 \geq 0,$$

so that $P(V_1 > 0) > 0$. Thus there *is* an arbitrage opportunity. Thus (10.2) is equivalent to the existence of an arbitrage portfolio with $\alpha > 0$.

Next consider the possibility of $\alpha < 0$. Arguing similarly to the above we must have *both* $s_{\pm 1} - (1+r)s_0 \leq 0$ for such an arbitrage portfolio to exist. Thus

$$s_{+1} - (1+r)s_0 \leq 0 \tag{10.3}$$

is necessary. Conversely if (10.3) holds the we can take $\alpha = -1$ and $\beta = s_0$ to produce a portfolio with $V_1 \geq 0$ and

$$V_1 = -(s_{-1} - (1+r)s_0) > -(s_{+1} - (1+r)s_0) \geq 0,$$

with probability $1 - p > 0$, so that we do indeed have an arbitrage portfolio.

Combining our conclusions we see that the nonexistence of arbitrage portfolios is equivalent to

$$s_{-1} < (1+r)s_0 < s_{+1},$$

as claimed. □

Suppose that in addition to S_t and B_t we allow a portfolio $V_t = \alpha S_t + \beta B_t$ constructed from them to be traded directly, i.e. as a new market asset. Could we create arbitrage opportunities in this way? A portfolio in this 3-asset market would be something of the form

$$\begin{aligned}\hat{V}_t &= \hat{\alpha} S_t + \hat{\beta} B_t + \gamma V_t \\ &= \hat{\alpha} S_t + \hat{\beta} B_t + \gamma(\alpha S_t + \beta B_t) \\ &= (\hat{\alpha} + \gamma\alpha) S_t + (\hat{\beta} + \gamma\beta) B_t.\end{aligned}$$

In other words any portfolio in the new 3-asset market is really the same as some portfolio in the original 2-asset market. So if there were no arbitrage portfolios in the 2-asset market there can't be any in the 3-asset market either.

Lemma 10.2. *If the single-branch market consisting of B_t and S_t is free of arbitrage and (α, β) is a portfolio with value V_t, then the market consisting of B_t, S_t and V_t is free of arbitrage.*

Let's look again at the formulas we found for the portfolio which replicated the contingent claim $\phi(S_1)$. The market value of the claim at $t = 0$ is given by

$$\begin{aligned}V_0 &= \alpha S_0 + \beta B_0 \\ &= \frac{\phi(s_{+1}) - \phi(s_{-1})}{s_{+1} - s_{-1}} S_0 + \frac{s_{+1}\phi(s_{-1}) - s_{-1}\phi(s_{+1})}{(1+r)(s_{+1} - s_{-1})} B_0 \\ &= \frac{1}{1+r}\left[(1+r)\frac{\phi(s_{+1}) - \phi(s_{-1})}{s_{+1} - s_{-1}} s_0 + \frac{s_{+1}\phi(s_{-1}) - s_{-1}\phi(s_{+1})}{s_{+1} - s_{-1}}\right] \\ &= \frac{1}{1+r}\left[\left(\frac{(1+r)s_0 - s_{-1}}{s_{+1} - s_{-1}}\right)\phi(s_{+1}) + \left(\frac{s_{+1} - (1+r)s_0}{s_{+1} - s_{-1}}\right)\phi(s_{-1})\right].\end{aligned}$$

We can write this as

$$V_0 = \frac{1}{1+r}\left[q\phi(s_{+1}) + (1-q)\phi(s_{-1})\right], \tag{10.4}$$

where

$$q = \frac{(1+r)s_0 - s_{-1}}{s_{+1} - s_{-1}}.$$

This is a very appealing way to write the formula. In fact the inequality $0 < q < 1$ is equivalent to our no-arbitrage conditions in Lemma 10.1 above:

$$s_{-1} < (1+r)s_0 < s_{+1} \text{ is equivalent to } 0 < q < 1.$$

Moreover formula (10.4) looks like the mean present value,

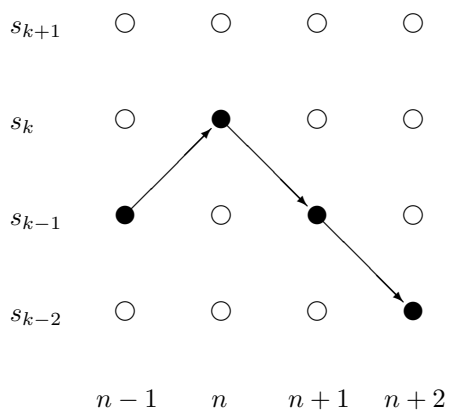
$$V_0 = \frac{1}{1+r} E[\phi(S_1)],$$

except that the expected value is computed using q in place of p. Note also that the value of q is uniquely determined by the equation

$$s_0 = q\frac{s_{+1}}{1+r} + (1-q)\frac{s_{-1}}{1+r}.$$

10.3.2 Pricing for the Random Walk Model

Now lets consider our full problem over several time steps $n = 0, 1, \ldots T$, and an arbitrary exercise value function $\phi(s)$. We are interested in the contingent claim whose exercise value is $\phi(S_T)$. We have the stock price process S_n given by the random walk on an (infinite) set of s_i values as described above, and the deposit certificate price process $B_n = (1+r)^n$, both for $n = 0, \ldots, T$. If we keep track of both time and price (n, S_n) follows a zig-zag path moving left to right through the lattice of (n, s) pairs in the plane. We have illustrated a section of this lattice in the following figure with a possible path for (n, S_n) drawn in.



At each time transition the path moves one step to the right (time advancement) and one step either up or down (price transition). Because S_n *must* move either up or down, S_n can*not* stay constant. The path is not allowed to move strictly horizontally.

With this picture in mind, for each lattice position (s, n) with $0 \le n \le T$ there should be a value $v(s, n)$ which gives the correct market price of the contingent claim when $S_n = s$. Our task is the calculate the values of this function $v(s, n)$.

Observe that each $n \to n+1$ time step is a version of our single branch model. If we know the values of $v(\cdot, n+1)$ we can use our single step calculation (10.4) to calculate the values of $v(\cdot, n)$. The calculation is not complicated. We first calculate the values

$$q_k = \frac{(1+r)s_k - s_{k-1}}{s_{k+1} - s_{k-1}}. \tag{10.5}$$

Then starting from the values of $v(s, T) = \phi(s)$ work backwards in time, calculating $v(s_k, n)$ from $v(s_{k\pm1}, n+1)$ according to

$$v(s_k, n) = \frac{1}{1+r}\left[q_k v(s_{k+1}, n+1) + (1-q_k)v(s_{k-1}, n+1)\right]. \tag{10.6}$$

We will see how to organize such a calculation in MATLAB below, but first we need to work out the rest of the mathematical ideas.

Having determined the values of $v(s,n)$ the market price of the contingent claim is the stochastic process derived from S_n by

$$V_n = v(S_n, n).$$

The justification for this again depends on the existence of a replicating portfolio for the claim. This is more complicated than the single branch setting. Suppose we are at $S_n = s_k$. Our single branch calculations give us portfolio values

$$\alpha(s_k,n) = \frac{v(s_{k+1},n+1) - v(s_{k-1},n+1)}{s_{k+1}-s_{k-1}}, \quad \beta(s_k,n) = \frac{s_{k+1}v(s_{k-1},n+1) - s_{k-1}v(s_{k+1},n+1)}{(1+r)^{n+1}(s_{k+1}-s_{k-1})} \tag{10.7}$$

with the property that

$$v(s_k,n) = \alpha(s_k,n)S_n + \beta(s_k,n)B_n,$$

and for both possible values of $S_{n+1} = s_{k\pm1}$,

$$v(s_{k\pm1},n+1) = \alpha(s_k,n)s_{k\pm1} + \beta(s_k,n)B_{n+1}. \tag{10.8}$$

(The reason for the higher power of $(1+r)$ in the denominator of β is that the bank portion of the portfolio goes from βB_n to βB_{n+1} instead of βB_0 to βB_1. So βB_n is what replaces the value of β in our single time-step calculation.) Restated, for each $n = 0,\dots,T-1$ we have

$$\alpha(S_{n-1},n-1)S_n + \beta(S_{n-1},n-1)B_n = v(S_n,n) = \alpha(S_n,n)S_n + \beta(S_n,n)B_n. \tag{10.9}$$

The implementation of (10.9) is as follows. The values $\alpha(S_{n-1},n-1)$, $\beta(S_{n-1},n-1)$ are the stock and bond holdings we use for the $n-1 \to n$ time step. The left equality of (10.9) says they produce a portfolio with value $v(S_n,n)$ at time n. Having reached time n we get ready for the $t = n \to n+1$ time step. The holdings we want to use for this transition are $\alpha(S_n,n)$, $\beta(S_n,n)$ because they maintain today's value (the right inequality in (10.9)) and by (10.8) will have value $v(S_{n+1},n+1)$ tomorrow for both possible values of S_{n+1}. The key feature is that there are *two* portfolios producing today's value $v(S_n,n)$: the one we just used for the $n-1 \to n$ time transition and the one we are about to use for the $n \to n+1$ time transition. At time n we *refinance* the portfolio by changing from $\alpha(S_{n-1},n-1), \beta(S_{n-1},n-1)$ to $\alpha(S_n,n), \beta(S_n,n)$. This involves buying and selling at time n *but the portfolio's value is the same before and after the transaction* (that's what (10.9) says) so the sales proceeds will exactly cover the purchase expenses. This is called a *self-financing* transaction. Now we are ready for the $n \to n+1$ time step.

$$\begin{aligned}
&\vdots \\
t = n-1:\ & V_{n-1} = v(S_{n-1},n-1) = S_{n-1}\alpha(S_{n-1},n-1) + B_{n-1}\beta(S_{n-1},n-1) \\
&\text{(time step)} \\
t = n:\ & V_n = v(S_n,n) = S_n\alpha(S_{n-1},n-1) + B_n\beta(S_{n-1},n-1) \\
& \quad \text{refinance to} \quad = S_n\alpha(S_n,n) + B_n\beta(S_n,n) \\
&\text{(time step)} \\
t = n+1:\ & V_{n+1} = v(S_{n+1},n+1) = S_{n+1}\alpha(S_n,n) + B_{n+1}\beta(S_n,n) \\
& \quad \text{refinance to} \quad = S_{n+1}\alpha(S_{n+1},n+1) + B_{n+1}\beta(S_{n+1},n+1) \\
&\text{(time step)} \\
t = n+2:\ & V_{n+2} = \cdots \\
&\vdots
\end{aligned}$$

You may notice that (10.7) does *not* produce $\alpha(s,n)$ and $\beta(s,n)$ for $n = T$; it only produces values for $n < T$. That is why we said (10.9) holds only for $n = 1,\dots,T-1$; the right side is undefined for $n = T$. But the left side is correct for $n = T$, giving the exercise values.

Observe that a portfolio which replicates the value of the claim at all times is not just a pair of constants α, β, but is something that changes as time evolves and in a random way that depends on the evolution of the price process S_n. Thus a replicating portfolio is a pair of stochastic processes X_n, Y_n constructed from the Markov chain using the functions $\alpha(\cdot,\cdot)$, $\beta(\cdot,\cdot)$ which we have computed:

$$X_n = \alpha(S_n, n), \quad Y_n = \beta(S_n, n).$$

The value of this portfolio always agrees with the value $V_n = v(n, S_n)$ that we calculated for the option:

$$V_n = X_n S_n + Y_n B_n, \text{ for } n = 0, \ldots, T.$$

But it also has the important self-financing property (10.9). Stated in terms of X_n, Y_n this means that the refinancing from X_{n-1}, Y_{n-1} to X_n, Y_n which occurs at time n does not change the portfolio value:

$$V_n = X_{n-1} S_n + Y_{n-1} B_n.$$

Subtracting the tow formulas for V_n the self-financing property can be expressed as

$$\begin{aligned} 0 &= (X_n - X_{n-1}) S_n + (Y_n - Y_{n-1}) B_n, \text{ or} \\ 0 &= \Delta X_n \frac{S_n}{B_n} + \Delta Y_n \end{aligned} \tag{10.10}$$

where $\Delta X_n = X_n - X_{n-1}$ and $\Delta Y_n = Y_n - Y_{n-1}$ are the backward differences. Another way to express it is

$$\begin{aligned} \Delta V_n &= V_n - V_{n-1} \\ &= (X_{n-1} S_n + Y_{n-1} B_n) - (X_{n-1} S_{n-1} + Y_{n-1} B_{n-1}) \\ &= X_{n-1} \cdot \Delta S_n + Y_{n-1} \cdot \Delta B_n. \end{aligned} \tag{10.11}$$

We can interpret this as saying that the changes in the portfolio's value are due entirely to changes in the prices S_n and B_n; the adjustments to the portfolio's holdings at each time $t = n$ are *not* producing any change in the portfolio's value. Once such a portfolio is created its management does not require or produce new funds as time proceeds (though it does take intelligence).

Self-financing portfolios play two important roles in our discussion. First, they are the way we can replicate the evolving value $V_n = v(S_n, n)$ of a contingent claim using a portfolio of stock and savings alone.. A *replicating portfolio* for such a claim $\phi(S_T)$ is a self-financing portfolio (X_n, Y_n) with the property that

$$\phi(S_T) = X_T S_T + Y_T B_T, \text{ for all possible values of } S_T.$$

It is the existence of replicating portfolios which determines the market value V_n of the claim; see the theorem below.

Secondly, self-financing portfolios allow us to define what we mean by an arbitrage opportunity. An *arbitrage portfolio* is a self-financing portfolio whose value process $V_n = X_n S_n + Y_n B_n$ satisfies

1. $V_0 = 0$,
2. $P(V_T \geq 0) = 1$, and
3. $P(V_T > 0) > 0$.

We maintain that the only reasonable market models are *arbitrage-free*, meaning that no arbitrage portfolios exist. We can state exactly what this requires of our random walk model.

Lemma 10.3. *The random walk market model is arbitrage-free if and only if*

$$s_{i-1} < (1 + r) s_i < s_{i+1}$$

for all states s_i which are accessible before time T.

By "accessible before time T" we mean that

$$P(S_n = s_i \text{ for some } n < T) > 0.$$

This depends on the distribution of S_0. If $S_0 = s_k$, a fixed starting price, the possible values of S_T are those s_i with $k - T < i < k + T$, sometimes called the *cone of influence* of s_k. We will come back to prove the lemma once we have talked about martingales in the next section, because they make the proof easier to describe. For now let's take the lemma for granted and see what else we can say about the random walk model.

Theorem 10.4. *If the random walk model is free of arbitrage, then every contingent claim with terminal value at time T given by an exercise value function $\phi(S_T)$ of the final stock price has a replicating portfolio (X_n, Y_n). The market price of the claim must agree at all times with the market value of the replicating portfolio: $V_n = X_n S_n + Y_n B_n$.*

Proof. The proof is the success of our construction above. □

Suppose you are the writer/issuer of a contract on this claim; you are paid V_0 at time $n = 0$. At time T you have to be prepared to deliver something of value $\phi(S_T)$. You have to invest that V_0 and then manage that investment portfolio as time proceeds so that when $n = T$ arrives the portfolio will be worth exactly the $\phi(S_T)$ you need to fulfill your obligation. This is called hedging your obligation and thus a replicating portfolio is also called a *hedging portfolio*. This is not hard to do, provided we have calculated all the $\alpha(n, s_k)$, $\beta(n, s_k)$ values. It simply consists of following the replicating portfolio $X_n = \alpha(n, S_n)$, $Y_n = \beta(n, S_n)$ as it evolves. Example 10.2 below will illustrate this.

Computing Examples

Given the s_i and r satisfying Lemma 10.3, an exercise value function $\phi(s)$, and a final time T, here are what the calculations need to do.

1. Calculate the q_i from (10.5).
2. Set $v(s_i, T) = \phi(s_i)$ for all i.
3. Iterate backwards from $n = T - 1$ down to $n = 0$ in (10.6) to find the value and portfolio values for each (s, n). If we are going to be responsible for managing a hedging portfolio we will also need to use (10.7) to calculate the values of $\alpha(s, n)$ and $\beta(s, n)$ as we go.

This is relatively simple to have MATLAB do, except for two practical issues. First, MATLAB doesn't let us use $n = 0$ as an array index. So if we use an array `v` we just need to remember that the time index is shifted by 1: `v(i,k+1)` holds the value $v(s_i, k)$, $k = 0, \ldots, T$.

The second issue is that the calculations can only use a finite list of states, `s`$= (s_1, \ldots, s_m)$. But the formulas in (10.6) and (10.7) for s_1 refer to s_0, and for s_n they refer to s_{n+1}. Mathematically the equations involve the full infinite collection of states. There are a couple ways around this. The simplest is to use some approximate formulas for $v(s_1, \cdot)$ and $v(s_m, \cdot)$ which can be worked out in advance, so that we only need to do the calculations for $i = 2, \ldots, n - 1$. When we get to Section 10.5.1 we will explain the simple linear approximation used in the code below. Regarding q for s_1 and s_n we will effectively treat those as absorbing states; indicated in the code by the `NaN` values in `q`. That interpretation allows us to fill in portfolio values for s_1 and s_n as well. Since $S_k = s_1 \to S_{k+1} = s_1$ with certainty, self-financing requires

$$\begin{aligned} \alpha(s_1, k)s_1 + \beta(s_1, k)B_k &= v(s_1, k), \text{ and} \\ \alpha(s_1, k)s_1 + \beta(s_1, k)B_{k+1} &= v(s_1, k+1). \end{aligned}$$

Solving these yields

$$\begin{aligned} \alpha(s_1, k) &= \frac{(1+r)v(s_1, k) - v(s_1, k+1)}{rs_1} \\ \beta(s_1, k) &= \frac{v(s_1, k+1) - v(s_1, k)}{rB_k}, \end{aligned}$$

and similarly for s_n. The code uses these (last two lines of the iteration loop) to fill in values for `alpha(1,k)`, `alpha(n,k)`, `beta(1,k)` and `beta(n,k)`.

Here then is a script to do these calculations for us.

EuroOpt.m

```
%This script assumes that s, phi, T and r are already defined.
%If s, phi are rows switch them to columns.
if dot([1,-1],size(s))<0
    s=s';
end
if dot([1,-1],size(phi))<0
    phi=phi';
end
%
%Calculate the array of s(i+1)-s(i-1) and list of Bank account values.
n=length(s);
dds=s(3:n)-s(1:n-2);
B=(1+r).^(0:T);
%
%Calculate the martingale probabilities.
q=[NaN;((1+r)*s(2:n-1)-s(1:n-2))./dds;NaN];
%
%Initialize v, alpha, beta
v=zeros(n,T+1);
alpha=v; beta=v;
v(:,T+1)=phi;
%
%Fill in v(1,:) and v(n,:) using linear approximation formula:
v(n,1:T)=s(n)*(phi(n)-phi(n-1))/(s(n)-s(n-1))+B(1:T)*(phi(n-1)*s(n)-phi(n)*s(n-1))/(B(T+1)*(s(n)-s(n-1)
v(1,1:T)=s(1)*(phi(2)-phi(1))/(s(2)-s(1))+B(1:T)*(phi(1)*s(2)-phi(2)*s(1))/(B(T+1)*(s(2)-s(1)));
%
%The iteration.
for k=T:-1:1
    v(2:n-1,k)=(q(2:n-1).*v(3:n,k+1)+(1-q(2:n-1)).*v(1:n-2,k+1))/(1+r);
    alpha(2:n-1,k)=(v(3:n,k+1)-v(1:n-2,k+1))./dds;
    beta(2:n-1,k)=(s(3:n).*v(1:n-2,k+1)-s(1:n-2).*v(3:n,k+1))./(B(k+1)*dds);
    %
    %Portfolio values for s(1) and s(n) as absorbing states
    alpha([1;n],k)=((1+r)*v([1;n],k)-v([1;n],k+1))./(r*s([1;n]));
    beta([1;n],k)=(v([1;n],k+1)-v([1;n],k))/(r*B(k));
end
%
%Display results, flipped top to bottom
disp('s and q:')
disp(flipud([s,q]))
disp('v:')
disp(flipud(v))
disp('alpha:')
disp(flipud(alpha))
disp('beta:')
disp(flipud(beta))
```

Example 10.2. As an example suppose the stock price states include $20, 40, 60, 80, 100, 120, 140$ and the interest rate is $r = .05$. Let's consider a call option with strike price $K = 70$:

$$\phi(s) = \begin{cases} s - 70 & \text{if } s \geq 70 \\ 0 & \text{if } s < 70. \end{cases}$$

and work it out for $T = 3$ time periods, from the initial stock price of $S_0 = 80$. Here are the results (calculated with the script `EuroOpt.m`).

```
s and q:
  140.0000       NaN
  120.0000    0.6500
  100.0000    0.6250
   80.0000    0.6000
   60.0000    0.5750
   40.0000    0.5500
   20.0000       NaN

v:
   79.5314   76.5079   73.3333   70.0000
   59.9849   56.5079   53.3333   50.0000
   41.1835   37.8685   33.3333   30.0000
   25.2154   21.1338   17.1429   10.0000
   12.7343    9.3878    5.4762         0
    4.9174    2.8685         0         0
         0         0         0         0

alpha:
    1.0000    1.0000    1.0000         0
    0.9660    1.0000    1.0000         0
    0.8844    0.9048    1.0000         0
    0.7120    0.6964    0.7500         0
    0.4566    0.4286    0.2500         0
    0.2347    0.1369         0         0
         0         0         0         0

beta:
  -60.4686  -60.4686  -60.4686         0
  -55.9335  -60.4686  -60.4686         0
  -47.2519  -50.1026  -60.4686         0
  -31.7460  -32.9338  -38.8727         0
  -14.6636  -15.5491   -8.6384         0
   -4.4704   -2.4835         0         0
         0         0         0         0
```

Let's follow this carefully through the sequence $S_0 = 80$, $S_1 = 100$, $S_2 = 120$, $S_3 = 100$ and see how the replicating portfolio evolves. We start with

$$V_0 = .712 \cdot S_0 - 31.746 \cdot B_0 = .712 \cdot 80 - 31.746 \cdot 1 = 25.2154.$$

(If you do the arithmetic you will find a discrepancy in the third decimal; that's because we have rounded of the values of α and β in the display above.) Now make the transition to $n = 1$.

$$V_1 = .712 \cdot S_1 - 31.746 \cdot B_1 = .712 \cdot 100 - 31.746 \cdot 1.05 = 37.8685.$$

We now refinance the portfolio to

$$V_1 = .9048 \cdot S_1 - 50.1026 \cdot B_1 = .9048 \cdot 100 - 50.1026 \cdot 1.05 = 37.8685.$$

With the new portfolio allocations we are ready for the transition to $n = 2$.

$$V_2 = .9048 \cdot S_2 - 50.1026 \cdot B_2 = .9048 \cdot 120 - 50.1026 \cdot 1.05^2 = 53.3333.$$

Refinance again.

$$V_2 = 1.0 \cdot S_2 - 60.4686 \cdot B_2 = 1.0 \cdot 120 - 60.4686 \cdot 1.05^2 = 53.3333.$$

Now make the final transition to $n = 3$.

$$V_3 = 1.0 \cdot S_3 - 60.4686 \cdot B_3 = 1.0 \cdot 100 - 60.4686 \cdot 1.05^3 = 30.0,$$

which agrees with the exercise value $\phi(S_3) = \phi(100) = 30$.

10.4 No-Arbitrage Pricing and Martingales

Now we will see that the relationships we have found for the random walk model can be expressed naturally in terms of martingales. The relationships we find here turn out to describe more complicated models and are foundational ideas of mathematical finance in general. In particular the martingale properties here will guide us when we talk about the Black-Scholes model in Chapter 12.

The first thing to observe is that the q_i values are determined by

$$s_i = \frac{1}{1+r}[q_i s_{i+1} + (1 - q_i)s_{i-1}].$$

If we think of q_i as replacing the p_i for the Markov chain S_n, these equations collectively say that

$$S_n/B_n = E^q[S_{n+1}/B_{n+1}|S_{0:n}].$$

In other words, when we use the q_i instead of the p_i, S_n/B_n is a martingale! For this reason the q_i are often referred to as *martingale probabilities* or *risk-neutral probabilities*. The "E^q" above means expectation for the Markov chain under the martingale probabilities. We found that **the absence of arbitrage in our model is equivalent to the the existence of martingale probabilities with** $0 < q_i < 1$. Every sequence of states S_n which is possible under the original p_i is possible under the q_i (although with a different probability), and conversely. If you read other discussions of mathematical finance you may see the phrase "equivalent martingale probabilities" or "equivalent probability measure". Here "equivalent" doesn't mean equal or the same, but only that the same paths have positive probabiliites under each. (In fact the *only* role of the original p_i in these pricing calculations is to determine which stock price paths can occur with positive probability.)

We will use

$$M_n = S_n/B_n$$

to denote this important q-martingale. This is not the only q-martingale hiding among our various formulas. Consider a self-financing portfolio (X_n, Y_n). The portfolio's value is $V_n = X_n S_n + Y_n B_n$. Let

$$C_n = V_n/B_n.$$

This is related to our M_n by

$$C_n = X_n M_n + Y_n.$$

Let's calculate C_n as a sum of its increments

$$C_n = C_0 + \sum_{k=1}^{n} \Delta C_k,$$

were $\Delta C_k = C_k - C_{k-1}$. Now we can work out this difference as a sort of discrete product rule.

$$\begin{aligned}
\Delta C_k &= (X_k M_k + Y_k) - (X_{k-1} M_{k-1} + Y_{k-1}) \\
&= (X_k - X_{k-1})M_k + (Y_k - Y_{k-1}) + X_{k-1}(M_k - M_{-1}) \\
&= \Delta X_k \, M_k + \Delta Y_k + X_{k-1} \Delta M_k.
\end{aligned}$$

Now notice that the self-financing property (10.11) says that

$$\Delta X_k\, M_k + \Delta Y_k = 0.$$

So we find that

$$\Delta C_k = X_{k-1}\Delta M_k$$

and therefore

$$C_n = C_0 + \sum_{k=1}^{n} X_{k-1}\Delta M_k.$$

This is what we called a discrete stochastic integral in Chapter 9, which we know produces another martingale. We find that C_n is also a q-martingale! In particular we find the *martingale pricing formula* (or *risk-neutral pricing formula*): if there is a replicating portfolio for $\phi(S_T)$ then the market value V_n of this portfolio divided by B_n is a q-martingale: for $0 \le n \le T$

$$V_n/B_n = C_n = E^q[C_T|S_{0:n}] = E^q[\phi(S_T)/B_T|S_{0:n}]. \tag{10.12}$$

This implies

$$\begin{aligned} v(s,n) &= B_n E^q[\phi(S_T)/B_T|\ S_n = s] \\ v(s,0) &= E^q_s[\phi(S_T)/B_T], \end{aligned}$$

the last line because $B_0 = 1$. This formulation assumes the existence of a replicating portfolio but otherwise makes no reference to what the replicating portfolio X_n, Y_n actually is; everything is in terms of the q_i.

This is a good place to go back and prove Lemma 10.3.

Proof. To say that $s_{i-1} < (1+r)s_i < s_{i+1}$ for all i is the same as saying there exists a set of equivalent martingale probabilities q_i. If that is true we claim that no arbitrage opportunity can exist. Suppose that there was a starting state s_0 and a self-financing portfolio so that $V_T = X_{T-1}S_T + Y_{T-1}B_T \ge 0$ and $V_T > 0$ with positive probability. If $V_T > 0$ has positive probability under the p_i then it has positive probability under the q_i. As above, $C_n = V_n/B_n$ is a q-martingale so

$$C_0 = E^q_{s_0}[C_T].$$

But $C_T \ge 0$ and $P^q(C_T > 0) > 0$ implies that the above expectation is strictly positive. So $V_0 = C_0 B_0 = 0$ is not possible. Therefore no arbitrage opportunities exist.

Now suppose that no arbitrage opportunities exist. Consider any particular branch $s_{k-1} \nwarrow s_k \nearrow s_{k+1}$. We know that if $s_{k-1} < (1+r)s_k < s_{k+1}$ were *not* true there would exist α and β so that

$$\begin{aligned} 0 &= \alpha s_k + \beta \\ 0 &\le \alpha s_{k\pm 1} + \beta(1+r), \text{ strictly in one case.} \end{aligned}$$

Pick a starting position s_0 from which it *is* possible to reach s_k at some time prior to the terminal time T. Now construct a portfolio as follows. Initially $X_0 = Y_0 = 0$. We keep $X_n = Y_n = 0$ up to the first time $\mathcal{K}$ when $S_{\mathcal{K}} = s_k$. At that time we make $X_{\mathcal{K}} = \alpha$ and $Y_{\mathcal{K}} = \beta/B_{\mathcal{K}}$. That makes

$$V_{\mathcal{K}} = \alpha s_k + \beta = 0.$$

We go one time step past $\mathcal{K}$ and now

$$V_{\mathcal{K}+1} = \alpha s_{k\pm 1} + \beta(1+r).$$

We know this is nonnegative, and strictly positive in one case. Now refinance by converting the portfolio's value to bonds: $X_{\mathcal{K}+1} = 0$ and

$$Y_{\mathcal{K}+1}B_{\mathcal{K}+1} = V_{\mathcal{K}+1}.$$

Now we just leave X_n and Y_n unchanged for the remaining $k+1 < n < T$. This is clearly self-financing. It has $V_0 = X_0 S_0 + Y_0 B_0 = 0$. $V_T \ge 0$ in all cases. And in the event that the price S_n does hit s_i at some time $n = \mathcal{K} < T$ we will have $V_{\mathcal{K}+1} > 0$ with positive probability, and will remain so until the final time. In other words $V_T > 0$ does have positive probability. Thus an arbitrage portfolio does exist. But we are assuming that this in not possible, so $s_{ik1} < (1+r)s_k < s_{k+1}$ must be true of all k. □

To summarize the description of the random walk model in martingale terms, we have found the following.

- The model is free of arbitrage if and only if the p_i can be replaced by equivalent "martingale probabilities" q_i under which
$$M_n = S_n/B_n$$
is a q-martingale. Lemma 10.3 characterizes this in terms of r and the s_i.

- Assuming the model is free of arbitrage, if $V_n = X_n S_n + Y_n B_n$ is the value process for a self-financing portfolio X_n, Y_n then
$$V_n/B_n$$
is also a q-martingale.

- Assuming the model is free of arbitrage given any exercise value function $\phi(s)$ we can always find a self-financing portfolio which replicates $V_T = \phi(S_T)$. This is described by saying that the market S_n, B_n is *complete*.

10.5 Pricing and Parity Relations Among Options

There are some simple relationships between the prices of various types of assets. These are easily derived from the martingale pricing formula (10.12). Forwards are elementary; for $\phi^{\text{forward}}(s) = s - K$ we have

$$\begin{aligned} V_n^{\text{forward}} &= B_n E^q \left[\frac{S_T - K}{B_T} \middle| S_{0:n} \right] \\ &= B_n E^q \left[\frac{S_T}{B_T} \middle| S_{0:n} \right] - K \frac{B_n}{B_T} \\ &= B_n \frac{S_n}{B_n} - K \frac{B_n}{B_T} \\ &= S_n - K(1+r)^{n-T}. \end{aligned}$$

The exercise value functions for the call and put (using the same "strike price "K) are

$$\phi^{\text{call}}(s) = \begin{cases} s - K & \text{if } s \geq K \\ 0 & \text{if } s < K \end{cases}, \quad \phi^{\text{put}}(s) = \begin{cases} 0 & \text{if } s \geq K \\ K - s & \text{if } s < K \end{cases}.$$

Explicit pricing expressions for these exist for specific models, such as CRR in Section 10.6.1 below. (For the continuous time model, the explicit pricing formula is the famous Black-Scholes formula; see Chapter 12.) But there is a simple relation between them:

$$\phi^{\text{call}}(s) - \phi^{\text{put}}(s) = s - K = \phi^{\text{forward}}(s).$$

Consequently

$$V_n^{\text{call}} - V_n^{\text{put}} = V_n^{\text{forward}} = S_n - K(1+r)^{n-T}.$$

This relation is called *put-call parity*. (Bear in mind that it assumes that the call and put have the same terminal time T and strike price K.) In particular if there is an explicit pricing formula for a call then there is for a put as well.

Other exercise value functions are considered in the literature, many with colorful names.

- A cash-or-nothing call: $\phi^{\text{c-n call}}(s) = \begin{cases} C & \text{if } s \geq K \\ 0 & \text{if } s < K \end{cases}$.

- A stock-or-nothing put: $\phi^{\text{s-n call}}(s) = \begin{cases} s & \text{if } s \geq K \\ 0 & \text{if } s < K \end{cases}$.

- A strangle: $\phi^{\text{strangle}}(s) = \begin{cases} K_1 - s & \text{if } s \geq K_1 \\ 0 & \text{if } K_1 < s < K_2 \\ s - K_2 & \text{if } K_2 \leq s \end{cases}$.

There are put versions of cash-or-nothings and stock-or-nothings. There are "straddles", "bear spreads", "bull spreads", "butterfly spreads", ... There are many parity-like relations among these, so only a couple nontrivial pricing formulas are needed to obtain formulas for the rest. In principle any piecewise linear exercise value function can be represented using a combination of calls and puts, all with the same exercise time T. Hull [31] discusses these and others. There are compound options, such as a put on a call. We won't pursue the endless varieties.

10.5.1 Approximations for Small or Large Price

If $\phi(s) = c_1 s + c_2$ then we can work out a simple expresson for $v(n, s)$. The martingale pricing formula makes this simple.

$$\begin{aligned} v(S_n, n)/B_n &= E^q\left[\frac{c_1 S_T + c_2}{B_T} \mid S_{0:n}\right] \\ &= c_1 E^q\left[\frac{S_T}{B_T} \mid S_{0:n}\right] + c_2 E^q\left[\frac{1}{B_T} \mid S_{0:n}\right] \\ &= c_1 \frac{S_n}{B_n} + c_2 \frac{1}{B_T}, \end{aligned}$$

so that

$$v(s, n) = c_1 s + c_2 (1 + r)^{n-T}. \tag{10.13}$$

Most of the exercise value functions $\phi(s)$ considered in the literature are piecewise linear, with $\phi(s) = c_1 s + c_2$ for all sufficiently large (or small) s. We might reason from this that for sufficiently large (or small) s the formula (10.13) should be a good approximation.

We used that in the code on page 162 to produce an approximate formula for $v(s_n, k)$ for the largest s_n carried by the calculations. The code simply fitted a linear formula to

$$\phi(s_n) = c_1 s_n + c_2 \text{ and } \phi(s_{n-1}) = c_1 s_{n-1} + c_2$$

to obtain

$$c_1 = \frac{\phi(s_n) - \phi(s_{n-1})}{s_n - s_{n-1}} \text{ and } c_2 = \frac{s_n \phi(s_{n-1}) - s_{n-1}\phi(s_n)}{s_n - s_{n-1}}$$

and then used these values in (10.13) to approximate $v(s_n, k)$. We did likewise for s_1 and s_2.

10.6 Generalizations

Our random walk model has served to introduce some key ideas in the mathematical modeling of financial markets. But it is rather simplistic compared to the complexities of real markets. To make it a bit more realistic there are many possible refinements, extensions, and generalizations which can be considered. In this section we briefly mention a few of them.

10.6.1 The Cox-Ross-Rubinstein Model

We have said little about how the states $\mathcal{S} = \{s_i\}$ and transition probabilities p_i ought to be chosen. In the mathematical finance literature random walk-like models typically assume that the stock price changes according to some prescribed ratios: S_{i+1}/S_i being either $u > 1$ for a price increase or $d < 1$ for a price decrease. This is usually called the *Cox-Ross-Rubenstein model*. In the special case that $u = \rho$ and $d = 1/\rho$ for a constant $\rho > 1$ this becomes an instance of our random walk model, with states consisting of all integer powers of ρ: $s_i = \rho^i$. Under this model at each transition the stock price either increases or deceases by a

factor of ρ. Although probabilities $p_i, 1-p_i$ for the transitions could be specified, there is no need because they are replaced by the martingale probability of (10.5)

$$q_i = q = \frac{(1+r)\rho - 1}{\rho^2 - 1}.$$

for all the pricing calculations. For an arbitrage-free model we can use any interest rate r with $\rho^{-1} < r < \rho$.

Under the martingale probabilities we see that $S_n = \rho^{X_n}$ where X_n is the random walk on the integers with $X_n \to X_{n+1} = X_n \pm 1$ with probabilities q and $1-q$ respectively. This reduces the pricing calculation to calculations for a random walk. With $s = \rho^x$

$$P_s^q(S_n = \rho^y) = P_x(X_n = y)$$

With initial value $X_0 = x$ the possible values of $X_n = y$ are

$$y = x + \ell - (n - \ell) = x + 2\ell - n \text{ for } \ell = 0, 1, \ldots, n$$

with probabilities

$$P_x^q(X_n = x + 2\ell - n) = \binom{n}{\ell} q^\ell (1-q)^{n-\ell}.$$

This allows us to write down price formulas. For instance in the case of a call option with exercise price $K = \rho^L$ exercise date T, if the initial stock price is $S_0 = s = \rho^x$ we have

$$\begin{aligned} v(s,0) &= (1+r)^{-T} E_s^q[(\rho^{X_T} - K)^+] \\ &= (1+r)^{-T} E_x^q[\rho^{X_T} - \rho^L; X_T \geq L] \\ &= (1+r)^{-T} \sum_{\ell=\cdots} \left[\rho^{x+2\ell-T} - \rho^L\right] \binom{T}{\ell} q^\ell (1-q)^{T-\ell}, \end{aligned}$$

where the sum "$\ell = \cdots$" is over those $\ell = 0, 1, \ldots, T$ for which $x + 2\ell - T \geq L$. This formula is not particularly illuminating. But it would be simple for a computer to evaluate this formula and not need to go through the time iteration of page 162.

10.6.2 Multiple Stocks

Suppose there are M different stocks with prices given by S_n^j for $j = 1, \ldots, M$ and $n = 0 \ldots, T$. We assume that these are all random variables taking only a countable number of possible values. We can assemble the prices as a vector $\mathbf{S}_n = (S_n^1, \ldots, S_n^M)$, along with our bank account process $B_n = (1+r)^n$. A portfolio consists of processes Y_n and $\mathbf{X}_n = (X_n^1, \ldots, X_n^M)$, these being $\mathbf{S}_{0:n}$-determined for each $n = 0, \ldots T-1$. The value of such a portfolio is the process

$$\begin{aligned} V_n &= \mathbf{X}_n \cdot \mathbf{S}_n + Y_n B_n \\ &= \left(\sum_{j=1}^M X_n^j S_n^j \right) + Y_n B_n. \end{aligned}$$

The portfolio $\mathbf{X}_n$, Y_n is self-financing if

$$\mathbf{X}_{n-1} \cdot \mathbf{S}_n + Y_{n-1} B_n = \mathbf{X}_n \cdot \mathbf{S}_n + Y_n B_n$$

for all $n = 1, \ldots, T-1$. An arbitrage portfolio is a self-financing portfolio with

- $V_0 = 0$,
- $V_T \geq 0$, and
- $P(V_T > 0) > 0$.

We say that the model is free of arbitrage if no arbitrage portfolios exist. This will imply that there is an "equivalent" assignment of probabilities which makes all of the S_n^j/B_n martingales. If we refer to the new assignment of probabilities as the "q-probabilities" then by "equivalent" is meant that the events which have positive probability under the original probabilities are the same as those which have positive probabilities under the q-probabilities. (There is no assumption that $\mathbf{S}_n$ is Markov here, either under the original probabilities or the q-probabilites.)

A contingent claim $\phi(\mathbf{S}_T)$ is replicated by a self-financing portfolio $\mathbf{X}_n$, Y_n if

$$V_T = \phi(\mathbf{S}_T).$$

If such a replicating portfolio exists, then the market value of the claim at time n is V_n, determined mathematically by the property that V_n/B_n must be a q-martingale:

$$V_n/B_n = E^q[\phi(\mathbf{S}_T)/B_T|\mathbf{S}_{0:n}].$$

The argument for this is just a vector version of the reasoning which led to (10.12).

10.6.3 Path-Dependent Claims

The exercise value V_T we have been considering is a function $\phi(S_T)$ or the final stock price. In some situations this needs to be generalized to allow the exercise value to be an $S_{0:T}$-dependent random variable $\Phi(S_{0:T})$. This means it may involve past as well as present values of the stock price. These are called *path-dependent* options because the final value depends on the path the stock price followed to reach its final value, not just the final value itself. An example is a *forward lookback put*, for which

$$V_T = (Z_T - K)^+, \text{ where } Z_T = \min_{k\le T} S_k.$$

The martingale pricing formula still holds for path-dependent options, although the calculations are more complicated in general because the value V_n at time n depends on the price history $S_{0:n}$ not just S_n itself. In the particular case of our lookback put we can turn it back in to a Markov chain situation by considering the chain consisting of the pair (S_n, Z_n) where $Z_n = \min_{k\le n} S_k$. We could then use an iterative calculation similar to that of page 161 to compute values $v(s, z, n)$ for s and z in some large but finite range of values. We are not going to pursue the details.

10.6.4 General Results about Finite Markets

The success of our pricing of contingent claims for the random walk model is based on three key features.

- The absence of arbitrage in the market (no arbitrage portfolios exist).
- The existence of martingale probabilities (S_n/B_n is a q-martingale).
- The market is complete (every contingent claim $\Phi(S_{0:T})$ can be replicated with a self-financing portfolio).

We have seen that all claims of the form $\phi(S_T)$ can be replicated. That remains true for the more general path-dependent case $\Phi(S_{0:T})$; the details are just more involved.

A couple possible ways to improve our stock price model were mentioned above. But with each new model we are faced with the question of whether our three key features continue to hold. So what can we say in general? Can these features fail if we replace the random walk with a more general vector-valued Markov chain?

Even in the random walk model we know that absence of arbitrage can fail if the interest rate r is outside a certain range. In general it turns out that the market is free of arbitrage if and only if there is an equivalent assignment of probabilities which makes all of the S_n^j/B_n martingales. (Again, "equivalent" means that the events which have positive probability under the original probabilities are the same as those which have positive probabilities under the q-probabilities.) In the random walk model we saw that the condition on r for absence of arbitrage was the same as for the existence of the martingale probability q. That the existence

of martingale probabilities is equivalent to the absence of arbitrage turns out to be true in the more general models.

However in some cases there can be more than one choice of martingale probability. See Problem 10.4 for a simple instance in which this occurs. This is one reason why our simple random walk model, with only two possible price transitions at each stage, works out so nicely. If there are different choices of martingale probabilities then the martingale pricing formula produces different option prices depending on which martingale probability is being used. This means the price will depend on considerations beyond those we have discussed.

What about completeness? If there is no replicating portfolio for an option then our reasoning breaks down. It turns out however that this can happen if and only if the martingale probabilities are not unique! In other words if there is only one way to choose martingale probabilites, then all contingent claims *can* be replicated with a self-financing portfolio and so our martingale pricing formula will prescribe the market value. (This is reminiscent of basic linear algebra, in which the non-solvability of some equations $\mathbf{Ax} = \mathbf{b}$ is equivalent to nonuniqueness of solutions for those equations which are solvable.)

For a development of this general theory see Musiela and Rutkowski [44], Chapter 3 specifically, and the other references cited there. These general features continue to hold in continuous time models, but there are more technicalities to deal with in that setting. We will encounter them again in Chapter 12 when we have a brief look at the Black-Scholes model.

10.6.5 American Options

Consider again our random walk model and an exercise value $\phi(S_T)$. We want to mention birefly another new feature: the possibility of *early exercise*. This means that the option owner is allowed to exercise the option at a time $\mathcal{T} \leq T$ of their choosing, instead of always waiting to the final time T. The option owner will need to have a strategy for deciding when to exercise, and that will likely depend on the stock price history, but of course can not take advantage of some prophetic knowledge of what the stock price will do in the future. In other words $\mathcal{T}$ must be a stopping time. Options with this feature are called *American* options. Without this early exercise feature (i.e. with the requirement that $\mathcal{T} = T$) they are called *European* options. As you might expect the market value of an American option depends on the optimal strategy for deciding when to exercise. This becomes an optimal stopping (under the martingale probabilities). Without drawing out all the details, the essential change in the calculations is that (10.6) is replaced by

$$v(s_k, n) = \max\left(\phi(s_k), \frac{1}{1+r}\left[q_k v(s_{k+1}, n+1) + (1-q_k)v(s_{k-1}, n+1)\right]\right).$$

The option holder's strategy must be to use the early exercise privilege the first time $\mathcal{T} = k$ that $v(S_k, k) = \phi(S_k)$. American options are very computable in this discrete-time setting. (They become considerably harder in continuous time.)

Problems

Problem 10.1
Consider the single branch model. Show that if $p = 1$ but $s_{+1} \neq (1+r)s_0$ then the market has an arbitrage opportunity and explain what transactions would take advantage of the opportunity. Do the same if $p = 0$ and $s_{-1} \neq (1+r)s_0$.

..... P01

Problem 10.2
Consider a random walk stock price model whose prices include

$$\ldots, s_0 = 20,\ s_1 = 40,\ s_2 = 60,\ s_3 = 80,\ s_4 = 100,\ s_5 = 120,\ s_6 = 140,\ \ldots$$

Based on this much of the model, what bounds must the interest rate r obey to insure the absence of arbitrage?

Now suppose that interest rate is $r = .10$ and consider the contingent claim with exercise value function

$$\phi(s) = \begin{cases} 50 & \text{if } 50 \le s \le 110 \\ 0 & \text{otherwise} \end{cases}$$

and exercise time $T = 3$. Compute the market value $v(80, 0)$ of this contingent claim at $t = 0$ assuming that $S_0 = 80$.

Suppose you write such a contract to a client. They pay you $v(80, 0)$ and now you need form and manage a hedging portfolio. What would your initial hedging portfolio X_0, Y_0 be? Describe the changes you would make to your hedging portfolio between $t = 0$ and $t = 3$ if the stock price evolves through the values $S_1 = 60$, $S_2 = 40$, $S_3 = 60$. (This should be something similar to what is on the top of page 167.)

You can do the calculations by hand or with MATLAB. But you should write out at least one step of the calculation by hand, for instance the calculation of $v(80, 0)$, $\alpha(80, 0)$, $\beta(80, 0)$ from $v(60, 1)$ and $v(100, 1)$ (just to be sure you understand what calculations take place at each stage).

..... `3step`

Problem 10.3

For the CRR model find an explicit pricing formula if the exercise value function is $\phi(s) = s^\gamma$ for a positive constant γ.

..... `CRRgamma`

Problem 10.4

Consider a one-step market model in which S_1 has *three* possible values instead of two. To be specific, consider the single branch of a "trinomial tree" illustrated in Figure 10.1 below. The values inside the nodes are the respective stock prices. Suppose that the interest rate is $r = 0$, so that $B_0 = B_1 = 1$.

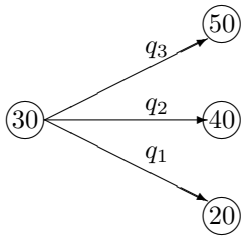


Figure 10.1: Trinomial Branch

a) Assuming that all three branches have nonzero probabilities, show that there are no arbitrage opportunities in this market.

b) Show that there is *more than one way* to assign probabilities q_1, q_2, q_3 (nonnegative, with $q_1 + q_2 + q_3 = 1$) to the three branches so that $S_0 = E^q[S_1]$; i.e. there is more than one martingale measure.

c) Show that there does *not* exist a replicating portfolio for the European call option with exercise price of 30 at time $t = 1$, i.e. for the contingent claim with

$$X = \begin{cases} S_1 - 30 & \text{if } S_1 \ge 30 \\ 0 & \text{if } S_1 < 30 \end{cases}$$

The point is that binomial trees are special. With three or more branches martingale probabilities will exist but be non-unique and even if there is no arbitrage, the market may fail to be complete.

..... `Trinomial`

Problem 10.5
Consider a random walk stock price model whose prices include

$$s_1 = 1,\ s_2 = 2,\ \ldots, s_{20} = 20$$

using interest rate $r = .03$ over $T = 50$ time periods. Consider a call option with exercise value function $\phi(s) = \max(0, s - 10)$. Using our script `EuroOpt.m` calculate the values $v(s, n)$. (You might want to disable the `disp ...` lines at the end of the script to keep it from printing all the results in the command window.) Now produce a surface plot of the resulting values of $v(s, n)$ as a function of s and n. You can do that with the following commands.

```
t=0:50;
[tt,ss]=meshgrid(t,s)
surf(tt,ss,v)
```

Now add the values of $\phi(s)$ the plot, as follows.

```
hold on
Phi=phi*ones(1,51)
surf(tt,ss,Phi)
```

Repeat the above for the corresponding put option: $\phi(s) = \max(0, 10 - s)$.

The values you have computed are for the European versions of the options (no early exercise). For one of the two options there would be no benefit from exercising the option early even if you could, while for the other there are *some circumstances* in which early exercise would be preferable (i.e. $\phi(s) > v(s, n)$) if it were allowed. Which one is which? How can you tell this from your graphs?

Just you you don't misunderstand, you will have only calculated the values of the European versions of these, not the values of their American counterparts. However you can tell from this simple comparison whether the American version would or would not be more valuable than the European version.

..... `EA-Comp`

For Further Study

A very good reference is Shreve [55]. The first volume provides a more thorough treatment of discrete models. The second volume develops the continuous time theory. The Cox-Ross-Rubinstein model and general theory of finite markets (Section 10.6) is developed in [44]. Hull [31] includes a lot details about actual markets.

Chapter 11

Continuous Time Markov Chains

This chapter considers a class of Markov chains Y_t for which time varies continuously, rather than discretely as in the preceding chapters. We will continue to assume that the state space $\mathcal{S}$ is countable. (The term "chain" is typically used to distinguish processes with countable state space from those with continuous state spaces, such as we will encounter in Chapter 12.) The Markov chains of this chapter are sometimes called *jump processes* because they move only by instantaneous jumps, with waiting times of random lengths between jumps.

The transition probabilities will be denoted

$$p_{i,j}(t) = P(Y_t = j | Y_0 = i),\ i, j \in \mathcal{S}.$$

The i, j range over the state space $\mathcal{S}$, and the time variable t can take any real value $0 \leq t$. Because there is no smallest possible time step the statement of the Markov property must involve two time values $0 \leq s < t$:

$$P(Y_t = k | Y_{[0,s]} = y_{[0,s]}) = p_{y_s,k}(t - s).$$

This should hold for all $0 \leq s < t$; all $k \in \mathcal{S}$; and all possible histories $y_{[0,s]}$. The difference from (3.19) is that we understand $y_{[0,s]}$ to be a function $y(u)$ of a continuous variable u defined for $0 \leq u \leq s$ and taking values in $\mathcal{S}$, so that "$Y_{[0,s]} = y_{[0,s]}$" is specifying the full history of Y over the time interval $[0, s]$. The Markov property is that the right side only depends on the most recent state y_s, not what happened prior to that. The Tower Law (part 8 of Proposition 3.8) says

$$P(Y_t = k | Y_0 = y_0) = E[P(Y_t = k | Y_{[0,s]} = y_{[0,s]}) | Y_0 = y_0].$$

Replacing $y_0 = i$ and $y_s = j$ this boils down to the Chapman Kolmogorov equation

$$p_{i,j}(k) = \sum_j p_{i,j}(s) p_{j,k}(t - s), \tag{11.1}$$

generalizing (2.4). We would expect that

$$\sum_j p_{i,j}(t) = 1$$

as well, but this turns out to be problematical in some cases. The difficulty is due to a new phenomenon which can occur for continuous time chains: that an infinite number of jumps occur in a finite amount of time. This is what we call explosion in finite time. If that occurs before t then it is not clear what the state Y_t should be. We see an example in Section 11.2.2 and discuss the possibility more in Section 11.6. Sections 11.3.2 and 11.3.1 describe two areas of application where chains of this type are being studied.

11.1 The Exponential Distribution and the Markov Property

Markov chains in continuous time depend on the memoryless property of the exponential distribution. Recall that a random variable W has an exponential distribution (with parameter $\lambda > 0$) if it has density $f(w) =$

$\lambda e^{-\lambda w}$ for $w \geq 0$ and $f(w) = 0$ for $w < 0$. In particular $W \geq 0$ with probability 1 and for $t \geq 0$

$$P(W > t) = \int_t^\infty \lambda e^{-\lambda w}\, dw = e^{-\lambda t}.$$

Observe that

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

In terms of conditional probabilities,

$$P(W > t + s | W > t) = P(W > s). \tag{11.2}$$

This is the *memoryless property* of the exponential distribution. It means that if you are waiting for W to happen and it has not happened yet ($W > t$) the conditional probability that you will wait at least s time units more is the same as the probability of waiting at least s in the first place. In brief, how long you have already waited does not affect how much more you can expect to wait.

Now consider the possibility of a Markov process Y_t which starts at $Y_0 = 0$ and waits there for an exponentially distributed random amount of time W, at which it jumps to $Y_t = 1$ where it stays ever after:

$$Y_t = \begin{cases} 0 & \text{for } t < W \\ 1 & \text{for } W \leq t. \end{cases} \tag{11.3}$$

We claim that this is a Markov chain. Clearly the state space is $\mathcal{S} = \{0, 1\}$ and 1 is an absorbing state. The Markov property would say that

$$P(Y_{s+t} = 0 | Y_u = 0 \text{ for all } 0 \leq u \leq t) = P(Y_s = 0 | Y_0 = 0).$$

In terms of W this reduces to

$$P(W > t + s | W > t) = P(W > s).$$

In other words if W is an exponential random variable then Y_t *does* have the Markov property. Conversely, *only* exponential random variables have the memoryless property (see Problem 11.1), so if Y_t of (11.3) is Markov then W *must* be an exponential random variable.

The construction (11.3) is only random with regard to *when* the jump occurs. We know with certainty *where* it will jump to: state 1. A similar construction will produce a continuous time Markov chain for which where it jumps to is also random. Let W_1, W_2 be two independent exponential random variables with parameters λ_1, λ_2 respectively. We view the W_i as timers. At time $t = 0$ we set $Y_0 = 0$ and start *both* timers. We wait to see which timer goes off first. If timer 1 goes off first ($W_1 < W_2$) then Y_t jumps to state 1 at time $t = W_1$ and remains there forever. If timer 2 goes off first Y_t jumps to state 2 at time $t = W_2$ and remains there forever.

$$Y_t = \begin{cases} 0 & \text{for } 0 \leq t < \min(W_1, W_2) \\ k & \text{if } W_k = \min(W_1, W_2) \leq t. \end{cases} \tag{11.4}$$

This process has state space $\mathcal{S} = \{0, 1, 2\}$. We claim that it is again a Markov process. Both states $1, 2$ are absorbing. The principal thing to check is that for either $k = 1, 2$

$$P(Y_{s+t} = k | Y_u = 0 \text{ for all } 0 \leq u \leq t) = P(Y_s = k | Y_0 = 0). \tag{11.5}$$

Let's check this for $k = 1$. Start with the right side.

$$P(Y_s = 1 | Y_0 = 0) = P(W_1 \leq s \text{ and } W_1 \leq W_2).$$

Because the W_1 and W_2 are independent their joint density is

$$f(w_1, w_2) = \lambda_1 e^{-\lambda_1 w_1} \lambda_2 e^{-\lambda_2 w_2}.$$

So

$$\begin{aligned} P(Y_s = k|Y_0 = 0) &= P(W_1 \le s \text{ and } W_1 \le W_2) \\ &= \int_0^s \left[\int_{w_1}^\infty f(w_1, w_2)\, dw_2\right] dw_1 \\ &= \int_0^s \lambda_1 e^{-\lambda_1 w_1} e^{-\lambda_2 w_1}\, dw_1 \\ &= \frac{\lambda_1}{\lambda_1+\lambda_2}(1 - e^{-(\lambda_1+\lambda_2)s}). \end{aligned} \tag{11.6}$$

The left side of (11.5) is

$$P(t < W_1 \le s + t \text{ and } W_1 \le W_2)/P(t < \min(W_1, W_2)).$$

Because of independence the denominator is

$$P(t < W_1 \text{ and } t < W_2) = P(t < W_1)P(t < W_2) = e^{-t\lambda_1}e^{-t\lambda_2} = e^{-t(\lambda_1+\lambda_2)}. \tag{11.7}$$

Using the joint density again the numerator is

$$\begin{aligned} P(t < W_1 \le s + t \text{ and } W_1 \le W_2) &= \int_t^{s+t} \left[\int_{w_1}^\infty f(w_1, w_2)\, dw_2\right] dw_1 \\ &= \int_t^{s+t} \lambda_1 e^{-\lambda_1 w_1} e^{-\lambda_2 w_1}\, dw_1 \\ &= \frac{\lambda_1}{\lambda_1+\lambda_2}(e^{(\lambda_1+\lambda_2)(s+t)} - e^{(\lambda_1+\lambda_2)t}). \end{aligned}$$

Dividing numerator by denominator now confirms (11.5) (for our test case of $k = 1$, $m = 2$).

There is an alternate way to understand this same construction. Let $J = \min(W_1, W_2)$ be when the first timer goes off and K the index of the winning timer: $K = 1$ if $J = W_1$ or $K = 2$ if $J = W_2$. We see from (11.7) that J is another exponential random variable, with parameter $\bar\lambda = \lambda_1 + \lambda_2$:

$$P(J > t) = P(t < W_1 \text{ and } t < W_2) = e^{-\bar\lambda t}.$$

The distribution of K follows by letting $s \to \infty$ in (11.6):

$$P(K = 1) = P(W_1 < W_2) = \frac{\lambda_1}{\lambda_1+\lambda_2} = \frac{\lambda_1}{\bar\lambda}.$$

Similarly,

$$P(K = 2) = \frac{\lambda_2}{\bar\lambda}.$$

Moreover J and K are independent. We see this by calculating again with the joint density of W_1, W_2 as follows.

$$\begin{aligned} P(a \le J \le b \text{ and } K = 1) &= P(a \le W_1 \le b \text{ and } W_1 < W_2) \\ &= \int_a^b \int_{w_1}^\infty f(w_1, w_2)\, dw_2\, dw_1 \\ &= \int_a^b \lambda_1 e^{-\lambda_1 w_1} e^{-\lambda_2 w_1}\, dw_1 \\ &= \frac{\lambda_1}{\bar\lambda} \int_a^b \bar\lambda e^{-\bar\lambda w_1}\, dw_1 \\ &= P(K = 1)P(a \le J \le b). \end{aligned}$$

So the the construction (11.4) could be alternately described by starting at $Y_0 = 0$, wait a $\bar\lambda$-exponentially distributed amount of time J, at which time Y jumps to a new state Y_J given by the random variable K:

$$Y_t = \begin{cases} 0 & \text{for } t < J \\ K & \text{for } J \le t. \end{cases}$$

11.2 Examples

The constructions we used above, (11.3) and (11.4), both produce processes which jump only once. We can extend the constructions by starting new (independent) timers after each jump. The examples of this section illustrate this with the most common examples of continuous time Markov chains.

11.2.1 The Poisson Process

We will describe the Poisson process N_t for initial state $N_0 = 0$. ($\tilde{N}_t = k + N_t$ will give a Poisson process with initial state $\tilde{N}_0 = k$.) At time $t = 0$ we start a timer W_1 which determines how long we wait before jumping to state 1. The instant we reach state 1 we start a new timer W_2. When it goes off we jump to state 2 and start a new timer W_3. The process continues in this way. The timers $W_1,\ W_2\ , \ldots$ are assumed to be an i.i.d. sequence of exponential random variables with parameter λ. The successive jump times are

$$J_1 = W_1,\ J_2 = W_1 + W_2,\ J_3 = W_1 + W_2 + W_3, \ldots, J_{n+1} = J_n + W_{n+1}.$$

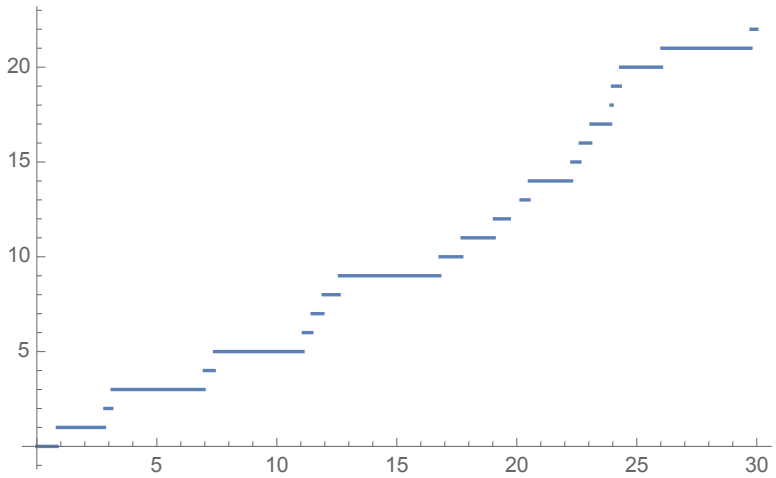
The resulting Markov chain N_t is called the *Poisson process* with parameter λ.

$$\begin{aligned} N_t = &\text{ the number of } J_i \text{ with } J_i \le t \\ &= \max_i\{i : J_i \le t\}. \end{aligned} \tag{11.8}$$

(For this to produce $N_t = 0$ correctly we should include $J_0 = 0$.) N_t is an integer-valued, piecewise constant, continuous time Markov chain. Each jump increases N_t by one unit.

A Poisson process is often used to describe the physical process of radioactive decay. The intermittent clicks of a Geiger counter are produced by alpha-particles emitted by a radioactive substance. They are the jump times J_n above. Each click is a jump in the number N_t of emitted particles detected by the counter up to time t. The number and times of the clicks heard in the past do not influence the probability distribution of the time to the next click - that's the Markov property. Poisson processes are also used to model breakdowns of machinery and arrivals of packets in an electronic communications system, among other things.

Here is a typical path of N_t (for $\lambda = 1$).



At the jump times themselves N_t is taken to be the *new* state, so that N_t is continuous from the right: $N_t = i$ for $J_i \le t < J_{i+1}$. So each horizontal line segment in the graph should be closed on the left and open on the right: $[J_i, J_{i+1})$.

We would typically denote the transition probabilities as

$$p_{i,j}(t) = P_i(N_t = j),$$

but for the Poisson process this only depends on the size of the increment: $j - i$. So we will simplify by writing

$$p_k(t) = P_0(N_t = k)\ (= p_{i,i+k}(t)).$$

Downward jumps are impossible, so $p_{i,j}(t) = 0$ for $i > j$; i.e. $p_k(t) = 0$ for $k < 0$. The Markov property is that (for any history $n_{[0,s]}$, $s < t$ and nonnegative integer k)

$$P(N_t = n_s + k | N_{[0,s]} = n_{[0,s]}) = p_k(t-s). \tag{11.9}$$

We won't write out the technical verification of this – it boils down the the memoryless property (11.2) again but in a more complicated calculation. Instead we will focus on calculating $p_k(t)$.

The first step is to find the density $f_n(\cdot)$ of $J_n = W_1 + \cdots + W_n$. Of course $f_1(t) = \lambda e^{-\lambda t}$ for $t \geq 0$. Since $J_{n+1} = J_n + W_{n+1}$ and J_n and W_{n+1} are independent, we obtain f_n as a convolution of f_n and the exponential density f_1:

$$f_{n+1}(t) = \int_0^t f_n(s)\lambda e^{-\lambda(t-s)}\, ds.$$

It is now straightforward to use the integration by parts to verify by induction that

$$f_n(t) = \lambda^n \frac{t^{n-1}}{(n-1)!} e^{-\lambda t}.$$

With this in hand we can calculate the transition probabilities. Observe that for $k \geq 1$ the event $N_t = k$ is equivalent to $J_k \leq t$ and $t < J_k + W_{k+1}$. Since the random variables J_k and W_{k+1} are independent their joint density is the product of their individual densities. So for $k \geq 1$ we find

$$\begin{aligned}
p_k(t) &= P_0(N_t = k) \\
&= \int_0^t \int_{t-u}^\infty f_k(u)\lambda e^{-\lambda v}\, dv\, du \\
&= \int_0^t f_k(u) e^{-\lambda(t-u)}\, du \\
&= \int_0^t \lambda^k \frac{u^{k-1}}{(k-1)!} e^{-\lambda u} e^{-\lambda(t-u)}\, du \\
&= \frac{\lambda^k t^k}{k!} e^{-\lambda t}.
\end{aligned} \tag{11.10}$$

For $k = 0$

$$\begin{aligned}
p_0(t) &= P_0(N_t = 0) \\
&= P(W_1 > t) \\
&= e^{-\lambda t} \\
&= \frac{\lambda^0 t^0}{0!} e^{-\lambda t},
\end{aligned}$$

conforming to formula (11.10) in the case of $k = 0$ as well. We can now check the Chapman-Kolmogorov equation explicitly, finding that it reduces to the Binomial Theorem (third line):

$$\begin{aligned}
\sum_{j=0}^k p_j(s) p_{k-j}(t) &= \sum_{j=0}^k \frac{\lambda^j s^j}{j!} e^{-\lambda s} \frac{\lambda^{k-j} t^{k-j}}{(k-j)!} e^{-\lambda t} \\
&= e^{-\lambda(s+t)} \frac{\lambda^k}{k!} \sum_{j=0}^k \frac{k!}{j!(k-j)!} s^j t^{k-j} \\
&= e^{-\lambda(s+t)} \frac{\lambda^k}{k!} (s+t)^k \\
&= p_k(s+t).
\end{aligned}$$

The (infinite) transition matrix would be

$$\mathbf{P}(t) = [p_{i,j}(t)] = \begin{bmatrix} p_0(t) & p_1(t) & p_2(t) & \cdots \\ 0 & p_0(t) & p_1(t) & \cdots \\ 0 & 0 & p_0(t) & \cdots \\ \vdots & & & \ddots \end{bmatrix}.$$

In this notation Chapman-Kolmogorov equation becomes

$$\mathbf{P}(s)\mathbf{P}(t) = \mathbf{P}(s+t). \tag{11.11}$$

This is the continuous time version of the matrix power formula (2.4) with the integer powers m, n replaced by continuous time variables s, t.

The following properties are simple consequences of what we have said about the Poisson process.

Proposition 11.1. *A* Poisson process *with intensity $\lambda > 0$ is a Markov process N_t, $t \geq 0$ taking values on $\mathbb{Z}^+$ with the following properties.*

a) $N_0 = 0$ and $N_s \leq N_t$ for $s \leq t$.

b) $N_t - N_s$ is independent of $N_{[0,s]}$ for $s \leq t$.

c)

$$P(N_{t+h} = n + m \,|\, N_t = n) = \begin{cases} 1 - \lambda h + \mathrm{o}(h) & \text{if } m = 0 \\ \lambda h + \mathrm{o}(h) & \text{if } m = 1 \end{cases}$$

The "o(h)" refers to some (unspecified) function $f(h)$ with the property that $\lim_{h\to 0} f(h)/h = 0$. I.e. when h is small o(h) is some quantity which is an order of magnitude smaller, a tiny fraction of h itself. So "$P(\cdots) = \lambda h + \mathrm{o}(h)$" simply means

$$\lim_{h\to 0} \frac{P(\cdots) - \lambda h}{h} = 0.$$

A simple consequence of c) is that

$$\begin{aligned} P(N_{t+h} \geq n+2 \,|\, N_t = n) &= 1 - (P(N_{t+h} = n+0 \,|\, N_t = n) + P(N_{t+h} = n+1 \,|\, N_t = n)) \\ &= 1 - (1 - \lambda h + \mathrm{o}(h) + \lambda h + \mathrm{o}(h)) \\ &= \mathrm{o}(h) + \mathrm{o}(h) \\ &= \mathrm{o}(h). \end{aligned}$$

The interpretation of c) is that over a very small time interval $[t, t+h]$ the probability of no jumps is approximately $1 - \lambda h$, the probability of exactly one jump is approximately λh, and the probability of 2 or more jumps is approximately 0. The error in these approximations is a tiny faction of h itself. This is an infinitesimal description of the behavior of the process. We will say more about it shortly, but first let's look at the very brief proof of the proposition.

Proof. Part a) is trivial since the process is only allowed to jump up.

Part b) is a consequence of the Markov property (11.9):

$$P(N_t - N_s = k | N_{[0,s]}) = P(N_t = k + N_s | N_{[0,s]}) = p_k(t-s)$$

does not depend on $N_{[0,s]}$, so by part 6 of Proposition 3.8 $N_t - N_s$ and $N_{[0,s]}$ are independent.

Since $P(N_{t+h} = n + m \,|\, N_t = n) = p_m(h)$ part c) follows from (11.10). For $m = 0$ we have

$$\frac{p_0(h) - 1 + \lambda h}{h} = \frac{e^{-\lambda h} - 1 + \lambda h}{h} \to 0.$$

For $m = 1$ we have

$$\frac{p_1(h) - \lambda h}{h} = \frac{\lambda h e^{-\lambda h} - \lambda h}{h} = \lambda(e^{-\lambda h} - 1) \to 0,$$

□

Kolmogorov Equations and the Generator

We have calculated the $p_k(t)$ from our wait-and-jump construction of the Poisson process and then used the explicit expressions (11.10) to justify the infinitesimal properties of Proposition 11.1. Most other texts use an infinitesimal approach. If we start with the Chapman-Kolmogorov equation (a consequence of the Markov property) and use part c) of the proposition,

$$\begin{aligned} p_j(t+h) &= \sum_{i=0}^{j} p_{j-i}(t)p_i(h) \\ &= p_j(t)p_0(h) + p_{j-1}(t)p_1(h) + \sum_{i=2}^{j} p_{j-i}(t)p_i(h) \\ &= p_j(t)[1-\lambda h + \mathrm{o}(h)] + p_{j-1}(t)[\lambda h + \mathrm{o}(h)] + \mathrm{o}(h). \end{aligned}$$

The last term is bounded above by $P(N_{t+h} \geq i+2 \,|\, N_t = i) = \mathrm{o}(h)$, using what we said above. So we have

$$p_j(t+h) = (1-\lambda h)p_j(t) + \lambda h p_{j-1}(t) + \mathrm{o}(h)$$

and therefore

$$\frac{p_j(t+h) - p_j(t)}{h} = \lambda[p_{j-1}(t) - p_j(t)] + \mathrm{o}(h).$$

Letting $h \to 0^+$ we find that

$$\begin{aligned} p_j'(t) &= \lambda p_{j-1}(t) - \lambda p_j(t) \text{ for } j \geq 1, \\ p_0'(t) &= -\lambda p_0(t) \text{ for } j = 0. \end{aligned} \tag{11.12}$$

(For $j = 0$ only the first term is present in the above calculation.) These are called the *Kolmogorov equations* for the Poisson process. Since $N_0 = 0$ we have $p_0(0) = 1$ and $p_j(0) = 0$ for $j \geq 1$. These initial conditions uniquely determine the solutions $p_j(t)$ given by the formulae (11.10). Thus if we start with the Chapman-Kolmogorov equation and the infinitesimal properties of Proposition 11.1 we can derive the equations (11.10) and then solve them to determine the transition probabilities. This is a natural approach to finding transition probabilities for continuous time Markov chains in general. This will be considered more carefully in Section 11.3 below.

The calculation above which led the differential equations can be summarized in matrix form as

$$\begin{aligned} \mathbf{P}(t+h) &= \mathbf{P}(t)\mathbf{P}(h) \\ \frac{1}{h}[\mathbf{P}(t+h) - \mathbf{P}(t)] &= \mathbf{P}(t)\ \frac{1}{h}[\mathbf{P}(h) - \mathbf{I}] \\ \mathbf{P}'(t) &= \mathbf{P}(t)\mathbf{P}'(0) \\ \mathbf{P}'(t) &= \mathbf{P}(t)\mathcal{A}, \end{aligned} \tag{11.13}$$

where

$$\mathcal{A} = \mathbf{P}'(0) = \begin{bmatrix} -\lambda & \lambda & 0 & 0 & \cdots \\ 0 & -\lambda & \lambda & 0 & \cdots \\ 0 & 0 & -\lambda & \lambda & \\ 0 & 0 & 0 & -\lambda & \\ \vdots & \vdots & & & \ddots \end{bmatrix}.$$

(Remember that all the diagonal entries of $\mathbf{P}(t)$ are $p_0(t)$ and all the superdiagonal entries are $p_1(t)$.)

The equations (11.13) are a differentiated or "infinitesimal time step" version of the Chapman-Kolmogorov equation. In discrete time we have (2.4), rearranged as a difference equation

$$\mathbf{P}^{n+1} - \mathbf{P}^n = \mathbf{P}^n(\mathbf{P} - \mathbf{I}) = \mathbf{P}^n\mathbf{A}.$$

But in continuous time we have (11.11) which leads to our differential equation

$$\mathbf{P}'(t) = \mathbf{P}(t)\mathcal{A}.$$

Note especially that the role of $\mathbf{A} = (\mathbf{P} - \mathbf{I})$ for discrete time is taken over by $\mathcal{A} = \mathbf{P}'(0)$ in continuous time. This $\mathcal{A}$ is called the *differential generator* or *infinitesimal generator* of the Poisson process. We can think of it as an infinitely big matrix, but for more general continuous time processes it is better to view it as an operator on functions $f(n)$ on the state space $\mathcal{S} = \{0, 1, 2, \ldots\}$ of the Poisson process:

$$\mathcal{A}f(n) = \lambda[f(n+1) - f(n)].$$

We can read off the dynamics of the process from here: from state n the process jumps to state $n+1$ with "rate" λ. See (11.18) for the general case.

Instead of differentiating with respect to h in $\mathbf{P}(t+h) = \mathbf{P}(t)\mathbf{P}(h)$ we could use $\mathbf{P}(t+h) = \mathbf{P}(h)\mathbf{P}(t)$. Following the reasoning of (11.13) this leads to

$$\mathbf{P}'(t) = \mathcal{A}\mathbf{P}(t). \tag{11.14}$$

For the Poisson process with $p_{i,j}(t) = p_{j-i}(t)$ this reduces to the same equations (11.12) but in general it leads to a different set of differential equations. We will see this in our next example.

11.2.2 Pure Birth

Suppose we generalize the Poisson process by allowing the exponential waiting time parameter λ to depend on the state n: we havea $\lambda_n > 0$ associated with each state n. The waiting times in the successive states, $W_n = J_{n+1} - J_n$, are independent exponential random variables, but with different parameters λ_n (so *not* identically distributed). N_t is constructed from such a sequence W_k in the same way as (11.8) above. The result is called a *pure birth process*. Again it will be a Markov chain with state space $\mathcal{S} = \{0, 1, 2, \ldots\}$ except for the possibility of a phenomenon that we haven't encountered before.

Suppose that $\lambda_n \to \infty$. Then in some rough sense the W_n will get smaller as $n \to \infty$. Could it be that $W_n \to 0$ so fast that $\sum_1^\infty W_n < \infty$? If this happened the sequence of jump times would have a finite limit J_∞ called the *explosion time*:

$$J_\infty = \lim_n J_n = \sum_1^\infty W_n < \infty.$$

This in turn would mean that $N_t \to \infty$ as $t \to J_\infty < \infty$, a phenomenon called *explosion in finite time*. If this happens (and it can!) then we have a problem in defining N_t for t beyond J_∞. For the pure jump processes there is a nice criteria for when this does or does not occur.

Lemma 11.2. *Let W_1, W_2, ... be a sequence of independent exponentially distributed random variables with parameters λ_1, λ_2, ... respectively. Let $J_\infty = \sum_1^\infty W_n$. Then*

$$P(J_\infty < \infty) = \begin{cases} 0 & \text{if } \sum_1^\infty 1/\lambda_n = \infty \\ 1 & \text{if } \sum_1^\infty 1/\lambda_n < \infty \end{cases}$$

It is remarkable that $J_\infty < \infty$ has probability either 0 or 1; no values in between are possible! Either J_∞ is certain to be finite or certain to be infinite. (If $J_\infty < \infty$ is certain that does *not* mean that the actual value of J_∞ is certain, only that it is certain to be some finite value.)

Proof. First suppose $\sum_1^\infty 1/\lambda_n < \infty$. For any k we have

$$E\left[\sum_1^k W_n\right] = \sum_1^k E[W_n] = \sum_1^k 1/\lambda_n \le \sum_1^\infty 1/\lambda_n.$$

Since $\sum_1^k W_n \uparrow J_\infty$ the Monotone Convergence Theorem tells us that

$$E[J_\infty] \le \sum_1^\infty 1/\lambda_n < \infty.$$

This is only possible if $P(J_\infty < \infty) = 1$.

Now suppose $\sum_1^\infty 1/\lambda_n = \infty$ and consider e^{-J_∞}. We know that

$$1 \geq e^{-\sum_1^k W_n} \downarrow e^{-J_\infty},$$

so we can use the Dominated Convergence Theorem to say

$$E[e^{-J_\infty}] = \lim_k E[e^{-\sum_1^k W_n}] = \lim_k \prod_1^k E[e^{-W_n}].$$

Now

$$E[e^{-W_n}] = \frac{\lambda_n}{1+\lambda_n} = (1+\lambda_n^{-1})^{-1}.$$

Therefore

$$\prod_1^k E[e^{-W_n}] = \left(\prod_1^k (1+\lambda_n^{-1})\right)^{-1}. \tag{11.15}$$

But if you think about multiplying the product out and discarding some nonnegative terms you will see that

$$\prod_1^k (1+\lambda_n^{-1}) \geq 1 + \sum_1^k 1/\lambda_n \to \infty.$$

Therefore

$$\left(\prod_1^k (1+\lambda_n^{-1})\right)^{-1} \to 0$$

and so

$$E[e^{-J_\infty}] = \lim_k \prod_1^k E[e^{-W_n}] = 0.$$

This is only possible if $P(J_\infty < \infty) = 0$. □

We call the pure birth process N_t *non-explosive* when $P(J_\infty < \infty) = 0$ ad *explosive* when $P(J_\infty < \infty) = 1$. According to the lemma N_t is non-explosive precisely when $\sum 1/\lambda_n = \infty$. In that case N_t is defined for all $0 \leq t < \infty$ and is a Markov chain. The infinitesimal description of Proposition 11.1 remains valid, but with part c) generalized to

$$P(N_{t+h} = n+m \,|\, N_t = n) = \begin{cases} 1 - \lambda_n h + \mathrm{o}(h) & \text{if } m = 0 \\ \lambda_n h + \mathrm{o}(h) & \text{if } m = 1 \\ \mathrm{o}(h) & \text{if } m > 1 \end{cases}$$

The transition probabilities $p_{i,j}(t)$ depend on both i and j, not just their difference $j-i$ as in the Poisson case. We could work out formulas for them by generalizing the calculations of page 177, but they get complicated.

The generator for the pure birth process is described by

$$\mathcal{A}f(n) = \lambda_n[f(n+1) - f(n)]$$

or as an infinite matrix

$$\mathcal{A} = \begin{bmatrix} -\lambda_0 & \lambda_0 & 0 & 0 & \cdots \\ 0 & -\lambda_1 & \lambda_1 & 0 & \cdots \\ 0 & 0 & -\lambda_2 & \lambda_2 & \\ 0 & 0 & 0 & -\lambda_3 & \\ \vdots & \vdots & & & \ddots \end{bmatrix}.$$

We will establish the validity of the Kolmogorov differential equations in Section 11.4. For now we want to observe the form these equations take for the pure birth process. The calculation (11.13) leads to the *forward* equations $\mathbf{P}'(t) = \mathbf{P}(t)\mathcal{A}$. Written out these are

$$\begin{aligned} p'_{i,j}(t) &= \lambda_{j-1}p_{i,j-1}(t) - \lambda_j p_{i,j}(t) \text{ for } 1 \le j, \\ p'_{i,0}(t) &= -\lambda_0 p_{i,0}(t) \text{ for } j = 0, \end{aligned} \tag{11.16}$$

with initial conditions $\mathbf{P}(0) = \mathbf{I}$. The *backward* equation $\mathbf{P}'(t) = \mathcal{A}\mathbf{P}(t)$ of (11.14) is also valid, but works out as

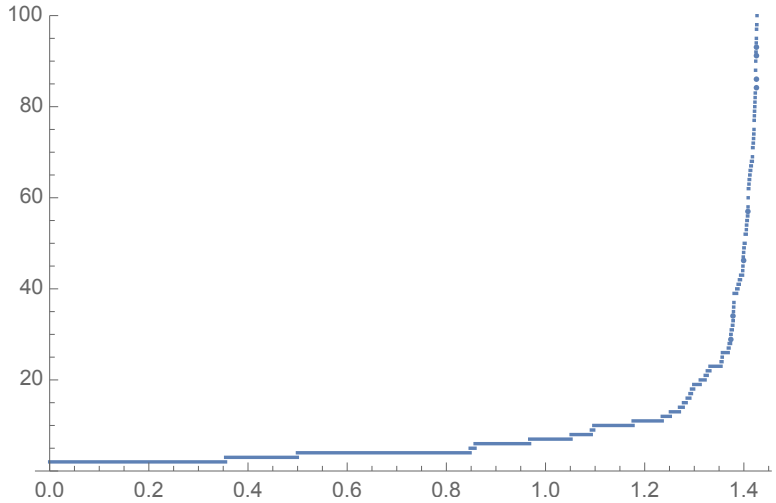
$$p'_{i,j}(t) = \lambda_i p_{i+1,j}(t) - \lambda_i p_{i,j}(t) \text{ for all } i, j \ge 0. \tag{11.17}$$

We see that the forward and backward equations are *not* the same (as they were for the Poisson process), although they both hold.

In the explosive case what we mean by $p_{i,j}(t)$ is unclear because N_t is not well-defined if $J_\infty \le t$. We will consider one way to deal with this in (11.21) below.

Example 11.1. The *simple birth process* has $\lambda_n = n\lambda$. Think of a population consisting of N_t individuals at time t. Each individual, independently of the others, waits an exponential–λ amount of time and then gives birth to one new individual. If there are n individuals present then the time to the next new birth is exponential with parameter $n\lambda$. This is non-explosive because $\sum_1^\infty 1/n = \infty$.

Example 11.2. Suppose the rates are rates $\lambda_n = n(n-1)/2$ and $N_0 = 2$. (Since both $\lambda_0 = \lambda_1 = 0$ the states 1 and 2 are absorbing. That's why we want to start with $N_0 = 2$.) We can think of a collection of particles moving around in a confined space. When two of them collide they survive but produce an additional new particle. The number of different pairs that can be formed from a population of n particles is $\frac{n(n-1)}{2}$, and so the likelihood of a collision in a small amount of time should be proportional to $\frac{n(n-1)}{2}$ if n particles are present. The process is explosive, since $\sum_2^\infty \frac{2}{n(n-1)} < \infty$. By plotting a simulation we can see that the sum of the inter-arrival times is converging to an explosion time J_∞, with $N_t \to \infty$ as $t \to J_\infty$, just as the lemma said.



In this simulation we see $J_\infty \approx 1.4$. This process is like a little nuclear reaction; the occurrence of collisions accelerates until an infinite number of them happen in the instant just before time J_∞.

11.2.3 Birth and Death Processes

Now consider a generalization of the construction (11.4). For each integer n suppose we have two rates: $\lambda_n^\pm$. If $Y_t = n$ we start a pair of exponential timers with rates $\lambda_n^\pm$. The process will jump by ± 1 when the first timer goes off, to $n+1$ if the λ_n^+ timer was first, to $n-1$ if the λ_n^- timer was first. Alternately we can start a single exponential timer with rate $\lambda_n = \lambda_n^+ + \lambda_n^-$. When it goes off we jump up to $n+1$ with probability $q_{n,n+1} = \lambda_n^+/\lambda_n$ and down to $n-1$ with probability $q_{n,n-1} = \lambda_n^-/\lambda_n$. In general this process Y_t will have

the full integers $\mathbb{Z}$ as state space. But if $\lambda_0^- = 0$ then it can't jump down from 0 so we could use just the nonnegative integers. Explosion in finite time is possible here if the $\lambda_n^\pm \to \infty$ in some way as $n \to \pm\infty$. Two types of explosion are possible, $Y_t \to \pm\infty$ as $t \to J_\infty$, depending of course on details of the $\lambda_n^\pm$. It is not so easy to write down simple equivalent conditions for non-explosion. A simple sufficient condition for non-explosion is that the $\lambda_n^\pm$ be bounded; see Problem 11.2. Except in special cases there is little hope of finding explicit formulas for the transition probabilities, but they do satisfy both the forward and backward equations with generator described by

$$\begin{aligned}\mathcal{A}f(n) &= \lambda_n^+[f(n+1) - f(n)] + \lambda_n^-[f(n-1) - f(n)] \\ &= \lambda_n \{q_{n,n+1}[f(n+1) - f(n)] + q_{n,n-1}[f(n-1) - f(n)]\}\end{aligned}$$

11.3 The General Case

We now want to describe the general process of this type. We assume a countable state space $\mathcal{S}$; we will use i, j, k, x, y to denote typical elements of S. For each state i there is a *jump rate* $\lambda_i \geq 0$ and a *jump distribution* $q_{i,j}$. The intuitive idea is that when Y_t arrives in state i, it stays there an exponential-λ_i amount of time, and then jumps to a new state j selected using the $q_{i,j}$ distribution. A state i with $\lambda_i = 0$ is an absorbing state because the waiting time to leave i is infinite. We will insist that $q_{i,i} = 0$ for all states; a state may not jump to itself. Such a jump would be completely undetectable to an observer and awkward to deal with mathematically. So in general,

$$q_{i,i} = 0, \quad q_{i,j} \geq 0, \quad \text{and } \sum_j q_{i,j} = 1.$$

We can describe the construction of Y_t from the parameters λ_i and $q_{i,j}$ in the following way. Observe that $\mathbf{Q} = [q_{i,j}]$ is the transition matrix of a discrete time Markov chain X_n, as in Chapter 4. (Give it the same initial state or initial distribution as Y_0.) This X-chain is the sequence of states Y_t jumps to, but does not account for the waiting times that Y_t spends in the various states. Sometimes X_n is called the *embedded chain* for Y_t. Given the outcomes of the X-chain let W_1 be an exponential-λ_{X_0} random variable, W_2 an exponential-λ_{X_1} random variable, $\dots W_n$ an exponential-$\lambda_{X_{n-1}}$ random variable, $\dots$ all independent of each other except for their shared dependence on the outcome of the X-chain. To be more explicit we could take a i.i.d. sequence $\tilde{W}_1, \tilde{W}_2, \dots$ of exponential-1 random variables, independent of the X-chain, and then take $W_n = \tilde{W}_n / \lambda_{X_{n-1}}$. These are the waiting times between jumps. The jump times are then

$$J_n = \sum_{k=1}^{n} W_k.$$

Specifically, $t = J_n$ is when the $X_{n-1} \to X_n$ jump occurs in Y_t.

$$Y_{J_n} = X_n.$$

We now complete the description of Y_t by making it constant between jumps:

$$Y_t = X_n \text{ for } J_n \leq t < J_{n+1}.$$

This is what we will call the *wait-and-jump* construction of Y_t starting with the jump rates λ_i and jump distributions $q_{i,j}$. This description is how we would simulate Y_t and will be our working understanding of Y_t. In practice it may be more natural to calculate the X_n and W_n one at a time: from $Y_t = i$ use a λ_i-exponential random variable W to determine the next jump time $t + W$ and then a single transition of the $\mathbf{Q}$-chain from i to find the next state j: $Y_{t+W} = j$. Then repeat until the desired final time is reached.

The generator, applied to a function $f : \mathcal{S} \to \mathbb{R}$, is described by the following formula:

$$\mathcal{A}f(i) = \lambda_i \sum_{j \neq i} q_{i,j}[f(j) - f(i)]. \tag{11.18}$$

(The summation is over $j \in \mathcal{S}$ but excluding $j = i$ because $f(i) - f(i) = 0$.) In general this is an infinite series so convergence is an issue. However if f is bounded then the infinite series *is* convergent, since $\sum_j q_{i,j} < \infty$. Equation (11.18) is the most intuitive form for the generator because we can see the $i \to j$ transitions multiplied by their rates and jump distributions. However other arrangements of the terms can sometimes be useful. If we separate out the subtracted terms we get

$$\mathcal{A}f(i) = -\lambda_i f(i) + \sum_{j \neq i} \lambda_i g_{i,j} f(j).$$

As an infinite matrix $\mathcal{A} = [\alpha_{i,j}]$ the entries are

$$\alpha_{i,j} = \begin{cases} \lambda_i q_{i,j} & \text{for } j \neq i \\ -\lambda_i & \text{for } j = i. \end{cases} \tag{11.19}$$

In that notation

$$\mathcal{A}f(i) = \sum_j \alpha_{i,j} f(j),$$

the summation now being over *all* $j \in \mathcal{S}$. Sometimes it is more convenient to specify the values of $\alpha_{i,j}$ rather than λ_i and $q_{i,j}$. See the Examples in Sections 11.3.2 and 11.3.1 below for instance. It is easy to determine λ_i and $q_{i,j}$ from the $\alpha_{i,j}$ for $i \neq j$:

$$\lambda_i = \sum_{j \neq i} \alpha_{i,j} \text{ and } q_{i,j} = \alpha_{i,j}/\lambda_i. \tag{11.20}$$

Explosion is possible if $J_\infty < 0$ has positive probability, where

$$J_\infty = \lim J_n = \sum_1^\infty W_n.$$

We will say that the Markov process Y_t is *non-explosive* if $P_y(\mathcal{T}_\infty < \infty) = 0$ for *all* initial states $y \in \mathcal{S}$. To be explosive *from initial state* y means $P_y(\mathcal{T}_\infty < \infty) > 0$. If we say the *process is explosive* we simply mean explosive *from some initial state*.

A generator of the form described above determines the jump rates and jump distributions. The wait-and-jump construction describes a process Y_t but only up to the explosion time $\mathcal{T}_\infty$. For $\mathcal{T}_\infty \leq t$ we consideer Y_t to be undefined. The literature calls this Y_t the *minimal* process for $\mathcal{A}$. There are various ways to extend the definition to keep the process running past $\mathcal{T}_\infty$. But these require specifying some additional structure beyond $\mathcal{A}$ alone and is a difficult and complicated subject. We will limit our considerations to the minimal process, i.e. those things which occur prior to $\mathcal{T}_\infty$. In particular we can't quite talk about $P_i(Y_t = j)$, but only $P_i(Y_t = j; t < \mathcal{T}_\infty)$. The transition probabilities for the minimal process are taken to be

$$p_{i,j}(t) = P_i(Y_t = j \text{ and } t < \mathcal{T}_\infty). \tag{11.21}$$

Thus the $p_{i,j}(t)$ are the probabilities that the transitions occur *before* explosion. As usual we assemble these into an $\mathcal{S} \times \mathcal{S}$ matrix

$$\mathbf{P}(t) = [p_{i,j}(t)].$$

Bear in mind that this is has infinitely many rows and columns if $\mathcal{S}$ is infinite. A consequence of our definition of the transition probabilities for the minimal chain is that

$$\mathbf{P}(t)[1](i) = \sum_j p_{i,j}(t) = P_i(t < \mathcal{T}_\infty),$$

which may be < 1 in the explosive case. The Markov property, inclusive of the explosion time $\mathcal{T}_\infty$, can be expressed as

$$E[\Phi(Y_{[t,\mathcal{T}_\infty)})|Y_{[0,t]} = y_{[0,t]}] \cdot 1_{t<\mathcal{T}_\infty} = E_{y(t)}[\Phi(Y_{[0,\mathcal{T}_\infty)})] \cdot 1_{t<\mathcal{T}_\infty} \tag{11.22}$$

See (3.22) for comparison with the discrete time case. To write out a detailed argument like that of Section 11.1 to prove the Markov property would be possible, but more tedious than instructive. We will take the Markov property (11.22) for granted and proceed.

If we take

$$\Phi(y_{[0,\zeta)}) = \begin{cases} 1 & \text{if } s < \zeta \text{ and } y_s = j \\ 0 & \text{otherwise,} \end{cases}$$

then (11.22) says that

$$\begin{aligned} P(Y_{t+s} = j \text{ and } t+s < \mathcal{T}_\infty | Y_{[0,t]} = y_{[0,t]}) \cdot 1_{t<\mathcal{T}_\infty} &= P_{y(t)}(Y_s = j \text{ and } s < \mathcal{T}_\infty) \cdot 1_{t<\mathcal{T}_\infty} \\ &= p_{y(t),j}(s) \cdot 1_{t<\mathcal{T}_\infty}. \end{aligned}$$

It follows from the Tower Law for conditional expectations that

$$\begin{aligned} p_{i,j}(t+s) &= E_i[P(Y_{t+s} = j \text{ and } t+s < \mathcal{T}_\infty | Y_{[0,t]}) \cdot 1_{t<\mathcal{T}_\infty}] \\ &= E_j[1_{t<\mathcal{T}_\infty} p_{Y_t,j}(s)] \\ &= \sum_k p_{i,k}(t) p_{k,j}(s). \end{aligned}$$

This is the Chapman-Kolmogorov equation for the minimal process. In matrix form it says

$$\mathbf{P}(t+s) = \mathbf{P}(t)\mathbf{P}(s)$$

11.3.1 A Chemical Kinetics Example

A contemporary application area is stochastic chemical kinetics. Imagine a collection of different chemical substances A, B, C, ... mixed together in a container where the different kinds of molecules may make contact with each other and react. Let Y_t^A, Y_t^B, Y_t^C, ... denote the numbers of molecules of the different types that are present at time t. (Because we are reserving the subscript position for the time variable we have put the "A", "B", "C" in the superscript position.) The vector of all these molecular counts $Y_t = (Y_t^A, Y_t^B, Y_t^C, \ldots)$ will be the state of the Markov chain. When a chemical reaction occurs these numbers change.

For instance consider a reaction which we will write as

$$A + B \to C.$$

In this reaction one molecule of A and one molecule of B combine to form one molecule of C. The time when this happens will be one of the jump times J_k, and when it does both Y^A and Y^B will decrease by one and Y^C will increase by one. To be more precise

$$\begin{aligned} (Y_{J_k}^A, Y_{J_k}^B, Y_{J_k}^C, \ldots) &= (Y_{J_k-}^A, Y_{J_k-}^B, Y_{J_k-}^C, \ldots) + (-1, -1, 1, 0 \cdots) \\ Y_{J_k} &= Y_{J_k-} + \nu \end{aligned}$$

Here $\nu = (-1, -1, 1, 0 \cdots)$ is the *state change vector* for this particular reaction. Given all the molecular counts the time until the next reaction *of this type* is random with a Poisson-like description:

$$P(\text{ this type of reaction occurs in } (t, t+h] | Y_t = y) \approx \alpha_{y,y+\nu} h,$$

the "$\approx$" is because some other reaction might occur first so the waiting time is not precisely exponential. This is b) of Lemma 11.4 below in the general case. The jump coefficient $\alpha_{y,y+\nu}$ (see (11.19)) has values determined by physical principles and the number of reactants involved. For our "second order" reaction $A + B \to C$ the form is

$$\alpha_{y,y+\nu} = c\, y^A y^B.$$

Here c is a parameter related to how likely the reaction is to occur when an A-molecule and B-molecule do actually collide. Each different reaction has a different state change vector ν, parameter c and formula for $\alpha_{y,y+\nu}$ depending on the numbers of reacting molecules involved. Together the combined reactions produce

a stochastic process Y_t of the general type considered in this chapter. The generator $\mathcal{A}f(y)$ consists of a sum of terms, one for each type of reaction.

$$\mathcal{A}f(y) = \sum_{\text{reactions } \nu} \alpha_{y,y+\nu}[f(y+\nu) - f(y)].$$

From here we could determine the jump rates

$$\lambda_y = \sum_{\text{reactions } \nu} \alpha_{y,y+\nu}$$

and then the jump distributions

$$q_{y,y+\nu} = \alpha_{y,y+\nu}/\lambda_y$$

as in (11.20). Most $q_{y,z}$ will be 0; only those z for which there is a reaction ν with $z = y + \nu$ will have a nonzero $q_{y,z}$. This is why the form of $\mathcal{A}f(y)$ above is more convenient than (11.18) for these examples.

The Michaelis-Menton model of enzyme kinetics is a well-studied example which illustrates the description above. In this model an enzyme B converts a "substrate" A into an end product D by means of an intermediate product C. This involves three reactions:

$$\begin{aligned}
&\text{reaction \#1: } A + B \to C;\ \nu_1 = (-1,-1,1,0)\\
&\text{reaction \#2: } C \to A + B;\ \nu_2 = (1,1,-1,0)\\
&\text{reaction \#3: } C \to B + D;\ \nu_3 = (0,1,-1,1).
\end{aligned}$$

In reaction #1 the enzyme B combines with the substate A to produce the intermediate product C. The intermediate product can disassociate back into its original components in reaction #2 or release the enzyme as it transforms to the final product D in reaction #3. The reaction "propensities" take the following forms:

$$\begin{aligned}
\alpha_{y,y+\nu_1} &= c_1 y^A y^B\\
\alpha_{y,y+\nu_2} &= c_2 y^C\\
\alpha_{y,y+\nu_3} &= c_3 y^C
\end{aligned}$$

The resulting generator is

$$\begin{aligned}
\mathcal{A}f(y) = \ & c_1 y^A y^B [f(y^A - 1, y^B - 1, y^C + 1, y^D) - f(y)]\\
&+ c_2 y^C [f(y^A + 1, y^B + 1, y^C - 1, y^D) - f(y)]\\
&+ c_3 y^C [f(y^A, y^B + 1, y^C - 1, y^D + 1) - f(y)]
\end{aligned}$$

If we start this process with $Y^A(0) > 0,\ Y^B(0) > 0,\ Y^C(0) = Y^D(0) = 0$ it will eventually convert all the substrate to end product with the total amount of enzyme preserved: $Y_t^A \to 0,\ Y_t^B \to Y^B(0),\ Y_t^C \to 0,\ Y_t^D \to Y^A(0)$ as $t \to \infty$.

Chemical systems underly most biological processes, and so understanding their properties is important to current research in cell biology. In most cases the number of molecules needs to be quite large for the chemical concentrations to be large enough to have a biological effect. It makes sense to introduce a scaling parameter $N = \text{vol} \cdot n_A$ where vol is the volume of the container in which the reaction is taking place and n_A is Avagadro's number ($\approx 6.023 \times 10^{23}$), the number of molecules in 1 mole. Then $Y^A = N$ corresponds to a concentration of 1 mole/liter of substance A. The constant c_1 in reaction #1 depends on N, because the density of molecules per unit volume clearly influences the probability of the A & B collision needed for reaction #1. (The other reactions don't depend on molecules coming together so don't depend on N.) This leads to an N-dependent choice of the constants, such as (taken from [27])

$$c_1 = \frac{10^6}{N},\ c_2 = 10^4,\ c_3 = 10^{-1}.$$

The initial values may be given in terms of concentrations (mole/liter) as well. This leads to a Markov chain with a parameter N in its specification. There has been considerable research on the Markov chain's

behavior for very large N-values (including descriptions in the limit as $N \to \infty$). Simulations become more challenging when parameters and state values are very large and/or small. The simulation idea described in Section 11.3 has come to be called the *Gillespie algorithm* in these applications. The basic mathematical idea behind the simulation is not new in Gillespie's work, but he and others introduced enhancements and approximations that made it more efficient and effective for use with complicated chemical kinetics processes involving large numbers of transitions over short time periods. Higham [27] provides a nice overview of this application area and references to the literature. Some simulations of the Michaelis-Menton example are presented there.

11.3.2 A Queueing Network Example

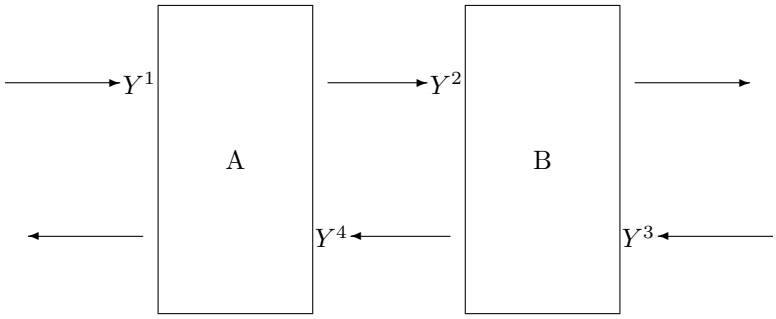
Markov jump processes are also important in queueing network theory. A queue is a waiting line. Picture the line waiting to check out at the grocery store. Suppose the times between when new customers join the waiting line are i.i.d. exponential random variables with parameter λ^+. The times needed to check out with the cashier are given by i.i.d. exponential random variables, λ^-. The resulting queue length process Y_t can jump either up by 1 or down by 1. It is a birth & death process as in Section 11.2.3. The only dependence of the rates on the state n is that if the queue is empty $Y_t = n = 0$ then it cannot jump down:

$$\lambda_n^+ = \lambda^+, \quad \lambda_n^- = \begin{cases} \lambda^- & \text{if } n > 0 \\ 0 & \text{if } n = 0. \end{cases}$$

This is called an M/M/1 queue in the literature.

Now imagine a network of such queues. When a customer is finished in one queue they go get in line at another queue, according to some routing mechanism. At each queue each item requires a new independent service time (exponential distribution, with a parameter depending on the queue). Moreover several queues may depend on the same server, which must follow some protocol to determine which of the queues delegated to it will receive its attention next. This is called a *queueing network*. Processes of this type describe the flow of work through a manufacturing facility or messages through a networked communication system, and can often be described using continuous time Markov chains of the type we are discussing.

We want to look at a particular example, often called the Kumar-Sideman network. There will be four queues and two servers (A and B). The state consists of a 4-vector of nonnegative integers $y = (y^1, y^2, y^3, y^4) \in (\mathbb{Z}^+)^4$ counting the numbers of items in each of the four queues. (Again we are using superscripts instead of subscripts to index the queues to reserve the subscript position for the time variable. Try to remember that the superscripts do *not* mean exponents here.) Server A attends to queues 1 and 4; server B attends to queues 2 and 3. New customers can arrive in both queues 1 and 3. When a customer is finished in queue 1 he joins queue 2. When finished in queue 2 he is done and leaves the system. When a customer in queue 3 is finished he joins queue 4, and when finished there leaves the system. This organization is illustrated with the following diagram.



The arrival rates for queues 1 and 3 are both 1. The service rates for queues 1 and 3 are 10. The service rates for queues 2 and 4 are 5/3. Finally the service protocol for A is that queue 4 has higher priority: A only serves queue 1 if queue 4 is empty. Similarly B only serves queue 3 if queue 2 is empty.

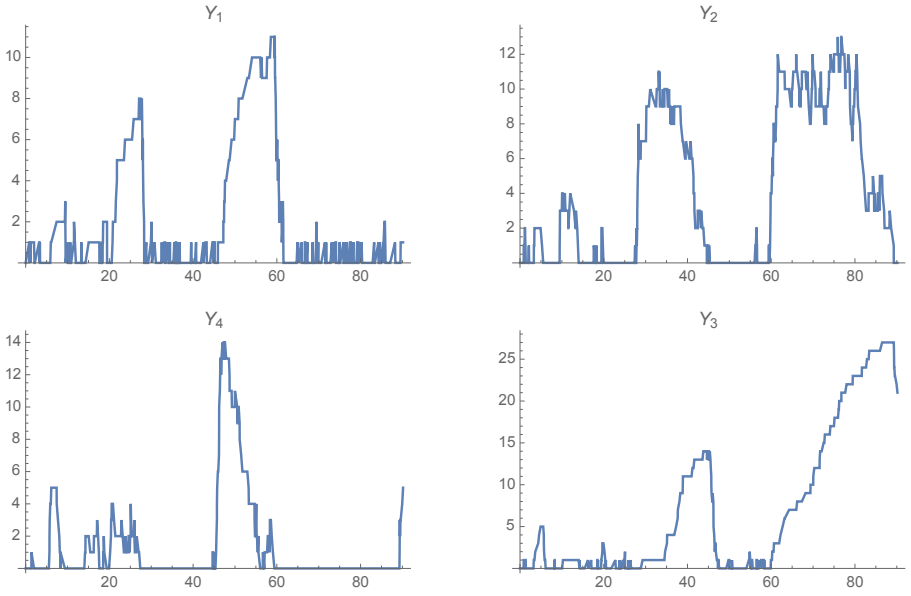
This again is a situation in which it is more convenient to describe the generator in terms of the combined jump rate and distribution, $\alpha_{y,y+\nu} = \lambda_y q_{y,y+\nu}$, rather than λ_y and $q_{y,y+\nu}$ separately. From a state y there

are 6 possible transitions $y \to y+\nu$, i.e. 6 possible choices of ν for which $\alpha_{y,y+\nu} = \lambda_y q_{y,y+\nu}$ could be nonzero, one for each arrival or service event.

- Arrival in queue #1:
$$\nu = (1,0,0,0), \quad \alpha_{y,y+\nu} = 1.$$
- A-Service of item in queue #1:
$$\nu = (-1,1,0,0), \quad \alpha_{y,y+\nu} = \begin{cases} 10 & \text{if } y^4 = 0 \text{ and } y^1 > 0 \\ 0 & \text{otherwise.} \end{cases}$$
- B-Service of item in queue #2:
$$\nu = (0,-1,0,0), \quad \alpha_{y,y+\nu} = \begin{cases} 5/3 & \text{if } y^2 > 0 \\ 0 & \text{otherwise.} \end{cases}$$
- Arrival in queue #3:
$$\nu = (0,0,1,0), \quad \alpha_{y,y+\nu} = 1.$$
- B-Service of item in queue #3:
$$\nu = (0,0,-1,1), \quad \alpha_{y,y+\nu} = \begin{cases} 10 & \text{if } y^2 = 0 \text{ and } y^3 > 0 \\ 0 & \text{otherwise.} \end{cases}$$
- A-Service of item in queue #4:
$$\nu = (0,0,0,-1), \quad \alpha_{y,y+\nu} = \begin{cases} 5/3 & \text{if } y^4 > 0 \\ 0 & \text{otherwise.} \end{cases}$$

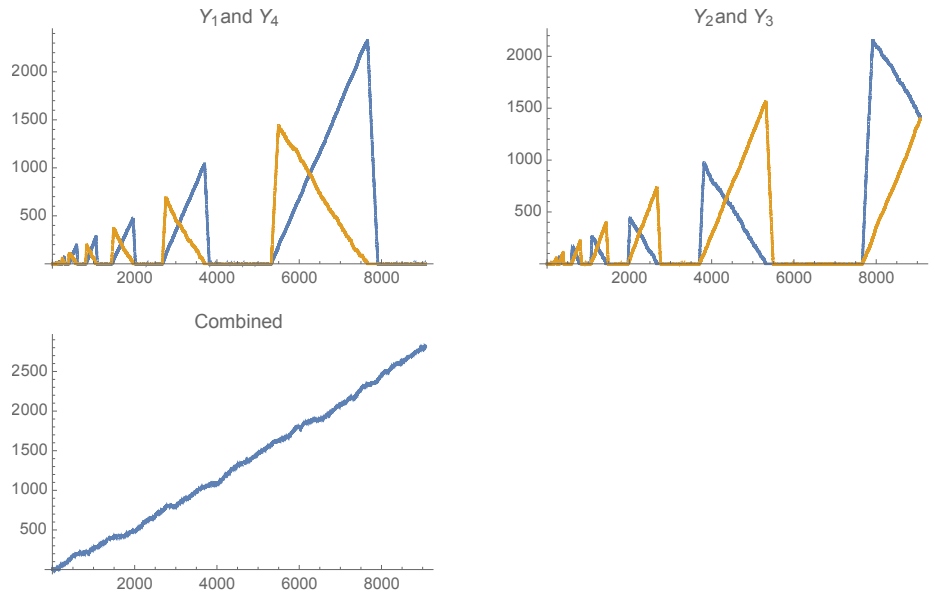
The generator is then
$$\mathcal{A}f(y) = \sum_{\nu} \alpha_{y,y+\nu}[f(y+\nu) - f(y)],$$
the summation being over the six choices of ν above.

Let's look at a simulation over a modest time interval of about 90 time units.



If you look closely you will see that Y^1 does not decrease while $Y^4 > 0$. Similarly Y^3 does not decrease while $Y^2 > 0$. This is because of the server priorities described above.

Now let's look at a simulation over a longer time scale, about 9000 time units. We have plotted Y^1 and Y^4 together (Y^1 is blue) and Y^2 and Y^3 together (Y^2 is blue). The third plot is the sum, $Y^1+Y^2+Y^3+Y^4$.



There is clearly a regular pattern emerging, and escalating as time proceeds. Y^4 gets a big boost up when Y^2 hits zero and releases the backed up items in queue 3. As Y^4 fills up it blocks service to items in Y^1 so Y^1 grows, until Y^4 reaches 0. Then Y^1 empties rapidly into Y^2, which causes Y^3 to fill up while Y^2 is emptying. When Y^2 reaches zero the cycle begins over again, but with more items in the system than before. The number of items in all the queues combined is climbing steadily. If this were a real system the queues would have finite capacities, which would eventually be exceeded and the system would fail. Based on our simulation it appears that $\mathbf{Y}_t$ is non-eplosive, but transient in that $|\mathbf{Y}_t| \to \infty$.

One lesson of this example is that careful design of a network, its service rates and protocols can be important for reliable performance. In particular methods to discern regular behavior that emerges over long time periods, like what we observed above, can be very useful. See Chen and Yao [12] for an introduction to queueing networks, and their Chapter 8 for more on the Kumar-Sideman example specifically.

11.4 Kolmogorov's Equations

We want to focus on two aspects of Markov jump processes that are substantially different from the discrete time case: the description of transition probabilities using differential equations (in this section) and the phenomenon of explosion (in Section 11.6). Other topics, such as recurrence and equilibrium, are important but we let readers consult other references for those.

We saw that the transition probabilities for the Poisson process satisfied a set of differential equations which could be expressed concisely in terms of the generator (11.13). We want to do this for a Markov jump processes in general. But we will have to face several technical difficulties. One is that in the general case the generator (11.18) involves an infinite series. We have to worry about convergence and use care when interchanging integrals and derivatives with infinite series expressions. Another is the possibility of explosion. Remember that we have a meaning for $p_{i,j}(t)$ in the case of explosion; see (11.21). But we can't assume $\sum_j p_{i,j}(t) = 1$, only that it is ≤ 1.

We are going to prove that in general $\mathbf{P}(t) = [p_{i,j}(t)]$ solves two different systems of differential equations. The first are called the *Kolmogorov forward equations*, written concisely as

$$\mathbf{P}'(t) = \mathbf{P}(t)\mathcal{A}. \tag{11.23}$$

We interpret this entry-by-entry, meaning that for each pair $i, j \in \mathcal{S}$

$$\begin{aligned} p'_{i,j}(t) &= (\mathbf{P}(t)\mathcal{A})_{i,j} \\ &= -p_{i,j}(t)\lambda_j + \sum_{k \neq j} p_{i,k}(t)\lambda_k q_{k,j}. \end{aligned} \tag{11.24}$$

The right side is in general an infinite series over k. Notice also that the right side involves those $p_{i,k}(t)$ with the same first index i but *all possible values of the second index* k. In other words it involves the i^{th} row of $\mathbf{P}(t)$.

The equations of the second system are called the *Kolmogorov backward equations*, expressed as

$$\mathbf{P}'(t) = \mathcal{A}\mathbf{P}(t). \tag{11.25}$$

Again this is to be understood entry-by-entry:

$$\begin{aligned} p'_{i,j}(t) &= (\mathcal{A}\mathbf{P}(t))_{i,j} \\ &= \sum_{k \neq i} \lambda_i q_{i,k}[p_{k,j}(t) - p_{i,j}(t)] \\ &= -\lambda_i p_{i,j}(t) + \sum_{k \neq i} \lambda_i q_{i,k} p_{k,j}(t). \end{aligned} \tag{11.26}$$

Now observe that the right side involves those $p_{k,j}$ with the same second index j but *the first index taking all possible values* k. In other words this system involves j^{th} column of $\mathbf{P}(t)$. Again the right side is an infinite series (over k) in general.

More generally, given a (bounded) function $\phi : \mathcal{S} \to \mathbb{R}$, the backward equations describe how $E_i[f(Y_t)]$ evolves over time. If we let $u(i,t) = E_i[\phi(Y_t)]$ then this will satisfy the system of backward equations

$$\mathbf{u}'(t) = \mathcal{A}\mathbf{u}(t).$$

with the initial condition $\mathbf{u}(0) = [\phi(i)]$. This is d) of the theorem below. There is a sort of time-reversal here, which is one reason to call these the "backward" equations. If we associate $\phi(\cdot)$ with a fixed time T then $u(\cdot,\cdot)$ is associated with an *earlier* time:

$$u(Y_t, T-t) = E[\phi(Y_T)|Y_{[0,t]}].$$

Increasing the value $(T-t)$ in the time position of g corresponds to working back to an earlier t in the conditional expectation. (The transition probabilities themselves are the special case $\phi(\cdot) = 1_j(\cdot)$.)

In brief, time dependence of the distribution of Y_t is described by the forward equation and time dependence of the expected values $E_i[\phi(Y_t)]$ of a function of Y_t is described by the backward equation.

Properties of $\mathcal{A}$

Before proceeding we need to make some observations about $\mathcal{A}$. If f is a bounded function on $\mathcal{S}$ then we can be sure that the infinite series in $\mathcal{A}f(i)$ is convergent. If $|f| \leq B$ then since $\sum_j q_{i,j} = 1$ we know that $\sum_j \lambda_i q_{i,j} B$ is convergent, so $\sum_j \lambda_i q_{i,j} f(j)$ converges (absolutely) by the dominated series test, and

$$|\mathcal{A}f(i)| \leq B\lambda_i.$$

The forward equation involves a multiplication on the left of $\mathcal{A}$. We will always be multiplying on the left by something interpreted as a distribution ν on $\mathcal{S}$, viewed as a row for purposes of multiplication on the left. We will use subscripts: ν_i for $i \in \mathcal{S}$ and will only consider $0 \leq \nu_i$. We understand the notation $\nu\mathcal{A}$ to refer to the row vector with components

$$\begin{aligned} (\nu\mathcal{A})_j &= \sum_i \nu_i \alpha_{i,j} \\ &= -\nu_j \lambda_j + \sum_{i \neq j} \nu_i \lambda_i q_{i,j}. \end{aligned}$$

Since $0 \le \nu_i$ only one term of the series can be negative. So we can always ascribe a value to this if we allow $+\infty$ when the series diverges. Since $0 \le q_{i,j} \le 1$ a simple sufficient condition for convergence is that $\sum_i \nu_i \lambda_i < \infty$. If $\lambda_i \le M$ then $\sum_i \nu_i < \infty$ is sufficient.

Here are the principal results we want to prove about the Kolmogorov equations, gathered into a single theorem.

Theorem 11.3. *Let Y_t be the* minimal *continuous time Markov chain with generator $\mathcal{A}$ on a countable state space $\mathcal{S}$ and $\mathbf{P}(t) = [p_{i,j}(t)]$ its matrix of transition probabilities,* (11.21).

a) $\mathbf{P}(t)$ satisfies the backward equations: $\mathbf{P}'(t) = \mathcal{A}\mathbf{P}(t)$. Specifically for each pair of states i, j the transition probability $p_{i,j}(t)$ is a continuously differentiable function of $t \ge 0$ with derivative given by

$$p'_{i,j}(t) = -\lambda_i p_{i,j}(t) + \sum_{k \ne i} \lambda_i q_{i,k} p_{k,j}(t).$$

b) $\mathbf{P}(t)$ satisfies the forward equations: $\mathbf{P}'(t) = \mathbf{P}(t)\mathcal{A}$. Specifically for each pair of states i, j the transition probability $p_{i,j}(t)$ is a continuously differentiable function of $t \ge 0$ with derivative given by

$$p'_{i,j}(t) = -p_{i,j}(t)\lambda_j + \sum_{k \ne j} p_{i,k}(t)\lambda_k q_{k,j}.$$

c) Any nonnegative continuously differentiable solution of the forward equations $\mathbf{Q}'(t) = \mathbf{Q}(t)\mathcal{A}$ with $\mathbf{Q}(0) = \mathbf{I}$ satisfies $\mathbf{P}(t) \le \mathbf{Q}(t)$ (componentwise). Likewise any nonnegative continuously differentiable solution of the backward equations $\mathbf{Q}'(t) = \mathcal{A}\mathbf{Q}(t)$ with $\mathbf{Q}(0) = \mathbf{I}$ satisfies $\mathbf{P}(t) \le \mathbf{Q}(t)$ (componentwise).

d) Let $\phi : \mathcal{S} \to \mathbb{R}$ be a bounded function, with $|\phi(i)| \le B$ for all $i \in \mathcal{S}$.. The function $\mathbf{u}(t) = [u(i,t)]$ where

$$u(i,t) = \sum_j p_{i,j}(t) f(j) = E_i[\phi(Y_t)]$$

satisfies the backward equations $\mathbf{u}'(t) = \mathcal{A}\mathbf{u}(t)$. Specifically for each $i \in \mathcal{S}$

$$\frac{d}{dt} u(i,t) = -\lambda_i u(i,t) + \sum_{k \ne i} \lambda_i q_{i,k} u(k,t). \tag{11.27}$$

The infinite series on the right converges to a continuous function satisfying the bound $|\frac{d}{dt} u(i,t)| \le \lambda_i B$.

e) Let $\phi : \mathcal{S} \to \mathbb{R}$ be a bounded function. The Markov chain is non-explosive if and only (11.27) *has a unique bounded solution with $\mathbf{u}(0) = [\phi(i)]$.*

Let's try to digest what this is saying. First keep in mind that the $p_{i,j}(t)$ are the transition probabilities for the *minimal* chain, defined as in (11.21). Parts a) and b) say that the backward and forward equations *do* hold and that the infinite series in $\mathbf{P}(t)\mathcal{A}$ and $\mathcal{A}\mathbf{P}(t)$ *are* convergent and yield continuous functions of t. Part d) is a generalization of a) to any bounded initial condition ϕ. We have included it because it will be used in the proof of part e).

The next question is uniqueness. If we find some solution $\mathbf{Q}(t)$ of either the forward or backward equations with the correct initial values $\mathbf{Q}(0) = \mathbf{I}$ can we be sure that $\mathbf{Q}(t) = \mathbf{P}(t)$? In the finite state case, yes, but in general no! So what can we say about the relation of $\mathbf{P}(t)$ to other possible solutions $\mathbf{Q}(t)$? Part c) says something about this, namely that *if* $\mathbf{Q}(t) \ge 0$ then $\mathbf{P}(t) \le \mathbf{Q}(t)$ for either the forward or backward equations. In other words $\mathbf{P}(t)$ is always the smallest among all *nonnegative* solutions $\mathbf{Q}(t)$. But can there exist solutions which are either larger than $\mathbf{P}$ or take some negative values? For the backward equation part e) says the answer is yes if the chain is explosive, but no (at least for *bounded* solutions) if the chain is non-explosive. For the forward equation this is a difficult question to answer. The next two examples will illustrate.

Example 11.3. Consider the forward equations for the pure birth process (11.16). We know that this can be either explosive or non-explosive depending on the choice of jump rates. But solutions of the forward equations are unique regardless. This is easy to see because the equations can be solved one j at a time. Given any set of initial values $p_i(0)$ the forward equation for $p_0(t)$ is

$$p_0'(t) = -\lambda_0 p_0(t),$$

which has the unique solution

$$p_0(t) = e^{-\lambda_0 t} p_0(0).$$

For $j > 0$ once $p_{j-1}(t)$ is determined we see that the forward equation

$$p_j'(t) = -\lambda_j p_j(t) + \lambda_{j-1} p_{j-1}(t)$$

has the unique solution

$$p_j(t) = e^{-\lambda_j t} p_j(0) + \int_0^t e^{-\lambda_j (t-s)} \lambda_{j-1} p_{j-1}(s)\, ds.$$

So even if the process is explosive there is still only one solution to the forward equations with given initial conditions.

Example 11.4. Suppose we take a pure birth process but *reverse the direction of the jumps*. When the process leaves state i it goes to state $i-1$ (instead of $i+1$). From any initial state $Y(0) = n$ the process jumps its way down through the positive integers until it reaches 0. If we make $\lambda_0 = 0$ then 0 is an absorbing state so that once 0 is reached the chain never jumps again. Perhaps we should call this a *pure death process*. Clearly this is a non-explosive chain. In Problem 11.8 you will show that the forward equation $\mathbf{q}'(t) = \mathbf{q}(t)\mathcal{A}$ has a non-zero solution (a row $\mathbf{q}(t) = [q_0(t), q_1(t), \ldots]$) with $\mathbf{q}(0) = \mathbf{0}$. This means that the solutions to $\mathbf{Q}'(t) = \mathbf{Q}(t)\mathcal{A}$ with $\mathbf{Q}(0) = \mathbf{I}$ are *not* unique. There are always infinitely many solutions.

We might hope that none of these extra solutions are nonnegative. But that is not true either! Suppose we choose jump rates so that $\sum_1^\infty 1/\lambda_i < \infty$. Let W_n be independent exponentially distributed waiting times with parameters λ_n. Then $\sum_1^\infty W_n < \infty$ with probability 1, by Lemma 11.2. Imagine constructing a process Z_t which starts at "$Z_0 = \infty$" and jumps from n to $n-1$ at time

$$J_n = \sum_{k=n}^{\infty} W_k.$$

So

$$Z_t = n \text{ for } J_{n-1} \le t < J_n \text{ and } Z_t = 0 \text{ for } J_1 = \sum_1^\infty W_n \le t.$$

This is just like our construction of the pure birth process except that this one starts at ∞, makes infinitely many downward jumps in the first fraction of a second and continues jumping down until it eventually reaches 0. We can define

$$q_n(t) = P(Z_t = n) = P(J_{n-1} \le t < J_n).$$

Clearly $q_n(0) = 0$ for all n. And although we won't write out a proof, it should not be hard to accept that the forward equations

$$q_n'(t) = -\lambda_n q_n(t) + \lambda_{n+1} q_{n+1}(t)$$

will be satisfied, just as their counterparts for the true pure birth process are. This describes a solution of $\mathbf{q}'(t) = \mathbf{q}(t)\mathcal{A}$ with $\mathbf{q}(0) = [0]$, $0 < q_n(t)$ for all n and all $t > 0$, and $\sum q_n(t) = 1$. We can add multiples of $\mathbf{q}$ to the rows of $\mathbf{P}$ to get infinitely many nonnegative solutions of $\mathbf{Q}'(t) = \mathbf{Q}(t)\mathcal{A}$ with $\mathbf{Q}(0) = \mathbf{I}$. Thus even though the pure death process Y_t is non-explosive, the forward equation for $\mathbf{P}(t)$ has many nonnegative solutions with the correct initial conditions.

The rest of this section is devoted to the proof of Theorem 11.3. This will take several pages and lots of work. Some readers may not want to go through all the details. Here is an overview in case you want to just skip over the details and go on to the next section.

- First we will assume that the jump rates are bounded: $\lambda_i \leq M$ for all i. Under this assumption we will prove a generalization of part c) of Proposition 11.1 which is more precise about the $o(h)$ terms: Lemma 11.4 just below.
- Using the lemma the proof of the backward and forward equations is not hard: Theorem 11.5.
- The case of a finite state space is covered by Theorem 11.5. The connection with matrix exponentials is described in Section 11.4.2.
- The assumption of bounded rates rules out explosive processes. To treat the general case we consider a stopped version of the process in which we make all the states outside a finite set K absorbing (jump rates of 0). This is what we call the K-process. It too falls within the scope of Theorem 11.5. This is described in Section 11.4.3.
- Finally we get the general case by passing to the limit as $K \uparrow \mathcal{S}$. Lemma 11.6 gives the basic convergence. The proof of Theorem 11.3 will finally come in Section 11.4.4.

11.4.1 Bounded Rates

We will say that the process has *bounded rates* if there is a constant M so that

$$\lambda_i \leq M \text{ for all } i \in \mathcal{S}.$$

The following lemma gives us a more precise version of Proposition 11.1 part c).

Lemma 11.4. *Assume the Markov jump process Y_t has bounded rates: $\lambda_i \leq M$ and $0 < h$.*

a) For $j \neq i$

$$p_{i,j}(h) = \lambda_i q_{i,j} h + o_{i,j}(h),$$

where $|o_{i,j}(h)| \leq M^2 h^2$.

b)

$$p_{i,i}(h) = 1 - \lambda_i h + o_{i,i}(h),$$

where $|o_{i,i}(h)| \leq M^2 h^2$.

c) For any bounded $\phi : \mathcal{S} \to \mathbb{R}$ with $|\phi| \leq B$,

$$E_i[\phi(Y_h)] = \phi(i) + \sum_{j \neq i} \lambda_i q_{i,j}[\phi(j) - \phi(i)] + o_\phi(h),$$

where $|o_\phi(h)| \leq 2BM^2 h^2$.

We recognize in part c) the generator as we defined it in (11.18).

$$\mathcal{A}\phi(i) = \sum_j \lambda_i q_{i,j}[\phi(j) - \phi(i)].$$

In fact part c) is the backward equation of part d) of Theorem 11.3 at $t = 0$.

Proof. For small h the most likely way for $Y(h) = j \neq i = Y(0)$ is for a single jump to occur between 0 and h: $0 < J_1 \leq h < J_2$ and $Y(J_1) = j$. But it is also possible to make the transition in 2 or more jumps, which would imply $J_2 \leq h$. We can bound $p_{i,j}(h)$ as follows.

$$P_i(Y_h = j; h < J_2) \leq p_{i,j}(h) \leq P_i(Y_h = j; h < J_2) + P_i(J_2 \leq h).$$

We will show that

$$P_i(J_2 \leq h) = o(h) \text{ where } 0 \leq o(h) \leq \frac{1}{2} M^2 h^2, \tag{11.28}$$

and for $i \neq j$

$$P_i(Y_h = j; h < J_2) = \lambda_i q_{i,j} h - q_{i,j} \mathrm{o}_j(h) \text{ where } 0 \leq \mathrm{o}_j(h) \leq M^2 h^2. \tag{11.29}$$

Part a) will then follow using $\mathrm{o}_{i,j} = \max(q_{i,j}\mathrm{o}_j(h), \mathrm{o}(h))$ because

$$-q_{i,j}\mathrm{o}_j(h) \leq p_{i,j}(h) - \lambda_i q_{i,j} h \leq -q_{i,j}\mathrm{o}_j(h) + \mathrm{o}(h).$$

Let's examine (11.28) first. Assume $Y_0 = i$. The time of the second jump is $J_2 = W_1 + W_2$ where W_1 is exponential with parameter λ_i and W_2 is exponential with parameter λ_j *given* that the outcome of the first jump is j, $P_i(Y_{W_1} = j) = q_{i,j}$. The probability of two or more jumps in $[0, h]$ is

$$\begin{aligned} P_i(J_2 \leq h) &= \sum_j q_{i,j} \int_0^h \int_0^{h-s} \lambda_i e^{-\lambda_i s} \lambda_j e^{-\lambda_j t} \, dt \, ds \\ &\leq M^2 \int_0^h \int_0^{h-s} 1 \, dt \, ds \\ &= M^2 \int_0^h h - s \, ds \\ &= M^2 h^2/2, \end{aligned}$$

as claimed.

Next consider (11.29). To say $Y_h = j; h < J_2$ means that $W_1 < h$, the first jump is to j and then $W_1 + W_2 > h$ so there is no additional jump before time h. The probability of this is the following.

$$\begin{aligned} P(W_1 < h < W_1 + W_2) &= q_{i,j} \int_0^h \int_{h-s}^\infty \lambda_i e^{-\lambda_i s} \lambda_j e^{-\lambda_j t} \, dt \, ds \\ &= q_{i,j} \int_0^h \lambda_i e^{-\lambda_i s} e^{-\lambda_j (h-s)} \, ds \\ &= \lambda_i q_{i,j} \left[h - \int_0^h 1 - e^{-\lambda_i h - \lambda_j (h-s)} \, ds \right]. \end{aligned}$$

Now $0 \leq 1 - e^{-\lambda_i h - \lambda_j (h-s)} \leq \lambda_i h + \lambda_j (h - s)$ so

$$\begin{aligned} 0 &\leq \mathrm{o}_j(h) \\ &= \lambda_i \int_0^h 1 - e^{-\lambda_i h - \lambda_j (h-s)} \, ds \\ &\leq \lambda_i \int_0^h \lambda_i h + \lambda_j (h - s) \, ds \\ &= \lambda_i (\lambda_i + \lambda_j) h^2/2 \\ &\leq M^2 h^2. \end{aligned}$$

This completes our proof of part a).

For part b) we argue similarly, using

$$P_i(Y_h = i; h < J_1) \leq p_{i,i}(h) \leq P_i(Y_h = i; h < J_1) + P_i(J_2 \leq h).$$

This is because if $Y(0) = i$ and there is a jump before time h then there has to be at least one more jump to bring $Y(h)$ back to i. We will use (11.28) again but also will need

$$P_i(Y_h = i; h < J_1) = 1 - \lambda_i h + \mathrm{o}_i(h) \text{ where } 0 \leq \mathrm{o}_i(h) \leq \frac{1}{2} M^2 h^2. \tag{11.30}$$

This and (11.28) imply that

$$\mathrm{o}_i(h) \leq p_{i,i}(h) - (1 - \lambda_i h) \leq \mathrm{o}_i(h) + \mathrm{o}(h),$$

from which b) follows.

To prove (11.30) observe that $Y_0 = i = Y_h$ with no jumps in $[0,h]$ simply means that $W_1 > h$. This has probability

$$\begin{aligned} P_i(Y_h = i; h < J_1) &= P_i(W_1 > h) \\ &= 1 - \int_0^h \lambda_i e^{-\lambda_i t}\, dt \\ &= 1 - \lambda_i h + \int_0^h \lambda_i \left[1 - e^{-\lambda_i t}\right]\, dt \end{aligned}$$

Using $0 \le 1 - e^{-\lambda_i t} \le \lambda_i t$ we find that

$$0 \le \mathrm{o}_i(h) = \int_0^h \lambda_i \left[1 - e^{-\lambda_i t}\right]\, dt \le \int_0^h \lambda_i^2 t\, dt \le M^2 h^2/2.$$

Finally, for c)

$$\begin{aligned} E_i[\phi(Y_h)] &= E_i[\phi(Y_h); J_2 \le h] + E_i[\phi(Y_h); h < J_2] \\ &= E_i[\phi(Y_h); J_2 \le h] + \sum_j \phi(j) P_i(\phi(Y_h = j; h < J_2) \\ &= E_i[\phi(Y_h); J_2 \le h] + \phi(i)[1 - \lambda_i h + \mathrm{o}_i(h)] + \sum_{j \ne i} \phi(j)[\lambda_i q_{i,j} h - q_{i,j}\mathrm{o}_j(h)] \\ &= E_i[\phi(Y_h); J_2 \le h] + \phi(i) + \sum_{j \ne i} \lambda_i q_{i,j}[\phi(j) - \phi(i)] - \phi(i)\mathrm{o}_i(h) + \sum_{j \ne i} \phi(j) q_{i,j} \mathrm{o}_j(h). \end{aligned}$$

So using the bounds from (11.28), (11.29), (11.30) we have

$$\begin{aligned} \left| E_i[\phi(Y_h)] - \left(\phi(i) + \sum_{j \ne i} \lambda_i q_{i,j}[\phi(j) - \phi(i)] \right) \right| &\le B\mathrm{o}(h) + B\mathrm{o}_i(h) + B \sum_{j \ne i} q_{i,j}\mathrm{o}_j(h) \\ &\le B \left[\frac{1}{2} M^2 h^2 + \frac{1}{2} M^2 h^2 + \sum_{j \ne i} q_{i,j} M^2 h^2 \right] \\ &= 2BM^2 h^2. \end{aligned}$$

□

Now we are ready to establish the Kolmogorov differential equations (both forward and backward) under the assumption of bounded rates. (Problem 11.2 shows that the process is nonexplosive in this case.)

Theorem 11.5. *Suppose Y_t is a Markov jump process with bounded transition rates ($\lambda_i \le M$). The matrix $\mathbf{P}(t) = [p_{i,j}(t)]$ satisfies both systems of differential equations:* (11.24) *and* (11.26).

Proof. To prove (11.24) start with the Chapman-Kolmogorov equation and use parts a) and b) of the lemma.

$$\begin{aligned} p_{i,j}(t+h) &= \sum_k p_{i,k}(t) p_{k,j}(h) \\ &= p_{i,j}(t)[1 - \lambda_j h + \mathrm{o}_{i,j}(h)] + \sum_{k \ne j} p_{i,k}(t)[\lambda_k q_{k,j} h + \mathrm{o}_{k,j}(h)] \\ &= p_{i,j}(t) + h \left[-\lambda_j p_{i,j}(t) + \sum_{k \ne j} \lambda_k q_{k,j} p_{i,k}(t) \right] + \sum_k p_{i,k}(t)\mathrm{o}_{k,j}(h). \\ \frac{p_{i,j}(t+h) - p_{i,j}(t)}{h} &= -\lambda_j p_{i,j}(t) + \sum_{k \ne j} \lambda_k q_{k,j} p_{i,k}(t) + \frac{1}{h} \sum_k p_{i,k}(t)\mathrm{o}_{i,k}(h). \end{aligned}$$

The subscripts on the $o_{\cdot,\cdot}(h)$ terms to remind us that they depend on the indices i and k. But according to the lemma they are all bounded by M^2h^2 so the last term above is bounded by

$$\left|\frac{1}{h}\sum_k p_{i,k}(t)o_k(h)\right| \le M^2 h \sum_k p_{i,k}(t) = M^2 h \to 0 \text{ as } h \to 0.$$

So we find that the forward equation holds:

$$\begin{aligned} p'_{i,j}(t) &= \lim_{h\to 0}\frac{p_{i,j}(t+h)-p_{i,j}(t)}{h} \\ &= -\lambda_j p_{i,j}(t) + \sum_{k\neq j}\lambda_k q_{k,j} p_{i,k}(t) \\ &= (\mathbf{P}(t)\mathcal{A})_{i,j}. \end{aligned}$$

(Technically we have only established the right-hand derivative since $h > 0$. But if we repeat the calculation starting from $p_{i,j}(t) = \sum_k p_{i,k}(t-h)p_{k,j}(h)$ we establish the left-hand derivative as well.)

To prove (11.26) we again start with the Chapman-Kolmogorov equation (but with t and h reversed) and apply part c) of the lemma using $\phi(k) = p_{k,j}(t)$. ($0 \le \phi(k) \le 1$ so the boundedness hypothesis is satisfied.)

$$\begin{aligned} p_{i,j}(t+h) &= \sum_k p_{i,k}(h)p_{k,j}(t) \\ &= E_i[\phi(Y_h)]) \\ &= \phi(i) + \sum_k \lambda_i q_{i,k}[\phi(k)-\phi(i)] + o_\phi(h) \\ &= p_{i,j}(t) + \sum_k \lambda_i q_{i,k}[p_{k,j}(t) - p_{i,j}(t)] + o_\phi(h). \end{aligned}$$

So we conclude that

$$\begin{aligned} p'_{i,j}(t) &= \lim_{h\to 0}\frac{p_{i,j}(t+h)-p_{i,j}(t)}{h} \\ &= \sum_k \lambda_i q_{i,k}[p_{k,j}(t) - p_{i,j}(t)] + \lim_{h\to 0}\frac{1}{h}o_\phi(h) \\ &= \sum_k \lambda_i q_{i,k}[p_{k,j}(t) - p_{i,j}(t)] \\ &= (\mathcal{A}\mathbf{P}(t))_{i,j}. \end{aligned}$$

(The left-hand derivative follows by the same revision as above.) □

11.4.2 The Finite State Case

When the state space $\mathcal{S}$ is finite there are only finitely many λ_i so there *is* a common bound on the λ_i and thus Theorem 11.5 applies. Now $\mathcal{A}$ is just a conventional (finite dimensional, square) matrix. From standard theory of ordinary differential equations we know that the initial value problem for the backward equations

$$\mathbf{P}'(t) = \mathcal{A}\mathbf{P}(t);\ \mathbf{P}(0) = \mathbf{I}$$

has a unique solution given by the matrix exponential:

$$\mathbf{P}(t) = e^{t\mathcal{A}} = \sum_0^\infty \frac{t^n}{n!}\mathcal{A}^n.$$

This is also the unique solution of the forward equations, since $\mathcal{A}e^{t\mathcal{A}} = e^{t\mathcal{A}}\mathcal{A}$.

Example 11.5. As a simple example consider the chain on $\mathcal{S} = \{1, 2\}$ with

$$\mathcal{A} = \begin{bmatrix} -\alpha & \alpha \\ \beta & -\beta \end{bmatrix}.$$

In other words $\lambda_1 = \alpha$, $\lambda_2 = \beta$ and

$$\mathbf{Q} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

The chain just jumps back and forth between the two states, spending $1/\alpha$ on average on each visit to 1 and $1/\beta$ on average on each visit to 2. To calculate $\mathbf{P}(t) = e^{t\mathcal{A}}$ the simplest thing to do is diagonalize: $\mathcal{A} = BDB^{-1}$ where

$$B = \begin{bmatrix} \alpha & 1 \\ -\beta & 1 \end{bmatrix}, \quad D = \begin{bmatrix} -(\alpha+\beta) & 0 \\ 0 & 0 \end{bmatrix}.$$

This leads to

$$\mathbf{P}(t) = Be^{Dt}B^{-1} = \frac{1}{\alpha+\beta}\begin{bmatrix} \alpha e^{-(\alpha+\beta)t} + \beta & \alpha(1 - e^{-(\alpha+\beta)t}) \\ \beta(1 - e^{-(\alpha+\beta)t}) & \alpha + \beta e^{-(\alpha+\beta)t} \end{bmatrix}$$

In particular, by letting $t \to \infty$ we find a unique stationary distribution

$$\pi = \frac{1}{\alpha+\beta}(\beta, \alpha).$$

11.4.3 The K-Process

Our approach for the general (possibly explosive) case is to obtain it as the limit of bounded-rate approximations, to which Theorem 11.5 applies. These bounded-rate approximations are the subject of this section. In the next section we will carry out the limit to give a proof of Theorem 11.3.

Let $K \subseteq \mathcal{S}$ be a *finite* subset of the state space. The idea is to let Y_t proceed as usual as long as it remains inside K. But as soon as it jumps to a state outside of K we freeze it at that state forever. We will call the result the *K-process* and write it as $Y_t^{[K]}$. If we let $\mathcal{T}_{K^c}$ be the first time that $Y_t \notin K$ then we can write

$$Y_t^{[K]} = Y_{t \wedge \mathcal{T}_{K^c}}. \tag{11.31}$$

Observe that if $J_\infty < \infty$ then it can *not* be that $Y(t)$ stayed in K through all those jumps, because the argument of Problem 11.2 would imply that $J_\infty = \sum W_n = \infty$. So $J_\infty \leq \mathcal{T}_{K^c}$ is not possible (has probability 0). Thus $Y_{t \wedge \mathcal{T}_{K^c}}$ does make sense.

The generator $\mathcal{A}^{[K]}$ for $Y^{[K]}$ is obtained simply by setting the jump rates for states outside K to 0. The $q_{i,j}$ are unchanged but

$$\lambda_i^{[K]} = \begin{cases} \lambda_i & \text{if } i \in K \\ 0 & \text{if } i \in K^c. \end{cases}$$

If we write $\mathcal{A}$ in block matrix form

$$\mathcal{A} = \begin{bmatrix} \mathcal{A}_{K,K} & \mathcal{A}_{K,K^c} \\ \mathcal{A}_{K^c,K} & \mathcal{A}_{K^c,K^c} \end{bmatrix} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix} \tag{11.32}$$

then

$$\mathcal{A}^{[K]} = \begin{bmatrix} \mathcal{A}_{K,K} & \mathcal{A}_{K,K^c} \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ 0 & 0 \end{bmatrix}.$$

We will use $p_{i,j}^{[K]}(t)$ to denote the transition probabilities for $Y_t^{[K]}$. Because the states outside K are absorbing we know that for $i \in K^c$

$$p_{i,j}^{[K]}(t) = \begin{cases} 1 & \text{if } j = i \\ 0 & \text{if } j \neq i. \end{cases} \tag{11.33}$$

Since there are only finitely many nonzero jump rates Theorem 11.5 *does* apply to $Y_t^{[K]}$. So let's concentrate on $i \in K$. In the forward equations (11.24) we can ignore all $k \notin K$ because those $\lambda_k^{[K]} = 0$. So the forward equations say that

$$p_{i,j}^{[K]\prime}(t) = -\lambda_j^{[K]} p_{i,j}^{[K]}(t) + \sum_{k\in K,\ k\neq i} \lambda_k^{[K]} q_{k,j} p_{i,k}^{[K]}(t). \tag{11.34}$$

This is a finite dimensional linear system for the $j \in K$. In matrix form and using $\mathbf{A}$ from (11.32),

$$\mathbf{P}_{K,K}^{[K]\prime}(t) = \mathbf{P}_{K,K}^{[K]}(t)\mathbf{A}.$$

The initial values are $\mathbf{P}_{K,K}^{[K]}(0) = \mathbf{I}$ so we know that

$$\mathbf{P}_{K,K}^{[K]}(t) = e^{\mathbf{A}t}.$$

For $p_{i,j}^{[K]}(t)$ when $i \in K$ but $j \in K^c$ see Problem 11.4.

For $i, j \in K$, if $Y_0^{[K]} = i = Y_0$ then $Y_t^{[K]} = j$ is equivalent to $Y_t = j$ and $t < \mathcal{T}_{K^c}$. So the transition probabilities $p_{i,j}^{[K]}(t)$ can be expressed in terms of the original process Y_t as

$$P_i(Y_t = j; t < \mathcal{T}_{K^c}) = p_{i,j}^{[K]}(t). \tag{11.35}$$

If $t < \mathcal{T}_{K^c}$ then $Y_t = Y_t^{[K]} \in K$, so we find

$$P_i(t < \mathcal{T}_{K^c}) = \sum_{j\in K} p_{i,j}^{[K]}(t) = \mathbf{P}_{K,K}^{[K]}(t)[1](i) = e^{\mathbf{A}t}[1](i).$$

We can differentiate this with respect to t:

$$\frac{d}{dt} e^{\mathbf{A}t}[1](i) = \mathbf{A} e^{\mathbf{A}t}[1](i).$$

To restate this, the function $f(i,t) = P_i(t < \mathcal{T}_{K^c})$, $i \in K$ is the solution of system

$$\frac{d}{dt} f(i,t) = \mathbf{A} f(i,t);\ f(i,0) = 1. \tag{11.36}$$

Now consider a second finite set $\tilde{K}$ which is larger than K: $K \subseteq \tilde{K}$. If $Y_t \in \tilde{K}^c$ then $Y_t \in K^c$. This means that $\mathcal{T}_{K^c} \leq \mathcal{T}_{\tilde{K}^c}$. So for $i, j \in K$

$$\begin{aligned} p_{i,j}^{[\tilde{K}]}(t) &= P_i(Y_t = j; t < \mathcal{T}_{\tilde{K}^c}) \\ &= P_i(Y_t = j; t < \mathcal{T}_{K^c}) + P_i(Y_t = j; \mathcal{T}_{K^c} \leq t < \mathcal{T}_{\tilde{K}^c}) \\ &\geq P_i(Y_t = j; t < \mathcal{T}_{K^c}) \\ &= p_{i,j}^{[K]}(t). \end{aligned} \tag{11.37}$$

In words, for $i, j \in K$ the value of $p_{i,j}^{[K]}(t)$ gets larger if we enlarge the set K to a bigger finite set $\tilde{K}$.

11.4.4 The Infinite State Case

The following lemma explains how we get the general case from a limit of the K-processes.

Lemma 11.6. *As $K \uparrow \mathcal{S}$ we have (with probability 1)*

$$\mathcal{T}_{K^c} \uparrow J_\infty,$$

and

$$\mathbf{P}^{[K]}(t) \uparrow \mathbf{P}(t).$$

Proof. We know that $\mathcal{T}_{K^c}$ is monotone in K and that $\mathcal{T}_{K^c} \leq J_\infty$ (with probability 1). Therefore $\lim_K \mathcal{T}_{K^c} \leq J_\infty$ (with probability 1). If K is a finite set with

$$K \supseteq \{Y_{J_i} : i = 0, \ldots, n\}$$

then $J_n < \mathcal{T}_{K^c}$. Now it follows from the properties of probability described in Section 3.1.1 that $P(K \supseteq \{Y_{J_i} : i = 0, \ldots, n\}) \uparrow 1$ as $K \uparrow \mathcal{S}$. Therefore $P(J_n < \lim_K \mathcal{T}_{K^c}) = 1$. Since $J_n \to J_\infty$ it follows that $J_\infty \leq \lim_K \mathcal{T}_{K^c}$ (with probabiity 1). Having proven inequality both ways this establishes that $J_\infty = \lim_K \mathcal{T}_{K^c}$ (with probabiity 1).

The convergence of $\mathbf{P}^{[K]}$ now follows from the monotonicity of probabilities (last bullet of page 33):

$$p_{i,j}^{[K]}(t) = P_i(Y_t = j; t < \mathcal{T}_{K^c}) \uparrow P_i(Y_t = j; t < \mathcal{T}_\infty) = p_{i,j}(t).$$

□

We are now ready to write the proof of Theorem 11.3 in the general case.

Proof. We begin with the backward equations in a). For a given i, j choose K large enough that $i, j \in K$. The backward equations for $Y^{[K]}$ say that

$$p_{i,j}^{[K]\prime}(t) = -\lambda_i p_{i,j}^{[K]}(t) + \sum_{k \neq i} \lambda_i q_{i,k} p_{k,j}^{[K]}(t).$$

(Since $i \in K$ we have $\lambda_i^{[K]} = \lambda_i$.) Expressing this in integrated form,

$$p_{i,j}^{[K]}(t) = p_{i,j}^{[K]}(0) - \int_0^t \lambda_i p_{i,j}^{[K]}(s)\, ds + \int_0^t \sum_{k \neq i} \lambda_i q_{i,k} p_{k,j}^{[K]}(s)\, ds.$$

Each of the two integrals converges as $K \uparrow \mathcal{S}$ by the Monotone Convergence Theorem. (We needed to separate out the $-\lambda_j$ integral to make this argument.) We find that

$$p_{i,j}(t) = p_{i,j}(0) - \int_0^t \lambda_i p_{i,j}(s)\, ds + \int_0^t \sum_{k \neq i} \lambda_i q_{i,k} p_{k,j}(s)\, ds. \tag{11.38}$$

It follows from this that $p_{i,j}(t)$ is continuous. Moreover since $\sum_k q_{i,k} = 1$ the series $\sum_{k \neq i} \lambda_i q_{i,k} p_{k,j}(s)$ converges uniformly and is therefore continuous. That allows us to deduce that $p_{i,j}(t)$ is in fact continuously differentiable with

$$p'_{i,j}(t) = -\lambda_i p_{i,j}(t) + \sum_{k \neq i} \lambda_i q_{i,k} p_{k,j}(t),$$

which is the backward equation for $\mathbf{P}(t)$. Before proceeding observe that the backward equation implies that

$$\frac{d}{dt}\left[e^{\lambda_i t} p_{i,j}(t)\right] = e^{\lambda_i t} \sum_{k \neq i} \lambda_i q_{i,k} p_{k,j}(t) \geq 0.$$

This means $e^{\lambda_i t} p_{i,j}(t)$ is monotone increasing, a fact we will need to prove b).

The argument for the forward equations b) starts in the same way. The forward equation (11.34) in integrated form says

$$p_{i,j}^{[K]}(t) = p_{i,j}^{[K]}(0) - \int_0^t \lambda_j p_{i,j}^{[K]}(s)\, ds + \int_0^t \sum_{k \in K,\ k \neq i} \lambda_k q_{k,j} p_{i,k}^{[K]}(s)\, ds.$$

(This is for K large enough to include j and we have omitted all $\lambda_k^{[K]} = 0$ for $k \notin K$ from the sum.) Both integrals converge as $K \uparrow \mathcal{S}$ by the monotone convergence theorem to give

$$p_{i,j}(t) = p_{i,j}(0) - \int_0^t \lambda_j p_{i,j}(s)\, ds + \int_0^t \sum_{k \neq i} \lambda_k q_{k,j} p_{i,k}(s)\, ds.$$

The final integral is trickier here because the convergence of the series is not obvious. However by rearranging we deduce a bound on the the final integral:

$$\int_0^t \sum_{k\neq i} \lambda_k q_{k,j} p_{i,k}(s)\, ds \leq 1 + \lambda_j t < \infty.$$

So the integral is finite, which is enough to imply that $p_{i,j}(t)$ is continuous. By the observation at the end of the argument for the backwards equation we know

$$e^{\lambda_i t} \sum_{k\neq i} \lambda_k q_{k,j} p_{i,k}(t)$$

is monotone increasing. So for $s \leq t$

$$\sum_{k\neq i} \lambda_k q_{k,j} p_{i,k}(s) \leq e^{\lambda_i (t-s)} \sum_{k\neq i} \lambda_k q_{k,j} p_{i,k}(t).$$

Since the series must converge (for its integral to be finite) this inequality implies that it converges uniformly in $s \leq t$. Therefore $\sum_{k\neq i} \lambda_k q_{k,j} p_{i,k}(s)$ is continuous, allowing us to conclude that $p_{i,j}(t)$ is continuously differentiable with

$$p'_{i,j}(t) = -\lambda_j p_{i,j}(s)\, ds + \sum_{k\neq i} \lambda_k q_{k,j} p_{i,k}(s)\, ds,$$

which is the forward equation.

Next consider part c) for the forward equation. Suppose $\mathbf{Q}'(t) = \mathbf{Q}(t)\mathcal{A}$ with $\mathbf{Q}(0) = \mathbf{I}$. Consider any finite K. Since $\mathbf{P}^{[K]}(t) \uparrow \mathbf{P}(t)$ it suffices to show $\mathbf{P}^{[K]}(t) \leq \mathbf{Q}(t)$. Using the block matrix components $\mathbf{A}$ and $\mathbf{C}$ from (11.32) above the forward equations for $\mathbf{Q}$ say that

$$\mathbf{Q}'_{K,K}(t) = \mathbf{Q}_{K,K}(t)\mathbf{A} + \mathbf{Q}_{K,K^c}(t)\mathbf{C}.$$

Now the variation of constants formula, or an integrating factor argument, leads to

$$\begin{aligned}\mathbf{Q}_{K,K}(t) &= e^{\mathbf{A}t} + \int_0^t \mathbf{Q}_{K,K^c}(s)\mathbf{C}e^{\mathbf{A}(t-s)}\, ds \\ &= \mathbf{P}^{[K]}_{K,K}(t) + \int_0^t \mathbf{Q}_{K,K^c}(s)\mathbf{C}e^{\mathbf{A}(t-s)}\, ds\end{aligned}$$

Since we are also assuming that $\mathbf{Q}(t) \geq \mathbf{0}$ the integral term is nonnegative. Therefore $\mathbf{Q}_{K,K}(t) \geq \mathbf{P}^{[K]}_{K,K}(t)$. This being true for any finite K we conclude that $\mathbf{Q}(t) \geq \mathbf{P}(t)$ as claimed. The argument for the backward equation works the same way, but with the matrix exponentials multiplied on the right.

To prove d) we back up to (11.38), multiply both sides by $\phi(j)$ and sum over j. (Dominated convergence justifies passing the sums through the integrals.) We obtain

$$u(i,t) = u(i,0) - \int_0^t \lambda_i u(i,s)\, ds + \int_0^t \sum_{k\neq i} \lambda_i q_{i,k} u(k,s)\, ds.$$

From here the argument is the same as above, to deduce that $g(i,t)$ is continuously differentiable in t with

$$\frac{\partial}{\partial t} u(i,t) = \mathcal{A}u(i,t).$$

Moreover since $0 \leq g(i,t) \leq B$ and $\sum_k q_{i,k} = 1$ it follows that

$$|\mathcal{A}u(i,t)| \leq \lambda_i B.$$

For part e) let $u(i,t) = \mathbf{P}(t)[1](i)$, which we know is $u(i,t) = P_i(\mathcal{T}_\infty < t)$ and a solution of the backward equation (by part d)), and therefore so is $\psi(i,t) = 1 - u(i,t)$. Now $\psi(i,0) = 0$ for all i, but in the explosive

case $\psi(i,t) > 0$ for some state i and $t > 0$, so the backward equation does have non-unique bounded solutions in the explosive case. Consider the non-explosive case and suppose $\mathbf{u}(t) = [u(i,t)]$ is a *bounded* solution of (11.27) with $\mathbf{u}(0) = [0]$ we need to show that $\mathbf{u}(t) = [0]$ for all $t > 0$. Consider any finite $K \subseteq \mathcal{S}$, with the block matrix components $\mathbf{A}$ and $\mathbf{B}$ as above. The backward equations say that on K

$$\mathbf{u}_K'(t) = \mathbf{A}\mathbf{u}_K(t) + \mathbf{B}\mathbf{u}_{K^c}(t).$$

Using $e^{-\mathbf{A}t}$ as an integrating factor we get

$$\begin{aligned}\mathbf{u}_K(t) &= e^{\mathbf{A}t}\mathbf{u}_K(0) + \int_0^t e^{\mathbf{A}(t-s)}\mathbf{B}\mathbf{u}_{K^c}(s)\,ds \\ &= \int_0^t e^{\mathbf{A}(t-s)}\mathbf{B}\mathbf{u}_{K^c}(s)\,ds.\end{aligned}$$

We claim this $\to 0$ as $K \uparrow \mathcal{S}$. Suppose $|u(j,t)| \le c$ for all $j \in \mathcal{S}$. Fix i. Once K is large enough to include i we have $0 \le b_{i,j}$ for all $j \in K^c$, so

$$|\mathbf{B}\mathbf{u}_{K^c}(s)| \le c\mathbf{B}\mathbf{1}_{K^c},$$

and since $e^{\mathbf{A}(t-s)} \ge 0$ we find by using Problem 11.4 that

$$\begin{aligned}\left|\int_0^t e^{\mathbf{A}(t-s)}\mathbf{B}\mathbf{u}_{K^c}(s)\,ds\right| &\le c\int_0^t e^{\mathbf{A}(t-s)}\mathbf{B}\mathbf{1}_{K^c}\,ds \\ &= c\mathbf{P}^{[K]}_{K,K^c}(t)\mathbf{1}_{K^c} \\ &= cP_i(\mathcal{T}_{K^c} < t).\end{aligned}$$

If the process is non-explosive then $P_i(\mathcal{T}_{K^c} < t) \to P_i(\mathcal{T}_\infty < t) = 0$. Therefore $u(i,t) = 0$ for all i. □

11.5 Martingales and the Generator

Next we want to make the connection between the generator and martingales for a jump process. We will not develop this as fully as we might, but will just discuss what we will need for use in the next section on conditions for explosion or non-explosion.

The generator $\mathcal{A}$ is the continuous time analogue of the matrix $\mathbf{A}$ for discrete time Markov chains. By analogy with Theorem 9.1 we might expect

$$M_t = f(Y_t) - \int_0^t \mathcal{A}f(Y_u)\,du$$

to be a martingale for a bounded function $f : S \to \mathbb{R}$. This is essentially true but we need to be careful in the continuous time setting because of the possibility of explosion as well as that $\mathcal{A}f$ might be unbounded, so that we are unsure about $E[\mathcal{A}f(Y_t)]$ being defined. Setting those issues aside for the moment, the basic idea is as follows. Observe that for $t < s$

$$M_s = M_t + f(Y_t) - f(Y_s) + \int_t^s \mathcal{A}f(Y_u)\,du.$$

So

$$E[M_s|Y_{[0,t]}] = M_t + \left\{f(Y_t) - E\left[f(Y_s) + \int_t^s \mathcal{A}f(Y_u)\,du\Big|Y_{[0,t]}\right]\right\}.$$

For M_t to be a martingale means that $\{\cdots\} = 0$. The Markov property says that when $Y_t = y$ we should replace $E[\cdots Y_u \cdots |Y_{[0,t]}]$ by $E_y[\cdots Y_{(u-t)} \cdots]$. Rearranging we find that the martingale property reduces to

$$E_y[f(Y_T)] = f(y) + E_y\left[\int_0^T \mathcal{A}f(Y_u)\,du\right].$$

But this is essentially the forward equation for $\mathbf{P}(t)$: $\frac{d}{dt}(\mathbf{P}(t)f(y)) = \mathbf{P}(t)\mathcal{A}f(y)$, or in integrated form,

$$\mathbf{P}(T)f(y) = f(y) + \int_0^T \mathbf{P}(u)\mathcal{A}f(y)\,du. \tag{11.39}$$

There is an additional technical issue here that our notation hides. We have understood the forward equation in a entry-by-entry sense: $\frac{d}{dt}p_{y,j}(t) = (\mathbf{P}(t)\mathcal{A})_{y,j}$. But $\mathbf{P}(t)f(y) = \sum_j p_{y,j}(t)f(j)$ is in general an infinite series. So we have to worry about the validity of differentiating the infinite series term-by-term. And then there is the need to justify $E_y[\int_0^T \cdots du] = \int_0^T E_y[\cdots]\,du$, which we used to arrive at (11.39). So we need to impose hypotheses and formulate a proof which allows us to navigate through all these issues. And if that is not enough, we are going to generalize to allow the function f to depend on *both* the state and time variables: $f(y,t)$. This will be useful for our discussion of conditions for non-explosion in Section 11.6.

Theorem 11.7. *Suppose Y_t is non-explosive and that $f(y,t)$ is a bounded function on $\mathcal{S}\times[0,\infty)$ for which both $\mathcal{A}f(y,t)$ and $\frac{\partial}{\partial t}f(y,t)$ are also bounded. Then*

$$M_t = f(Y_t,t) - \int_0^t \frac{\partial}{\partial u}f(Y_u,u) + \mathcal{A}f(Y_u,u)\,du$$

is a martingale.

Compare this to Theorem 9.1.

Proof. Rearranging as above (and making an elementary change of time variable) we find that the martingale property reduces to

$$E_y[f(Y_T,T)] = f(y,0) + \int_0^T E_y\left[\frac{\partial}{\partial t}f(Y_u,u) + \mathcal{A}f(Y_u,u)\right]\,du. \tag{11.40}$$

Suppose that there is a finite set K such that $f(i,t) = 0$ if $i \notin K$. Then

$$\mathbf{P}(t)f(y,t) = \sum_{j\in K} p_{y,j}(t)f(j,t)$$

is a *finite* sum so we *can* differentiate term-by-term and use the forward equation and the product rule to conclude that

$$\mathbf{P}(T)f(y,T) = f(y,0) + \int_0^T \mathbf{P}(u)\left[\frac{\partial}{\partial t}f(y,u) + \mathcal{A}f(y,u)\right]\,du.$$

That both $\frac{\partial}{\partial t}f$ and $\mathcal{A}f$ are bounded is enough to justify the interchange $E_y[\int_0^T \cdots du] = \int_0^T E_y[\cdots]\,du$, so (11.40) does hold under the K assumption.

For a general f satisfying the hypotheses of the theorem we can consider $f_K = 1_K f$. According to what we just showed equation (11.40) does apply to f_K. We want to pass to the limit as $K \uparrow \mathcal{S}$ to get the general case. But to justify that requires more information about $\mathcal{A}f_K$ than is conveniently available to us. But suppose in addition that Y_t has bounded rates: $\lambda_k \le M$. In that case $\mathcal{A}f$ and $\mathcal{A}f_K$ are both bounded,

$$|\mathcal{A}f| \le 2MB \text{ and } |\mathcal{A}f_K| \le 2MB \text{ where } |f| \le B,$$

as well as $\frac{\partial}{\partial t}f(y,u)$. Therefore the dominated convergence theorem applies as $K \uparrow \mathcal{S}$ to tell us that

$$\int_0^T \mathbf{P}(u)\left[\frac{\partial}{\partial t}f_K(y,u) + \mathcal{A}f_K(y)\right]\,du \to \int_0^T \mathbf{P}(u)\left[\frac{\partial}{\partial t}f(y,u) + \mathcal{A}f(y)\right]\,du,$$

as well as the other terms in (11.40). This proves the theorem in the case of bounded rates.

To establish the general case, our stopped process $Y_t^{[K]}$ has bounded rates so the theorem holds for it. Observe that

$$\mathcal{A}^{[K]}f(i) = \begin{cases} \mathcal{A}f(i) & \text{if } i \in K \\ 0 & \text{if } i \notin K \end{cases} = 1_K(i)\mathcal{A}f(i).$$

So the martingale property for $Y_t^{[K]}$ says that

$$E_y[f(Y_T^{[K]},T)] = f(y,0) + E_y\left[\int_0^T \frac{\partial}{\partial t} f(Y_u^{[K]},u) + 1_K(Y_u^{[K]})\mathcal{A}f(Y_u^{[K]},u)\,du\right].$$

With the assumption that $\mathcal{A}f$ and $\frac{\partial}{\partial t}f$ are bounded we can let $K \uparrow \mathcal{S}$, which entails $\mathcal{T}_{K^c} \uparrow \infty$, $1_K(Y_u^{[K]}) \to 1$ for a non-explosive process, and by dominated convergence again we conclude

$$E_y[f(Y_T,T)] = f(y,0) + E_y\left[\int_0^T \frac{\partial}{\partial t} f(Y_u,u) + \mathcal{A}f(Y_u,u)\,du\right],$$

as desired. □

Before moving on we want to apply the theorem to the particular equation we will use in the next section. Suppose φ is a bounded function satisfying

$$\mathcal{A}\varphi = \alpha\varphi \text{ for some constant } \alpha > 0.$$

Let $f(i,t) = \varphi(i)e^{-\alpha t}$. This f satisfies all the hypotheses of the theorem: $\mathcal{A}f = -\frac{\partial}{\partial t}f = \alpha\varphi e^{-\alpha t}$ is bounded since φ is. Observe that $\mathcal{A}f + \frac{\partial}{\partial t}f = 0$ so if $Y(t)$ is not explosive it follows that

$$M_t = e^{-\alpha t}\varphi(Y(t))$$

is a martingale. In particular for any state y

$$\varphi(y) = E_y[e^{-\alpha t}\varphi(Y(t))]. \tag{11.41}$$

We want to generalize this in a couple different ways. First, even if $Y(t)$ is explosive we can apply the same reasoning to the K-process $Y^{[K]}(t)$ (for any finite subset $K \subseteq \mathcal{S}$). Note that applying the generator for $Y^{[K]}(t)$ we get

$$\mathcal{A}^{[K]}f(i,t) = 1_K(i)\mathcal{A}f(i,t)$$

and so is still bounded. But

$$\mathcal{A}^{[K]}f + \frac{\partial}{\partial t}f = \begin{cases} 0 & \text{if } i \in K \\ \frac{\partial}{\partial t}f(i,t) & \text{if } i \notin K. \end{cases}$$

So applying the theorem to the K-process we obtain

$$\begin{aligned}
\varphi(y) &= E_y\left[f(Y^{[K]}(t),t) - \int_{t\wedge\mathcal{T}_{K^c}}^t \frac{\partial}{\partial t} f(Y^{[K]}(s),s)\,ds\right] \\
&= E_y\left[f(Y(t\wedge\mathcal{T}_{K^c}),t) - \int_{t\wedge\mathcal{T}_{K^c}}^t \frac{\partial}{\partial t} f(Y(\mathcal{T}_{K^c}),s)\,ds\right] \\
&= E_y\left[f(Y(t\wedge\mathcal{T}_{K^c}),t) - f(Y(t\wedge\mathcal{T}_{K^c}),t) + f(Y(t\wedge\mathcal{T}_{K^c}),t\wedge\mathcal{T}_{K^c})\right] \\
&= E_y\left[f(Y(t\wedge\mathcal{T}_{K^c}),t\wedge\mathcal{T}_{K^c})\right] \\
&= E_y\left[e^{-\alpha(t\wedge\mathcal{T}_{K^c})}\varphi(Y(t\wedge\mathcal{T}_{K^c}))\right].
\end{aligned}$$

This is optional stopping applied to the martingale M_t. Compare to Theorem 9.6.

Now suppose we replace φ with a function ψ, still assumed bounded but only satisfying the inequality $\mathcal{A}\psi \geq \alpha\psi$ and take $f(i,t) = \psi(i)e^{-\alpha t}$ like before. We still have that $\frac{\partial}{\partial t}f$ is bounded but without additional assumptions we can't say that $\mathcal{A}f$ is bounded. However it does follow that $\mathcal{A}^{[K]}f$ is bounded so we can apply the theorem using $Y^{[K]}$. We proceed as above, but using

$$\mathcal{A}^{[K]}f + \frac{\partial}{\partial t}f \geq \begin{cases} 0 & \text{if } i \in K \\ \frac{\partial}{\partial t}f(i,t) & \text{if } i \notin K. \end{cases}$$

The conclusion is that

$$\psi(y) \le E_y\left[e^{-\alpha(t\wedge\mathcal{T}_{K^c})}\psi(Y(t\wedge\mathcal{T}_{K^c}))\right].$$

If in fact Y is non-explosive from $Y(0)=y$ then we can let $K\uparrow\mathcal{S}$, $\mathcal{T}_{K^c}\to\infty$ and use dominated convergence (since ψ is bounded) to conclude that

$$\psi(y) \le E_y\left[e^{-\alpha t}\psi(Y(t))\right]. \tag{11.42}$$

If we reverse the inequality and assume $\mathcal{A}\psi\le\alpha\psi$ (but still that ψ is bounded) we simply reverse the inequality in the preceeding result:

$$\psi(y) \ge E_y\left[e^{-\alpha(t\wedge\mathcal{T}_{K^c})}\psi(Y(t\wedge\mathcal{T}_{K^c}))\right].$$

We want to weaken the hypotheses again to assume $\mathcal{A}\psi\le\alpha\psi$ and $0\le\psi$ but not that ψ is necessarily bounded above. First choose the finite set K. Next replace ψ by

$$\psi_c(i) = c\wedge\psi(i) = \begin{cases}\psi(i) & \text{if } \psi(i)\le c\\ c & \text{if } c<\psi(i)\end{cases}$$

and take $f_c(i,t)=e^{-\alpha t}\psi_c(i)$. We want to apply the theorem using f_c and the K-process $Y^{[K]}$. Since ψ_c is bounded and $Y^{[K]}$ has bounded rates all the hypotheses of the theorem hold. If c is large enough ($c\ge\max_{i\in K}\psi(i)$) then $\psi_c=\psi$ on K. So for $i\in K$ we have

$$\begin{aligned}\mathcal{A}^{[K]}\psi_c(i) &= -\lambda_i\psi_c(i)+\sum_{j\ne i}\lambda_i q_{i,j}\psi_c(j)\\ &= -\lambda_i\psi(i)+\sum_{j\ne i}\lambda_i q_{i,j}\psi_c(j)\\ &\le -\lambda_i\psi(i)+\sum_{j\ne i}\lambda_i q_{i,j}\psi(j)\\ &= \mathcal{A}\psi(i)\\ &\le \alpha\psi(i)\\ &= \alpha\psi_c(i).\end{aligned}$$

So we again have an inequality

$$\mathcal{A}^{[K]}f_c+\frac{\partial}{\partial t}f_c \le \begin{cases}0 & \text{if } i\in K\\ \frac{\partial}{\partial t}f_c(i,t) & \text{if } i\notin K.\end{cases}$$

This leads to

$$\psi_c(y) \ge E_y\left[e^{-\alpha(t\wedge\mathcal{T}_{K^c})}\psi_c(Y(t\wedge\mathcal{T}_{K^c}))\right].$$

as above. Finally let $c\uparrow\infty$ and use monotone convergence to conclude

$$\psi(y) \ge E_y\left[e^{-\alpha(t\wedge\mathcal{T}_{K^c})}\psi(Y(t\wedge\mathcal{T}_{K^c}))\right]. \tag{11.43}$$

as desired.

11.6 Explosion

One of our themes has been to show how various probabilistic properties of a Markov process can be established in terms of equations involving the generator. In this section we are going to do that for the phenomena of explosion and non-explosion. Lemma 11.2 provided a simple test for the pure birth process. The proof of that lemma focused on $E[e^{-J_\infty}]$. Here we consider

$$\varphi(y) = E_y[e^{-\alpha J_\infty}]$$

for a constant $\alpha>0$. If the process is non-explosive then $\varphi(y)=0$ but in the explosive case $\varphi(y)>0$. The next lemma says that φ must solve the equation $\mathcal{A}\varphi=\alpha\varphi$.

11.6.1 The Distribution of J_∞

Lemma 11.8. *Let $\alpha > 0$. The function $\varphi(y) = E_y[e^{-\alpha J_\infty}]$ is a nonnegative solution of*

$$\mathcal{A}\varphi = \alpha\varphi$$

bounded by $\varphi(y) \le 1$.

Proof. To begin,

$$P_i(t < J_\infty) = \sum_j p_{i,j}(t) = u(i,t) = \mathbf{P}(t)[1](i),$$

as considered in part d) of Theorem 11.3. We know from there that

$$\mathbf{u}'(t) = \mathcal{A}\mathbf{u}(t), \quad \mathbf{u}(0) = [1].$$

So we can calculate as follows.

$$\begin{aligned}
e^{-\alpha J_\infty} &= 1 - \int_0^{J_\infty} \alpha e^{-\alpha t}\, dt \\
&= 1 - \int_0^{\infty} 1_{t<J_\infty} \alpha e^{-\alpha t}\, dt \\
\varphi(i) &= E_i\left[1 - \int_0^{\infty} 1_{t<J_\infty} \alpha e^{-\alpha t}\, dt\right] \\
&= 1 - \int_0^{\infty} E_i[1_{t<J_\infty}] \alpha e^{-\alpha t}\, dt \\
&= 1 - \int_0^{\infty} u(i,t) \alpha e^{-\alpha t}\, dt.
\end{aligned}$$

We want to calculate $\mathcal{A}\varphi(i)$ using the right side of this expression. Since $\mathcal{A}[1] = [0]$ the first term vanishes. We have

$$\mathcal{A}\varphi(i) = \lambda_i \int_0^{\infty} u(i,t)\alpha e^{-\alpha t}\, dt - \sum_{j\neq i} \lambda_i \int_0^{\infty} q_{i,j} u(j,t) \alpha e^{-\alpha t}\, dt$$

In the last term ($\sum\int$) we want to interchange the integral and summation. If the sum involved only a finite number of terms this would be no problem. In general, since $0 \le u \le 1$ and $\sum_{j\neq i} q_{i,j} = 1$, the series $\sum_{j\neq i} q_{i,j} u(i,t)$ converges *uniformly*. That justifies the interchange in general. So we have

$$\begin{aligned}
\mathcal{A}\varphi(i) &= 0 + \int_0^{\infty} \lambda_i u(i,t)\alpha e^{-\alpha t}\, dt - \int_0^{\infty} \sum_{j\neq i} \lambda_i q_{i,j} u(j,t) \alpha e^{-\alpha t}\, dt \\
&= -\int_0^{\infty} \mathcal{A}u(i,t)\alpha e^{-\alpha t}\, dt \\
&= -\int_0^{\infty} \frac{\partial}{\partial t} u(i,t)\alpha e^{-\alpha t}\, dt \\
&= -u(i,t)\alpha e^{-\alpha t}\big]_0^{\infty} - \alpha \int_0^{\infty} u(i,t)\alpha e^{-\alpha t}\, dt \\
&= \alpha\left[1 - \int_0^{\infty} u(i,t)\alpha e^{-\alpha t}\, dt\right] \\
&= \alpha\varphi(i).
\end{aligned}$$

□

11.6.2 Conditions for Explosion and Non-Explosion

Here is a nice necessary and sufficient condition for non-explosion in terms of solutions to $\mathcal{A}\varphi = \alpha\varphi$.

Theorem 11.9. *Let $\alpha > 0$. Y_t is non-explosive if and only if the only bounded solution φ of $\mathcal{A}\varphi = \alpha\varphi$ is $\varphi \equiv 0$.*

Proof. Suppose Y_t is nonexplosive from y and φ is *any* bounded solution of $\mathcal{A}\varphi = \alpha\varphi$. From (11.41) we know that

$$\varphi(y) = E_y[e^{-\alpha t}\varphi(Y_t)].$$

Letting $t \to \infty$ and using the boundedness of φ we deduce that $\varphi(y) = 0$.

For the converse assume that the only bounded solution of $\mathcal{A}\varphi = \alpha\varphi$ is $\varphi \equiv 0$. Since we know $\varphi(y) = E_y[e^{-\alpha J_\infty}]$ is bounded and solves the equation, it follows that $E_y[e^{-\alpha J_\infty}] = 0$ for every initial state y. Therefore for every initial state $J_\infty = \infty$ implying nonexplosion. □

We have stated this for "non-explosive" in general, meaning non-explosive *from every initial state*. The theorem remains true if we focus on a specific initial state y: non-explosion *from* y is equivalent to saying every bounded solution of $\mathcal{A}\varphi = \alpha\varphi$ has $\varphi(y) = 0$.

Similar to Section 4.3.2, inequality versions of $\mathcal{A}\varphi = \alpha\varphi$ can be easier to work with, and sometimes provide sufficient conditions for explosion or non-explosion.

Theorem 11.10. *Suppose $\psi : S \to \mathbb{R}$ is nonegative bounded but not identically 0, and satisfies $\mathcal{A}\psi \geq \alpha\psi$ for a constant $\alpha > 0$. Then Y_t is explosive from any y with $\psi(y) > 0$.*

Proof. If the process were nonexplosive from y we could use (11.42):

$$\psi(y) \leq E_y[e^{-\alpha t}\psi(Y(t))].$$

Since ψ is bounded, letting $t \to \infty$ would imply the $\psi(y) = 0$. So if $\psi(y) > 0$ we see that $Y(t)$ must be explosive from y. □

Theorem 11.11. *Suppose $\psi : S \to \mathbb{R}$ satisfies $\mathcal{A}\psi \leq \alpha\psi$ for a constant $\alpha > 0$ and $\psi(x) \to +\infty$ as $|x| \to \infty$. Then Y_t is nonexplosive (from all initial states).*

Proof. This time we want to use (11.43). For that we need $\psi \geq 0$, which is not assumed. But we can fix that. Because $\psi(x) \to \infty$ implies that ψ must be that ψ s bounded below. So there is a constant $c > 0$ for which

$$\bar{\psi} = c + \psi \geq 0.$$

Now observe that

$$\mathcal{A}\bar{\psi} = \mathcal{A}(\psi + c) = \mathcal{A}\psi \leq \alpha\psi \leq \alpha(\psi + c) = \alpha\bar{\psi}.$$

We can now apply (11.43) to conclude that for any finite K

$$\bar{\psi}(y) \geq E_y\left[e^{-\alpha(t\wedge\mathcal{T}_{K^c})}\bar{\psi}(Y(t \wedge \mathcal{T}_{K^c}))\right].$$

Now if $Y(t)$ were explosive then for some t we have $P_y(\mathcal{T}_{K^c} < t) \geq P_y(J_\infty < t) > 0$. By hypothesis $\bar{\psi}(Y(\mathcal{T}_{K^c})) \to \infty$ as $K \uparrow \mathcal{S}$. Taking this limit in the above would give

$$E_y\left[e^{-\alpha(t\wedge\mathcal{T}_{K^c})}\bar{\psi}(Y(t \wedge \mathcal{T}_{K^c}))\right] \to \infty.$$

This is not possible since $\bar{\psi}(y) < \infty$. So the process must be non-explosive. □

Finally let's apply these conditions to our examples. The queueing example of Section 11.3.2 has bounded rates so is non-explosive by Problem 11.2. For the chemical kinetics example of Section 11.3.1 observe that $Y^A + Y^B + 2Y^C + Y^D$ never changes; this quantity is unaltered by each of the three reactions. That means that

$$\psi(y) = y^A + y^B + 2y^C + y^D$$

has $\mathcal{A}\psi = 0$. So

$$\mathcal{A}\psi = 0 \leq \psi.$$

Since $\psi \geq 0$ on the state space and $\psi(y) \to \infty$ as $|y| \to \infty$ Theorem 11.11 tells us that the process is non-explosive.

Consider a pure birth process (Section 11.2.2) with rates $\lambda_n > 0$ for $n \geq 0$. The equation $\mathcal{A}\varphi = \alpha\varphi$ is the system

$$\lambda_n(\varphi(n+1) - \varphi(n)) = \alpha\varphi(n).$$

which we can rearrange as

$$\varphi(n+1) = \frac{\alpha + \lambda_n}{\lambda_n}\varphi(n).$$

To analyze the nontrivial bounded solutions (as needed to use Theorem 11.9) means studying the infinite product

$$\lim_{k\to\infty} \prod_1^k \frac{\alpha + \lambda_n}{\lambda_n}.$$

Notice that for $\alpha = 1$ this is the same as what we encountered in (11.15). For particular cases we can exhibit solutions to either the equation of Theorem 11.9 or the inequalities of Theorems 11.10. Consider the simple birth process of Example 11.1: $\lambda_n = n\lambda$. Take $\psi(n) = n$ and use $\alpha = \lambda$. This in fact satisfies $\mathcal{A}\psi = \alpha\psi$ exactly, and has $\psi \geq 0$ and $\psi \to \infty$ so we deduce non-explosion by Theorem 11.11.

Next consider the explosive pure birth process of Example 11.2: $\lambda_n = n(n-1)/2$. Explosion will follow from Theorem 11.10 if we can exhibit a nontrivial bounded solution of $\mathcal{A}\psi \geq \frac{1}{2}\psi$:

$$\psi(n+1) \geq \frac{1/2 + n(n-1)/2}{n(n-1)/2}\psi(n).$$

Try $\psi(n) = e^{-\frac{1}{n-1}}$ for $n \geq 2$ and $\psi(0) = \psi(1) = 0$. You can check the inequality for $n = 0, 1$. For $n \geq 2$ we have

$$\begin{aligned}
\psi(n+1) &= e^{\frac{1}{n-1} - \frac{1}{n}}\psi(n) \\
&= e^{\frac{1}{n(n-1)}}\psi(n) \\
&\geq \left(1 + \frac{1}{n(n-1)}\right)\psi(n) \\
&= \frac{1/2 + n(n-1)/2}{n(n-1)/2}\psi(n),
\end{aligned}$$

as desired. Theorem 11.10 now implies the process is explosive from any initial state ≥ 2, but not from 0 or 1.

11.7 Extensions and Further Reading

We have not covered all the standard material on continuous time Markov chains. In particular we have neglected recurrence and equilibrium. You can find these and other aspects discussed in references such as Norris [45], Grimmett & Stirzaker [25]. See also Brémaud [10], Chung [15], Karlin [33], Kemeny & Snell [34], and Stroock [57].

We have only developed the relation of jump processes to martingales to the extent needed for Section 11.6. In particular we have not stated a full martingale characterization analogous to Theorem 9.1. The problem

of showing the converse of that theorem is usually referred to as "the martingale problem". It is developed for jump processes in Ethier & Kurtz [21] and in Stroock [58].

Conditions for explosion or non-explosion are developed by many authors. See Norris [45], Brémaud [10], Chow and Khasminskii [13], Has'minskiĭ [26], Stroock [57], and Varadhan [64].

Continuing Past Explosion

A full treatment of solutions to the Kolmogorov equations is difficult. It involves ways of extending the definition of an explosive process beyond its explosion time and boundary conditions "at infinity" for the solutions. This is an interesting but complicated subject. See Feller [22] XVII.10 for some commentary and references on this. Most approaches use the theory of semigroups. Rogers & Williams [51] presents some of that theory in vol. 1 but it requires some graduate level background.

There are other types of Markov processes which move only by jumps but which are more complicated than what we have described above. We close this chapter with just the briefest indication of what a couple of them are.

The Cauchy Process

If we allow a continuous state space $\mathcal{S} = \mathbb{R}$ then it is possible to have Markov processes which make infinitely many jumps in a small amount of time, provided most of them are so small that the sum of their spatial increments is finite. The best known example is the *Cauchy Process*. Its generator, applied to a smooth function f, is

$$\mathcal{A}f(x) = \int_{-\infty}^{\infty} \left[f(y) - f(x) - f'(x) \frac{y-x}{1+(y-x)^2} \right] \frac{1}{\pi(y-x)^2}\, dy.$$

Here the $\frac{1}{\pi(y-x)^2}\, dy$ plays a role like that of our $\lambda_x q_{x,y}$. In fact if $\frac{1}{\pi(y-x)^2}$ were integrable w.r.t. y (it is not, but if it were) then the f' term would integrate to 0 by symmetry and the above would reduce to $\int_{-\infty}^{\infty} [f(y) - f(x)] \frac{1}{\pi(y-x)^2}\, dy$, which looks a lot like our (11.18). The additional term in $\mathcal{A}f(x)$ is necessary for the integral to even exist. The effect is that small jumps ($y \approx x$) occur at a faster rate, becoming infinitely fast in the limit as $y \to x$. Every time interval contains infinitely many jumps, most quite small in size. Our wait and jump description is not adequate to describe it. The Cauchy process belongs a general class of Markov processes called *stable processes*. Beriman [9] is a good introduction to these, but be advised that this is a topic requiring a graduate-level background in analysis.

Interacting Particle Systems

Imagine a collection of individuals, one located at each integer point on the line. Each individual can be in one of two states, healthy or infected. The state of this system is an infinite sequence of 0s and 1s:

$$X(t) = (\ldots X_{-2}(t), X_{-1}(t), X_0(t), X_1(t), X_2(t), \ldots).$$

$X_n(t) = 1$ means that the individual at location n is *infected* at time t; $X_n(t) = 0$ means that the individual at location n is *healthy* at time t. Each individual's status jumps back and forth between 0 and 1 like a Markov jump process. If $X_n(t) = 1$ (infected) then it jumps to the healthy state at rate of 1. If $X_n(t) =$ (healthy) then it jumps to the infected state at rate of $\lambda[X_{n-1}(t) + X_{n+1}(t)]$, i.e. proportional to the number of infected neighbors it has. So each individual's status jumps back and forth at rates which depend on the status of it's neighbors. This is called the *Contact Process* on $\mathbb{Z}$. We could consider the same thing with individuals located at the integer lattice points $\mathbb{Z}^2$, or in higher dimensions. This example is interesting as a simple model for the spread of a communicable disease. A natural question is whether the disease eventually dies out, i.e. the process reaches the state of all 0s, or if it can survive ($X_n(t) = 1$ for some n) forever. It turns out that there is a critical value λ_f so that if $\lambda \le \lambda_f$ then the infection will die out with probability 1, but for $\lambda_f < \lambda$ the infection can continue forever. (What happens when $\lambda = \lambda_f$ was unresolved for many years until it was finally solved in 1989.) See Durrett [19] for more on this interesting example. Markov processes of this general type are called *interacting particle systems*. There are several important examples in statistical physics.

Problems

Problem 11.1
Suppose $W \geq 0$ is a random variable with the memoryless property (11.2) holding for all $0 \leq s, t$. The purpose of this problem is to show that W *must* be an exponential random variable. Let $g(t)$ denote the function

$$g(t) = P(W > t).$$

Explain why $g(t)$ has these properties:

- $0 \leq g(t) \leq 1$,
- nonincreasing,
- right continuous: $g(t) = \lim_{s \to t^+} g(s)$,
- $\lim_{t \to \infty} g(t) = 0$,
- and satisfies

$$g(s+t) = g(t)g(s) \text{ for all } 0 \leq s, t.$$

An extreme case would be if $g(t) = 0$ for all $t > 0$. In that case explain why $W \equiv 0$. (This might be considered an exponential random variable with $\lambda = \infty$.) Let's dismiss that case and assume $g(t) > 0$ for some $t > 0$. Show that this implies $g(t) > 0$ for all $t > 0$. Similarly use the last bullet above to show that $g(t) < 1$ for all $t > 0$. Since we now know that $0 < g(t) < 1$ for all $t > 0$ we can consider its logarithm: $\phi(t) = \ln(g(t))$. The last bullet above says that

$$\phi(s+t) = \phi(s) + \phi(t).$$

Show that $\phi(t)$ has these properties:

- $\phi(nt) = n\phi(t)$ for all positive integers n and all $t > 0$,
- $\phi(t) = \frac{1}{n}\phi(nt)$ for all positive integers n and all $t > 0$,
- $\phi(\frac{n}{m}) = \frac{1}{m}\phi(n) = \frac{n}{m}\phi(1)$ for all positive integers n, m,
- and $\phi(t) = t\phi(1)$ for all $t > 0$.

So if we let $\lambda = -\phi(1)$ then $\phi(t) = -\lambda t$. Note that $g(1) < 1$ implies $\lambda > 0$. So we have found that for some parameter $\lambda > 0$ the distribution of W must be described by

$$P(t < W) = e^{-\lambda t}.$$

In other words W can only be exponentially distributed if it has the memoryless property.

ExpNoMem

Problem 11.2
Assume that as in Section 11.4.1 the jump rates obey a common bound

$$\lambda_i \leq M \text{ for all } i \in \mathcal{S}.$$

Prove that the associated jump process is nonexplosive. You can do this using the construction $W_n = \frac{1}{\lambda_{X_n}}\tilde{W}_n$ from Section 11.3 along with Lemma 11.2.

Do this a second way using Theorem 11.9. Hint: if $|\varphi(i)| \leq c$ for all i then, by working with the equation $\mathcal{A}\varphi = \alpha\varphi$ deduce that $|\varphi(i)| \leq \frac{\lambda_i}{\alpha+\lambda_i}c \leq \frac{M}{\alpha+M}c$. Conclude that $\varphi \equiv 0$.

Now suppose $\mathcal{S} = \mathbb{N}$ and that there is a bound on the mean jump size.

$$\sum_{j \neq i} q_{i,j}|j - i| \leq B.$$

Show that you can choose α so that $\psi(i) = i+1$ satisfies the hypotheses of Theorem 11.11, giving yet another proof, under the assumption of bounded mean jump size.

BddNExp

Problem 11.3
Explain why the forward equations for $p_{i,j}^{[K]}(t)$ imply (11.33). Then show that it follows from the backward equations as well.

CKC

Problem 11.4
In Section 11.4.3 we did not produce a formula for $p_{i,j}^{[K]}(t)$ when $i \in K$ but $j \in K^c$. These would be the values in the submatrix $\mathbf{P}_{K,K^c}^{[K]}(t)$. Explain why the forward equations say that

$$\mathbf{P}_{K,K^c}^{[K]}{}'(t) = \mathbf{P}_{K,K}^{[K]}(t)\mathbf{B}$$

and so $\mathbf{P}_{K,K^c}^{[K]}(t)$ must be given by the formula

$$\mathbf{P}_{K,K^c}^{[K]}(t) = \int_0^t e^{\mathbf{A}s}\mathbf{B}\,ds.$$

K-Kc

Problem 11.5
The monotonicity property (11.37) can be proven from the forward equations. Let $L = \tilde{K} \setminus K$ and

$$\mathbf{G} = \mathcal{A}_{L,K}.$$

Check that the forward equations say that

$$\frac{d}{dt}\mathbf{P}_{K,K}^{[\tilde{K}]}(t) = \mathbf{P}_{K,K}^{[\tilde{K}]}(t)\mathbf{A} + \mathbf{P}_{K,L}^{[\tilde{K}]}(t)\mathbf{G}.$$

Explain why the solution of this is

$$\mathbf{P}_{K,K}^{[\tilde{K}]}(t) = e^{\mathbf{A}t} + \int_0^t \mathbf{P}_{K,L}^{[\tilde{K}]}(s)\mathbf{G}e^{\mathbf{A}(t-s)}\,ds.$$

(If you right-multiply both sides by $e^{\mathbf{A}t}$ this is an integrating factor calculation.) We know that all entires of $e^{\mathbf{A}(t-s)}$ are nonnegative because they are probabilities. Explain why all entries of the integral term are nonnegative, and therefore

$$\mathbf{P}_{K,K}^{[\tilde{K}]}(t) \geq \mathbf{P}_{K,K}^{[K]}(t),$$

the inequality meaning entry-by-entry.

K-mono

Problem 11.6
Show that if a Markov chain with generator $\mathcal{A}$ is non-explosive then $\mathbf{P}(t)$ is the *only* solution of the forward equations $\mathbf{Q}'(t) = \mathbf{Q}(t)\mathcal{A}$ which has initial values $\mathbf{Q}(0) = \mathbf{I}$, is nonnegative and satisfies $\mathbf{Q}(t)[1] = [1]$.

FNExp

Problem 11.7
Suppose that every nonnegative solution $\mathbf{Q}(t)$ of the forward equations with $\mathbf{Q}(0) = \mathbf{I}$ has $\mathbf{Q}(t)[1] = [1]$. Show that the chain is nonexplosive.

FE1

Problem 11.8
Consider a "pure death process": $i \to i-1$ with rate $\lambda_i > 0$ and $\lambda_0 = 0$. This is non-explosive because the process will reach 0 after a finite number of steps and then never jump again. Take $q_0(t)$ to be a nontrivial infinitely differentiable function with $0 = q_0(0) = q_0'(0) = q_0''(0) = \cdots = q_0^{(n)}(0) = \cdots$. Explain how this determines a solution of $\mathbf{q}'(t) = \mathbf{q}(t)\mathcal{A}$ with $q_i(0) = 0$ for all i.

. . . . . NUF

Problem 11.9
Here is another sufficient condition for explosion. Suppose $\psi : S \to \mathbb{R}$ is nonnegative and satisfies $\mathcal{A}\psi + \alpha \le 0$ for a constant $\alpha > 0$. Applying Theorem 11.7 show that

$$\begin{aligned} E_y[\psi(Y_{t\wedge\mathcal{T}_{K^c}})] &= \psi(y) + E_y[\int_0^{t\wedge\mathcal{T}_{K^c}} \mathcal{A}\psi(Y_s)\,ds] \\ &\le \psi(y) - \alpha E_y[t \wedge \mathcal{T}_{K^c}]. \end{aligned}$$

Explain how, if the process was nonexplosive, we could let $K \uparrow \mathcal{S}$ to conclude

$$0 \le \psi(y) - \alpha t,$$

which would be a contradiction for large t. (See Theorem 4.3.6 of Stroock [57].)

. . . . . StroockCond

Chapter 12

Brownian Motion

The Markov processes we have considered so far all move by making discontinuous jumps. It is remarkable that there are also Markov processes which produce continuous paths. These are generally called *diffusions*. The premier example is Brownian motion. It is named after R. Brown, a botanist who in 1827 observed the erratic motion of small particles suspended in water when observed under a microscope. The idea arose again in the context of financial applications in the Ph.D. thesis of Bachelier [3] in 1900. It is also called the Wiener process in honor of M.I.T. mathematician N. Wiener, who proved it's mathematical existence in 1923. This chapter is an introduction to Brownian motion, its most basic properties, its use in the Black-Scholes model of mathematical finance, and a look at the formalism of Itô calculus. What we offer is only an introductory sampler of properties of Brownian motion and a beginner's guide to working with Itô calculus. We will not attempt to prove the various features we describe.

12.1 Definition and Properties

We will introduce Brownian motion W_t by considering a limit ($n \to \infty$) of scaled random walks. Start with a standard symmetric random walk X_k with initial state $X_0 = 0$. It makes transitions $X_k \to X_{k+1} = X_k \pm 1$, each with probability $1/2$. Thus X_k is defined for integer times $k \geq 0$ and takes integer values. Next we want to rescale time and space in just the right way: let

$$\delta t = \frac{1}{n} \text{ and } \delta x = \frac{1}{\sqrt{n}}$$

and define

$$W_t^{(n)} = \delta x X_{t/\delta t}.$$

Thus in $W_t^{(n)}$ the transitions happen every $1/n$ units of time are of size $1/\sqrt{n}$. Brownian motion W_t is the limit of $W_t^{(n)}$ as $n \to \infty$ (in an appropriate sense).

Consider a single value of t. In that amount of time approximately $k = nt$ jumps will have occurred. If we write $X_k = \sum_1^k Y_i$ where Y_i are i.i.d. with $P(Y_i = \pm 1) = 1/2$ we recognize the Central Limit Theorem as $n \to \infty$:

$$W_t^{(n)} = \delta x X_{t/\delta t} = \frac{1}{\sqrt{n}} \sum_{i=1}^{nt} Y_k = \sqrt{t} \cdot \left(\frac{1}{\sqrt{nt}} \sum_{i=1}^{nt} Y_k \right) \Rightarrow \sqrt{t}\, Z,$$

where Z is a standard normal random variable. So $W_t = \lim_{n\to\infty} W_t^{(n)}$ should be a normal random variable with mean 0 and variance t. We need to remember that the Central Limit Theorem 3.7 only implies convergence "$\Rightarrow$" in distribution, i.e. convergence of the *probabilities* of $W_t^{(n)}$, not convergence of the values of $W_t^{(n)}$ themselves:

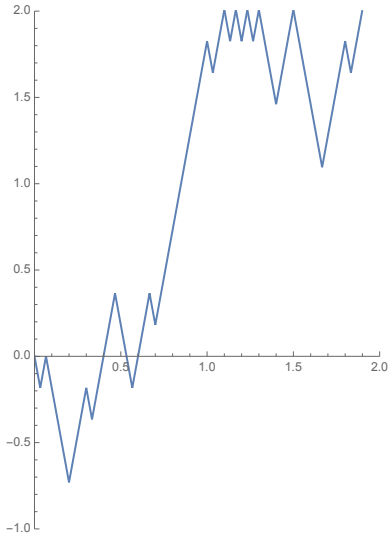
$$P(a \leq W_t \leq b) = \lim_{n\to\infty} P(a \leq W_t^{(n)} \leq b) = \int_a^b \frac{1}{\sqrt{2\pi t}} e^{-\frac{x^2}{2t}}\, dx.$$

Another way to say this is that

$$E[\phi(W_t^{(n)})] \to E[\phi(W_t)] \text{ as } n \to \infty \tag{12.1}$$

for every bounded continuous function $\phi(\cdot)$.

The above describes what happens in the limit for a single t. The remarkable thing is that this limit still exists if we consider $W_t^{(n)}$ as a function of t over an interval $t \in [0, T]$, i.e. all t at once rather than one t at a time. But before we proceed there is a technical issue we should address. We have only defined $W_t^{(n)}$ for those t which are multiples of δt: $t = \frac{k}{n} = k\delta t$ for some k. As n changes the t for which $W_t^{(n)}$ is defined also change. This is awkward, but we can remedy that by using linear interpolation to define $W_t^{(n)}$ for t between multiples of δt. In other words we "connect the dots" to get the graph of $W_t^{(n)}$. Here is a sample of this for $n = 30$. (A plot like this is simple to generate with MATLAB. Simply specify a value for `n` and then enter the command `plot(0:1/n:1,[0,cumsum(randn(1,n))/sqrt(n)])`.)



This makes $W_t^{(n)}$ defined for *all* $0 \le t \le T$, and a continuous function of t. If t is not a multiple of δt, its difference from $W_t^{(n)}$ and the nearest multiple will be small: if $(m-1)\delta t \le t \le m\delta t$ then

$$|W_t^{(n)} - W_{(m-1)\delta t}^{(n)}| \le \delta x \quad \text{and} \quad |W_t^{(n)} - W_{m\delta t}^{(n)}| \le \delta x.$$

We won't go through the details, but the upshot is that in the limit as $n \to \infty$ we can proceed as if all t were multiples of δt; the discrepancy is negligible as $n \to \infty$.

Our main assertion is that there *is* a stochastic process W_t to which $W_t^{(n)}$ converges (in distribution) *as a process*. This means that (12.1) generalizes if $\phi(W_t^{(n)})$ is replaced by $\Phi(W_\cdot^{(n)})$, where $\Phi(\cdot)$ is any bounded continuous "functional" defined on $C([0, T])$ (any $0 < T < \infty$). Basically $\Phi(f(\cdot))$ can be any quantity we can construct from the values of the function $f(t)$ over $t \in [0, T]$ and which is bounded and depends continuously on the choice of $f(\cdot)$. For instance

$$\begin{aligned}
\Phi(f(\cdot)) &= \int_0^T \phi(f(t))\, dt, \\
\Phi(f(\cdot)) &= \phi(f(t_1), f(t_2), \ldots, f(t_N)), \\
\Phi(f(\cdot)) &= \max_{0 \le t \le T} \phi(f(t)),
\end{aligned}$$

using any bounded and continuous ϕ. The assertion is that for any such Φ

$$\lim_{n \to \infty} E[\Phi(W_\cdot^{(n)})] = E[\Phi(W_\cdot)].$$

The W_t whose probabilities occur in the limit above discussion is Brownian motion with initial state $W_0 = 0$. In general a *Brownian motion starting at* $W_0 = w_0$ consists of a collection of random variables W_t, $t \ge 0$ defined on some probability space (Ω, P) with the following essential features:

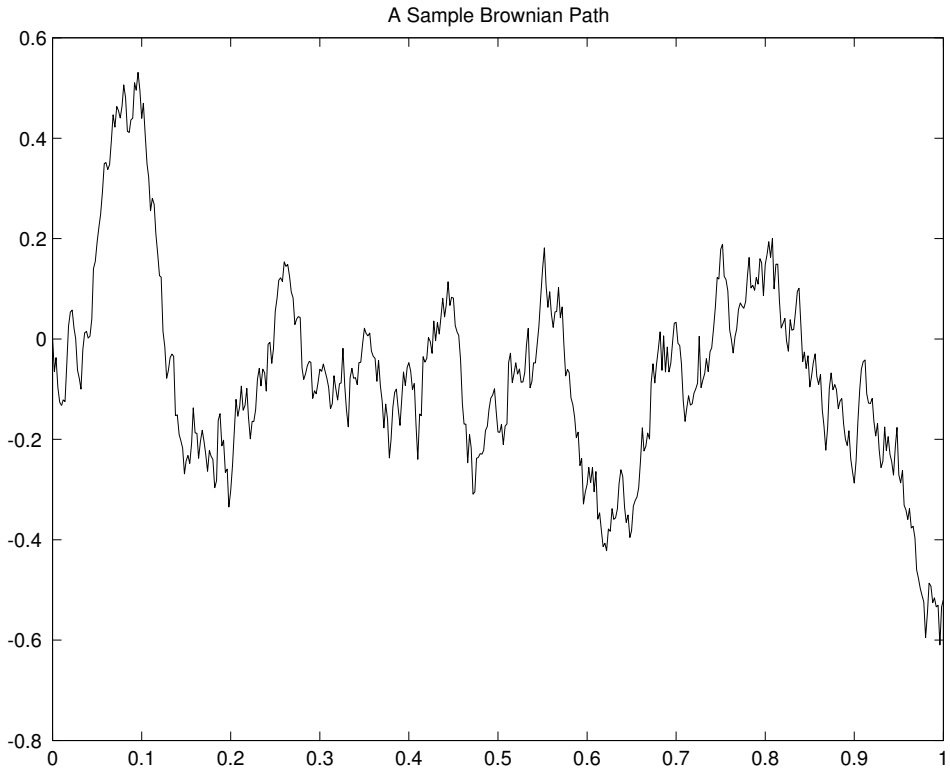
1) $W_0 = w_0$;

2) For each pair $0 \le s < t$, $W_t - W_s$ is independent of $W_{0:s}$ (the history of W_u for all $0 \le u \le s$);

3) For each pair $0 \le s < t$, $W_t - W_s$ has a normal distribution with mean 0 and variance $t - s$.

4) W_t is continuous in t.

There are a number of other equivalent characterizations, but this is most natural for us. You can see what a typical path looks like by plotting $W_t^{(n)}$ for a large n (say $n = 500$).



Let's consider how our description of W_t as the limit of $W_t^{(n)}$ leads to properties 1)–4) above, with $w_0 = 0$. Since $X_0 = 0$ we have $W_0^{(n)} = 0$, so $W_0 = 0$ in part 1) above says. For part 2), consider $s = m\delta t < t = k\delta t$. Then $W_t^{(n)} - W_s^{(n)}$ depends on $Y_{m+1}, \ldots, Y_k$, while all $W_u^{(n)}$ for $u \le s$ depend only on $Y_1, \ldots, Y_m$. Since the Y_i are all independent of each other, this makes it clear that $W_u^{(n)}$ for $u \le s$ is independent of $W_t^{(n)} - W_s^{(n)}$. The independence passes through to the limit as $n \to \infty$. That is because for any $s_1 < s_2 < \cdots < s_M \le s < t$ and bounded continuous functions ϕ, and ψ

$$\begin{aligned}
E[\phi(W_{s_1}, \ldots, W_{s_M})\psi(W_t - W_{s_M})] &= \lim_{n\to\infty} E[\phi(W_{s_1}^{(n)}, \ldots, W_{s_M}^{(n)})\psi(W_t^{(n)} - W_s^{(n)})] \\
&= \lim_{n\to\infty} E[\phi(W_{s_1}^{(n)}, \ldots, W_{s_M}^{(n)})]E[\psi(W_t^{(n)} - W_s^{(n)})] \\
&= \lim_{n\to\infty} E[\phi(W_{s_1}^{(n)}, \ldots, W_{s_M}^{(n)})] \lim_{n\to\infty} E[\psi(W_t^{(n)} - W_s^{(n)})] \\
&= E[\phi(W_{s_1}, \ldots, W_{s_M})]E[\psi(W_t - W_s)].
\end{aligned}$$

Part 3) is a modest generalization of our earlier calculation:

$$\begin{aligned} W_t^{(n)} - W_s^{(n)} &= \delta x \sum_{k=ns+1}^{nt} Y_i \\ &= \sqrt{t-s}\left(\frac{1}{\sqrt{n(t-s)}}\sum_{k=1}^{n(t-s)} Y_{ns+k}\right) \\ &\Rightarrow \sqrt{t-s}\,Z, \end{aligned}$$

where, by the Central Limit Theorem, Z is a standard normal random variable. So $W_t - W_s$ is normal with mean 0 and variance $t-s$, as claimed.

The continuity of W_t in part 4) is the difficult and amazing part. The process W_t is an assemblage of infinitely many independent normal random variables, and yet continuity is a sort of dependence among the different W_t. These ideas seem at odds with each other. (The proof of 4) is quite technical and well beyond what we can describe here.) On the other hand, W_t is *only* continuous. You can see from the picture above that a typical Brownian path is very irregular, although it is continuous. We will say more about this in Section 12.2.4.

12.2 Properties

Brownian motion has many remarkable and important properties. We summarize just a few of them below.

12.2.1 Markov Property and Expected Values

Property 2) above implies the Markov property of W_t. Specifically the independence of $W_t - W_s$ from $W_{[0,s]}$ allows us to drop the conditional expectation in the third line below.

$$\begin{aligned} E[\phi(W_t)|W_{[0,s]} = w_{[0,s]}] &= E[\phi((W_t - W_s) + W_s)|W_{[0,s]} = w_{[0,s]}] \\ &= E[\phi((W_t - W_s) + w_s)|W_{[0,s]} = w_{[0,s]}] \\ &= E[\phi((W_t - W_s) + w_s)] \\ &= \int_{-\infty}^{\infty} \phi(v + w_s)\frac{1}{\sqrt{2\pi(t-s)}}e^{-v^2/2(t-s)}\,dv \\ &= \int_{-\infty}^{\infty} \phi(y)\frac{1}{\sqrt{2\pi(t-s)}}e^{-(y-w_s)^2/2(t-s)}\,dy \\ &= \int_{-\infty}^{\infty} \phi(y)p(w_s, t-s, y)\,dy. \end{aligned}$$

We have used property 3) to write in the density of $W_t - W_s$ in the fourth line. The Markov property is the fact that this depends only on w_s not the full history $w_{[0,s]}$. We have then the transition density for Brownian motion:

$$p(x,h,y) = \frac{1}{\sqrt{2\pi h}}e^{-(y-x)^2/2h}.$$

Here x is the starting state, y is the end state and h is the elapsed time between start and end. Since the state ranges over all real numbers it is no longer reasonable to think of the $p(x,h,y)$ values as making up an infinite matrix. For that reason we are no longer writing the state variables as subscripts as in previous chapters but as arguments to p. One way to express the above is the conditional expectation formula

$$E[\phi(W_t) \mid W_{0:s}] = g(W_s), \quad s < t,$$

where $g(x)$ is the function determined from $\phi(y)$ by

$$g(x) = \int \phi(y)p(x, t-s, y)\,dy. \tag{12.2}$$

This is the analogue of (3.20) for Brownian Motion.

If we combine this with the Tower Law we can calculate as follows.

$$\begin{aligned}\int_{-\infty}^{\infty}\phi(y)p(x,t,y) &= E_x[\phi(W_t)]\\ &= E_x[E[\phi(W_t)|W_{[0,s]}]]\\ &= E_x[g(W_s)]]\\ &= \int_{-\infty}^{\infty} p(x,s,v)g(v)\,dv\\ &= \int_{-\infty}^{\infty} p(x,s,v)\left[\int_{-\infty}^{\infty} p(v,t-s,y)\phi(y)\,dy\right]\,dv\\ &= \int_{-\infty}^{\infty} \phi(y)\left[\int_{-\infty}^{\infty} p(x,s,v)p(v,t-s,y)\,dv\right]\,dy,\end{aligned}$$

from which we extract the Chapman-Kolmogorov formula:

$$p(x,t,y) = \int_{-\infty}^{\infty} p(x,s,v)p(v,t-s,y)\,dv.$$

Since we have an explicit formula for $p(x,h,y)$ this can also be checked directly. (It is just the fact that the sum of independent normal random variables is also normal.)

If you check you will find that for $0 < t$

$$\frac{\partial}{\partial t}p(x,t,y) = \frac{1}{2}\frac{\partial^2}{\partial x^2}p(x,t,y).$$

This is the backward equation for the transition density

$$\frac{\partial}{\partial t}p(x,t,y) = \mathcal{A}p(x,t,y)$$

where the generator is the partial derivative operator

$$\mathcal{A} = \frac{1}{2}\frac{\partial^2}{\partial x^2}$$

acting on the "initial variable" x. From here it should be no suprize that for a bounded continuous function $\phi(\cdot)$,

$$u(x,t) = \int p(x,t,y)\phi(y)\,dy$$

is solves the partial differential equation

$$u_t(x,t) = \frac{1}{2}u_{xx}(x,t) \text{ with } u(x,0) = \phi(x). \tag{12.3}$$

This follows the same pattern as part d) of Theorem 11.3. One way to think of this is that in order to find

$$u(x,t) = E_x[\phi(W_t)]$$

we need to solve (12.3) using $u(x,0) = \phi(x)$. This is now a partial differential equation. Its analogue for Markov chains is that

$$u(i,n) = E_i[\phi(X_n)]$$

is obtained by solving

$$\mathbf{u}(n+1) - \mathbf{u}(n) = \mathbf{A}\mathbf{u}(n)$$

starting from $\mathbf{u}(0) = [\phi(i)]$. This is just a way of writing $\mathbf{u} = \mathbf{P}^n[\phi(i)]$.

12.2.2 Martingale

Brownian motion has several martingale properties. It is itself a martingale, as can be checked using properties 2) and 3) above. Suppose $0 \le s < t \le T$. Then

$$\begin{aligned} E[W_t \mid W_{0:s}] &= E[(W_t - W_s) + W_s \mid W_{0:s}] \\ &= E[(W_t - W_s) \mid W_{0:s}] + W_s \\ &= E[W_t - W_s] + W_s \\ &= 0 + W_s \\ &= W_s. \end{aligned}$$

Another martingale is

$$M_t = e^{\theta W_t - \frac{\theta^2}{2} t}.$$

The verification is again an integral calculation; see Problem 12.2. In fact for M_t to be a martingale for all $\theta \in \mathbb{R}$ implies that W_t is Brownian motion. The reason is that

$$E[e^{\theta(W_t - W_s)} | W_{[0:s]}] = e^{\theta^2/2t - \theta W_s} E[M_t | W_{[0:s]}] = e^{\theta^2/2t - \theta W_s} M_s = e^{\theta^2 (t-s)}.$$

This is giving us the conditional moment generating function of $W_t - W_s$ given $W_{0:s}$. From here it can be proven that $W_t - W_s$ is normal with mean 0 and variance $t - s$ and independent of $W_{0:s}$. By using this idea over $0 < t_1 < t_2 < \cdots < t_n$ you can eventually deduce 2) and 3) of the definition of Brownian motion. This line of reasoning eventually leads to a proof that of the following.

Theorem 12.1. *If W_t is a continuous process with $W_0 = 0$ and*

$$e^{\theta W_t - \theta^2 t/2}$$

is a martingale for all $\theta \in \mathbb{R}$, then W_t is a Brownian motion.

This characterization will be useful in the Black-Scholes section below.

More generally the following characterizes Brownian motion in terms of martingales.

Theorem 12.2. *Suppose W_t is a continuous process with $W_0 = 0$ and that the following is a martingale*

$$M_t = f(W_t, t) - \int_0^t f_t(W_s, s) + \frac{1}{2} f_{xx}(W_s, s)\, ds$$

whenever f, f_t, f_x, f_{xx} are bounded continuous functions. Then W_t is a Brownian motion. Conversely, if W_t is a Brownian motion then M_t is a martingale for all f as described.

Observe how this follows the pattern of Theorems 11.7 and 9.1 above with $\mathcal{A} = \frac{1}{2} \frac{\partial^2}{\partial x^2}$.

12.2.3 Scaling

It is sometimes said that "Brownian motion looks the same on any scale." This is only true if interpreted correctly.

If we take a function $f(t)$ (lets say $f(0) = 0$ for simplicity) and look at its graph under a magnifying glass. A point with coordinates (s, y) on the "viewing screen" of our magnifier corresponds to the point $(t, x) = (s/c, y/c)$ on the sample being magnified, where $c > 1$ is the magnification factor. (If $c < 1$ we would have a "reducer".) So (s, y) will appear on our magnified graph if $y/c = f(s/c)$ is on the original graph. In other words we will see the graph of $y = cf(s/c)$ on the magnifier's viewing screen. This is what happens with a magnifier which rescales the time and space axes by the same factor c. For a differentiable function f with $f(0) = 0$, under high magnification (large c) we will see essentially the graph of the tangent line at 0: $y = f'(0)s$.

If we look at the graph of Brownian motion, with $W_0 = 0$, under the same magnifier we will see the graph of $cW_{s/c}$. Now if you check the definition you will see that $\tilde{W}_s = \sqrt{c}\, W_{s/c}$ is also a Brownian motion,

with s as the time variable. So what we will see under the magnifier is the graph of $\sqrt{c}\tilde{W}_s$, a graph of the Brownian motion $\tilde{W}_s$ but with the vertical scale enlarged by a factor of $\sqrt{c}$. The ragged nature of the graph will be enhanced. If we rescale the time and space axes *differently*, $(t, x) = (s/c, y/\sqrt{c})$, then the graph of $x = W_t$ would appear as the graph of $y = \tilde{W}_s$, so would again be the graph of a Brownian motion. But a standard magnifying glass does not behave this way.

There are other transformations of Brownian motion that result in new Brownian motions:

- $-W_t$,
- $W_{t+t_0} - W_{t_0}$, for any $t_0 \geq 0$;
- $tW_{1/t}$.

The last one is particularly interesting because it reverses the direction of the time axis!

12.2.4 Irregularity

The definition of Brownian motion W_t says that the sample paths are continuous. However, as pictures of the sample paths suggest, they are rather ragged functions. For one thing, they are *never* differentiable in t:

$$P(\ \frac{d}{dt}W_t \text{ exits for some } t) = 0.$$

So we can never talk about W_t' in the usual sense of $' = \frac{d}{dt}$. It is important to remember this when we encounter "dW_t" in stochastic differential expressions below; it will *not* mean $W_t'\,dt$ as you might expect from change of variable calculations in calculus.

The *Law of the Iterated Logarithm* explains more about just how continuous W_t is. It says that for any s, the following two limits hold with probability 1:

$$\limsup_{h\downarrow 0} \frac{W_{s+h} - W_s}{\sqrt{2h\ln(\ln(1/h))}} = 1, \quad \text{and } \liminf_{h\downarrow 0} \frac{W_{s+h} - W_s}{\sqrt{2h\ln(\ln(1/h))}} = -1.$$

This means that if a Brownian path passes through $x = W_s$, then the extremes of its up and down motion for t just beyond s will be described approximately by

$$x \pm \sqrt{2(t-s)\ln(\ln(\frac{1}{t-s}))} \tag{12.4}$$

in the limit as $t \downarrow s$. If we could watch W_t, as we decrease t down to s we would see it move between the top and bottom half of this curve infinitely many times. In particular this means that W_t oscillates dramatically, recrossing $W_t = x$ infinitely many times on any time interval $s < t < s + \delta$, if $W_s = x$. Thus W_t oscillates so frantically in the vertical direction that once it hits a level x it will hit it again infinitely many times in the next split-second.

12.3 Itô Calculus

Section 9.3 pointed out how discrete parameter martingales could be used to form new martingales in an integral-like summation procedure called discrete stochastic integration. With Brownian motion this idea blossoms more fully. Once we define stochastic integrals we also find that there is a "calculus" involving integrals and differentials and a form of the chain rule which can then be used to calculate with stochastic processes in the same way as we do with ordinary functions in freshman calculus. In this section we outline the main features of this calculus.

The key ingredient is the stochastic integral (also called the Itô integral)

$$I_t = \int_0^t \psi_s\, dW_s,$$

where ψ_s is another stochastic process. We pointed out above that Brownian motion has no derivative in the conventional sense, so the dW_s above cannot be interpreted using $dW_s = W_s' \, ds$ as change of variable based on ordinary techniques from calculus. Moreover the Brownian paths have infinite arc length, so the integral cannot be defined as a kind of integral along paths either. There is a way to make sense of the above integral, but it depends very much on stochastic properties of Brownian motion, especially its martingale properties.

Simple Integrands

Integrals are usually defined by means of a limiting process: first identify a special type of integrand ψ_s for which we can write down what $\int_0^t \psi_s \, dW_s$ should be directly using finite sums. Then for a more general ψ_s define

$$\int_0^t \psi_s \, dW_s = \lim_{n\to\infty} \int_0^t \psi_s^{(n)} \, dW_s$$

where $\psi_s^{(n)} \to \psi_s$ and the $\psi_s^{(n)}$ are integrands of the special type. For instance the Riemann integral of calculus $\int f(x) \, dx$ is defined in this way, where the special type of integrand is a piecewise constant function (whose integral is a Riemann sum). A similar idea is used here. The special type of integrand is called a *simple* process: piecewise constant, $W_{0:t}$-determined and square-integrable. This means that there exist some $0 = t_0 < t_1 < \ldots < t_m = t$ and random variables $X_1, \ldots, X_m$ so that

- $\psi_s = X_i$ for s in $(t_{i-1}, t_i]$,
- each X_i is $W_{0:t_{i-1}}$-determined,
- each $E[X_i^2] < \infty$.

For such a ψ_s we define

$$I_t = \int_0^t \psi_s \, dW_s = \sum_{i=1}^m X_i (W_{t_i} - W_{t_{i-1}}).$$

When ψ_s is not simple want to approximate it using a sequence $\psi_s^{(n)}$ of simple processes, form $I_t^{(n)} = \int_0^t \psi_s^{(n)} \, dW_s$ for each of them (as above), and then take the limit

$$I_t = \int_0^t \psi_s \, dW_s = \lim_{n\to\infty} I_t^{(n)}.$$

To form simple approximates $\psi_s^{(n)}$ we can divide $[0, t]$ up using $t_i = \frac{i}{n} t$ for $i = 0, \ldots, n$ and in each interval just "freeze" ψ_s at the start of each interval.

$$\psi_s^{(n)} = \psi_{t_{i-1}} \text{ for } t_{i-1} < s \le t_i. \tag{12.5}$$

An Example

As an example of the above strategy for defining stochastic integrals of non-simple ψ_t, we will work out the case of $\psi_t = W_t$:

$$\int_0^t W_s \, dW_s.$$

Our approximation strategy (above) produces

$$\psi_s^{(n)} = W_{t_{i-1}} \text{ for } t_{i-1} < s \le t_i.$$

To write out the approximating integrals we will use the notations

$$\Delta W_{t_i} = W_{t_i} - W_{t_{i-1}} \text{ and } \Delta(W_{t_i}^2) = W_{t_i}^2 - W_{t_{i-1}}^2.$$

We have

$$I_t^{(n)} = \int_0^t \psi_s^{(n)}\, dW_s = \sum_{i=1}^{n} W_{t_{i-1}} \Delta W_{t_i}$$

To simplify this, observe that

$$\begin{aligned}\frac{1}{2}[\Delta(W_{t_i}^2) - (\Delta W_{t_i})^2] &= \frac{1}{2}[W_{t_i}^2 - W_{ti-1}^2 - (W_{t_i}^2 - 2W_{t_i}W_{t_{i-1}} + W_{t_{i-1}}^2)] \\ &= \frac{1}{2}[2W_{t_i}W_{t_{i-1}} - 2W_{t_{i-1}}^2] \\ &= W_{t_{i-1}}\Delta W_{t_i}.\end{aligned}$$

So we find that

$$\begin{aligned}I_t^{(n)} &= \frac{1}{2}\sum_{i=1}^{n}[\Delta(W_{t_i}^2) - (\Delta W_{t_i})^2] \\ &= \frac{1}{2}W_t^2 - \frac{1}{2}\sum_{i=1}^{n}(\Delta W_{t_i})^2\end{aligned}$$

To understand the last summation, the ΔW_{t_i} are independent normal random variables with mean 0 and variance t/n. We might write $\Delta W_{t_i} = \sqrt{t/n}Y_i$ where the Y_i are standard normal. So

$$\sum_{i=1}^{n}(\Delta W_{t_i})^2 = \frac{t}{n}\sum_{i=1}^{n} Y_i^2.$$

This looks like the Law of Large Numbers as $n \to \infty$. Since $E[Y_i^2] = 1$ this suggests that

$$\sum_{i=1}^{n}(\Delta W_{t_i})^2 \to t. \tag{12.6}$$

Its not actually this simple, because as we change n we change the ΔW_{t_i} by breaking the Brownian path up into smaller increments, so that the Y_i themselves change with n. In other words we are not working with a single i.i.d. sequence of Y_i. Nonetheless, it turns out that (12.6) is still correct under a different notion of convergence:

$$E\left[\left(\sum_{i=1}^{n}(\Delta W_{t_i})^2 - t\right)^2\right] \to 0.$$

We find then that with an appropriate notion of convergence

$$\int_0^t W_s\, dW_s = \lim_{n\to\infty} I_t^{(n)} = \frac{1}{2}(W_t^2 - t). \tag{12.7}$$

Contrast this with the conventional integral $\int_0^t s\, ds = \frac{1}{2}t^2$.

A second example that can be worked out by hand (see Problem 12.3) is

$$\int_0^t s\, dW_s = tW_t - \int_0^t W_s\, ds, \tag{12.8}$$

the integral on the right being a conventional Riemann integral.

There are several comments to be made about this calculation and the approach we have described to obtain $\int_0^t \psi_s\, dW_s$ as $I_t = \lim I_t^{(n)}$ in general.

- We set up the partition t_i so that the upper limit of integration t was a partition point: $t = t_n$. That is just a convenience to keep our calculations above as simple as possible. If t fell between two partition points $t_n < t < t_{n+1}$ we would need to include an extra term in the above to account for the last bit of the integral from t_n to t. A different way to handle it would be to start with a partition t_i and then if the t we are interested in is not one of the partition points, just insert it as a new partition point.

- The convergence (12.6) is not simple almost sure convergence, but so-called "mean-square" convergence. That turns out to be the appropriate sense in which $I_t^{(n)} \to I_t$ in general.
- Our freeze-at-the-left-endpoint approach (12.5) for constructing approximating $\psi_s^{(n)}$ for a given partition will work if ψ_s is continuous (or left-continuous), is $W_{0:s}$-determined for each s (usually described by saying that ψ_s is *adapted* to the Brownian motion) and satisfies some integrability hypothesis. A full treatment would need to specify the sense in which we need $\psi_s^{(n)} \to \psi_s$. That turns out to be an integrated mean-square sense. But we are not going to develop those details.
- In the usual Riemann integral $\int_0^t f(s)\,ds$ if we approximate $f(s)$ on an interval $t_{i-1} \le s \le t_i$ using a single value $f(s_i)$, it won't matter in the limit whether we use the left endpoint $s_i = t_{i-1}$ or right endpoint or midpoint; we will get the same limit $\int_0^t f(s)\,ds$ regardless. But for the stochastic integral it *does* matter. The choice of the left endpoint in (12.5) is what insures that the value of $\psi_s^{(n)}$ is $W_{0:t_{i-1}}$-determined on $[t_{i-1}, t_i]$. That in turn is important for the martingale properties of $\int_0^t \psi_s\,dW_s$. If, for instance, you use the right endpoint instead in our example above, you will get a different limit in (12.7)!

When the theory is worked out in general (which we are *not* doing here) the integrands ψ_s which can be allowed are those for which

- ψ_s is adapted to the Brownian motion (i.e. is $W_{0:s}$-determined for each s),
- $E\left[\int_0^T \psi_t^2\,dt\right] < \infty$,
- is "progressively measurable". This is a form of regularity in it's s-dependence which is too technical for us to describe. A sufficient condition which is adequate for our purposes is that ψ_s is continuous.

We will call these *admissible* integrands. For an admissible integrand the resulting stochastic integral

$$I_t = \int_0^t \psi_s\,dW_s$$

is best considered as a stochastic process in its own right, like the indefinite integral of calculus, rather than for a fixed t alone. Here are some of the most important properties.

1. $I_0 = 0$;
2. I_t is is $W_{0:t}$-determined and continuous in t;
3. I_t is a martingale;
4. $I_t^2 - \int_0^t \psi_s^2\,ds$ is a martingale.

If ψ_s and ϕ_s are two admissible integrands then the following hold.

5. $\int_0^t a\psi_s + b\phi_s\,dW_s = a\int_0^t \psi_s\,dW_s + b\int_0^t \phi_s\,dW_s$ for any two constants a, b;
6. $(\int_0^t \psi_s\,dW_s)(\int_0^t \phi_s\,dW_s) - \int_0^t \psi_s\phi_s\,ds$ is a martingale.
7. $E[(\int_0^t \psi_s\,dW_s)(\int_0^t \phi_s\,dW_s)] = E[\int_0^t \psi_s\phi_s\,ds]$.

These are verificatied by first proving them for simple integrands, and then passing to the limit to get the general case. Suppose that $\psi_s = X_i$ on $(t_{i-1}, t_i]$. If t falls in the interval $t_k < t \le t_{k+1}$. Then we can write

$$I_t = \sum_{i=1}^{k-1} X_i(W_{t_i} - W_{t_{i-1}}) \quad + \quad X_k(W_t - W_{t_k}).$$

From this you should be able to convince yourself that I_t is indeed continuous in t as claimed in 2, because W_t is. What about the martingale property, 3? Considered at just at the discrete set of times t_i, we see that

I_{t_i} is a martingale, because it is just an instance of discrete stochastic integration as in (9.3). To check that $I_s = E[I_t|\ W_{0:s}]$ when s and t are not among the t_i, observe that we can just insert them into the list of the t_i and using some duplicate copies of the X_i on the new intervals, thereby re-expressing ψ_t as a simple process with an enlarged set of t_i that now *does* include both s and t.

The assertion 4 above is especially important. If $0 \le s < t$, we want to show that

$$E[I_t^2 - \int_0^t \psi_u^2\, du \mid W_{0:s}] = I_s^2 - \int_0^s \psi_u^2\, du$$

Consider a simple integrand with $\psi_s = X_i$ for $t_{i-1} < s \le t_i$. We can assume s and t are among the t_i. Now the calculations are essentially the same as for the martingale B_n of Problem 9.1. Observe that

$$\Delta I_{t_i} = I_{t_i} - I_{t_{i-1}} = X_i(W_{t_i} - W_{t_{i-1}}).$$

Adopting the notation from Problem 9.1 we have

$$\begin{aligned}\overline{(\Delta I_{t_i})^2} &= E[(\Delta I_{t_i})^2 | W_{0:t_{i-1}}] \\ &= E[X_i^2 (W_{t_i} - W_{t_{i-1}})^2 | W_{0:t_{i-1}}] \\ &= X_i^2 E[(W_{t_i} - W_{t_{i-1}})^2 | W_{0:t_{i-1}}] \\ &= X_i^2 \Delta t_i.\end{aligned}$$

So we have

$$\begin{aligned}I_{t_n}^2 - \sum_1^n \overline{(\Delta I_{t_i})^2} &= I_{t_n}^2 - \sum_1^n X_i^2 \Delta t_i \\ &= I_{t_n}^2 - \int_0^{t_n} \psi_s^2\, ds.\end{aligned}$$

12.3.1 The Formal Structure of Itô Calculus

Just as with freshman calculus, it is rare that we actually calculate a stochastic integral from the definition. We usually work from a few elementary known integrals together with various rules for manipulation, such as the technique of substitution (which is really the Chain Rule). We will describe these rules in this subsection.

The stochastic processes we want to work with are all ones that can be written as sums of Riemann and stochastic integrals, i.e. X_t for which

$$X_t = X_0 + \int_0^t \phi_s\, ds + \int_0^t \psi_s\, dW_s,$$

for some admissible stochastic integrands ϕ_t and ψ_t. This is what is meant when we write the stochastic differential relationship

$$dX_t = \phi_t\, dt + \psi_t\, dW_t.$$

The differentials dX_t and dW_t have no direct meaning. The differential relationship really refers to its integrated form,

$$X_t - X_0 \ \left(= \int_0^t dX_t\right) \ = \int_0^t \phi_s\, ds + \int_0^t \psi_s\, dW_s.$$

It is important to note here that the second (stochastic) integral is a martingale but the first (Riemann) integral is not (unless $\phi_s \equiv 0$). This is a very important practical observation: **we can recognize martingales because their stochastic differential will contain no dt term, only a dW_t term!** (Actually this is not quite right: $dX_t = 0\, dt + \psi_t\, dW_t$ makes X_t a *local* martingale. A little more is needed to insure that it is a true martingale. A sufficient condition is that $E[\int_0^T \psi_s^2\, ds] < \infty$. But we don't want to tackle these technicalities.)

Stochastic differential expressions can be manipulated following a couple basic rules. For instance suppose we have differential expressions for two stochastic processes,

$$dX_t = \phi_t^X\, dt + \psi_t^X\, dW_t, \quad dY_t = \phi_t^Y\, dt + \psi_t^Y\, dW_t.$$

The stochastic differential for the product is obtained from the *stochastic product rule*

$$d(X_tY_t) = X_t\, dY_t + Y_t\, dX_t + dX_t\, dY_t. \tag{12.9}$$

This differs from the conventional product rule because we retain the product of the differentials: $dX_t\, dY_t$. To work out $dX_t\, dY_t$ substitute in the differentials dX_t and dY_t in terms of dt and dW_t and use the basic differential multiplication formulas

$$(dt)^2 = 0, \quad dt \cdot dW_t = 0, \quad (dW_t)^2 = dt. \tag{12.10}$$

to reduce the right side of (12.9).

For instance,

$$\begin{aligned} d(W_t^2) &= W_t\, dW_t + W_t\, dW_t + dW_t\, dW_t \\ &= 2W_t\, dW_t + dt, \end{aligned}$$

which agrees with our example (12.7). In the case of (12.8) we get

$$d(tW_t) = t\, dW_t + W_t\, dt + dt\, dW_t = t\, dW_t + W_t\, dt + 0.$$

The same method produces a correct stochastic differential expression from (12.9) in general. Actually there are some technical qualifications, related to whether the product of two admissible integrands is also an admissible integrand. We will just ignore those issues, since our concern here is to learn how to manipulate formulas using stochastic differentials. If you study this stochastic calculus again with a more rigorous approach, those issues will have to be dealt with.

Itô's Formula

We can continue to build up more stochastic differential formulas. For instance, using our formula for $d(W_t^2)$ from above,

$$\begin{aligned} d(W_t^3) &= d(W_t\, W_t^2) \\ &= W_t\, d(W_t^2) + W_t^2\, dW_t + d(W_t^2) \cdot dW_t \\ &= W_t[dt + 2W_t\, dW_t] + W_t^2\, dW_t + [dt + 2W_t\, dW_t] \cdot dW_t \\ &= 3W_t^2\, dW_t + 3W_t\, dt. \end{aligned}$$

If you continue working up to higher powers you will find that

$$d(W_t^n) = nW_t^{n-1}\, dW_t + \frac{n(n-1)}{2} W_t^{n-2}\, dt. \tag{12.11}$$

You will recognize the two terms on the right as $f'(W_t)$ and $\frac{1}{2}f''(W_t)$, for $f(x) = x^n$. We are starting to see the emergence of Itô's formula.

For conventional calculus the Chain Rule can be exposed as follows. if $x'(t) = \psi(t)$ and $f(x)$ is a continuously differentiable function, then

$$\frac{d}{dt} f(x(t)) = f'(x(t))\psi(t)$$

so that

$$f(x(b)) = f(x(a)) + \int_a^b f'(x(t))\psi(t)\, dt.$$

In a differential notation, the conventional Chain Rule could be written this way. Assuming

$$dx_t = \psi_t\,dt,$$

and $f(x)$ is a continuously differentiable then

$$df(x_t) = f'(x_t)\psi_t\,dt.$$

Itô's formula is an extension of this form of the Chain Rule to stochastic differentials for which dX_t includes a dW_t term as well.

Theorem 12.3 (Itô's Formula)**.** *Suppose X_t has a stochastic differential using admissible integrands, and $f(x)$ is twice continuously differentiable. Then $f(X_t)$ has the stochastic differential*

$$df(X_t) = f'(X_t)\,dX_t + \frac{1}{2}f''(X_t)\,(dX_t)^2,$$

provided all the resulting integrands are admissible.

Of course we need to expand and simplify the right side using (12.10). Notice how much the formula resembles the first two terms of a Tayor expansion. That's a good way to remember it. Itô calculus is sometimes called a second order calculus because the second order terms are important.

Example 12.1. As an example consider $\zeta_t = e^{\theta W_t - \frac{1}{2}\theta^2 t}$, our familiar exponential martingale. With $X_t = \theta W_t - \frac{1}{2}\theta^2 t$ and $f(x) = e^x$ we can write $\zeta_t = f(X_t)$ and then work out its differential using Itô's formula.

$$\begin{aligned} d\zeta_t &= e^{X_t}\,dX_t + \frac{1}{2}e^{X_t}\,(dX_t)^2 \\ &= \zeta_t\,[-\frac{1}{2}\theta^2\,dt + \theta\,dW_t + \frac{1}{2}\theta^2\,dt] \\ &= \theta\zeta_t\,dW_t. \end{aligned}$$

This explains why we should expect it to be a martingale: there is no dt term. This would be a proof that it's a martingale if we took the trouble to verify that $\theta\zeta_t$ is an admissible stochastic integrand. That's not hard to do, but we won't.

Suppose $f(x_1, \ldots, x_n)$ is a function of several variables, with all second order partial derivatives continuous, and we have several stochastic processes $X_t^{(i)}$ each with stochastic differentials. For brevity, lets write $f(X_t)$ for $f(X_t^{(1)}, \ldots, X_t^{(n)})$. *Itô's formula for functions of several variables* says that

$$df(X_t) = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(X_t^{(\cdot)})\,dX_t^{(i)} + \frac{1}{2}\sum_{i,j=1}^{n} \frac{\partial^2 f}{\partial x_i \partial x_j}(X_t)\,(dX_t^{(i)}\,dX_t^{(i)}), \tag{12.12}$$

again with the technical qualification that all resulting integrands be admissible. One important clarification: The double sum $\sum_{i,j=1}^{n}$ is a sum of a sum:

$$\sum_{i,j=1}^{n} \ldots = \sum_{i=1}^{n} \left[\sum_{j=1}^{n} \ldots \quad \right].$$

This means that the "mixed" second order partial derivatives ($i \neq j$) will each occur *twice*. For instance $\frac{\partial^2 f}{\partial x_1 \partial x_2}$ occurs once for $i = 1, j = 2$ and again for $i = 2, j = 1$. The "diagonal" terms ($i = j$) only occur once.

To illustrate this we will compute the stochastic differential for

$$Y_t = \sin(W_t^2)W_t.$$

The simplest approach would be to apply Itô's formula for functions of a single variable to $f(x) = x\sin(x^2)$. However we want to illustrate (12.12). With

$$f(x_1, x_2) = \sin(x_1)x_2,$$

and

$$X_t^{(1)} = W_t^2, \quad dX_t^{(1)} = 2W_t\,dW_t + dt;$$
$$X_t^{(2)} = W_t, \quad dX_t^{(2)} = dW_t,$$

we have

$$Y_t = f(X_t^{(1)}, X_t^{(2)}),$$

to which we apply Itô's formula. We will use f_{x_i} and f_{x_i,x_j} to denote the various partial derivatives. These will always be assumed to be evaluated at $X_t^{(1)}, X_t^{(2)}$ in the dY_t calculation below. Itô's formula says that

$$dY_t = f_{x_1}\,dX_t^{(1)} + f_{x_2}\,dX_t^{(2)} + \frac{1}{2}\left[f_{x_1,x_1}(dX_t^{(1)})^2 + 2f_{x_1,x_2}dX_t^{(1)}dX_t^{(2)} + f_{x_2,x_2}(dX_t^{(2)})^2\right].$$

Here are the various pieces:

$$\begin{aligned}
f_{x_1} &= \cos(x_1)x_2 \\
f_{x_2} &= \sin(x_1) \\
f_{x_1x_1} &= -\sin(x_1)x_2 \\
f_{x_1x_2} &= \cos(x_1) \\
f_{x_2x_2} &= 0 \\
(dX_t^{(1)})^2 &= 4W_t^2\,dt \\
dX_t^{(1)}dX_t^{(2)} &= 2W_t\,dt
\end{aligned}$$

Now assemble the pieces and make the substitutions for $X_t^{(1)}$ and $X_t^{(2)}$ to obtain

$$\begin{aligned}
dY_t &= \cos(W_t^2)W_t\,(2W_t\,dW_t + dt) + \sin(W_t^2)\,dW_t + \frac{1}{2}\left[-\sin(W_t^2)W_t\,4W_t^2 + 2\cos(W_t^2)2W_t\right]dt \\
&= \left[\cos(W_t^2)2W_t^2 + \sin(W_t^2)\right]\,dW_t + \left[3\cos(W_t^2)W_t - 2\sin(W_t^2)W_t^3\right]\,dt.
\end{aligned}$$

12.4 The Black-Scholes Model and Option Pricing

We return to the topic of mathematical finance to illustrate the application of martingale properties of Brownian Motion and stochastic calculus. Just as in Chapter 10 we will consider a bank process

$$B_t = e^{rt}.$$

Here the constant $r > 0$ is the *continuously compounded* interest rate. We understand B_t as the value at time t of \$1 deposited at time 0. Its dynamics are described in differential notation by

$$dB_t = rB_t\,dt.$$

For the stock price process S_t we want a process with a stochastic differential

$$dS_t = \mu S_t\,dt + \sigma S_t\,dW_t. \tag{12.13}$$

The idea is that $\mu > 0$ is the mean relative growth rate of the value of a share of stock, similar to the interest rate for B_t. But we want the stock price to have some random fluctuation. The $\sigma S_t\,dW_t$ term is intended to provide that. The dW_t is viewed as providing up-down fluctuations from an underlying Brownian motion and the σS_t makes the magnitude of the fluctuations proportional to the price itself. The constant σ (usually called the *volatility*) determines how strongly the dW_t fluctuations influence dS_t.

Equation 12.13 does not tell us directly what S_t. It is only a *stochastic differential equation* that S_t must satisfy. Given an initial value S_0 our first task is see if we can solve for S_t. Looking back at Example 12.1 suggests that by picking constants c_1 and c_2 we might be able to get

$$S_t = S_0 e^{c_1 t + c_2 W_t}$$

to work. Calculating as in the example,

$$\begin{aligned} dS_t &= S_t(c_1\,dt + c_2\,dW_t) + \frac{1}{2}S_t(c_1\,dt + c_2\,dW_t)^2 \\ &= S_t(c_1\,dt + c_2\,dW_t) + \frac{1}{2}S_t c_2^2\,dt \\ &= (c_1 + \frac{1}{2}c_2^2)S_t\,dt + c_2 S_t\,dW_t. \end{aligned}$$

We can now easliy pick the constants to fit (12.13):

$$c_2 = \sigma \text{ and } c_1 = \mu - \frac{1}{2}\sigma^2.$$

So the desired stock price process is

$$S_t = S_0 e^{(\mu - \frac{1}{2}\sigma^2)t + \sigma W_t}. \tag{12.14}$$

This with B_t as above is the *Black-Scholes* model of a market with a single stock in continuous time.

We will work out the option pricing formulas for this model by following the basic theoretical structure we discovered in the discrete time setting. The first step is to find an "equivalent" probability measure Q with respect to which $M_t = S_t/B_t$ is a Q-martingale. Observe that

$$M_t = S_0 e^{\sigma W_t + (\mu - \frac{1}{2}\sigma^2)t} e^{-rt}.$$

Following the lead of Section 9.7 we expect this to be described by a nonnegative martingale ζ_t with $\zeta_0 = 1$. If we fix the time horizon T, Q will be related to P by

$$Q(A) = E[\zeta_T; A]$$

for $S_{0:T}$-determined events A. In fact our basic exponential martingale

$$\zeta_t = e^{\theta W_t - \frac{1}{2}\theta^2 t}$$

will be exactly what we need, for a carefully selected value of θ. But how can we recognize when a process M_t is a Q-martingale? We thought about this in Section 9.7 and found that we need $M_t\zeta_t$ to be a P-martingale. That means the following should be a P-martingale.

$$\begin{aligned} M_t\zeta_t &= S_0 e^{\sigma W_t + (\mu - \frac{1}{2}\sigma^2)t} e^{-rt} e^{\theta W_t - \frac{1}{2}\theta^2 t} \\ &= S_0 e^{(\sigma+\theta)W_t - (\frac{1}{2}\theta^2 + r + \frac{1}{2}\sigma^2 - \mu)t}. \end{aligned}$$

By Theorem 12.1 this will be a martingale if

$$\frac{1}{2}(\sigma+\theta)^2 = \frac{1}{2}\theta^2 + r + \frac{1}{2}\sigma^2 - \mu,$$

which reduces to

$$\theta = (r - \mu)/\sigma.$$

So this value of θ accomplishes what we have been seeking: it gives us an equivalent probability measure Q which makes $M_t = S_t/B_t$ a martingale.

The next step is to recognize that with respect to Q the following is a Brownian motion (for $0 \le t \le T$),

$$W_t^Q = W_t - \theta t.$$

In other words under the new Q the original W_t is no longer a Brownian motion, but a modified version is. The simplest way to confirm this is to use Theorem 12.1: check that for any γ

$$\xi_t = e^{\gamma W_t^Q - \frac{1}{2}\gamma^2 t}$$

is a Q-martingale (for $t \le T$). For that we check that $\xi_t \zeta_t$ is a P-martingale. This is simple, because

$$\begin{aligned}\xi_t \zeta_t &= e^{\gamma(W_t - \theta t) - \frac{1}{2}\gamma^2 t + \theta W_t - \frac{1}{2}\theta^2 t} \\ &= e^{(\gamma+\theta)W_t - \frac{1}{2}(\gamma+\theta)^2 t},\end{aligned}$$

just another version of our original exponential P-martingale.

Calculations with respect to Q are more natural if we rewrite S_t in terms of W_t^Q rather than W_t:

$$\begin{aligned}S_t &= S_0 e^{\sigma W_t + (\mu - \frac{1}{2}\sigma^2)t} \\ &= S_0 e^{\sigma(W_t - \theta t) + (\sigma\theta + \mu - \frac{1}{2}\sigma^2)t} \\ &= S_0 e^{\sigma W_t^Q + (r - \frac{1}{2}\sigma^2)t}.\end{aligned}$$

We now can price options using the risk-neutral formula from Chapter 10:

$$v(t,s)/B_t = E^Q[\phi(S_T)/B_T \,|\, S_t = s].$$

Said otherwise,

$$v(S_t, t)e^{-rt} = E^Q[\phi(S_T)e^{-rT}|S_{0:t}]. \tag{12.15}$$

So the calculation boils down to

$$\begin{aligned}v(s,t) &= e^{r(T-t)} E^Q[\phi(S_T)|S_t = s] \\ &= e^{r(T-t)} \int \phi(se^{\sigma y + (r - \frac{1}{2}\sigma^2)(T-t)}) \frac{1}{\sqrt{2\pi(T-t)}} e^{-y^2/2(T-t)}\, dy.\end{aligned}$$

It looks nasty, but it is explicit. (It would be simple to write a MATLAB m-file to compute the right hand side by numerical integration for given s, σ, r, T, t and $\phi(\cdot)$.)

We should make it clear that we have just blindly followed the patterns from Chapter 10 to obtain these formulas. Their justification really depends on making clear what we mean by self-financing portfolios in this setting, and to verify that there *does exist* a self-financing portfolio which replicates $\phi(S_T)$ for any exercise value function $\phi(s)$ (subject to technical hypotheses). A more sophisticated analysis is needed to address these issues properly, so we will not attempt it. But we will at least point out what "self-financing" is usually taken to mean in terms of Itô calculus in Section 12.4.2 below. First however let's work out the most famous particular example of (12.15).

12.4.1 The Black-Scholes Formula

Consider the case of a call option: $\phi(s) = \max(s - K, 0)$. We will work out the calculation of (12.15), which will produce the famous Black-Scholes formula. Some notation will help clean things up.

$$\begin{aligned}\tau &= T - t \\ \tilde{r} &= r - \frac{1}{2}\sigma^2 \\ z_0 &= \frac{\ln(K/s) - \tilde{r}\tau}{\sigma\sqrt{\tau}}.\end{aligned}$$

With the change of variable $y = \sqrt{\tau} z$ our calculation becomes

$$v(s,t) = e^{-r\tau} \int_{-\infty}^{\infty} \phi(se^{\tilde{r}\tau + \sigma\sqrt{\tau}z}) p(z)\, dz,$$

where $p(z)$ is the standard normal density,

$$p(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}.$$

For

$$\phi(s) = \begin{cases} s - K & \text{if } s - K \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

we want to integrate over those z for which

$$se^{\tilde{r}\tau + \sigma\sqrt{\tau}z} \geq K,$$

which is equivalent to

$$z \geq z_0 \text{ as defined above.}$$

So

$$\begin{aligned} \int_{-\infty}^{\infty} \phi(se^{\tilde{r}\tau+\sigma\sqrt{\tau}z})p(z)\,dz &= \int_{z_0}^{\infty} (se^{\tilde{r}\tau+\sigma\sqrt{\tau}z} - K)p(z)\,dz \\ &= s\int_{z_0}^{\infty} e^{\tilde{r}\tau+\sigma\sqrt{\tau}z}p(z)\,dz - K\int_{z_0}^{\infty} p(z)\,dz \end{aligned}$$

The integral in the second term is

$$\begin{aligned} \int_{z_0}^{\infty} p(z)\,dz &= \int_{-\infty}^{-z_0} p(\hat{z})\,d\hat{z}, \text{ using } \hat{z} = -z \\ &= \mathcal{N}(-z_0), \end{aligned}$$

where

$$\mathcal{N}(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

is the standard normal distribution function. The integral in the first term can be rewritten as follows.

$$\begin{aligned} \int_{z_0}^{\infty} e^{\tilde{r}\tau+\sigma\sqrt{\tau}z}p(z)\,dz &= \frac{e^{\tilde{r}\tau}}{\sqrt{2\pi}} \int_{z_0}^{\infty} e^{-z^2/2+\sigma\sqrt{\tau}z}\,dz \\ &= e^{(\tilde{r}+\sigma^2/2)\tau} \int_{z_0}^{\infty} e^{-\frac{1}{2}(z-\sigma\sqrt{\tau})^2} \frac{1}{\sqrt{2\pi}}\,dz \\ &= e^{r\tau} \int_{-\infty}^{\sigma\sqrt{\tau}-z_0} p(\hat{z})\,d\hat{z}, \text{ using } \hat{z} = \sigma\sqrt{\tau} - z \\ &= e^{r\tau}\mathcal{N}(\sigma\sqrt{\tau} - z_0). \end{aligned}$$

Putting the pieces back together, we have

$$v(s,t) = s\mathcal{N}(\sigma\sqrt{\tau} - z_0) - e^{-r\tau}K\mathcal{N}(-z_0).$$

The two arguments of $\mathcal{N}(\cdot)$ are usually expressed as follows.

$$\begin{aligned} \sigma\sqrt{\tau} - z_0 &= \sigma\sqrt{\tau} - \frac{\ln(K/s) - \tilde{r}\tau}{\sigma\sqrt{\tau}} \\ &= \frac{\ln(s/K/) + (\tilde{r} + \sigma^2)\tau}{\sigma\sqrt{\tau}} \\ &= \frac{\ln(s/K/) + (r + \frac{1}{2}\sigma^2)(T-t)}{\sigma\sqrt{T-t}} \end{aligned}$$

and

$$-z_0 = (\sigma\sqrt{\tau} - z_0) - \sigma\sqrt{T-t}.$$

Here then is the usual form of the *Black-Scholes formula* for a European call:

$$v^{\text{call}}(s,t) = s\mathcal{N}(d_1(s,t)) - e^{-r(T-t)}K\mathcal{N}(d_2(s,t)), \tag{12.16}$$

where

$$\begin{aligned} d_1(s,t) &= \frac{\ln(s/K) + (r + \frac{1}{2}\sigma^2)(T-t)}{\sigma\sqrt{T-t}} \\ d_2(s,t) &= d_1(s,t) - \sigma\sqrt{T-t} \\ &= \frac{\ln(s/K) + (r - \frac{1}{2}\sigma^2)(T-t)}{\sigma\sqrt{T-t}}. \end{aligned}$$

12.4.2 Itô Calculus and Self-Financing Portfolios

Our next goal is to see how Itô calculus reveals the features we have discussed for the Black-Scholes model. Let's start with the change of probability measure from P to Q. This was designed so that $M_t = S_t/B_t$ would be a Q-martingale. We also recognized that it had the effect that

$$W_t^Q = W_t - \theta t$$

is a Q Brownian motion. In terms of differentials

$$dW_t^Q = dW_t - \theta\, dt.$$

For working with respect to Q we will want to replace all the dW_t with dW_t^Q by means of the above relationship. For instance the natural way to arrange the stochastic differential of S_t for consideration under Q is as follows.

$$\begin{aligned} dS_t &= \mu S_t\, dt + \sigma S_t\, dW_t \\ &= \mu S_t\, dt + \sigma S_t\, (dW_t^Q + \theta\, dt) \\ &= (\mu + \sigma\theta) S_t\, dt + \sigma S_t\, dW_t^Q \\ &= r S_t\, dt + \sigma S_t\, dW_t^Q \end{aligned}$$

Comparing this to (12.13) we see that the effect of changing to Q is that the growth rate μ has been replaced with the interest rate r.

Next let's see why, based on stochastic differentials, S_t/B_t is a Q-martingale. Since $B_t^{-1} = e^{-rt}$ we have $dB_t^{-1} = -rB_t^{-1}\, dt$ and so (remembering that $\theta = (r-\mu)/\sigma$)

$$\begin{aligned} d(S_t B_t^{-1}) &= S_t\, dB_t^{-1} + B_t^{-1}\, dS_t + dS_t\, dB_t^{-1} \\ &= -rS_t B_t^{-1}\, dt + B_t^{-1}\, (\mu S_t\, dt + \sigma S_t\, dW_t) + 0 \\ &= S_t B_t^{-1} \sigma(-\theta\, dt + d\, W_t) \\ &= \sigma S_t B_t^{-1}\, dW_t^Q. \end{aligned}$$

So once we convert to the Q-Brownian motion W^Q there are no dt terms, so this should be a Q-martingale.

The Black-Scholes Equation

Next notice that (12.15) says that the pricing function $v(s,t)$ should be such that $v(S_t,t)e^{-rt}$ is a Q-martingale. We can use the Q-stochastic differential of S_t see what this means about the function $v(s,t)$. According to Itô's formula,

$$\begin{aligned} d[v(S_t,t)e^{-rt}] &= v_s(S_t,t)e^{-rt}\, dS_t + \frac{1}{2} v_{ss}(S_t,t)e^{-rt}\, (dS_t)^2 + v_t(S_t,t)e^{-rt}\, dt - rv(S_t,t)e^{-rt}\, dt \\ &= e^{-rt}\left[v_s(S_t,t)\, (rS_t\, dt + \sigma S_t\, dW_t^Q) + \frac{1}{2} v_{ss}(S_t,t)\sigma^2 S_t^2\, dt + v_t(S_t,t)\, dt - rv(S_t,t)\, dt \right] \\ &= e^{-rt}\left[\frac{1}{2}\sigma^2 S_t^2 v_{ss}(S_t,t) + rS_t v_s(S_t,t) - rv(S_t,t) + v_t(S_t,t) \right]\, dt + e^{-rt}\sigma S_t\, dW_t^Q \end{aligned}$$

For $v(S_t,t)e^{-rt}$ to be a Q-martingale we need the $[\cdots]\,dt$ term to vanish, i.e. for $v(s,t)$ to satisfy the partial differential equation

$$\frac{1}{2}\sigma^2 s^2 v_{ss} + rsv_s - rv + v_t = 0, \text{ for } t < T,\ 0 < s. \tag{12.17}$$

This is called the *Black-Scholes equation*. If we are given an exercise value function $\phi(s)$ then we can find the associated option's pricing function by solving (12.17) with the terminal data $v(s,T) = \phi(s)$. Then the option's market price at time t will be $v(S_t,t)$. In (12.16) we worked out the solution for $\phi(s) = \max(s-K,0)$ in particular. For ϕ in general an explicit solution may not be possible. In those cases a common approach is to employ numerical methods for partial differential equations to compute $v(s,t)$ from (12.17).

Self-Financing Portfolios

A portfolio consists of a pair X_t, Y_t of $S_{0:t}$-determined processes, representing holdings of stock and bank shares at time t. The value of this portfolio is the stochastic process

$$V_t = X_t S_t + Y_t B_t.$$

In the discrete time setting all these processes were constant over the time intervals (t_{i-1}, t_i), making discontinuous changes at the t_i. Using the backward difference notation $\Delta F_{t_i} = F_{t_i} - F_{t_{i-1}}$ we had several equivalent ways to express the self-financing property, one of which was equation (10.11):

$$\Delta V_i = X_{t_{i-1}} \Delta S_{t_i} + Y_{t_{i-1}} \Delta B_{t_i}.$$

With the understanding that the processes X_t and Y_t are constant over $[t_{i-1}, t_i)$ we might view this as in an integrated form as

$$V_t - V_0 = \int_0^t X_s\, dS_s + Y_s\, dB_s.$$

In differential form it would say

$$dV_t = X_t\, dS_t + Y_t\, dB_t. \tag{12.18}$$

This is what we will take as our definition of self-financing in continuous time. Although X_t, Y_t are not assumed piece-wise constant, self-financing means that their differentials do not contribute to dV_t. Itô's formula tells us that in general dV_t has some additional terms. Self-financing means that those additional terms cancel each other out. Observe that the self-financing property makes V_t/B_t a martingale, because

$$\begin{aligned} d[V_t B_t^{-1}] &= V_t\, dB_t^{-1} + B_t^{-1}\, dV_t + 0 \\ &= -r(X_t S_t + Y_t B_t) B_t^{-1}\, dt + B_t^{-1}(X_t\, dS_t + Y_t\, dB_t) \\ &= X_t B_t^{-1}(-rS_t\, dt + dS_t) \\ &= X_t B_t^{-1} \sigma\, dW_t^Q \end{aligned}$$

Suppose that we have solved (12.17) using a particular exercise value $\phi(s)$ and want to construct a self-financing portfolio X_t, Y_t which replicates the option:

$$v(S_t,t) = X_t S_t + Y_t B_t.$$

From Itô's formula and (12.17) we can work out that

$$\begin{aligned} dv(S_t,t) &= v_s\, dS_t + \left[v_t + \frac{1}{2}\sigma^2 S_t^2 v_{ss}\right] dt \\ &= v_s\, dS_t + \frac{1}{rB_t}\left[v_t + \frac{1}{2}\sigma^2 S_t^2 v_{ss}\right] dB_t \\ &= X_t\, dS_t + Y_t\, dB_t, \end{aligned}$$

where

$$X_t = v_s$$
$$Y_t = \frac{1}{rB_t}\left[v_t + \frac{1}{2}\sigma^2 S_t^2 v_{ss}\right].$$

Using (12.17) we can rewrite

$$Y_t = \frac{1}{B_t}[v - S_t v_s],$$

so that

$$X_t S_t + Y_t B_t = v(S_t, t).$$

This confirms that the value of the self-financing portfolio X_t, Y_t is indeed $v(S_t, t)$ as desired. Once again we see that (12.17) is intimately connected with the self-financing property.

Technicalities

Back in Section 9.3 we considered the possibility of doubling strategies by which a gambler can insure an eventual profit, provided there are no limits on how many times he is able to play or how deep in debt he is able to go before his eventual big win. Mathematically the same type of thing is possible in the Black-Scholes model. A mathematically complete treatment of mathematical finance in continuous time needs to impose technical conditions which eliminate such strategies from consideration, and address other technical issues. We of course have not attempted to do that. But you should be aware that there are such details that we have not dealt with.

For Further Study

There are many references on this material at various levels of sophistication and generality, see for instance Øksendal [46], Karatzas and Shreve [34], Stroock and Varadhan [59], and Rogers and Williams [51]. For the applications to mathematical finance in particular see Mikosch [42], Shreve [56]&[55] and Musiela and M. Rutkowski [44].

Problems

Problem 12.1
Show, for any $0 < s < t$, $tW_{1/t} - sW_{1/s}$ is a normally distributed random variable with mean 0 and variance $t - s$. (You may find it helpful to use the fact that the sum of a pair of independent normal random variables is again normal. The mean of the sum is the sum of the means and the variance of the sum is the sum of the variances.)

. . . . . BrMo1

Problem 12.2
Show that $N_t = W_t^2 - t$ and (for any $\theta \in \mathbb{R}$) $M_t = e^{\theta W_t - \frac{1}{2}\theta^2 t}$ are both martingales.

. . . . . BrMo2

Problem 12.3
Compute the stochastic integral $\int_0^T t\, dW_t$. (For each n describe a partition $t_i^{(n)}$ of $[0, T]$. Let $\psi_t^{(n)} = t_{k-1}^{(n)}$ on $(t_{k-1}^{(n)}, t_k^{(n)}]$. Explain why $E[\int_0^T (\psi_t^{(n)} - t)^2\, dt] \to 0$. Use Problem the "discrete product rule"

$$\Delta(x_i y_i) = x_{i-1}\Delta y_i + y_{i-1}\Delta x_i + (\Delta x_i)(\Delta y_i).$$

to rewrite $\int_0^T \psi_t^{(n)}\, dW_t$ and take the limit to determine $\int_0^T t\, dW_t$.)

. . . . . tWt

Problem 12.4

a) Verify the following formula, for each integer $n \geq 1$:

$$\int_0^t s^n \, dW_s = t^n W_t - \int_0^t n s^{n-1} W_s \, ds$$

b) Find a similar formula for $\int_0^t W_s^n \, dW_s$.

. . . . . Wpowers

Problem 12.5
Write out the induction argument to verify (12.11).

. . . . . Wn

Problem 12.6
For ordinary integration the n-fold iterated integrals of $f(t) = 1$ are the familiar monomials from Taylor series:

$$\int_0^t \int_0^{s_n} \int_0^{s_{n-1}} \cdots \int_0^{s_2} 1 \, ds_1 \cdots ds_{n-1} \, ds_n = t^n/n! \, .$$

The analogous formula for stochastic integrals is different. Verify the following formulas.

$$\int_0^t \int_0^{s_2} 1 \, dW_{s_1} dW_{s_2} = \frac{1}{2}(W_t^2 - t)$$

$$\int_0^t \int_0^{s_3} \int_0^{s_2} 1 \, dW_{s_1} dW_{s_2} dW_{s_3} = \frac{1}{6}(W_t^3 - 3tW_t)$$

$$\int_0^t \int_0^{s_4} \int_0^{s_3} \int_0^{s_2} 1 \, dW_{s_1} dW_{s_2} dW_{s_3} dW_{s_4} = \frac{1}{24}(W_t^4 - 6tW_t^2 + 3t^2)$$

. . . . . ItIntg

Problem 12.7
Use Problem 12.6 to compute $E[W_t^4]$.

. . . . . EW4

Problem 12.8
Use Itô's formula to verify that $e^{\frac{1}{2}\theta^2 t} \sin(\theta W_t)$ is a martingale for any $\theta \in \mathbb{R}$.

. . . . . ExpMg

Problem 12.9

a) Show that if $\frac{\partial}{\partial t} f(t,x) + \frac{1}{2} \frac{\partial^2}{\partial x^2} f(t,x) = 0$ then

$$df(t, W_t) = \frac{\partial}{\partial x} f(t, W_t) \, dW_t.$$

Thus if f satifies the PDE above, then $f(t, W_t)$ should be a martingale (subject to integrability conditions which we have been neglecting).

b) For our discounted stock price process, $M_t = S_t/B_t$ (considered with respect to the "risk neutral" probability Q), we know that $dM_t = \sigma M_t\, dW_t^Q$. Similar to a), find a partial differential equation for a function $w(t,z)$ which would imply that $w(t,M_t)$ is a martingale with respect to Q (again assuming the appropriate integrability condition can be verified).

c) Suppose we form a portfolio $\phi_t = f(t,M_t)$ and $\psi_t = g(t,M_t)$ using a pair of functions $ft,z)$, $g(t,z)$. For (ϕ_t,ψ_t) to be self-financing would require that $U_t = \phi_t M_t + \psi_t$ satisfy $dE_t = \phi_t\, dM_t$. Find an equation or equations that f and g would need to satisfy in order for (ϕ_t,ψ_t) to be self-financing.

. . . . . HeatEqn

Problem 12.10
Based on the stock price process (12.14) show that

$$E[S_t] = S_0 e^{(\mu+\sigma^2/2)t}.$$

Find a density for S_t, a function $p(s)$ so that

$$P(S_t \le c) = \int_{-\infty}^{c} p(s)\, ds.$$

Using $S_0 = 1$, $\mu = 2$, and $\sigma = 1/2$ produce a plot of $p(s)$ for $-1 \le s \le 10$.

. . . . . BSmean

Problem 12.11
Using the Black-Scholes formula, calculate the value of a call option at time $t = 0$ assuming $\sigma = .05$, $r = .03$, $K = 10$, $T = 10$, $S_0 = 8$.

. . . . . BSFeval

Problem 12.12
In this problem you are asked to establish some properties of the Black-Scholes formula (12.16). Remember also that v^{call} is given by an expectation:

$$v^{\text{call}}(s,t) = e^{-r(T-t)} E^Q[(S_T - K)^+ | S_t = s].$$

In each part below one or another of these may be more convenient. Also remember that v^{call} depends on the parameter values K, r, σ.

a) Show that $v^{\text{call}}(s,t) \ge (s-K)^+$. An implication of this is that there is never an advantage to exercising a call option early. (Hint: $(s-K)^+ \ge s-K$.)

b) Show that $v^{\text{call}}(s,t)$ is decreasing in K.

c) Show that $\lim_{t\to-\infty} v^{\text{call}}(s,t) = s$

d) What is the limit of $v^{\text{call}}(s,t)$ as the volatility increases without bound: i.e. $\sigma \to \infty$?

. . . . . BSFprop

Problem 12.13
A digital contract with strike price k is one whose final value is

$$V_T = \begin{cases} 1 & \text{if } S_T \ge k \\ 0 & \text{if } S_T < k \end{cases}.$$

Show that for our familiar Black-Scholes market model (the box on pg. 83) the market price of this contract is given by the formula

$$V_t = e^{-r(T-t)} \mathcal{N}\left(\frac{\log(S_t/k) + (r - \frac{1}{2}\sigma^2)(T-t)}{\sigma\sqrt{T-t}} \right).$$

. . . . . W

Appendix A: Random Variables

Chapter 3 gave a brief introduction to random variables and the basic ideas of the Kolmogorov model of probability theory. Here we collect some supplemental information.

A.1 Common Distributions

The *distribution* of a random variable refers to the collection of probabilities for its different possible outcomes. We often describe the type of a random variable by identifying its distribution. Here are the common distributions for discrete random variables which come up in our discussions. (The parameters below are $0 < p < 1$, $0 < \lambda$, $n \in \mathbb{N}$.)

- Bernoulli with parameter p: $P(X = 1) = p$, $P(X = 0) = 1 - p$.
- Uniform on $\{a_1, \ldots, a_n\}$: $P(X = a_i) = \frac{1}{n}$ (the same for all $i = 1, \ldots, n$).
- Binomial with parameters (n, p): $P(X = i) = \binom{n}{i} p^i (1-p)^{n-i}$ for $i = 0, \ldots n$.
- Geometric with parameter p: $P(X = n) = p(1-p)^n$, $n = 0, 1, \ldots$.
- Poisson with parameter λ: $P(X = n) = \frac{\lambda^n}{n!} e^{-\lambda}$, $n = 0, 1, \ldots$.

Here are the densities for the continuous distributions which we will encounter. (The parameters are $\alpha < \beta$; $\lambda > 0$; $\sigma, \mu \in \mathbb{R}$.)

- Uniform on $[\alpha, \beta]$: $f(x) = \frac{1}{\beta - \alpha}$ for $\alpha \leq x \leq \beta$ (and $f(x) = 0$ otherwise).
- Exponential with parameter λ: $f(x) = \lambda e^{-\lambda x}$ for $x \geq 0$ (and $f(x) = 0$ for $x < 0$)
- Normal with parameters (μ, σ^2): $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$.

A.2 Distribution Functions

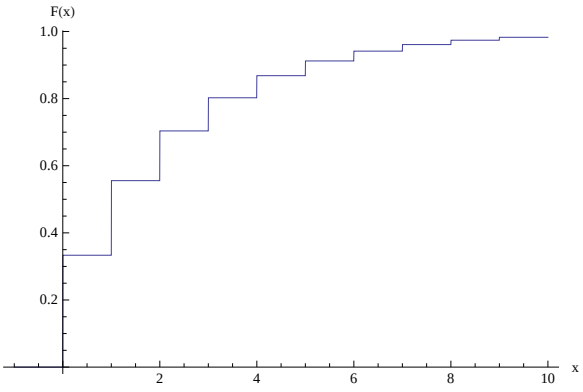
There exist distributions that are neither discrete nor continuous. A general approach to describing the distribution of a random variable X is based on its *distribution function*:

$$F_X(y) = P(X \leq y).$$

When X is discrete with $P(X = a_i) = p_i$ the distribution function is constant on the intervals between the a_i and with discontinuities of size p_i at a_i. Assuming the a_i are labeled in order, $a_i < a_{i+1}$,

$$F_X(y) = \sum_{a_i \leq y} p_i.$$

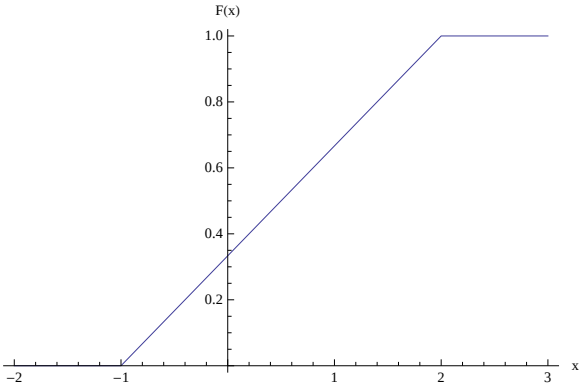
Example A.2. For a geometric random variable with parameter $p = 1/3$ the distribution function looks like this (except that software has not rendered the discontinuities properly).



When X is continuous the distribution function can be expressed in terms of the integral of the density $f(x)$

$$F_X(y) = \int_{-\infty}^{y} f(x)\,dx.$$

Example A.3. The distribution function for a uniform random variable on $[-1, 2]$ is this.



A distribution function always has the following properties.

Proposition A.4. *The distribution function $F(\cdot) = F_X(\cdot)$ of a random variable X has the following properties.*

a) $F(\cdot)$ *is nondecreasing: if* $a \le b$ *then* $F(a) \le F(b)$.

b) $F(\cdot)$ *is right continuous, i.e.* $F(c) = \lim_{y \to c^+} F(y)$ *for any* c.

c) $\lim_{y \to c^-} F(y)$ *(denoted* $F(c-)$*) always exists and is* $\le F(c)$.

d) $\lim_{y \to -\infty} F(y) = 0$ *and* $\lim_{y \to +\infty} F(y) = 1$.

e) $P(X < c) = F(c-)$.

f) $P(X = c) = F(c) - F(c-)$

g) $P(X > c) = 1 - F(c)$.

h) $P(a < X \le b) = F(b) - F(a)$, *for* $a < b$.

i) $P(a \le X \le b) = F(b) - F(a-)$, *for* $a < b$.

j) $P(a < X < b) = F(b-) - F(a)$.

These properties can be proven using the properties of probabilities in Section 3.1. As an example let's look at b): $F_X(c) = \lim_{y \to c^+} F_X(y)$. Consider any decreasing sequence $y_1 > y_2 > \cdots \to c$. We want to show that $\lim_n F_X(y_n) = F_X(c)$. The values of the distribution function are given by $F_X(y_n) = P(A_n)$ where A_n are the events

$$A_n = \{\omega \in \Omega \,:\, X(\omega) \leq y_n\}.$$

Because the y_n are decreasing this is a diminishing sequence of sets: $A_1 \supseteq A_2 \supseteq \cdots$ and so by the fifth bullet on page 33

$$\lim P(A_n) = P(A),$$

where (since $y_n \downarrow c$)

$$A = \cap A_n = \{\omega \in \Omega \,:\, X(\omega) \leq c\}.$$

Since $P(A) = F_X(c)$, this proves that $F_X(y_n) \to F_X(c)$. The same thing does *not* work if $y_1 < y_2 < \cdots \to c$ because in that case $A_1 \subseteq A_2 \subseteq \cdots$ but

$$A = \cup A_n = \{\omega \in \Omega \,:\, X(\omega) < c\} \neq \{\omega \in \Omega \,:\, X(\omega) \leq c\}.$$

We won't pursue such proofs any further. Our point is simply that rigorous proofs can be given based on the mathematical properties of the Kolmogorov model.

Distribution functions provide a unified approach to random variables of *all* types: discrete, continuous or neither. One important practical use for them is in the inverse method for computer simulation in Section A.3.2 below.

If $X \geq 0$ with probability 1, then its expected value can be calculated from its distribution function using the formula

$$E[X] = \int_0^\infty 1 - F(x-)\, dx.$$

This is the general version of equation (3.8) and Problem 3.4.

A.3 Random Number Generation

We have used experiments in which we examined a collection of computer-generated samples of a random variable to illustrate it's properties. In this section we want to discuss how we can produce such a set of sample values of a random variable X with a prescribed distribution. We will be interested specifically in how to do this in MATLAB.

A.3.1 Random and Pseudo-Random Numbers

For certain discrete random variables (with only a finite number of possible outcomes) we can build a physical device that behaves in the desired way: a dice, a roulette wheel or "spinner" with colored regions of specified relative sizes. There are a number of physical processes that are fundamentally random which can be used to construct such devices. One is the decay of radioactive substances. Such substances decay by emitting sub-atomic particles intermittently over a period of time. The times between particle emissions are exponentially distributed random variables. (This has to do with the quantum-mechanical description of what goes on inside the atoms of such unstable substances.) Think of the "clicks" produced by a Geiger counter; the spacings between the clicks are i.i.d. exponential random variables. There is an internet random number service that works along these lines: `http://www.fourmilab.ch/hotbits/` . They will supply sets of random numbers generated by the decay of a sample of Cæsium-137. Atmospheric noise in radio transmissions is another source. At `Random.org` you can download files of random numbers produced in that way.

The most famous "physical" random number generator is probably the one constructed by the Rand Corporation in the late 1940s, and used to produce the book A MILLION RANDOM DIGITS WITH 100,000 NORMAL DEVIATES [50]. They built a machine which used some known noisy electronic phenomena to produce random digits. The book's introduction describes the machine as follows. "In principle the machine was a 32-place roulette wheel which made, on the average, about 3000 revolutions per trial and produced

one number per second. A binary-to-decimal converter was used which converted 20 of the 32 numbers (the other twelve were discarded) and retained only the final digit of two-digit numbers; this final digit was fed into an IBM punch to produce finally a punched card table of random digits." The book is available on-line: [50]. Their list of a million random digits (i.e. numbers 0, 1, ..., 9, all with equal probability of 1/10) can be downloaded from there as a text file if you wish. For instance, here are the first 500 random digits of their table. (The first column is just an index.)

00000	10097	32533	76520	13586	34673	54876	80959	09117	39292	74945
00001	37542	04805	64894	74296	24805	24037	20636	10402	00822	91665
00002	08422	68953	19645	09303	23209	02560	15953	34764	35080	33606
00003	99019	02529	09376	70715	38311	31165	88676	74397	04436	27659
00004	12807	99970	80157	36147	64032	36653	98951	16877	12171	76833
00005	66065	74717	34072	76850	36697	36170	65813	39885	11199	29170
00006	31060	10805	45571	82406	35303	42614	86799	07439	23403	09732
00007	85269	77602	02051	65692	68665	74818	73053	85247	18623	88579
00008	63573	32135	05325	47048	90553	57548	28468	28709	83491	25624
00009	73796	45753	03529	64778	35808	34282	60935	20344	35273	88435

There are several physical random number generators in use today. For instance a Swiss firm (IDQ) sells a random number generator using quantum phenomena that will plug into the USB port of a computer. But physical random number generators are slow compared to the processing speed of modern computers. To get around this many approaches have been developed to get the computer itself to produce random numbers, so they will be available at a much higher rate than physical generators like those described above. But there is nothing random about a contemporary computer — it is a purely deterministic device. So random number generating software doesn't produce true random numbers but rather *pseudorandom* numbers. These are numbers which come from a deterministic algorithm but which have statistical properties which make them a suitable substitute for many purposes.

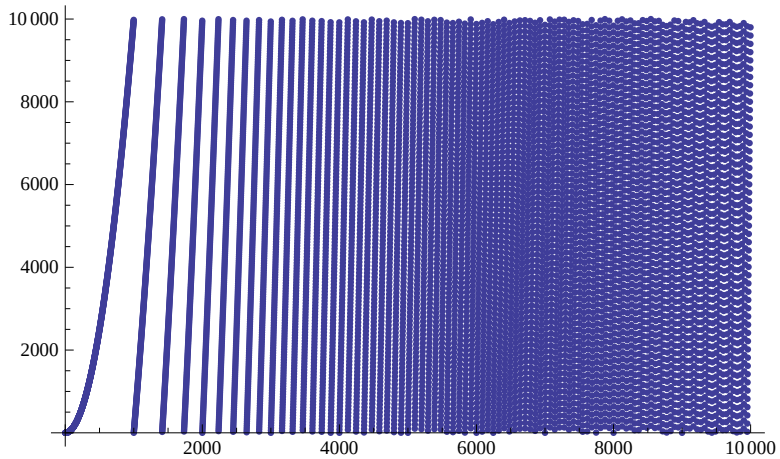
Pseudo-random number generators typically use some deterministic function $\Phi(x)$ for which the values $y = \Phi(x)$ vary in a highly sensitive and irregular way as a function of the input x. We select a starting value x_0, called the *seed*, and iterate:

$$x_1 = \Phi(x_0),\ x_2 = \Phi(x_1),\ \ldots,\ x_{n+1} = \Phi(x_n), \ldots$$

These values $x_1, x_2, \ldots, x_n, \ldots$ are taken as the i.i.d. samples. One of the first such methods to be proposed was J. von Neumann's middle square method. If we want random integers from 0000 to 9999 (i.e. four decimal digits), here is how Φ for the middle square method would work. Starting with a four digit number, 4234 for instance, we would square it and then extract the four digits from the middle of the result. To illustrate,

$$4234^2 = 17926756$$

So $\Phi(4234) = 9267$. To see how irregular the dependence of $\Phi(x)$ on x is we can look at its graph.



But look what happens if work out a long section of the sequence starting with our seed, $x_0 = 4234$:

$$\ldots x_{88} = 8100,\ x_{89} = 6100,\ x_{90} = 2100,\ x_{91} = 4100,\ x_{92} = 8100,\ \ldots$$

We see that the numbers start repeating. So after a while they are hardly random-like at all.

All pseudo-random number generators will eventually repeat, but good ones take an insanely long time before that starts. Although the middle square method works somewhat better if implemented for a larger number of digits than four, it still turns out to be rather poor. Many more effective methods have been developed. Since the 1950's most programming languages (at least those used for mathematical calculation) have included some pseudo-random number generating algorithm.

The selection of a good pseudo-random number generator takes care. D. Knuth says in [36],"The moral to this story is that random numbers should not be generated by a method chosen at random. Some theory should be used." It is not our purpose to pursue that theory further. We just want to use a pseudo-random number generator to simulate random variables and stochastic processes. Below we will talk about MATLAB commands to do this. But we will trust that the developers of MATLAB have considered the issues and made good choices for the algorithms they built into their software. However, we should keep in mind that we are using pseudo-random numbers, instead of true random numbers, and that could conceivably introduce some unexpected feature into our calculations.

If you are interested in the algorithms used for pseudo-random number generation, your software's documentation will probably give you some references that the developers used. A couple helpful but older references are Anderson [2] and Knuth [36].

A.3.2 Conversion of Uniform to Other Distributions

Virtually all true or pseudo-random number generators produce a sequence of samples of a discrete random variable Y taking values in $\{0, 2, \ldots, n-1\}$ with equal probabilities, for some n that depends on the choice of generator. Most software will also produce pseudo-random samples of the uniform distribution on $[0, 1]$ or a standard normal distribution for us. But if we want something other than uniform or standard normal distribution and the software does not have a built in command for that distribution we may have to do a conversion ourselves. By this we mean to start with Y of one distribution and apply some function to it to get a random variable $X = \phi(Y)$ with a different distribution. For instance if Y is standard normal (from `randn`) then $X = \sigma Y + \mu$ will be normal with parameters (μ, σ^2). If U is uniform on $[0, 1]$ (from `rand`) then $X = cU + a$ will be uniform on $[a, a + c]$.

The *inversion method* is a general way to convert a uniform – $[0, 1]$ random variable to a random variable with a different distribution. Let F be the distribution function for the desired random variable X. We know $F : \mathbb{R} \to [0, 1]$ is right continuous and nondecreasing. We need a function $F^* : (0, 1) \to \mathbb{R}$ with the property that

$$F^*(u) \le a \text{ if and only if } u \le F(a). \tag{A.19}$$

With such a function consider $X = F^*(U)$, where U is uniform on $[0, 1]$. (Notice that we have not required $F^*(0)$ or $F^*(1)$ to be defined. That is not a problem because $P(U = 0 \text{ or } 1) = 0$ so we will not encounter $F^*(0)$ or $F^*(1)$. More on this shortly.) By virtue of (A.19) the distribution function of X is

$$P(X \le a) = P(F^*(U) \le a) = P(U \le F(a)) = F(a).$$

so X has the desired distribution F.

Such a function F^* does always exist; it is called the *generalized inverse* of F and is defined in general by

$$F^*(u) = \min\{x \,:\, u \le F(x)\}. \tag{A.20}$$

Because F is right continuous the set on the right does have a minimum for any $0 < u < 1$. Visually the definition means that we look at the part of the graph of F which is at or above the level u. Then $F^*(u)$ is the smallest (leftmost) x value for this portion of the graph. When there is a unique x with $F(x) = u$ then we just get $x = F^*(u)$ as with a typical inverse function. But suppose there is a gap $F(c-) < F(c)$ in the graph of F. (This means F is not onto.) If u falls in the gap, $F(c-) \le u \le F(c)$, then we find that $F^*(u) = c$. This produces a flat section in the graph of F^*. Suppose u corresponds to a flat section

in the graph of F: $u = F(x)$ for all $x \in [a, b)$ or $x \in [a, b]$. (This means F is not one-to-one.) Then we get $F^*(u) = a$, the leftmost point of the flat section. But for $F(a) = F(b-) < u$ we get $b \le F^*(u)$. This produces a gap or discontinuity in the graph of F^*.

We see (and can prove) that F^* is nondecreasing but is *left*-continuous. In some cases (A.20) would produce $F^*(0) = -\infty$ or $F^*(1) = +\infty$. That is why we have not required $F^*(0)$ or $F^*(1)$ to be defined. As pointed out, because $P(U = 0) = 0 = P(U = 1)$ we have no practical need for those values.

If U_n are i.i.d. uniform on $[0, 1]$ then $X_n = F^*(U_n)$ will be i.i.d. with distribution F. So to generate a set of independent samples X_n we use `rand` to produce i.i.d. uniform–$[0, 1]$ samples and then take F^* of them to obtain i.i.d. samples of X. This is illustrated in Examples A.4 and A.5 below.

A.4 Matlab

Let's look at how we can use MATLAB to simulate random variables. Before using any of the commands described below we need to "seed" the random number generator. It is possible to specify a specific seed. You might want to do this if you will want to repeat your calculations with exactly the same sequence of randomly generated values. But we generally won't be doing that. The simplest thing to do is use the command `rng('shuffle')` just once at the beginning of your MATLAB session. That will "randomly" select a seed for you. You don't need to repeat it until the next time you start up MATLAB.

There are three basic random number producing commands: `randi` produces pseudo-random integers in a specified range $\{m, m+1, \dots n\}$, uniformly distributed; `rand` produces uniformly distributed pseudo-random numbers in $[0, 1]$; and `randn` produces pseudo-random real numbers with a standard normal distribution. (There is also `randperm` to produce random permutations, but we won't be using that.) In addition MATLAB's Statistics Toolbox[1] includes commands for simulating many other common distributions. For instance there are `binornd` for binomial, `geornd` for geometric, `poissrnd` for Poisson, `exprnd` for exponential as well as many others. You can consult MATLAB's help to learn details about the syntax for these. Note that if you want an array of pseudo-random values you don't need to use `for`-loops to fill up an array; these commands will allow you to specify the dimensions for an array of results.

We will need to be able write our own pseudo-random number generating m-files to simulate various types of stochastic processes in later chapters. To gain some experience with that the next examples illustrate how we can use the F^*-technique to write MATLAB m-files various distributions.

Example A.4. First consider an exponentially distributed random variable X with parameter $\lambda > 0$. For $0 < x$ the distribution function is

$$F(x) = \int_0^x \lambda e^{-\lambda t}\, dt = 1 - e^{-\lambda x}.$$

Solving $u = F(x)$ for $0 < u < 1$ we find the generalized inverse to be

$$F^*(u) = \frac{-\ln(1-u)}{\lambda}.$$

So our m-file will use `rand` produce a value of a uniform random variable U and then use

$$X = \frac{-\ln(1-U)}{\lambda}.$$

That's pretty simple. We just need to write the m-file to accept a parameter `lambda` to specify λ, and an optional second parameter `n` which specifies how many such values to produce.

[1] To see if you have the Statistics Toolbox, just enter the command `ver` and you will get a listing which includes the toolboxes that you have. For a list of all the commands provided by the toolbox, see `http://www.mathworks.com/help/stats/functionlist.html` .

randexpn.m

```
function x = randexpn(lambda,n)
%randexpn(lambda,n) produces a 1xn array of exponentially distributed
%pseudo-random variables, with parameter lambda.
%
if nargin==1
n=1;
end;
x=-log(1-rand([1,n]))/lambda;
end
% M. Day, August 7, 2014
```

Example A.5. Next let's write an m-file to produce samples of a random variable X with values in $\{1, 2, 3, \ldots n\}$ with a specified set of probabilities $p_i = P(X = i)$. (MATLAB does not provide a command to do this!) The distribution function is piecewise constant,

$$F(x) = \sum_{1}^{k} p_i \text{ for } k \leq x < k + 1.$$

For F^* we need the values

$$c_0 = 0,\ c_1 = p_1,\ c_2 = p_1 + p_2,\ c_3 = p_1 + p_2 + p_3, \ldots, c_n = 1.$$

It turns out that

$$F^*(u) = k \text{ for } c_{k-1} < x \leq c_k.$$

In other words, $F^*(u)$ is the smallest $k \geq 1$ with $u \leq c_k$. Here is an implementation of this as an m-file.

randd.m

```
function x = randd(pmf,n)
%randd(pmf,n) produces a 1xn array of pseudo-random integers with
%probabilities specified in the vector pmf.  If n is absent a single
%such vale is produced.
%
if nargin==1
n=1;
end
c=cumsum(pmf);
pick=@(u) find(u<=c,1); %Anonymous function for first i with u<=c(i).
x=arrayfun(pick,rand([1,n])); %Apply to each entry of i.i.d. unform array.
end
% M. Day, August 7, 2014
```

A.4.1 List of Relevant Commands

To help you out with MATLAB we want to provide a short list of commands you might use for simulating random variables. For more details of their syntax consult the documentation[2].

`rng('shuffle')` seeds the random number generator. Use it at the start of each session.

The following commands each produce an array with dimensions specified by `size` of pseudo-random numbers from the specified distribution. Specifying `[m,n]` for `size` produces an $m \times n$ array of values. If `size` is a single value n an $n \times n$ array will be produced. If `size` is omitted a single such value is produced.

[2] MATLAB's Help Menu, or `http://www.mathworks.com/help/matlab/index.html`

`rand(size)` for the uniform distribution on $(0, 1)$.
`randn(size)` for the standard normal distribution.
`randi(k,size)`, `randi([m,n],size)` for the uniform distribution on $\{1, \ldots, k\}$ or $\{m, n\}$.
`binornd(n,p,size)` for the binomial distribution with parameters (n, p) (*Statistics Toolbox*).
`geornd(p,size)` for the geometric distribution (*Statistics Toolbox*).
`poissrnd(lambda,size)` for the Poisson distribution with parameter $\lambda =$`lambda` (*Statistics Toolbox*).
`exprnd(lambda,size)` for the exponential distribution with parameter $\lambda =$`lambda` (*Statistics Toolbox*).
`nchoose(n,m)` calculates the binomial coefficient $\binom{n}{m} = \frac{n!}{(n-m)!m!}$.

If you have two lists (vectors) of data `X` and `Y` of the same size, the following may be useful to explore them.

`mean(X)`, `var(X)` produce the sample mean and variance (respectively) of the data in `X`. (Note that `var` uses "$n - 1$ normalization.")
`histogram(X,m)` produces a "histogram" plot showing the number of terms from `X` separated into `m` equal sized "bins." There are many options for specifying the bins other than just their total number.
`histcounts(X,m)` will give you the numbers of terms falling the the respective bins.
`scatter(X,Y)` or `plot(X,Y,'.')` produces a plot of the (`X(i),Y(i)`) pairs. (Don't leave `'.'` out of `plot` or it will connect the dots!)
`plot(X,'.')` plots the points (`i,X(i)`).

`save('x.mat','X','-ASCII')` will save the contents of `X` in an ASCII file named `x.mat` in the default directory. You can use `load 'x.mat'` to read that data back in on a future occasion.
Another way to import data from a file is to select `Import Data ...` from the file menu to start the Import Wizard.

Appendix B: Mathematical Supplements

This appendix provides summaries of mathematical topics you may not have encountered in your previous courses.

A.1 Convex Functions and Jensen's Inequality

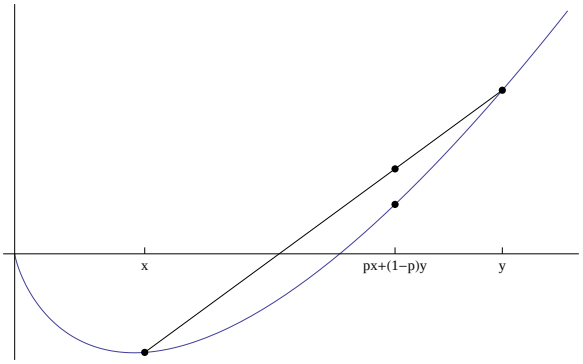
A convex function is essentially what is called "concave up" in a freshman calculus class.

Definition. *Suppose $I \subseteq \mathbb{R}$ is an interval. A function $f : I \to \mathbb{R}$ is called* convex *on I if for any $x, y \in I$ and any $0 \leq p \leq 1$*

$$f(px + (1-p)y) \leq pf(x) + (1-p)f(y).$$

If the above inequality is strict ($<$) whenever $x \neq y$ and $0 < p < 1$ we say f is strictly *convex.*

You should view $px + (1-p)y$ as a (weighted) average of x and y. The definition says that if you first average two points and then evaluate the function you should get no more than if you first evaluated the function at the two points and then averaged the function values: "f of average $\leq$ average of f." The following picture illustrates this. The point on the straight line corresponds to $pf(x) + (1-p)f(y)$.



Convex functions do not need to be differentiable. The definition above is phrased in a way that avoids any reference to derivatives. For instance the absolute value function $f(x) = |x|$ is convex. (They do have to be continuous, however.) For functions that are twice differentiable $f'' \geq 0$ is a sufficient condition for f to be convex.

Lemma A.5. *Suppose $f : I \to \mathbb{R}$ where I is an interval. If f is continuous on I and twice differentiable with $f''(x) \geq 0$ at all interior points of I, then f is convex on I. If $f''(x) > 0$ at all interior points then f is strictly convex.*

Proof. Suppose f is as hypothesized in the lemma, and $x, y \in I$ and $0 \leq p \leq 1$. We need to verify the inequality of the definition of convex function. If $x = y$ or $p = 0$ or $p = 1$ then the inequality is trivial. So

suppose $x < y$ and $0 < p < 1$. Let $z = px + (1-p)y$. Then $x < z < y$. (Because I is an interval z must be in I as well.) By the usual Mean Value Theorem there exist points a and b with $x < a < z < b < y$ for which

$$\frac{f(z) - f(x)}{z - x} = f'(a), \quad \frac{f(y) - f(z)}{y - z} = f'(b).$$

Because $f'' \geq 0$ on $[a, b]$ we know that $f'(a) \leq f'(b)$ and therefore

$$\frac{f(z) - f(x)}{z - x} \leq \frac{f(y) - f(z)}{y - z}.$$

Our choice of z implies that $z - x = (1-p)(y-x)$ and $y - z = p(y-x)$. Making those replacements and multiplying both sides by $p(1-p)(y-z)$ we find that the above inequality reduces to $f(z) \leq pf(x) + (1-p)f(y)$. □

As an example, consider $f(x) = -\sqrt{x}$ on $I = [0, \infty)$. This function is continuous but not differentiable at the left endpoint $x = 0$. But for all interior points ($x > 0$) we have $f''(x) = \frac{1}{4}x^{-3/2} > 0$, so by the lemma this is a convex function. Using $p = 1/2$ in particular leads (after rearrangement) to the inequality

$$\sqrt{x} + \sqrt{y} \leq \sqrt{2}\sqrt{x+y} \text{ for any } x, y \geq 0.$$

Many useful inequalities can be derived this way using different convex functions. We will see several applications in Chapter 7. Jensen's Inequality says that the inequality in the definition must also hold for averages of any finite number of points.

Theorem A.6 (Jensen's Inequality)**.** *Suppose $f : I \to \mathbb{R}$ is a convex function on an interval I. For any finite set of points $x_i \in I$, $i = 1, \ldots, n$ and probabilities $p_i \geq 0$, $\sum_1^n p_i = 1$,*

$$f\left(\sum_1^n p_i x_i\right) \leq \sum_1^n p_i f(x_i).$$

If f is strictly convex then the above inequality is strict whenever those x_i for which $p_i > 0$ are not identical.

Proof. We use induction on $n \geq 2$. The definition is the case of $n = 2$. Suppose it holds for n (and any choice of x_i, p_i with $\sum_1^n p_i = 1$). Let y_i, $i = 1, \ldots n+1$ and $q_i \geq 0$ with $\sum_1^{n+1} q_i = 1$ be given. We can reduce this to the case of n terms as follows. Let $p_i = q_i$ $i = 1, \ldots n-1$ and $p_n = q_n + q_{n+1}$. Then take $x_i = y_i$ for $i = 1, \ldots n-1$ and $x_n = \frac{1}{p_n}(q_n y_n + q_{n+1} y_{n+1})$. Observe that

$$\sum_1^{n+1} q_i y_i = \sum_1^n p_i x_i.$$

The assumed validity for n points implies that

$$f\left(\sum_1^n p_i x_i\right) \leq \sum_1^n p_i f(x_i) = \sum_1^{n-1} q_i f(y_i) + p_n f\left(\frac{q_n}{p_n} y_n + \frac{q_{n+1}}{p_n} y_{n+1}\right)$$

But since $\frac{q_n}{p_n} + \frac{q_{n+1}}{p_n} = 1$ we can use convexity for the last term as well:

$$f\left(\frac{q_n}{p_n} y_n + \frac{q_{n+1}}{p_n} y_{n+1}\right) \leq \frac{q_n}{p_n} f(y_n) + \frac{q_{n+1}}{p_n} f(y_{n+1}).$$

Assembling the pieces, we have Jensen's Inequality for $n+1$ terms:

$$f\left(\sum_1^{n+1} q_i y_i\right) \leq \sum_1^{n+1} q_i f(y_i).$$

That completes the proof by induction. □

The definition of convex functions extends to functions $f : \mathbb{R}^m \to \mathbb{R}$. If f is convex then in fact Jensen's Inequality holds for any random variable X taking values in $\mathbb{R}^m$ provided both X and $f(X)$ are integrable:

$$f(E[X]) \leq E[f(X)].$$

A.2 Inf and Sup

For a finite (nonempty) set of numbers $A \subseteq \mathbb{R}$ we can always find a largest value, which we call the maximum, and likewise a smallest value with we call the minimum. For instance

$$A = \{2, 4, 6.6, 45\}$$

has $\min A = 2$ and $\max A = 45$. But when A is infinite there may or may not be a largest or smallest value. For instance the half-open interval

$$A = (0, 1]$$

has largest value $\max A = 1$ but there is no smallest value. We might say something like "A has a lower limit of 0, although 0 does not actually belong to A". This idea of lower limit is what we call the *infimum* of a set, denoted "$\inf A$". It's sort of where the minimum ought to be if there were one. So we would say

$$\inf(0, 1] = 0.$$

When a set has a minimum then the minimum and infimum are the same, but sets can have an infimum even if they don't have a minimum. That's essentially because the infimum is not required to belong to the set. Any nonempty set which is bounded below *always* has an infimum. This very general existence property is what makes the concept useful.

The counterpart for largest element is the *supremum*, denoted "$\sup A$". When a set has a maximum the maximum and supremum agree, but $\sup A$ will always exist for any nonempty set which is bounded above. $\liminf$

A.3 Order in Infinite Series

We say that the infinite series $\sum_{n=0}^{\infty} a_n$ of real numbers a_n converges to a value A when the sequence of partial sums converges

$$\lim_{k\to\infty} \sum_{n=0}^{k} a_n = A.$$

In particular the limit A must be a finite value for the series to be called convergent. (If $\lim_{k\to\infty} \sum_{n=0}^{k} a_n = +\infty$ we might write "$\sum_{n=0}^{\infty} a_n = +\infty$" but we would not call this a convergent series.)

The convergence of a series can be influenced by cancellation between positive and negative values among the a_n. The usual example is the harmonic series $\sum_1^\infty \frac{1}{n}$ which is divergent, and the alternating harmonic series $\sum_1^\infty \frac{(-1)^{n-1}}{n}$ which is convergent. When the series converges after removing any negative signs, i.e. when

$$\sum_0^\infty |a_n| \text{ converges,}$$

we call the series *absolutely convergent*. An absolutely convergent series is always convergent, but it is possible to be *conditionally convergent* which means that the series converges but not absolutely. In other words its convergence depends on some cancellation between the positive and negative terms. The alternating harmonic series is an example. One reason absolute convergence is important has to do with changing the order in which the terms are summed. Suppose $\sum_0^\infty b_n$ uses the same terms as $\sum_0^\infty a_n$ but in a different order, i.e. the a_n are permuted to obtain the b_n. If $\sum_0^\infty a_n$ is absolutely convergent, then

$$\sum_0^\infty a_n = \sum_0^\infty b_n.$$

(Norris [45] gives a version of this in his Lemma 6.1.1, but only for series with nonnegative terms. Convergence is the same as absolute convergence in that case. For nonnegative terms the equality of the two series also holds if they both diverge, i.e. "$= \infty$".) But this is *not true* for conditionally convergent series. In fact there is a famous theorem which says that if $\sum_0^\infty a_n$ is conditionally convergent, then for any B, finite or not, it is

possible to find a rearrangement with $\sum_0^\infty b_n = B$. I.e. you can reorder a conditionally convergent series to make it converge to anything you want, or even to diverge. Order of summation matters for conditionally convergent series! (See Rudin [53].)

With that in mind we might consider this approach to analyzing a given series $\sum_0^\infty a_n$. First separate the terms into the positive ones and negative ones: let

$$a_n^+ = \begin{cases} a_n & \text{if } a_n \geq 0 \\ 0 & \text{if } a_n < 0, \end{cases} \quad \text{and} \quad a_n^- = \begin{cases} 0 & \text{if } a_n \geq 0 \\ -a_n & \text{if } a_n < 0. \end{cases}$$

This creates two series, $\sum_0^\infty a_n^-$ and $\sum_0^\infty a_n^+$ both of which have only nonnegative terms. Moreover $a_n = a_n^+ - a_n^-$. So we can try to understand $\sum_0^\infty a_n$ by writing it as

$$\sum_0^\infty a_n = \left(\sum_0^\infty a_n^+\right) - \left(\sum_0^\infty a_n^-\right).$$

You could view this as rearranging the original series into two series, one with all the nonnegative terms and one with all the nonpositive terms (and lots of extra zeros thrown in). For both of the separated series $\sum_0^\infty a_n^\pm$ to converge is equivalent to saying $\sum_0^\infty a_n$ is absolutely convergent. (That's because $\sum |a_n| = \sum a_n^+ + \sum a_n^-$.) But when $\sum_0^\infty a_n$ is conditionally convergent, the left side of the above equation is finite but both terms on the right are $+\infty$ so that their difference is undefined. Our point is that only for absolutely convergent series can you sum the positive and negative terms separately!

When X is a discrete random variable but $\sum a_i P(X = a_i)$ is conditionally convergent we will have trouble defining $E[X] = \sum a_i P(X = a_i)$, because the result will depend on the order of summation. Only when $\sum |a_i| P(X = a_i) < \infty$, which is to say $E[|X|] < \infty$, can we define $E[X]$ without worrying about the order in which the sum is taken. When $E[|X|] = \infty$ we have something analogous to the conditionally convergent situation for infinite series, and we consider $E[X]$ to be undefined.

The issue of ordering the terms also comes up when we multiply two convergent series $A = \sum_{n=0}^\infty a_n$ and $B = \sum_{m=0}^\infty b_m$. We expect their product

$$AB = \left(\sum_{n=0}^\infty a_n\right)\left(\sum_{n=0}^\infty b_n\right)$$

to be a double series of all $a_n b_m$. But any effort to multiply the product out and gather the terms back together into a single series involves some shifting around of the order of summation. When both original series are absolutely convergent then again the order turns out not to matter; see [1] page 73. One particularly important ordering results from grouping the terms according to the value of $k = n + m$. This leads to

$$AB = \sum_{k=0}^\infty \left(\sum_{n=0}^k a_n b_{k-n}\right) = \sum_{k=0}^\infty c_k,$$

where c_k is the convolution sequence:

$$c_k = a_0 b_k + \cdots a_k b_0.$$

A.3.1 Interchanging Limits

We occasionally need to take limits of series, or series of limits. Suppose the terms $a_{n,m}$ depend on two parameters n. We want to sum over n and take the limit with respect to m. The basic problem is whether or not it is valid to say

$$\lim_{m\to\infty}\left(\sum_{n=1}^\infty a_{n,m}\right) = \sum_{n=1}^\infty \left(\lim_{m\to\infty} a_{n,m}\right).$$

These are sometimes equal, but not always.

Example A.6. Suppose $a_{n,m} = 1$ when $n = m$ and $= 0$ otherwise. Then $\sum_{n=1}^\infty a_{n,m} = 1$ for all m and therefore $\lim_{m\to\infty}(\sum_{n=1}^\infty a_{n,m}) = 1$. But $\lim_m a_{n,m} = 0$ so $\sum_{n=1}^\infty (\lim_m a_{n,m}) = 0$.

Here are the two best-known results which insure equality. (Neither of them apply to the preceding example!) These are both cousins to the results of the same names in Section 3.2.1: Theorems 3.3 and 3.2. Here the setting is simple enough that we can write out proofs.

Theorem A.7 (Dominated Convergence for Series)**.** *Suppose there is an absolutely convergent series* $\sum_0^\infty c_n$ *so that for every* m

$$|a_{n,m}| \le c_n.$$

If $\lim_{m\to\infty} a_{n,m} = b_n$ *for each* n*, then*

$$\lim_{m\to\infty}\left(\sum_{n=0}^{\infty} a_{n,m}\right) = \sum_{n=0}^{\infty} b_n.$$

Proof. Since $\sum_{n=0}^\infty c_n$ converges and $|b_n| \le c_n$, the comparison test implies that $\sum_{n=0}^\infty b_n$ is convergent, and that $\sum_{n=0}^\infty a_{n,m}$ is convergent for every m. Let $B = \sum_{n=0}^\infty b_n$.

Consider any $\epsilon > 0$. There exists N so that

$$\sum_{n=N+1}^{\infty} c_n < \epsilon/3.$$

So we have

$$\begin{aligned}\left|\sum_{n=0}^{\infty} a_{n,m} - \sum_{n=0}^{\infty} b_n\right| &\le \sum_{n=0}^{N} |a_{n,m} - b_n| + \sum_{n=N+1}^{\infty} |a_{n,m}| + \sum_{n=N+1}^{\infty} |b_n| \\ &\le \sum_{n=0}^{N} |a_{n,m} - b_n| + 2\sum_{n=N+1}^{\infty} c_n \\ &\le 2\epsilon/3 + \sum_{n=0}^{N} |a_{n,m} - b_n|\end{aligned}$$

Now the limit of the right side as $m \to \infty$ is simply $2\epsilon/3$. It follows that for sufficiently large m that

$$\left|\sum_{n=0}^{\infty} a_{n,m} - \sum_{n=0}^{\infty} b_n\right| < \epsilon,$$

which proves the theorem. □

Theorem A.8 (Monotone Convergence for Series)**.** *Suppose* $0 \le a_{n,m}$ *and for each* n

$$a_{n,1} \le a_{n,2} \le \cdots a_{n,m} \le a_{n,m+1} \cdots \text{ with } \lim_m a_{n,m} = b_n.$$

Then

$$\lim_{m\to\infty}\left(\sum_{n=0}^{\infty} a_{n,m}\right) = \sum_{n=0}^{\infty} b_n$$

This holds even if $\sum_{n=0}^\infty b_n = \infty$.

Proof. Let $B = \sum_{n=0}^\infty b_n$ and consider any $\beta < B$. There exists N with

$$\beta < \sum_{n=0}^{N} b_n.$$

Now

$$\lim_m \sum_{n=0}^{\infty} a_{n,m} \ge \lim_m \sum_{n=0}^{N} a_{n,m} = \sum_{n=0}^{N} \lim_m a_{n,m} = \sum_{n=0}^{N} b_n > \beta.$$

Considering all possible $\beta < B$ this implies

$$\lim_m \sum_{n=0}^{\infty} a_{n,m} \geq B.$$

But since $a_{n,m} \leq b_n$ we also know

$$\lim_m \sum_{n=0}^{\infty} a_{n,m} \leq \lim_m \sum_{n=0}^{\infty} b_n = B.$$

□

A.4 About Greatest Common Divisors

The d greatest common divisor of a set $W \subseteq \mathbb{N}$ of positive integers is by definition a positive integer g which is a common divisor of W (i.e. g divides every $n \in W$) and which is divisible by any other common divisor. The standard argument for the existence of a $d = \gcd(W)$ is to consider all finite linear combinations of W with integer coefficients:

$$L = \left\{ k > 0 : k = \sum_{i=1}^{m} \alpha_i n_i \text{ for some } n_i \in W \text{ and } \alpha_i \in \mathbb{Z} \right\}.$$

As a set of positive integers L has a smallest element g which can then be proven to be the greatest common divisor of W. This also shows that $g = \gcd(W)$ can be written as a (finite) linear combination of integers from W:

$$g = \sum_{i=1}^{m} \alpha_i n_i \tag{A.21}$$

for some $n_1, \ldots, n_m \in W$ and $\alpha_i \in \mathbb{Z}$. Clearly all multiples of g belong to L. An additional fact which we need in Lemma 2.4 is that all *sufficiently large* multiples of g can be written as linear combinations of W with using *nonnegative* coefficients.

Lemma A.9. *Suppose $W \subseteq \mathbb{N}$ and $g = \gcd(W)$. There exists $K \geq 0$ so that for all $k \geq K$ it is possible to express kg as*

$$kg = \sum_{i=1}^{m} \alpha_i n_i$$

for some $n_i \in W$ and $\alpha_i \in \mathbb{N}$.

Proof. Starting with (A.21) divide the terms into two sets with a_i being the nonnegative coefficients and $-b_j$ being the negative ones (so $b_j > 0$):

$$g = \sum_i a_i n_i \ - \sum_j b_j n_j.$$

(If all the c_i in (A.21) are positive the lemma is trivial, so we can assume that there is at least one positive b_j.) Consider $M = \sum_j b_j n_j$. Since g is a divisor of all the n_j it is also a divisor of M: $M = \beta g$ for some integer $\beta \geq 1$. Let $K = \beta M$. For any $k \geq K$ we can write $k = mM + \ell$ with $m \geq M$ and $0 \leq \ell < M$. So

$$\begin{aligned} kg &= mgM + \ell g \\ &= mg \sum_j b_j n_j \ + \ell \left(\sum_i a_i n_i \ - \sum_j b_j n_j \right) \\ &= \sum_i (\ell a_i) n_i + \sum_j (mg - \ell) b_j n_j \end{aligned}$$

Since $k \geq K = \beta M$ it follows that $m \geq \beta$. Therefore

$$mg \geq \beta g = M > \ell.$$

Thus all the n_j coefficients in the above representation of kg are nonnegative. $\square$

Bibliography

[1] S. Abbot, Understanding Analysis, Springer-Verlag NY, 2001.

[2] S. L. Anderson, *Random number generators on vector supercomputers and other architectures*, SIAM Review, v.32 no. 2 (1990), pp. 221-251.

[3] L. Bachelier, Theorie de la speculation, *Ann. Sci. Ecole Norm. Sup.* 17 (1900), pp. 21–86. [English translation in **The Random Character of Stock Market Process**, P. H. Coonter (ed.), MIT Press, Cambridge MA, 1964, pp. 17–78.]

[4] A. Berman and R. Plemmons, Nonnegative Matricies in the Mathematical Sciences, SIAM, Philadephia PA, 1994,

[5] D. P. Bertsekas, Dynamic Programming and Optimal Control vol. 1& 2, Athena Scientific, Belmont, Mass., 1995.

[6] P. Billingsley, Probability and Measure (second ed.), J. Wiley & Sons, New York, 1986.

[7] P. Billingsley, Ergodic Theory and Information, J. Wiley & Sons, NY, 1965.

[8] R. E. Blahut, Principles and Practice of Information Theory, Addison-Wesley, Reading MA, 1987.

[9] L. Breiman, Probability, Addison-Wesley, Reading, MA, 1968.

[10] Pierre Brémaud, Markov Chains: Gibbs Fields, MonteCarlo Simulation, and Queues, Springer-Verlag, NY, 1999.

[11] D. M. Chance, An Introduction to Derivatives (4th edition), Dryden Press, Ft. Worth, TX, 1998.

[12] H. Chen and D. D. Yao, Fundamentals of Queueing Networks: Performance, Asymptotics and Optimization, Springer-Verlag, N.Y., 2001.

[13] P. L. Chow and R. Z. Khasminskii, *Method of Lyapunov functions for analysis of absorption and explosion in Markov chains*, Prob. Info. Transmission **47** (2011), pp. 232–250.

[14] R. W. Cottle, J. S. Pang, and R. E. Stone, The Linear Complementarity Problem, Academic Press, Boston, 1992.

[15] K. L. Chung, Markov Chains, Springer-Verlag, NY, 1967.

[16] J. L. Doob, Stochastic Processes,

[17] L. E. Dubins and L. J. Savage, Inequalities for Stohastic Processes: How to Gamble if You Must, Dover Publications, NY, 1976.

[18] R. Durrett, Essentials of Stochastic Processes, Springer, NY, 2012.

[19] R. Durrett, *The Contact Process: 1974–1989*, in Mathemaitcs of Random Media, AMS Lectures in Applied Mathematics, v. 27, 1991.

[20] E. B. Dynkin & A. A. Yushkevtich, Markov Processes: Theorems and Problems, Plenum Press, NY, 1969.

[21] S. N. Ethier and T. G. Kurtz, Markov Processes: Characterization and Convergence, J. Wiley & Sons, New York, 1986.

[22] W. Feller, An Introduction to Probability Theory and Its Applications vol. 1 (second ed.), J. Wiley & Sons, New York, 1957.

[23] T. S. Ferguson, Optimal Stopping and Applications,

http://www.math.ucla.edu/~tom/Stopping/Contents.html

[24] W. N. Francis and H. Kucera, A Standard Corpus of Present-Day Edited American English, Providence, Brown University Press, 1967.

[25] G. Grimmett and D. Stirzaker, Probability and Random Processes (3rd ed.), Oxford U. Press, Oxford, 2001.

[26] R. Z. Has'minskiĭ, Stochastic Stability of Differential Equaitons, Sijthoff & Noordhoff, Rockville MD, 1980.

[27] D. J. Higham, *Modeling and Simulation of Chemical Reactions*, SIAM Review **50** (2008), pp. 347-386.

[28] P. Hoel, S. Port, C. Stone, Intorduction to Stochastic Processes, Houghton Mifflin, Boston, 1972.

[29] P. Hoffman, The Man Who Loved Only Numbers, Hyperion, NY, 1998.

[30] R. A. Howard, Dynamic programming and Markov processes, Technology Press of M.I.T., Cambridge, 1960.

[31] J. C. Hull, Options, Futures, and Other Derivative Securities (third edition), Prentice Hall, Englewood Cliffs, NJ, 1997.

[32] I. Karatzas and S. Shreve, Brownian Motion and Stochastic Calculus (2nd edition), Springer, NY, 1991.

[33] S. Karlin, A First Course i Stochastic Processes, Academic Press, NY, 1969.

[34] J. Kemeny and J. Snell, Finite Markov Chains, Springer-Verlag, New York, 1976.

[35] A. I. Khinchin, Information Theory, Dover, NY, 1957.

[36] D. E. Knuth, The Art of Computer Programming — Vol.2: Seminumerical Algorithms, Addison-Wesley, Reading, MA, 1981.

[37] A. G. Konheim, Cryptography, A Primer, J. Wiley & Sons, New York, 1981.

[38] S. Lang, Linear Algebra (second ed.), Addison-Wesley, Reading Mass., 1971.

[39] A. Langville and C. Meyer, *A Survey of Eigenvector Methods for Web Information Retrieval*, SIAM Review **47** (2005), pp. 135–161.

[40] G. Lawler, Introduction to Stochastic Processes, Chapman & Hall/CRC, Boca Raton, 2006.

[41] J. Marsden and A. Tromba, Vector Calculus (fifth ed.), W. H. Freeman and Co., New York, 2003.

[42] T. Mikosch, Elementary Stochastic Calculus, with Finance in View, World Scientific, Singapore, 1998.

[43] K. G. Murty and F.-T. Yu, Linear Complementarity, Linear and Nonlinear Programming, Internet Edition:

http://ioe.engin.umich.edu/people/fac/books/murty/linear_complementarity_webbook/

[44] M. Musiela and M. Rutkowski, Martingale Methods in Financial Modeling : Theory and Applications, Springer, New York, 2005.

[45] J. R. Norris, Markov Chains, Cambridge Univ. Press, Cambridge, UK, 1997.

[46] B. Øksendal, Stochastic Differential Equations: An Introduction with Applications (6th ed.), Springer, NY, 2013.

[47] P. E. Pfiffer, Concepts of Probability Theory, Dover, 1978.

[48] S. C. Port, *A simple probabilistic proof of the discrete generalized renewal theorem*, Ann. Math. Stat. **36** (1965), pp. 1294–1297.

[49] M. L. Puterman, Markov Decision Processes: Discrete Stochastic Dynamic Programming, J. Wiley & Sons, NY, 1994.

[50] Rand Corporation, A Million Random Digits with 100,000 Normal Deviates, Free Press, Glencoe, Ill., 1955. http://www.rand.org/pubs/monograph_reports/MR1418.html

[51] L. C. G. Rogers and D. Williams,Diffusions, Markov Processes and Martingales (2 vol.s) (2nd ed.), Cambridge Univ. Press, Cambridge, UK, 2000.

[52] G. S. Ross, Introduction to Probability Models (10th edition), Academic Press, Amsterdam, 2010.

[53] W. Rudin, Principles of Mathematical Analysis (3rd ed.), McGraw-Hill, NY, 1976.

[54] N. Shokhirev, *Hidden Markov models*, http://www.shokhirev.com/nikolai/abc/alg/hmm/hmm.html

[55] S. E. Shreve, Stochastic Calculus for Finance (2 vol.s), Springer, New York, 2004.

[56] S. E. Shreve, Methods of Mathematical Finance (Stochastic Modelling and Applied Probability), Springer, NY, 1998.

[57] D. W. Stroock, An Introduction to Markov Processes, Springer, Berlin, 2005.

[58] D. W. Stroock, *Diffusion Processes Associated with Levy Generators*, Z. Wahr. **32** (1975), pp. 209–244.

[59] D. W. Stroock and S. R. S. Varadhan, Multidimensional Diffusion Processes, Springer,-Verlag, New York, 1979.

[60] F. Spitzer, Principles of Random Walk, Springer, New York, 1976.

[61] H. M. Taylor, *Optimal stopping in a Markov Process*, Ann. Math. Stat. **39** (1968), pp.1333–1344.

[62] W. Trappe and L. C. Washington, Introduction to Cryptography with Coding Theory (2nd ed.), Pearson Prentice Hall, Upper Saddle River NJ, 2006

[63] S. R. S. Varadhan, Probability Theory, Courant Lecture Notes #7, AMS, Providence, RI, 2001.

[64] S. R. S. Varadhan, Stochastic Processes, Courant Lecture Notes #16, AMS, Providence, RI, 2007.

[65] P. Whittle, Probability via Expectation (third ed.), Springer-Verlag, NY, 1992.