

Contents

1	Markov chains	5
1.1	Specifying and simulating a Markov chain	5
1.2	The Markov property	8
1.3	“It’s all just matrix theory”	9
1.4	The basic limit theorem of Markov chains	10
1.5	Stationary distributions	12
1.6	Irreducibility, periodicity, and recurrence	15
1.7	An aside on coupling	25
1.8	Proof of the Basic Limit Theorem	27
1.9	A SLLN for Markov chains	31
1.10	Exercises	36
2	Markov Chains: Examples and Applications	43
2.1	Branching Processes	43
2.2	Time Reversibility	46
2.3	More on Time Reversibility: A Tandem Queue Model	49
2.4	The Metropolis method	53
2.5	Simulated annealing	57
2.5.1	Description of the method	58
2.5.2	The Main Theorem	65
2.6	Ergodicity Concepts	67
2.6.1	The Ergodic Coefficient	68
2.6.2	Sufficient Conditions for Weak and Strong Ergodicity	69
2.7	Proof of Main Theorem of Simulated Annealing	71
2.8	Card Shuffling	73
2.8.1	“Top-in-at-random” Shuffle	74
2.8.2	Threshold Phenomenon	74
2.8.3	A random time to exact stationarity	76
2.8.4	Strong Stationary Times	77
2.8.5	Proof of threshold phenomenon in shuffling	78
2.9	Exercises	81

3	MRFs and HMMs	87
3.1	MRF's on Graphs and HMM's	87
3.2	Bayesian Framework	89
3.3	Hammersley-Clifford Theorem	90
3.4	Long range dependence in the Ising model	97
3.5	Hidden Markov chains	99
3.5.1	Description of the model	100
3.5.2	How to calculate likelihoods	102
3.5.3	Maximum Likelihood and the EM algorithm	104
3.5.4	Applying the EM algorithm to a hidden Markov chain	106
3.6	The Gibbs Sampler	111
3.7	Exercises	112
4	Martingales	117
4.1	Why "martingale"?	117
4.2	Definitions	118
4.3	Examples	119
4.4	Optional sampling	121
4.5	Stochastic integrals and option pricing	126
4.6	Martingale convergence	135
4.7	Stochastic approximation	138
4.8	Exercises	142
5	Brownian motion	149
5.1	The definition	150
5.2	Visualizing Brownian motion	153
5.3	The reflection principle	156
5.4	Conditional distributions	157
5.5	Existence and Construction	159
5.6	The Brownian bridge	163
5.6.1	A boundary crossing probability	164
5.6.2	Application to testing for uniformity	165
5.7	Boundary Crossing	166
5.7.1	Differential Equations	168
5.7.2	Martingales	170
5.8	Some confusing questions (or answers)	172
5.9	Exercises	175
6	Diffusions and Stochastic Calculus	179
6.1	Specifying a diffusion	180
6.2	A calculation with diffusions	184
6.3	Infinitesimal parameters of a function of a diffusion	186
6.4	Backward and forward equations	187
6.5	Stationary distributions	191
6.6	Probability flux for diffusions	192

6.7	Quadratic Variation of Brownian Motion	194
6.8	Stochastic Differential Equations	196
6.9	Simple examples	199
6.10	The Black-Scholes Formula	201
6.11	Stochastic Integrals	204
7	Likelihood Ratios	209
7.1	The idea of likelihood ratios	209
7.2	The idea of importance sampling	210
7.3	A gambler's ruin problem	213
7.4	Importance sampling for the gambler's ruin	216
7.5	Brownian motion	217
7.6	The Sequential Probability Ratio Test	221
8	Extremes and Poisson clumping	223
8.1	The Poisson clumping heuristic	223
A	Convergence theorems	227
B	Conditioning	229
B.1	Definitions	229
B.2	Summary of some rules	231
B.3	Conditional probabilities and expectations	232

1. Markov chains

Section 1. What is a Markov chain? How to simulate one.

Section 2. The Markov property.

Section 3. How matrix multiplication gets into the picture.

Section 4. Statement of the Basic Limit Theorem about convergence to stationarity. A motivating example shows how complicated random objects can be generated using Markov chains.

Section 5. Stationary distributions, with examples. Probability flux.

Section 6. Other concepts from the Basic Limit Theorem: irreducibility, periodicity, and recurrence. An interesting classical example: recurrence or transience of random walks.

Section 7. Introduces the idea of coupling.

Section 8. Uses coupling to prove the Basic Limit Theorem.

Section 9. A Strong Law of Large Numbers for Markov chains.

Markov chains are a relatively simple but very interesting and useful class of random processes. A Markov chain describes a system whose state changes over time. The changes are not completely predictable, but rather are governed by probability distributions. These probability distributions incorporate a simple sort of dependence structure, where the conditional distribution of future states of the system, given some information about past states, depends only on the most recent piece of information. That is, what matters in predicting the future of the system is its present state, and not the path by which the system got to its present state. Markov chains illustrate many of the important ideas of stochastic processes in an elementary setting. This classical subject is still very much alive, with important developments in both theory and applications coming at an accelerating pace in recent decades.

1.1 Specifying and simulating a Markov chain

What is a Markov chain*? One answer is to say that it is a sequence $\{X_0, X_1, X_2, \dots\}$ of random variables that has the “Markov property”; we will discuss this in the next section. For now, to get a feeling for what a Markov chain is, let’s think about how to *simulate* one, that is, how to use a computer or a table of random numbers to generate a typical “sample

* Unless stated otherwise, when we use the term “Markov chain,” we will be restricting our attention to the subclass of *time-homogeneous* Markov chains. We’ll do this to avoid monotonous repetition of the phrase “time-homogeneous.” I’ll point out below the place at which the assumption of time-homogeneity enters.

path.” To start, how do I tell you which particular Markov chain I want you to simulate? There are three items involved: to specify a Markov chain, I need to tell you its

- State space \mathcal{S} .

\mathcal{S} is a finite or countable set of *states*, that is, values that the random variables X_i may take on. For definiteness, and without loss of generality, let us label the states as follows: either $\mathcal{S} = \{1, 2, \dots, N\}$ for some finite N , or $\mathcal{S} = \{1, 2, \dots\}$, which we may think of as the case “ $N = \infty$ ”.

- Initial distribution π_0 .

This is the probability distribution of the Markov chain at time 0. For each state $i \in \mathcal{S}$, we denote by $\pi_0(i)$ the probability $\mathbb{P}\{X_0 = i\}$ that the Markov chain starts out in state i . Formally, π_0 is a function taking \mathcal{S} into the interval $[0, 1]$ such that

$$\pi_0(i) \geq 0 \text{ for all } i \in \mathcal{S}$$

and

$$\sum_{i \in \mathcal{S}} \pi_0(i) = 1.$$

Equivalently, instead of thinking of π_0 as a function from \mathcal{S} to $[0, 1]$, we could think of π_0 as the vector whose i th entry is $\pi_0(i) = \mathbb{P}\{X_0 = i\}$.

- Probability transition rule.

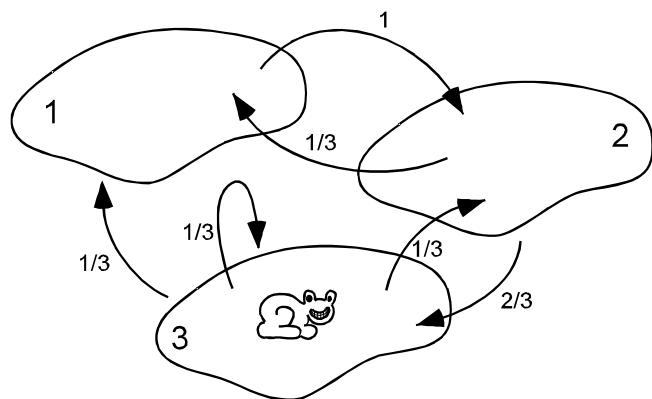
This is specified by giving a matrix $P = (P_{ij})$. If \mathcal{S} contains N states, then P is an $N \times N$ matrix. The interpretation of the number P_{ij} is the conditional probability, given that the chain is in state i at time n , say, that the chain jumps to the state j at time $n + 1$. That is,

$$P_{ij} = \mathbb{P}\{X_{n+1} = j \mid X_n = i\}.$$

We will also use the notation $P(i, j)$ for the same thing. Note that we have written this probability as a function of just i and j , but of course it could depend on n as well. The **time homogeneity** restriction mentioned in the previous footnote is just the assumption that this probability does not depend on the time n , but rather remains constant over time.

Formally, a **probability transition matrix** is an $N \times N$ matrix whose entries are all nonnegative and whose rows sum to 1.

Finally, you may be wondering why we bother to arrange these conditional probabilities into a matrix. That is a good question, and will be answered soon.

(1.1) FIGURE. *The Markov frog.*

We can now get to the question of how to simulate a Markov chain, now that we know how to specify what Markov chain we wish to simulate. Let's do an example: suppose the state space is $\mathcal{S} = \{1, 2, 3\}$, the initial distribution is $\pi_0 = (1/2, 1/4, 1/4)$, and the probability transition matrix is

$$(1.2) \quad P = \begin{matrix} & \begin{matrix} 1 & 2 & 3 \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ 3 \end{matrix} & \begin{pmatrix} 0 & 1 & 0 \\ 1/3 & 0 & 2/3 \\ 1/3 & 1/3 & 1/3 \end{pmatrix} \end{matrix}.$$

Think of a frog hopping among lily pads as in Figure 1.1. How does the Markov frog choose a path? To start, he chooses his initial position X_0 according to the specified initial distribution π_0 . He could do this by going to his computer to generate a uniformly distributed random number $U_0 \sim \text{Unif}(0, 1)$, and then taking[†]

$$X_0 = \begin{cases} 1 & \text{if } 0 \leq U_0 \leq 1/2 \\ 2 & \text{if } 1/2 < U_0 \leq 3/4 \\ 3 & \text{if } 3/4 < U_0 \leq 1 \end{cases}$$

For example, suppose that U_0 comes out to be 0.8419, so that $X_0 = 3$. Then the frog chooses X_1 according to the probability distribution in row 3 of P , namely, $(1/3, 1/3, 1/3)$; to do this, he paws his computer again to generate $U_1 \sim \text{Unif}(0, 1)$ independently of U_0 , and takes

$$X_1 = \begin{cases} 1 & \text{if } 0 \leq U_0 \leq 1/3 \\ 2 & \text{if } 1/3 < U_0 \leq 2/3 \\ 3 & \text{if } 2/3 < U_0 \leq 1. \end{cases}$$

[†]Don't be distracted by the distinctions between " $<$ " and " \leq " below—for example, what we do if U_0 comes out be exactly $1/2$ or $3/4$ —since the probability of U_0 taking on any particular precise value is 0.

Suppose he happens to get $U_1 = 0.1234$, so that $X_1 = 1$. Then he chooses X_2 according to row 1 of P , so that $X_2 = 2$; there's no choice this time. Next, he chooses X_3 according to row 2 of P . And so on. . . .

1.2 The Markov property

Clearly, in the previous example, if I told you that we came up with the values $X_0 = 3$, $X_1 = 1$, and $X_2 = 2$, then the conditional probability distribution for X_3 is

$$\mathbb{P}\{X_3 = j \mid X_0 = 3, X_1 = 1, X_2 = 2\} = \begin{cases} 1/3 & \text{for } j = 1 \\ 0 & \text{for } j = 2 \\ 2/3 & \text{for } j = 3, \end{cases}$$

which is also the conditional probability distribution for X_3 given only the information that $X_2 = 2$. In other words, given that $X_0 = 3$, $X_1 = 1$, and $X_2 = 2$, the only information relevant to the distribution to X_3 is the information that $X_2 = 2$; we may ignore the information that $X_0 = 3$ and $X_1 = 1$. This is clear from the description of how to simulate the chain! Thus,

$$\mathbb{P}\{X_3 = j \mid X_2 = 2, X_1 = 1, X_0 = 3\} = \mathbb{P}\{X_3 = j \mid X_2 = 2\} \text{ for all } j.$$

This is an example of the Markov property.

(1.3) DEFINITION. A process X_0, X_1, \dots satisfies the **Markov property** if

$$\begin{aligned} \mathbb{P}\{X_{n+1} = i_{n+1} \mid X_n = i_n, X_{n-1} = i_{n-1}, \dots, X_0 = i_0\} \\ = \mathbb{P}\{X_{n+1} = i_{n+1} \mid X_n = i_n\} \end{aligned}$$

for all n and all $i_0, \dots, i_{n+1} \in \mathcal{S}$.

The issue addressed by the Markov property is the *dependence structure* among random variables. The simplest dependence structure for X_0, X_1, \dots is no dependence at all, that is, independence. The Markov property could be said to capture the next simplest sort of dependence: in generating the process X_0, X_1, \dots sequentially, the “next” state X_{n+1} depends only on the “current” value X_n , and not on the “past” values X_0, \dots, X_{n-1} . The Markov property allows much more interesting and general processes to be considered than if we restricted ourselves to independent random variables X_i , without allowing so much generality that a mathematical treatment becomes intractable.

- ▷ The idea of the Markov property might be expressed in a pithy phrase, “Conditional on the present, the future does not depend on the past.” But there are subtleties. Exercise [1.1] shows the need to think carefully about what the Markov property does and does not say. [The exercises are collected in the final section of the chapter.]

The Markov property implies a simple expression for the probability of our Markov chain taking any specified path, as follows:

$$\begin{aligned}
 & \mathbb{P}\{X_0 = i_0, X_1 = i_1, X_2 = i_2, \dots, X_n = i_n\} \\
 &= \mathbb{P}\{X_0 = i_0\} \mathbb{P}\{X_1 = i_1 \mid X_0 = i_0\} \mathbb{P}\{X_2 = i_2 \mid X_1 = i_1, X_0 = i_0\} \\
 & \quad \cdots \mathbb{P}\{X_n = i_n \mid X_{n-1} = i_{n-1}, \dots, X_1 = i_1, X_0 = i_0\} \\
 &= \mathbb{P}\{X_0 = i_0\} \mathbb{P}\{X_1 = i_1 \mid X_0 = i_0\} \mathbb{P}\{X_2 = i_2 \mid X_1 = i_1\} \\
 & \quad \cdots \mathbb{P}\{X_n = i_n \mid X_{n-1} = i_{n-1}\} \\
 &= \pi_0(i_0) P(i_0, i_1) P(i_1, i_2) \cdots P(i_{n-1}, i_n).
 \end{aligned}$$

So, to get the probability of a path, we start out with the initial probability of the first state and successively multiply by the matrix elements corresponding to the transitions along the path.

The Markov property of Markov chains can be generalized to allow dependence on the previous several values. The next definition makes this idea precise.

(1.4) DEFINITION. We say that a process X_0, X_1, \dots is ***r*th order Markov** if

$$\begin{aligned}
 & \mathbb{P}\{X_{n+1} = i_{n+1} \mid X_n = i_n, X_{n-1} = i_{n-1}, \dots, X_0 = i_0\} \\
 &= \mathbb{P}\{X_{n+1} = i_{n+1} \mid X_n = i_n, \dots, X_{n-r+1} = i_{n-r+1}\}
 \end{aligned}$$

for all $n \geq r$ and all $i_0, \dots, i_{n+1} \in \mathcal{S}$.

- ▷ Is this generalization general enough to capture everything of interest? No; for example, Exercise [1.6] shows that an important type of stochastic process, the “moving average process,” is generally not r^{th} order Markov for any r .

1.3 “It’s all just matrix theory”

Recall that the vector π_0 having components $\pi_0(i) = \mathbb{P}\{X_0 = i\}$ is the initial distribution of the chain. Let π_n denote the distribution of the chain at time n , that is, $\pi_n(i) = \mathbb{P}\{X_n = i\}$. Suppose for simplicity that the state space is finite: $\mathcal{S} = \{1, \dots, N\}$, say. Then the Markov chain has an $N \times N$ probability transition matrix

$$P = (P_{ij}) = (P(i, j)),$$

where $P(i, j) = \mathbb{P}\{X_{n+1} = j \mid X_n = i\} = \mathbb{P}\{X_1 = j \mid X_0 = i\}$. The law of total probability gives

$$\begin{aligned}
 \pi_{n+1}(j) &= \mathbb{P}\{X_{n+1} = j\} \\
 &= \sum_{i=1}^N \mathbb{P}\{X_n = i\} \mathbb{P}\{X_{n+1} = j \mid X_n = i\} \\
 &= \sum_{i=1}^N \pi_n(i) P(i, j),
 \end{aligned}$$

which, in matrix notation, is just the equation

$$\pi_{n+1} = \pi_n P.$$

Note that here we are thinking of π_n and π_{n+1} as *row vectors*, so that, for example,

$$\pi_n = (\pi_n(1), \dots, \pi_n(N)).$$

Thus, we have

$$\begin{aligned} (1.5) \quad \pi_1 &= \pi_0 P \\ \pi_2 &= \pi_1 P = \pi_0 P^2 \\ \pi_3 &= \pi_2 P = \pi_0 P^3, \end{aligned}$$

and so on, so that by induction

$$(1.6) \quad \pi_n = \pi_0 P^n.$$

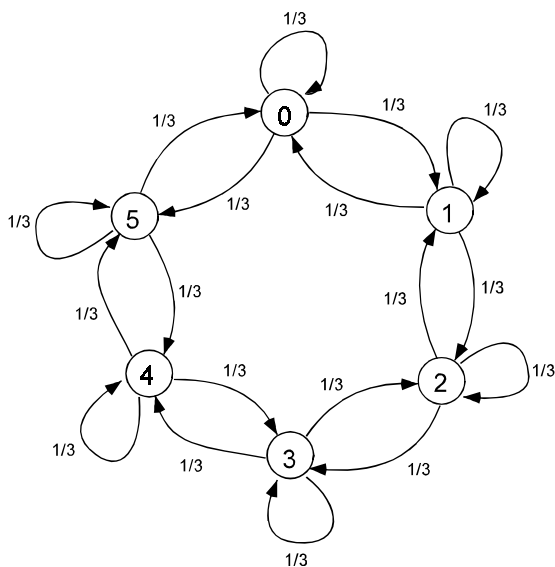
We will let $P^n(i, j)$ denote the (i, j) element in the matrix P^n .

▷ *Exercise [1.7] gives some basic practice with the definitions.*

So, in principle, we can find the answer to any question about the probabilistic behavior of a Markov chain by doing matrix algebra, finding powers of matrices, etc. However, what is viable in practice may be another story. For example, the state space for a Markov chain that describes repeated shuffling of a deck of cards contains $52!$ elements—the permutations of the 52 cards of the deck. This number $52!$ is large: about 80 million million million million million million million million million million. The probability transition matrix that describes the effect of a single shuffle is a $52!$ by $52!$ matrix. So, “all we have to do” to answer questions about shuffling is to take powers of such a matrix, find its eigenvalues, and so on! In a practical sense, simply reformulating probability questions as matrix calculations often provides only minimal illumination in concrete questions like “how many shuffles are required in order to mix the deck well?” Probabilistic reasoning can lead to insights and results that would be hard to come by from thinking of these problems as “just” matrix theory problems.

1.4 The basic limit theorem of Markov chains

As indicated by its name, the theorem we will discuss in this section occupies a fundamental and important role in Markov chain theory. What is it all about? Let’s start with an example in which we can all see intuitively what is going on.



(1.7) FIGURE. A random walk on a clock.

(1.8) EXAMPLE [RANDOM WALK ON A CLOCK]. For ease of writing and drawing, consider a clock with 6 numbers on it: 0,1,2,3,4,5. Suppose we perform a random walk by moving clockwise, moving counterclockwise, and staying in place with probabilities $1/3$ each at every time n . That is,

$$P(i, j) = \begin{cases} 1/3 & \text{if } j = i - 1 \bmod 6 \\ 1/3 & \text{if } j = i \\ 1/3 & \text{if } j = i + 1 \bmod 6. \end{cases}$$

Suppose we start out at $X_0 = 2$, say. That is,

$$\pi_0 = (\pi_0(0), \pi_0(1), \dots, \pi_0(5)) = (0, 0, 1, 0, 0, 0).$$

Then of course

$$\pi_1 = (0, \frac{1}{3}, \frac{1}{3}, \frac{1}{3}, 0, 0),$$

and it is easy to calculate

$$\pi_2 = (\frac{1}{9}, \frac{2}{9}, \frac{1}{3}, \frac{2}{9}, \frac{1}{9}, 0)$$

and

$$\pi_3 = (\frac{3}{27}, \frac{6}{27}, \frac{7}{27}, \frac{6}{27}, \frac{3}{27}, \frac{2}{27}).$$

Notice how the probability is spreading out away from its initial concentration on the state 2. We could keep calculating π_n for more values of n , but it is intuitively clear what will happen: the probability will continue to spread out, and π_n will approach the uniform distribution:

$$\pi_n \rightarrow (\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6})$$

as $n \rightarrow \infty$. Just imagine: if the chain starts out in state 2 at time 0, then we close our eyes while the random walk takes 10,000 steps, and then we are asked to guess what state the random walk is in at time 10,000, what would we think the probabilities of the various states are? I would say: “ $X_{10,000}$ is for all practical purposes uniformly distributed over the 6 states.” By time 10,000, the random walk has essentially “forgotten” that it started out in state 2 at time 0, and it is nearly equally likely to be anywhere.

Now observe that the starting state 2 was not special; we could have started from anywhere, and over time the probabilities would spread out away from the initial point, and approach the same limiting distribution. Thus, π_n approaches a limit that does not depend upon the initial distribution π_0 . \square

The following “Basic Limit Theorem” says that the phenomenon discussed in the previous example happens quite generally. We will start with a statement and discussion of the theorem, and then prove the theorem later.

(1.9) THEOREM [BASIC LIMIT THEOREM]. *Let X_0, X_1, \dots be an irreducible, aperiodic Markov chain having a stationary distribution $\pi(\cdot)$. Let X_0 have the distribution π_0 , an arbitrary initial distribution. Then $\lim_{n \rightarrow \infty} \pi_n(i) = \pi(i)$ for all states i .*

We need to define the words “irreducible,” “aperiodic,” and “stationary distribution.” Let’s start with “stationary distribution.”

1.5 Stationary distributions

Suppose a distribution π on \mathcal{S} is such that, if our Markov chain starts out with initial distribution $\pi_0 = \pi$, then we also have $\pi_1 = \pi$. That is, if the distribution at time 0 is π , then the distribution at time 1 is still π . Then π is called a **stationary distribution** for the Markov chain. From (1.5) we see that the definition of stationary distribution amounts to saying that π satisfies the equation

$$(1.10) \quad \pi = \pi P,$$

that is,

$$\pi(j) = \sum_{i \in \mathcal{S}} \pi(i)P(i, j) \quad \text{for all } j \in \mathcal{S}.$$

[In the case of an infinite state space, (1.10) is an infinite system of equations.] Also from equations (1.5) we can see that if the Markov chain has initial distribution $\pi_0 = \pi$, then we have not only $\pi_1 = \pi$, but also $\pi_n = \pi$ for all n . That is, a Markov chain started out in a stationary distribution π stays in the distribution π forever; that’s why the distribution π is called “stationary.”

(1.11) EXAMPLE. If the $N \times N$ probability transition matrix P is symmetric, then the uniform distribution $[\pi(i) = 1/N \text{ for all } i]$ is stationary. More generally, the uniform distribution is stationary if the matrix P is *doubly stochastic*, that is, the column-sums of P are 1 (we already know the row-sums of P are all 1). \square

It should not be surprising that π appears as the limit in Theorem (1.9). It is easy to see that if π_n approaches a limiting distribution as $n \rightarrow \infty$, then that limiting distribution must be stationary. To see this, suppose that $\lim_{n \rightarrow \infty} \pi_n = \tilde{\pi}$, and let $n \rightarrow \infty$ in the equation $\pi_{n+1} = \pi_n P$ to obtain $\tilde{\pi} = \tilde{\pi} P$, which says that $\tilde{\pi}$ is stationary.

▷ *The argument just stated goes through clearly and easily when the state space is finite—there are no issues of mathematical analysis that arise in taking the limits. I'll leave it as Exercise [1.10] for the mathematically inclined among you to worry about the details of carrying through the above argument in the case of a countably infinite state space.*

Computing stationary distributions is an algebra problem.

(1.12) EXAMPLE. Let's find the stationary distribution for the frog chain, whose probability transition matrix was given in (1.2). Since most people are accustomed to solving linear systems of the form $Ax = b$, let us take the transpose of the equation $\pi(P - I) = 0$, obtaining the equation $(P^T - I)\pi^T = 0$. In our example, this becomes

$$\begin{pmatrix} -1 & 1/3 & 1/3 \\ 1 & -1 & 1/3 \\ 0 & 2/3 & -2/3 \end{pmatrix} \begin{pmatrix} \pi(1) \\ \pi(2) \\ \pi(3) \end{pmatrix} = 0,$$

or

$$\begin{pmatrix} -1 & 1/3 & 1/3 \\ 0 & -2/3 & 2/3 \\ 0 & 2/3 & -2/3 \end{pmatrix} \begin{pmatrix} \pi(1) \\ \pi(2) \\ \pi(3) \end{pmatrix} = 0,$$

which has solutions of the form $\pi = \text{const}(2/3, 1, 1)$. For the unique solution that satisfies the constraint $\sum \pi(i) = 1$, take the constant to be $3/8$, so that $\pi = (1/4, 3/8, 3/8)$.

As an alternative approach, here is another way, aside from solving the linear equations, to address the problem of finding a stationary distribution; this idea can work particularly well with computers. If we believe the Basic Limit Theorem, we should see the stationary distribution in the limit as we run the chain for a long time. Let's try it: Here are some calculations of powers of the transition matrix P from (1.2):

$$P^5 = \begin{pmatrix} 0.246914 & 0.407407 & 0.345679 \\ 0.251029 & 0.36214 & 0.386831 \\ 0.251029 & 0.366255 & 0.382716 \end{pmatrix},$$

$$P^{10} = \begin{pmatrix} 0.250013 & 0.37474 & 0.375248 \\ 0.249996 & 0.375095 & 0.374909 \\ 0.249996 & 0.375078 & 0.374926 \end{pmatrix},$$

$$P^{20} = \begin{pmatrix} 0.2500000002 & 0.3749999913 & 0.3750000085 \\ 0.2499999999 & 0.375000003 & 0.374999997 \\ 0.2499999999 & 0.3750000028 & 0.3749999973 \end{pmatrix}.$$

So we don't really have to solve equations; in this example, any of the rows of the matrix P^{20} provides a very accurate approximation for π . No matter what state we start from, the

distribution after 20 steps of the chain is very close to $(.25, .375, .375)$. This is the Basic Limit Theorem in action. \square

(1.13) EXAMPLE [EHRENFEST CHAIN]. The Ehrenfest chain is a simple model of “mixing” processes. This chain can shed light on perplexing questions like “Why aren’t people dying all the time due to the air molecules bunching up in some odd corner of their bedrooms while they sleep?” The model considers d balls distributed among two urns, and results in a Markov chain $\{X_0, X_1, \dots\}$ having state space $\{0, 1, \dots, d\}$, with the state X_n of the chain at time n being the number of balls in urn #1 at time n . At each time, we choose a ball at random uniformly from the d possibilities, take that ball out of its current urn, and drop it into the other urn. Thus, $P(i, i-1) = i/d$ and $P(i, i+1) = (d-i)/d$ for all i .

- ▷ *What is the stationary distribution of the Ehrenfest chain? Exercise [1.9] asks you to discover and explain the answer, which turns out to be a distribution that is one of your old friends.*

\square

A Markov chain might have no stationary distribution, one stationary distribution, or infinitely many stationary distributions. We just saw examples with one stationary distribution. A trivial example with infinitely many is when P is the identity matrix, in which case all distributions are stationary. To find an example without any stationary distribution, we need to consider an infinite state space. [We will see later that any finite-state Markov chain has at least one stationary distribution.] An easy example of this has $\mathcal{S} = \{1, 2, \dots\}$ and $P(i, i+1) = 1$ for all i , which corresponds to a Markov chain that moves deterministically “to the right.” In this case, the equation $\pi(j) = \sum_{i \in \mathcal{S}} \pi(i)P(i, j)$ reduces to $\pi(j) = \pi(j-1)$, which clearly has no solution satisfying $\sum \pi(j) = 1$. Another interesting example is the *simple, symmetric random walk on the integers*: $P(i, i-1) = 1/2 = P(i, i+1)$. Here the equations for stationarity become

$$\pi(j) = \frac{1}{2}\pi(j-1) + \frac{1}{2}\pi(j+1).$$

Again it is easy to see [how?] that these equations have no solution π that is a probability mass function.

Intuitively, notice the qualitative difference: in the examples without a stationary distribution, the probability doesn’t settle down to a limit probability distribution—in the first example the probability moves off to infinity, and in the second example it spreads out in both directions. In both cases, the probability on any fixed state converges to 0; one might say the probability escapes off to infinity (or $-\infty$).

- ▷ *Exercise [1.8] analyzes an example of a Markov chain that moves around on all of the integers, while no probability escapes to infinity, and the chain has a stationary distribution.*

A Markov chain in its stationary distribution π is at peace with itself; its distribution stays constant, with no desire to change into anything else. This property is explored further in terms of the idea of “probability flux.”

(1.14) DEFINITION. For subsets A and B of the state space, define the **probability flux from the set A into the set B** to be

$$\text{flux}(A, B) = \sum_{i \in A} \sum_{j \in B} \pi(i)P(i, j)$$

A fundamental balancing property occurs when we consider the probability flux between a set A and its complement A^c , in which case

$$(1.15) \quad \text{flux}(A, A^c) = \text{flux}(A^c, A).$$

▷ Exercise [1.11] supplies some hints to help you prove this.

The left side of (1.15) is the “probability flux flowing out of A into A^c .” The equality says that this must be the same as the flux from A^c back into A . This has the suggestive interpretation that the stationary probabilities describe a stable system in which all the probability is happy where it is, and does not want to flow to anywhere else, so that the net flow from A to A^c must be zero. We can say this in a less mysterious way as follows. Think of $\pi(i)$ as the long run fraction of time that the chain is in state i . [We will soon see a theorem (“a strong law of large numbers for Markov chains”) that supports this interpretation.] Then $\pi(i)P(i, j)$ is the long run fraction of times that a transition from i to j takes place. But clearly the long run fraction of times occupied by transitions going from a state in A to a state in A^c must equal the long run fraction of times occupied by transitions going the opposite way. [In fact, along any sample path, the numbers of transitions that have occurred in the two directions up to any time n may differ by at most 1!]

1.6 Irreducibility, periodicity, and recurrence

We’ll start by introducing some convenient notation to be used throughout the remainder of this chapter, then we’ll define irreducibility and related terms.

(1.16) NOTATION. We will use the shorthand “ \mathbb{P}_i ” to indicate a probability taken in a Markov chain started in state i at time 0. That is, “ $\mathbb{P}_i(A)$ ” is shorthand for “ $\mathbb{P}\{A \mid X_0 = i\}$.” We’ll also use the notation “ \mathbb{E}_i ” in an analogous way for expectation.

(1.17) DEFINITION. Let i and j be two states. We say that j **is accessible from i** if it is possible [with positive probability] for the chain ever to visit state j if the chain starts in state i , or, in other words,

$$\mathbb{P}_i \left\{ \bigcup_{n=0}^{\infty} \{X_n = j\} \right\} > 0.$$

Clearly an equivalent condition is

$$(1.18) \quad \sum_{n=0}^{\infty} P^n(i, j) \triangleq \sum_{n=0}^{\infty} \mathbb{P}_i\{X_n = j\} > 0.$$

We say i **communicates with** j if j is accessible from i and i is accessible from j . We say that the Markov chain is **irreducible** if all pairs of states communicate.

- ▷ In Exercise [1.15] you are asked to show that the relation “communicates with” is an equivalence relation. That is, you will show that the “communicates with” relation is reflexive, symmetric, and transitive.

Recall that an equivalence relation on a set induces a partition of that set into equivalence classes. Thus, by Exercise [1.15], the state space \mathcal{S} may be partitioned into what we will call “communicating classes,” or simply “classes.” The chain is irreducible if there is just one communicating class, that is, the whole state space \mathcal{S} . Note that whether or not a Markov chain is irreducible is determined by the state space \mathcal{S} and the transition matrix $(P(i, j))$; the initial distribution π_0 is irrelevant. In fact, all that matters is the pattern of zeroes in the transition matrix.

Why do we require irreducibility in the “Basic Limit Theorem” (1.9)? Here is a trivial example of how the conclusion can fail if we do not assume irreducibility. Let $\mathcal{S} = \{0, 1\}$ and let $P = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$. Clearly the resulting Markov chain is not irreducible. Also, clearly the conclusion of the Basic Limit Theorem does not hold; that is, π_n does not approach any limit that is independent of π_0 . In fact, $\pi_n = \pi_0$ for all n .

Next, to discuss periodicity, let’s begin with another trivial example: take $\mathcal{S} = \{0, 1\}$ again, and let $P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$. The conclusion of the Basic Limit Theorem does not hold here: for example, if $\pi_0 = (1, 0)$, then $\pi_n = (1, 0)$ if n is even and $\pi_n = (0, 1)$ if n is odd. So in this case $\pi_n(1)$ alternates between the two values 0 and 1 as n increases, and hence does not converge to anything. The problem in this example is not lack of irreducibility; clearly this chain is irreducible. So, assuming the Basic Limit Theorem is true, the chain must not be aperiodic! That is, the chain is **periodic**. The trouble stems from the fact that, starting from state 1 at time 0, the chain can visit state 1 only at even times. The same holds for state 2.

(1.19) DEFINITION. Given a Markov chain $\{X_0, X_1, \dots\}$, define the **period** of a state i to be the greatest common divisor (gcd)

$$d_i = \gcd\{n : P^n(i, i) > 0\}.$$

Note that both states 1 and 2 in the example $P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ have period 2. In fact, the next result shows that if two states i and j communicate, then they must have the same period.

(1.20) THEOREM. If the states i and j communicate, then $d_i = d_j$.

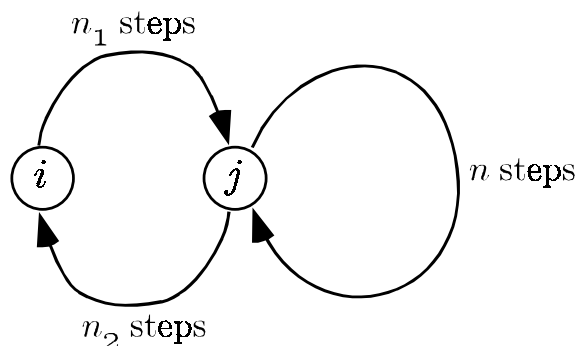
PROOF: Since j is accessible from i , by (1.18) there exists an n_1 such that $P^{n_1}(i, j) > 0$. Similarly, since i is accessible from j , there is an n_2 such that $P^{n_2}(j, i) > 0$. Noting that $P^{n_1+n_2}(i, i) > 0$, it follows that

$$d_i \mid n_1 + n_2,$$

that is, d_i divides $n_1 + n_2$, which means that $n_1 + n_2$ is an integer multiple of d_i . Now suppose that $P^n(j, j) > 0$. Then $P^{n_1+n+n_2}(i, i) > 0$, so that

$$d_i \mid n_1 + n + n_2.$$

Subtracting the last two displays gives $d_i \mid n$. Since n was an arbitrary integer satisfying $P^n(j, j) > 0$, we have found that d_i is a common divisor of the set $\{n : P^n(j, j) > 0\}$. Since d_j is defined to be the *greatest* common divisor of this set, we have shown that $d_j \geq d_i$. Interchanging the roles of i and j in the previous argument gives the opposite inequality $d_i \geq d_j$. This completes the proof. \square



It follows from Theorem (1.20) that all states in a communicating class have the same period. We say that the period of a state is a “class property.” In particular, all states in an irreducible Markov chain have the same period. Thus, we can speak of *the period of a Markov chain* if that Markov chain is irreducible: the period of an irreducible Markov chain is the period of any of its states.

(1.21) DEFINITION. An irreducible Markov chain is said to be **aperiodic** if its period is 1, and **periodic** otherwise.

▷ A simple sufficient (but not necessary) condition for an irreducible chain to be aperiodic is that there exist a state i such that $P(i, i) > 0$. This is Exercise [1.16].

We have now discussed all of the words we need in order to understand the statement of the Basic Limit Theorem (1.9). We will need another concept or two before we can

get to the proof, and the proof will then take some time beyond that. So I propose that we pause to discuss an interesting example of an application of the Basic Limit Theorem; this will help us build up some motivation to carry us through the proof, and will also give some practice that should be helpful in assimilating the concepts of irreducibility and aperiodicity. We'll also use the next example to introduce the important idea of using the Basic Limit Theorem, in a sense, *in reverse*, to generate random objects from specified distributions. This idea underlies many of the modern uses of Markov chains.

(1.22) EXAMPLE [GENERATING A RANDOM TABLE WITH FIXED ROW AND COLUMN SUMS]. Consider the 4×4 table of numbers that is enclosed within the rectangle below. The four numbers along the bottom of the table are the column sums, and those along the right edge of the table are the row sums.

68	119	26	7	220
20	84	17	94	215
15	54	14	10	93
5	29	14	16	64
108	286	71	127	

Suppose we want to generate a random, uniformly distributed, 4×4 table of nonnegative integers that has the same row and column sums as the table above. To make sure the goal is clear, define \mathcal{S} to be the set of all nonnegative 4×4 tables that have the given row and column sums. Let $\#(\mathcal{S})$ denote the cardinality of \mathcal{S} , that is, the number of elements in \mathcal{S} . Remember, each element of \mathcal{S} is a 4×4 table! We want to generate a random element, that is, a random 4×4 table, from \mathcal{S} , with each element having equal probability—that's the “uniform” part. That is, each of the $\#(\mathcal{S})$ tables in \mathcal{S} should have probability $1/\#(\mathcal{S})$ of being the table actually generated.

In spirit, this problem is the same as the much simpler problem of drawing a uniformly distributed state from our random walk on a clock as described in Example (1.8). This much simpler problem is merely to generate a uniformly distributed random element X from the set $\mathcal{S} = \{1, 2, 3, 4, 5, 6\}$, and we can do that without any fancy Markov chains. Just generate a random number $U \sim U[0, 1]$, and then take $X = i$ if U is between $(i-1)/6$ and $i/6$.

Although the two problems may be spiritually the same, there is a crucial practical difference. The set \mathcal{S} for the clock problem has only 6 elements. The set \mathcal{S} for the 4×4 tables is much larger, and in fact we don't know how many elements it has!

So an approach that works fine for $\mathcal{S} = \{1, 2, 3, 4, 5, 6\}$ —generate a $U \sim U[0, 1]$ and chop up the interval $[0, 1]$ into the appropriate number of pieces—cannot be used to generate a random 4×4 table in our example. However, the Basic Limit Theorem suggests another general approach: start from any state in \mathcal{S} , and run an appropriate Markov chain [such as the random walk on the clock] for a sufficiently long time, and take whatever state the chain finds itself in. This approach is rather silly if \mathcal{S} is very simple, like $\mathcal{S} = \{1, 2, 3, 4, 5, 6\}$, but in many practical problems, it is the only approach that has a hope of working. In our 4×4 table problem, we can indeed generate an approximate solution, that is, a random

table having a distribution arbitrarily close to uniform, by running a Markov chain on \mathcal{S} , our set of tables.

Here is one way to do it. Start with any table having the correct row and column sums; so of course the 4×4 table given above will do. Denote the entries in that table by a_{ij} . Choose a pair $\{i_1, i_2\}$ of rows at random, that is, uniformly over the $\binom{4}{2} = 6$ possible pairs. Similarly, choose a random pair of columns $\{j_1, j_2\}$. Then flip a coin. If you get heads: add 1 to $a_{i_1 j_1}$ and $a_{i_2 j_2}$, and subtract 1 from $a_{i_1 j_2}$ and $a_{i_2 j_1}$ if you can do so without producing any negative entries—if you cannot do so, then do nothing. Similarly, if the coin flip comes up tails, then subtract 1 from $a_{i_1 j_1}$ and $a_{i_2 j_2}$, and add 1 to $a_{i_1 j_2}$ and $a_{i_2 j_1}$, with the same nonnegativity proviso, and otherwise do nothing. This describes a random transformation of the original table that results in a new table in the desired set of tables \mathcal{S} . Now repeat the same random transformation on the new table, and so on.

- ▷ In this example, a careful check that the conditions allowing application of the Basic Limit Theorem hold constitutes a challenging exercise, which you are asked to do in Exercise [1.17]. Exercise [1.18] suggests an alternative Markov chain for the same purpose, and Exercise [1.19] introduces a fascinating connection between two problems: generating an approximately uniformly distributed random element of a set, and approximately counting the number of elements in the set. My hope is that these interesting applications of the Basic Limit Theorem are stimulating enough to whet your appetite for digesting the proof of that theorem!



For the proof of the Basic Limit Theorem, we will need one more concept: *recurrence*. Analogously to what we did with the notion of periodicity, we will begin by saying what a recurrent state is, and then show [in Theorem (1.24) below] that recurrence is actually a class property. In particular, in an irreducible Markov chain, either all states are recurrent or all states are *transient*, which means “not recurrent.” Thus, if a chain is irreducible, we can speak of the chain being either recurrent or transient.

The idea of recurrence is this: a state i is recurrent if, starting from the state i at time 0, the chain is sure to return to i eventually. More precisely, define the *first hitting time* T_i of the state i by

$$T_i = \inf\{n > 0 : X_n = i\},$$

and make the following definition.

(1.23) DEFINITION. The state i is **recurrent** if $\mathbb{P}_i\{T_i < \infty\} = 1$. If i is not recurrent, it is called **transient**.

The meaning of recurrence is this: state i is recurrent if, when the Markov chain is started out in state i , the chain is *certain* to return to i at some finite future time. Observe the difference in spirit between this and the definition of “accessible from” [see the paragraph containing (1.18)], which requires only that it be *possible* for the chain to hit a state j . In terms of the first hitting time notation, the definition of “accessible from” may be

restated as follows: for distinct states $i \neq j$, we say that j is accessible from i if and only if $\mathbb{P}_i\{T_j < \infty\} > 0$. [Why did I bother to say “for distinct states $i \neq j$ ”?]

Here is the promised result that implies that recurrence is a class property.

(1.24) THEOREM. *Let i be a recurrent state, and suppose that j is accessible from i . Then in fact all of the following hold:*

- (i) $\mathbb{P}_i\{T_j < \infty\} = 1$;
- (ii) $\mathbb{P}_j\{T_i < \infty\} = 1$;
- (iii) *The state j is recurrent.*

PROOF: The proof will be given somewhat informally; it can be rigorized. Suppose $i \neq j$, since the result is trivial otherwise.

Firstly, let us observe that (iii) follows from (i) and (ii): clearly if (ii) holds [that is, starting from j the chain is certain to visit i eventually] and (i) holds [so that starting from i the chain is certain to visit j eventually], then (iii) must also hold [since starting from j the chain is certain to visit i , after which it will definitely get back to j].

To prove (i), let us imagine starting the chain in state i , so that $X_0 = i$. With probability one, the chain returns at some time $T_i < \infty$ to i . For the same reason, continuing the chain after time T_i , the chain is sure to return to i for a second time. In fact, by continuing this argument we see that, with probability one, the chain returns to i infinitely many times. Thus, we may visualize the path followed by the Markov chain as a succession of infinitely many “cycles,” where a cycle is a portion of the path between two successive visits to i . That is, we’ll say that the first cycle is the segment X_1, \dots, X_{T_i} of the path, the second cycle starts with X_{T_i+1} and continues up to and including the second return to i , and so on. The behaviors of the chain in successive cycles are independent and have identical probabilistic characteristics. In particular, letting $I_n = 1$ if the chain visits j sometime during the n th cycle and $I_n = 0$ otherwise, we see that I_1, I_2, \dots is an *iid* sequence of Bernoulli trials. Let p denote the common “success probability”

$$p = \mathbb{P}\{\text{visit } j \text{ in a cycle}\} = \mathbb{P}_i \left[\bigcup_{k=1}^{T_i} \{X_k = j\} \right]$$

for these trials. Clearly if p were 0, then with probability one the chain would not visit j in any cycle, which would contradict the assumption that j is accessible from i . Therefore, $p > 0$. Now observe that in such a sequence of *iid* Bernoulli trials with a positive success probability, with probability one we will eventually observe a success. In fact,

$$\mathbb{P}_i\{\text{chain does not visit } j \text{ in the first } n \text{ cycles}\} = (1 - p)^n \rightarrow 0$$

as $n \rightarrow \infty$. That is, with probability one, eventually there will be a cycle in which the chain does visit j , so that (i) holds.

It is also easy to see that (ii) must hold. In fact, suppose to the contrary that $\mathbb{P}_j\{T_i = \infty\} > 0$. Combining this with the hypothesis that j is accessible from i , we see that it is

possible with positive probability for the chain to go from i to j in some finite amount of time, and then, continuing from state j , never to return to i . But this contradicts the fact that starting from i the chain must return to i infinitely many times with probability one. Thus, (ii) holds, and we are done. \square

The “cycle” idea used in the previous proof is powerful and important; we will be using it again.

The next theorem gives a useful equivalent condition for recurrence. The statement uses the notation N_i for the total number of visits of the Markov chain to the state i , that is,

$$N_i = \sum_{n=0}^{\infty} I\{X_n = i\}.$$

(1.25) THEOREM. *The state i is recurrent if and only if $\mathbb{E}_i(N_i) = \infty$.*

PROOF: We already know that if i is recurrent, then

$$\mathbb{P}_i\{N_i = \infty\} = 1,$$

that is, starting from i , the chain visits i infinitely many times with probability one. But of course the last display implies that $\mathbb{E}_i(N_i) = \infty$. To prove the converse, suppose that i is transient, so that $q := \mathbb{P}_i\{T_i = \infty\} > 0$. Considering the sample path of the Markov chain as a succession of “cycles” as in the proof of Theorem (1.24), we see that each cycle has probability q of never ending, so that there are no more cycles, and no more visits to i . In fact, a bit of thought shows that N_i , the total number of visits to i [including the visit at time 0], has a geometric distribution with “success probability” q , and hence expected value $1/q$, which is finite, since $q > 0$. \square

(1.26) COROLLARY. *If j is transient, then $\lim_{n \rightarrow \infty} P^n(i, j) = 0$ for all states i .*

PROOF: Supposing j is transient, we know that $\mathbb{E}_j(N_j) < \infty$. Starting from an arbitrary state $i \neq j$, we have

$$\mathbb{E}_i(N_j) = \mathbb{P}_i\{T_j < \infty\} \mathbb{E}_i(N_j \mid T_j < \infty).$$

However, $\mathbb{E}_i(N_j \mid T_j < \infty) = \mathbb{E}_j(N_j)$; this is clear intuitively since, starting from i , if the Markov chain hits j at the finite time T_j , then it “probabilistically restarts” at time T_j . [Exercise: give a formal argument.] Thus, $\mathbb{E}_i(N_j) \leq \mathbb{E}_j(N_j) < \infty$, so that in fact we have $\mathbb{E}_i(N_j) = \sum_{n=1}^{\infty} P^n(i, j) < \infty$, which implies the conclusion of the Corollary. \square

(1.27) EXAMPLE [“A DRUNK MAN WILL FIND HIS WAY HOME, BUT A DRUNK BIRD MAY GET LOST FOREVER,” OR, RECURRENCE AND TRANSIENCE OF RANDOM WALKS]. The quotation is from Yale’s own professor Kakutani, as told by R. Durrett in his probability book. We’ll consider a certain model of a random walk in d dimensions, and show that the walk is recurrent if $d = 1$ or $d = 2$, and the walk is transient if $d \geq 3$.

In one dimension, our random walk is the “simple, symmetric” random walk on the integers, which takes steps of $+1$ and -1 with probability $1/2$ each. That is, letting X_1, X_2, \dots be *iid* taking the values ± 1 with probability $1/2$, we define the position of the random walk at time n to be $S_n = X_1 + \dots + X_n$. What is a random walk in d dimensions? Here is what we will take it to be: the position of such a random walk at time n is

$$S_n = (S_n(1), \dots, S_n(d)) \in \mathbb{Z}^d,$$

where the coordinates $S_n(1), \dots, S_n(d)$ are independent simple, symmetric random walks in \mathbb{Z} . That is, to form a random walk in \mathbb{Z}^d , simply concatenate d independent one-dimensional random walks into a d -dimensional vector process.

Thus, our random walk S_n may be written as $S_n = X_1 + \dots + X_n$, where X_1, X_2, \dots are *iid* taking on the 2^d values $(\pm 1, \dots, \pm 1)$ with probability 2^{-d} each. This might not be the first model that would come to your mind. Another natural model would be to have the random walk take a step by choosing one of the d coordinate directions at random (probability $1/d$ each) and then taking a step of $+1$ or -1 with probability $1/2$. That is, the increments X_1, X_2, \dots would be *iid* taking the $2d$ values

$$(\pm 1, 0, \dots, 0), (0, \pm 1, \dots, 0), \dots, (0, 0, \dots, \pm 1)$$

with probability $1/2d$ each. This is indeed a popular model, and can be analyzed to reach the conclusion “recurrent in $d \leq 2$ and transient in $d \geq 3$ ” as well. But the “concatenation of d independent random walks” model we will consider is a bit simpler to analyze. Also, for all you Brownian motion fans out there, our model is the random walk analog of d -dimensional Brownian motion, which is a concatenation of d independent one-dimensional Brownian motions.

We’ll start with $d = 1$. It is obvious that S_0, S_1, \dots is an irreducible Markov chain. Since recurrence is a class property, to show that every state is recurrent it suffices to show that the state 0 is recurrent. Thus, by Theorem (1.25) we want to show that

$$(1.28) \quad \mathbb{E}_0(N_0) = \sum_n P^n(0, 0) = \infty.$$

But $P^n(0, 0) = 0$ if n is odd, and for even $n = 2m$, say, $P^{2m}(0, 0)$ is the probability that a Binomial($2m, 1/2$) takes the value m , or

$$P^{2m}(0, 0) = \binom{2m}{m} 2^{-2m}.$$

This can be closely approximated in a convenient form by using Stirling’s formula, which says that

$$k! \sim \sqrt{2\pi k} (k/e)^k,$$

where the notation “ $a_k \sim b_k$ ” means that $a_k/b_k \rightarrow 1$ as $k \rightarrow \infty$. Applying Stirling’s formula gives

$$P^{2m}(0, 0) = \frac{(2m)!}{(m!)^2 2^{2m}} \sim \frac{\sqrt{2\pi(2m)} (2m/e)^{2m}}{2\pi m (m/e)^{2m} 2^{2m}} = \frac{1}{\sqrt{\pi m}}.$$

Thus, from the fact that $\sum(1/\sqrt{m}) = \infty$ it follows that (1.28) holds, so that the random walk is recurrent.

Now it's easy to see what happens in higher dimensions. In $d = 2$ dimensions, for example, again we have an irreducible Markov chain, so we may determine the recurrence or transience of chain by determining whether the sum

$$(1.29) \quad \sum_{n=0}^{\infty} \mathbb{P}_{(0,0)}\{S_{2n} = (0,0)\}$$

is infinite or finite, where S_{2n} is the vector (S_{2n}^1, S_{2n}^2) , say. By the assumed independence of the two components of the random walk, we have

$$\mathbb{P}_{(0,0)}\{S_{2m} = (0,0)\} = \mathbb{P}_0\{S_{2m}^1 = 0\}\mathbb{P}_0\{S_{2m}^2 = 0\} \sim \left(\frac{1}{\sqrt{\pi m}}\right) \left(\frac{1}{\sqrt{\pi m}}\right) = \frac{1}{\pi m},$$

so that (1.29) is infinite, and the random walk is again recurrent. However, in $d = 3$ dimensions, the analogous sum

$$\sum_{n=0}^{\infty} \mathbb{P}_{(0,0,0)}\{S_{2n} = (0,0,0)\}$$

is finite, since

$$\mathbb{P}_{(0,0,0)}\{S_{2m} = (0,0,0)\} = \mathbb{P}_0\{S_{2m}^1 = 0\}\mathbb{P}_0\{S_{2m}^2 = 0\}\mathbb{P}_0\{S_{2m}^3 = 0\} \sim \left(\frac{1}{\sqrt{\pi m}}\right)^3,$$

so that in three [or more] dimensions the random walk is transient.

The calculations are simple once we know that in one dimension $\mathbb{P}_0\{S_{2m} = 0\}$ is of order of magnitude $1/\sqrt{m}$. In a sense it is not very satisfactory to get this by using Stirling's formula and having huge exponentially large titans in the numerator and denominator fighting it out and killing each other off, leaving just a humble \sqrt{m} standing in the denominator after the dust clears. In fact, it is easy to guess without any unnecessary violence or calculation that the order of magnitude is $1/\sqrt{m}$ —note that the distribution of S_{2m} , having variance $2m$, is “spread out” over a range of order \sqrt{m} , so that the probabilities of points in that range should be of order $1/\sqrt{m}$. Another way to see the answer is to use a Normal approximation to the binomial distribution. We approximate the Binomial($2m, 1/2$) distribution by the Normal distribution $N(m, m/2)$, with the usual continuity correction:

$$\begin{aligned} \mathbb{P}\{\text{Binomial}(2m, 1/2) = m\} &\sim \mathbb{P}\{m - 1/2 < N(m, m/2) < m + 1/2\} \\ &= \mathbb{P}\{-(1/2)\sqrt{2/m} < N(0, 1) < (1/2)\sqrt{2/m}\} \\ &\sim \phi(0)\sqrt{2/m} = (1/\sqrt{2\pi})\sqrt{2/m} = 1/\sqrt{\pi m}. \end{aligned}$$

Although this calculation does not follow as a direct consequence of the usual Central Limit Theorem, it is an example of a “local Central Limit Theorem.” \square

▷ Do you feel that the 3-dimensional random walk we have considered was not the one you would have naturally defined? Would you have considered a random walk that at each time moved either North or South, or East or West, or Up or Down? Exercise [1.20] shows that this random walk is also transient. The analysis is somewhat more complicated than that for the 3-dimensional random walk we have just considered.

We'll end this section with a discussion of the relationship between recurrence and the existence of a stationary distribution. The results will be useful in the next section.

(1.30) PROPOSITION. Suppose a Markov chain has a stationary distribution π . If the state j is transient, then $\pi(j) = 0$.

PROOF: Since π is stationary, we have $\pi P^n = \pi$ for all n , so that

$$(1.31) \quad \sum_i \pi(i) P^n(i, j) = \pi(j) \quad \text{for all } n.$$

However, since j is transient, Corollary (1.26) says that $\lim_{n \rightarrow \infty} P^n(i, j) = 0$ for all i . Thus, the left side of (1.31) approaches 0 as n approaches ∞ , which implies that $\pi(j)$ must be 0. \square

The last bit of reasoning about equation (1.31) may look a little strange, but in fact $\pi(i) P^n(i, j) = 0$ for all i and n . In light of what we now know, this is easy to see. First, if i is transient, then $\pi(i) = 0$. Otherwise, if i is recurrent, then $P^n(i, j) = 0$ for all n , since if not, then j would be accessible from i , which would contradict the assumption that j is transient.

(1.32) COROLLARY. If an irreducible Markov chain has a stationary distribution, then the chain is recurrent.

PROOF: Being irreducible, the chain must be either recurrent or transient. However, if the chain were transient, then the previous Proposition would imply that $\pi(j) = 0$ for all j , which would contradict the assumption that π is a probability distribution, and so must sum to 1. \square

The previous Corollary says that for an irreducible Markov chain, the existence of a stationary distribution implies recurrence. However, we know that the converse is not true. That is, there are irreducible, recurrent Markov chains that do not have stationary distributions. For example, we have seen that the simple symmetric random walk on the integers in one dimension is irreducible and recurrent but does not have a stationary distribution. This random walk is recurrent all right, but in a sense it is “just barely recurrent.” That is, by recurrence we have $\mathbb{P}_0\{T_0 < \infty\} = 1$, for example, but we also have $\mathbb{E}_0(T_0) = \infty$. The name for this kind of recurrence is **null recurrence**: the state i is null recurrent if it is recurrent and $\mathbb{E}_i(T_i) = \infty$. Otherwise, a recurrent state is called **positive recurrent**: the state i is positive recurrent if $\mathbb{E}_i(T_i) < \infty$. A positive recurrent state i is not just barely recurrent, it is recurrent by a comfortable margin—when started at i , we have not only that T_i is finite almost surely, but also that T_i has finite expectation.

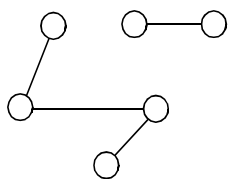
Positive recurrence is in a sense the right notion to relate to the existence of a stationary distribution. For now let me state just the facts, ma'am; these will be justified later. Positive recurrence is also a class property, so that if a chain is irreducible, the chain is either transient, null recurrent, or positive recurrent. It turns out that an irreducible chain has a stationary distribution if and only if it is positive recurrent. That is, strengthening “recurrence” to “positive recurrence” gives the converse to Corollary (1.32).

1.7 An aside on coupling

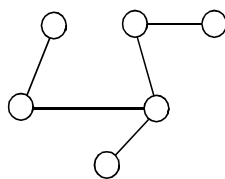
Coupling is a powerful technique in probability. It has a distinctly probabilistic flavor. That is, using the coupling idea entails thinking probabilistically, as opposed to simply applying analysis or algebra or some other area of mathematics. Many people like to prove assertions using coupling and feel happy when they have done so—a probabilistic assertion deserves a probabilistic proof, and a good coupling proof can make obvious what might otherwise be a mysterious statement. For example, we will prove the Basic Limit Theorem of Markov chains using coupling. As I have said before, we could do it using matrix theory, but the probabilist tends to find the coupling proof much more appealing, and I hope you do too.

It is a little hard to give a crisp definition of coupling, and different people vary in how they use the word and what they feel it applies to. Let's start by discussing a very simple example of coupling, and then say something about what the common ideas are.

(1.33) EXAMPLE [CONNECTIVITY OF A RANDOM GRAPH]. A graph is said to be *connected* if for each pair of distinct nodes i and j there is a path from i to j that consists of edges of the graph.



Not connected



Connected

Consider a random graph on a given finite set of nodes, in which each pair of nodes is joined by an edge independently with probability p . We could simulate, or “construct,” such a random graph as follows: for each pair of nodes $i < j$, generate a random number $U_{ij} \sim U[0, 1]$, and join nodes i and j with an edge if $U_{ij} \leq p$. Here is a problem: show that the probability of the resulting graph being connected is nondecreasing in p . That is, for

$p_1 < p_2$, we want to show that

$$\mathbb{P}_{p_1}\{\text{graph connected}\} \leq \mathbb{P}_{p_2}\{\text{graph connected}\}.$$

I would say that this is intuitively obvious, but we want to give an actual *proof*. Again, the example is just meant to illustrate the idea of coupling, not to give an example that can be solved only with coupling!

One way that one might approach this problem is to try to find an explicit expression for the probability of being connected as a function of p . Then one would hope to show that that function is increasing, perhaps by differentiating with respect to p and showing that the derivative is nonnegative.

That is conceptually a straightforward approach, but you may become discouraged at the first step—I don’t think there is an obvious way of writing down the probability the graph is connected. Anyway, doesn’t it seem somehow very inefficient, or at least “overkill,” to have to give a precise expression for the desired probability if all one desires is to show the intuitively obvious monotonicity property? Wouldn’t you hope to give an argument that somehow simply formalizes the intuition that we all have?

One nice way to show that probabilities are ordered is to show that the corresponding events are ordered: if $A \subseteq B$ then $\mathbb{P}A \leq \mathbb{P}B$. So let’s make two events by making two random graphs G_1 and G_2 on the same set of nodes. The graph G_1 is constructed by having each possible edge appear with probability p_1 . Similarly, for G_2 , each edge is present with probability p_2 . We could do this by using two sets of $U[0, 1]$ random variables: $\{U_{ij}\}$ for G_1 and $\{V_{ij}\}$ for G_2 . OK, so now we ask: is it true that

$$(1.34) \quad \{G_1 \text{ connected}\} \subseteq \{G_2 \text{ connected}\}?$$

The answer is no; indeed, the random graphs G_1 and G_2 are independent, so that clearly

$$\mathbb{P}\{G_1 \text{ connected}, G_2 \text{ not connected}\} = \mathbb{P}\{G_1 \text{ connected}\}\mathbb{P}\{G_2 \text{ not connected}\} > 0.$$

The problem is that we have used different, independent random numbers in constructing the graphs G_1 and G_2 , so that, for example, it is perfectly possible to have simultaneously $U_{ij} \leq p_1$ and $V_{ij} > p_2$ for all $i < j$, in which the graph G_1 would be completely connected and the graph G_2 would be completely disconnected.

Here is a simple way to fix the argument: use the *same random numbers* in defining the two graphs. That is, draw the edge (i, j) in graph G_1 if $U_{ij} \leq p_1$ and the edge (i, j) in graph G_2 if $U_{ij} \leq p_2$. Now notice how the picture has changed: with the modified definitions it is obvious that, if an edge (i, j) is in the graph G_1 , then that edge is also in G_2 . From this, it is equally obvious that (1.34) now holds. This establishes the desired monotonicity of the probability of being connected. Perfectly obvious, isn’t it? \square

So, what characterizes a coupling argument? In our example, we wanted to establish a statement about two distributions: the distributions of random graphs with edge probabilities p_1 and p_2 . To do this, we showed how to “construct” [i.e., *simulate* using uniform random numbers!] random objects having the desired distributions in such a way that the desired conclusion became obvious. The trick was to make appropriate use of the same

uniform random variables in constructing the two objects. I think this is a general feature of coupling arguments: somewhere in there you will find the same set of random variables used to construct two different objects about which one wishes to make some probabilistic statement. The term “coupling” reflects the fact that the two objects are related in this way. \square

- ▷ *Exercise [1.24] uses this type of coupling idea, proving a result for one process by comparing it with another process.*

1.8 Proof of the Basic Limit Theorem

The Basic Limit Theorem says that if an irreducible, aperiodic Markov chain has a stationary distribution π , then for each initial distribution π_0 , as $n \rightarrow \infty$ we have $\pi_n(i) \rightarrow \pi(i)$ for all states i . Let me start by pointing something out, just in case the wording of the statement strikes you as a bit strange. Why does the statement read “... *a* stationary distribution”? For example, what if the chain has two stationary distributions? The answer is that this is impossible: the assumed conditions imply that a stationary distribution is in fact unique. In fact, once we prove the Basic Limit Theorem, we will know this to be the case. Clearly if the Basic Limit Theorem is true, an irreducible and aperiodic Markov chain cannot have two different stationary distributions π and $\tilde{\pi}$, since obviously $\pi_n(i)$ cannot approach both $\pi(i)$ and $\tilde{\pi}(i)$ for all i .

An equivalent but conceptually useful reformulation is to define a distance between probability distributions, and then to show that as $n \rightarrow \infty$, the distance between the distribution π_n and the distribution π converges to 0. The notion of distance that we will use is called “total variation distance.”

(1.35) DEFINITION. *Let λ and μ be two probability distributions on the set \mathcal{S} . Then the **total variation distance** $\|\lambda - \mu\|$ between λ and μ is defined by*

$$\|\lambda - \mu\| = \sup_{A \subset \mathcal{S}} [\lambda(A) - \mu(A)].$$

(1.36) PROPOSITION. *The total variation distance $\|\lambda - \mu\|$ may also be expressed in the alternative forms*

$$\|\lambda - \mu\| = \sup_{A \subset \mathcal{S}} |\lambda(A) - \mu(A)| = \frac{1}{2} \sum_{i \in \mathcal{S}} |\lambda(i) - \mu(i)| = 1 - \sum_{i \in \mathcal{S}} \min\{\lambda(i), \mu(i)\}.$$

- ▷ *The proof of this simple Proposition is Exercise [1.25].*

Two probability distributions λ and μ assign probabilities to all possible events. The total variation distance between λ and μ is the largest possible discrepancy between the

probabilities assigned by λ and μ to any event. For example, let π_7 denote the distribution of the ordering of a deck of cards after 7 shuffles, and let π denote the uniform distribution on all $52!$ permutations of the deck, which corresponds to the result of perfect shuffling (or “shuffling infinitely many times”). Suppose, for illustration, that the total variation distance $\|\pi_7 - \pi\|$ happens to be 0.17. This tells us that the probability of any event — for example, the probability of winning any specified card game — using a deck shuffled 7 times differs by at most 0.17 from the probability of the same event using a perfectly shuffled deck.

To introduce the coupling method, let Y_0, Y_1, \dots be a Markov chain with the same probability transition matrix as X_0, X_1, \dots , but let Y_0 have the distribution π ; that is, we start the Y chain off in the initial distribution π instead of the initial distribution π_0 of the X chain. Note that $\{Y_n\}$ is a stationary Markov chain, and, in particular, that Y_n has the distribution π for all n . Further let the Y chain be independent of the X chain.

Roughly speaking, we want to show that for large n , the probabilistic behavior of X_n is close to that of Y_n . The next result says that we can do this by showing that for large n , the X and Y chains will have met with high probability by time n . Define the *coupling time* T to be the first time at which X_n equals Y_n :

$$T = \inf\{n : X_n = Y_n\},$$

where of course we define $T = \infty$ if $X_n \neq Y_n$ for all n .

(1.37) LEMMA [“THE COUPLING INEQUALITY”]. *For all n we have*

$$\|\pi_n - \pi\| \leq \mathbb{P}\{T > n\}.$$

PROOF: Define the process $\{Y_n^*\}$ by

$$Y_n^* = \begin{cases} Y_n & \text{if } n < T \\ X_n & \text{if } n \geq T. \end{cases}$$

It is easy to see that $\{Y_n^*\}$ is a Markov chain, and it has the same probability transition matrix $P(i, j)$ as $\{X_n\}$ has. [To understand this, start by thinking of the X chain as a frog carrying a table of random numbers jumping around in the state space. The frog uses his table of *iid* uniform random numbers to generate his path as we described earlier in the section about specifying and simulating Markov chains. He uses the first number in his table together with his initial distribution π_0 to determine X_0 , and then reads down successive numbers in the table to determine the successive transitions on his path. The Y frog does the same sort of thing, except he uses his own, different table of uniform random numbers so he will be independent of the X frog, and he starts out with the initial distribution π instead of π_0 . How about the Y^* frog? Is he also doing a Markov chain? Well, is he choosing his transitions using uniform random numbers like the other frogs? Yes, he is; the only difference is that he starts by using Y ’s table of random numbers (and hence he follows Y) until the coupling time T , after which he stops reading numbers from Y ’s table and switches to X ’s table. But big deal; he is still generating his path by using

uniform random numbers in the way required to generate a Markov chain.]] The chain $\{Y_n^*\}$ is stationary: $Y_0^* \sim \pi$, since $Y_0^* = Y_0$ and $Y_0 \sim \pi$. Thus, $Y_n^* \sim \pi$ for all n . so that for $A \subseteq \mathcal{S}$ we have

$$\begin{aligned}\pi_n(A) - \pi(A) &= \mathbb{P}\{X_n \in A\} - \mathbb{P}\{Y_n^* \in A\} \\ &= \mathbb{P}\{X_n \in A, T \leq n\} + \mathbb{P}\{X_n \in A, T > n\} \\ &\quad - \mathbb{P}\{Y_n^* \in A, T \leq n\} - \mathbb{P}\{Y_n^* \in A, T > n\}.\end{aligned}$$

However, on the event $\{T \leq n\}$, we have $Y_n^* = X_n$, so that the two events $\{X_n \in A, T \leq n\}$ and $\{Y_n^* \in A, T \leq n\}$ are the same, and hence they have the same probability. Therefore, the first and third terms in the last expression cancel, yielding

$$\pi_n(A) - \pi(A) = \mathbb{P}\{X_n \in A, T > n\} - \mathbb{P}\{Y_n^* \in A, T > n\}.$$

Since the last difference is obviously bounded by $\mathbb{P}\{T > n\}$, we are done. \square

Note the significance of the coupling inequality: it reduces the problem of showing that $\|\pi_n - \pi\| \rightarrow 0$ to that of showing that $\mathbb{P}\{T > n\} \rightarrow 0$, or equivalently, that $\mathbb{P}\{T < \infty\} = 1$. To do this, we consider the “bivariate chain” $\{Z_n = (X_n, Y_n) : n \geq 0\}$. A bit of thought confirms that Z_0, Z_1, \dots is a Markov chain on the state space $\mathcal{S} \times \mathcal{S}$. Since the X and Y chains are independent, the probability transition matrix P_Z of the Z chain can be written as

$$P_Z(i_x i_y, j_x j_y) = P(i_x, j_x)P(i_y, j_y).$$

It is easy to check that the Z chain has stationary distribution

$$\pi_Z(i_x i_y) = \pi(i_x)\pi(i_y).$$

Watch closely now; we’re about to make an important reduction of the problem. Recall that we want to show that $\mathbb{P}\{T < \infty\} = 1$. Stated in terms of the Z chain, we want to show that with probability one, the Z chain hits the “diagonal” $\{(j, j) : j \in \mathcal{S}\}$ in $\mathcal{S} \times \mathcal{S}$ in finite time. To do this, it is sufficient to show that the Z chain is irreducible and recurrent [why?]. However, since we know that the Z chain has a stationary distribution, by Corollary (1.32), to prove the Basic Limit Theorem, it suffices to show that the Z chain is irreducible.

This is, strangely[†], the hard part. This is where the aperiodicity assumption comes in. For example, consider a Markov chain $\{X_n\}$ having the “type A frog” transition matrix $P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ started out in the condition $X_0 = 0$. Then the stationary chain $\{Y_n\}$ starts out in the uniform distribution: probability 1/2 on each state 0,1. The bivariate chain $\{(X_n, Y_n)\}$ is not irreducible: for example, from the state (0,0), we clearly cannot reach the state (0,1). And this ruins everything. For example, if $Y_0 = 1$, which happens with probability 1/2, the X and Y chains can never meet, so that $T = \infty$. Thus, $\mathbb{P}\{T < \infty\} < 1$.

[†]Or maybe not so strangely, in view of Exercise [1.17].

A little number-theoretic result will help us establish irreducibility of the Z chain.

(1.38) LEMMA. *Suppose A is a set of positive integers that is closed under addition and has greatest common divisor (gcd) one. Then there exists an integer N such that $n \in A$ for all $n \geq N$.*

PROOF: First we claim that A contains at least one pair of consecutive integers. To see this, suppose to the contrary that the minimal “spacing” between successive elements of A is $s > 1$. That is, any two distinct elements of A differ by at least s , and there exists an integer n_1 such that both $n_1 \in A$ and $n_1 + s \in A$. Let $m \in A$ be such that s does not divide m ; we know that such an m exists because $\gcd(A) = 1$. Write $m = qs + r$, where $0 < r < s$. Now observe that, by the closure under addition assumption, the two numbers $a_1 = (q+1)(n_1 + s)$ and $a_2 = (q+1)n_1 + m$ are both in A . However, $a_1 - a_2 = s - r \in (0, s)$, which contradicts the definition of s . This proves the claim.

Thus, A contains two consecutive integers, say, c and $c+1$. Now we will finish the proof by showing that $n \in A$ for all $n \geq c^2$. If $c = 0$ this is trivially true, so assume that $c > 0$. We have, by closure under addition,

$$\begin{aligned} c^2 &= (c)(c) \in A \\ c^2 + 1 &= (c-1)c + (c+1) \in A \\ &\vdots \\ c^2 + c - 1 &= c + (c-1)(c+1) \in A. \end{aligned}$$

Thus, $\{c^2, c^2 + 1, \dots, c^2 + c - 1\}$, a set of c consecutive integers, is a subset of A . Now we can add c to all of these numbers to show that the next set $\{c^2 + c, c^2 + c + 1, \dots, c^2 + 2c - 1\}$ of c integers is also a subset of A . Repeating this argument, clearly all integers c^2 or above are in A . \square

Let $i \in \mathcal{S}$, and retain the assumption that the chain is aperiodic. Then since the set $\{n : P^n(i, i) > 0\}$ is clearly closed under addition, and, by the aperiodicity assumption, has greatest common divisor 1, the previous lemma applies to give that $P^n(i, i) > 0$ for all sufficiently large n . From this, for any $i, j \in \mathcal{S}$, since irreducibility implies that $P^m(i, j) > 0$ for some m , it follows that $P^n(i, j) > 0$ for all sufficiently large n .

Now we complete the proof of the Basic Limit Theorem by showing that the chain $\{Z_n\}$ is irreducible. Let $i_x, i_y, j_x, j_y \in \mathcal{S}$. It is sufficient to show, in the bivariate chain $\{Z_n\}$, that $(j_x j_y)$ is accessible from $(i_x i_y)$. To do this, it is sufficient to show that $P_Z^n(i_x i_y, j_x j_y) > 0$ for some n . However, by the assumed independence of $\{X_n\}$ and $\{Y_n\}$,

$$P_Z^n(i_x i_y, j_x j_y) = P^n(i_x, j_x) P^n(i_y, j_y),$$

which, by the previous paragraph, is positive for all sufficiently large n . Of course, this implies the desired result, and we are done.

▷ Exercises [1.27] and [1.28] give you a chance to think about the coupling idea used in this proof.

1.9 A SLLN for Markov chains

The usual Strong Law of Large Numbers for independent and identically distributed (*iid*) random variables says that if X_1, X_2, \dots are *iid* with mean μ , then the average $(1/n) \sum_{t=1}^n X_t$ converges to μ with probability 1 as $n \rightarrow \infty$.

Some fine print: It is possible to have $\mu = +\infty$, and the SLLN still holds. For example, supposing that the random variables X_t take their values in the set of nonnegative integers $\{0, 1, 2, \dots\}$, the mean is defined to be $\mu = \sum_{k=0}^{\infty} k \mathbb{P}\{X_0 = k\}$. This sum could diverge, in which case we define μ to be $+\infty$, and we have $(1/n) \sum_{t=1}^n X_t \rightarrow \infty$ with probability 1.

For example, if X_0, X_1, \dots are *iid* with values in the set \mathcal{S} , then the SLLN tells us that

$$(1/n) \sum_{t=1}^n I\{X_t = i\} \rightarrow \mathbb{P}\{X_0 = i\}$$

with probability 1 as $n \rightarrow \infty$. That is, the fraction of times that the *iid* process takes the value i in the first n observations converges to $\mathbb{P}\{X_0 = i\}$, the probability that any given observation is i .

We will do a generalization of this result for Markov chains. This law of large numbers will tell us that the fraction of times that a Markov chain occupies state i converges to a limit.

It is possible to view this result as a consequence of a more general and rather advanced *ergodic theorem* (see, for example, Durrett's *Probability: Theory and Examples*). However, I do not want to assume prior knowledge of ergodic theory. Also, the result for Markov chains is quite simple to derive as a consequence of the ordinary law of large numbers for *iid* random variables. Although the successive states of a Markov chain are not independent, of course, we have seen that certain features of a Markov chain are independent of each other. Here we will use the idea that the path of the chain consists of a succession of independent “cycles,” the segments of the path between successive visits to a recurrent state. This independence makes the treatment of Markov chains simpler than the general treatment of stationary processes, and it allows us to apply the law of large numbers that we already know.

(1.39) THEOREM. *Let X_0, X_1, \dots be a Markov chain starting in the state $X_0 = i$, and suppose that the state i communicates with another state j . The limiting fraction of time that the chain spends in state j is $1/\mathbb{E}_j T_j$. That is,*

$$\mathbb{P}_i \left\{ \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{t=1}^n I\{X_t = j\} = \frac{1}{\mathbb{E}_j T_j} \right\} = 1.$$

PROOF: The result is easy if the state j is transient, since in that case $\mathbb{E}_j T_j = \infty$ and (with probability 1) the chain visits j only finitely many times, so that

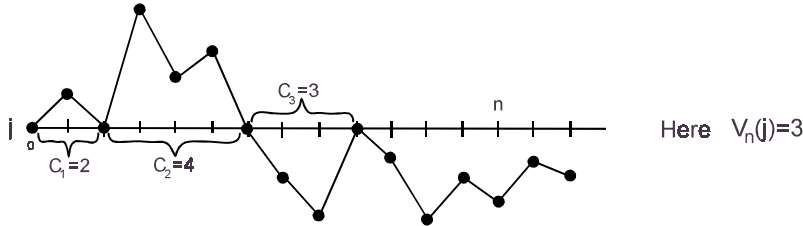
$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{t=1}^n I\{X_t = j\} = 0 = \frac{1}{\mathbb{E}_j T_j}$$

with probability 1. So we assume that j is recurrent. We will also begin by proving the result in the case $i = j$; the general case will be an easy consequence of this special case. Again we will think of the Markov chain path as a succession of *cycles*, where a cycle is a segment of the path that lies between successive visits to j . The cycle lengths C_1, C_2, \dots are *iid* and distributed as T_j ; here we have already made use of the assumption that we are starting at the state $X_0 = j$. Define $S_k = C_1 + \dots + C_k$ and let $V_n(j)$ denote the number of visits to state j made by X_1, \dots, X_n , that is,

$$V_n(j) = \sum_{t=1}^n \{X_t = j\}.$$

A bit of thought [see also the picture below] shows that $V_n(j)$ is also the number of cycles completed up to time n , that is,

$$V_n(j) = \max\{k : S_k \leq n\}.$$



To ease the notation, let V_n denote $V_n(j)$. Notice that

$$S_{V_n} \leq n < S_{V_n+1},$$

and divide by V_n to obtain

$$\frac{S_{V_n}}{V_n} \leq \frac{n}{V_n} < \frac{S_{V_n+1}}{V_n}.$$

Since j is recurrent, $V_n \rightarrow \infty$ with probability one as $n \rightarrow \infty$. Thus, by the ordinary Strong Law of Large Numbers for *iid* random variables, we have both

$$\frac{S_{V_n}}{V_n} \rightarrow \mathbb{E}_j(T_j)$$

and

$$\frac{S_{V_n+1}}{V_n} = \left(\frac{S_{V_n+1}}{V_n+1} \right) \left(\frac{V_n+1}{V_n} \right) \rightarrow \mathbb{E}_j(T_j) \times 1 = \mathbb{E}_j(T_j)$$

with probability one. Note that the last two displays hold whether $\mathbb{E}_j(T_j)$ is finite or infinite. Thus, $n/V_n \rightarrow \mathbb{E}_j(T_j)$ with probability one, so that

$$\frac{V_n}{n} \rightarrow \frac{1}{\mathbb{E}_j T_j}$$

with probability one, which is what we wanted to show.

Next, to treat the general case where i may be different from j , note that $P_i\{T_j < \infty\} = 1$ by Theorem 1.24. Thus, with probability one, a path starting from i behaves as follows. It starts by going from i to j in some finite number T_j of steps, and then proceeds on from state j in such a way that the long run fraction of time that $X_t = j$ for $t \geq T_j$ approaches $1/\mathbb{E}_j(T_j)$. But clearly the long run fraction of time the chain is at j is not affected by the behavior of the chain on the finite segment X_0, \dots, X_{T_j-1} . So with probability one, the long run fraction of time that $X_n = j$ for $n \geq 0$ must approach $1/\mathbb{E}_j(T_j)$. \square

The following result follows directly from Theorem (1.39) by the Bounded Convergence Theorem from the Appendix. [That is, we are using the following fact: if $Z_n \rightarrow c$ with probability one as $n \rightarrow \infty$ and the random variables Z_n all take values in the same bounded interval, then we also have $\mathbb{E}(Z_n) \rightarrow c$. To apply this in our situation, note that we have

$$Z_n := \frac{1}{n} \sum_{t=1}^n I\{X_t = j\} \rightarrow \frac{1}{\mathbb{E}_j T_j}$$

with probability one as $n \rightarrow \infty$, and also each Z_n lies in the interval $[0,1]$. Finally, use the fact that the expectation of an indicator random variable is just the probability of the corresponding event.]

(1.40) COROLLARY. *For an irreducible Markov chain, we have*

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{t=1}^n P^t(i, j) = \frac{1}{\mathbb{E}_j(T_j)}$$

for all states i and j .

There's something suggestive here. Consider for the moment an irreducible, aperiodic Markov chain having a stationary distribution π . From the Basic Limit Theorem, we know that, $P^n(i, j) \rightarrow \pi(j)$ as $n \rightarrow \infty$. However, it is a simple fact that if a sequence of numbers converges to a limit, then the sequence of "Cesaro averages" converges to the same limit; that is, if $a_t \rightarrow a$ as $t \rightarrow \infty$, then $(1/n) \sum_{t=1}^n a_t \rightarrow a$ as $n \rightarrow \infty$. Thus, the Cesaro averages of $P^n(i, j)$ must converge to $\pi(j)$. However, the previous Corollary shows that the Cesaro averages converge to $1/\mathbb{E}_j(T_j)$. Thus, it follows that

$$\pi(j) = \frac{1}{\mathbb{E}_j(T_j)}.$$

It turns out that the aperiodicity assumption is not needed for this last conclusion; we'll see this in the next result. Incidentally, we could have proved this result much earlier; for example we don't need the Basic Limit Theorem in the development.

(1.41) THEOREM. *An irreducible, positive recurrent Markov chain has a unique stationary distribution π given by*

$$\pi(j) = \frac{1}{\mathbb{E}_j(T_j)}.$$

PROOF: For the uniqueness, let π be a stationary distribution. We start with the relation

$$\sum_i \pi(i) P^t(i, j) = \pi(j),$$

which holds for all t . Averaging this over values of t from 1 to n gives

$$\sum_i \pi(i) \frac{1}{n} \sum_{t=1}^n P^t(i, j) = \pi(j).$$

By Corollary 1.40 [and the Dominated Convergence Theorem], the left side of the last equation approaches

$$\sum_i \pi(i) \frac{1}{\mathbb{E}_j(T_j)} = \frac{1}{\mathbb{E}_j(T_j)}$$

as $n \rightarrow \infty$. Thus, $\pi(j) = 1/\mathbb{E}_j(T_j)$, which establishes the uniqueness assertion.

We begin the proof of existence by doing the proof in the special case where the state space is finite. The proof is simpler here than in the general case, which involves some distracting technicalities.

So assume for the moment that the state space is finite. We begin again with Corollary 1.40, which says that

$$(1.42) \quad \frac{1}{n} \sum_{t=1}^n P^t(i, j) \rightarrow \frac{1}{\mathbb{E}_j(T_j)}.$$

However, the sum over all j of the left side of (1.42) is 1, for all n . Therefore,

$$\sum_j \frac{1}{\mathbb{E}_j(T_j)} = 1.$$

That's good, since we want our claimed stationary distribution to be a probability distribution.

Next we write out the matrix equation $P^t P = P^{t+1}$ as follows:

$$(1.43) \quad \sum_k P^t(i, k) P(k, j) = P^{t+1}(i, j).$$

Averaging this over $t = 1, \dots, n$ gives

$$\sum_k \left[\frac{1}{n} \sum_{t=1}^n P^t(i, k) \right] P(k, j) = \frac{1}{n} \sum_{t=1}^n P^{t+1}(i, j).$$

Taking the limit as $n \rightarrow \infty$ of the last equation and using (1.42) again gives

$$\sum_k \left(\frac{1}{\mathbb{E}_k T_k} \right) P(k, j) = \frac{1}{\mathbb{E}_j T_j}.$$

Thus, our claimed stationary distribution is indeed stationary.

Finally, let's see how to handle the infinite state space case. Let $A \subset \mathcal{S}$ be a finite subset of the state space. Summing (1.42) over $j \in A$ gives the inequality

$$\sum_{j \in A} \frac{1}{\mathbb{E}_j(T_j)} \leq 1.$$

Therefore, since this is true for all subsets A , we get

$$\sum_{j \in \mathcal{S}} \frac{1}{\mathbb{E}_j(T_j)} =: C \leq 1.$$

By the assumption of positive recurrence, we have $C > 0$; in a moment we'll see that $C = 1$. The same sort of treatment of (1.43) [i.e., sum over $k \in A$, average over $t = 1, \dots, n$, let $n \rightarrow \infty$, and then take supremum over subsets A of \mathcal{S}] gives the inequality

$$(1.44) \quad \sum_k \left(\frac{1}{\mathbb{E}_k T_k} \right) P(k, j) \leq \frac{1}{\mathbb{E}_j T_j}.$$

However, the sum over all j of the left side of (1.44) is

$$\sum_k \left(\frac{1}{\mathbb{E}_k T_k} \right) \sum_j P(k, j) = \sum_k \left(\frac{1}{\mathbb{E}_k T_k} \right),$$

which is the same as the sum of the right side of (1.44). Thus, the left and right sides of (1.44) must be the same for all j . From this we may conclude that the distribution

$$\tilde{\pi}(j) = \frac{1}{C} \left(\frac{1}{\mathbb{E}_j(T_j)} \right)$$

is stationary, so that, in particular, we know that our chain does have a stationary distribution. Thus, by the uniqueness assertion we proved above, we must have $C = 1$, and we are done. \square

▷ You might like to try Exercise [1.29] at this point. I hope you can play chess.

1.10 Exercises

[1.1] Let X_0, X_1, \dots be a Markov chain, and let A and B be subsets of the state space.

- (a) Is it true that $\mathbb{P}\{X_2 \in B \mid X_1 = x_1, X_0 \in A\} = \mathbb{P}\{X_2 \in B \mid X_1 = x_1\}$? Give a proof or counterexample.
- (b) Is it true that $\mathbb{P}\{X_2 \in B \mid X_1 \in A, X_0 = x_0\} = \mathbb{P}\{X_2 \in B \mid X_1 \in A\}$? Give a proof or counterexample.

[[The moral: be careful about what the Markov property says!]]

[1.2] Let X_0, X_1, \dots be a Markov chain on the state space $\{-1, 0, 1\}$, and suppose that $P(i, j) > 0$ for all i, j . What is a necessary and sufficient condition for the sequence of absolute values $|X_0|, |X_1|, \dots$ to be a Markov chain?

▷ *Exercise [1.3] uses a basic and important technique: conditioning on what happens in the first step of the chain. And then in Exercise [1.4] you get to use this to do something interesting.*

[1.3] Let $\{X_n\}$ be a finite-state Markov chain and let A be a subset of the state space. Suppose we want to determine the expected time until the chain enters the set A , starting from an arbitrary initial state. That is, letting $\tau_A = \inf\{n \geq 0 : X_n \in A\}$ denote the first time to hit A [[defined to be 0 if $X_0 \in A$]], we want to determine $\mathbb{E}_i(\tau_A)$. Show that

$$\mathbb{E}_i(\tau_A) = 1 + \sum_k P(i, k) \mathbb{E}_k(\tau_A)$$

for $i \notin A$.

[1.4] You are tossing a coin repeatedly. Which pattern would you expect to see faster: HH or HT? For example, if you get the sequence TTHHHTH..., then you see “HH” at the 4th toss and “HT” at the 6th. Letting N_1 and N_2 denote the times required to see “HH” and “HT”, respectively, can you guess intuitively whether $\mathbb{E}(N_1)$ is smaller than, the same as, or larger than $\mathbb{E}(N_2)$? Go ahead, make a guess [[and my day]]. Why don’t you also simulate some to see how the answer looks; I recommend a computer, but if you like tossing real coins, enjoy yourself by all means. Finally, you can use the reasoning of the Exercise [1.3] to solve the problem and evaluate $\mathbb{E}(N_i)$. A hint is to set up a Markov chain having the 4 states HH, HT, TH, and TT.

[1.5] Here is a chance to practice formalizing some typical “intuitively obvious” statements. Let X_0, X_1, \dots be a finite-state Markov chain.

- a. We start with an observation about conditional probabilities that will be a useful tool

throughout the rest of this problem. Let F_1, \dots, F_m be disjoint events. Show that if $\mathbb{P}(E|F_i) = p$ for all $i = 1, \dots, m$ then $\mathbb{P}(E | \bigcup_{i=1}^m F_i) = p$.

b. Show that

Typo mistake here -->
on the 2nd line (in the \mathbb{P}_j prob),
the X subscripts should be 1,...,r
instead of n+1,...,n+r

$$\begin{aligned} & \mathbb{P}\{X_{n+1} \in A_1, \dots, X_{n+r} \in A_r \mid X_n = j, X_{n-1} \in B_{n-1}, \dots, X_0 \in B_0\} \\ &= \mathbb{P}_j\{X_{n+1} \in A_1, \dots, X_{n+r} \in A_r\}. \end{aligned}$$

- c. Recall the definition of hitting times: $T_i = \inf\{n > 0 : X_n = i\}$. Show that $\mathbb{P}_i\{T_i = n + m \mid T_j = n, T_i > n\} = \mathbb{P}_j\{T_i = m\}$, and conclude that $\mathbb{P}_i\{T_i = T_j + m \mid T_j < \infty, T_i > T_j\} = \mathbb{P}_j\{T_i = m\}$. This is one manifestation of the statement that the Markov chain “probabilistically restarts” after it hits j .
- d. Show that $\mathbb{P}_i\{T_i < \infty \mid T_j < \infty, T_i > T_j\} = \mathbb{P}_j\{T_i < \infty\}$. Use this to show that if $\mathbb{P}_i\{T_j < \infty\} = 1$ and $\mathbb{P}_j\{T_i < \infty\} = 1$, then $\mathbb{P}_i\{T_i < \infty\} = 1$.
- e. Let i be a recurrent state and let $j \neq i$. Recall the idea of “cycles,” the segments of the path between successive visits to i . For simplicity let’s just look at the first two cycles. Formulate and prove an assertion to the effect that whether or not the chain visits state j during the first and second cycles can be described by *iid* Bernoulli random variables.

- [1.6] [A moving average process] Moving average models are used frequently in time series analysis, economics and engineering. For these models, one assumes that there is an underlying, unobserved process $\dots, Y_{-1}, Y_0, Y_1, \dots$ of *iid* random variables. A **moving average process** takes an average (possibly a weighted average) of these *iid* random variables in a “sliding window.” For example, suppose that at time n we simply take the average of the Y_n and Y_{n-1} , defining $X_n = (1/2)(Y_n + Y_{n-1})$. Our goal is to show that the process X_0, X_1, \dots defined in this way is not Markov. As a simple example, suppose that the distribution of the *iid* Y random variables is $\mathbb{P}\{Y_i = 1\} = 1/2 = \mathbb{P}\{Y_i = -1\}$.

- (a) Show that X_0, X_1, \dots is not a Markov chain.
- (b) Show that X_0, X_1, \dots is not an r th order Markov chain for any finite r .

- [1.7] Let $P^n(i, j)$ denote the (i, j) element in the matrix P^n , the n th power of P . Show that $P^n(i, j) = \mathbb{P}\{X_n = j \mid X_0 = i\}$. Ideally, you should get quite confused about what is being asked, and then straighten it all out.

- [1.8] Consider a Markov chain on the integers with

$$\begin{aligned} P(i, i+1) &= .4 \text{ and } P(i, i-1) = .6 \text{ for } i > 0, \\ P(i, i+1) &= .6 \text{ and } P(i, i-1) = .4 \text{ for } i < 0, \\ P(0, 1) &= P(0, -1) = 1/2. \end{aligned}$$

This is a chain with infinitely many states, but it has a sort of probabilistic “restoring force” that always pushes back toward 0. Find the stationary distribution.

[1.9] Recall the definition the Ehrenfest chain from Example (1.13).

- (a) What is the stationary distribution? You might want to solve the problem for a few small values of d . You should notice a pattern, and come up with a familiar answer.
- (b) Can you explain without calculation why this distribution is stationary? That is, supposing you start the Ehrenfest chain at time 0 by choosing a state according to the distribution that you claim is stationary, you should argue without calculation that the state at time 1 should also have this same distribution.

[1.10] On page 13 we argued that a limiting distribution must be stationary. This argument was clear in the case of a finite state space. For you fans of mathematical analysis, what happens in the case of a countably infinite state space? Can you still make the limiting argument work?

[1.11] Consider a partition of the state space \mathcal{S} of a Markov chain into two complementary subsets A and A^c . Suppose the Markov chain has stationary distribution π . Show that $\text{flux}(A, A^c) = \text{flux}(A^c, A)$. As a hint, here is an outline of steps you might follow.

- (i) Show that the flux function has the following sort of linearity properties: If B and C are disjoint,

$$\begin{aligned}\text{flux}(A, B \cup C) &= \text{flux}(A, B) + \text{flux}(A, C) \\ \text{flux}(B \cup C, A) &= \text{flux}(B, A) + \text{flux}(C, A)\end{aligned}$$

- (ii) Show that $\text{flux}(\mathcal{S}, \{k\}) = \text{flux}(\{k\}, \mathcal{S})$ for all singleton sets $\{k\}$.
- (iii) Using the first two steps, show that $\text{flux}(\mathcal{S}, A) = \text{flux}(A, \mathcal{S})$.
- (iv) By subtracting a certain flux quantity from both sides, conclude that $\text{flux}(A, A^c) = \text{flux}(A^c, A)$.

[1.12] Show by example that for general subsets A and B , the equality $\text{flux}(A, B) = \text{flux}(B, A)$ does not necessarily hold.

[1.13] Use Exercise [1.11] to re-do Exercise [1.9], by writing the equations produced by (1.15) with the choice $A = \{0, 1, \dots, i\}$ for various i . The calculation should be easier.

[1.14] [Renewal theory, the residual, and length-biased sampling] Let X_1, X_2, \dots be *iid* taking values in $\{1, \dots, d\}$. You might, for example, think of these random variables as lifetimes of light bulbs. Define $S_k = X_1 + \dots + X_k$, $\tau(n) = \inf\{k : S_k \geq n\}$, and $R_n = S_{\tau(n)} - n$. Then R_n is called the *residual lifetime* at time n . This is the amount of lifetime remaining in the light bulb that is in operation at time n .

- (a) The sequence R_0, R_1, \dots is a Markov chain. What is its transition matrix? What is the stationary distribution?

- (b) Define the *total lifetime* L_n at time n by $L_n = X_{\tau(n)}$. This is the total lifetime of the light bulb in operation at time n . Show that L_0, L_1, \dots is not a Markov chain. But L_n still has a limiting distribution, and we'd like to find it. We'll do this by constructing a Markov chain by enlarging the state space and considering the sequence of random vectors $(R_0, L_0), (R_1, L_1), \dots$. This sequence does form a Markov chain. What is its probability transition function and stationary distribution? Now, assuming the Basic Limit Theorem applies here, what is the limiting distribution of L_n as $n \rightarrow \infty$? This is the famous "length-biased sampling" distribution.
- [1.15] Show that the relation "communicates with" is an equivalence relation. That is, show that the "communicates with" relation is reflexive, symmetric, and transitive.
- [1.16] Show that if an irreducible Markov chain has a state i such that $P(i, i) > 0$, then the chain is aperiodic. Also show by example that this sufficient condition is not necessary.
- [1.17] [[Generating a random 4×4 table of numbers satisfying given restrictions]] Show that if we run the process described in Example (1.22) for a sufficiently long time, then we will end up with a random table having probability distribution arbitrarily close to the desired distribution (that is, uniform on \mathcal{S}). In order to do this, you need to demonstrate that the conditions of the Basic Limit Theorem are satisfied in this example, by showing that
- (a) The procedure generates a Markov chain whose state space is \mathcal{S} ,
 - (b) that Markov chain is irreducible,
 - (c) that Markov chain is aperiodic, and
 - (d) that Markov chain has the desired distribution as its stationary distribution.
- [1.18] [[More on 4×4 tables]] Refer to the description of the Markov chain in Example (1.22). Imagine that we have already chosen a random pair of rows $\{i_1, i_2\}$ and a random pair of columns $\{j_1, j_2\}$. The Markov chain described in Example (1.22) takes very small steps, adding ± 1 to $a_{i_1 j_1}$ and $a_{i_2 j_2}$, and subtracting ± 1 from $a_{i_1 j_2}$ and $a_{i_2 j_1}$, when doing so produces no negative entries. We could make larger changes by choosing uniformly from all possible modifications of the form: add m to $a_{i_1 j_1}$ and $a_{i_2 j_2}$, and subtract m from $a_{i_1 j_2}$ and $a_{i_2 j_1}$, where m is any integer that does not cause any table entries to become negative. Describe in a more explicit way (explicit enough to make it clear how to write a computer program to do this) how to run this Markov chain. Show that the Basic Limit Theorem applies here to guarantee convergence to the uniform distribution on \mathcal{S} . If you feel inspired and/or your instructor asks you to do so, simulate this chain in our example and show the world some random tables from \mathcal{S} .
- [1.19] [[A computing project: Approximate counting]] In Example (1.22), we don't know the cardinality of the state space, $\#(\mathcal{S})$. How many such tables are there? About a million? A billion? A trillion? Hey, we don't even know approximately *how many digits* the cardinality has! In some problems there is a nice connection between being able to generate a nearly

uniformly distributed element of a set and the problem of approximating the number of elements in the set. You can try the idea out in the setting of Example (1.22). This is stated in a somewhat open-ended way; there are many variations in how you might approach this, some more or less efficient than the others, and there will be lots of details to work out. The basic idea of the connection between random generation and approximate counting is use the approximate uniform generation to reduce the original approximate counting problem recursively to smaller and smaller counting problems. For example, suppose we knew the fraction, f_{11} , of elements of \mathcal{S} that have a “68” as their (1,1) [upper left-hand corner] entry. Then we have reduced the problem to counting a smaller set, namely, the subset $\mathcal{S}_{11} = \{a \in \mathcal{S} : a_{11} = 68\}$ of \mathcal{S} meeting this additional restriction, because $\#(\mathcal{S}) = \#(\mathcal{S}_{11})/f_{11}$. How do we estimate f_{11} ? Well, f_{11} is the probability of a uniformly distributed $A \in \mathcal{S}$ satisfying the extra restriction $A_{11} = 68$. Now you see where the uniform generation comes in: you can estimate f_{11} by generating many nearly uniformly distributed tables from \mathcal{S} and taking the fraction of those that have “68” in their upper left corner. The same idea may be applied recursively in this example. Estimating $\#(\mathcal{S}_{11})$ involves adding an extra restriction, say on the (1,2) entry of the table, which defines a further subset $\mathcal{S}_{11,12}$ of \mathcal{S}_{11} . Estimating the fraction $\#(\mathcal{S}_{11,12})/\#(\mathcal{S}_{11})$ involves running a Markov chain in the smaller state space \mathcal{S}_{11} . And so on.

Note: as a practical matter and to preserve your sanity, before applying your methodology to the original large problem, it’s a good idea to test it on some much smaller version of the problem (smaller than a 4×4 table) where you know the answer.

- [1.20] [The other 3-dimensional random walk] Consider a random walk on the 3-dimensional integer lattice; at each time the random walk moves with equal probability to one of the 6 nearest neighbors, adding or subtracting 1 in just one of the three coordinates. Show that this random walk is transient.

Hint: You want to show that some series converges. An upper bound on the terms will be enough. How big is the largest probability in the Multinomial($n; 1/3, 1/3, 1/3$) distribution?

▷ Here are three additional problems about a simple symmetric random walk $\{S_n\}$ in one dimension starting from $S_0 = 0$ at time 0.

- [1.21] Let a and b be integers with $a < 0 < b$. Defining the hitting times $\tau_c = \inf\{n \geq 0 : S_n = c\}$, show that the probability $\mathbb{P}\{\tau_b < \tau_a\}$ is given by $(0 - a)/(b - a)$.
- [1.22] Let S_0, S_1, \dots be a simple, symmetric random walk in one dimension as we have discussed, with $S_0 = 0$. Show that

$$\mathbb{P}\{S_1 \neq 0, \dots, S_{2n} \neq 0\} = \mathbb{P}\{S_{2n} = 0\}.$$

Now you can do a calculation that explains why the expected time to return to 0 is infinite.

- [1.23] As in the previous exercise, consider a simple, symmetric random walk started out at 0. Letting $k \neq 0$ be any fixed state, show that the expected number of times the random walk visits state k before returning to state 0 is 1.
- [1.24] Consider a Markov Chain on the nonnegative integers $\mathbb{S} = \{0, 1, 2, \dots\}$. Defining $P(i, i + 1) = p_i$ and $P(i, i - 1) = q_i$, assume that $p_i + q_i = 1$ for all $i \in \mathbb{S}$, and also $p_0 = 1$, and $0 < p_i \leq 1/2$ for all $i \geq 1$. Use what you know about the simple, symmetric random walk to show that the given Markov chain is recurrent.
- [1.25] Prove Proposition (1.36).
- [1.26] Let π_0 and ρ_0 be probability mass functions on \mathbb{S} , and define $\pi_1 = \pi_0 P$ and $\rho_1 = \rho_0 P$, where P is a probability transition matrix. Show that $\|\pi_1 - \rho_1\| \leq \|\pi_0 - \rho_0\|$. That is, in terms of total variation distance, π_1 and ρ_1 are closer to each other than π_0 and ρ_0 were.
- [1.27] Here is a little practice with the coupling idea as used in the proof of the Basic Limit Theorem. Consider a Markov chain $\{X_n\}$ having probability transition matrix

$$P = \begin{pmatrix} 1/2 & 1/4 & 1/4 \\ 1/4 & 1/2 & 1/4 \\ 1/4 & 1/4 & 1/2 \end{pmatrix}.$$

Note that $\{X_n\}$ has stationary distribution $\pi = (1/3, 1/3, 1/3)$. Using the kind of coupling we did in the proof of the Basic Limit Theorem, show that, no matter what the initial distribution π_0 of X_0 is, we have

$$\|\pi_n - \pi\| \leq \frac{2}{3} \left(\frac{11}{16} \right)^n$$

for all n .

- [1.28] Do you think the bound you just derived in Exercise [1.27] is a good one? In particular, is $11/16$ the smallest we can get, or can we do better? What is the actual rate of geometric decrease of $\|\pi_n - \pi\|$? You could think about this in your head, investigate numerically by matrix multiplication, or both.

[[Hint about coupling: Try to think of a more “aggressive” coupling to get a better bound. What does this mean? The coupling we used in the proof of the Basic Limit Theorem was not very aggressive, in that it let the two chains evolve independently until they happened to meet, and only then started to use the same uniform random numbers to generate the paths. No attempt was made to get the chains together as fast as possible. A more aggressive coupling would somehow make use of some random numbers in common to both chains in generating their paths right from the beginning.]]

- [1.29] Consider a knight sitting on the lower left corner square of an ordinary 8×8 chess board. The knight has residual frog-like tendencies, left over from an old spell an older witch cast

upon him. So he performs a random walk on the chess board, at each time choosing a random move uniformly distributed over the set of his possible knight moves. What is the expected time until he first returns to the lower left corner square?

- [1.30] Recall the definition of positive recurrence on page 24. Show that positive recurrence is a class property.
- [1.31] Suppose a Markov chain has a stationary distribution π and the state j is null recurrent. Show that $\pi(j) = 0$.
- [1.32] **[[Birth-collapse chain]]** Consider a Markov chain on $\mathcal{S} = \{0, 1, 2, \dots\}$ having $P(i, i+1) = p_i$, $P(i, 0) = 1 - p_i$ for all i , with $p_0 = 1$ and $0 < p_i < 1$ for all $i > 0$. Show that
- (i) The chain is recurrent if and only if $\lim_{n \rightarrow \infty} \prod_{i=1}^n p_i = 0$. **[[This, in turn, is equivalent to the condition $\sum_{i=1}^{\infty} (1 - p_i) = \infty$. (This was just for interest; not a problem or a hint.)]]**
 - (ii) The chain is positive recurrent if and only if $\sum_{n=1}^{\infty} \prod_{i=1}^n p_i < \infty$.
 - (iii) What is the stationary distribution if $p_i = 1/(i+1)$?
- [1.33] Consider an irreducible Markov chain $\{X_0, X_1, \dots\}$ on a state space having $n < \infty$ states. Let π denote the stationary distribution of the chain, and suppose X_0 is distributed according to π . Define τ to be the first return time to the initial state, that is, $\tau = \inf\{k > 0 : X_k = X_0\}$. What is the expectation of τ ?

2. More on Markov chains, Examples and Applications

Section 1. Branching processes.

Section 2. Time reversibility.

Section 3. Application of time reversibility: a tandem queue model.

Section 4. The Metropolis method.

Section 5. Simulated annealing.

Section 6. Ergodicity concepts for time-inhomogeneous Markov chains.

Section 7. Proof of the main theorem of simulated annealing.

Section 8. Card shuffling: speed of convergence to stationarity.

2.1 Branching Processes

The branching process model we will study was formulated in 1873 by Sir Francis Galton,^{*} who was interested in the survival and extinction of family names. Suppose children inherit their fathers' names, so we need only keep track of fathers and sons. Consider a male who is the only member of his generation to have a given family name, so that the responsibility of keeping the family name alive falls upon him—if his line of male descendants terminates, so does the family name. Suppose for simplicity that each male has probability $f(0)$ of producing no sons, $f(1)$ of producing one son, and so on. Here is a question: What is the probability that the family name eventually becomes extinct?

Galton brought the problem to his mathematician friend, Rev. H. W. Watson, who devised the method of analysis using probability generating functions that is still used today. However, a minor mathematical slip caused Galton and Watson to get the answer to the main question wrong. They believed that the extinction probability is 1 — all names are doomed to eventual extinction. We will see below that this is false: if the expected number of sons is greater than 1, the branching process model produces lines of descent that have positive probability of going on forever.

Let us begin with a more formal description of the branching process. Thinking of G_t as the number of males in generation t , start with $G_0 = 1$. If $G_t = i$ then write $G_{t+1} = X_{t1} + X_{t2} + \cdots + X_{ti}$; here X_{tj} denotes the number of sons fathered by the j th man in generation t . Assume the random variables $\{X_{tj} : t \geq 0, j \geq 1\}$ are *iid* with probability mass function f , so that $\mathbb{P}\{X_{tj} = k\} = f(k)$ for $k = 0, 1, \dots$. To avoid trivial cases we

^{*}See Jagers (1975) and Guttorp (1991) for more on the history.

assume that $f(0) > 0$ and $f(0) + f(1) < 1$. [Why are these trivial?] We are interested in the extinction probability $\rho = \mathbb{P}_1\{G_t = 0 \text{ for some } t\}$.

It is clear from the verbal description of the process that $\{G_t : t \geq 0\}$ is a Markov chain. We can say a few interesting things about the process directly from general results of the previous chapter. Clearly state 0 is absorbing. Therefore, for each $i > 0$, since $\mathbb{P}_i\{G_1 = 0\} = (f(0))^i > 0$, the state i must be transient—this follows from Theorem (1.24). Consequently, we know that with probability 1, each state $i > 0$ is visited only a finite number of times. From this, a bit of thought shows that, with probability 1, the chain must either get absorbed at 0 eventually or approach ∞ . [EXERCISE: *Think a bit and show this.*]

We can obtain an equation for ρ by the idea we have used before a number of times—e.g. see exercise ([1.3])—namely, conditioning on what happens at the first step of the chain. This gives

$$\rho = \sum_{k=0}^{\infty} \mathbb{P}\{G_1 = k \mid G_0 = 1\} \mathbb{P}\{\text{eventual extinction} \mid G_1 = k\}.$$

Evidently, since the males all have sons independently (in the terminology above, the random variables X_{tj} are independent), we have $\mathbb{P}\{\text{eventual extinction} \mid G_1 = k\} = \rho^k$. This holds because the event of eventual extinction, given $G_1 = k$, requires each of the k male lines starting at time 1 to reach eventual extinction; this is the intersection of k independent events, each of which has probability ρ . Thus, ρ satisfies

$$(2.1) \quad \rho = \sum_{k=0}^{\infty} f(k) \rho^k =: \psi(\rho).$$

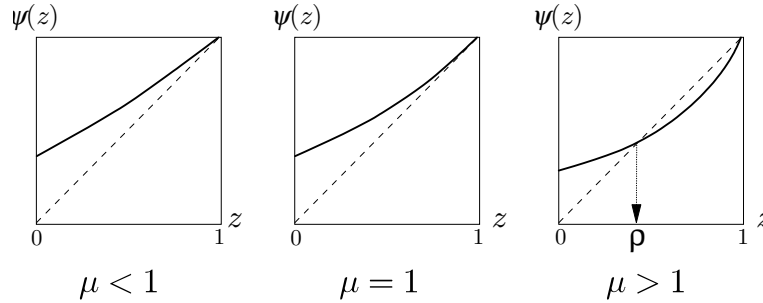
The last sum is a function of ρ ; for each distribution f there is a corresponding function of ρ , which we have denoted by ψ . So ρ satisfies $\psi(\rho) = \rho$: the extinction probability ρ is a fixed point of ψ .

So the function ψ , which is called the *probability generating function* of the probability mass function f , arises in a natural and interesting way in this problem. Let us pause to collect a few of its properties. The first two derivatives are given by

$$\psi'(z) = \sum_{k=1}^{\infty} k f(k) z^{k-1}, \quad \psi''(z) = \sum_{k=2}^{\infty} k(k-1) f(k) z^{k-2}$$

for $z \in (0, 1)$. Since these are positive, the function ψ is strictly increasing and convex on $(0, 1)$. Also, clearly $\psi(0) = f(0)$ and $\psi(1) = 1$. Finally, notice that $\psi'(1) = \sum k f(k) = \mu$, where μ denotes $E(X)$, the expected number of sons for each male.

These properties imply that the graph of ψ over $[0, 1]$ must look like one of the three following pictures, depending on the value of $\mu = \psi'(1)$.



So you can see what happens. Since $\psi(1) = 1$, the equation $\psi(\rho) = \rho$ always has a trivial solution at $\rho = 1$. When $\mu \leq 1$, this trivial solution is the only solution, so that, since the probability ρ of eventual extinction satisfies $\psi(\rho) = \rho$, it must be the case that $\rho = 1$. When $\mu > 1$, there is one additional solution, indicated by the arrow in the picture. This solution was missed by Watson and Galton (1875), leading them to believe that the probability of extinction would be 1 in this case as well. We will show that this was incorrect, and that the probability of extinction is the smaller solution of the equation $\psi(\rho) = \rho$.

Thus, assuming $\mu > 1$ and defining r to be the smaller solution of $\psi(r) = r$, we want to show that $\rho = r$. Since $\psi(\rho) = \rho$, we know that ρ must be either r or 1. Defining $p_t = \mathbb{P}_1\{G_t = 0\}$, observe that as $t \rightarrow \infty$,

$$p_t \uparrow \mathbb{P}_1 \left[\bigcup_{n=1}^{\infty} \{G_n = 0\} \right] = \rho.$$

Therefore, to rule out the possibility that $\rho = 1$, it is sufficient to prove the following statement:

$$p_t \leq r \text{ for all } t.$$

To prove this by induction, observe that $p_0 = 0$, so that the statement holds for $t = 0$. Next observe that

$$p_{t+1} = \mathbb{P}_1\{G_{t+1} = 0\} = \sum_{i=0}^{\infty} \mathbb{P}_1\{G_1 = i\} \mathbb{P}_1\{G_{t+1} = 0 \mid G_1 = i\} = \sum_{i=0}^{\infty} f(i)(p_t)^i,$$

that is, $p_{t+1} = \psi(p_t)$. Thus, using the induction hypothesis $p_t \leq r$ and the fact that the function ψ is increasing, we obtain $p_{t+1} \leq \psi(r) = r$, which completes the proof.

(2.2) EXAMPLE. Suppose each man has 3 children, with each child having probability $1/2$ of being male, and different children being independent. What is the probability that a particular man's line of male descendants will eventually become extinct? Here the distribution f is the binomial distribution $\text{Bin}(3, 1/2)$, so that $\mu = 3/2 > 1$. Thus, we know that the probability ρ of extinction is less than 1. Here $f(0) = 1/8$, $f(1) = 3/8$, $f(2) = 3/8$,

and $f(3) = 1/8$, so that the equation $\psi(r) = r$ becomes

$$\frac{1}{8} + \frac{3}{8}r + \frac{3}{8}r^2 + \frac{1}{8}r^3 = r,$$

or $r^3 + 3r^2 - 5r + 1 = 0$. Fortunately, $r = 1$ is a solution (as it must be!), so we can factor it out, getting the equation $(r - 1)(r^2 + 4r - 1) = 0$. Solving the quadratic equation gives $\rho = \sqrt{5} - 2 = 0.2361$. The man can rest assured that with probability $1 - \rho = 0.7639$ his glorious family name will live on forever. \square

2.2 Time Reversibility

Let X_0, X_1, \dots be a Markov chain having probability transition matrix $P = P(i, j)$. Imagine that I recorded a movie of the sequence of states (X_0, \dots, X_n) , and I am showing you the movie on my fancy machine that can play the tape forward or backward equally well. Can you tell by watching the sequence of transitions on the movie whether I am showing it forward or backward?

Of course, we are assuming that you know the transition matrix P ; otherwise, this would be an unreasonable request. There are cases in which distinguishing the direction of time is very easy. For example, if the state space is $\{1, 2, 3\}$ and $P(1, 2) = P(2, 3) = P(3, 1) = 1$ [one of our standard periodic examples], observing just one transition of the chain is enough to tell you for sure the direction of time; for example, a “movie” in which we observe 3 followed by 2 must be running backward.

That one was easy. Let’s consider another example: do you think a stationary Ehrenfest chain is time-reversible? Here the state space is $\{0, 1, \dots, d\}$, say, and $X_0 \sim \text{Bin}(d, 1/2)$, the stationary distribution of the chain. It is clear in this case that you will not be able to tell *for sure* from observing any finite movie (X_0, \dots, X_n) which direction the movie is being shown—a sequence has positive probability if and only if its reversal also has positive probability. But we are asking whether or not you can get *any sort of probabilistic hint* about the direction in which the movie is being shown, and I am willing to show you as long a segment as you would like to request. So you can have plenty of data to look at. One might suspect that it should be possible to make this sort of distinction. For example, we know that the Ehrenfest chain has a “restoring force” that pulls it toward the level $d/2$, where half the balls are in each of the two urns. So, for instance, if we observe a long sequence that moves from $(3/4)d$ down toward $d/2$, we might favor the explanation that the movie is being shown forward, since otherwise we are observing a long sequence moving against the restoring force.

Did you buy that? I hope not, because in fact we will see that the Ehrenfest chain is time-reversible: no movie, no matter how long, will give you any probabilistic information that is useful in distinguishing the direction of time. [And the argument suggested above really didn’t make much sense — what comes down must have gone up.]

Here is a definition that captures the concept.

(2.3) DEFINITION. We say that a Markov chain $\{X_n\}$ is time-reversible if, for each n ,

$$(X_0, X_1, \dots, X_n) \stackrel{\mathcal{D}}{=} (X_n, X_{n-1}, \dots, X_0)$$

that is, the joint distribution of (X_0, X_1, \dots, X_n) is the same as the joint distribution of $(X_n, X_{n-1}, \dots, X_0)$.

Suppose $\{X_n\}$ is time-reversible. As a particular consequence of the definition, we see that $(X_0, X_1) \stackrel{\mathcal{D}}{=} (X_1, X_0)$. This, in turn, implies that $X_1 \stackrel{\mathcal{D}}{=} X_0$, that is, $\pi_1 = \pi_0$. Thus, in view of the relation $\pi_1 = \pi_0 P$, we obtain $\pi_0 = \pi_0 P$, so that the initial distribution π_0 is stationary. Not surprisingly, we have found that a time-reversible chain must be stationary.

We will write π for the distribution π_0 to emphasize that it is stationary. So $X_n \sim \pi$ for all n . The condition $(X_0, X_1) \stackrel{\mathcal{D}}{=} (X_1, X_0)$ says that $\mathbb{P}\{X_0 = i, X_1 = j\} = \mathbb{P}\{X_1 = i, X_0 = j\}$ for all i, j ; that is,

$$(2.4) \quad \pi(i)P(i, j) = \pi(j)P(j, i) \quad \text{for all } i, j.$$

We have shown that the condition (2.4) together with $X_0 \sim \pi$ is necessary for a chain to be reversible. In fact, these two conditions are also sufficient for reversibility.

(2.5) PROPOSITION. The Markov chain $\{X_n\}$ is time-reversible if and only if the distribution π of X_0 satisfies $\pi P = \pi$ and the condition (2.4) holds.

To see this, observe that (2.4) gives, for example,

$$\begin{aligned} \mathbb{P}\{X_0 = i, X_1 = j, X_2 = k\} &= [\pi(i)P(i, j)]P(j, k) \\ &= [P(j, i)\pi(j)]P(j, k) \\ &= P(j, i)[\pi(j)P(j, k)] \\ &= P(j, i)[P(k, j)\pi(k)] \\ &= \pi(k)P(k, j)P(j, i) \\ &= \mathbb{P}\{X_0 = k, X_1 = j, X_2 = i\} \\ &= \mathbb{P}\{X_2 = i, X_1 = j, X_0 = k\}, \end{aligned}$$

that is, $(X_0, X_1, X_2) \stackrel{\mathcal{D}}{=} (X_2, X_1, X_0)$. Notice how (2.4) allowed the π factor to propagate through the product from the left end to the right, reversing the direction of all of the transitions along the way. The same trick allows us to deduce the general equality required in the definition (2.3).

The equalities in (2.4) have a nice interpretation in terms of probability flux. Recall [as discussed in one of your homework problems] that the flux from i to j is defined as $\pi(i)P(i, j)$. So (2.4) says that the flux from i to j is the same as the flux from j to i —flux balances between each pair of states. These are called the “detailed balance” (or “local balance”) equations; they are more detailed than the “global balance equations” $\pi(j) = \sum_i \pi(i)P(i, j)$ that characterize stationarity. Global balance, which can be rewritten as $\sum_i \pi(j)P(j, i) = \sum_i \pi(i)P(i, j)$ says that the total flux leading out of state j is the same

as the total flux into state j . If we think of a system of containers of fluid connected by tubes, one container for each state, and we think of probability flux as fluid flowing around the system, global balance says that the flow out of container j is balanced by the flow into j , so that the fluid level in container j stays constant, neither rising nor falling. This is a less stringent requirement than detailed balance, which requires flow to balance between each pair of containers.

A more probabilistic interpretation is this: think of $\pi(i)P(i, j)$ as the limiting long run fraction of transitions made by the Markov chain that go from state i to state j . Time reversibility requires that the long run fraction of i -to- j transitions is the same as that of the j -to- i transitions, for all i and j . This is a more stringent requirement than stationarity, which equates the long run fraction of transitions that go out of state i to the long run fraction of transitions that go into state i .

The mathematical formulation of this relationship is simple.

(2.6) PROPOSITION. *If the local balance equations (2.4) hold, then the distribution π is stationary.*

PROOF: Summing the local balance equations (2.4) over i gives the global balance equations

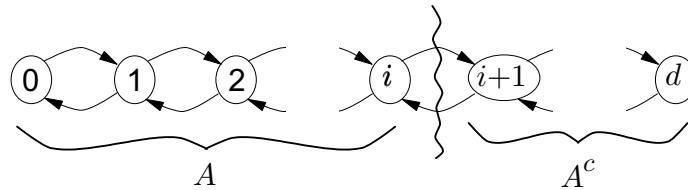
$$\sum_i \pi(i)P(i, j) = \sum_i \pi(j)P(j, i) = \pi(j).$$

□

So why is the Ehrenfest chain time-reversible? The Ehrenfest chain is an example of a *birth and death chain*, which is defined to be a Markov chain whose states consist of nonnegative integers and whose transitions increase or decrease the state by at most 1. That is, interpreting the current state of the chain as the population count of living individuals, the population can change by at most 1 in a transition, which might represent a birth, a death, or no change. The time reversibility of the Ehrenfest chain is an example of a more general fact.

(2.7) CLAIM. *All stationary birth and death chains are reversible.*

To show this, consider a stationary birth and death chain on the state space $\mathcal{S} = \{0, 1, \dots, d\}$.



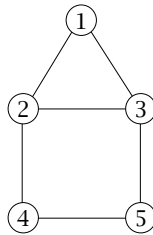
We ask: does $\pi(i)P(i, j) = \pi(j)P(j, i)$ for all i, j ? Since a birth and death chain has $P(i, j) = 0$ if $|i - j| > 1$, we need only consider the case where $j = i + 1$. Recall from

Exercise [1.11] (the exercise on probability flux) that for any subset $A \subset \mathcal{S}$, the flux from A to A^c must equal the flux from A^c to A . Taking $A = \{0, \dots, i\}$ as in the picture gives just what we want: $\pi(i)P(i, i+1) = \pi(i+1)P(i+1, i)$. This establishes the claim. \square

(2.8) EXAMPLE. Another important class of examples of time-reversible Markov chains is the **random walk on a graph**. Defining $d(i)$ to be the degree of node i , the random walk moves according to the matrix $P(i, j) = 1/(d(i))$ for each neighbor j of node i , and $P(i, j) = 0$ otherwise. Then it is easy to check that the distribution

$$\pi(i) = \frac{d(i)}{\sum_{j \in \mathcal{S}} d(j)}$$

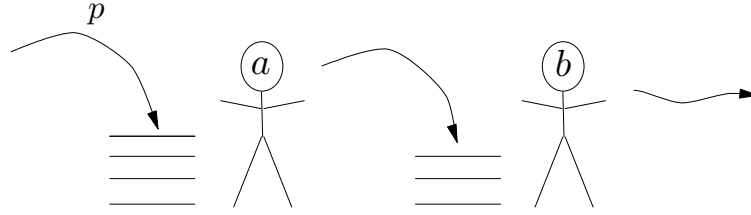
satisfies the detailed balance equations. Thus, the random walk is time-reversible, and π is its stationary distribution.



For example, consider a random walk on the house graph above. The degrees are $(d_1, d_2, d_3, d_4, d_5) = (2, 3, 3, 2, 2)$. So the stationary distribution is $(\pi_1, \pi_2, \pi_3, \pi_4, \pi_5) = (2/12, 3/12, 3/12, 2/12, 2/12)$. \square

2.3 More on Time Reversibility: A Tandem Queue Model

Consider two office workers Andrew and Bertha who have a lot of paper work to do. When a piece of paper arrives at the office, it goes first to Andrew for processing. When he completes his task, he puts the paper on Bertha's pile. When she completes her processing of the paper, it is sent out of the office.

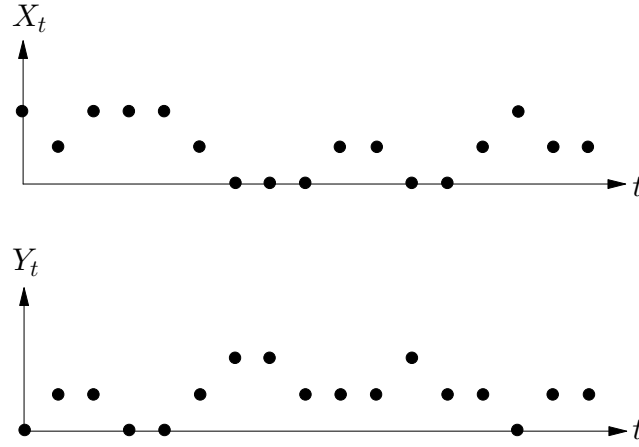


Let's specify a model of this in more detail. Consider a Markov chain whose state at time n is (X_n, Y_n) , where X_n is the number of papers in Andrew's work pile and Y_n is the number of papers in Bertha's pile. Suppose that $X_n = i \geq 0$. Then with probability p , a new piece of work enters the office, resulting in $X_{n+1} = i + 1$. For definiteness, it is helpful to paint a detailed picture, so let's suppose that any new arrival is placed onto the *top* of Andrew's current pile of papers, preempting any other task he might have been working on at the time. Thus, if a new paper arrives, Andrew cannot complete his processing of any previously received papers that period. If Andrew receives no new arrival in a period and $i > 0$, then with probability a Andrew completes the processing of the paper on top of his pile, resulting in $X_{n+1} = i - 1$. Thus, in summary, Andrew's pile evolves as follows. Given $X_n = i > 0$,

$$X_{n+1} = \begin{cases} i + 1 & \text{with probability } p \\ i & \text{with probability } (1 - p)(1 - a) \\ i - 1 & \text{with probability } (1 - p)a, \end{cases}$$

and if $X_n = 0$, then $X_{n+1} = 1$ with probability p and $X_{n+1} = 0$ with probability $1 - p$.

Let us assume that Bertha's pile changes in the same way as Andrew's, except that she gets her new work from Andrew's completed papers rather than from the outside, and her service-completion probability is b rather than a . A sample path of the $\{(X_n, Y_n)\}$ process is shown in the picture. Notice that in each period in which the X process decreases, the Y process increases: work completed by Andrew is sent to Bertha.



The $\{X_n\}$ process is a birth and death chain in its own right. Letting q_a denote the downward transition probability $q_a = (1 - p)a$, a stationary distribution π_a for Andrew's process X exists if $p < q_a$, in which case familiar probability flux reasoning gives $\pi_a(i)p = \pi_a(i + 1)q_a$, or $\pi_a(i + 1)/\pi_a(i) = p/q_a$, so that

$$\pi_a(i) = \left(\frac{p}{q_a}\right)^i \left(1 - \frac{p}{q_a}\right) \quad \text{for } i = 0, 1, \dots$$

Here is where time reversibility allows us to make an interesting and useful observation. Assume $X_0 \sim \pi_a$. Then we know that $\{X_n\}$, being a stationary birth and death process, is time reversible. Define A_n to be 1 if Andrew has an “arrival” at time n and 0 otherwise, so that $A_n = 1$ occurs when $X_n = X_{n-1} + 1$. Define another indicator random variable D_n to be 1 if Andrew has a “departure” at time n , that is, when $X_n = X_{n-1} - 1$. Considering k to be the present time, clearly the present queue size X_k is independent of the future arrivals A_{k+1}, A_{k+2}, \dots . This obvious fact, when applied to the reversed process, gives something interesting. In the reversed process, if k again represents the “present,” then “future arrivals” correspond to the departures D_k, D_{k-1}, \dots . Therefore, we can say that the departures (D_1, D_2, \dots, D_k) are independent of the queue size X_k . This is quite surprising, isn't it? Also, since reversibility implies that arrivals in the reversed process have the same probabilistic behavior as arrivals in the forward process, we see that the departures D_1, D_2, \dots are *iid* Bernoulli(p). Thus, the output process of Andrew's queue is the same probabilistically as the input process of his queue. Isn't that interesting? For example, we have found that the departure process does not depend on the service completion probability a .

Bertha's queue size $\{Y_n\}$ is also a birth and death chain, with a similar structure as Andrew's. We have just shown that Bertha's input consists of *iid* Bernoulli(p) random variables, just as Andrew's input does. Defining the downward transition probability $q_b =$

$(1-p)b$ for $\{Y_n\}$, if $p < q_b$ the stationary distribution π_b is given by

$$\pi_b(i) = \left(\frac{p}{q_b}\right)^i \left(1 - \frac{p}{q_b}\right) \quad \text{for } i = 0, 1, \dots$$

Now we are ready to show a surprising property of the stationary distribution of (X_k, Y_k) : the two queue sizes X_k and Y_k are independent! That is, we claim that the stationary distribution π of $\{(X_n, Y_n)\}$ takes the product form

$$\pi(i, j) = \pi_a(i)\pi_b(j).$$

I'll try to say the idea loosely in words first, then more carefully. It is helpful to imagine that at each time, Andrew flips a coin with probability a of heads, and if he gets heads he completes a piece of work, if that is possible—i.e. if there is work in his pile and if he has not received a new arrival. Bertha does the same, only with her coin having probability b . Back to the question: Supposing that X_0 and Y_0 are independent with distributions π_a and π_b , we want to see that X_n and Y_n are also independent with distributions π_a and π_b . We know the marginal distributions of X_n and Y_n are π_a and π_b ; independence is the real question. The key is the observation made above that X_n is independent of Andrew's departures up to time n , which are the same as Bertha's arrivals up to time n . So since Y_n is determined by Y_0 , Bertha's arrivals up to time n , and Bertha's service coin flips, all of which are independent of X_n , we should have Y_n independent of X_n .

To establish this more formally, assuming that $(X_0, Y_0) \sim \pi$, we want to show that $(X_1, Y_1) \sim \pi$. Since π_a and π_b are stationary for $\{X_n\}$ and $\{Y_n\}$, we know that $X_1 \sim \pi_a$ and $Y_1 \sim \pi_b$, so our task is to show that X_1 and Y_1 are independent. Let A_k^X denote the indicator of an arrival to Andrew's desk at time k . Let $S_k^X = 1$ if at time k Andrew's "service completion coin flip" as described in the previous paragraph comes up heads, and $S_k^X = 0$ otherwise. Define S_k^Y analogously for Bertha. We are assuming that the random variables X_0, Y_0, A_1^X, S_1^X , and S_1^Y are independent. But we can write X_1 and A_1^Y as functions of (X_0, A_1^X, S_1^X) . [The precise functional forms are not important, but just for fun,

$$X_1 = \begin{cases} X_0 + 1 & \text{if } A_1^X = 1 \\ X_0 & \text{if } A_1^X = 0 \text{ and } S_1^X = 0 \\ X_0 - 1 & \text{if } A_1^X = 0 \text{ and } S_1^X = 1 \text{ and } X_0 > 0 \\ 0 & \text{if } A_1^X = 0 \text{ and } X_0 = 0 \end{cases}$$

and

$$A_1^Y = \begin{cases} 1 & \text{if } X_0 > 0 \text{ and } A_1^X = 0 \text{ and } S_1^X = 1 \\ 0 & \text{otherwise} \end{cases}$$

is one way to write them.] So Y_0 is independent of (X_1, A_1^Y) . But we know that X_1 is independent of whether there is a departure from Andrew's queue at time 1, which is just the indicator A_1^Y . Therefore, the 3 random variables Y_0, X_1 , and A_1^Y are independent. Finally, observe that S_1^Y is independent of (Y_0, X_1, A_1^Y) , so that the 4 random variables Y_0, X_1, A_1^Y , and S_1^Y are all independent. Thus, since Y_1 is a function of (Y_0, A_1^Y, S_1^Y) , it follows that X_1 and Y_1 are independent.

2.4 The Metropolis method

This is a very useful general method for using Markov chains for simulation. The idea is a famous one due to Metropolis et al. (1953), and is known as the *Metropolis method*. Suppose we want to simulate a random draw from some distribution π on a finite set \mathcal{S} . By the Basic Limit Theorem above, one way to do this (approximately) is to find an irreducible, aperiodic probability transition matrix P satisfying $\pi P = \pi$, and then run a Markov chain according to P for a sufficiently long time.

Suppose we have chosen some connected graph structure $\mathcal{G} = (\mathcal{S}, \mathcal{E})$ on our set \mathcal{S} . That is, we think of each element of \mathcal{S} as a node, and we imagine a collection \mathcal{E} of edges, where each edge joins some pair of nodes. If nodes i and j are joined by an edge, we say i and j are *neighbors*. Let $\mathcal{N}(i)$ denote the set of neighbors of node i . [We'll assume that $i \notin \mathcal{N}(i)$ for all i .] Just to make sure the situation is clear, I want to emphasize that this graph structure is not imposed on us, and there is not a single, unique, magic choice that will work. The set of edges \mathcal{E} is ours to choose; we have a great deal of freedom here, and different choices will lead to different matrices P satisfying $\pi P = \pi$.

As a preliminary observation, recall from Example (2.8) that a random walk on \mathcal{G} , which has probability transition matrix

$$(2.9) \quad P_{\text{rw}}(i, j) = \begin{cases} \frac{1}{d(i)} & \text{if } j \in \mathcal{N}(i) \\ 0 & \text{otherwise} \end{cases}$$

has stationary distribution

$$\pi_{\text{rw}}(i) = \frac{d(i)}{\sum_{j \in \mathcal{S}} d(j)},$$

where $d(i)$ is the degree of node i . To reduce typographical and conceptual clutter, let us write this as

$$\pi_{\text{rw}}(i) \propto d(i),$$

by omitting the denominator, which is simply a normalization constant [constant in that it does not depend on i] that makes the probabilities add up to 1. The Basic Limit Theorem tells us that (assuming aperiodicity holds) if we run the random walk for sufficiently long, then we get arbitrarily close to achieving the distribution π_{rw} .

Thus, simply running a random walk on \mathcal{G} would solve our simulation problem if we happened to want to simulate from π_{rw} . In general, however, we will want to simulate from some different, arbitrary distribution π , which we will write in the form

$$(2.10) \quad \pi(i) \propto d(i)f(i).$$

That is, we are interested in modifying the relative probabilities of the natural random walk stationary distribution by some multiplicative function f . Our goal here is a simple way to modify the random walk transition probabilities in such a way that the modified probability transition matrix has stationary distribution π . The Metropolis method solves

this problem by defining the probability transition matrix

$$(2.11) \quad P(i, j) = \begin{cases} \frac{1}{d(i)} \min\{1, \frac{f(j)}{f(i)}\} & \text{if } j \in \mathcal{N}(i) \\ 1 - \sum_{k \in \mathcal{N}(i)} P(i, k) & \text{if } j = i \\ 0 & \text{otherwise.} \end{cases}$$

The verification that this method works is simple.

(2.12) CLAIM. For π defined by (2.10) and $(P(i, j))$ defined by (2.11), we have $\pi P = \pi$.

PROOF: For $j \in \mathcal{N}(i)$,

$$\pi(i)P(i, j) \propto f(i) \min\{1, \frac{f(j)}{f(i)}\} = \min\{f(i), f(j)\}$$

is symmetric in i and j , so we have

$$(2.13) \quad \pi(i)P(i, j) = \pi(j)P(j, i).$$

In fact, (2.13) holds for all i and j , since it is trivial to verify (2.13) in the cases when $i = j$ or $j \notin \mathcal{N}(i)$. Summing (2.13) over i gives

$$\sum_i \pi(i)P(i, j) = \pi(j) \sum_i P(j, i) = \pi(j),$$

so that $\pi P = \pi$. □

Notice that the last proof showed that the “detailed balance” equations $\pi(i)P(i, j) = \pi(j)P(j, i)$ that characterize time-reversible Markov chains hold.

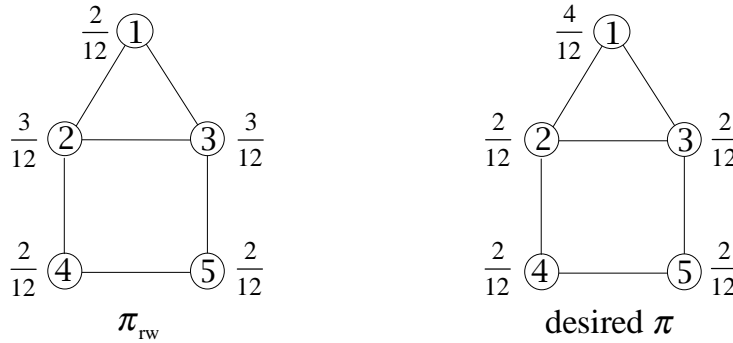
▷ As we know from our discussion of the Basic Limit Theorem, the fact that a Markov chain has stationary distribution π does not in itself guarantee that the Markov chain will converge in distribution to π . Exercise [2.10] gives conditions under which this convergence holds.

Running the Metropolis chain (using the P from (2.11)) is actually a simple modification of performing a random walk (using P_{rw} from (2.9)). To run the random walk, at each time, we choose a random neighbor and go there. We can think of running the Metropolis chain as follows. Suppose we are currently at state i and we are about to generate our next random transition. We start out, in the same way as the random walk, by choosing a random neighbor of i ; let’s call our choice j . The difference between the Metropolis chain and the random walk is that in the Metropolis chain, we might move to j , or we might stay at i . So let’s think of j as our “candidate” state, and we next make a probabilistic decision about whether to “accept the candidate” and move to j , or “reject the candidate” and stay at i . The probability that we accept the candidate is the extra factor

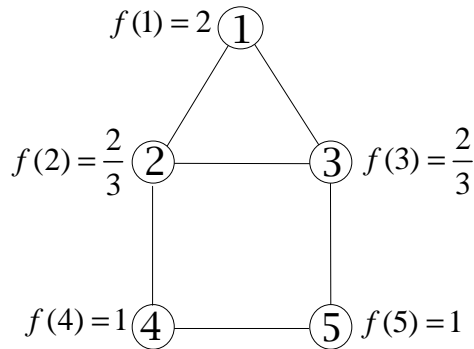
$$(2.14) \quad \min\{1, \frac{f(j)}{f(i)}\}$$

that appears in (2.11). Thus, having nominated a candidate j , we look at the values $f(j)$ and $f(i)$. If $f(j) \geq f(i)$, the minimum in (2.14) is 1, and we definitely move to j . If $f(j) < f(i)$, the minimum in (2.14) is $f(j)/f(i)$, and we move to j with probability $f(j)/f(i)$. This makes qualitative sense: for example, if $f(j)$ is much smaller than $f(i)$, this means that, relative to the random walk stationary distribution π_{rw} , our desired distribution π places much less probability on j than on i , so that we should make a transition from i to j much less frequently than the random walk does. This is accomplished in the Metropolis chain by usually rejecting the candidate j .

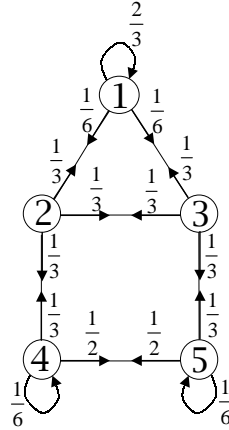
(2.15) EXAMPLE. To illustrate the Metropolis method in a simple way, we'll discuss an artificial toy example where we don't really need the method. Suppose the distribution π on $\mathcal{S} = \{1, 2, 3, 4, 5\}$ is $(4/12, 2/12, 2/12, 2/12, 2/12)$, and suppose the graph structure we choose on \mathcal{S} is the “house” graph from Example (2.8). Thus, we want to be on the roof of the house (state 1) with probability $4/12$, and at each of the other states with equal probability, $2/12$.



Comparing with the distribution π_{rw} that we found in Example (2.8) [and is reproduced in the figure above] we can calculate the f ratios to be $f(1) = (4/12)/(2/12) = 2$, and, similarly, $f(2) = 2/3 = f(3)$ and $f(4) = 1 = f(5)$.



And now the Metropolis formula for P shows that to modify the random walk to have the desired stationary distribution, we run the process depicted in the figure below.



These transition probabilities were calculated as follows. For example,

$$P(1, 2) = \frac{1}{2} \min \left\{ 1, \frac{2/3}{2} \right\} = \frac{1}{6} = P(1, 3),$$

$$P(1, 1) = 1 - \frac{1}{6} - \frac{1}{6} = \frac{2}{3},$$

$$P(2, 1) = \frac{1}{3} \min \left\{ 1, \frac{2}{2/3} \right\} = \frac{1}{3},$$

$$P(2, 3) = \frac{1}{3} \min \left\{ 1, \frac{2/3}{2/3} \right\} = \frac{1}{3},$$

$$P(2, 4) = \frac{1}{3} \min \left\{ 1, \frac{1}{2/3} \right\} = \frac{1}{3},$$

$$P(2, 2) = 1 - \frac{1}{3} - \frac{1}{3} - \frac{1}{3} = 0,$$

and so on. □

The Metropolis method has a nice property: actually we do not even quite have to be able to write down or compute the probabilities $\pi(i)$ in order to be able to use the method to simulate from π ! That is, as is clear from (2.11), to run the Metropolis chain, all we need to know are ratios of the form $f(j)/f(i)$, or, equivalently, ratios of the form $\pi(j)/\pi(i)$ [these are equivalent by (2.10), and because we assume we know the degrees $d(j)$ and $d(i)$]. That is, we do not have to know any of the $\pi(i)$'s explicitly to use the method; all we have to know are the *ratios* $\pi(j)/\pi(i)$. Now this may not seem like a big deal, but there are cases in which we would like to simulate from a distribution π for which the ratios $\pi(j)/\pi(i)$ are easy to compute, while the individual probabilities $\pi(j)$ and $\pi(i)$ are extremely difficult to compute. This happens when π is known only up to a complicated multiplicative normalization constant. One simple example of this you have already seen: in our problem of simulating a uniformly distributed 4×4 table with given row and column sums, the desired probability of any given table is the reciprocal of the number of tables

satisfying the given restrictions—a number that we do not know! (Remember, in fact, a bonus of being able to generate a nearly uniformly distributed table is that it leads to a method for approximating the number of such tables.) So in this problem, we do not know the individual probabilities of the form $\pi(i)$. But the ratio $\pi(j)/\pi(i)$ is simply 1 for a uniform distribution! Now, simulating from a uniform distribution is admittedly a special case, and a symmetric probability transition matrix will do the trick. For a more general class of examples, in the Bayesian approach to statistics, suppose the unknown parameter of interest is $\theta \in \Theta$, where Θ is a finite parameter space. Suppose our prior distribution [probability mass function] for θ is $\mu(\theta)$ and the likelihood is $P(x | \theta)$. Both $\mu(\theta)$ and $P(x | \theta)$ are known to us because we specify them as part of our probability model. The posterior distribution for θ given x is

$$P(\theta | x) = \frac{\mu(\theta)P(x | \theta)}{G},$$

where

$$G = \sum_{\theta' \in \Theta} \mu(\theta')P(x | \theta').$$

The sum G may be very difficult to compute; in statistical mechanics it is the infamous “partition function.” However, for given x , if we want to simulate from the posterior distribution $P(\cdot | x)$, we can do so using the Metropolis method; although the distribution itself may be hard to compute because G is hard to compute, the ratios

$$\frac{P(\theta_1 | x)}{P(\theta_2 | x)} = \frac{\mu(\theta_1)P(x | \theta_1)}{\mu(\theta_2)P(x | \theta_2)}$$

are easy to compute.

2.5 Simulated annealing

Simulated annealing is a recent and powerful technique for addressing large, complicated optimization problems. Although the idea is so simple it may sound naive, the simulated annealing method has enabled people in some cases to find better answers to bigger problems than any previously known method.

Suppose we would like to minimize some “cost function” $c(\cdot)$ defined on a set \mathcal{S} . For example, $c(\cdot)$ might be a function of d variables defined on the simplest interesting domain \mathcal{S} , namely, the domain $\mathcal{S} = \{0, 1\}^d$, in which each of the d variables may take on only the two values 0 and 1. That is, this \mathcal{S} is the set of 2^d d -tuples $i = (i_1, \dots, i_d)$; we could think of these as the vertices of the d -dimensional unit cube. So for $d = 10$ variables, \mathcal{S} contains $2^{10} \approx 1000$ points. If we want to solve a problem with $d = 20, 30$, or 40 variables, the number of points in \mathcal{S} rises to about one million, one billion, and one trillion, respectively. Have our computers gotten fast enough that we can just about handle a trillion points now? Well, if we then just add 20 more variables to the problem, all of a sudden our computers are a million times too slow again. So even though computers are getting faster all the time, clearly our appetite for solving larger and more complex problems grows much, much faster. “But come now, who really deals with functions of 60 or 100 variables?” you

may be wondering. Well, consider, for example, an image processing problem, in which we want to calculate an optimal guess at an image, given a noisy, corrupted version. If we are dealing with a black and white image, so that each pixel can be encoded as a 0 or a 1, then our state space is exactly of the form $\{0,1\}^d$ that we have been discussing. How big are typical values of d ? Very big: d is the number of pixels in the image, so if we have a (quite crude!) 200×200 image, then $d = 40,000$. This is *much* bigger than 60 or 100! I hope this gives you some inkling of the inconceivable vastness of many quite ordinary combinatorial optimization problems, and a feeling for the utter hopelessness of *ever* solving such problems by exhaustive enumerative search of the state space.

Stuart and Donald Geman used simulated annealing in their approach to image restoration. Another famous example of a difficult optimization problem is the traveling salesman problem in which a salesman is given a set of cities he must visit, and he wants to decide which city to visit first, second, and so on, in such a way that his total travel distance is minimized. For us, in addition to its practical importance, simulated annealing provides a nice illustration of some of the Markov chain ideas we have discussed so far, as well as an excuse to learn something about time-inhomogeneous Markov chains.

2.5.1 Description of the method

The method is a combination of the familiar idea of using Markov chains for simulation and a new idea (“annealing”) that provides the connection to optimization. We have already discussed the very general Metropolis method that allows us to simulate approximately from any desired distribution on \mathcal{S} . But what does simulation have to do with optimization or “annealing” or whatever?

Our goal is to find an $i \in \mathcal{S}$ minimizing $c(i)$, where c is a given cost function defined on the set of nodes \mathcal{S} of a graph. As discussed above, an exact solution of this problem may be an unattainable ideal in practice, but we would like to come as close as possible. For each $T > 0$, define a probability distribution $\alpha_T = \{\alpha_T(i) : i \in \mathcal{S}\}$ on \mathcal{S} by

$$(2.16) \quad \alpha_T(i) = \frac{d(i)e^{-c(i)/T}}{G(T)},$$

where again $d(i)$ is the degree of node i and of course

$$G(T) = \sum_{i \in \mathcal{S}} d(i)e^{-c(i)/T}$$

is just the normalizing constant that makes (2.16) a probability distribution. The letter “ T ” stands for “temperature.”

We have defined a family of probability distributions on \mathcal{S} ; corresponding to each positive T there is a distribution α_T . These distributions have an important property that explains why we are interested in them. To state this property, let \mathcal{S}^* denote the set of global minimizers of $c(\cdot)$, that is,

$$\mathcal{S}^* = \{i^* \in \mathcal{S} : c(i^*) = \min_{i \in \mathcal{S}} c(i)\}.$$

Our goal is to find an element of \mathcal{S}^* . Define a distribution α^* on \mathcal{S}^* by

$$(2.17) \quad \alpha^*(i) = \frac{d(i)}{\sum_{j \in \mathcal{S}^*} d(j)}$$

if $i \in \mathcal{S}^*$, and $\alpha^*(i) = 0$ otherwise. The important thing to keep in mind about the distribution α^* is that it puts positive probability only on globally optimal solutions of our optimization problem.

$$(2.18) \text{ FACT. As } T \downarrow 0, \text{ we have } \alpha_T \xrightarrow{\mathcal{D}} \alpha^*, \text{ that is, } \alpha_T(i) \rightarrow \alpha^*(i) \text{ for all } i \in \mathcal{S}.$$

The symbol “ $\xrightarrow{\mathcal{D}}$ ” stands for convergence in distribution.

▷ *Exercise [2.16] asks you to prove (2.18).*

Thus, we have found that, as $T \downarrow 0$, the distributions α_T converge to the special distribution α^* . If we could somehow simulate from α^* , our optimization problem would be solved: We would just push the simulate from α^* button, and out would pop a random element of \mathcal{S}^* , which would make us most happy. Of course, we cannot do that, since we are assuming that we do not already have the answer to the minimization problem that we are trying to solve! However, we *can* do something that seems as if it should be nearly as good: simulate from α_T . If we do this for a value of T that is pretty close to 0, then since α_T is pretty close to α^* for that T , presumably we would be doing something pretty good.

So, fix a $T > 0$. How do we simulate from α_T ? We can use the Metropolis idea to create a probability transition matrix $A_T = (A_T(i, j))$ such that $\alpha_T A_T = \alpha_T$, and then run a Markov chain according to A_T .

[A note on notation: I hope you aren’t bothered by the use here of α and A for stationary distributions and probability transition matrices related to simulated annealing. The usual letters π and P have been so overworked that using different notation for the special example of simulated annealing should be clearer ultimately. Although they don’t look much like π and P , the letters α and A might be remembered as mnemonic for “simulated Annealing” at least.]

A glance at the definition of α_T in (2.16) shows that we are in the situation of the Metropolis method as described in (2.10) with the choice $f(i) = e^{-c(i)/T}$. So as prescribed by (2.11), for $j \in \mathcal{N}(i)$ we take

$$(2.19) \quad \begin{aligned} A_T(i, j) &= \frac{1}{d(i)} \min \left\{ 1, \frac{e^{-c(j)/T}}{e^{-c(i)/T}} \right\} \\ &= \frac{1}{d(i)} \begin{cases} 1 & \text{if } c(j) \leq c(i) \\ e^{-(c(j)-c(i))/T} & \text{if } c(j) > c(i). \end{cases} \end{aligned}$$

The specification of A_T is completed by the obvious requirements that $A_T(i, i) = 1 - \sum_{j \in \mathcal{N}(i)} A_T(i, j)$ and $A_T(i, j) = 0$ if $j \notin \mathcal{N}(i)$.

For any fixed temperature $T_1 > 0$, if we run a Markov chain $\{X_n\}$ having probability transition matrix A_{T_1} for “a long time,” then the distribution of $\{X_n\}$ will be very close to

α_{T_1} . If our goal is to get the chain close to the distribution α^* , then continuing to run this chain will not do much good, since the distribution of X_n will get closer and closer to α_{T_1} , not α^* . So the only way we can continue to “make progress” toward the distribution α^* is to decrease the temperature to T_2 , say, where $T_2 < T_1$, and continue to run the chain, but now using the probability transition matrix A_{T_2} rather than A_{T_1} . Then the distribution of X_n will approach the distribution α_{T_2} . Again, after the distribution gets very close to α_{T_2} , continuing to run the chain at temperature T_2 will not be an effective way to get closer to the desired distribution α^* , so it makes sense to lower the temperature again.

It should be quite clear intuitively that, if we lower the temperature slowly enough, we can get the chain to converge in distribution to α^* . For example, consider “piecewise constant” schedules as discussed in the last paragraph. Given a decreasing sequence of positive temperatures $T_1 > T_2 > \dots$ such that $T_n \downarrow 0$, we could start by running a chain for a time n_1 long enough so that the distribution of the chain at time n_1 is within a distance of $1/2$ [in total variation distance, say] from α_{T_1} . Then we could continue to run at temperature T_2 until time n_2 , at which the distribution of the chain is within $1/3$ of α_{T_2} . Then we could run at temperature T_3 until we are within $1/4$ of α_{T_3} . And so on. Thus, as long as we run long enough at each temperature, the chain should converge in distribution to α^* . We must lower the temperature slowly enough so that the chain can always “catch up” and remain close to the stationary distribution for the current temperature.

Once we have had this idea, we need not lower the temperature in such a piecewise constant manner, keeping the temperature constant for many iterations and only then changing it. Instead, let us allow the temperature to change at each at each step of the Markov chain. Thus, each time n may have its own associated temperature T_n , and hence its own probability transition matrix A_{T_n} . The main theoretical result that has been obtained for simulated annealing says that for any problem there is a cooling schedule of the form

$$(2.20) \quad T_n = \frac{a}{\log(n+b)}$$

[where a and b are constants] such that, starting from any state $i \in \mathcal{S}$, the [time-inhomogeneous] Markov chain $\{X_n\}$ will converge in distribution to α^* , the uniform distribution on the set of global minima.

Accordingly, a simulated annealing procedure may be specified as follows. Choose a “cooling schedule” T_0, T_1, \dots ; the schedules we will discuss later will have the property that $T_n \downarrow 0$ as $n \rightarrow \infty$. Choose the initial state X_0 according to a distribution ν_0 . Let the succession of states X_0, X_1, X_2, \dots form a time-inhomogeneous Markov chain with probability transition matrices $A_{T_0}, A_{T_1}, A_{T_2}, \dots$, so that

$$\mathbb{P}\{X_{n+1} = j \mid X_n = i\} = A_{T_n}(i, j)$$

and

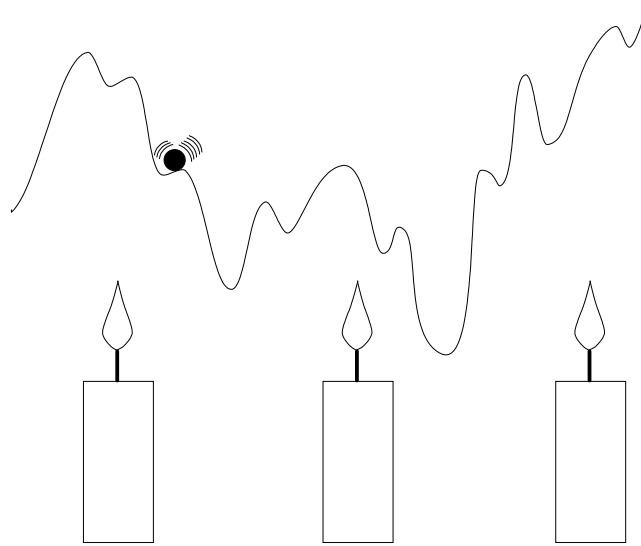
$$(2.21) \quad X_n \sim \nu_n = \nu_0 A_{T_0} A_{T_1} \cdots A_{T_{n-1}}.$$

By the way, what is all this talk about “temperature,” “cooling schedule,” “annealing,” and stuff like that? I recommend you consult the article by Kirkpatrick et al., but I’ll try to say a few words here. Let’s see, I’ll start by looking up the word “anneal” in my dictionary. It gives the definition

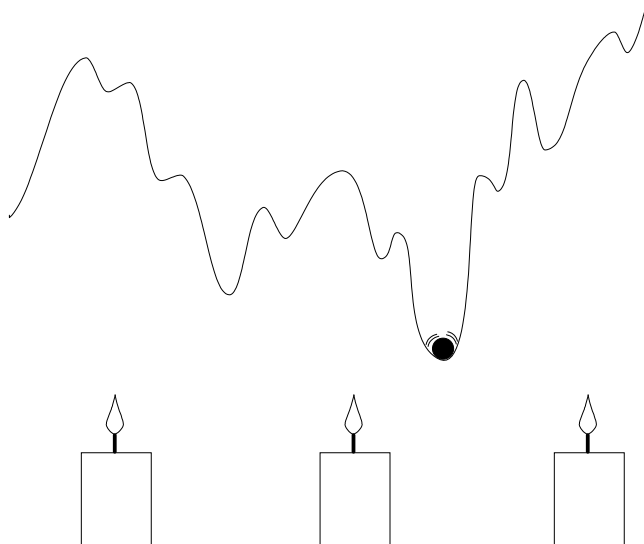
To free (glass, metals, etc.) from internal stress by heating and gradually cooling.

The idea, as I understand it, is this. Suppose we want to have a metal that is as “stress-free” as possible. That corresponds to the metal being “relaxed,” in a low energy state. The metal is happy when it has low energy, and it wants to get to such a state, just as a ball will roll downhill to decrease its potential energy. In the liquid state, the atoms of a metal are all sliding around in all sorts of random orientations. If we quickly freeze the metal into a solid by cooling it quickly, its atoms will freeze into a configuration that has all sorts of haphazard orientations that give the metal unnecessary excess potential energy. In contrast, if we start with the metal as a liquid and then cool it extremely slowly, then the atoms have plenty of time to explore around, find, and work themselves into a nice, happy, ordered, low-energy configuration. Thus, nature can successfully address an optimization problem—minimization of the energy of the metal—by slow cooling. I hope that gives the flavor of the idea at least.

For another image that captures certain aspects of simulated annealing, imagine a bumpy frying pan containing a marble made of popcorn. At high temperature, the marble is jumping around quite wildly, ignoring the ups and downs of the pan.



As we lower the temperature, the popcorn begins to settle down into the lower parts of the pan.



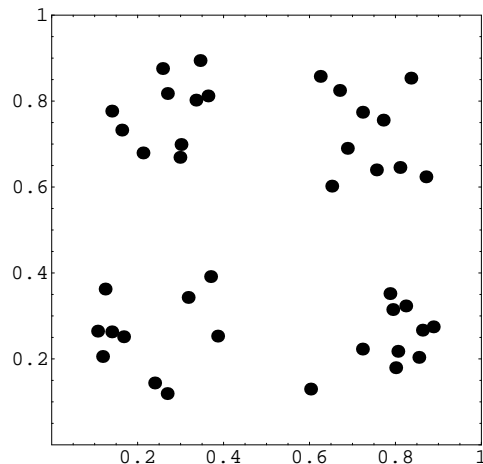
I'd like to make sure it is clear to you how to run the method, so that you will be able to implement it yourself and experiment with it if you'd like. So let us pause for a moment to imagine what it is like to move around according to a Markov chain having probability transition matrix A_T specified by (2.19). Here is one way to describe it. Suppose that we are now at time n sitting at node i of the graph, and we are trying to decide where to go next. First we choose at random one of the d neighbors of i , each neighbor being chosen with the same probability $1/d(i)$. Say we choose node j . Then j becomes out "candidate" for where to go at time $n + 1$; in fact, at time $n + 1$ we will either move to j (accept the candidate) or stay at i . To decide between these two alternatives, we must look at the values of $c(i)$ and $c(j)$. If $c(j) \leq c(i)$, we definitely move to j . If $c(j) > c(i)$, we might or might not move to j ; in fact we move to j with probability $e^{-(c(j)-c(i))/T}$ and stay at i with the complementary probability $1 - e^{-(c(j)-c(i))/T}$. At high temperature, even when $c(j) > c(i)$, the probability $e^{-(c(j)-c(i))/T}$ is close to 1, so that we accept all candidates with high probability. Thus, at high temperature, the process behaves nearly like a random walk on \mathcal{S} , choosing candidates as a random walk would, and almost always accepting them. At lower temperatures, the process still always accepts "downhill moves" [those with $c(j) \leq c(i)$], but has a lower probability of accepting uphill moves. At temperatures very close to zero, the process very seldom moves uphill.

Note that when $c(j) \leq c(i)$, moving to j "makes sense"; since we are trying to minimize $c(\cdot)$, decreasing c looks like progress. However, simulated annealing is not just another "descent method," since we allow ourselves positive probability of taking steps that *increase* the value of c . This feature of the procedure prevents it from getting stuck in local minima.

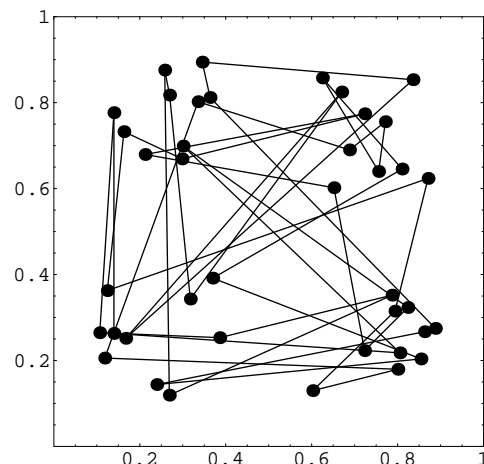
You may have asked yourself the question, "If we want to bring the temperature down to 0, then why don't we just run the chain at temperature 0?" But now you can see that at temperature $T = 0$, the process will never make an uphill move. Thus, running at temperature 0 is a descent method, which will get stuck in local minima, and therefore will not approach global minima. At temperature 0, the chain all of a sudden loses the nice properties it has at positive temperatures—it is no longer irreducible, and so the

basic limit theorem doesn't apply.

(2.22) EXAMPLE [TRAVELING SALESMAN PROBLEM]. The figure below shows the location of 40 cities. A traveling salesman who lives in one of the cities wants to plan a tour in such a way that he visits all of the cities and then returns home while traveling the shortest possible total distance.



To solve this problem by simulated annealing, we start with some legal tour. We'll start in a very dumb way, with a tour that is simply a random permutation of the 40 cities.



As we know, simulated annealing gradually modifies and, we hope, improves the tour by proposing random modifications of the tour, and accepting or rejecting those modifications randomly based on the current value of the temperature parameter and the amount of improvement or deterioration produced by the proposed modification.

Here our state space \mathcal{S} is the set of all possible traveling salesman tours. Each such tour can be specified by a permutation of the cities; for example, if there are 8 cities, the tour $(1, 3, 5, 7, 2, 4, 6, 8)$ means that we start at city 1, then go to 3, 5, 7, 2, 4, 6, and 8 in that order, and finally return from 8 back to 1. [Note that since the tours are closed circuits because the salesman returns to his starting point, we could consider, for example, the tour $(3, 5, 7, 2, 4, 6, 8, 1)$ to be the same tour as $(1, 3, 5, 7, 2, 4, 6, 8)$. So in this way different

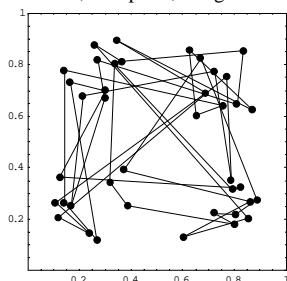
permutations can represent the same tour. We could say these two equivalent permutations are “rotations” of each other.]]

Recall that the Metropolis method, and simulated annealing, require us to choose a neighborhood structure on \mathcal{S} . That is, for each tour i in \mathcal{S} , we must say which other tours are the neighbors of i , and then the simulated annealing method chooses a random candidate neighbor at each iteration. There are many possible ways to do this. One way that seems to move around \mathcal{S} nicely and is also easy to work with is to choose two cities on the tour randomly, and then reverse the portion of the tour that lies between them. For example, if we are currently at the tour $(1, 2, 3, 4, 5, 6, 7, 8)$, we might choose the two cities 4 and 6, and change from the tour $(1, 2, 3, 4, 5, 6, 7, 8)$ to the neighboring tour $(1, 2, 3, \underbrace{6, 5, 4}, 7, 8)$.

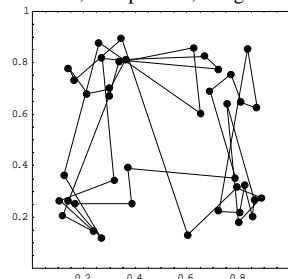
Another convenient feature of this definition of neighbors is that it is easy to calculate the difference in total lengths between neighboring tours; no matter how many cities we are considering, the *difference* in tour lengths is calculated using only four intercity distances — the edges that are broken and created as we break out part of the tour and then reattach it in the reversed direction.

Using the type of moves just described, it was quite easy to write a computer program, and fun to watch it run. Here is how it went:

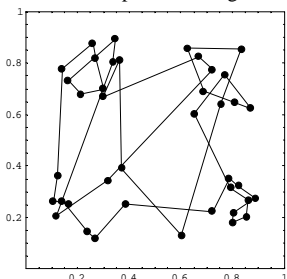
Time 1000, Temp 0.5, Length 15.5615



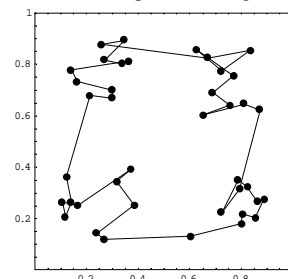
Time 2000, Temp 0.212, Length 10.0135



Time 5000, Temp 0.089, Length 7.6668

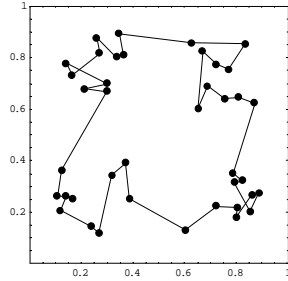


Time 10000, Temp 0.058, Length 5.07881

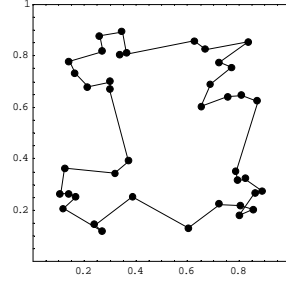


If we were watching a movie of this, we would see the path slithering around and gradually untangling itself. The slithering is violent at high temperatures. It starts to get the gross features right, removing those silly long crossings from one corner of the picture to another. And now let's watch it continue.

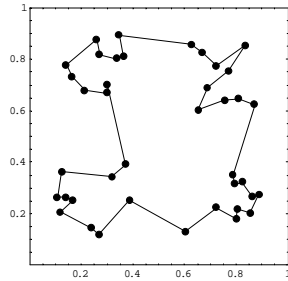
Time 20000, Temp 0.038, Length 4.52239



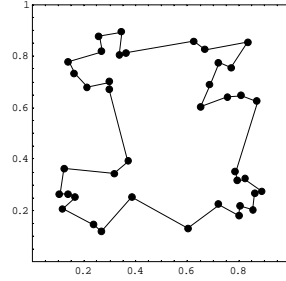
Time 30000, Temp 0.0245, Length 4.11739



Time 40000, Temp 0.0245, Length 4.00033



Time 66423, Temp 0.0104, Length 3.96092



Our salesman ends up with a very nice tour. I checked to make sure this tour is a local optimum — no neighboring tour is better. Maybe it's even the globally optimal tour. \square

Doesn't this method of random searching seem extraordinarily dumb? No particular knowledge or strategic ideas about the particular problem are built in. The Markov property means that there is no memory; any successes or failures from the past are forgotten. But then again, evolution is “dumb” in the same way, accumulating random changes, gently guided by natural selection, which makes favorable changes more likely to persist. And look at all of the marvelous solutions evolution has produced.

2.5.2 The Main Theorem

First some notation. Define the matrix $A^{(n)}$ by

$$A^{(n)} = \prod_{k=0}^{n-1} A_{T_k}.$$

Recalling that ν_n denotes the distribution of X_n , (2.21) says that $\nu_n = \nu_0 A^{(n)}$. Let $A^{(n)}(i, \cdot)$ denote the i th row of the matrix $A^{(n)}$. This row may be interpreted as the distribution of X_n given that $X_0 = i$. We will continue using the notation $\|\mu - \nu\|$ for the total variation distance

$$\|\mu - \nu\| = \sup_{S \subset \mathcal{S}} |\mu(S) - \nu(S)| = \frac{1}{2} \sum_{i \in \mathcal{S}} |\mu_i - \nu_i|$$

between the distributions μ and ν .

We also need some definitions related to the structure of the graph \mathcal{G} and the cost function $c(\cdot)$. For two nodes $i, j \in \mathcal{G}$, define the *distance* $\rho(i, j)$ to be the length [i.e., number of edges] of a path from i to j of minimal length. Define the *radius* r of \mathcal{G} by

$$r = \min_{i \in \mathcal{S}_c} \max_{j \in \mathcal{S}} \rho(i, j),$$

where

$$\mathcal{S}_c = \{i \in \mathcal{S} : c(j) > c(i) \text{ for some } j \in \mathcal{N}(i)\},$$

that is, \mathcal{S}_c is the set of all nodes that are not local maxima of the function c . [Note that r actually depends on the function c in addition to the graph \mathcal{G} , but we won't indicate this in the notation.] Define the number L to be the largest “local fluctuation” of $c(\cdot)$, that is,

$$L = \max_{i \in \mathcal{S}} \max_{j \in \mathcal{N}(i)} |c(j) - c(i)|.$$

Finally, recall the definition of α^* from (2.17).

After only 38 million definitions (counting multiplicity), we are ready to state the main theorem.

(2.23) THEOREM [MAIN THEOREM]. *For any cooling schedule T_0, T_1, \dots satisfying*

- (i) $T_{n+1} < T_n$ for all $n \geq 0$
- (ii) $T_n \rightarrow 0$ as $n \rightarrow \infty$
- (iii) $\sum_k \exp(-rL/T_{kr-1}) = \infty$,

we have

$$\|A^{(n)}(i, \cdot) - \alpha^*\| \rightarrow 0$$

as $n \rightarrow \infty$, for all $i \in \mathcal{S}$.

The proof of this theorem is a rather easy application of some theory of time-inhomogeneous Markov chains. We'll develop this theory in the next section, and then come back to simulated annealing to prove Theorem (2.23).

Here is a specific family of cooling schedules that satisfy the conditions of the theorem: taking $\gamma \geq rL$, let

$$T_n = \frac{\gamma}{\log(n)}$$

for $n > 1$. [And let T_0 and T_1 be arbitrary, subject to the monotonicity condition (i).] It is easy to check that the conditions of the theorem hold; (iii) boils down to the statement that $\sum_k k^{-p} = \infty$ if $p \leq 1$.

Note that the “convergence in distribution” sort of conclusion of the theorem is weaker than almost sure (“a.s.”) convergence [i.e., convergence with probability 1]. There is good reason for this: a.s. convergence indeed does not hold in general for simulated annealing. Intuitively, we should not expect to be able to get a.s. convergence—if our procedure has

the nice property of always escaping eventually from local minima, then we must live with the fact that it will also eventually “escape” from global minima. That is, the process will not get “stuck” in \llbracket that is, converge to \rrbracket a global minimum. It is true that, along a typical sample path, the process will spend a larger and larger fraction of its time at or near global minima as n increases. However, it will also make infinitely many (although increasingly rare) excursions away from global minima.

▷ *I would say that the theorem is certainly elegant and instructive. But if you come away thinking it is only a beginning and is not yet answering the “real” question, I would consider that a good sign. Exercise [2.15] asks for your thoughts on the matter.*

2.6 Ergodicity Concepts for Time-Inhomogeneous Markov chains

In this section we will work in the general setting of a time-inhomogeneous Markov chain on a countably infinite state space \mathcal{S} ; we will revert to finite \mathcal{S} when we return to the specific problem of simulated annealing.

(2.24) NOTATION. For a time-inhomogeneous Markov chain $\{X_n\}$, let P_n denote the probability transition matrix governing the transition from X_n to X_{n+1} , that is, $P_n(i, j) = \mathbb{P}\{X_{n+1} = j \mid X_n = i\}$. Also, for $m < n$ define $P^{(m,n)} = \prod_{k=m}^{n-1} P_k$, so that $P^{(m,n)}(i, j) = \mathbb{P}\{X_n = j \mid X_m = i\}$.

(2.25) DEFINITION. $\{X_n\}$ is strongly ergodic if there exists a probability distribution π^* on \mathcal{S} such that

$$\lim_{n \rightarrow \infty} \sup_{i \in \mathcal{S}} \|P^{(m,n)}(i, \cdot) - \pi^*\| = 0 \quad \text{for all } m.$$

For example, we will prove Theorem (2.23) by showing that under the stated conditions, the simulated annealing chain $\{X_n\}$ is strongly ergodic, with limiting distribution α^* as given in (2.17).

The reason for the modifier “strongly” is to distinguish the last concept from the following weaker one.

(2.26) DEFINITION. $\{X_n\}$ is weakly ergodic if

$$\lim_{n \rightarrow \infty} \sup_{i, j \in \mathcal{S}} \|P^{(m,n)}(i, \cdot) - P^{(m,n)}(j, \cdot)\| = 0 \quad \text{for all } m.$$

To interpret these definitions a bit, notice that weak ergodicity is a sort of “loss of memory” concept. It says that at a large enough time n , the chain has nearly “forgotten” its state at time m , in the sense that the distribution at time n would be nearly the same no matter what the state was at time m . However, there is no requirement that the distribution be converging to anything as $n \rightarrow \infty$. The concept that incorporates convergence in addition to loss of memory is strong ergodicity.

What is the role of the “for all m ” requirement? Why not just use $\lim_{n \rightarrow \infty} \sup_{i \in \mathcal{S}} \|P^{(0,n)}(i, \cdot) - \pi^*\| = 0$ for strong ergodicity and $\lim_{n \rightarrow \infty} \sup_{i,j \in \mathcal{S}} \|P^{(0,n)}(i, \cdot) - P^{(0,n)}(j, \cdot)\| = 0$ for weak ergodicity? Here are a couple of examples to show that these would not be desirable definitions. Let $\mathcal{S} = \{1, 2\}$ and $P_0 = \begin{pmatrix} 1/4 & 3/4 \\ 1/4 & 3/4 \end{pmatrix}$. Then with these definitions, $\{X_n\}$ would be strongly ergodic even if $P_k = I$ for all $k \geq 1$ and $\{X_n\}$ would be weakly ergodic for *any* sequence of probability transition matrices P_1, P_2, \dots . This seems silly. We want more “robust” concepts that cannot be determined just by one matrix P_0 , but rather depend on the whole sequence of transition matrices.

Incidentally, since our goal with respect to simulated annealing is the main theorem (2.23) above, we are really interested in proving strong ergodicity. However, we will find weak ergodicity to be a useful stepping stone on the way toward that goal.

2.6.1 The Ergodic Coefficient

This will be a useful quantity in formulating sufficient conditions for weak and strong ergodicity. For a probability transition matrix $P = (P(i, j))$, the *ergodic coefficient* $\delta(P)$ of P is defined to be the maximum total variation distance between pairs of rows of P , that is,

$$\begin{aligned} \delta(P) &= \sup_{i,j \in \mathcal{S}} \|P(i, \cdot) - P(j, \cdot)\| \\ &= \frac{1}{2} \sup_{i,j \in \mathcal{S}} \sum_{k \in \mathcal{S}} |P(i, k) - P(j, k)| \\ &= \sup_{i,j \in \mathcal{S}} \sum_{k \in \mathcal{S}} (P(i, k) - P(j, k))^+. \end{aligned}$$

The basic idea here is that $\delta(P)$ being small is “good” for ergodicity. For example, the extreme case is $\delta(P) = 0$, in which case all of the rows of P are identical, and so P would cause a Markov chain to lose its memory completely in just one step: $\nu_1 = \nu_0 P$ does not depend on ν_0 .

Here is a useful lemma.

(2.27) LEMMA. $\delta(PQ) \leq \delta(P)\delta(Q)$ for probability transition matrices P, Q .

PROOF: By definition, $\delta(PQ) = \sup_{i,j \in \mathcal{S}} \sum_{k \in \mathcal{S}} [(PQ)_{ik} - (PQ)_{jk}]^+$, where here and throughout this proof, for readability let us use subscripts to denote matrix entries. Fix a pair of

states i, j , and let $A = \{k \in \mathcal{S} : (PQ)_{ik} > (PQ)_{jk}\}$. Then

$$\begin{aligned}
 \sum_{k \in \mathcal{S}} [(PQ)_{ik} - (PQ)_{jk}]^+ &= \sum_{k \in A} [(PQ)_{ik} - (PQ)_{jk}] \\
 &= \sum_{k \in A} \sum_{l \in \mathcal{S}} [P_{il}Q_{lk} - P_{jl}Q_{lk}] \\
 &= \sum_{l \in \mathcal{S}} [P_{il} - P_{jl}] \sum_{k \in A} Q_{lk} \\
 &\leq \sum_{l \in \mathcal{S}} (P_{il} - P_{jl})^+ \left[\sup_l \sum_{k \in A} Q_{lk} \right] - \sum_{l \in \mathcal{S}} (P_{il} - P_{jl})^- \left[\inf_l \sum_{k \in A} Q_{lk} \right] \\
 &= \sum_{l \in \mathcal{S}} (P_{il} - P_{jl})^+ \left[\sup_l \sum_{k \in A} Q_{lk} - \inf_l \sum_{k \in A} Q_{lk} \right] \\
 &\leq \left[\sum_{l \in \mathcal{S}} (P_{il} - P_{jl})^+ \right] \delta(Q) \\
 &\leq \delta(P) \delta(Q),
 \end{aligned}$$

where the last equality uses the fact that $\sum_l (P_{il} - P_{jl})^+ = \sum_l (P_{il} - P_{jl})^-$, which holds because P is a probability transition matrix. Since i and j were arbitrary in \mathcal{S} , we are done. \square

▷ Exercises [2.17] and [2.18] lead to an alternative proof of Lemma (2.27) using coupling.

2.6.2 Sufficient Conditions for Weak and Strong Ergodicity

Sufficient conditions are given in the next two propositions.

(2.28) PROPOSITION. *If there exist $n_0 < n_1 < n_2 < \dots$ such that $\sum_k [1 - \delta(P^{(n_k, n_{k+1})})] = \infty$, then $\{X_n\}$ is weakly ergodic.*

(2.29) PROPOSITION. *If $\{X_n\}$ is weakly ergodic and if there exist π_0, π_1, \dots such that π_n is a stationary distribution for P_n for all n and $\sum_n \|\pi_n - \pi_{n+1}\| < \infty$, then $\{X_n\}$ is strongly ergodic. In that case, the distribution π^* in the definition (2.25) is given by $\pi^* = \lim_{n \rightarrow \infty} \pi_n$.*

Recall that strong ergodicity is like weak ergodicity [loss of memory] together with convergence. The extra condition $\sum_n \|\pi_n - \pi_{n+1}\| < \infty$ is giving this convergence in (2.29).

PROOF OF PROPOSITION (2.28). It follows directly from the definitions we have given that weak ergodicity is equivalent to the condition that $\lim_{n \rightarrow \infty} \delta(P^{(m, n)}) = 0$ for all m . By assumption, $\sum_{k \geq K} [1 - \delta(P^{(n_k, n_{k+1})})] = \infty$ for all K . We will use the following little fact about real numbers: if $0 \leq a_n \leq 1$ for all n and $\sum_k a_k = \infty$, then $\prod_k (1 - a_k) = 0$. [Proof: under the assumed conditions, $0 \leq \prod (1 - a_k) \leq \prod e^{-a_k} = e^{-\sum a_k} = 0$.] From this

we obtain $\prod_{k \geq K} \delta(P^{(n_k, n_{k+1})}) = 0$ for all K . That is, $\lim_{L \rightarrow \infty} \prod_{k=K}^{L-1} \delta(P^{(n_k, n_{k+1})}) = 0$ for all K . However, from Lemma (2.27),

$$\delta(P^{(n_K, n_L)}) \leq \prod_{k=K}^{L-1} \delta(P^{(n_k, n_{k+1})}).$$

Therefore, $\lim_{L \rightarrow \infty} \delta(P^{(n_K, n_L)}) = 0$ for all K . Clearly this implies that $\lim_{n \rightarrow \infty} \delta(P^{(m, n)}) = 0$ for all m . \square

PROOF OF PROPOSITION (2.29). Suppose that $\{X_n\}$ is weakly ergodic and $\sum \|\pi_n - \pi_{n+1}\| < \infty$, where each π_n is stationary for P_n . Let $\pi^* = \lim \pi_n$; clearly the limit exists by the assumption $\sum \|\pi_n - \pi_{n+1}\| < \infty$. Let k be fixed. Then for any $l > k$ and $m > l$ we have

$$(2.30) \quad \begin{aligned} \|P^{(k, m)}(i, \cdot) - \pi^*\| &\leq \|P^{(k, m)}(i, \cdot) - \pi_l P^{(l, m)}\| \\ &\quad + \|\pi_l P^{(l, m)} - \pi_m\| + \|\pi_m - \pi^*\|. \end{aligned}$$

Let $\epsilon > 0$. We will show that the right-hand side can be made less than ϵ if m is large enough; we'll do this by making a judicious choice of l . The last term is the simplest; clearly there exists M_3 such that $\|\pi_m - \pi^*\| \leq \epsilon/3$ for all $m \geq M_3$. For the second term, note that

$$\begin{aligned} \pi_l P^{(l, m)} &= \pi_l P_l P^{(l+1, m)} = \pi_l P^{(l+1, m)} \\ &= \pi_{l+1} P^{(l+1, m)} + [\pi_l - \pi_{l+1}] P^{(l+1, m)}, \end{aligned}$$

so that

$$\pi_l P^{(l, m)} - \pi_m = [\pi_{l+1} P^{(l+1, m)} - \pi_m] + [\pi_l - \pi_{l+1}] P^{(l+1, m)}.$$

Applying this relation recursively gives

$$\pi_l P^{(l, m)} - \pi_m = \sum_{n=l}^{m-1} [\pi_n - \pi_{n+1}] P^{(n+1, m)}$$

So $\|\pi_l P^{(l, m)} - \pi_m\| \leq \sum_{n=l}^{m-1} \|\pi_n - \pi_{n+1}\|$. [Why? **Exercise.**] Therefore, since $\sum \|\pi_n - \pi_{n+1}\| < \infty$, we can make $\|\pi_l P^{(l, m)} - \pi_m\| \leq \epsilon/3$ by taking m and l large enough, say, $m \geq l \geq L_2$.

Finally, for the first term on the right-hand side of (2.30), note that

$$\|P^{(k, m)}(i, \cdot) - \pi_l P^{(l, m)}\| = \|[P^{(k, l)}(i, \cdot) - \pi_l] P^{(l, m)}\|.$$

However, for any given l , we can make the last expression less than $\epsilon/3$ by taking m large enough—by weak ergodicity, at a large enough time m , the chain doesn't "remember" whether its distribution at time l was $P^{(k, l)}(i, \cdot)$ or π_l ! So, for all l , there is an $M_1(l)$ such that if $m \geq M_1(l)$ then $\|P^{(k, m)}(i, \cdot) - \pi_l P^{(l, m)}\| < \epsilon/3$. So we are done: if $m \geq \max\{M_3, M_1(L_2)\}$, then $\sup_i \|P^{(k, m)}(i, \cdot) - \pi^*\| \leq \epsilon$. \square

Notice how the hypotheses of the last result were used to get the terms on the right-hand side of (2.30) small: weak ergodicity took care of the first term, and $\sum \|\pi_n - \pi_{n+1}\| < \infty$ took care of the other two.

2.7 Proof of the Main Theorem of Simulated Annealing

We will show that if conditions (i)–(iii) of Theorem (2.23) hold, then the time-inhomogeneous Markov chain $\{X_n\}$ for simulated annealing satisfies the sufficient conditions for strong ergodicity.

We have used two sets of parallel notation in the last two sections; one (involving α 's and A 's) that we introduced specifically to discuss simulated annealing and one (involving π 's and P 's) used in the general theory of time-inhomogeneous Markov chains. Let's relate them and make sure there is no confusion. The general notation P_n refers to the probability transition matrix used in going from X_n to X_{n+1} . In the case of simulated annealing, since the chain is operating at temperature T_n during that transition, we have $P_n = A_{T_n}$. Similarly, π_n , the stationary distribution associated with P_n , is α_{T_n} for simulated annealing. We'll also continue to use the notation $P^{(m,n)} = \prod_{k=m}^{n-1} P_k = \prod_{k=m}^{n-1} A_{T_k}$.

To establish weak ergodicity, we want to show that the sufficient condition $\sum_k [1 - \delta(P^{(n_k, n_{k+1})})] = \infty$ holds for some sequence $\{n_k\}$. In fact, we will show that the condition holds for the sequence $n_k = kr$, that is,

$$(2.31) \quad \sum_k [1 - \delta(P^{(kr-r, kr)})] = \infty.$$

We will do this by finding an upper bound for $\delta(P^{(kr-r, kr)})$ that will guarantee that $\delta(P^{(kr-r, kr)})$ does not get too close to 1.

Recall the definition of the radius $r = \min_{i \in \mathcal{S}_c} \max_{j \in \mathcal{S}} \rho(i, j)$. Let i_0 denote a *center* of the graph, that is, a node at which the minimum in the definition of r is assumed. Also recall the definition $L = \max_{i \in \mathcal{S}} \max_{j \in \mathcal{N}(i)} |c(j) - c(i)|$.

(2.32) CLAIM. *For all sufficiently large m we have*

$$P^{(m-r, m)}(i, i_0) \geq D^{-r} \exp(-rL/T_{m-1}) \quad \text{for all } i \in \mathcal{S},$$

where $D = \max_j d(j)$.

PROOF: It follows immediately from the definitions of $P_{ij}(n)$, D , and L that for all n , $i \in \mathcal{S}$, and $j \in \mathcal{N}(i)$, we have

$$P_{ij}(n) = \frac{1}{d(i)} \min\{1, e^{-(c(j)-c(i))/T_n}\} \geq D^{-1} e^{-L/T_n}.$$

Since we did not allow i_0 to be a local maximum of $c(\cdot)$, we must have $\{j \in \mathcal{N}(i_0) : c(j) > c(i_0)\}$ nonempty, so that as $n \rightarrow \infty$,

$$P_{i_0, i_0}(n) = 1 - \sum_{j \in \mathcal{N}(i_0)} P_{i_0 j}(n) \rightarrow 1 - \sum_{\substack{j \in \mathcal{N}(i_0) \\ c(j) \leq c(i_0)}} \frac{1}{d(i_0)} = \sum_{\substack{j \in \mathcal{N}(i_0) \\ c(j) > c(i_0)}} \frac{1}{d(i_0)} > 0.$$

Therefore, $P_{i_0, i_0}(n) \geq D^{-1} \exp(-L/T_n)$ clearly holds for large enough n .

Let $i \in \mathcal{S}$, and consider $P^{(m-r, m)}(i, i_0) = \mathbb{P}\{X_m = i_0 \mid X_{m-r} = i\}$. Clearly this probability is at least the conditional probability that, starting from i at time $m-r$, the

chain takes a specified shortest path from i to i_0 [which is of length at most r , by the definition of r], and then holds at i_0 for the remaining time until m . However, by the previous paragraph, if m is large enough, this probability is in turn bounded below by

$$\prod_{n=m-r}^{m-1} D^{-1} e^{-L/T_n} \geq D^{-r} e^{-rL/T_{m-1}},$$

where we have used the assumption that T_n is decreasing in n here. Thus, $P^{(m-r,m)}(i, i_0) \geq D^{-r} e^{-rL/T_{m-1}}$ for large enough m . \square

Taking $m = kr$ in the last claim, we get for all sufficiently large k that

$$P^{(kr-r,kr)}(i, i_0) \geq D^{-r} e^{-rL/T_{kr-1}} \quad \text{for all } i.$$

Thus, $P^{(kr-r,kr)}$ is a probability transition matrix having a column [namely, column i_0] all of whose entries are at least $D^{-r} \exp(-rL/T_{kr-1})$. Next we use the general observation that if a probability transition matrix Q has a column all of whose entries are at least a , then $\delta(Q) \leq 1 - a$. [Exercise 2.19 asks you to prove this.] Therefore, for large enough k , we have $1 - \delta(P^{(kr-r,kr)}) \geq D^{-r} \exp(-rL/T_{kr-1})$, which, by assumption (iii) of the main theorem, gives (2.31). This completes the proof of weak ergodicity.

Finally, we turn to the proof of strong ergodicity. Recall that the stationary distribution π_n for P_n is given by $\pi_n(i) = (G(n))^{-1} d(i) \exp[-c(i)/T_n]$. By our sufficient conditions for strong ergodicity, we will be done if we can show that $\sum \|\pi_{n+1} - \pi_n\| < \infty$. This will be easy to see from the following monotonicity properties of the stationary distributions π_n .

(2.33) LEMMA. *If $i \in \mathcal{S}^*$ then $\pi_{n+1}(i) > \pi_n(i)$ for all n . If $i \notin \mathcal{S}^*$ then there exists \tilde{n}_i such that $\pi_{n+1}(i) < \pi_n(i)$ for all $n \geq \tilde{n}_i$.*

Thus, as the temperature decreases, the stationary probabilities of the optimal states increase. Also, for each nonoptimal state, as the temperature decreases to 0, the stationary probability of that state decreases eventually, that is, when the temperature is low enough. The proof is calculus; just differentiate away. I'll leave this as an exercise.

From this nice, monotonic behavior, the desired result follows easily. Letting \tilde{n} denote $\max\{\tilde{n}_i : i \notin \mathcal{S}^*\}$,

$$\begin{aligned} \sum_{n=\tilde{n}}^{\infty} \|\pi_{n+1} - \pi_n\| &= \sum_{n=\tilde{n}}^{\infty} \sum_{i \in \mathcal{S}} (\pi_{n+1} - \pi_n)^+ \\ &= \sum_{n=\tilde{n}}^{\infty} \sum_{i \in \mathcal{S}^*} (\pi_{n+1} - \pi_n) \\ &= \sum_{i \in \mathcal{S}^*} \sum_{n=\tilde{n}}^{\infty} (\pi_{n+1} - \pi_n) \\ &= \sum_{i \in \mathcal{S}^*} (\pi^*(i) - \pi_{\tilde{n}}(i)) \leq \sum_{i \in \mathcal{S}^*} \pi^*(i) = 1, \end{aligned}$$

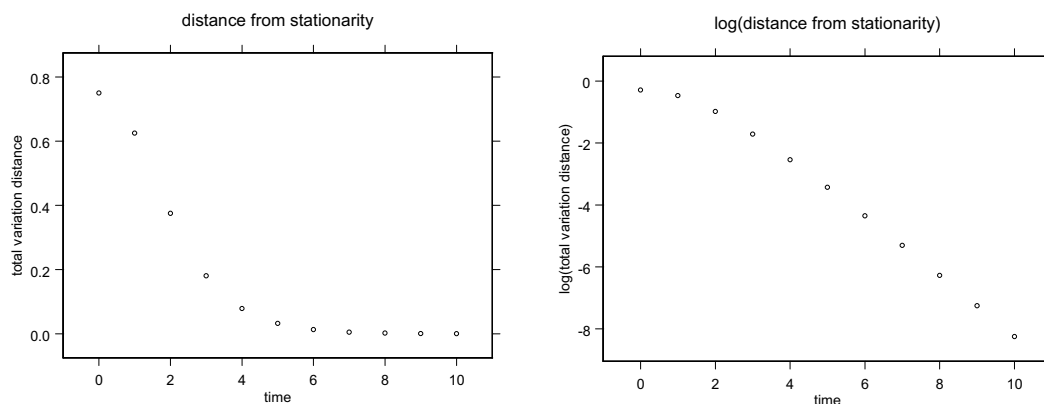
so that clearly $\sum_{n=\tilde{n}}^{\infty} \|\pi_{n+1} - \pi_n\| < \infty$. This completes the proof of strong ergodicity.

REFERENCES: This proof came from a paper of Mitra, Romeo, and Sangiovanni-Vincentelli called “Convergence and finite-time behavior of simulated annealing” [*Advances in Applied Probability*, **18**, 747–771 (1986)]. The material on time-inhomogeneous Markov chains is covered very nicely in the book *Markov chains: Theory and Applications* by Isaacson and Madsen (1976).

2.8 Card Shuffling: Speed of Convergence to Stationarity

We have seen that for an irreducible, aperiodic Markov chain $\{X_n\}$ having stationary distribution π , the distribution π_n of X_n converges to π in the total variation distance. For example, this was used in Example (1.22), which showed how to generate a nearly uniformly distributed 4×4 table having given row and column sums by simulating a certain Markov chain for a long enough time. The inevitable question is: How long is “long enough”? We could ask how close (in total variation, say) is the Markov chain to being uniformly distributed after 100 steps? How about 1000? 50 billion?

In certain simple Markov chain examples we have discussed, it is easy to figure out the rate of convergence of π_n to π . For instance, for the Markov frog example (1.2), starting in the initial distribution $\pi_0 = (1, 0, 0)$, say, we can compute the distributions π_1, π_2, \dots by matrix multiplication and compare them to the stationary distribution $\pi = (1/4, 3/8, 3/8)$, getting the results shown in Figure (2.34).



(2.34) FIGURE. *Speed of convergence to stationarity in Markov frog example from Chapter 1.*

Notice the smooth geometric rate of decrease: the log distance decreases linearly, which means that the distance decreases to 0 geometrically. Coupling is a technique that can sometimes shed some light on questions of this sort. For example, in Exercise (1.27) we showed that $\|\pi_n - \pi\| \leq \frac{2}{3} \left(\frac{11}{16}\right)^n$ for all n , [and, in fact, $\|\pi_n - \pi\| \leq \frac{2}{3} \left(\frac{1}{4}\right)^n$] which gives much more information than just saying that $\|\pi_n - \pi\| \rightarrow 0$.

In this section we will concentrate on a simple shuffling example considered by Aldous and Diaconis in their article “Shuffling cards and stopping times.” Again, the basic question

is: How close is the deck to being “random” (i.e. uniformly distributed over the $52!$ possible permutations) after n shuffles? Or, put another way, how many shuffles does it take to shuffle well? Of course, the answer depends on the details of the method of shuffling; Aldous and Diaconis say that for the riffle shuffle model, which is rather realistic, the answer is “about 7.” In this sort of problem, in contrast to the smooth and steady sort of decrease depicted in Figure (2.34), we will see that an interesting *threshold phenomenon* arises.

2.8.1 “Top-in-at-random” Shuffle

This shuffle, which is the model of shuffling we will analyze in detail, is simpler than the riffle shuffle for which Aldous and Diaconis gave the answer “about 7 is enough.” The analysis we will discuss here also comes from their paper.

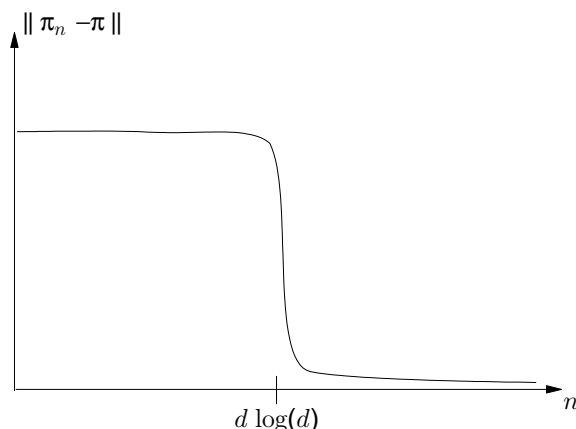
The top-in-at-random method would be a really silly way to shuffle a deck of cards in practice, but it is an appealingly simple model for a first example to study. One such shuffle consists of taking the top card off the deck and then inserting it back into the deck in a random position. “Random” here means that the top card could end up back on top (in which case the shuffle didn’t change the deck at all), or it could end up second from top, third from top, ..., or at the bottom of the deck: altogether 52 equally likely positions.

Repeated performances of this shuffle on a deck of cards produces a sequence of “states” of the deck. This sequence of states forms a Markov chain $\{X_n\}$ having state space \mathcal{S}_{52} , the group of permutations of the cards. This Markov chain is irreducible, aperiodic, and has stationary distribution $\pi = \text{Uniform on } \mathcal{S}_{52}$ (i.e. probability $1/(52!)$ for each permutation); therefore, by the Basic Limit Theorem, we may conclude that $\|\pi_n - \pi\| \rightarrow 0$ as $n \rightarrow \infty$.

▷ *Explaining each assertion in the previous sentence is Exercise [2.20].*

2.8.2 Threshold Phenomenon

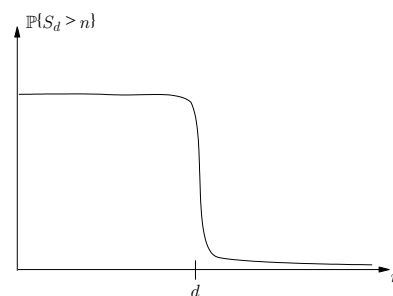
Suppose we are working with a fresh deck of d cards, which we know are in the original order: card 1 on top, card 2 just below card 1, ..., and card d on the bottom. Then $\|\pi_0 - \pi\| = 1 - (1/d!)$. We also know that $\|\pi_n - \pi\| \rightarrow 0$ as $n \rightarrow \infty$, by the Basic Limit Theorem. It is natural to presume that the distance from stationarity $\|\pi_n - \pi\|$ decreases to 0 in some smooth, uneventful manner. However, the fact of the matter is that $\|\pi_n - \pi\|$ stays close to 1 for a while, then it undergoes a rather abrupt decrease from nearly 1 to nearly 0, and this abrupt change happens in a relatively small neighborhood of the value $n = d \log d$. That is, for large d the graph of $\|\pi_n - \pi\|$ versus n looks rather like the next picture.



The larger the value of the deck size d , the sharper (relative to $d \log d$) the drop in $\|\pi_n - \pi\|$ near $n = d \log d$.

Well, this is hardly a gradual or uneventful decrease! This interesting behavior of $\|\pi_n - \pi\|$ has been called the *threshold phenomenon*. The phenomenon is not limited to this particular shuffle or even to shuffling in general, but rather seems to occur in a wide variety of interesting Markov chains. Aldous and Diaconis give a number of examples in which threshold phenomena can be shown to occur, but they point out that the phenomenon is not yet very well understood, in the sense that general conditions under which the phenomenon does or does not occur are not known.

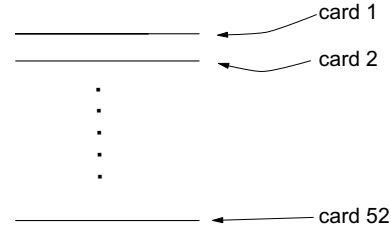
This seems weird, doesn't it? As a partial antidote to the uneasy feelings of mystery that tend to accompany a first exposure to the threshold phenomenon idea, let's think about a simple problem that is familiar to all of us, in which a threshold of sorts also occurs. Suppose X_1, X_2, \dots are *iid* random variables with mean 1 and variance 1. In fact, for simplicity, let's suppose that they have the Normal distribution $N(1, 1)$. Thinking of d as a large number,



let $S_d = X_1 + \dots + X_d$, and consider the probability $\mathbb{P}\{S_d > n\}$ as a function of n . Since $S_d \sim N(d, d)$, it is easy to see that the graph of $\mathbb{P}\{S_d > n\}$ has a “threshold” near $n = d$ if d is large. In fact, the length of a neighborhood about $n = d$ in which $\mathbb{P}\{S_d > n\}$ decreases from nearly 1 to nearly 0 is of order \sqrt{d} . However, if d is large, then \sqrt{d} is vanishingly small compared with d . Thus, for large d , the graph of $\mathbb{P}\{S_d > n\}$ versus n (plotted on a scale in which $n = d$ is at some moderate, fixed distance from $n = 0$) will indeed appear as a sharp dropoff near $n = d$. In particular, note that the existence of a threshold for large d does *not* say that the dropoff from near 1 to near 0 takes place over a shorter and shorter time interval as d increases; it is just that the length (here $O(\sqrt{d})$) of that dropoff interval is smaller and smaller in comparison with the location (here around d) of that interval.

2.8.3 A random time to exact stationarity

Let's give a name to each card in the deck: say "card 1" is $2\heartsuit$, "card 2" is $3\heartsuit$, ..., "card 51" is $K\spadesuit$, "card 52" is $A\spadesuit$. Suppose we start with the deck in the pristine order shown to the right. Wouldn't it be nice if we could say, "After 1000 shuffles the deck will be exactly random," or maybe "After 1,000,000 shuffles the deck will be exactly random"? Well, sorry. We can't. We know π_n gets closer and closer to the uniform distribution as n increases, but unfortunately π_n will *never* become *exactly* random, even if n is 53 bezillion.



However, it *is* possible to find a *random* time T at which the deck becomes exactly uniformly distributed, that is, $X_T \sim \text{Unif}(\mathcal{S}_{52})$! Here is an example of such a random time. To describe it, let's agree that "card i " always refers to the same card [e.g. card 52 = $A\spadesuit$], while terms like "top card," "card in position 2," and so on just refer to whatever card happens to be on top, in position 2, and so on at the time under consideration. Also note that we may describe a sequence of shuffles simply by a sequence of *iid* random variables U_1, U_2, \dots uniformly distributed on $\{1, 2, \dots, 52\}$: just say that the i th shuffle moves the top card to position U_i . Define the following random times:

$$\begin{aligned} T_1 &= \inf\{n : U_n = 52\} = \text{1st time a top card goes below card 52,} \\ T_2 &= \inf\{n > T_1 : U_n \geq 51\} = \text{2nd time a top card goes below card 52,} \\ T_3 &= \inf\{n > T_2 : U_n \geq 50\} = \text{3rd time a top card goes below card 52,} \\ &\vdots \\ T_{51} &= \inf\{n > T_{50} : U_n \geq 2\} = \text{51st time a top card goes below card 52,} \end{aligned}$$

and

$$(2.35) \quad T = T_{52} = T_{51} + 1.$$

It is not hard to see that T has the desired property and that X_T is uniformly distributed. To understand this, start with T_1 . At time T_1 , we know that some card is below card 52; we don't know which card, but that will not matter. After time T_1 we continue to shuffle until T_2 , at which time another card goes below card 52. At time T_2 , there are 2 cards below card 52. Again, we do not know which cards they are, but *conditional on which 2 cards are below card 52, each of the two possible orderings of those 2 cards is equally likely*. Similarly, we continue to shuffle until time T_3 , at which time there are some 3 cards below card 52, and, whatever those 3 cards are, each of their $(3!)$ possible relative positions in the deck is equally likely. And so on. At time T_{51} , card 52 has risen all the way up to become the top card, and the other 51 cards are below card 52 (now we *do* know which cards they are), and those 51 cards are in random positions (i.e. uniform over $51!$ possibilities). Now all we have to do is shuffle one more time to get card 52 in random position, so that at time $T = T_{52} = T_{51} + 1$, the whole deck is random.

Let us find ET . Clearly by the definitions above we have $T_1 \sim \text{Geom}(1/52)$, $(T_2 - T_1) \sim \text{Geom}(2/52)$, ..., $(T_{51} - T_{50}) \sim \text{Geom}(51/52)$, and $(T_{52} - T_{51}) \sim \text{Geom}(52/52) = 1$. Therefore,

$$\begin{aligned} ET &= E(T_1) + E(T_2 - T_1) + \cdots + (T_{51} - T_{50}) + E(T_{52} - T_{51}) \\ &= 52 + (52/2) + (52/3) + \cdots + (52/51) + (52/52) \\ &= 52 \left(1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{51} + \frac{1}{52} \right) \approx 52 \log 52. \end{aligned}$$

Analogously, if the deck had d cards rather than 52, we would have obtained $ET \sim d \log d$ (for large d), where T is now a random time at which the whole deck of d cards becomes uniformly distributed on \mathcal{S}_d .

2.8.4 Strong Stationary Times

The main new idea in the analysis of the shuffling example is that of a strong stationary time. As we have observed, the random variable T that we just constructed has the property that $X_T \sim \pi$. T also has two other important properties. First, X_T is independent of T . Second, T is a **stopping time**, that is, for each n , one can determine whether or not $T = n$ just by looking at the values of X_0, \dots, X_n . In particular, to determine whether or not $T = n$ it is not necessary to know any “future” values X_{n+1}, X_{n+2}, \dots .

A random variable having the three properties just enumerated is called a strong stationary time.

(2.36) DEFINITION. *A random variable T satisfying*

- (i) *T is a stopping time,*
- (ii) *X_T is distributed as π , and*
- (iii) *X_T is independent of T*

*is called a **strong stationary time**.*

So what’s so good about strong stationary times? For us, the answer is contained in the following theorem, which says that strong stationary times satisfy the same inequality that we derived for coupling times in (1.37).

(2.37) LEMMA. *If T is a strong stationary time for the Markov chain $\{X_n\}$, then $\|\pi_n - \pi\| \leq \mathbb{P}\{T > n\}$ for all n .*

PROOF: Letting $A \subseteq \mathcal{S}$, we will begin by showing that

$$(2.38) \quad \mathbb{P}\{T \leq n, X_n \in A\} = \mathbb{P}\{T \leq n\}\pi(A).$$

To see this, let $k \leq n$ and write $\mathbb{P}\{T = k, X_n \in A\} = \sum_i \mathbb{P}\{T = k, X_k = i, X_n \in A\} = \sum_i \mathbb{P}\{T = k, X_k = i\} \mathbb{P}\{X_n \in A \mid T = k, X_k = i\}$. But $\mathbb{P}\{X_n \in A \mid T = k, X_k = i\} = \mathbb{P}\{X_n \in A \mid X_k = i\} =: P^{n-k}(i, A)$ by the Markov property and the assumption that T

is a stopping time. Also, $\mathbb{P}\{T = k, X_k = i\} = \mathbb{P}\{T = k, X_T = i\} = \mathbb{P}\{T = k\}\mathbb{P}\{X_T = i\} = \mathbb{P}\{T = k\}\pi(i)$ by properties (ii) and (iii) of the definition of strong stationary time. Substituting gives $\mathbb{P}\{T = k, X_n \in A\} = \mathbb{P}\{T = k\} \sum_i \pi(i) P^{n-k}(i, A) = \mathbb{P}\{T = k\}\pi(A)$ by the stationarity of π . Summing this over $k \leq n$ gives (2.38).

Next, similarly to the proof of the coupling inequality, we decompose according to whether $T \leq n$ or $T > n$ to show that for $A \subseteq \mathcal{S}$

$$\begin{aligned} \pi_n(A) - \pi(A) &= \mathbb{P}\{X_n \in A\} - \pi(A) = \mathbb{P}\{X_n \in A, T \leq n\} + \mathbb{P}\{X_n \in A, T > n\} - \pi(A) \\ &= \pi(A)\mathbb{P}\{T \leq n\} + \mathbb{P}\{X_n \in A, T > n\} - \pi(A) \quad \text{by (2.38)} \\ &= \mathbb{P}\{X_n \in A, T > n\} - \pi(A)\mathbb{P}\{T > n\}. \end{aligned}$$

Since each of the last two quantities lies between 0 and $\mathbb{P}\{T > n\}$, we conclude that $|\pi_n(A) - \pi(A)| \leq \mathbb{P}\{T > n\}$, so that $\|\pi_n - \pi\| \leq \mathbb{P}\{T > n\}$. \square

2.8.5 Proof of threshold phenomenon in shuffling

Let $\Delta(n)$ denote $\|\pi_n - \pi\|$. The proof that the threshold phenomenon occurs in the top-in-at-random shuffle consists of two parts: Roughly speaking, the first part shows that $\Delta(n)$ is close to 0 for n slightly larger than $d \log d$, and the second part shows that $\Delta(n)$ is close to 1 for n slightly smaller than $d \log d$, where in both cases the meaning of “slightly” is “small relative to $d \log d$.”

The first part is addressed by the next result.

(2.39) THEOREM. For T as defined in (2.35), we have

$$\Delta(d \log d + cd) \leq \mathbb{P}\{T > d \log d + cd\} \leq e^{-c} \quad \text{for all } c \geq 0.$$

Note that for each fixed c , cd is small relative to $d \log d$ if d is large enough.

PROOF: The first inequality in (2.39) is just Lemma (2.37), so our task is to prove the second inequality. Recall that, as discussed above, $T_1 \sim \text{Geom}(1/d)$, $T_2 - T_1 \sim \text{Geom}(2/d)$, $T_3 - T_2 \sim \text{Geom}(3/d)$, ..., and $T - T_{d-1} = T_d - T_{d-1} \sim \text{Geom}(d/d) = 1$. It is also clear that $T_1, T_2 - T_1, T_3 - T_2, \dots$, and $T - T_{d-1}$ are independent. Thus,

$$T \sim \text{Geom}\left(\frac{1}{d}\right) \oplus \text{Geom}\left(\frac{2}{d}\right) \oplus \cdots \oplus \text{Geom}\left(\frac{d-1}{d}\right) \oplus 1,$$

where the symbol “ \oplus ” indicates a sum of independent random variables. However, observe that the distribution $1 \oplus \text{Geom}[(d-1)/d] \oplus \cdots \oplus \text{Geom}[1/d]$ is also the distribution that arises in the famous *coupon collector's problem*. [To review the coupon collector's problem: Suppose that each box of Raisin Bran cereal contains one of d possible coupons numbered $\{1, \dots, d\}$, with the coupons in different boxes being independent and uniformly distributed on $\{1, \dots, d\}$. The number of cereal boxes a collector must buy in order to obtain a complete set of d coupons has the distribution $1 \oplus \text{Geom}[(d-1)/d] \oplus \cdots \oplus \text{Geom}[1/d]$.]

To find a bound on $\mathbb{P}\{T > n\}$, let us adopt this coupon collecting interpretation of T . For each $i = 1, \dots, d$ define an event

$$B_i = \{\text{coupon } i \text{ does not appear among the first } n \text{ cereal boxes}\}.$$

Then the event $T > n$ is just the union $\bigcup_{i=1}^d B_i$, so that

$$\mathbb{P}\{T > n\} \leq \sum_{i=1}^d \mathbb{P}(B_i) = \sum_{i=1}^d \left(\frac{d-1}{d}\right)^n = d \left(1 - \frac{1}{d}\right)^n \leq de^{-n/d},$$

where the last inequality uses the fact that $1 - x \leq e^{-x}$ for all numbers x . Setting $n = d \log d + cd$ gives

$$\mathbb{P}\{T > n\} \leq de^{-(\log d + c)} = e^{-c},$$

as desired. □

For the second part of the proof, let us temporarily be a bit more fastidious about notation: Instead of just writing π_n and π , let us write $\pi_n^{(d)}$ and $\pi^{(d)}$ to indicate explicitly the dependence of the various distributions on the deck size d .

(2.40) THEOREM. *Let $k(d) = d \log d - c_d d$, where $\{c_d\}$ is a sequence of numbers that approaches infinity as $d \rightarrow \infty$. Then*

$$\|\pi_{k(d)}^{(d)} - \pi^{(d)}\| \rightarrow 1 \quad \text{as } d \rightarrow \infty.$$

NOTE: The case of interest for establishing the threshold at $d \log d$ is when $c_d = o(\log d)$, since in that case $k(d) \sim d \log d$.

PROOF: Let's start with some fumbling around, intended to provide some glimmer of hope that we might have been able to think of this proof ourselves. The proof proceeds by bounding $\|\pi_{k(d)}^{(d)} - \pi^{(d)}\|$ below by something that is close to 1 for large d . By the definition of the total variation distance $\|\cdot\|$, this may be done by finding events A_d such that $\|\pi_{k(d)}^{(d)}(A_d) - \pi^{(d)}(A_d)\|$ is close to 1. OK, now let's drop those pesky d 's from the notation, and say that for large d , *we want to find events A such that $\pi_k(A)$ is large (close to 1) while $\pi(A)$ is small (close to 0)*. Fumbling time...

- How about $A = \{\text{card } d \text{ is still on the bottom}\}$?
 - Is $\pi(A)$ small? Yes: $\pi(A) = 1/d$.
 - Is $\pi_k(A)$ large? No: since $k \gg d$, clearly $\mathbb{P}\{T_1 > k\} = \mathbb{P}\{\text{Geom}(1/d) > k\}$ is not large, so that $\mathbb{P}\{\text{card } d \text{ is still on the bottom at time } k\}$ is not large.
- How about $A = \{\text{cards } d-1 \text{ and } d \text{ are still in their original positions}\}$?
 - Is $\pi(A)$ small? Yes: $\pi(A) = 1/[d(d-1)]$.
 - Is $\pi_k(A)$ large? No: in fact, it is smaller than the previous $\pi_k(A)$ we just considered. You should be ashamed of yourself for that suggestion!

- How about just requiring cards $d - 1$ and d still be in their original *order*, that is, $A = \{\text{card } d - 1 \text{ still above card } d \text{ in the deck}\}$?
 - Is $\pi_k(A)$ large? Maybe; this doesn't seem very obvious.
 - Is $\pi(A)$ small? No: $\pi(A) = 1/2$.
- Well, that may look discouraging. But with a little more thought we can at least extend the previous idea to get $\pi(A)$ small while keeping a “maybe” for $\pi_k(A)$ being large, as follows. How about

$$A = A_{d,a} = \{\text{cards } d - a + 1, d - a + 2, \dots, d \text{ still in their original order}\}?$$

- Is $\pi_k(A)$ large? Still maybe.
- Is $\pi(A)$ small? $\pi(A) = 1/(a!)$, so **yes** if a increases with d .

Let's review what we are doing. We are given a sequence of numbers $\{c_d\}$ such that $c_d \rightarrow \infty$ as $d \rightarrow \infty$. [As noted, we are interested in the case where we also have $c_d = o(\log d)$, but the proof will not require this.] For each d , we have also defined a number $k = k(d) = d \log d - c_d d$. For each d and each $a \leq d$ we have identified an event $A = A_{d,a}$. What we want to show is that there is a sequence $\{a_d\}$ of values of a such that, as $d \rightarrow \infty$, we have $\pi_k(A) \rightarrow 1$ and $\pi(A) \rightarrow 0$. [Actually we should write these statements as $\pi_{k(d)}^{(d)}(A_{d,a_d}) \rightarrow 1$ and $\pi^{(d)}(A_{d,a_d}) \rightarrow 0$, but I doubt any of us really wants that.]

As for getting the second statement, since $\pi(A) = 1/(a!)$, any sequence $\{a_d\}$ that tends to infinity as $d \rightarrow \infty$ will suffice.

To get the first statement to hold we need a little more analysis. Suppose that in k shuffles, card $d - a + 1$ has not yet “risen to the top of the deck.” In this case, clearly the event A occurs. Letting U denote the number of shuffles required for card $d - a + 1$ to rise to the top of the deck, we thus have

$$\pi_k(A) \geq \mathbb{P}\{U > k\}.$$

Note that

$$U \sim \text{Geom}\left(\frac{a}{d}\right) \oplus \text{Geom}\left(\frac{a+1}{d}\right) \oplus \dots \oplus \text{Geom}\left(\frac{d-1}{d}\right).$$

The plan now is to use Chebyshev's inequality to show that we can cause $\mathbb{P}\{U > k\} \rightarrow 1$ to hold by choosing a_d appropriately. This will show that $\pi_k(A) \rightarrow 1$, and hence prove the theorem.

The ingredients needed to use Chebyshev are $\mathbb{E}(U)$ and $\text{Var}(U)$. Since $\mathbb{E}[\text{Geom}(p)] = 1/p$, we have

$$\mathbb{E}(U) = d \left(\frac{1}{a} + \frac{1}{a+1} + \dots + \frac{1}{d-1} \right) = d(\log d - \log a + o(1))$$

where the second equality assumes only that $a_d \rightarrow \infty$. Next, since $\text{Var}[\text{Geom}(p)] = (1-p)/p^2 \leq 1/p^2$, using independence gives

$$\text{Var}(U) \leq d^2 \left(\frac{1}{a^2} + \frac{1}{(a+1)^2} + \dots \right) =: \epsilon(a)d^2,$$

where $\epsilon(a) \rightarrow 0$ as $a \rightarrow \infty$ (or $d \rightarrow \infty$).

So here it is in a nutshell. Since $\text{Var}(U) = o(d^2)$, so that U has standard deviation $\text{SD}(U) = o(d)$, all we have to do is choose a_d so that the difference

$$\mathbb{E}(U) - k = d(\log d - \log a_d + o(1)) - d(\log d - c_d) \sim d(c_d - \log a_d)$$

is large compared with $\text{SD}(U)$, that is, at least the order of magnitude of d . But that's easy; for example, if we choose $a_d = e^{c_d/2}$, then $\mathbb{E}(U) - k \sim d(c_d/2)$, whose order of magnitude is larger than d .

To say this in a bit more detail, choose $a_d = e^{c_d/2}$, say. [Many other choices would also do.] Then of course $a_d \rightarrow \infty$. So we have

$$\begin{aligned} \mathbb{P}\{U > k(d)\} &= \mathbb{P}\{U > d(\log d - c_d)\} \\ &= \mathbb{P}\{U - \mathbb{E}(U) > d(\log d - c_d) - d(\log d - \log a_d + o(1))\} \\ &= \mathbb{P}\{U - \mathbb{E}(U) > -d(c_d - \log a_d + o(1))\}. \end{aligned}$$

Substituting $a_d = e^{c_d/2}$, this becomes

$$\begin{aligned} \mathbb{P}\{U > k(d)\} &= \mathbb{P}\{U - \mathbb{E}(U) > -d(c_d/2 + o(1))\} \\ &\geq \mathbb{P}\{|U - \mathbb{E}(U)| < d(c_d/2 + o(1))\} \\ &\geq 1 - \frac{\text{Var}(U)}{d^2(c_d/2 + o(1))^2} \\ &\geq 1 - \frac{\epsilon(a_d)}{(c_d/2 + o(1))^2}. \end{aligned}$$

Since the last expression approaches 1 as $d \rightarrow \infty$, we are done. \square

This completes our analysis of the top-in-at-random shuffle. There are lots of other interesting things to look at in the Aldous and Diaconis paper as well as some of the other references. For example, a book of P. Diaconis applies group representation theory to this sort of problem. The paper of Brad Mann is a readable treatment of the riffle shuffle.

2.9 Exercises

- [2.1] For a branching process $\{G_t\}$ with $G_0 = 1$, define the probability generating function of G_t to be ψ_t , that is,

$$\psi_t(z) = \mathbb{E}(z^{G_t}) = \sum_{k=0}^{\infty} z^k \mathbb{P}\{G_t = k\}.$$

With ψ defined as in (2.1), show that $\psi_1(z) = \psi(z)$, $\psi_2(z) = \psi(\psi(z))$, $\psi_3(z) = \psi(\psi(\psi(z)))$, and so on.

- [2.2] With ψ_t defined as in Exercise [2.1], show that $\mathbb{P}\{G_t = 1\} = \psi'_t(0)$.

- [2.3] Consider a branching process with offspring distribution $\text{Poisson}(2)$, that is, Poisson with mean 2. Calculate the extinction probability ρ to four decimal places.
- [2.4] As in the previous exercise, consider again a branching process with offspring distribution $\text{Poisson}(2)$. We know that the process will either go extinct or diverge to infinity, and the probability that it is any fixed finite value should converge to 0 as $t \rightarrow \infty$. In this exercise you will investigate how fast such probabilities converge to 0. In particular, consider the probability $\mathbb{P}\{G_t = 1\}$, and find the limiting ratio

$$\lim_{t \rightarrow \infty} \frac{\mathbb{P}\{G_{t+1} = 1\}}{\mathbb{P}\{G_t = 1\}}.$$

This may be interpreted as a rate of geometric decrease of $\mathbb{P}\{G_t = 1\}$.

[Hint: use the result of Exercise [2.2](#).]

- [2.5] Consider a branching process $\{G_t\}$ with $G_0 = 1$ and offspring distribution $f(k) = q^k p$ for $k = 0, 1, \dots$, where $q = 1 - p$. So f is the probability mass function of $X - 1$, where $X \sim \text{Geom}(p)$.

(a) Show that

$$\frac{\psi(z) - (p/q)}{\psi(z) - 1} = \frac{p}{q} \left(\frac{z - (p/q)}{z - 1} \right).$$

(b) Derive the expressions

$$\psi_t(z) = \begin{cases} \frac{p[(q^t - p^t) - qz(q^{t-1} - p^{t-1})]}{q^{t+1} - p^{t+1} - qz(q^t - p^t)} & \text{if } p \neq 1/2 \\ \frac{t - (t-1)z}{t+1 - tz} & \text{if } p = 1/2. \end{cases}$$

[Hint: The first part of the problem makes this part quite easy. If you are finding yourself in a depressing, messy calculation, you are missing the easy way. For $p \neq 1/2$, consider the fraction $[\psi_t(z) - (p/q)]/[\psi_t(z) - 1]$.]

(c) What is the probability of ultimate extinction, as a function of p ?

[Hint: Observe that $\mathbb{P}\{G_t = 0\} = \psi_t(0)$.]

- [2.6] Let $\{G_t\}$ be a supercritical (i.e. $\mu = \mathbb{E}(X) > 1$) branching process with extinction probability $\rho \in (0, 1)$. Let $B = \bigcup \{G_t = 0\}$ denote the event of eventual extinction.

(a) Show that $\mathbb{E}(z^{G_t} \mid B) = (1/\rho)\psi_t(\rho z)$.

(b) Consider again the example of Exercise [2.5](#), with $p < 1/2$. Let $\{\tilde{G}_t\}$ be a branching process of the same form as $\{G_t\}$, except with the probabilities p and q interchanged. So $\{\tilde{G}_t\}$ is subcritical, and goes extinct with probability 1. Show that the G process, conditional on the event B , behaves like the \tilde{G} process, in the sense that $\mathbb{E}(z^{G_t} \mid B) = E(z^{\tilde{G}_t})$.

(c) Isn't that interesting?

- [2.7] Consider an irreducible, time-reversible Markov chain $\{X_t\}$ with $X_t \sim \pi$, where the distribution π is stationary. Let A be a subset of the state space. Let $0 < \alpha < 1$, and define on the same state space a Markov chain $\{Y_t\}$ having probability transition matrix Q satisfying, for $i \neq j$,

$$Q(i, j) = \begin{cases} \alpha P(i, j) & \text{if } i \in A \text{ and } j \notin A \\ P(i, j) & \text{otherwise.} \end{cases}$$

Define the diagonal elements $Q(i, i)$ so that the rows of Q sum to 1.

- (a) What is the stationary distribution of $\{Y_t\}$, in terms of π and α ?
 (b) Show that the chain $\{Y_t\}$ is also time-reversible.
 (c) Show by example that the simple relationship of part (1) need not hold if we drop the assumption that X is reversible.
- [2.8] Let $\{X_t\}$ have probability transition matrix P and initial distribution π_0 . Imagine observing the process until time n , seeing $X_0, X_1, \dots, X_{n-1}, X_n$. The time reversal of this sequence of random variables is $X_n, X_{n-1}, \dots, X_1, X_0$, which we can think of as another random process \tilde{X} . That is, given the Markov chain X , define the reversed process $\{\tilde{X}_t\}$ by $\tilde{X}_t = X_{n-t}$.

- (a) Show that

$$\mathbb{P}\{X_t = j \mid X_{t+1} = i, X_{t+2} = x_{t+2}, \dots, X_{t+n} = x_{t+n}\} = \frac{\pi_t(j)P(j, i)}{\pi_{t+1}(i)}$$

- (b) Use part (a) to show that the process $\{\tilde{X}_t\}$ is a Markov chain, although it is not time homogeneous in general.
 (c) Suppose $\{X_t\}$ has stationary distribution π , and suppose X_0 is distributed according to π . Show that the reversed process $\{\tilde{X}_t\}$ is a time-homogeneous Markov chain.
- [2.9] Let $p = (p_1, \dots, p_d)$ be a probability mass function on $\{1, \dots, d\}$. Consider the residual lifetime chain, discussed in Exercise [1.14], which has probability transition matrix

$$P = \begin{matrix} & \begin{matrix} 0 & 1 & 2 & \cdots & d-2 & d-1 \end{matrix} \\ \begin{matrix} 0 \\ 1 \\ 2 \\ \vdots \\ d-1 \end{matrix} & \begin{pmatrix} p_1 & p_2 & p_3 & \cdots & p_{d-1} & p_d \\ 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & & & & & \\ 0 & 0 & 0 & \cdots & 1 & 0 \end{pmatrix} \end{matrix}$$

and stationary distribution $\pi(i) = \mathbb{P}\{X > i\}/\mathbb{E}(X)$, where X denotes a random variable distributed according to p .

- (a) Find \tilde{P} , the probability transition matrix for the reversed chain.

- (b) In renewal theory, the time since the most recent renewal is called the *age*, and the process $\{A_t\}$, whose state A_t is the age at time t , is called the *age process*. Show that the matrix \tilde{P} that you have just found is the probability transition matrix of the age process. That is, the time-reversed residual lifetime chain is the age chain.

[2.10] Let's think about the irreducibility and aperiodicity conditions of the Basic Limit Theorem as applied to the Metropolis method. Suppose that the graph structure on \mathcal{S} is a connected graph. Let π be any distribution other than π_{rw} , the stationary distribution of a random walk on the graph. Show that the Basic Limit Theorem implies that the Metropolis chain converges in distribution to π .

[2.11] Why was the condition $\pi \neq \pi_{\text{rw}}$ needed in Exercise [2.10]?

[2.12] [Metropolis-Hastings method] For simplicity, let us assume that π is positive, so that we won't have to worry about dividing by 0. Choose any probability transition matrix $Q = (Q(i, j))$ [again, suppose it is positive], and define $P(i, j)$ for $i \neq j$ by

$$P(i, j) = Q(i, j) \min \left(1, \frac{\pi(j)Q(j, i)}{\pi(i)Q(i, j)} \right),$$

and of course define $P(i, i) = 1 - \sum_{j \neq i} P(i, j)$. Show that the probability transition matrix P has stationary distribution π . Show how the Metropolis method we have discussed is a special case of this Metropolis-Hastings method.

[2.13] [Computing project: traveling salesman problem] Make up an example of the traveling salesman problem; it could look similar to the first figure in Example (2.22) if you'd like. Write a program to implement simulated annealing and produce a sequence of figures showing various improving traveling salesman tours. You could even produce a slithering snake movie if you are so inspired.

[2.14] For simulated annealing, temperature schedules of the form (2.20) decrease excruciatingly slowly. It is reasonable to ask whether we could decrease the temperature faster and still retain a guarantee of convergence in distribution to global optima. Let c be a positive number, and consider performing simulated annealing with the cooling schedule $T_n = bn^{-c}$. Of course, this schedule decreases faster than (2.20), no matter how small c is. Can you give an example that shows that such a schedule decreases too fast, in the sense that the process has positive probability of getting stuck in a local minimum forever? Thus, even $T_n = n^{-.0001}$ cools "too fast"!

[2.15] [A creative writing, essay-type question] Do you care about convergence in distribution to a global minimum? Does this property of simulated annealing make you happy?

[2.16] Prove (2.18).

- [2.17] Here is yet another interpretation of total variation distance. Let μ and ν be distributions on a finite set \mathcal{S} . Show that

$$\|\mu - \nu\| = \min \mathbb{P}\{X \neq Y\},$$

where the minimum is taken over all \mathbb{P} , X , and Y such that X has distribution μ and Y has distribution ν .

- [2.18] Prove Lemma (2.27) using coupling.

Hint: Defining $R = PQ$, we want to show that for all i and j ,

$$\|R_{i\cdot} - R_{j\cdot}\| \leq \sup_{k,l} \|P_{k\cdot} - P_{l\cdot}\| \sup_{k,l} \|Q_{k\cdot} - Q_{l\cdot}\|.$$

Construct Markov chains $X_0 \xrightarrow{P} X_1 \xrightarrow{Q} X_2$ and $Y_0 \xrightarrow{P} Y_1 \xrightarrow{Q} Y_2$ with $X_0 = i$ and $Y_0 = j$. Take (X_1, Y_1) to be a coupling achieving the total variation distance $\|P_{i\cdot} - P_{j\cdot}\|$. Then, conditional on (X_1, Y_1) , take X_2 and Y_2 to achieve the total variation distance $\|Q_{X_1\cdot} - Q_{Y_1\cdot}\|$.

- [2.19] Show that if a probability transition matrix Q has a column all of whose entries are at least a , then $\delta(Q) \leq 1 - a$.
- [2.20] Repeated performances of the top-in-at-random shuffle on a deck of cards produces a Markov chain $\{X_n\}$ having state space \mathcal{S}_{52} , the group of permutations of the cards. Show that this Markov chain is irreducible, aperiodic, and has stationary distribution $\pi = \text{Uniform on } \mathcal{S}_{52}$ (i.e. probability $1/(52!)$ for each permutation), so that, by the Basic Limit Theorem, we may conclude that $\|\pi_n - \pi\| \rightarrow 0$ as $n \rightarrow \infty$.
- [2.21] Why do we require the “strong” in strong stationary times? That is, in Definition (2.36), although I’m not so inclined to question the requirement $X_T \sim \pi$, why do we require X_T to be independent of T ? It is easy to see where this is used in the proof of the fundamental inequality $\|\pi_n - \pi\| \leq \mathbb{P}\{T > n\}$, but that is only a partial answer. The real question is whether the fundamental inequality could fail to hold if we do not require X_T to be independent of T . Can you find an example?

Things to do

- Add a section introducing optimal stopping, dynamic programming, and Markov decision problems.

3. Markov Random Fields and Hidden Markov Models

Section 1. MRF's on graphs and HMM's.
Section 2. Bayesian framework
Section 3. The Hammersley-Clifford Theorem and Gibbs Distributions
Section 4. Phase transitions in the Ising model.
Section 5. Likelihood and data analysis in hidden Markov chains.
Section 6. Simulating MRF's, the Gibbs' sampler.

In this chapter we will look at some aspects of Markov random fields (MRF's), hidden Markov models (HMM's), and their applications. These models have been successfully used in an impressive variety of applications. For example, state of the art speech recognition systems are based on hidden Markov models. Other examples of application areas include image processing, evolutionary tree reconstruction, DNA sequence alignment, cryptography, modeling ion channels of neurons, and spatial statistics.

3.1 MRF's on Graphs and HMM's

A stochastic process is a collection of random variables $\{X_t : t \in \mathcal{T}\}$ indexed by some subset \mathcal{T} of the real line \mathbb{R} . The elements of \mathcal{T} are often interpreted as times, in which case X_t represents the state at time t of the random process under consideration. The term *random field* refers to a generalization of the notion of a stochastic process: a random field $\{X_s : s \in \mathcal{G}\}$ is still a collection of random variables, but now the index set \mathcal{G} need not be a subset of \mathbb{R} . For example, \mathcal{G} could be a subset of the plane \mathbb{R}^2 ; such random fields are naturally of interest in certain image processing problems, in which an observed image might be modeled as an unknown “true” image plus some random “noise”. In this chapter, we will be considering \mathcal{G} to be the set of nodes of a graph. (This set of nodes will be finite, or countably infinite at most.) Important aspects of the dependence among the random variables will be determined by the edges of the graph through a generalization of the Markov property.

(3.1) NOTATION. *Given a graph \mathcal{G} , we say two nodes s and t are neighbors, denoted $s \sim t$, if s and t are joined by an edge of the graph. We do not consider a node to be a neighbor of itself. For a node $t \in \mathcal{G}$, let $\mathcal{N}(t)$ denote the set of neighbors of t , that is, $\mathcal{N}(t) = \{s \in \mathcal{G} : s \sim t\}$.*

(3.2) DEFINITION. Suppose we are given a graph \mathcal{G} having set of nodes $\{1, \dots, n\}$ and a given neighborhood structure $\mathcal{N}(t) = \{\text{neighbors of node } t\}$. The collection of random variables (X_1, \dots, X_n) is a **Markov random field** on \mathcal{G} if

$$(3.3) \quad \mathbb{P}\{X_t = x_t \mid X_s = x_s \text{ for } s \neq t\} = \mathbb{P}\{X_t = x_t \mid X_s = x_s \text{ for } s \in \mathcal{N}(t)\}$$

for all nodes $t \in \{1, \dots, n\}$.

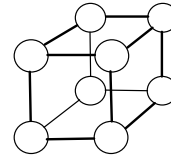
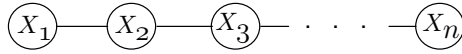
Some compact notation will be convenient, even though, strictly speaking, much of it is sloppy and an abuse of notation. For a subset of nodes $A \subset \mathcal{G}$, let x_A denote the vector $(x_s : s \in A)$. We will lazily write $p(x_t \mid x_{\mathcal{N}(t)})$ in place of the more long-winded expression $\mathbb{P}\{X_t = x_t \mid X_s = x_s \text{ for } s \in \mathcal{N}(t)\}$. Also, we will write $(x_s : s \neq t)$ as $x_{\neq t}$. Thus, the Markov property (3.3) may be written as

$$(3.4) \quad p(x_t \mid x_{\neq t}) = p(x_t \mid x_{\mathcal{N}(t)}).$$

At times we may also use loose language like “the graph $\mathcal{G} = \{1, \dots, n\}$ ” and “for all nodes $s \in \mathcal{G}$,” even though the graph \mathcal{G} is more than just its set of nodes.

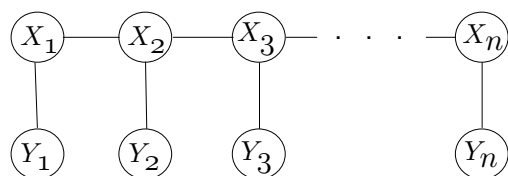
We allow different random variables to take values in different state spaces; say the set of possible values for X_t is $\mathcal{S}_t = \{0, 1, \dots, m_t\}$. The state space for the whole random field (X_1, \dots, X_n) is the product $\mathcal{S} = \mathcal{S}_1 \times \dots \times \mathcal{S}_n$.

Warning: in the Markov business, one sees two different kinds of graphs in common use, and it is easy to get confused.

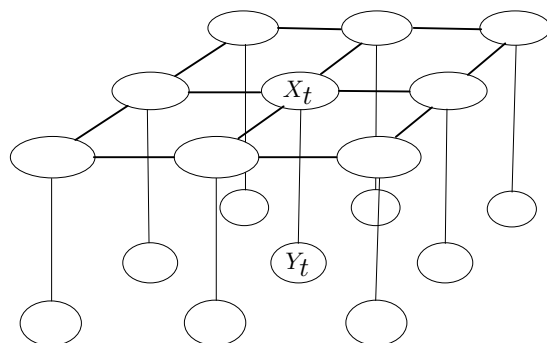


The issue is that some graphs depict the index set (or “time” set) of the Markov random field, while others depict the state space of the Markov random field. The graph \mathcal{G} referred to in the definition of MRF above is of the first type; an example of such a graph might be the graph on the left above. The graphs that we had in mind earlier when we were discussing Markov chains (e.g. random walks on graphs and simulated annealing) were of the second type: the state space of a Markov chain. For example, for optimizing a function defined on a cube, the graph might look like the graph on the right above, since our Markov chain is then searching among the vertices of the cube. However, the corresponding graph of the index set would still look like the graph on the left. This has nothing to do with the particular Markov chain we have in mind; the “time set graph” of *every* MC has this appearance. This graph tells us, for example, that $P(x_2 \mid x_{\neq 2}) = P(x_2 \mid x_1, x_3)$.

A subtopic of Markov random fields is the study of *hidden Markov models* (HMM’s). A HMM is a Markov random field in which some of the random variables are observable and others are not (that is, they are “hidden”). For example, in speech recognition HMM’s of interest might resemble

(3.5) FIGURE. *The hidden Markov chain model.*

The graph above models noisy observation of a Markov chain; the “hidden” Markov chain $\{X_t\}$ runs along the top, and the observed process is $\{Y_t\}$. We will adopt the convention of using the letter X for hidden random variables and Y for observed random variables. Similarly, the graph below represents noisy observation of a two-dimensional Markov random field.



In image processing, the hidden Markov random field $\{X_t\}$ along the top might represent the true, unknown image, and we observe the random field $\{Y_t\}$, which has been corrupted by noise. A simple model of noise in a black and white image, say, might be that a demon goes through the image, pixel by pixel, and flips pixel t (from black to white or white to black) independently with some probability p . In this case,

$$\mathbb{P}\{Y_t = y_t \mid X = x, Y_{\neq t} = y_{\neq t}\} = \mathbb{P}\{Y_t = y_t \mid X_t = x_t\} = p \quad \text{if } y_t \neq x_t.$$

We could also model *blurring* by having each Y_t depend not only on the X_t directly above it, but rather on several nearby X 's. Such a Markov random field model would have more connections from the top layer to the bottom one (and also possibly more within the top layer).

3.2 Bayesian Framework

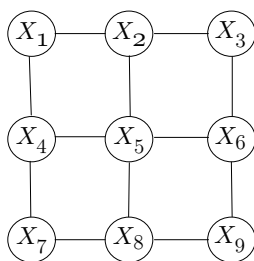
What do we get out of these models? How can we use them? One approach is Bayesian: HMM's fit nicely in the Bayesian framework. Here X is the object of interest; it is unknown. For example, in modeling a noisy image, X could be the true image. We consider the unknown X to be random, and we assume it has a certain *prior distribution*. This distribution, our probabilistic model for X , is assumed to be a Markov random field. We

also postulate a certain probabilistic model for Y conditional on X . This conditional distribution of Y given X reflects our ideas about the noise or blurring or whatever transforms the hidden true image X into the image Y that we observe. Given our assumed prior distribution of X and conditional distribution of $(Y | X)$, Bayes' formula gives the posterior distribution of $(X | Y)$. Thus, given an observed value $Y = y$, in principle at least we get a *posterior distribution* $P\{X = \cdot | Y = y\}$ over all possible true images, so that (again in principle at least) we could make a variety of reasonable choices of our estimator of X . For example, we could choose the x that maximizes $P\{X = x | Y = y\}$. This is called “MAP estimation,” where MAP stands for “maximum *a posteriori*.”

3.3 The Hammersley-Clifford Theorem and Gibbs Distributions

We started our discussion of Markov chains in Chapter 1 by saying how to specify a Markov chain, which requires specifying a state space, initial distribution, and a probability transition structure. How do we specify a Markov random field? A look at (3.4) suggests that a specification be done in terms of conditional distributions of the form $p(x_t | x_{\mathcal{N}(t)})$. Let's think about an example that suggests what goes wrong with this approach.

(3.6) EXAMPLE. An attractive feature of Markov random field models for images is that, despite their simplicity (which may be their most attractive feature), they can capture the idea that images tend to have a degree of “cohesiveness” — pixels located near to each other tend to have the same or similar colors. Suppose we were designing a Markov random field model (perhaps a prior distribution to use in a Bayesian analysis) for images on the three by three lattice shown below.



(3.7) FIGURE. A three-by-three lattice graph as a toy “image” model.

A configuration of just 9 pixels hardly deserves to be called an “image,” but we should start small to get the idea. Here is an example of how we might contemplate an appropriate joint distribution for (X_1, \dots, X_9) . For each pixel, let us specify the conditional distribution of its color given the colors of its neighbors. Suppose there are two colors: 0 and 1. For the

center pixel, a specification like

$$(3.8) \quad \mathbb{P}\{X_5 = 1 | X_2, X_4, X_6, X_8\} = \begin{cases} .1 & \text{if } X_2 + X_4 + X_6 + X_8 = 0 \\ .3 & \text{if } X_2 + X_4 + X_6 + X_8 = 1 \\ .5 & \text{if } X_2 + X_4 + X_6 + X_8 = 2 \\ .7 & \text{if } X_2 + X_4 + X_6 + X_8 = 3 \\ .9 & \text{if } X_2 + X_4 + X_6 + X_8 = 4, \end{cases}$$

for example, might appeal to us. We might feel that such a specification captures our idea that if a pixel has neighbors that are mostly of one color, then that pixel is likely to be of the same color. There are four pixels on the edges that have three neighbors. We might like these pixels to have the conditional distribution

$$(3.9) \quad \mathbb{P}\{X_k = 1 | \text{neighbors}\} = \begin{cases} .2 & \text{if sum of neighbors}=0 \\ .4 & \text{if sum of neighbors}=1 \\ .6 & \text{if sum of neighbors}=2 \\ .8 & \text{if sum of neighbors}=3, \end{cases}$$

where k is 2, 4, 6, or 8. Finally, for the four corner pixels $k = 1, 3, 7$, and 9 that have two neighbors each, we might choose

$$\mathbb{P}\{X_k = 1 | \text{neighbors}\} = \begin{cases} .25 & \text{if sum of neighbors}=0 \\ .5 & \text{if sum of neighbors}=1 \\ .75 & \text{if sum of neighbors}=2. \end{cases}$$

Again, let's just suppose we felt like specifying these distributions, say because they seemed to have about the right amount of "cohesiveness" or whatever to reflect our opinions about the true image.

There's a problem, folks — we have specified this distribution out of existence! These conditional distributions are not consistent with each other; that is, there is no joint distribution of X_1, \dots, X_9 having the given distributions as its conditional distributions. So our careful introspection has literally led to nothing.

We can see the incompatibility of the specified conditional distributions by assuming that they hold and then deriving conclusions that contradict each other. We could start with the three ratios of probabilities

$$\frac{\mathbb{P} \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}}{\mathbb{P} \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}} = \frac{0.1}{0.9}, \quad \frac{\mathbb{P} \begin{bmatrix} 0 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}}{\mathbb{P} \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}} = \frac{0.4}{0.6}, \quad \text{and} \quad \frac{\mathbb{P} \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}}{\mathbb{P} \begin{bmatrix} 0 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}} = \frac{0.7}{0.3},$$

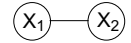
which follow from (3.8), (3.9), and (3.8), respectively. Multiplying the last three equations gives another ratio

$$\frac{\mathbb{P} \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}}{\mathbb{P} \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}} = \frac{(0.1)(0.4)(0.7)}{(0.9)(0.6)(0.3)} = 0.1728,$$

whereas (3.9) implies that the last ratio should be $(0.2)/(0.8) = 0.25$. How sad. \square

In general, we cannot hope to specify a full set of conditional distributions as we tried to do above, where by a “full set of conditional distributions” for (X_1, \dots, X_n) , say, we mean the n conditional distributions of the form $p(x_t \mid x_{\neq t})$ for $t = 1, \dots, n$. A quick way to see that there should be some kind of problem here is to “count equations and unknowns.”

Let’s do this for the simplest nontrivial example of a Markov random field:



That is, we are just looking at the joint distribution of two random variables X_1 and X_2 ; the Markov property doesn’t really say anything here. Again for simplicity, we’ll suppose X_1 and X_2 can take on only the values 0 and 1. In this case, specifying a joint distribution of X_1 and X_2 involves choosing 3 numbers: we have to specify 3 of the 4 values $p(x_1, x_2)$, and the fourth is then determined since the sum must be one. On the other hand, a specification of a full set of conditional distributions involves choosing 4 numbers; for example, we might choose the conditional distribution of X_1 given X_2 to be

$$(3.10) \quad \mathbb{P}\{X_1 = 1 \mid X_2\} = \begin{cases} .3 & \text{if } X_2 = 0 \\ .6 & \text{if } X_2 = 1 \end{cases}$$

and the conditional distribution of X_2 given X_1 to be

$$(3.11) \quad \mathbb{P}\{X_2 = 1 \mid X_1\} = \begin{cases} .2 & \text{if } X_1 = 0 \\ .8 & \text{if } X_1 = 1. \end{cases}$$

That clearly looks like trouble; we have only 3 numbers that we can play with in the joint distribution to try to fulfill 4 specified conditions. Just as we cannot generally solve a system of 4 equations in 3 unknowns, in general we cannot find a joint distribution of (X_1, X_2) that is consistent with the specified conditional distributions.

▷ Exercise [3.1] asks you to prove that the trouble here is real.

Thus, innocently writing down some seemingly reasonable conditional distributions generally produces contradictions. These considerations make conditional distributions look unfriendly and difficult to work with. One might worry that MRF’s also do not look so promising at this point, since the Markov property is defined in terms of the unfriendly conditional distributions. Also, it is embarrassing to be caught analyzing things that do not exist.

Fortunately, the **Hammersley-Clifford Theorem** says that a random field's having the Markov property is equivalent to its having a **Gibbs distribution**, which is a friendly sort of distribution. Thus, instead of worrying about specifying our MRF's in terms of consistent conditional distributions, we can just consider Gibbs distributions, which are simple to write down and work with.

To state the H-C theorem, we need a definition from graph theory.

(3.12) DEFINITION. A set of nodes C is **complete** if all distinct nodes in C are neighbors of each other. That is, C is not complete if it contains two nodes that are not neighbors. A **clique** is a maximal complete set of nodes, that is, C is a clique if C is complete and there is no complete set of nodes D that strictly contains C .

(3.13) DEFINITION. Let \mathcal{G} be a finite graph. A **Gibbs distribution** with respect to \mathcal{G} is a probability mass function that can be expressed in the form

$$(3.14) \quad p(x) = \prod_{C \text{ complete}} V_C(x),$$

where each V_C is a function that depends only on the values $x_C = (x_s : s \in C)$ of x at the nodes in the clique C . That is, the function V_C satisfies $V_C(x) = V_C(y)$ if $x_C = y_C$.

By combining functions V_C for sets C that are subsets of the same clique, we see that we can further reduce the product in the definition of Gibbs distribution to

$$p(x) = \prod_{C \text{ a clique}} V_C(x),$$

(3.15) THEOREM [HAMMERSLEY-CLIFFORD]. Suppose that $X = (X_1, \dots, X_n)$ has positive joint probability mass function. X is a Markov random field on \mathcal{G} if and only if X has a Gibbs distribution with respect to \mathcal{G} .

The history of this result is interesting. Hammersley and Clifford discovered the theorem in 1968, but kept delaying publication because they kept thinking they should be able to remove or relax the unattractive positivity assumption. For some years, through less formal means of communication, the world became familiar with this important, unpublished theorem, and in the meantime, a number of proofs appeared by others. Clifford published a proof in 1990.

PROOF: First a note on the notation we will use. Each random variable X_t takes its values in some finite set \mathcal{S}_t , and the whole Markov random field $X = (X_1, \dots, X_n)$ takes values in the state space $\mathcal{S} = \mathcal{S}_1 \times \dots \times \mathcal{S}_n$. Because the names of the elements of the sets \mathcal{S}_t are irrelevant, and because it will be convenient in the proof below, let $\mathcal{S}_t = \{0, 1, \dots, m_t\}$. With this convention, note in particular that from each set \mathcal{S}_t , we have arbitrarily chosen one element to call "0."

One direction of the proof is easy. Suppose that X has a Gibbs distribution. It is sufficient to show that the ratio

$$\frac{\mathbb{P}\{X_t = x_t \mid X_{\neq t} = x_{\neq t}\}}{\mathbb{P}\{X_t = 0 \mid X_{\neq t} = x_{\neq t}\}} \triangleq \frac{p(x_t \mid x_{\neq t})}{p(0_t \mid x_{\neq t})} = \frac{p(x_t, x_{\neq t})}{p(0_t, x_{\neq t})}$$

depend only on $x_{\mathcal{N}(t)}$. This ratio is

$$\frac{p(x_t, x_{\neq t})}{p(0_t, x_{\neq t})} = \frac{\prod_{t \in C} V_C(x_t, x_{\neq t})}{\prod_{t \in C} V_C(0_t, x_{\neq t})} \left[\frac{\prod_{t \notin C} V_C(x_t, x_{\neq t})}{\prod_{t \notin C} V_C(0_t, x_{\neq t})} \right].$$

But the fraction in square brackets is 1, since changing x_t to 0_t does not change V_C if node t is not in C . So the ratio involves only functions V_C , where node t is in the complete set C , so that every other node in C is a neighbor of t . Thus, such functions V_C depend only on $x_{\mathcal{N}(t)}$.

For the converse, suppose X is a Markov random field. We want to show that we may write the distribution p of X in the form

$$(3.16) \quad p(x) = \prod_A V_A(x),$$

where $V_A \equiv 1$ whenever A is not complete. In fact, we will define the functions V_A in such a way that

$$(3.17) \quad p(x_D, 0_{D^c}) = \prod_{A \subseteq D} V_A(x),$$

holds for all $D \subseteq \{1, \dots, n\}$, with $V_A \equiv 1$ when A is not complete, and (3.16) will follow from (3.17) by taking $D = \{1, \dots, n\}$. We find functions V_D satisfying (3.17) recursively, starting with $D = \emptyset$, then singleton sets D , and so on for larger sets D . For $D = \emptyset$, (3.17) says that $p(0) = V_\emptyset(x)$. [Note V_\emptyset is the constant function taking just the value $p(0)$, which is good, since V_\emptyset is not allowed to depend on any variables x_t !] For singleton $D = \{t\}$, (3.17) says

$$p(x_t, 0_{\neq t}) = V_\emptyset(x) V_{\{t\}}(x) = p(0) V_{\{t\}}(x),$$

so that

$$V_{\{t\}}(x) = \frac{p(x_t, 0_{\neq t})}{p(0)}.$$

This pattern may be continued to express each function V_D in terms of previously determined functions, with the general recursion

$$(3.18) \quad V_D(x) = \frac{p(x_D, 0_{D^c})}{\prod_{A \subset D} V_A(x)}.$$

[A notational reminder: “ $A \subset D$ ” means “ A is a proper subset of D ,” so that A is strictly contained in D . When A is allowed to equal D , we will write $A \subseteq D$.] These definitions guarantee that (3.17) holds for all D .

To finish the proof of the theorem, we will show that if D is not complete, then $V_D(x) = 1$ for all x . This statement will be proved by induction on the number of nodes in D , denoted $\#(D)$. The statement is vacuously true for $\#(D) \leq 1$, since all such sets D are complete. Supposing the desired statement is true for D with $\#(D) \leq k$, we will show that it is also true for $\#(D) = k + 1$. Suppose $\#(D) = k + 1$ and D is not complete, so that D contains two nodes t and u that are not neighbors. Write $D = \{t, u\} \cup B$, where $\#(B) = k - 1$. [For

example, for the case $k = 1$, B is the empty set, and you should make sure for yourself that the following argument works fine in that case.]] By (3.18), our remaining task is to show that

$$(3.19) \quad p(x_D, 0_{D^c}) = \prod_{A \subset D} V_A(x).$$

Start with

$$p(x_D, 0_{D^c}) = p(x_t, x_u, x_B, 0_{D^c}) = \left[\frac{p(x_t, x_u, x_B, 0_{D^c})}{p(0_t, x_u, x_B, 0_{D^c})} \right] p(0_t, x_u, x_B, 0_{D^c}).$$

Since t and u are not neighbors, by the Markov property we obtain

$$\frac{p(x_t, x_u, x_B, 0_{D^c})}{p(0_t, x_u, x_B, 0_{D^c})} = \frac{p(x_t \mid x_u, x_B, 0_{D^c})}{p(0_t \mid x_u, x_B, 0_{D^c})} = \frac{p(x_t \mid 0_u, x_B, 0_{D^c})}{p(0_t \mid 0_u, x_B, 0_{D^c})} = \frac{p(x_t, 0_u, x_B, 0_{D^c})}{p(0_t, 0_u, x_B, 0_{D^c})}.$$

Thus,

$$\begin{aligned} p(x_D, 0_{D^c}) &= \left[\frac{p(x_t, 0_u, x_B, 0_{D^c})}{p(0_t, 0_u, x_B, 0_{D^c})} \right] p(0_t, x_u, x_B, 0_{D^c}) \\ &= \frac{\left(\prod_{A \subseteq B \cup \{t\}} V_A(x) \right) \left(\prod_{A \subseteq B \cup \{u\}} V_A(x) \right)}{\prod_{A \subseteq B} V_A(x)} = \prod_{\substack{A \subset D \\ \{t, u\} \not\subseteq A}} V_A(x). \end{aligned}$$

However, by the induction hypothesis, $V_A \equiv 1$ if $\{t, u\} \subseteq A \subset D$. Therefore,

$$p(x_D, 0_{D^c}) = \prod_{\substack{A \subset D \\ \{t, u\} \not\subseteq A}} V_A(x) = \left(\prod_{\substack{A \subset D \\ \{t, u\} \not\subseteq A}} V_A(x) \right) \left(\prod_{\substack{A \subset D \\ \{t, u\} \subseteq A}} V_A(x) \right) = \prod_{A \subset D} V_A(x),$$

which proves (3.19). □

(3.20) EXAMPLE. A Markov chain X_0, X_1, \dots, X_n has joint distribution of the form

$$p(x_0, x_1, \dots, x_n) = \pi_0(x_0) P_1(x_0, x_1) P_2(x_1, x_2) \cdots P_n(x_{n-1}, x_n).$$

By defining $V_{\{0,1\}}(x_0, x_1) = \pi_0(x_0) P_1(x_0, x_1)$ and $V_{\{k-1,k\}}(x_{k-1}, x_k) = P_k(x_{k-1}, x_k)$ for $k > 0$, we see that this product is a Gibbs distribution on the graph



□

(3.21) EXAMPLE. Remember how the 3-by-3 lattice in Figure (3.7) gave us such trouble in Example (3.6)? There we had difficulty constructing a legitimate Markov random field by specifying conditional distributions at nodes given information about the neighbors. But it

is easy to specify Gibbs distributions. Here the cliques are simply pairs of neighbors that are joined by an edge. We may freely specify a Gibbs distribution on this graph simply by choosing, for each edge $\{i, j\}$ in the graph, four numbers $V_{i,j}(0, 0)$, $V_{i,j}(0, 1)$, $V_{i,j}(1, 0)$, and $V_{i,j}(1, 1)$. \square

▷ Exercises [3.3] and [3.4] ask you to think some more about this small example.

(3.22) EXAMPLE. Consider a hidden Markov model (X, Y) . The Hammersley-Clifford Theorem helps us see some nice properties. For example, if X has a Markov random field prior with a certain neighborhood structure, then the posterior distribution of $(X | Y)$ will also be a Markov random field, with the same neighborhoods as those in the prior.

It is also interesting to ask what kind of process Y is. That is, what is the marginal distribution of Y ? For example, in the “hidden Markov chain” picture (3.5), one might wonder if Y is in fact also a Markov chain. If it is, then instead of bothering with the extra complication of a hidden Markov chain model, why not just model Y directly as a Markov chain? In other words, if the class of phenomena we could model as realizations of the “ Y ” part of a hidden Markov chain model were themselves limited to being Markov chains, that would seem a bit disappointing. As it turns out, Y need not be a Markov chain; in fact, in the hidden Markov chain diagram above, in general the smallest graph on which the Y process can be considered a Markov random field is *fully connected*. To understand this heuristically, we ask ourselves, “If we wanted to guess the value of Y_3 , say, and somebody offered to tell us the values of all of the other Y ’s, would we want to know all of that information, or would we be willing to throw away any of it?” The answer is that in general we should not be willing to give up any of that information. The random variable most relevant to the value of Y_3 is X_3 . If we knew X_3 then we would not care about knowing any other random variables. But we do not know X_3 . So, for example, we would be eager to know X_2 , since it could shed light on X_3 , which in turn is informative for guessing Y_3 . But we do not know X_2 , so we would like to know X_1 . But we cannot know X_1 , so we would be eager to know Y_1 , which in fact we can observe. So, for example, we have argued that the conditional distribution of Y_3 given the values of the other Y random variables, depends on Y_1 , which violates the Markov property.

The general result of which this is a special case is the following. As usual, we assume our probability mass functions are positive, not merely nonnegative.

(3.23) PROPOSITION. Suppose (X, Y) is a Markov random field on the graph \mathcal{G} with the neighborhood structure \mathcal{N} . Write $\mathcal{G} = \mathcal{G}_X \cup \mathcal{G}_Y$, where \mathcal{G}_X and \mathcal{G}_Y are the sets of nodes in \mathcal{G} corresponding to the X and Y random variables, respectively. Then the marginal distribution of Y is a Markov random field on \mathcal{G}_Y , where two nodes $y_1, y_2 \in \mathcal{G}_Y$ are neighbors if either

1. They were neighbors in the original graph; that is, $y_1 \sim y_2$, or
2. There are nodes $x_1, x_2, \dots, x_k \in \mathcal{G}_X$ such that $y_1 \sim x_1 \sim x_2 \sim \dots \sim x_k \sim y_2$.

The conditional distribution of X given Y is a Markov random field on the graph \mathcal{G}_X , where nodes x_1 and x_2 are neighbors if $x_1 \sim x_2$, that is, if x_1 and x_2 were neighbors in the original graph. □

▷ The proof is left to you, as Exercise [3.5].

3.4 Long range dependence in the Ising model

We will work in the integer lattice in d dimensions, denoted \mathbb{Z}^d . So \mathbb{Z}^1 is simply the set of integers, \mathbb{Z}^2 is the set of points in the plane with both coordinates being integers, and so on. For each $t \in \mathbb{Z}^d$ there is a binary random variable X_t taking values in $\{-1, 1\}$, say. The Ising model gives a joint probability distribution for these random variables. We will consider a special case of the Ising model that may be written as follows. For x a configuration of $+1$'s and -1 's at the nodes of a finite subset of \mathbb{Z}^d , let $b(x)$ denote the number of “odd bonds” in x , that is, the number of edges $\{t, u\}$ such that $x_t \neq x_u$. Then, under the Ising model, a configuration x has probability proportional to $\alpha^{b(x)}$, where α is a positive parameter of the distribution. Typically $\alpha < 1$. The choice $\alpha = 1$ corresponds to the uniform distribution, giving equal probability to all configurations. Distributions with small α strongly discourage odd bonds, placing large probability on configurations with few odd bonds.

For the case $d = 1$, the model corresponds to a stationary Markov chain with probability transition matrix

$$P_\alpha = \begin{pmatrix} 1/(1+\alpha) & \alpha/(1+\alpha) \\ \alpha/(1+\alpha) & 1/(1+\alpha) \end{pmatrix}.$$

The basic limit theorem (or an explicit computation) tells us that

$$P_\alpha^n \rightarrow \begin{pmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{pmatrix} \quad \text{as } n \rightarrow \infty.$$

So, for example,

$$\begin{aligned} \lim_{n \rightarrow \infty} \mathbb{P}\{X_0 = +1 \mid X_{-n} = +1\} &= \lim_{n \rightarrow \infty} P_\alpha^n(+1, +1) = 1/2 \\ &= \lim_{n \rightarrow \infty} P_\alpha^n(-1, +1) = \lim_{n \rightarrow \infty} \mathbb{P}\{X_0 = +1 \mid X_{-n} = -1\} \end{aligned}$$

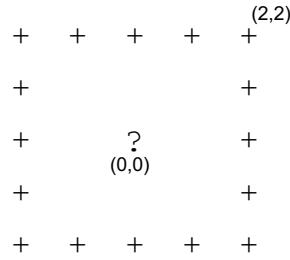
Thus, in the limit, the state of X_0 is unaffected by information about states in the remote past. For a similar statement that is more symmetrical in time, observe that

$$\begin{aligned} \lim_{n \rightarrow \infty} \mathbb{P}\{X_0 = +1 \mid X_{-n} = +1, X_n = +1\} &= \lim_{n \rightarrow \infty} \frac{\mathbb{P}\{X_0 = +1, X_n = +1 \mid X_{-n} = +1\}}{\mathbb{P}\{X_n = +1 \mid X_{-n} = +1\}} \\ &= \lim_{n \rightarrow \infty} \frac{P_\alpha^n(+1, +1)P_\alpha^n(+1, +1)}{P_\alpha^{2n}(+1, +1)} = \frac{1}{2} \\ &= \lim_{n \rightarrow \infty} \frac{P_\alpha^n(-1, +1)P_\alpha^n(+1, -1)}{P_\alpha^{2n}(-1, -1)} = \lim_{n \rightarrow \infty} \mathbb{P}\{X_0 = +1 \mid X_{-n} = -1, X_n = -1\} \end{aligned}$$

Thus, the state X_0 is asymptotically independent of information at nodes far away from 0.

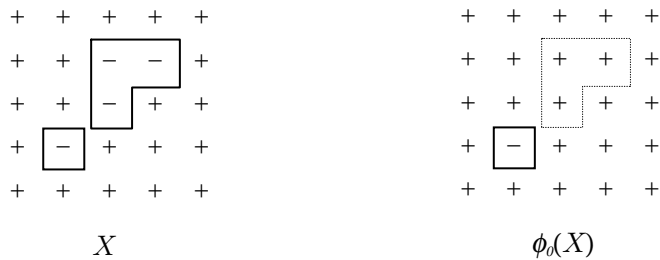
In the remainder of this section we will prove that the situation in $d = 2$ (and higher d) is qualitatively different, in that the effect of states at remote nodes does not disappear in two dimensions. To state the result, let's introduce a bit of notation. Imagine a "cube" K_n of side length $2n$ centered at 0 in \mathbb{Z}^d ; that is, K_n is the set $\{t \in \mathbb{Z}^d : |t_i| \leq n \text{ for all } i = 1, \dots, d\}$ consisting of lattice points whose d coordinates all lie between $-n$ and n . Let B_n denote the "boundary" points of the cube K_n , that is, the points of \mathbb{Z}^2 having at least one coordinate equal to n . Let $P_+^{(n)}\{X = x\}$ denote the Ising probability of $X = x$, conditional on $X_t = +1$ for all $t \in B_n$. Similarly, $P_-^{(n)}(x)$ will denote probabilities conditional on having -1 's on the boundary. Both of these probability measures are functions of α ; for convenience, this dependence is just not indicated in the notation.

The figure below indicates the sort of picture you should have in your mind for the probability $P_+^{(2)}\{X_0 = +1\}$, for example.



(3.24) THEOREM. *For the Ising model on \mathbb{Z}^2 , the effect of the boundary does not disappear. In particular, for example, there exists α such that $P_+^{(n)}\{X_0 = -1\}$ remains below 0.4 for all n , no matter how large.*

PROOF: For any configuration X , imagine drawing little line segments of length 1 cutting through each of the odd bonds in X , as pictured in the left portion of the following figure.



Note that

$$P_+^{(n)}(X) \propto \alpha^{\text{total length of all segments drawn}}.$$

If $X_0 = -1$, that is, the configuration X has a negative spin at the origin, then the line segments we have drawn will form a circuit around the origin; let us denote this circuit by $\gamma_0 = \gamma_0(X)$. So

$$P_+^{(n)}\{X_0 = -1\} = \sum_{\text{circuits } \gamma \text{ about } 0} P_+^{(n)}\{x : \gamma_0(x) = \gamma\}$$

For any given x with $\gamma_0(x) = \gamma$, let $\phi_0(x)$ denote the configuration obtained by flipping the -1 's inside the circuit γ to $+1$'s, as shown in the right part of the previous figure. Let $\ell(\gamma)$ denote the length of γ . Noting that the picture for x has $\ell(\gamma_0(x))$ more segments drawn than the picture for $\phi_0(x)$, we see that

$$P_+^{(n)}(x) = \alpha^{\ell(\gamma_0(x))} P_+^{(n)}(\phi_0(x))$$

for each x with $x_0 = -1$. So

$$\begin{aligned} P_+^{(n)}\{X_0 = -1\} &= \sum_{\gamma} P_+^{(n)}\{x : \gamma_0(x) = \gamma\} \\ &= \sum_{\gamma} \sum_{x: \gamma_0(x) = \gamma} P_+^{(n)}(x) \\ &= \sum_{\gamma} \alpha^{\ell(\gamma)} \sum_{x: \gamma_0(x) = \gamma} P_+^{(n)}(\phi_0(x)) \\ &\leq \sum_{\gamma} \alpha^{\ell(\gamma)}, \end{aligned}$$

where in the last inequality we have used the crude upper bound

$$\sum_{x: \gamma_0(x) = \gamma} P_+^{(n)}(\phi_0(x)) \leq 1,$$

which holds because if x and y are distinct configurations with $\gamma_0(x) = \gamma = \gamma_0(y)$, then $\phi_0(x)$ and $\phi_0(y)$ are also distinct. Let $\nu(\ell)$ denote the number of circuits about 0 of length ℓ . Thus,

$$P_+^{(n)}\{X_0 = -1\} \leq \sum_{\gamma} \alpha^{\ell(\gamma)} = \sum_{\ell=4}^{\infty} \nu(\ell) \alpha^{\ell}.$$

But it is easy to give a crude upper bound $\nu(\ell) \leq \ell 3^{\ell}$. To see this, first observe that each circuit of length ℓ must at least cut through the positive horizontal axis at some point $(k + 1/2, 0)$ for k less than ℓ . Then note that having chosen the first j segments in a circuit, we have at most 3 choices for the $(j + 1)$ st segment, since we cannot backtrack to the previous point. So we have obtained

$$P_+^{(n)}\{X_0 = -1\} \leq \sum_{\ell=4}^{\infty} \ell 3^{\ell} \alpha^{\ell} = \sum_{\ell=4}^{\infty} \ell (3\alpha)^{\ell},$$

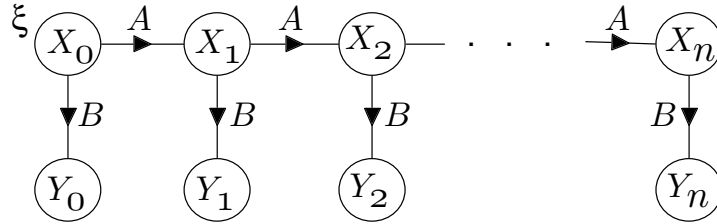
a bound that does not depend on n , and which clearly may be made smaller than 0.4 by choosing α sufficiently small. \square

3.5 Hidden Markov chains

The hidden Markov chain model is the most successfully applied HMM model structure. For example, it is the model that current speech recognition systems use, and the basis of important techniques in bioinformatics.

3.5.1 Description of the model

Like all hidden Markov models, the hidden Markov chain is a MRF in which some of the random variables are observed and others are not — they are hidden. In the graph structure for the hidden Markov chain



the hidden Markov chain is X_0, X_1, \dots, X_n and the observed process is Y_0, Y_1, \dots, Y_n . Edges join X_t to both Y_t and X_{t+1} . The model is parametrized by a marginal distribution ξ of X_0 and, if we assume time-homogeneity of the transition matrices, by two transition matrices A and B , where $A(i, j) = \mathbb{P}\{X_{t+1} = j \mid X_t = i\}$ and $B(i, j) = \mathbb{P}\{Y_t = j \mid X_t = i\}$. Let us write θ for the vector of all parameters: $\theta = (\xi, A, B)$. If there are u states possible for each of the hidden X random variables and v outcomes possible for the observed Y random variables, then ξ is a vector of u probabilities, A is a $u \times u$ probability transition matrix, and B is a $u \times v$ matrix, each of whose rows is a probability mass function.

Compared to the Markov chain model we have studied, the hidden Markov chain is more general — it includes Markov chains as a special case. To get a feeling for this let's look at a few examples. For example, consider the case $u = 1$, that is, there is just one hidden state possible for each X . Then A is simply the 1×1 probability transition matrix $A = (1)$, B is a $1 \times v$ matrix, so that it has just one row, and the Y process is simply an *iid* sequence, where each Y_t has probability mass function given by B . So *iid* sequences are a special case of the hidden Markov chain model. If $u = v$ and B is the identity matrix, then $Y_t = X_t$ for all t , so that we are observing the Markov chain X with no random error, and the process Y is also a Markov chain. So Markov chains are indeed a special case of the hidden Markov chain.

What kind of more general processes can we get from the model? For example, we could model a process where there are two people tossing coins, and we are observing the outcomes of the tosses, a sequence of H 's and T 's. Suppose person 1 has a fair coin having $P(H) = 1/2 = P(T)$, and person 2 has a biased coin, with $P(H) = 0.2 = 1 - P(T)$. Who tosses his coin at time t may be decided in a random manner, with person 1 starting at time 0 and going for a random length of time, then switching to person 2, who tosses for another random length of time, then switching back to person 1, and so on. For example, a referee could roll a pair of dice at each time, and the coin tosser could be switched just at the times when the referee gets a pair of 6's. In this case, we would tend to see a rather long sequence of fair tosses (average length 36 tosses), followed by another long sequence of tosses most of which are T 's, followed by another stretch of fair tosses, and so on. Imagining that we only observe the sequence of H and T outcomes, but we do not know whether person 1 or person 2 is tossing the coin at any given moment, our sequence of observations is a Hidden

Markov chain, with parameters

$$(3.25) \quad \xi = (1, 0), \quad A = \frac{1}{2} \begin{pmatrix} 35/36 & 1/36 \\ 1/36 & 35/36 \end{pmatrix}, \quad \text{and } B = \frac{1}{2} \begin{pmatrix} 0.5 & 0.5 \\ 0.2 & 0.8 \end{pmatrix}$$

You can probably imagine what such a sequence might look like. But just for fun, I'll simulate one and show it to you. Here is the sequence of observations (the Y process) that I got:

THTHTHTTHTHHHTHHHTHTHHHTHTHHHTTTHTTTHTTTTHTTTHTHTTTHT
TTTTTHTHTTTTTTTTTTTHTTHTHTTTHTTTTTTTTTHTHHHTTTTTTHTT
TTHHHHHHHHTTHTTTTTHTHHHTHHHHHTTTHHHTTTHHTHTHHHTTTHTHHHH
TTHHHHTTTTHTTTTHTTHTTTHHTHHHTTHTHTHTHTTTHHHHTTTTTTTHHH
THTHTTTHHHHTHTHTTTTTTTTTTTTTTHHTTTTTHTTTHTTTTTTTTTHTTTHTT
HHHHTTTTHTTTTHTTTHHHHTHHHTTTHTHTTHTTTHHHHTHTTTTTTTTTHTTTT
TTTTTTTTTTTTHTTTTHTTHTTTHHTTTHHTTTTTTTTTHTTTTTTTHHTHTHTHTH
THTTTTTHTTTTTHTTTTTTTTTHTTTTHTHTTTTTTHTHTTTTTTHTTHTHTHT

(3.26) *The observed sequence from a simulated hidden Markov chain.*

The above sequence is Y , the data we get to see. We do not get to see the X sequence of hidden states. In this example, all we know is that $X_0 = 1$ (we know this because $\xi = (1, 0)$); if we know only the Y sequence, we cannot say for sure what any of the other X_t values are. On the other hand, I simulated this sequence myself, and of course I needed to generate the X sequence as part of the simulation. I imagine you are curious what those hidden states are. I wouldn't show this to just anybody, but, for you, I'll make an exception. The sequence below is again the Y sequence, except Y_t is written in lower case if $X_t = 1$, and Y_t is in upper case if $X_t = 2$.

t h t h t h t t h t h h t h h t h t h h t h t h h t h h t t t h T T H T T T T H T T T H T H T T H T
T T T T T h t h T T T T T T T T T T H T T H T H T T H T T T T T T T t t h t h h t t T T T T H T T
T T h h h h h h h t t h t t t t h t h h t h h h h h t t t h h h t t h h t h t h h t t t h t h h h
t t h h h t h t t t h h t t t h h t h t t h h t h h h t t h t h t h t t h h h t t t t t h h h
t h t h t t h h h t h t h T T T T T T T T T T T T H H T T T T T H T T T H T T T T T T T H T t h t t
h h h h t t t t h t t t h h t h h h h t h h t t t h t h t t h t h h h h t h T T T T T T T H T T T T
T T T T T T T T T T H T T T H T T H T T H H T T H H T T T t t t h t t t t t h h t h t h t h h t h
T H T T T T T H T T T T H T T T T T T t h h t t t h T H T T T T t t h t h t t t t t h t t h t h h t

This should all seem somewhat reasonable. The sequence alternates between regions in lower case, in which Heads and Tails are relatively balanced, and regions in upper case, in which there are substantially fewer Heads than Tails.

Having understood the basic structure of the Hidden Markov chain model, you can probably easily imagine a variety of contexts in which the model could be applied and

appreciate how the model has been so useful. In speech recognition, the hidden state might be the phoneme or the word currently being spoken. Each time the same word is spoken, the received signal will be somewhat different, and the model captures this random variation. In modeling the stock market, we might imagine a hidden state having values “bull market” or “bear market.” Investors would be interested in estimating the current hidden state of the market. Biological sequences, such as DNA and proteins, have been modeled with hidden Markov chains, where the hidden states have names like “match,” “insert,” and “delete.” The idea is that if we look at analogous proteins in different species such as human and mouse, for example, the amino acid sequences will be similar but not identical, since in evolving from a common ancestor, the sequences have experienced various substitutions of one amino acid for another, insertions of new material, and deletions.

3.5.2 How to calculate likelihoods

The *likelihood function* $L = L(\theta)$ is the probability of the observed data, as a function of the parameters of the model. The tricky aspect here is that we observe only the Y 's, so that

$$L(\theta) = p_\theta(y_0, y_1, \dots, y_n) = \sum_{x_0} \sum_{x_1} \cdots \sum_{x_n} p_\theta(x_0, x_1, \dots, x_n, y_0, y_1, \dots, y_n) =: \sum_x p_\theta(x, y)$$

This is an intimidating looking sum! For example, if the size of the state space of the hidden variables is just $u = 2$, the sum still has 2^{n+1} terms. That looks like trouble — computational complexity that is exponential in terms of the amount of data. Without a way around this computational issue, the hidden Markov chain model would be of little practical use. Fortunately, by organizing these calculations intelligently in a recursive form, this problem that looks to be of exponential complexity at first glance may actually be done in time that is linear in n . For now, we'll simply derive the recursions and be done with it, but the idea may be viewed from other angles, for example, as an example of dynamic programming, or, alternatively, simply as a conversion of the calculations to matrix multiplications.

Let us denote the state space for the hidden variables X_t by \mathcal{X} ; that is, \mathcal{X} is a finite set of possible hidden states. We are thinking of the observed Y values as fixed here — we know them, and we'll denote them by y_0, y_1, \dots, y_n . For each $t = 0, 1, \dots, n$ and for each $x_t \in \mathcal{X}$, define

$$(3.27) \quad \alpha_t(x_t) = p_\theta(y_0, \dots, y_t, x_t).$$

[[We are using our usual lazy shorthand; written out in full, this would be

$$\alpha_t(x_t) = \mathbb{P}_\theta\{Y_0 = y_0, \dots, Y_t = y_t, X_t = x_t\}.$$

We can calculate the function α_0 right away:

$$(3.28) \quad \alpha_0(x_0) = p_\theta(y_0, x_0) = \xi(x_0)B(x_0, y_0).$$

Also, there is a simple recursion that expresses α_{t+1} in terms of α_t :

$$\begin{aligned}
 \alpha_{t+1}(x_{t+1}) &= p_\theta(y_0, \dots, y_{t+1}, x_{t+1}) \\
 &= \sum_{x_t \in \mathcal{X}} p_\theta(y_0, \dots, y_t, x_t, x_{t+1}, y_{t+1}) \\
 (3.29) \quad &\stackrel{(a)}{=} \sum_{x_t \in \mathcal{X}} p_\theta(y_0, \dots, y_t, x_t) p_\theta(x_{t+1} \mid x_t) p_\theta(y_{t+1} \mid x_{t+1}) \\
 &= \sum_{x_t \in \mathcal{X}} \alpha_t(x_t) A(x_t, x_{t+1}) B(x_{t+1}, y_{t+1}).
 \end{aligned}$$

We have simplified the conditional probabilities in (a) above by using the Markov property of the MRF $(X_0, Y_0, \dots, X_n, Y_n)$ to say that

$$p_\theta(x_{t+1} \mid y_0, \dots, y_t, x_t) = p_\theta(x_{t+1} \mid x_t)$$

and

$$p_\theta(y_{t+1} \mid y_0, \dots, y_t, x_t, x_{t+1}) = p_\theta(y_{t+1} \mid x_{t+1}).$$

Note that the sum in the recursion

$$\alpha_{t+1}(x_{t+1}) = \sum_{x_t \in \mathcal{X}} \alpha_t(x_t) A(x_t, x_{t+1}) B(x_{t+1}, y_{t+1})$$

is very modest; for example, if $\mathcal{X} = \{0, 1\}$, so that our model has just two states possible at the hidden nodes, then this is just the sum of two terms. In this case calculating the function α_{t+1} entails just calculating two numbers $\alpha_{t+1}(0)$ and $\alpha_{t+1}(1)$, each of which is just the sum of two products of three known numbers. That is, using the recursion to calculate the function α_{t+1} from the function α_t involves just a fixed amount of work — the task gets no harder as t increases. Thus, the amount of work to calculate all of the probabilities $\alpha_t(x_t)$ for $t = 0, \dots, n$ and $x_t \in \mathcal{X}$ is linear in n .

Having completed the recursion to calculate the function α_n , the likelihood is simply

$$(3.30) \quad L(\theta) = p_\theta(y_0, \dots, y_n) = \sum_{x_n} p_\theta(y_0, \dots, y_n, x_n) = \sum_{x_n} \alpha_n(x_n).$$

The above probabilities are called “forward” probabilities. In a similar manner, we can calculate the “backward probabilities”

$$(3.31) \quad \beta_t(x_t) = p_\theta(y_{t+1}, \dots, y_n \mid x_t) = \mathbb{P}_\theta\{Y_{t+1} = y_{t+1}, \dots, Y_n = y_n \mid X_t = x_t\}$$

by using the recursion

$$(3.32) \quad \beta_{t-1}(x_{t-1}) = \sum_{x_t} A(x_{t-1}, x_t) B(x_t, y_t) \beta_t(x_t).$$

▷ *Justifying this recursion is left to you; see Exercise [3.9].*

3.5.3 Maximum Likelihood and the EM algorithm

Now that we know how to calculate the likelihood of the observed data at any given $\theta = (\xi, A, B)$, we can, in principle, use this to search for the θ that maximizes the likelihood. We could simply search by trial and error, trying various settings for the parameters and evaluating the likelihood for each. However, without an organized method for changing θ , this search would be a daunting task. Even with just 2 hidden states and 2 output states, there are $2 + 4 + 4 = 10$ parameters in θ , in which there are $1 + 2 + 2 = 5$ degrees of freedom — a 5 dimensional search. We do not want to be blindly fiddling with 5 free parameters.

The EM algorithm is a method for finding maximum likelihood estimates that is applicable to many statistical problems, including hidden Markov chains. The algorithm gives a simple way to increase the likelihood. It is an iterative method; starting from any parameter settings θ_0 , the algorithm gives a sequence of parameters $\theta_1, \theta_2, \dots$ with $L(\theta_0) < L(\theta_1) < L(\theta_2) < \dots$. So we are not blindly guessing new values of θ to try; we are systematically climbing the likelihood function, and the method converges to a local optimum of the likelihood function.

The method is based on knowing how to solve a simple optimization problem, and doing it over and over again. To illustrate that simple problem, here is an example.

(3.33) EXAMPLE. For $q = (q_1, q_2, q_3)$ a probability mass function, so that $q_i \geq 0$ and $q_1 + q_2 + q_3 = 1$, define

$$f(q) = 0.3 \log(q_1) + 0.5 \log(q_2) + 0.2 \log(q_3).$$

Find the vector q that maximizes $f(q)$.

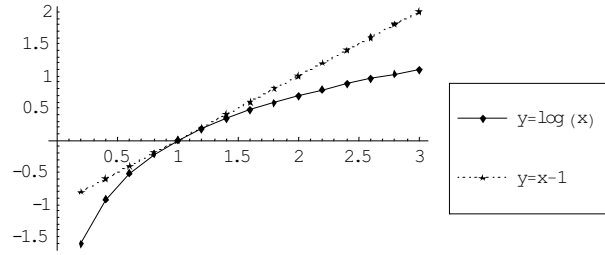
SOLUTION: First let's think for a moment. For example, if we take $q = (1/3, 1/3, 1/3)$, we get $f(q) = \log(1/3)$. Clearly this is not optimal; since $\log(q_2)$ has the largest coefficient, we should be able to increase f by increasing q_2 at the expense of q_1 and q_3 . But we do not want to increase q_2 too much; for example, as $q_2 \rightarrow 1$ (and $q_1 \rightarrow 0$ and $q_3 \rightarrow 0$), we get $f(q) \rightarrow -\infty$. It turns out that the problem has a beautiful answer: to maximize f , take $q = (0.3, 0.5, 0.2)$, the coefficients of the logs in f . A more general formulation and proof are provided by the next proposition. \square

(3.34) PROPOSITION. For $p = (p_1, \dots, p_k)$ and $q = (q_1, \dots, q_k)$ two probability mass functions on $\{1, \dots, k\}$, we have

$$(3.35) \quad \sum_i p_i \log p_i \geq \sum_i p_i \log q_i.$$

PROOF: First note that both sums in (3.35) are unchanged if we restrict the sums to $\{i : p_i > 0\}$, and (3.35) is automatically true if there is some i such that $p_i > 0$ and $q_i = 0$. So we may assume that p and q are both strictly positive, and we want to show that

$\sum_i p_i \log(q_i/p_i) \leq 0$. But note that the log function satisfies the inequality $\log x \leq x - 1$ for all $x > 0$; see the picture and note that the slope of $\log x$ is 1 at $x = 1$.



So

$$\sum_i p_i \log \frac{q_i}{p_i} \leq \sum_i p_i \left(\frac{q_i}{p_i} - 1 \right) = \sum_i (q_i - p_i) = 1 - 1 = 0.$$

□

This proposition is fundamental in the field of Statistics, and also in Information Theory. It leads to a definition of the “Kullback-Leibler distance” $D(p\|q)$ from one probability mass function $p = (p_1, \dots, p_k)$ to another probability mass function $q = (q_1, \dots, q_k)$:

$$D(p\|q) = \sum_i p_i \log \left(\frac{p_i}{q_i} \right).$$

The proposition shows that this “distance” is always nonnegative, and you can show that $D(p\|q) = 0$ if and only if $p = q$.

To see how the optimization problem we have just solved relates to maximum likelihood, consider the following example.

(3.36) EXAMPLE. Let Y_1, \dots, Y_{10} be *iid* with values in $\{1, 2, 3\}$ and probability mass function $\mathbb{P}\{Y_1 = i\} = \theta_i$ for $i = 1, 2, 3$. Suppose, for example, we observe

$$(Y_1, \dots, Y_{10}) = (2, 3, 2, 3, 2, 2, 1, 3, 2, 1).$$

The likelihood function is

$$L(\theta) = \theta_2 \theta_3 \theta_2 \theta_3 \theta_2 \theta_2 \theta_1 \theta_3 \theta_2 \theta_1 = (\theta_1)^2 (\theta_2)^5 (\theta_3)^3,$$

so the log likelihood is

$$\begin{aligned} l(\theta) &= 2 \log \theta_1 + 5 \log \theta_2 + 3 \log \theta_3 \\ &= 10[0.2 \log(\theta_1) + 0.5 \log \theta_2 + 0.3 \log \theta_3]. \end{aligned}$$

Example (3.33) shows that $l(\theta)$ is maximized by the choice $\hat{\theta}_1 = 0.2$, $\hat{\theta}_2 = 0.5$, and $\hat{\theta}_3 = 0.3$; this gives the maximum likelihood estimators of θ . □

Here is a brief description of the EM algorithm in the abstract. We’ll assume for simplicity that X and Y are discrete random variables or vectors, as they are in our particular

application. So probabilities are given by sums rather than integrals. Recall the problem: we want to find the θ maximizing $\log L(\theta)$ where $L(\theta) = p_\theta(y) = \sum_x p_\theta(x, y)$. The EM algorithm repeats the following update that is guaranteed to increase the likelihood at each iteration. Let θ_0 denote the current value of θ . Replace θ_0 by θ_1 , the value of θ that maximizes $E_{\theta_0}[\log p_\theta(X, y) \mid Y = y]$.

Incidentally, “EM” stands for “expectation maximization”: note that at each iteration, we maximize an expected log likelihood.

Why does it work? For a given θ_0 , define $g(\theta) = E_{\theta_0}[\log p_\theta(X, y) \mid Y = y]$. We will see that, in order to have $p_{\theta_1}(y) > p_{\theta_0}(y)$, in fact we do not really need to find θ_1 maximizing g , but rather it is enough to find a θ_1 such that $g(\theta_1) > g(\theta_0)$.

(3.37) PROPOSITION. *If $E_{\theta_0}[\log p_{\theta_1}(X, y) \mid Y = y] > E_{\theta_0}[\log p_{\theta_0}(X, y) \mid Y = y]$, then $p_{\theta_1}(y) > p_{\theta_0}(y)$.*

PROOF: We have

$$\begin{aligned}
 0 &\stackrel{(a)}{<} E_{\theta_0} \left[\log \frac{p_{\theta_1}(X, y)}{p_{\theta_0}(X, y)} \mid Y = y \right] \\
 &= \sum_x p_{\theta_0}(x \mid y) \log \frac{p_{\theta_1}(x, y)}{p_{\theta_0}(x, y)} \\
 &= \sum_x p_{\theta_0}(x \mid y) \log \frac{p_{\theta_1}(y)}{p_{\theta_0}(y)} - \sum_x p_{\theta_0}(x \mid y) \log \frac{p_{\theta_0}(x \mid y)}{p_{\theta_1}(x \mid y)} \\
 &\stackrel{(b)}{=} \log \frac{p_{\theta_1}(y)}{p_{\theta_0}(y)} - \sum_x p_{\theta_0}(x \mid y) \log \frac{p_{\theta_0}(x \mid y)}{p_{\theta_1}(x \mid y)} \\
 &\stackrel{(c)}{\leq} \log \frac{p_{\theta_1}(y)}{p_{\theta_0}(y)},
 \end{aligned}$$

where (a) holds by assumption, (b) holds because $\sum_x p_{\theta_0}(x \mid y) = 1$, and (c) follows from Proposition (3.34). \square

This simple little inequality, proved in 5 lines, is considered to be a major discovery.

3.5.4 Applying the EM algorithm to a hidden Markov chain

Here, as summarized above, we are considering a hidden Markov chain model for $(X, Y) = (X_0, \dots, X_n, Y_0, \dots, Y_n)$, where the random variables Y_t are observed and the random variables X_t are hidden. The model is parametrized by $\theta = (\xi, A, B)$, where ξ is the distribution of X_0 , A is the probability transition matrix of the hidden X chain, and B is the matrix that describes the transitions from X_t to Y_t . To describe one iteration of the EM method, we will imagine that our current guess for θ is $\theta_0 = (\xi_0, A_0, B_0)$, and we want a new guess $\theta_1 = (\xi_1, A_1, B_1)$ that has higher likelihood, that is, such that $p_{\theta_1}(y) > p_{\theta_0}(y)$.

Consider $E_{\theta_0}[\log p_\theta(X, y) \mid Y = y]$ where $\theta = (\xi, A, B)$. We know that

$$p_\theta(x, y) = \xi(x_0) \prod_{t=0}^{n-1} A(x_t, x_{t+1}) \prod_{t=0}^n B(x_t, y_t),$$

so

$$\log p_\theta(X, y) = \log \xi(X_0) + \sum_{t=0}^{n-1} \log A(X_t, X_{t+1}) + \sum_{t=0}^n \log B(X_t, y_t),$$

and

$$\begin{aligned} E_{\theta_0}[\log p_\theta(X, y) \mid Y = y] &= \sum_i \mathbb{P}_{\theta_0}\{X_0 = i \mid y\} \log \xi(i) && \leftarrow (\text{term 1}) \\ &+ \sum_{t=0}^{n-1} \sum_{i,j} \mathbb{P}_{\theta_0}\{X_t = i, X_{t+1} = j \mid y\} \log A(i, j) && \leftarrow (\text{term 2}) \\ &+ \sum_{t=0}^n \sum_i \mathbb{P}_{\theta_0}\{X_t = i \mid y\} \log B(i, y_t). && \leftarrow (\text{term 3}) \end{aligned}$$

We want to maximize the sum of these 3 terms over the variables (ξ, A, B) . However, of these variables, term 1 involves only ξ , term 2 involves only A , and term 3 involves only B . So the sum is maximized by maximizing the 3 terms separately. By Proposition (3.34), we maximize term 1 by the choice

$$(3.38) \quad \xi_1(i) = \mathbb{P}_{\theta_0}\{X_0 = i \mid y\}.$$

Similarly,

$$\text{term 2} = \sum_i \left[\sum_j \left(\sum_{t=0}^{n-1} \mathbb{P}_{\theta_0}\{X_t = i, X_{t+1} = j \mid y\} \right) \log A(i, j) \right],$$

and the i th summand (in large square brackets) involves only the i th row $A(i, \cdot)$, so that we may maximize these summands separately, and again by Proposition (3.34) we are led to choose

$$\begin{aligned} (3.39) \quad A_1(i, j) &= \frac{\sum_{t=0}^{n-1} \mathbb{P}_{\theta_0}\{X_t = i, X_{t+1} = j \mid y\}}{\sum_{j \in \mathcal{X}} \sum_{t=0}^{n-1} \mathbb{P}_{\theta_0}\{X_t = i, X_{t+1} = j \mid y\}} \\ &= \frac{\sum_{t=0}^{n-1} \mathbb{P}_{\theta_0}\{X_t = i, X_{t+1} = j \mid y\}}{\sum_{t=0}^{n-1} \mathbb{P}_{\theta_0}\{X_t = i \mid y\}} \end{aligned}$$

Finally,

$$\begin{aligned} \text{term 3} &= \sum_i \sum_{t=0}^n \mathbb{P}_{\theta_0}\{X_t = i \mid y\} \log B(i, y_t) \\ &= \sum_i \sum_j \left[\sum_{\{t: y_t=j\}} \mathbb{P}_{\theta_0}\{X_t = i \mid y\} \right] \log B(i, j) \end{aligned}$$

is maximized by

$$(3.40) \quad B_1(i, j) = \frac{\sum_{\{t: y_t=j\}} \mathbb{P}_{\theta_0}\{X_t = i \mid y\}}{\sum_j \sum_{\{t: y_t=j\}} \mathbb{P}_{\theta_0}\{X_t = i \mid y\}} = \frac{\sum_{\{t: y_t=j\}} \mathbb{P}_{\theta_0}\{X_t = i \mid y\}}{\sum_{t=0}^n \mathbb{P}_{\theta_0}\{X_t = i \mid y\}}.$$

The solutions for ξ_1 , A_1 , and B_1 given in (3.38), (3.39), and (3.40) have natural interpretations. For example, $\xi_1(i)$ is the probability, under the current guess θ_0 for the parameters, that $X_0 = i$, conditional on the observed data y . This seems natural as a guess for $\xi(i)$, the probability that $X_0 = i$.

To interpret $A_1(i, j)$, first think: we are trying to estimate the probability of $X_{t+1} = j$ given $X_t = i$. If we knew the states X_0, \dots, X_n [which we do not—they are hidden] a natural estimator for $A(i, j)$ would be a simple fraction. The numerator would be the number of $i \rightarrow j$ transitions among the X_t 's, that is, the number of occurrences of the event $\{X_t = i, X_{t+1} = j\}$ among times $t = 0, 1, \dots, n-1$. The denominator would be the number of visits of the X chain to the state i , that is, the number of occurrences of $\{X_t = i\}$ among $t = 0, 1, \dots, n-1$. That is, *if we knew the X_t 's*, we would like to use the estimator

$$\hat{A}(i, j) = \frac{\sum_{t=0}^{n-1} I\{X_t = i, X_{t+1} = j\}}{\sum_{t=0}^{n-1} I\{X_t = i\}}$$

But, as noted, we do not get to observe the X_t 's. The EM updated guess $A_1(i, j)$ does the next best thing, replacing the numerator and denominator of the desired but unattainable estimator $\hat{A}(i, j)$ by their expected values, conditional on what we know—the observed data y .

An analogous interpretation applies to estimating $B(i, j)$, the probability that $Y_t = j$ given $X_t = i$. If we knew the X_t 's, a natural estimator for $B(i, j)$ would be

$$\hat{B}(i, j) = \frac{\sum_{t=0}^n I\{X_t = i, Y_t = j\}}{\sum_{t=0}^n I\{X_t = i\}},$$

and the formula (3.40) is obtained from this by taking $\mathbb{E}_{\theta_0}(\cdots \mid Y = y)$ of the numerator and denominator.

OK, enough interpreting. How does one calculate these quantities? Notice that it is enough to be able to calculate probabilities of the form

$$\gamma_t(i, j) := \mathbb{P}_{\theta_0}\{X_t = i, X_{t+1} = j \mid y\},$$

since probabilities of precisely this form appear in the numerator of $A_1(i, j)$, and simpler probabilities appear in the expressions for ξ_1 and B_1 . [For example, the probabilities in B_1 are of the form

$$\mathbb{P}_{\theta_0}\{X_t = i \mid y\} = \sum_j \gamma_t(i, j).]$$

And of course, to calculate $\gamma_t(x_t, x_{t+1}) = p_{\theta_0}(x_t, x_{t+1} \mid y)$ [sliding into our lazy concise notation here] it is enough to be able to calculate probabilities of the form $p_{\theta_0}(x_t, x_{t+1}, y)$. But probabilities of this last form may be calculated simply in terms of the “forward” probabilities α_t [see (3.27)] and “backward” probabilities β_t [see (3.31)] discussed above, as follows. Imagine that we have calculated all of the α_t 's and β_t 's at $\theta = \theta_0$. Letting y_r^s denote $(y_r, y_{r+1}, \dots, y_s)$,

$$\begin{aligned} p_{\theta_0}(x_t, x_{t+1}, y) &= p_{\theta_0}(y_0^t, x_t, x_{t+1}, y_{t+1}^n) \\ &= p_{\theta_0}(y_0^t, x_t) p_{\theta_0}(x_{t+1} \mid x_t) p_{\theta_0}(y_{t+1}^n \mid x_{t+1}) \\ &= \alpha_t(x_t) A_0(x_t, x_{t+1}) p_{\theta_0}(y_{t+1}^n \mid x_{t+1}). \end{aligned}$$

But

$$\begin{aligned}
 p_{\theta_0}(y_{t+1}^n \mid x_{t+1}) &= p_{\theta_0}(y_{t+1}, y_{t+2}^n \mid x_{t+1}) \\
 &= p_{\theta_0}(y_{t+1} \mid x_{t+1}) p_{\theta_0}(y_{t+2}^n \mid y_{t+1}, x_{t+1}) \\
 &\stackrel{(a)}{=} p_{\theta_0}(y_{t+1} \mid x_{t+1}) p_{\theta_0}(y_{t+2}^n \mid x_{t+1}) \\
 &= B_0(x_{t+1}, y_{t+1}) \beta_{t+1}(x_{t+1}),
 \end{aligned}$$

again using the Markov property at (a) [Exercise: justify this in detail]. Thus,

$$(3.41) \quad p_{\theta_0}(x_t, x_{t+1}, y) = \alpha_t(x_t) A_0(x_t, x_{t+1}) B_0(x_{t+1}, y_{t+1}) \beta_{t+1}(x_{t+1}).$$

Normalizing this by using the definition of conditional probability, we get the required ingredients to perform the EM iteration:

$$(3.42) \quad \gamma_t(i, j) = \frac{\alpha_t(i) A_0(i, j) B_0(j, y_{t+1}) \beta_{t+1}(j)}{\sum_{k, l \in \mathcal{X}} \alpha_t(k) A_0(k, l) B_0(l, y_{t+1}) \beta_{t+1}(l)}.$$

In summary, and in case you'd like to be told how to do this in a structured way, here is the EM recipe for ascending the likelihood function in our hidden Markov chain model:

- Start with some choice of parameter values $\theta_0 = (\xi_0, A_0, B_0)$.
- Calculate forward and backward probabilities $\alpha_t(i)$ and $\beta_t(i)$ for $t = 0, 1, \dots, n$ and $i \in \mathcal{X}$ using recursions (3.29) and (3.32) with initial conditions (3.28) and (??) [all with (ξ, A, B) taken to be (ξ_0, A_0, B_0)]. If the number of hidden states in \mathcal{X} is u , you can store α and β as two $u \times (n+1)$ arrays.
- Calculate the quantities $\gamma_t(i, j)$ for $t \in \{0, \dots, n-1\}$, $i \in \mathcal{X}$ and $j \in \mathcal{X}$ using formula (3.42). These could all be stored in a $u \times u \times n$ array.
- Define

$$\begin{aligned}
 \xi_1(i) &= \sum_j \gamma_0(i, j) \\
 A_1(i, j) &= \frac{\sum_{t=0}^{n-1} \gamma_t(i, j)}{\sum_l \sum_{t=0}^{n-1} \gamma_t(i, l)} \\
 B_1(i, j) &= \frac{\sum_{t=0}^{n-1} \sum_l \gamma_t(i, l) I\{y_t = j\} + \sum_m \gamma_{n-1}(m, i) I\{y_n = j\}}{\sum_{t=0}^{n-1} \sum_l \gamma_t(i, l) + \sum_m \gamma_{n-1}(m, i)}.
 \end{aligned}$$

- Replace θ_0 by $\theta_1 = (\xi_1, A_1, B_1)$ and repeat, taking the new θ_0 to be the θ_1 we just calculated.

We could calculate and print out the likelihood at each stage [see (3.30)] to make sure that it is increasing, as it is supposed to be, and to decide when to stop the iterations, presumably when the increase becomes sufficiently small. Since we may have converged to

a local but not global maximum of the likelihood function, we might want to try the whole thing from several different starting locations.

(3.43) EXAMPLE. Back to our sequence from (3.26). Let's pretend that somebody gave us this sequence and told us only that it was generated from a hidden Markov chain model, without giving any of the details. That is, just try to forget that I already told you the information and parameter values in (3.25). So, for example, we don't know that the state space of the hidden X chain has $v = 2$ states. Let's consider the possibility that v might be 1, 2, or 3, and see whether our methods could suggest to us that in fact $v = 2$. And we want to give our best guess for the unknown parameters; in particular, the $u \times u$ probability transition matrix A and the $u \times v$ emission matrix B .

It is a modest programming task to write *Mathematica* programs to implement the recursions that calculate the likelihood and run the EM algorithm. For the simplest case $v = 1$, the estimates came out to be $B = (.37, .63)$, that is, we estimate that the Y 's are *iid* from the distribution $(.37, .63)$. The log likelihood of the data Y under these optimized parameters comes out to be $\ell_1 = -263.582$. Moving on to consider $v = 2$, the case of two hidden states, gave an optimized log likelihood of $\ell_2 = -258.137$, with parameter estimates

$$(3.44) \quad \xi = (1, 0), \quad A = \begin{array}{c} 1 \quad 2 \\ 1 \left(\begin{array}{cc} .976 & .024 \\ .018 & .982 \end{array} \right), \quad \text{and } B = \begin{array}{c} H \quad T \\ 1 \left(\begin{array}{cc} .515 & .485 \\ .243 & .757 \end{array} \right)$$

For $v = 3$, the algorithm gave a maximized log likelihood of $\ell_3 = -257.427$, with parameter estimates

$$\xi = (0, 0, 1), \quad A = \begin{array}{c} 1 \quad 2 \quad 3 \\ 1 \left(\begin{array}{ccc} .345 & .645 & .010 \\ .630 & .345 & .025 \\ .015 & .008 & .977 \end{array} \right), \quad \text{and } B = \begin{array}{c} H \quad T \\ 1 \left(\begin{array}{cc} .473 & .527 \\ .003 & .997 \\ .515 & .485 \end{array} \right)$$

How do we give our best guess for the number of hidden states, v ? This is an interesting question of statistical inference, and we can just give a brief indication here. It is a logical necessity that, if we did our optimization correctly, we must have $\ell_3 \geq \ell_2 \geq \ell_1$ — more complicated models with more parameters to fit will give higher likelihood. Our decision should be based on *how much* improvement there is in the log likelihood as we increase from $v = 1$ to $v = 2$ and then to $v = 3$. The increase in going from $v = 1$ to $v = 2$ is $\ell_2 - \ell_1 = 5.445$. How impressed we are with an improvement of this size depends on how many extra free parameters are added as we go from $v = 1$ to $v = 2$, that is, the increase in “degrees of freedom.” The model with $v = 1$ has just 1 degree of freedom, while the model with $v = 2$ has 5, so the difference is 4 degrees of freedom. So, in deciding between $v = 1$ and $v = 2$, we have added 4 degrees of freedom to the model and increased the log likelihood by 5.445. Standard statistical practice would suggest looking at the quantile of $2 \times 5.445 = 10.89$ in a chi-square distribution with 4 degrees of freedom, and this suggests believing that the model with $v = 2$ is preferable. Similarly, in considering $v = 2$ versus $v = 3$, we have a log likelihood increase of only $\ell_3 - \ell_2 = 0.71$. Since the number of degrees of freedom in the model with $v = 3$ is 11, we have increased the degrees of freedom by

$11 - 5 = 6$. So the model with $v = 3$ does not look better than the model with $v = 2$. On the basis of this, let's guess the model (3.44) with $v = 2$. Sneaking a peek back at the answer (3.25), we see that we've done rather well! \square

▷ *I had already let the cat out of the bag for this example, since I already revealed the answer before we started. Exercise [3.10] gives you a new bag, with a cat still inside.*

3.6 Simulating a Markov random field: the Gibbs Sampler

Terms like “Markov chain Monte Carlo” and “Markov sampling” refer to methods for generating random samples from given distributions by running Markov chains. Although such methods have quite a long history, they have become the subject of renewed interest in the last decade, particularly with the introduction of the “Gibbs sampler” by Geman and Geman (1984), who used the method in a Bayesian approach to image reconstruction. The Gibbs sampler itself has enjoyed a recent surge of intense interest within statistics community, spurred by Gelfand and Smith (1990), who applied the Gibbs sampler to a wide variety of inference problems.

Recall that a distribution π being “stationary” for a Markov chain X_0, X_1, \dots means that, if $X_0 \sim \pi$, then $X_n \sim \pi$ for all n . The basic phenomenon underlying all Markov sampling methods is the convergence in distribution of a Markov chain to its stationary distribution: If a Markov chain X_0, X_1, \dots has stationary distribution π , then under the conditions of the Basic Limit Theorem, the distribution of X_n for large n is close to π . Thus, in order to generate an observation from a desired distribution π , we find a Markov chain X_0, X_1, \dots that has π as its stationary distribution. The Basic Limit Theorem then suggests that running or simulating the chain until a large time n will produce a random variable X_n whose distribution is close to the desired π . By taking n large enough, in principle we obtain a value that may for practical purposes be considered a random draw from the distribution π .

The Gibbs sampler is a way of constructing a Markov chain having a desired stationary distribution. A simple setting that illustrates the idea involves a probability mass function π of the form $\pi(x, y)$. Suppose we want to generate a random vector $(X, Y) \sim \pi$. Denote the conditional probability distributions by $\pi(\cdot \mid X = \cdot)$ and $\pi(\cdot \mid Y = \cdot)$. To perform a Gibbs sampler, start with any initial point (X_0, Y_0) . Then generate X_1 from the conditional distribution $\pi(\cdot \mid Y = Y_0)$, and generate Y_1 from the conditional distribution $\pi(\cdot \mid X = X_1)$. Continue on in this way, generating X_2 from the conditional distribution $\pi(\cdot \mid Y = Y_1)$ and Y_2 from the conditional distribution $\pi(\cdot \mid X = X_2)$, and so on. Then the distribution π is stationary for the Markov chain $\{(X_n, Y_n) : n = 0, 1, \dots\}$. To see this, suppose $(X_0, Y_0) \sim \pi$. In particular, Y_0 is distributed according to the Y -marginal of π , so that, since X_1 is drawn from the conditional distribution of X given $Y = Y_0$, we have $(X_1, Y_0) \sim \pi$. Now we use the same reasoning again: X_1 is distributed according to the X -marginal of π , so that $(X_1, Y_1) \sim \pi$. Thus, the Gibbs sampler Markov chain $\{(X_n, Y_n) : n \geq 0\}$ has the property that if $(X_0, Y_0) \sim \pi$ then $(X_1, Y_1) \sim \pi$ —that is, the distribution π is stationary.

Simulating a Markov chain is technically and conceptually simple. We just generate the random variables in the chain, in order, and we are done. However, the index set of a Markov random field has no natural ordering in general. This is what causes iterative methods such as the Gibbs sampler to be necessary.

To use the Gibbs sampler to generate a Markov random field, we can start with an arbitrary starting state. For instance, in the example pictured below, the starting state was just *iid* Bernoulli(1/2) random variables—pure noise. One iteration of the Gibbs sampler then consists of visiting each of the $16 \times 16 = 256$ sites, and making a draw from the conditional distribution of that site given the current values for all of the rest of the sites. The Gibbs sampler is well suited to Markov random fields, since it works by repeatedly sampling from the conditional distribution at one node given the values at the remaining nodes, and the Markov property is precisely the statement that these conditional distributions are simple, depending only on the neighbors of the node.

3.7 Exercises

- [3.1] Show that the conditional distributions in (3.10) and (3.11) are indeed inconsistent with each other. Also argue that, in general, if we have specified one conditional distribution $\mathcal{L}(X_1 | X_2)$, say, then we are free to choose only one more distribution of the form $\mathcal{L}(X_2 | X_1 = a)$, where a is just a single, fixed possible value of X_1 .
- [3.2] For $D \subseteq \{1, \dots, n\}$ define V_D as in (3.18). Show that if $x_t = 0$ for some $t \in D$, then $V_D(x) = 1$.
- [3.3] In Example (3.21), suppose we define $V_{\{i,j\}}(0,0) = 1 = V_{\{i,j\}}(1,1)$ and $V_{\{i,j\}}(0,1) = 0.5 = V_{\{i,j\}}(1,0)$ for all neighboring nodes i and j . Calculate the following conditional probabilities:
- (a) $\mathbb{P}\{X_5 = 1 | X_2 = X_4 = X_6 = X_8 = 0\}$.
 - (b) $\mathbb{P}\{X_9 = 1 | X_6 = X_8 = 0\}$.

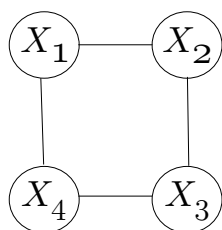
▷ *The next exercise asks you to count dimensions, and it requires some thought... If you cannot find the precise dimension, then just do your best to give whatever upper and lower bounds you can find. This is supposed to be a fun puzzle, mainly! Also, such dimension counting is needed for various practical statistical reasons, such as doing a likelihood ratio hypothesis test.*

- [3.4] Consider the set of all Gibbs distributions on the 3×3 lattice graph pictured in Figure (3.7). What is the dimension of this set? That is, how many degrees of freedom are there in this set of distributions?

[[Hint: Recall we observed that we could specify a Gibbs distribution by freely choosing 4

numbers for each edge. Do all such choices lead to different Gibbs distributions? The result of Exercise [3.2] may also be useful.]

- [3.5] Use the Hammersley-Clifford Theorem to verify Proposition (3.23).
- [3.6] The simplicity of the Hammersley-Clifford Theorem makes nontrivial use of the assumption that the joint probability mass function is strictly positive. To see how things can go wrong without this assumption, consider a joint distribution on a vector of four binary random variables $X = (X_1, X_2, X_3, X_4)$. Of the $2^4 = 16$ possible values for X , suppose the 8 values $(0, 0, 0, 0)$, $(1, 0, 0, 0)$, $(1, 1, 0, 0)$, $(1, 1, 1, 0)$, $(0, 0, 0, 1)$, $(0, 0, 1, 1)$, $(0, 1, 1, 1)$ and $(1, 1, 1, 1)$ each have probability $1/8$, with the remaining 8 patterns having probability 0 each. Show that the joint distribution of X satisfies the Markov property on the graph



but the probability mass function does not factorize into Gibbs form.

- [3.7] Imagine trying to use the method in Theorem (3.24) to prove the false statement that the effect of the boundary does not disappear in \mathbb{Z}^1 . Where does the above method of proof fail?
- [3.8] [Ising model on binary tree] Imagine an infinite bifurcating tree, as pictured below. On the top of the tree there is a root of degree 2. All other nodes have degree 3. Consider an Ising model on the tree: the probability of a configuration x of $+$'s and $-$'s on a finite subtree is proportional to $\alpha^{\nu(x)}$, where α is a parameter of the distribution (a number between 0 and 1, say), and $\nu(x)$ is the number of “odd bonds” in x . Let's use the notation L_n for the 2^n nodes on “level n ,” and let r denote the root of the tree. We'll write “ $X_{L_n} = +$ ” to mean “ $X_s = +$ for all $s \in L_n$,” and “ $X_{L_n} = -$ ” to mean “ $X_s = -$ for all $s \in L_n$.” Define

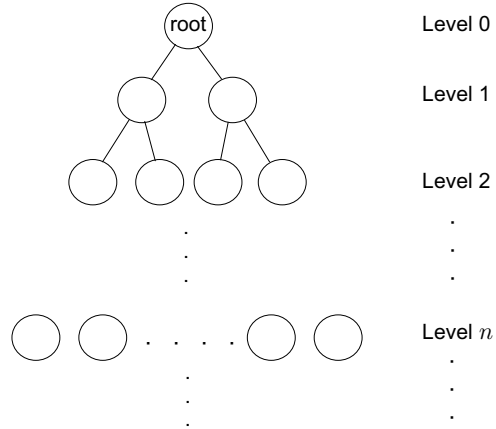
$$u_n = \mathbb{P}\{X_{L_n} = + \mid X_r = +\}$$

and

$$v_n = \mathbb{P}\{X_{L_n} = - \mid X_r = +\} = \mathbb{P}\{X_{L_n} = + \mid X_r = -\}.$$

- Write a recursion for u_n and v_n . That is, write u_{n+1} as a function of u_n and v_n , and write v_{n+1} as a function of u_n and v_n . (These functions will also involve α , of course.)
- Defining $\rho_n = v_n/u_n$, write a recursion for ρ_n ; that is, write ρ_{n+1} as a function of ρ_n (and α).

- (c) We say that the model has a *phase transition* if, as $n \rightarrow \infty$, the limit of $\mathbb{P}\{X_r = + \mid X_{L_n} = +\}$ is different from the limit of $\mathbb{P}\{X_r = + \mid X_{L_n} = -\}$. Show that there is a phase transition when ρ_n does not approach 1 as $n \rightarrow \infty$.
- (d) Observe that $\rho_0 = 0$. Use your recursion (and a computer, if you want!) to investigate the limiting behavior of ρ_n for several different values of α . Determine for which values of α there is a phase transition, and for which values of α there is no phase transition.



- [3.9] Derive the recursion (3.32) for the “backward” probabilities. Show that it is appropriate to start the calculations by setting

$$\beta_n(x_n) = 1 \quad \text{for all } x_n \in \mathcal{X}.$$

[3.10] [[A computing project]] Here are three sequences, each of length 500.

Sequence 1:

```
13311313111112111313121133332313323233213232332223331221222313133222113
233333222112323311332322333121322213233313123131113223112133113311332331
233323233111323313333333111233331311133133113131213112321111222221222212
221123112213211121112212222231213222312222313311133112223221121131133113
32313321333221321323311231322333332132222333113113311112111131131231333
33223212132132232213112112331111311131122122222212212232112123321112321
12132211312112122312112312122332223223112233122232312331322221331113
```

Sequence 2:

```
231311333111321131312132313231333333312331222331233232133313323131122222
132311211312112113133231322121111233311131123311113331332331232331211211
313113323111132211111331121132131111332332112331112113121111123211113131
21111233113213313333113321331331333221323213312313311311113331112322222
133111313122111212113332213113131331133331231231323122211221133111133321
113123131232133112321311322122111123312233313323321323321323123113112321
3222122333113113111332231111322133111311321113333321111123133331132
```

Sequence 3:

```
113331311333323333231331332123313312231333331233323333313333233333132323
122331231333332333322122113111211211132211312111211211113221121321111332
332331332331233313332332232321323313333323311211112211313332211111123331
313332332312222112111111321132121131313111111323233233333311333333213
33331333323333113311231111111111321333311333111113131321111313211111311
11222111112211333322213333112312311333332123332133133323313133333333333
321313333311331113111111211233321111111111113212113311111133111211
```

Your mission is to guess how I generated them. I'll give you some hints. Each sequence is the observed sequence "Y" simulated from a hidden Markov chain model. The corresponding hidden "X" chains have state spaces of different sizes: one sequence has just one hidden state, another has 2 states, and the remaining sequence has 3 hidden states. That is, one of the sequences is *iid* from some distribution on $\mathcal{Y} = \{1, 2, 3\}$, another sequence alternates between two distributions on \mathcal{Y} , and the remaining sequence switches among three different distributions on \mathcal{Y} . Which sequence is which? For each sequence, what is your best guess for the transition matrix A of the hidden chain and the symbol emission matrix B that I used to simulate the sequence?

[[You don't need to type these sequences; they are on the web at
<http://www.stat.yale.edu/~jtc5/251/hmmData.txt>]]

▷ To prepare your mind for the next three exercises, first consider the question of how to simulate a binary Markov chain X_1, \dots, X_n such that $\mathbb{P}\{X_{t+1} \neq X_t\} = 0.2$ for $t = 1, \dots, n-1$. That's easy: simply generate the random variables sequentially, flipping between the values 0 and 1 with probability 0.2. Next, let's change the question by requiring not only $\mathbb{P}\{X_{t+1} \neq X_t\} = 0.2$ for $t = 1, \dots, n-1$, but also $\mathbb{P}\{X_n \neq X_1\} = 0.2$. Now the simple sequential generation of a Markov chain we just described will not work; for example, if n is large, then clearly $\mathbb{P}\{X_n \neq X_1\}$ will be close to 0.5, not to 0.1. So how do we do it?

[3.11] Consider an Ising model on the “cycle” graph with n nodes $\{1, 2, \dots, n\}$ and edges $\{1, 2\}$, $\{2, 3\}$, ..., $\{n-1, n\}$, and $\{n, 1\}$. The Ising model with parameter α gives probability $p(x) \propto \alpha^{b(x)}$ to configuration x , where $b(x)$ is the number of “odd bonds” in x . Show how to use the Gibbs sampler to simulate such a model. For $n = 50$ and $\alpha = 0.9$, generate 5 independent realizations (approximately, using the Gibbs sampler).

[3.12] For an Ising model with parameter α on the cycle on $\{1, 2, \dots, n\}$, let δ denote the probability that any pair of neighboring random variables differ, that is, $\delta = \mathbb{P}\{X_1 \neq X_2\}$. Express δ in terms of α and n . For $n = 10$, use the relationship you just found to calculate the α corresponding to $\delta = 0.2$. Having found the appropriate α to use, simulate 5 realizations of the corresponding Ising model.

[Answer: Find the relationship by expressing the expected number of odd bonds, $n\delta$, as a sum (over even k) of k times the probability of having k odd bonds, and then dividing by n to obtain

$$\delta = \frac{\sum_{k \text{ even}, k \leq n} k \binom{n}{k} \alpha^k}{n \sum_{k \text{ even}, k \leq n} \binom{n}{k} \alpha^k} = \alpha \frac{(1 + \alpha)^{n-1} - (1 - \alpha)^{n-1}}{(1 + \alpha)^n + (1 - \alpha)^n}$$

[3.13] For the Ising model on a cycle graph as discussed in the previous two problems, prove that the process X_1, \dots, X_n is not a Markov chain in general.

4. Martingales

Section 1. Where did the name come from?
Section 2. Definition and examples.
Section 3. Optional sampling.
Section 4. Stochastic integrals and option pricing in discrete time.
Section 5. Martingale convergence.
Section 6. Stochastic approximation.

Imagine spending the day in a mythical “fair” casino, placing bets on various games as the day progresses, and watching your total wealth rise and fall randomly as you win and lose various amounts of money. The casino is “fair” in the sense that, whenever you play a game there, the expected change in your total wealth is always 0, no matter what the history of the process has been. A martingale is a stochastic process that models the time evolution of your total wealth. The theory of martingales and their relatives—submartingales and supermartingales—is one of the pillars of modern probability theory, and a basic tool in applications.

4.1 Why “martingale”?

There are a number of stories about this interesting name, some of which seem somewhat implausible. Here are two definitions the dictionary gives:

1. A device for steadying a horse’s head or checking its upward movement that consists essentially of a strap...
2. Any of several systems of betting in which a player increases his stake usually by doubling each time he loses a bet.

Some say that the reason for the name is actually the first definition above; it turns out that roughly speaking, martingales cannot oscillate around too much [cf., for example, the “upcrossing inequality”], so it is almost as if they are being restrained somehow. In fact, there is a type of stochastic process called the “harness,” which gives support to the equestrian interpretation. As for the second definition, the term “martingale” does have a long history in gambling circles, referring to a seductive sort of strategy in which the gambler continues to double his bet as long as he keeps losing, until he ends up on the winning side. A pitfall, of course, is that eventually the gambler will experience huge losses, using up his initial fortune. He’ll try to say, “But I’m not done with my martingale yet; let me keep playing until I get my dollar!” and he will be told “Sorry; no money, no

more playing.” Apparently the origin of that term is the name of the French community Martigues. Or perhaps the martingale was named in honor of the famous nurse Florence.

4.2 Definitions

Let us adopt some notation for convenience.

(4.1) NOTATION. *Given a process $W = \{W_k\}$, let $W_{m,n}$ denote the portion W_m, W_{m+1}, \dots, W_n of the process from time m up to time n .*

(4.2) DEFINITION. *A process M_0, M_1, \dots is a **martingale** if*

$$E[M_{n+1} \mid M_{0,n}] = M_n \text{ for each } n \geq 0.$$

Sometimes we will also use the following generalization of the last definition.

(4.3) DEFINITION. *A process M_0, M_1, \dots is a **martingale** with respect to another process W_0, W_1, \dots if*

$$E[M_{n+1} \mid W_{0,n}] = M_n \text{ for each } n \geq 0.$$

Definition (4.2) is a special case of Definition (4.3), obtained by taking W to be the same as M . That is, M is a martingale in the sense of Definition (4.2) if M is a martingale *with respect to itself* in the sense of Definition (4.3). Sometimes, however, it will be convenient to allow M and W to be different processes.

The crux of the definition is the condition $E[M_{n+1} \mid W_{0,n}] = M_n$, which is a “fair game” sort of requirement. If we are playing a fair game, then we expect neither to win nor to lose money on the average. Given the history of our fortunes up to time n , our expected fortune M_{n+1} at the future time $n+1$ should just be the fortune M_n that we have at time n .

In addition to the main idea of fairness just described, there is also a more minor technical condition that is implicitly assumed: We also require $E|M_n| < \infty$ for all n so that the conditional expectations in the definition are guaranteed to be well-defined.

▷ To solidify your grasp of the definition try Exercise [4.1].

How about those submartingales and supermartingales? These are processes that are “better than fair” and “worse than fair,” respectively.

(4.4) DEFINITION. *A process X_0, X_1, \dots is a **submartingale** with respect to a process W_0, W_1, \dots if $E[X_{n+1} \mid W_{0,n}] \geq X_n$ for each $n \geq 0$. We say $\{X_n\}$ is a **supermartingale** with respect to $\{W_n\}$ if $E[X_{n+1} \mid W_{0,n}] \leq X_n$ for each $n \geq 0$.*

Let’s discuss these names for a moment to help them stick. Which would you rather bet on: a submartingale or a supermartingale? Doesn’t the term “supermartingale” sound more attractive? But look at the definition: if you like to make money, you would rather bet on a submartingale. So people encountering these terms usually think the names sound backward, and they end up using the following sort of algorithm to decide which to say on

any given occasion: first think to yourself which word you would like to say, then switch and say the other one. This works satisfactorily, but why did these processes get these names anyway? Of course, there are two ways to view the inequalities in the definition. If you want to like the names, you'll remember the inequalities this way:

$$\begin{aligned}\text{submartingale:} \quad & X_n \leq E[X_{n+1} \mid W_{0,n}] \\ \text{supermartingale:} \quad & X_n \geq E[X_{n+1} \mid W_{0,n}].\end{aligned}$$

At each time, a *submartingale* is *below* its future expected value, whereas a *supermartingale* is *above* its future expected value. So the question is: how do you feel now? With a submartingale, you are below what you can expect if you continue playing, and with a supermartingale, you feel that things are best right now, and you should take your money and run.

▷ Exercise [4.2] may shed more light on the mystery of sub versus super.

4.3 Examples

(4.5) EXAMPLE [RANDOM WALKS]. Suppose X_1, X_2, \dots are *iid* and define $S_n = \sum_{k=1}^n X_k$ for $k = 0, 1, \dots$, with $S_0 = 0$. If the random variables X_t have mean 0, then $\{S_n : n \geq 0\}$ is a martingale with respect to itself:

$$E(S_{n+1} \mid S_{0,n}) = E(S_n + X_{n+1} \mid S_{0,n}) = S_n + E(X_{n+1} \mid S_{0,n}) = S_n + E(X_{n+1}) = S_n.$$

Similarly, if X_n has positive mean, $\{S_n\}$ is a submartingale, and if X_n has negative mean, $\{S_n\}$ is a supermartingale. \square

(4.6) EXAMPLE [BRANCHING PROCESSES]. Let X_0, X_1, \dots be a branching process: $X_{n+1} = \sum_{i=1}^{X_n} Z_{ni}$, where the offspring numbers $\{Z_{ni}\}$ are *iid* with mean μ , say.

$$\begin{aligned}E(X_{n+1} \mid X_{0,n} = x_{0,n}) &= E\left(\sum_{i=1}^{X_n} Z_{ni} \mid X_{0,n} = x_{0,n}\right) \\ &= E\left(\sum_{i=1}^{x_n} Z_{ni} \mid X_{0,n} = x_{0,n}\right) \\ &= E\left(\sum_{i=1}^{x_n} Z_{ni}\right) = x_n \mu,\end{aligned}$$

that is, $E(X_{n+1} \mid X_{0,n}) = X_n \mu$. Thus, defining $M_n = \mu^{-n} X_n$, we see that

$$E(M_{n+1} \mid X_{0,n}) = \mu^{-(n+1)}(X_n \mu) = M_n,$$

so that the process $\{M_n\}$ is a martingale. \square

(4.7) EXAMPLE [POLYA'S URN]. Suppose we start out at time 2 with one black ball and one white ball in an urn. Then at each time we draw a ball at random from the urn, and replace it together with a new ball of the same color. Let X_n denote the number of white balls at time n . Thus, given that $X_n = k$, with probability k/n we draw a white ball so that $X_{n+1} = k + 1$, and with the remaining probability $1 - (k/n)$ we draw a black ball so that $X_{n+1} = k$. Letting $M_n = X_n/n$, the fraction of white balls at time n , we have

$$\begin{aligned} E(M_{n+1} | X_{2,n}) &= E\left(\frac{X_{n+1}}{n+1} | X_n\right) = \frac{1}{n+1} \left[(X_n + 1) \frac{X_n}{n} + X_n \left(1 - \frac{X_n}{n}\right) \right] \\ &= X_n/n = M_n. \end{aligned}$$

□

(4.8) EXAMPLE [CONDITIONAL EXPECTATIONS OF A FIXED RANDOM VARIABLE GIVEN INCREASING AMOUNTS OF INFORMATION]. Let X_0, X_1, \dots be any sequence of random variables, and let Y be another random variable. Define $M_n = E(Y | X_{0,n})$. Then $\{M_n\}$ is a martingale with respect to the process $\{X_n\}$.

This has a simple interpretation. Imagine that you are to receive some future reward Y . Also imagine that you are going to observe the random variables X_0, X_1, \dots sequentially — at time n you observe the value of the random variable X_n . You do not know the value of the random variable Y . But we assume that from the beginning you know the joint distribution of the random variables, so that you can calculate expectations and so on. At time n , if you had to guess the value of Y , then your best guess (in the sense of minimizing your expected squared error) would be the conditional expectation of Y given all of the information at your disposal so far, that is, $E(Y | X_0, \dots, X_n)$. The claim is that your sequence of guesses forms a martingale. This makes some intuitive sense. For example, you do not expect tomorrow's guess to be systematically higher than today's; if you did expect this, that would mean that you think today's guess is too low, and it would not be your best guess!

The mathematical verification that $\{M_n\}$ is a martingale is an immediate application of one of the properties (currently (A.8??)) of conditional expectation given in the appendix. Just write it out and you see it:

$$\begin{aligned} E(M_{n+1} | X_{0,n}) &= E(E(Y | X_{0,n+1}) | X_{0,n}) \\ &= E(E(Y | X_{0,n}, X_{n+1}) | X_{0,n}) \\ &\stackrel{(*)}{=} E(Y | X_{0,n}) = M_n. \end{aligned}$$

Property (A.8??) was used for equality (*).

[[DRAW A TREE PICTURE WITH REWARDS (NUMBERS) ASSIGNED TO THE TERMINAL NODES, AND SHOW THE CONDITIONAL EXPECTATIONS AT THE INTERNAL NODES...]] □

(4.9) EXAMPLE [LIKELIHOOD RATIOS]. Suppose random variables X_1, X_2, \dots are independent with probability density function f . Imagine we are considering the alternative

hypothesis that these random variables are independent with a different probability density function g (but they are *really* distributed according to the density f). For simplicity (to eliminate worries related to dividing by 0) suppose that $\{x : f(x) > 0\} = \{x : g(x) > 0\}$. Define the likelihood ratio process $\{M_t\}$ by $M_0 = 1$ and

$$M_t = \frac{g(X_1) \cdots g(X_t)}{f(X_1) \cdots f(X_t)}.$$

Then since $M_{t+1} = M_t g(X_{t+1})/f(X_{t+1})$, we have

$$\begin{aligned} \mathbb{E}(M_{t+1} \mid X_1, \dots, X_t) &\stackrel{(a)}{=} M_t \mathbb{E}\left(g(X_{t+1})/f(X_{t+1}) \mid X_1, \dots, X_t\right) \\ &\stackrel{(b)}{=} M_t \mathbb{E}\left(g(X_{t+1})/f(X_{t+1})\right) \\ &\stackrel{(c)}{=} M_t \int \left(g(x)/f(x)\right) f(x) dx \\ &= M_t \int g(x) dx = M_t, \end{aligned}$$

where (a) holds because M_t is a function of X_1, \dots, X_t , (b) follows from the independence assumption, and (c) uses the assumption that the X_{t+1} has probability density f . \square

4.4 Optional sampling

Probably the most important property of martingales is a “conservation of fairness” property, or “you can’t beat the system” property, technically known as *optional sampling*. Let M be a martingale with respect to W . By the “fair game” property, $E\{M_{n+1}\} = E\{E[M_{n+1} \mid W_{0,n}]\} = E\{M_n\}$ for all n . This implies that

$$EM_n = EM_0 \text{ for all times } n \geq 0.$$

That is, I can say “stop” at any predetermined time t , like $t = 8$, say, and my winnings will be “fair”: $EM_8 = EM_0$.

Fairness is also conserved in many cases—but not in all cases—if I say “stop” at a time that is not a predetermined number, but *random*, that is, depending on the observed sample path of the game. The issue of optional sampling is this:

If T is a *random* time, that is, T is a nonnegative random variable, does the equality $EM_T = EM_0$ still hold?

It would be too much to hope for such a simple result also to hold for all random times $T \geq 0$. There are two sorts of things I should not be allowed to do if we want fairness to be conserved.

1. I should not be allowed to take an indefinitely long time to say “stop.” If I am able just to keep waiting until I see something I like, that seems clearly unfair, doesn’t it?

You would get impatient: “Come on—do something!” Here is an obvious example: consider the simple symmetric random walk $M_n = X_1 + \cdots + X_n$ where X_1, X_2, \dots are *iid* with $P\{X_i = \pm 1\} = 1/2$. If I say “stop” at time $T_1 = \inf\{n : M_n = 1\}$, then clearly $EM_{T_1} = 1 > 0 = EM_0$.

2. Another obvious loophole that violates fairness in games is when a player is allowed “take back” moves, changing his mind about something he did in the past. In our context, that would amount to my using the information available up to time t , say, and going back to some past time $s < t$ and claiming, “I meant to say ‘stop’ then!” That violates the spirit of the game: I am supposed to say “stop” using only the information available up to that time, without peeking ahead into the future. For example, again letting the martingale M be the simple symmetric random walk, consider the random time $T_{\max} \in [0, 3]$ at which M takes on its maximum value $\max\{M_n : 0 \leq n \leq 3\}$ (we could take the last such time if M takes its maximum value more than once). Then clearly $M_{T_{\max}} > 0$ with positive probability; indeed, since $M_1 = 1$ with probability $1/2$, clearly $P\{M_{T_{\max}} \geq 1\} \geq 1/2$. Therefore, $E[M_{T_{\max}}] > 0 = EM_0$. Notice that this sort of failure is indeed conceptually distinct from the previous type. In our example, I am not potentially taking too long to say stop, since T_{\max} is nicely bounded by 3.

Ruling out these two sorts of unfair behavior leaves a class of random times T at which we can be assured that the optional sampling statement $EM_T = EM_0$ holds. Disallowing arbitrarily long times is done by assuming T to be *bounded*, that is, there is a finite number b such that $T \leq b$ holds with probability 1. Random times that disallow the gambler from peeking ahead into the future are called *stopping times*.

(4.10) DEFINITION. A random variable T taking values in the set $\{0, 1, 2, \dots, \infty\}$ is a **stopping time** with respect to the process W_0, W_1, \dots if for each integer k , the indicator random variable $I\{T = k\}$ is a function of $W_{0,k}$.

Just in case you’re wondering: Do we ever really want to let a stopping time take the value “ ∞ ”? Yes, it is more convenient and less abstract than it might appear at first. For example, consider a random walk S_0, S_1, \dots on the integers, starting at $S_0 = 0$, with $S_n = X_1 + \cdots + X_n$ and X_1, X_2, \dots *iid* with $P\{X_i = 1\} = p < 1/2$ and $P\{X_i = -1\} = 1 - p > 1/2$. The random walk $\{S_n\}$ drifts downward (since $p < 1/2$), and it approaches $-\infty$ as $n \rightarrow \infty$. If we were interested in the first time that the random walk hits the value 3, we would be led to consider a definition like $T = \inf\{n : S_n = 3\}$. This random variable is clearly defined for those sample paths that hit the state 3 eventually, but what about the sample paths that never hit the state 3? This is not just idle speculation; there is positive probability that the process never hits the state 3 on its way down toward $-\infty$. In that case, the set $\{n : S_n = 3\}$ is empty, so our definition reduces to the infimum of the empty set, which is ∞ !

▷ You might have a look at Exercises [4.3] and [4.4] at this point.

(4.11) EXAMPLE. As discussed above, let $\{M_n\}$ be the simple, symmetric random walk starting at 0, and consider the random variable $T_1 = \inf\{n : M_n = 1\}$. Show that T_1 is a stopping time.

SOLUTION: Let $n > 0$. Clearly $T_1 = n$ if and only if $M_1 < 1, \dots, M_{n-1} < 1$, and $M_n = 1$, so that $I\{T_1 = n\}$ may be expressed as $I\{M_1 < 1, \dots, M_{n-1} < 1, M_n = 1\}$, which is a function of $M_{0,n}$. \square

(4.12) EXAMPLE. Show that the random variable $T_{\max} = \sup\{n \leq 3 : M_n = \max_{0 \leq k \leq 3} M_k\}$ from above is not a stopping time.

SOLUTION: We want to show that we do not necessarily know by time k whether or not $T = k$. For example, if $1 = X_1 = X_2 = X_3$ then $I\{T = 1\} = 0$ (since in fact $T = 3$), whereas if $X_1 = 1$ and $X_2 = X_3 = -1$ then $I\{T = 1\} = 1$. However, in both cases $M_0 = 0$ and $M_1 = 1$. Thus, the value of the indicator $I\{T = 1\}$ is not determined by $M_{0,1}$. \square

The next theorem is the main optional sampling result.

(4.13) THEOREM. Let M_0, M_1, \dots be a martingale with respect to W_0, W_1, \dots , and let T be a bounded stopping time. Then $EM_T = EM_0$.

PROOF: Suppose that T is bounded by n , that is, $T(\omega) \leq n$ holds for all ω . We can write M_T as M_0 plus the sum of increments of M as

$$\begin{aligned} M_T &= M_0 + \sum_{k=1}^T [M_k - M_{k-1}] \\ &= M_0 + \sum_{k=1}^n [M_k - M_{k-1}] I\{k \leq T\}. \end{aligned}$$

From this, by taking expectations of both sides,

$$(4.14) \quad EM_T = EM_0 + \sum_{k=1}^n E\left([M_k - M_{k-1}] I\{T \geq k\}\right),$$

where we have simply used the fact that the expected value of a finite sum of random variables is the sum of the expected values. [Remember, this interchange of sum and expectation is not necessarily valid for an infinite sum of random variables! So here we have used the assumption that T is bounded.] However, note that since $I\{T \geq k\} = 1 - I\{T \leq k-1\}$ is a function of $W_{0,k-1}$,

$$\begin{aligned} E\left[M_k I\{T \geq k\}\right] &\stackrel{(a)}{=} E\left[E\left(M_k I\{T \geq k\} \mid W_{0,k-1}\right)\right] \\ &\stackrel{(b)}{=} E\left[E\left(M_k \mid W_{0,k-1}\right) I\{T \geq k\}\right] \\ &\stackrel{(c)}{=} E\left[M_{k-1} I\{T \geq k\}\right], \end{aligned}$$

where we have used some of the rules of conditional expectation from Section A.2: (a) uses the iterated expectation rule 3, (b) uses rule 2, and (c) uses the definition of M as a martingale. Thus,

$$E\left([M_k - M_{k-1}]I\{T \geq k\}\right) = 0,$$

so that $EM_T = EM_0$, by (4.14). \square

(4.15) EXAMPLE. The game “Say Red” is played with a shuffled deck of 52 cards, 26 of which are red and 26 of which are black. You start with the cards in a pile, face down. At times $1, 2, \dots, 52$ you turn over a new card and look at it. You must choose one and only one time $\tau \in \{0, 1, \dots, 51\}$ to say “The next card will be red,” and you win the game if the next card is indeed red. Your choice τ is to be a stopping time, so that it may depend in any way on the information you have available up to that time.

Although one might think that it would be possible to play the game strategically and achieve a probability of winning that is greater than $1/2$, we claim that in fact this is impossible: the probability of winning is $1/2$ for all stopping times $\tau \in \{0, 1, \dots, 51\}$. To see this, let R_n denote the number of red cards remaining in the pile after n cards have been turned over. Defining $M_n = R_n/(52-n)$, a simple calculation shows that M_0, M_1, \dots, M_{51} is a martingale. Letting G denote the event that you win, clearly $P(G \mid \tau, R_\tau) = R_\tau/(52-\tau) = M_\tau$, so that $P(G) = EM_\tau$. Thus, by Theorem (4.13), $P(G) = EM_0 = 26/52 = 1/2$. \square

The same ideas are used to prove the next theorem, which generalizes the previous result by

- applying to general supermartingales rather than just martingales, and
- by replacing the two times 0 and T by two stopping times S and T satisfying $S \leq T$.

For a supermartingale X evaluated at nonrandom times $s \leq t$, we have $\mathbb{E}X_t \leq \mathbb{E}X_s$. The following optional sampling results generalizes this to stopping times.

(4.16) THEOREM. *Let X_0, X_1, \dots be a supermartingale with respect to W_0, W_1, \dots , and let S and T be a bounded stopping times with $S \leq T$. Then $\mathbb{E}X_T \leq \mathbb{E}X_S$.*

PROOF: Suppose that T is bounded by n , so that $S(\omega) \leq T(\omega) \leq n$ holds for all ω . Write

$$\begin{aligned} X_T &= X_S + \sum_{k=S+1}^T [M_k - M_{k-1}] \\ &= X_S + \sum_{k=1}^n [X_k - X_{k-1}] I\{k > S\} I\{k \leq T\}. \end{aligned}$$

Taking expectations of both sides, we see that it is sufficient to show that

$$(4.17) \quad E\left([X_k - X_{k-1}]I\{k > S\}I\{k \leq T\}\right) \leq 0.$$

However, note that since

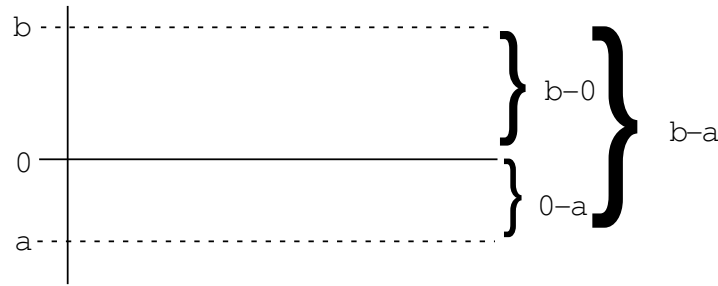
$$I\{k > S\}I\{k \leq T\} = I\{S \leq k-1\}(1 - I\{T \leq k-1\})$$

is a function of $W_{0,k-1}$, applying the rules as above gives

$$\begin{aligned} E\left[X_k I\{k > S\}I\{k \leq T\}\right] &= E\left[E\left(X_k I\{k > S\}I\{k \leq T\} \mid W_{0,k-1}\right)\right] \\ &= E\left[E\left(X_k \mid W_{0,k-1}\right)I\{k > S\}I\{k \leq T\}\right] \\ &\leq E\left[X_{k-1}I\{k > S\}I\{k \leq T\}\right], \end{aligned}$$

which is equivalent to (4.17). □

The optional sampling theorems stated above apply only to bounded stopping times. It would be disappointing if our analysis were really restricted in this way—the boundedness assumption is not satisfied in many natural problems. The next example is a simple case in which the stopping time of interest is not bounded. Its analysis illustrates a standard trick that allows Theorem (4.13) to be applied.



(4.18) EXAMPLE. Let S_0, S_1, \dots be a simple, symmetric random walk on the integers starting from $S_0 = 0$. Let a and b be integers with $a < 0 < b$, and define $T = \inf\{n : S_n = a \text{ or } S_n = b\}$. T is a stopping time, the first time that the random walk hits either a or b . However, T is not bounded; T has positive probability of exceeding any given number. The trick that enables us to apply the Theorem (4.13) is to define a new stopping time

$$(4.19) \quad T_m = \min\{T, m\} = T \wedge m$$

for each m . Then T_m is a bounded stopping time [it is bounded by m , and part (a) of Exercise ([4.5]) shows that it is a stopping time], so that our theorem gives $ES_{T_m} = 0$ for all m . However, $T < \infty$ with probability 1 [remember the random walk is recurrent!]. For each ω such that $T(\omega) < \infty$, we have $T_m(\omega) = T(\omega)$ for sufficiently large m [“sufficiently large” here means that $m \geq T(\omega)$, which depends on ω , but that doesn’t matter]. Therefore, for each ω satisfying $T(\omega) < \infty$, we have $S_{T_m} = S_T$ for sufficiently large m . This clearly implies that $S_{T_m} \rightarrow S_T$ with probability 1. This together with the Bounded Convergence

Theorem [see Appendix A] implies that $ES_{T_m} \rightarrow ES_T$. Thus, since $ES_{T_m} = 0$ for all m , we must have $ES_T = 0$. From here the rest is easy. The random variable S_T takes on two values: a with probability $P\{S_T = a\}$ and b with probability $P\{S_T = b\}$. So we get the equation

$$0 = ES_T = aP\{S_T = a\} + bP\{S_T = b\} = a[1 - P\{S_T = b\}] + bP\{S_T = b\},$$

which gives

$$P\{S_T = b\} = \frac{0 - a}{b - a}.$$

□

(4.20) EXAMPLE. Retain the definitions of the random walk process $\{S_n\}$ and the stopping time T from the previous example. Now we will find the expected value $E(T)$. To do this, we will use another martingale associated with the same random walk: define $M_n = S_n^2 - n$. Writing $S_{n+1} = S_n + X_{n+1}$,

$$\begin{aligned} E\{S_{n+1}^2 \mid S_{0,n}\} &= E\{S_n^2 + 2S_nX_{n+1} + X_{n+1}^2 \mid S_{0,n}\} \\ &= S_n^2 + 2S_nE\{X_{n+1} \mid S_{0,n}\} + E(X_{n+1}^2 \mid S_{0,n}) \\ &= S_n^2 + 2S_nE\{X_{n+1}\} + E(X_{n+1}^2) = S_n^2 + 1, \end{aligned}$$

so that $E\{S_{n+1}^2 - (n+1) \mid S_{0,n}\} = S_n^2 - n$; that is, $\{M_n\}$ is a martingale. Define $T_m = T \wedge m$ as in (4.19). Then since T_m is a bounded stopping time,

$$E\{S_{T_m}^2 - T_m\} = E\{M_{T_m}\} = EM_0 = 0,$$

or $E\{S_{T_m}^2\} = E(T_m)$ for all m . However, $P\{T < \infty\} = 1$, so that $T_m \rightarrow T$ with probability 1 as $m \rightarrow \infty$. By the Bounded Convergence Theorem, $E\{S_{T_m}^2\} \rightarrow E(S_T^2)$ as $m \rightarrow \infty$, and, by Monotone Convergence, $E(T_m) \rightarrow E(T)$. Thus, $E(T) = E(S_T^2)$. However, we just found the distribution of S_T in the previous example: $P\{S_T = a\} = b/(b-a)$ and $P\{S_T = b\} = -a/(b-a)$. Therefore,

$$ES_T^2 = a^2 \frac{b}{b-a} + b^2 \frac{-a}{b-a} = -ab.$$

That is, $E(T) = |a|b$. This is a handy result to remember. For example, the expected time until a random walk wanders 100 units in either direction away from its starting position is $100 \times 100 = 10000$. □

4.5 Stochastic integrals and option pricing in discrete time

Among the most voracious consumers of martingale theory in recent years have been mathematical economists and the “rocket scientist” types on Wall Street. In this section we’ll get a glimpse into what the attraction is.

Everyone has heard of common financial assets such as stocks and bonds. Probability theory enters naturally in the study of the unpredictable price changes of such securities. For fun and profit, people have invented many types of *derivative securities*, including “put” and “call” *options*, for example. A derivative security based on an underlying stock would pay off various amounts at various times depending on the behavior of the price of the stock.

In this section we will discuss the major theory that predicts and explains the prices of such derivative securities. Here we will consider discrete time models; an analogous development in the context of Brownian motion models in continuous time leads to the famous *Black-Scholes formula*, to be presented in Chapter ???.

Denote the stock price at time t by S_t ; the stock price process is S_0, S_1, S_2, \dots . We assume that at an agreed-upon future time n , say, the derivative security pays an amount X that is some function $X = g(S_{0,n})$ of the stock price history up to that time. The derivative security is specified by giving the underlying security, the time n , and the function g .

(4.21) EXAMPLE. A *call option* on a given underlying stock is the right to buy a share of the stock at a certain fixed price c (the “strike price”) at a certain fixed time n in the future (the “maturity date”). If I buy a call option from you, I am paying you some money in return for the right to force you to sell me a share of the stock, if I want it, at the strike price on the maturity date. If $S_n > c$, then the buyer of the option will exercise his right at time n , buying the stock for c and selling it for S_n , gaining a net $S_n - c$. If $S_n \leq c$, then it is not profitable to buy the stock at price c , so the option is not exercised, and the gain at time n is 0. In summary, for the call option, $g(S_{0,n}) = (S_n - c)_+$, where the subscript “+” denotes the “positive part” function. \square

A key concept behind the development in this section is *arbitrage*. An arbitrage is a transaction that makes money without risk, that is, with no chance of losing money. Such free lunches are not supposed to exist, or at least should be rare and short-lived. The basic reason for believing this is that many people are looking for such opportunities to make money. If the price of commodity A were so low, for example, that some clever set of transactions involving buying commodity A and perhaps selling some others were guaranteed to make a riskless profit, then many eager arbitrage seekers would try to perform the transaction many times. The resulting increased demand for commodity A would cause its price to increase, thereby destroying the arbitrage opportunity.

Accordingly, *we will assume that arbitrage opportunities do not exist*. A simple consequence of this assumption is the following principle.

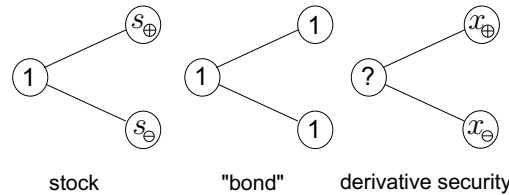
If there are two portfolios that give the same sets of payoffs at the same times (for example, two different combinations of securities that produce the same rewards under all circumstances), then those portfolios must have the same price.

The trick that allows the no-arbitrage assumption to identify the “right” price of a derivative security X based on a stock is this. We will show that the derivative security is actually *redundant*, in the sense that there is a portfolio involving just the stock and a “bond” that produces exactly the same payoffs as X does. Supposing we have done this,

since we are given the prices of the stock and the bond, then we can calculate the price of the reproducing portfolio, which (by the principle above) must be the same as the price of the derivative security.

Let's start simple: a problem involving just one period. We will consider a stock and just two states of nature, which correspond to the stock price rising or falling. Suppose that at time 0 the stock price is 1, and at time 1 the stock price must be either s_{\oplus} or s_{\ominus} , say, where $s_{\ominus} < 1 < s_{\oplus}$. The other security we may use in our portfolio is called a "bond." In fact, for simplicity—the ultimate goal here—let us assume the interest rate is zero. [In case this assumption bothers you, note that simply by redefining the unit of money in different periods, it turns out to be easy to convert a problem with a nonzero interest rate into an equivalent problem with interest rate zero.] In this case, the bond is really a dull investment: investing \$1 at time 0 returns exactly \$1 at time 1. We can think of it this way: "buying b shares of the bond" (where b is a real number) corresponds to lending out $\$b$ for one period: we lose $\$b$ at time 0 but then gain back $\$b$ at time 1. If $b < 0$ this corresponds to borrowing $\$b$ for one period. In other words, assuming an interest rate of zero means that we can lend or borrow money with no cost.

Like the stock, the redundant security X will have payoffs that are a function of the state of nature. So let's say that X pays x_{\oplus} if the stock price goes up to s_{\oplus} and x_{\ominus} if the stock price goes down to s_{\ominus} . The three financial instruments we may work with are shown below: the stock, the boring bond, and the option whose no-arbitrage price we wish to determine.



Finding such a "reproducing portfolio" is easy. We assume the stock and bond can be traded in continuous amounts, so that one can buy 2.718 shares of stock or sell 3.14 bonds, for example. Letting a and b denote the number of stock and bond shares in the portfolio, we want to solve for a and b . Since the payoffs at time 1 from such a portfolio are $as_{\oplus} + b$ if the stock goes up and $as_{\ominus} + b$ if the stock goes down, the requirement that the portfolio reproduce the payoffs of the redundant security consists of the two equations $as_{\oplus} + b = x_{\oplus}$ and $as_{\ominus} + b = x_{\ominus}$, or, in other words, the vector equation $a \begin{pmatrix} s_{\oplus} \\ s_{\ominus} \end{pmatrix} + b \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} x_{\oplus} \\ x_{\ominus} \end{pmatrix}$. We are assured that there is a solution for any given x_{\oplus} and x_{\ominus} , because the vectors $\begin{pmatrix} s_{\oplus} \\ s_{\ominus} \end{pmatrix}$ and $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$ are linearly independent, and therefore span \mathbb{R}^2 . Solving, we obtain

$$a = \frac{x_{\oplus} - x_{\ominus}}{s_{\oplus} - s_{\ominus}}, \quad b = \frac{x_{\ominus}s_{\oplus} - x_{\oplus}s_{\ominus}}{s_{\oplus} - s_{\ominus}};$$

in particular, the price π that we pay for this portfolio at time 0 is

$$\pi = a + b = x_{\oplus} \left(\frac{1 - s_{\ominus}}{s_{\oplus} - s_{\ominus}} \right) + x_{\ominus} \left(\frac{s_{\oplus} - 1}{s_{\oplus} - s_{\ominus}} \right).$$

Finally, the familiar arbitrage reasoning: If the price of the redundant security were anything other than π , we would have two investments (the redundant security and the reproducing portfolio) that have different prices but exactly the same payoffs, which would provide a clear recipe for arbitrage. Thus, the price for the redundant security implied by a no-arbitrage assumption is π .

There is a nice interpretation of π that makes it easy to remember and points toward the connection with martingales. Letting

$$p = \frac{1 - s_{\ominus}}{s_{\oplus} - s_{\ominus}},$$

we have found that

$$(4.22) \quad \pi = x_{\oplus} p + x_{\ominus} (1 - p).$$

Thus, if, for some reason, we assumed that the the stock price had a probability p of rising, then the price π would simply be the expected value of the payoff of the redundant security. Notice that the arbitrage argument goes through to give the same price π no matter what the true probabilities for the states of nature might be,^{*} hence the magic and mystery of option pricing: the “probability” p has nothing to do with any true probabilities. It does, however, have an interesting and useful interpretation: *p is the probability that makes the stock price a martingale.* Indeed, we could solve for p by this specification—if the probability of the stock’s rising is p , then the expected value of the stock price at time 1 is $ps_{\oplus} + (1-p)s_{\ominus}$. For the stock price to be a martingale, this last expression must be the same as the stock price at 0, which is 1. This happens precisely when $p = (1 - s_{\ominus})/(s_{\oplus} - s_{\ominus})$. The equality (4.22) says that *the price of the redundant security is its expected payoff, where the expectation is taken under the “equivalent martingale measure.”*

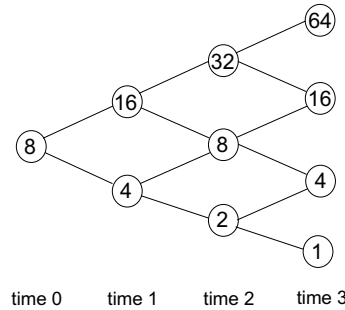
It may seem that bringing in martingales to explain a couple of simple bookkeeping calculations involving only simple arithmetic is underilluminating overkill. But the martingale story will continue to summarize the situation when we deal with Brownian motion and diffusions later on, where the arithmetic bookkeeping becomes tenuous. Even in simple discrete-time problems, martingales provide a slick approach, in multiperiod problems summarizing in one pithy statement what might otherwise seem to be a mass of mindless calculations.

Turning to multiperiod problems, we continue to assume a simple, stylized model for the stock price. Imagine that the stock price process can be described as a bifurcating tree, where for each possible history of the stock price up to time t , there are just two possible values for the price at time $t + 1$. We describe a path, or history, in the stock price tree

^{*}Well, except for something funny that happens if either state has probability 0. However, in fact such a situation is ruled out by the no-arbitrage assumption. For example, if the probability of the stock rising is 1, then the two portfolios “buy $1/s_{\oplus}$ shares of stock” and “buy 1 bond” have the same payoffs with probability 1 at time 1, but have different prices.

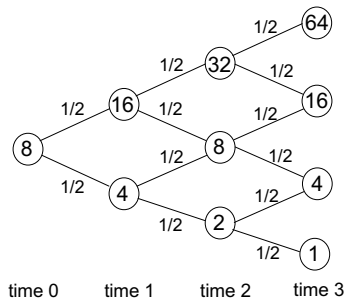
up to time t by a sequence of binary variables W_1, W_2, \dots, W_t , where $W_k = \oplus$ means that the price took the larger of the two possible values at time k , and $W_k = \ominus$ means that the price took the smaller of the two possible values at time k . Let's assume the initial stock price s_0 is known, and W_1, W_2, \dots is a random sequence. So the stock price S_t at time t is determined by the history $W_{1,t} = (W_1, \dots, W_t)$; let's write it as a function $S_t = S_t(W_{1,t})$.

(4.23) EXAMPLE. As a simple illustration, we might assume that in each period the stock price must either double or half. If, for example, the stock starts out at $s_0 = 8$ and it goes up in the first 3 periods, then at time 3 the price is $S_3(\oplus, \oplus, \oplus) = 64$. The price after 3 “down” periods would be $S_3(\ominus, \ominus, \ominus) = 1$. In general, $S_t(W_{1,t}) = s_0 2^{W_1 + \dots + W_t}$. The possible paths of the stock price for the first few periods are as follows:



This model looks very artificial—one would not expect a stock to either double or half each period. But the same sort of model becomes more realistic looking when there are many very short periods and the stock price can gain or lose some very small percentage each period.

We have not yet specified probabilities of various stock price paths. In the simple model of the stock price, the probabilities in question would be the probabilities of the 8 paths $(\ominus, \ominus, \ominus), (\ominus, \ominus, \oplus), (\ominus, \oplus, \ominus), \dots, (\oplus, \oplus, \oplus)$. For example, we might assume that all 8 paths are equally likely, which is equivalent to assuming a probability measure \mathbb{P} under which the random variables W_1, W_2 , and W_3 are independent with $\mathbb{P}\{W_i = \oplus\} = \mathbb{P}\{W_i = \ominus\} = 1/2$, as depicted below.



The probability \mathbb{P} .

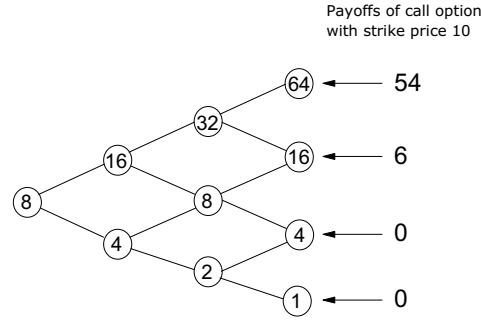
For the sake of discussion, let us suppose from now on that \mathbb{P} gives the true probabilities in this example. But as hinted above and as we will soon see again, a major surprise of the theory is that *the probabilities basically don't matter*. □

Given the price $S_t = S_t(W_1, \dots, W_t)$ at time t , the price S_{t+1} has the two possible values $S_{t+1}(W_1, \dots, W_t, \oplus)$ and $S_{t+1}(W_1, \dots, W_t, \ominus)$. Let us assume that the current stock price

is always strictly between the two possible prices at the next period, that is,

$$(4.24) \quad S_{t+1}(W_1, \dots, W_t, \ominus) < S_t(W_1, \dots, W_t) < S_{t+1}(W_1, \dots, W_t, \oplus).$$

We want to price a derivative security X whose value at time n is a function of $W_{1,n}$. For example, for the stock price process of Example (4.23), a call option with strike price 10 at time 3 would have the payoffs shown below.



Again the key is to show that the derivative security is redundant—its payoffs can be duplicated using the stock and a bond. A *trading strategy* consists of a specification of the number of stock shares that we hold at each time period. Let H_t denote the number of shares held at time t . Think of it this way: at time t we buy H_t shares to hold over the interval from t to $t + 1$. This choice is based on the history of the stock price process up to time t , and it cannot depend on future information. This is a reasonable model of a non-clairvoyant investor. For example, if we allowed H_t to depend on W_{t+1} , the investor could look into the future to see if the stock goes up or down in the next period and hold 1,000,000 shares if $W_{t+1} = \oplus$ and hold $-1,000,000$ shares (that is, sell a million shares) if $W_{t+1} = \ominus$.

The payoff of the strategy H in the first period is $H_0(S_1 - S_0)$, that is, the number of shares bought at time 0 times the amount the stock price rises. Accumulating such gains over the first n periods, we see that the gain at time n from the strategy H_0, H_1, \dots is

$$(4.25) \quad (H \cdot S)_n = H_0(S_1 - S_0) + H_1(S_2 - S_1) + \cdots + H_{n-1}(S_n - S_{n-1}).$$

It might be helpful to think about this in more detail. At time 0, we buy H_0 shares of stock, which costs $H_0 S_0$. To make things very clean, let's imagine that we borrow $H_0 S_0$ at time 0 to finance the stock purchase. So in fact we neither gain nor lose money at time 0, since we gain $H_0 S_0$ by borrowing it and then spend that same amount of money to buy the stock. At time 1, we sell our H_0 shares of stock (remember that H_0 is the number of shares of stock we hold just for the first period), gaining $H_0 S_1$, and we pay off the money we borrowed, namely $H_0 S_0$. So the strategy $H \cdot S$ produces payoffs $H_0(S_1 - S_0)$ at time 1, and the strategy costs nothing to perform. For practice, you should tell yourself a detailed story of how to implement the strategy $(H \cdot S)_3$, producing the payoffs $H_0(S_1 - S_0) + H_1(S_2 - S_1) + H_2(S_3 - S_2)$ at time 3, with no expenditures along the way.

The process $H \cdot S$ is a discrete-time *stochastic integral*. Generally speaking, an integral is a sum of products, or a limit of sums of products. For example, the “Riemann-Stieltjes”

integral $\int_a^b f dg$ can be defined to be a limit of sums of the form $\sum_{i=0}^{k-1} f(x_i)[g(x_{i+1}) - g(x_i)]$ as $k \rightarrow \infty$, where $a = x_0 < x_1 < \dots < x_k = b$ and $\max_{i < k} (x_{i+1} - x_i) \rightarrow 0$ as $k \rightarrow \infty$. [The ordinary Riemann integral $\int_a^b f(x) dx$ is a special case of this where g is simply the identity function $g(x) = x$.] That is, $\int f dg$ is a sum of products of values of f with changes in values of g . Analogously, (4.25) is a sensible way to define an integral $\int H dS$ for discrete-time processes.

We will see that in the type of bifurcating tree models of stock prices that we are considering, the payoffs given by any derivative security X with maturity date n may be reproduced precisely by the sum of a constant and some trading strategy; that is,

$$(4.26) \quad X = x_0 + (H \bullet S)_n$$

for some number x_0 and trading strategy H . This is not so obvious, but for now let us assume it is true and see what we can do with it.

(4.27) CLAIM. *In addition to assuming (4.24), suppose that each path $w_{1,t}$ has positive probability. Suppose X satisfies (4.26). Then the no-arbitrage price of X is x_0 .*

PROOF: The assumptions imply the existence of a martingale measure \mathbb{Q} . [Why?] We want to show that for each price other than x_0 , there is an arbitrage opportunity. For example, suppose X has price $y < x_0$ (a similar argument may be constructed if $y > x_0$). Then buy X for y , perform trading strategy $-H$, and borrow x_0 until period n . Then at time 0 you get $x_0 - y > 0$, whereas no money changes hands in any future periods — in particular, in period n your gains and losses of $x_0 + (H \bullet S)_n$ cancel out.

To show there is no arbitrage at price x_0 , suppose, to the contrary, that there is a trading strategy J such that buying X for x_0 and performing strategy J give a riskless profit, with

$$-x_0 + X + (J \bullet S)_n \begin{cases} \geq 0 & \text{for all } w_{1,n} \\ > 0 & \text{for some } w_{1,n}. \end{cases}$$

But

$$\mathbb{E}_{\mathbb{Q}}(-x_0 + X + (J \bullet S)_n) = 0,$$

which gives a contradiction, as desired. □

The claim reduces the problem of pricing X to that of finding x_0 in the representation (4.26).

A stochastic process, such as the price process for the stock, takes various possible paths with various probabilities. Different probability measures will allocate probability among paths differently. For discrete processes of the type we have been discussing, some paths will have positive probability and some will have zero probability.

(4.28) DEFINITION. *Two probability measures \mathbb{P} and \mathbb{Q} for a process are called **equivalent** if they agree on which sets of paths have zero probability and which sets of paths have positive probability.*

This is a weak notion of equivalence—two probability measures can be very different but still equivalent. So maybe it's not such a good name. But it's firmly entrenched in probability and measure theory, so let's not fight it.

(4.29) EXAMPLE. For the example considered in (4.23) with the probability measure \mathbb{P} pictured above, any equivalent measure \mathbb{Q} will simply reassign probabilities among the set of paths already taken by \mathbb{P} . An equivalent measure \mathbb{Q} will not go off and blaze new trails, but just reallocates probability among the old paths. \square

(4.30) DEFINITION. Given a process and a probability measure \mathbb{P} , a probability measure \mathbb{Q} is an **equivalent martingale measure** if \mathbb{Q} is equivalent to \mathbb{P} and the process is a martingale under the measure \mathbb{Q} .

A martingale measure \mathbb{Q} makes the identification of the desired price x_0 in (4.26) easy. Since the martingale property gives $\mathbb{E}_{\mathbb{Q}}(S_{t+1} - S_t \mid W_{1,t}) = 0$, we have

$$\begin{aligned} \mathbb{E}_{\mathbb{Q}}\left(H_t(S_{t+1} - S_t)\right) &= \mathbb{E}_{\mathbb{Q}}\left[\mathbb{E}_{\mathbb{Q}}\left(H_t(S_{t+1} - S_t) \mid W_{1,t}\right)\right] \\ &= \mathbb{E}_{\mathbb{Q}}\left[H_t \mathbb{E}_{\mathbb{Q}}(S_{t+1} - S_t \mid W_{1,t})\right] = 0, \end{aligned}$$

so that

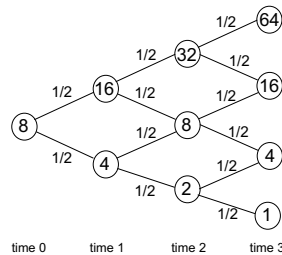
$$\mathbb{E}_{\mathbb{Q}}(H \cdot S)_n = \mathbb{E}_{\mathbb{Q}}[H_0(S_1 - S_0)] + \cdots + \mathbb{E}_{\mathbb{Q}}[H_{n-1}(S_n - S_{n-1})] = 0.$$

Thus, taking the expectation under \mathbb{Q} of both sides of (4.26), the $\mathbb{E}_{\mathbb{Q}}$ knocks out the stochastic integral part of the right-hand side, just leaving the desired price x_0 :

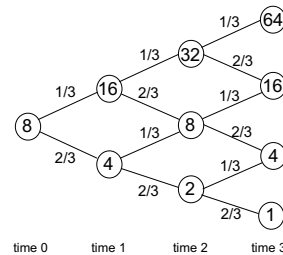
$$(4.31) \quad \mathbb{E}_{\mathbb{Q}}(X) = x_0.$$

In summary: We suppose the stock price is governed by a probability measure \mathbb{P} on the paths in a bifurcating tree. Letting \mathbb{Q} be a probability measure equivalent to \mathbb{P} under which the stock price is a martingale, the no-arbitrage price of the derivative security X is $\mathbb{E}_{\mathbb{Q}}X$, that is, the expectation of X , taken under the measure \mathbb{Q} .

(4.32) EXAMPLE. For the setting of Example (4.23), \mathbb{P} is not a martingale measure. For example, if the current price is 4, under \mathbb{P} the price at the next period will be either 2 or 8 with equal probability, giving an expected value of 5. The martingale measure \mathbb{Q} has $\mathbb{Q}\{W_k = \oplus\} = 1/3$ and $\mathbb{Q}\{W_k = \ominus\} = 2/3$.



Not a martingale under \mathbb{P}



A martingale under \mathbb{Q}

The no-arbitrage price of the call option with strike price 10 is

$$54[(1/3)^3] + 6[3(1/3)^2(2/3)] = 10/3$$

□

Finally, why is it that we can reproduce the payoffs of X by trading the stock? That is, under the assumed conditions, how do we know that there is a trading strategy H such that (4.26) holds? The development rests on a simple martingale representation result.


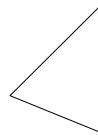
(4.33) THEOREM [MARTINGALE REPRESENTATION]. *Assuming $\{S_t\}$ is a martingale with respect to \mathbb{Q} , for any other martingale $\{M_t\}$ with respect to \mathbb{Q} there is a trading strategy $\{H_t\}$ such that $M_t = M_0 + (H \cdot S)_t$.*


WHY? We want to show that there is a process $\{H_t\}$ such that

$$M_{t+1} - M_t = H_t(S_{t+1} - S_t).$$

The content of this statement is that increments of the M process are scaled-up or scaled-down increments of the S process. For example, if a particular branch of the S

tree looks like , then the corresponding branch of the M tree might look like

 [[scaling factor $H_t = 0.5$]] or  [[scaling factor $H_t = 1.5$]], but not

 [[not a scaling]]. But this is clearly true if M and S are both martingales with respect to the same probability measure!

PROOF: The martingale property assumed of S may be written as

$$\mathbb{Q}\{W_{t+1} = \oplus \mid W_{1,t}\}S_{t+1}(W_{1,t}, \oplus) + \mathbb{Q}\{W_{t+1} = \ominus \mid W_{1,t}\}S_{t+1}(W_{1,t}, \ominus) = S_t(W_{1,t}),$$

or

$$(4.34) \quad \begin{aligned} & \mathbb{Q}\{W_{t+1} = \oplus \mid W_{1,t}\} \left(S_{t+1}(W_{1,t}, \oplus) - S_t(W_{1,t}) \right) \\ & = -\mathbb{Q}\{W_{t+1} = \ominus \mid W_{1,t}\} \left(S_{t+1}(W_{1,t}, \ominus) - S_t(W_{1,t}) \right). \end{aligned}$$

The assumption (4.24) implies that both sides of (4.34) are positive. Since $\{M_t\}$ is also a martingale under \mathbb{Q} , it satisfies the same type of equality as $\{S_t\}$ does:

$$(4.35) \quad \begin{aligned} & \mathbb{Q}\{W_{t+1} = \oplus \mid W_{1,t}\} \left(M_{t+1}(W_{1,t}, \oplus) - M_t(W_{1,t}) \right) \\ & = -\mathbb{Q}\{W_{t+1} = \ominus \mid W_{1,t}\} \left(M_{t+1}(W_{1,t}, \ominus) - M_t(W_{1,t}) \right). \end{aligned}$$

Dividing (4.35) by (4.34) gives

$$(4.36) \quad \frac{M_{t+1}(W_{1,t}, \oplus) - M_t(W_{1,t})}{S_{t+1}(W_{1,t}, \oplus) - S_t(W_{1,t})} = \frac{M_{t+1}(W_{1,t}, \ominus) - M_t(W_{1,t})}{S_{t+1}(W_{1,t}, \ominus) - S_t(W_{1,t})} =: H_t(W_{1,t}),$$

that is, we have defined $H_t(W_{1,t})$ to be either of the two fractions in (4.36). Thus, no matter whether $W_{t+1} = \oplus$ or $W_{t+1} = \ominus$, we have $M_{t+1}(W_{1,t+1}) - M_t(W_{1,t}) = H_t(W_{1,t})(S_{t+1}(W_{1,t+1}) - S_t(W_{1,t}))$. So this definition of H satisfies our desired conditions: H_t depends only on $W_{1,t}$, and $M_{t+1} - M_t = H_t(S_{t+1} - S_t)$. \square

Now we want to apply the theorem to establish the representation (4.26). The trick is to define a martingale $\{M_t\}$ such that $M_n = X$. But this is easy: define

$$M_t = \mathbb{E}_{\mathbb{Q}}(X \mid W_{1,t}).$$

Note that $M_n = X$, since we have assumed that X is a function of $W_{1,n}$, and $M_0 = \mathbb{E}_{\mathbb{Q}}X$. Thus, applying the martingale representation theorem to M , there is a trading strategy H such that

$$X = M_0 + \sum_{t=0}^{n-1} (M_{t+1} - M_t) = \mathbb{E}_{\mathbb{Q}}X + \sum_{t=0}^{n-1} H_t(S_{t+1} - S_t).$$

This is what we wanted to show.

4.6 Martingale convergence

Martingales and their relatives tend to be rather tame and well-behaved. In particular, they often converge. Of course, not *all* martingales converge. For example our old friend, the simple symmetric random walk on the integers, is a martingale, and we know that it does not converge to anything. In fact, no matter where the random walk is at any given time, we can be sure it will visit each integer some time thereafter. But convergence does occur under a number of reasonable conditions. For example, we will show that a nonnegative supermartingale converges with probability 1. This makes a bit of intuitive sense: it is a stochastic analog of the well-known and intuitively obvious statement that a nonnegative, nonincreasing sequence of numbers must converge.

The basic tool in proving these results are optional sampling results as discussed in the previous section. Here is what we will use.

(4.37) PROPOSITION. *Let X_0, X_1, \dots be a nonnegative supermartingale, and let $X_0 \leq a$, where a is a nonnegative number. Then for $b > a$, defining $T_b = \inf\{t : X_t \geq b\}$, we have $P\{T_b < \infty\} \leq a/b$.*

WHY? Suppose we are playing a supermartingale, starting with an initial fortune $X_0 \leq a$. Suppose we adopt the strategy of stopping as soon as our fortune exceeds b , if that ever happens. Our expected reward from playing this strategy is at least $bP\{T_b < \infty\}$, since we stop having won at least b with probability $P\{T_b < \infty\}$. But we expect to lose money on

a supermartingale, so this expected reward should be no larger than our initial fortune, so that $bP\{T_b < \infty\} \leq a$.

PROOF: For typographical convenience, let's drop the subscript b on T_b , so that $T = T_b$. By optional sampling, for each finite t , $E(X_{T \wedge t}) \leq E(X_0) \leq a$. Note that $X_{T \wedge t}$ is at least b if $T \leq t$, and of course $X_{T \wedge t}$ is nonnegative by the assumption that X is a nonnegative supermartingale. We can express those last two statements together in one concise inequality: $X_{T \wedge t} \geq bI\{T \leq t\}$. Taking expected values gives $a \geq EX_{T \wedge t} \geq bP\{T \leq t\}$, so that $P\{T \leq t\} \leq a/b$. Since t was arbitrary we get $P\{T < \infty\} \leq a/b$, as desired. \square

And now for the basic (super)martingale convergence theorem.

(4.38) THEOREM. *A nonnegative supermartingale converges with probability 1.*

WHY? A sequence of nonnegative numbers x_0, x_1, \dots has only 3 possible behaviors: (1) it can converge to a finite number, (2) it can converge to ∞ , or (3) it can “oscillate,” with $\liminf x_t < \limsup x_t$. We want to see that with probability 1, our sample path $X_0(\omega), X_1(\omega), \dots$ will not exhibit either behavior (2) or behavior (3). Convergence to ∞ is ruled out by the supermartingale property. That is, if $X_t \rightarrow \infty$ on a set of positive probability, we would have $E(X_t) \rightarrow \infty$. (Note that since the random variables are nonnegative, we cannot have $X_t \rightarrow -\infty$ on some other set of positive probability to compensate and keep the expectation bounded). But since $E(X_t) \leq E(X_0)$ for all t , it cannot be the case that $E(X_t) \rightarrow \infty$. So (2) is ruled out. For (3), let a and b be arbitrary nonnegative numbers, with $a < b$, say. We want to see that the path cannot oscillate infinitely many times, being below a infinitely many times and also above b infinitely many times. But we know that whenever the process gets below a , the probability that it ever goes above b after that is only a/b at most. So each time the process gets below a , there is a probability at least $1 - a/b$ that it will never again attain the level b . So with probability 1, eventually the process must stop making “upcrossings” from below a to above b , so that with probability 1, the process does not oscillate as in case (3).

PROOF: Let $0 \leq a < b$. Define stopping times

$$\begin{aligned} T_0 &= 0 \\ S_1 &= \inf\{t \geq T_0 : X_t \leq a\} \\ T_1 &= \inf\{t \geq S_1 : X_t \geq b\} \\ &\vdots \\ S_k &= \inf\{t \geq T_{k-1} : X_t \leq a\} \\ T_k &= \inf\{t \geq S_k : X_t \geq b\} \\ &\vdots \end{aligned}$$

Since $S_k \wedge n$ and $T_k \wedge n$ are bounded stopping times with $S_k \wedge n \leq T_k \wedge n$, the Optional Sampling Theorem (4.16) gives $\mathbb{E}(X_{T_k \wedge n}) \leq \mathbb{E}(X_{S_k \wedge n})$, or

$$(4.39) \quad \mathbb{E}(X_{T_k \wedge n} - X_{S_k \wedge n}) \leq 0.$$

Observe that

- if $T_k \leq n$, then $X_{T_k \wedge n} \geq b$, and
- if $T_k > n$, then $T_k \wedge n = n$, so that $X_{T_k \wedge n} = X_n$.

This observation implies the following more concise statement phrased in terms of indicator random variables:

$$X_{T_k \wedge n} \geq bI\{T_k \leq n\} + X_n I\{T_k > n\}.$$

The same kind of reasoning gives

$$X_{S_k \wedge n} \leq aI\{S_k \leq n\} + X_n I\{S_k > n\},$$

so that

$$\begin{aligned} X_{T_k \wedge n} - X_{S_k \wedge n} &\geq bI\{T_k \leq n\} - aI\{S_k \leq n\} + X_n(I\{T_k > n\} - I\{S_k > n\}) \\ &\geq bI\{T_k \leq n\} - aI\{S_k \leq n\}, \end{aligned}$$

where the last inequality uses the nonnegativity of X_n and the fact that $T_k \geq S_k$. Taking expectations and using (4.39) gives

$$0 \geq bP\{T_k \leq n\} - aP\{S_k \leq n\},$$

from which, letting $n \rightarrow \infty$, it follows that

$$P\{T_k < \infty\} \leq \left(\frac{a}{b}\right) P\{S_k < \infty\}.$$

Thus, since $\{S_k < \infty\} \subseteq \{T_{k-1} < \infty\}$ clearly holds for all $k \geq 1$, we obtain

$$(4.40) \quad P\{T_k < \infty\} \leq \left(\frac{a}{b}\right) P\{T_{k-1} < \infty\},$$

and iterating this relationship gives

$$P\{T_k < \infty\} \leq \left(\frac{a}{b}\right)^k P\{T_0 < \infty\} = \left(\frac{a}{b}\right)^k.$$

Defining $K = \sup\{k : T_k < \infty\}$, we have shown that $P\{K \geq k\} \leq (a/b)^k$, and taking limits as $k \rightarrow \infty$, we see that $P\{K = \infty\} = 0$. That is, with probability 1, the process eventually stops crossing the interval (a, b) . As explained above, since this holds for arbitrary $a < b$, with probability 1 the path must converge. \square

4.7 Stochastic approximation

Stochastic approximation is a method of addressing two very fundamental problems: solving equations and optimizing functions. Here let us think about solving equations; evidently the problem of maximizing functions is closely related (for example, in calculus we learn to maximize a function g by solving the equation $g'(x) = 0$). So, here's a problem: solve the equation $f(x) = 0$. Hmm... you say, that shouldn't be so hard; after all, there are lots of techniques for solving equations, right? Oh, sorry — I forgot to mention that you have to solve the equation without knowing what f is! You are limited to gathering some partial, noisy information about f : you will see only observations of the unknown function f that have been corrupted by random noise. So here is the game: you get to choose a value X_0 , say, and ask me the question, “What is $f(X_0)$?” However, in response, I will not give you the true answer, but rather the true answer plus some “noise” random variable η_0 having mean 0. That is, what I tell you is not $f(X_0)$, but rather the sum $Y_0 = f(X_0) + \eta_0$. For example, η_0 might be a $N(0, 1)$ -distributed random variable. Having been given the noisy observation Y_0 , you get to choose another value X_1 , and I will generate a new, independent noise random variable η_1 and tell you the value of the sum Y_1 , defined by $Y_1 = f(X_1) + \eta_1$. And so on: at each stage n you get to choose a value X_n based on the noisy observations you have been given so far.

Thus, the stochastic twist to the problem we will consider here is that we hope to solve the equation $f(x) = 0$ *without knowing what the function f is*, but rather given only randomly perturbed, *noisy observations* of f .

Let us suppose that we know certain qualitative information about f . For example, we assume that f has a unique but unknown zero x^* . Let us also assume that $f(x) < 0$ for $x < x^*$ and $f(x) > 0$ for $x > x^*$; for example, f might be monotone increasing, but need not be so to satisfy this assumption. There will also be some other assumptions to be specified later. Recall we want to find x^* , and we have to do this somehow by choosing a sequence of values X_0, X_1, \dots and asking the sequence of questions “What is $f(X_t)$?” for $t = 0, 1, \dots$. It would be wonderful if we had a method for choosing the values X_t in such a way that we could be sure that $X_t \rightarrow x^*$ as $t \rightarrow \infty$. That such a method should exist at all does not seem immediately apparent.

In this section we will discuss a beautifully simple method proposed by Robbins and Monro. This *stochastic approximation* method has become the basis of a number of interesting algorithms in a variety of applications, such as clustering data and training neural networks. This sort of algorithm is likely to be useful in cases where we expect to observe a large amount of data over time and we want to refine our estimate of some parameter, “learning from experience” in a reasonable way. It captures a familiar sort of idea. In the beginning of the learning process, having observed little data, we adjust our opinions quickly. Later on, having already observed a great deal of data, we become less willing to change our opinions hastily, and each new piece of information receives less weight. Over time, the algorithm gradually responds less to each noisy piece of information. [In fact, the recursive formula for the sample mean is a simple example of this idea:

$$\bar{X}_{n+1} = \bar{X}_n + \frac{1}{n+1}(X_{n+1} - \bar{X}_n).$$

Notice how the sample mean moves from its old value \bar{X}_n toward the new observation X_{n+1} , but only by the fraction $1/(n+1)$.]

Here is our model in general. We observe the random variables

$$(4.41) \quad Y_n = f(X_n) + \eta_n,$$

where the random variables X_1, X_2, \dots are subject to our choice and η_1, η_2, \dots are assumed to be *iid* “noise” random variables with mean 0 and variance 1. The function f is unknown, and we do not know or observe the noise random variables, only the Y ’s.

The Robbins-Monro iteration gives a simple way of choosing X_{n+1} given the previously observed Y values. Suppose for now that we are given a suitable sequence of positive numbers a_0, a_1, \dots ; we will shortly discuss criteria for a suitable choice. Then, given X_0 , we define X_1, X_2, \dots according to the recursion

$$(4.42) \quad X_{n+1} = X_n - a_n Y_n.$$

The iteration (4.42) is qualitatively reasonable, in the sense that it pushes us in a reasonable direction. For example, if Y_n is positive, since $E(Y_n) = f(X_n)$, we would tend to believe that $f(X_n)$ is more likely to be positive than negative. By our assumptions about f , then, we would guess (just on the basis of this bit of information) that $X_n > x^*$. This would cause us to favor making our next guess X_{n+1} less than X_n , which is what the iteration (4.42) does: if $Y_n > 0$ then $X_{n+1} < X_n$. Similarly, if $Y_n < 0$, then (4.42) makes $X_{n+1} > X_n$, which is in accordance with a suspicion that $f(X_n) < 0$, which corresponds to $X_n < x^*$.

How should we choose the sequence a_0, a_1, \dots ? Since we want X_n to converge to something (in particular, x^*) as $n \rightarrow \infty$, we must have $a_n \rightarrow 0$ as $n \rightarrow \infty$. Is that clear? Obviously, no matter what values the X_n ’s take, the sequence of Y_n ’s is a randomly varying sequence that does not approach 0. At best, even if X_n were x^* for all n , so that each $f(X_n) = 0$, we would have $Y_n = \eta_n$, so that $\{Y_n\}$ is an independent sequence of random variables having variance 1. So if a_n does not approach 0, the variability of the $\{Y_n\}$ sequence does not approach 0, and so the increments $a_n Y_n$ in the $\{X_n\}$ sequence do not converge to 0, and the sequence of X_n ’s cannot converge to anything. So, we know at least we will want our sequence a_n to converge to 0. On the other hand, if a_n converges to 0 too rapidly, then clearly we will not achieve our goal of having X_n converge to x^* starting from an arbitrary X_0 . For example, if a_n were in fact 0 for each n , then the $\{X_n\}$ sequence would be constant, with $X_n = X_0$ for all n ; the $\{X_n\}$ sequence gets stuck. If a_n is positive but $a_n \rightarrow 0$ too rapidly, we have similar difficulty.

So, conceptually, we want to choose a sequence $\{a_n\}$ that converges to 0 rapidly enough to damp out the randomness in the noise so that the resulting sequence of X_n ’s will converge, but slowly enough so that the sequence of X_n ’s is capable of moving the potentially arbitrarily large distance from X_0 to x^* .

(4.43) THEOREM. *Let $f : \mathbb{R} \rightarrow \mathbb{R}$ and let $E[(X_0)^2] < \infty$. Consider the sequence $\{X_n\}$ generated by the recursion*

$$\begin{aligned} Y_n &= f(X_n) + \eta_n \\ X_{n+1} &= X_n - a_n Y_n, \end{aligned}$$

where we assume the following conditions:

- (i) The random variables $X_0, \eta_0, \eta_1, \dots$ are independent, with η_0, η_1, \dots iid having mean 0 and variance 1.
- (ii) For some $1 < c < \infty$, we have $|f(x)| \leq c|x|$ for all x . *[[Note in particular that this incorporates the assumption $f(0) = 0$.]]*
- (iii) For all $\delta > 0$, $\inf_{|x| > \delta} (xf(x)) > 0$.
- (iv) Each a_n is a nonnegative number, and $\sum a_n = \infty$.
- (v) $\sum a_n^2 < \infty$.

Then $X_n \rightarrow 0$ with probability 1 as $n \rightarrow \infty$.

Condition (ii) is a growth condition on f . It is easy to imagine that if f grows extremely fast, we could have X_n making wilder and wilder oscillations over time. Also, note that just for notational convenience, we have assumed that $x^* = 0$, without loss of generality. The Robbins-Monro iteration makes no use of the knowledge that $x^* = 0$, and we can legitimately pretend we do not know that $x^* = 0$ and be happy without pretense when we show that $X_n \rightarrow 0$ with probability 1. Condition (iii) implies that $f(x) < 0$ for $x < x^*$ and $f(x) > 0$ for $x > x^*$, but otherwise allows quite a wide range of behavior. Conditions (iv) and (v) are a formulation of the idea, previously discussed, that the sequence $\{a_n\}$ needs to converge to 0 fast enough but not too fast.

PROOF: We start with some conditional expectation calculations.

$$\begin{aligned} E(X_{n+1}^2 \mid X_{0,n}) &= E\left\{(X_n - a_n[f(X_n) + \eta_n])^2 \mid X_{0,n}\right\} \\ &= X_n^2 - 2a_n E\{X_n[f(X_n) + \eta_n] \mid X_{0,n}\} + a_n^2 E\{[f(X_n) + \eta_n]^2 \mid X_{0,n}\}. \end{aligned}$$

However,

$$(4.44) \quad E\{X_n[f(X_n) + \eta_n] \mid X_{0,n}\} = X_n f(X_n) + X_n E\{\eta_n \mid X_{0,n}\} = X_n f(X_n)$$

since $E\{\eta_n \mid X_{0,n}\} = E\{\eta_n\} = 0$ by the independence of η_n and $X_{0,n}$. Also,

$$\begin{aligned} E\{[f(X_n) + \eta_n]^2 \mid X_{0,n}\} &= E\{f(X_n)^2 + 2f(X_n)\eta_n + \eta_n^2 \mid X_{0,n}\} \\ &= f(X_n)^2 + 2f(X_n)E\{\eta_n \mid X_{0,n}\} + E\{\eta_n^2 \mid X_{0,n}\} \\ &= f(X_n)^2 + 1, \end{aligned}$$

since $E\{\eta_n^2 \mid X_{0,n}\} = E\{\eta_n^2\} = 1$. Thus,

$$E(X_{n+1}^2 \mid X_{0,n}) = X_n^2 - 2a_n X_n f(X_n) + a_n^2 (f(X_n)^2 + 1).$$

By assumption, $|f(x)| \leq c|x|$, so that

$$a_n^2 (f(X_n)^2 + 1) \leq a_n^2 (c^2 X_n^2 + 1) \leq a_n^2 c^2 (X_n^2 + 1),$$

which gives

$$(4.45) \quad E(X_{n+1}^2 \mid X_{0,n}) \leq X_n^2(1 + a_n^2 c^2) - 2a_n X_n f(X_n) + a_n^2 c^2.$$

What now? This is a job for supermartingale! We want a supermartingale, but I don't feel like driving all the way to the supermartingale supermarket. The process $\{X_n^2\}$ is not a supermartingale, but in a sense it is “almost” a supermartingale. Let's fiddle with it to manufacture a supermartingale.

Here's a way to fiddle.

Define $W_n = b_n(X_n^2 + 1)$, where $b_n = 1/\prod_{k=1}^{n-1}(1 + a_k^2 c^2)$. Then $\{W_n\}$ is a supermartingale.

The verification is simple: From (4.45), by dropping the term $2a_n X_n f(X_n)$, which is non-negative by the assumption that $xf(x) \geq 0$ for all x , we obtain

$$E(X_{n+1}^2 \mid X_{0,n}) \leq X_n^2(1 + a_n^2 c^2) + a_n^2 c^2,$$

so that

$$\begin{aligned} E(W_{n+1} \mid X_{0,n}) &= b_{n+1}E(X_{n+1}^2 \mid X_{0,n}) + b_{n+1} \\ &\leq b_{n+1}\{X_n^2(1 + a_n^2 c^2) + a_n^2 c^2\} + b_{n+1} \\ &= b_{n+1}(1 + a_n^2 c^2)(X_n^2 + 1) = b_n(X_n^2 + 1) = W_n. \end{aligned}$$

This is useful: since $\{W_n\}$ is a nonnegative supermartingale, it converges with probability 1. Thus, since the assumption that $\sum a_n^2 < \infty$ implies that the infinite product $\prod_1^\infty (1 + a_n^2 c^2)$ converges [that is, the limit of $\prod_{k=1}^{n-1}(1 + a_k^2 c^2)$ as $n \rightarrow \infty$ is finite], we see that $b_n \rightarrow b$, say, as $n \rightarrow \infty$, where $0 < b < 1$. So since W_n converges with probability 1, it follows that $\lim_{n \rightarrow \infty} X_n^2$ exists with probability 1.

We want to show that in fact $X_n^2 \rightarrow 0$ with probability 1.

Recall that in showing that $\{W_n\}$ is a supermartingale, we dropped a term $-2a_n x_n f(X_n)$ that only helps things. Putting this term back in, we see that, in fact, $\{W_n\}$ is a supermartingale “with room to spare,” in the sense that

$$E(W_{n+1} \mid X_{0,n}) \leq W_n - 2a_n b_{n+1} X_n f(X_n).$$

That is, to have a supermartingale, we need only $E(W_{n+1} \mid X_{0,n}) \leq W_n$, whereas we have an extra negative contribution $-2a_n b_{n+1} X_n f(X_n)$ to the right-hand side. So the process $\{W_n\}$ is better than just a supermartingale; it is like a supermartingale and a batmartingale put together.

Let $\delta > 0$ and define $D = \{x : |x| > \delta\}$. We know that X_n^2 approaches a limit. To show that the limit is 0 with probability 1, it is sufficient to show that $\liminf X_n^2 = 0$ with probability 1. That is, it is sufficient to show that for each m , the set $B_m = \cap_{n=m}^\infty \{X_n \in D\}$ has probability 0. However, letting $\epsilon = \inf_{x \in D} xf(x)$, which is positive [by an assumption in the theorem], for $n \geq m$ we have

$$E(X_n f(X_n)) \geq E(X_n f(X_n) I(B_m)) \geq \epsilon P(B_m),$$

so that

$$EW_{n+1} \leq EW_n - 2a_nb_{n+1}E[X_nf(X_n)] \leq EW_n - 2a_nb_{n+1}\epsilon P(B_m).$$

Iterating this, for $n \geq m$ we have

$$E(W_n) \leq E(W_m) - 2\epsilon P(B_m) \sum_{k=m}^{n-1} a_k b_{k+1}.$$

Since $W_n \geq 0$, so that $E(W_n) \geq 0$, this implies

$$P(B_m) \leq \frac{E(W_m)}{2\epsilon \sum_{k=m}^{n-1} a_k b_{k+1}}.$$

However, since $\sum a_k = \infty$ and $b_k \rightarrow b > 0$, the last expression approaches 0 as $n \rightarrow \infty$. Thus, we must have $P(B_m) = 0$, so we are done. \square

Miscellaneous quick thoughts about things to do (and, eventually, for me to add to these notes):

- For more good clean fun, have your computer simulate and draw pictures of some sample paths from stochastic approximation. It's easy to program: look at the simple recursion! You can explore the effects of various choices of the sequence $\{a_n\}$ for various functions f .
- For further interesting applications of stochastic approximation, look at neural network training algorithms and clustering. Also see the “reinforcement learning” literature: Look for the terms “temporal differences” and “Q-learning.”

4.8 Exercises

[4.1] People first meeting the concepts of “Markov process” and “martingale” sometimes tend to be fuzzy about the distinction, whether one implies the other, and so on. To make sure this doesn't happen to you, give an example of

- a Markov chain that is not a martingale, and
- a martingale that is not a Markov chain.

One of the two questions should be an insult; the other may require some thought.

[4.2] Define a function $f : \mathbb{R} \rightarrow \mathbb{R}$ to be *harmonic* if it satisfies the equation $f((1-\lambda)x + \lambda y) = (1-\lambda)f(x) + \lambda f(y)$ for all real x and y and for all $\lambda \in [0, 1]$. Define f to be *subharmonic* if $f((1-\lambda)x + \lambda y) \leq (1-\lambda)f(x) + \lambda f(y)$ and *superharmonic* if $f((1-\lambda)x + \lambda y) \geq (1-\lambda)f(x) + \lambda f(y)$. In this setting, there are some synonyms that may be more familiar: harmonic functions are linear, subharmonic functions are convex, and superharmonic functions are concave. The sub and super terminology even feels right here; in a graph, a subharmonic function lies below any “chord,” and a superharmonic function stays above its chords. Let

X be a martingale, and let f be a function with $\mathbb{E}|f(X_t)| < \infty$ for all t . Show that if f is subharmonic, then the process $\{f(X_t)\}$ is a submartingale, and if f is superharmonic, then the process $\{f(X_t)\}$ is a supermartingale.

- [4.3] In your best lawyerly manner, marshal arguments that would persuade a jury that the infimum of the empty set of real numbers should be defined to be ∞ .
- [4.4] In the definition of stopping time, show that the requirement that $I\{T = k\}$ be a function of $W_{0,k}$ for each k is equivalent to requiring that $I\{T \leq k\}$ be a function of $W_{0,k}$ for each k .
- [4.5] Suppose that T and U are stopping times with respect to a process W_0, W_1, \dots . Show that each of the following is also a stopping time: (a) $\min\{T, U\}$, (b) $\max\{T, U\}$, and (c) $T + U$.
- [4.6] **[[Wright-Fisher process]]** This is a famous urn models from genetics. An urn contains d balls, some black and some white. **[[These might represent two forms of a gene. We might ask questions like: as genes are sampled randomly to form successive generations, what is the probability that the “white” gene will take over and completely replace the other form?]]** Here is a way to form the next “generation.” Sample a ball at random from the urn d times, *with replacement*. We get a new, random set of d balls, containing $0, 1, \dots$, or d white balls, with various probabilities. Call the resulting sample generation 1. Then we sample in the same way from generation 1 to form generation 2, and so on. Let X_t denote the number of white balls in generation t . After many generations, the population will have become “fixed,” that is, it will consist of just one color. Suppose that $X_0 = x_0$, some number between 0 and d .
- (a) Show that the process $\{X_t\}$ is a martingale.
- (b) Use the martingale to show that the probability that the population eventually becomes all white is x_0/d .
- [4.7] Let M_n denote the fraction of white balls at time n in Polya’s urn, as in Example (4.7). Show that as the time $n \rightarrow \infty$, the fraction M_n approaches a limit with probability 1. What is the distribution of this limiting fraction?
- [[To get an idea of the distribution of the limit, figure out the distribution of M_n for finite n . You can do a few small values of n by hand and see the pattern of your answers.]]**
- [4.8] Consider modifying the definition of the Polya urn from Example (4.7) so that, instead of adding one new ball of the same color at each step of the process, we add c new balls of the same color, where $c \geq 1$. Show that the fraction of white balls is still a martingale. Is the process still a martingale if we add c new balls of the same color and d new balls of the opposite color at each step?

[4.9] [[Asymmetric random walk on the integers]] Consider a random walk $\{S_t\}$ satisfying $S_0 = 0$ and $S_n = X_1 + \dots + X_n$, where X_1, X_2, \dots are *iid* taking values $+1$ and -1 with probabilities p and $1 - p$, respectively. Suppose $p \neq 1/2$.

- (a) There is a special value $\theta \neq 1$ (expressible in terms of p) such that the process $\{M_t\}$ defined by $M_t = \theta^{S_t}$ is a martingale. What is θ ?
- (b) Let a and b be integers satisfying $a < 0 < b$. Let T be the first time the random walk reaches either a or b . Use the martingale you discovered in the previous part to find $\mathbb{P}\{S_T = b\}$.

[4.10] [[Doob's inequality for submartingales]] Let X_0, X_1, \dots be a nonnegative submartingale, and let b be a positive number. Prove that

$$\mathbb{P}\{\max(X_0, \dots, X_n) \geq b\} \leq \frac{\mathbb{E}(X_n)}{b},$$

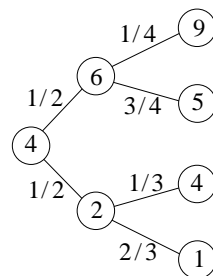
using the following steps as hints.

- (a) Define τ to be the first time t such that $X_t \geq b$, or n , whichever comes first; that is, $\tau = \inf\{t : X_t \geq b\} \wedge n$. Argue that $\{\max(X_0, \dots, X_n) \geq b\} = \{X_\tau \geq b\}$.
- (b) Apply Markov's inequality, and use an Optional Sampling theorem.

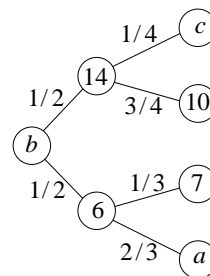
[4.11] upcrossing inequality...

[4.12] Consider a martingale S with respect to the probability measure \mathbb{Q} , and let M be another martingale with respect to \mathbb{Q} , as shown below.

- (a) Solve for the numerical values of a , b , and c .
- (b) What is the trading strategy such that $M_t = M_0 + (H \cdot S)_n$?



Probability measure \mathbb{Q} ,
stock process S



Martingale M
with respect to \mathbb{Q}

- [4.13] In Theorem (4.38) there is nothing special about assuming nonnegativity, that is, a lower bound of zero. Show that if $\{X_t\}$ is a supermartingale and there is a random variable X with $E(|X|) < \infty$ and $X_t \geq X$ for all t , then X_t converges with probability 1 as $t \rightarrow \infty$.
- [4.14] Let X_0, X_1, \dots be a branching process: $X_{t+1} = \sum_{i=1}^{X_t} Z_{ti}$, where the offspring numbers $\{Z_{ti}\}$ are iid with mean $\mu > 1$ and variance $\sigma^2 < \infty$. Define $M_t = \mu^{-t} X_t$.
- (a) Find a formula for the variance of M_t .
 - (b) Show that as $t \rightarrow \infty$, M_t converges with probability 1 to a limit random variable M_∞ .
 - (c) Show that $\mathbb{E}(M_\infty) = 1$.
- [4.15] This is a continuation of the previous problem. If you have not already completed the previous problem but still want to work on this one, you may assume the result of part (b) of the previous problem. This problem shows that the branching process either becomes extinct (hits the value 0) or grows geometrically fast; the probability of any other sort of behavior is 0.
- (a) For each integer b , show that $P\{0 < X_t < b \text{ for all } t\} = 0$.
 - (b) Define $\tau_b = \inf\{t : X_t \geq b\}$. Show that $P\{\tau_b < \infty, X_t = 0 \text{ for some } t\} \rightarrow 0$ as $b \rightarrow \infty$.
 - (c) Show that $P\{\tau_b < \infty, M_\infty = 0\} \rightarrow 0$ as $b \rightarrow \infty$.
 - (d) Show that $P\{\limsup X_t = \infty, M_\infty = 0\} = 0$.
 - (e) Show that $P\{X_t > 0 \text{ for all } t, M_\infty = 0\} = 0$.
- [4.16] Let $\{M_t\}$ be a likelihood ratio martingale as discussed in Example 4.9.
- (a) Show that $\mathbb{E}M_t = 1$ for all t .
 - (b) Show that as $t \rightarrow \infty$, we have $M_t \rightarrow 0$ with probability 1.
- [4.17] Comment on what you found in exercise [4.16]. Don't the results of parts (a) and (b) seem somewhat at odds with each other? How is it possible for them both to hold simultaneously? From a statistical point of view, explain why we should be happy that the result of part (b) holds.
- [4.18] Let X_1, X_2, \dots be independent with
- $$X_t = \begin{cases} -1 & \text{with probability } 1 - \frac{1}{t^2}, \\ t^2 - 1 & \text{with probability } \frac{1}{t^2}. \end{cases}$$
- Define $M_0 = 0$ and $M_t = X_1 + \dots + X_t$ for $t \geq 1$. Show that $\{M_t\}$ is a martingale, and $M_t \rightarrow -\infty$ with probability 1. The moral: Even martingales can be pretty unfair!
- [4.19] Suppose T is a stopping time and $\{X_t\}$ is a submartingale. Define $Y_t = X_{t \wedge T}$. Show that $\{Y_t\}$ is a submartingale.

- [4.20] Give an example to show that the conclusion of Exercise [4.19] can be false without the assumption that T is a stopping time.
- [4.21] Use Exercise [4.19] and the martingale convergence theorem to show, if we start a simple symmetric random walk at 1, with probability 1 it will eventually hit 0.

▷ *The next four exercises are discrete-time versions of results that are considered fundamental and important in the theory of continuous time processes, stochastic integration, and so on. In discrete time, they are elementary and simple; in the occult realm of continuous time, they are much more technical and mysterious.*

We'll use the following definition. We say a process $A = \{A_t : t \geq 0\}$ is **predictable** with respect to a process $W = \{W_t : t \geq 0\}$ if, for each t , the random variable A_t is a function of $W_{0,t-1} = (W_0, \dots, W_{t-1})$. That is, if we think of W_t as representing information available at time t , then we can perfectly predict the random variable A_t **before** time t , that is, at time $t - 1$.

- [4.22] [[Doob decomposition]] Let $\{X_t : t = 0, 1, \dots\}$ be a stochastic process with $\mathbb{E}|X_t| < \infty$ for all t . Let $W = \{W_t\}$ be another stochastic process; the role of W is that the terms “martingale” and “predictable” will be understood to be with respect to W . Show that there exists a martingale M , starting at $M_0 = 0$, and a predictable process A , such that $X_t = X_0 + A_t + M_t$ for all t .

[[Hint: If you just use the required properties of A and M , you will eventually be forced to the right definitions. In case you get stuck and want a hint, though, you can define M by $M_t = \sum_{s=1}^t (X_s - \mathbb{E}(X_s | W_{0,s-1}))$.]]

- [4.23] [[Uniqueness of Doob decomposition]] Show that the Doob decomposition in Exercise [4.22] is unique. That is, show that if \tilde{A} and \tilde{M} satisfy the same conditions as A and M , respectively, then $\tilde{A}_t = A_t$ and $\tilde{M}_t = M_t$ with probability 1, for all t .

- [4.24] [[Doob decomposition for functions of a Markov chain]] Let $\{X_t\}$ be a Markov chain on a finite state space \mathcal{S} with probability transition matrix $P = (P(i, j))$. Given a function $f : \mathcal{S} \rightarrow \mathbb{R}$, define the function $(Pf) : \mathcal{S} \rightarrow \mathbb{R}$ by $(Pf)(i) = \sum_j P(i, j)f(j)$. That is, here we are thinking of f as a column vector and simply doing matrix multiplication. Consider the Doob decomposition of the process $\{f(X_t)\}$ with respect to the process $\{X_t\}$. Show that the martingale part of this Doob decomposition equals

$$(4.46) \quad f(X_t) - f(X_0) - \sum_{s=0}^{t-1} ((P - I)f)(X_s),$$

where I is the identity matrix.

- [4.25] Let $\{X_t\}$ be a stochastic process on a state space \mathcal{S} containing n states. X is *not assumed* to be a Markov chain. Let P be an $n \times n$ probability transition matrix. (This simply means that P has nonnegative entries and each row of P sums to 1. Up to this point, there is no relationship assumed between P and X .) We say that the process X “satisfies the martingale problem for P ” if, for each function $f : \mathcal{S} \rightarrow \mathbb{R}$, the expression (4.46) is a martingale with respect to X . Show that if X satisfies the martingale problem for P , then in fact X is a Markov chain with transition matrix P .

5. Brownian motion

Section 1. The definition and some simple properties.
Section 2. Visualizing Brownian motion. Discussion and demystification of some strange and scary pathologies.
Section 3. The reflection principle.
Section 4. Conditional distribution of Brownian motion at some point in time, given observed values at some other times.
Section 5. Existence of Brownian motion. How to construct Brownian motion from familiar objects.
Section 6. Brownian bridge. Application to testing for uniformity.
Section 7. A boundary crossing problem solved in two ways: differential equations and martingales.
Section 8. Discussion of some issues about probability spaces and modeling.

Brownian motion is one of the most famous and fundamental of stochastic processes. The formulation of this process was inspired by the physical phenomenon of Brownian motion, which is the irregular jiggling sort of movement exhibited by a small particle suspended in a fluid, named after the botanist Robert Brown who observed and studied it in 1827. A physical explanation of Brownian motion was given by Einstein, who analyzed Brownian motion as the cumulative effect of innumerable collisions of the suspended particle with the molecules of the fluid. Einstein's analysis provided historically important support for the atomic theory of matter, which was still a matter of controversy at the time—shortly after 1900. The mathematical theory of Brownian motion was given a firm foundation by Norbert Wiener in 1923; the mathematical model we will study is also known as the “Wiener process.”

Admittedly, it is possible that you might not share an all-consuming fascination for the motion of tiny particles of pollen in water. However, there probably are any number of things that you do care about that jiggle about randomly. Such phenomena are candidates for modeling via Brownian motion, and the humble Brownian motion process has indeed come to occupy a central role in the theory and applications of stochastic processes. How does it fit into the big picture? We have studied Markov processes in discrete time and having a discrete state space. With continuous time and a continuous state space, the prospect arises that a process might have continuous sample paths. To speak a bit roughly for a moment, Markov processes that have continuous sample paths are called *diffusions*. Brownian motion is the simplest diffusion, and in fact other diffusions can be built up from Brownian motions in various ways. Brownian motion and diffusions are used all the time in models in all sorts of fields, such as finance [in modeling the prices of stocks, for

example]], economics, queueing theory, engineering, and biology. Just as a pollen particle is continually buffeted by collisions with water molecules, the price of a stock is buffeted by the actions of many individual investors. Brownian motion and diffusions also arise as approximations for other processes; many processes converge to diffusions when looked at in the right way. In fact, in a sense [[does non-sense count?]], Brownian motion is to stochastic processes as the standard normal distribution is to random variables: just as the normal distribution arises as a limit distribution for suitably normalized sequences of random variables, Brownian motion is a limit in distribution of certain suitably normalized sequences of stochastic processes. Roughly speaking, for many processes, if you look at them from far away and not excessively carefully, they look nearly like Brownian motion or diffusions, just as the distribution of a sum of many *iid* random variables looks nearly normal.

5.1 The definition

Let's scan through the definition first, and then come back to explain some of the words in it.

(5.1) DEFINITION. A **standard Brownian motion (SBM)** $\{W(t) : t \geq 0\}$ is a stochastic process having

- (i) *continuous paths*,
- (ii) *stationary, independent increments*, and
- (iii) $W(t) \sim N(0, t)$ for all $t \geq 0$.

The letter “ W ” is often used for this process, in honor of Norbert Wiener. [[Then again, there is also an object called the “Wiener sausage” studied in physics.]]

The definition contains a number of important terms. First, it is always worth pointing out that a **stochastic process** W is really a function $W : \mathbb{R} \times \Omega \rightarrow \mathbb{R}$, and thus may be viewed in two ways: as a “collection of random variables” and as a “random function” [[or “random path”]]. That is, $W = W(t, \omega)$. For each fixed t , $W(t, \cdot)$ is a real-valued function defined on Ω , that is, a random variable. So W is a collection of random variables. We will use both notations $W(t)$ and W_t for the random variable $W(t, \cdot)$. For fixed ω , $W(\cdot, \omega)$ is a real-valued function defined on \mathbb{R}_+ ; such a function could be viewed as a “path.” Thus W is a random function, or random path. Which brings us to the next item: **continuous paths**. By this we mean that

$$P\{\omega \in \Omega : W(\cdot, \omega) \text{ is a continuous function}\} = 1.$$

Next, the **independent increments** requirement means that for each n and for all choices of times $0 \leq t_0 < t_1 < \dots < t_n < \infty$, the random variables $W(t_1) - W(t_0), W(t_2) - W(t_1), \dots, W(t_n) - W(t_{n-1})$ are independent. The term **stationary increments** means that the distribution of the increment $W(t) - W(s)$ depends only on $t - s$. Note that from

the requirement that $W(t) \sim N(0, t)$, we can see that $W(0) = 0$ with probability one. From this, using (ii) and (iii) of the definition,

$$W(t) - W(s) \sim W(t - s) - W(0) = W(t - s) \sim N(0, t - s)$$

for $s \leq t$. Thus:

The increment that a standard Brownian motion makes over a time interval of length h is normally distributed with mean 0 and variance h .

Here is a very useful alternative characterization of standard Brownian motion. While describing this characterization we will also introduce two important definitions. First, W is a **Gaussian process**, which means that for all numbers n and times t_1, \dots, t_n the random vector $(W(t_1), \dots, W(t_n))$ has a joint normal distribution. An equivalent characterization of the Gaussianity of W is that the sum

$$a_1 W(t_1) + \dots + a_n W(t_n)$$

is normally distributed for each n , all t_1, \dots, t_n , and all real numbers a_1, \dots, a_n . Being a Gaussian process having mean 0, the joint distribution of all finite collections of random variables $W(t_1), \dots, W(t_n)$ are determined by the **covariance function**

$$r(s, t) = \text{Cov}(W_s, W_t).$$

For standard Brownian motion, $\text{Cov}(W_s, W_t) = s \wedge t$. To see this, suppose that $s \leq t$, and observe that

$$\begin{aligned} \text{Cov}(W_s, W_t) &= \text{Cov}(W_s, W_s + (W_t - W_s)) \\ &= \text{Var}(W_s) + \text{Cov}(W_s, W_t - W_s) \\ &= s + 0 = s, \end{aligned}$$

where we have used the independence of increments to say that

$$\text{Cov}(W_s, W_t - W_s) = \text{Cov}(W_s - W_0, W_t - W_s) = 0.$$

It is easy to see [Exercise!] that a process W is Gaussian with mean 0 and covariance function $r(s, t) = s \wedge t$ if and only if (ii) and (iii) of Definition (5.1) hold for W . Thus:

(5.2) A Gaussian process having continuous paths, mean 0, and covariance function $r(s, t) = s \wedge t$ is a standard Brownian motion.

The characterization (5.2) of Brownian motion can be a convenient and powerful tool.

(5.3) EXAMPLE. Suppose that W is a SBM, and define a process X by $X(t) = tW(1/t)$ for $t > 0$, and define $X(0) = 0$. Then we claim that X is also a SBM.

To check this, we'll check that X satisfies the conditions in the last characterization. To start we ask: is X a Gaussian process? Given n, t_1, \dots, t_n , and a_1, \dots, a_n , we have

$$a_1 X(t_1) + \dots + a_n X(t_n) = a_1 t_1 W(1/t_1) + \dots + a_n t_n W(1/t_n),$$

which, being a linear combination of W evaluated at various times, has a normal distribution. Thus, the fact that W is a Gaussian process implies that X is also. Next, observe that the path continuity of X is also a simple consequence of the path continuity of W : if $t \mapsto W(t)$ is continuous, then so is $t \mapsto tW(1/t)$. [Well, this proves that with probability one $X = X(t)$ is continuous for all positive t . For $t = 0$, if you believe that $\lim_{s \rightarrow \infty} W(s)/s = 0$ with probability one—which is eminently believable by the SLLN—then making the substitution $s = 1/t$ gives $\lim_{t \rightarrow 0} tW(1/t) = 0$ with probability 1, so that X is also continuous at $t = 0$. Let's leave it at this for now.] The fact that $X(t)$ has mean 0 is trivial. Finally, to check the covariance function of X , let $s \leq t$ and observe that

$$\begin{aligned} \text{Cov}(X(s), X(t)) &= \text{Cov}(sW(\frac{1}{s}), tW(\frac{1}{t})) = st \text{Cov}(W(\frac{1}{s}), W(\frac{1}{t})) \\ &= st \left(\frac{1}{s} \wedge \frac{1}{t} \right) = st \frac{1}{t} = s. \end{aligned}$$

Thus, X is a SBM. □

The previous example may appear to be at first sight to be a rather odd thing to want to know. However, as we will see, there are times when this particular property of Brownian motion provides just the right insight, relating the behavior of the process on the time interval $(0, 1)$ to its behavior on the time interval $(1, \infty)$.

The next property, sometimes called “Brownian scaling,” is used all the time.

(5.4) THEOREM [BROWNIAN SCALING]. *Suppose that W is a standard Brownian motion, and let $c > 0$. Then the process X defined by $X(t) = c^{-1/2}W(ct)$ is also a standard Brownian motion.*

- ▷ The proof is left to you, as Exercise [5.4]. Having seen Example (5.3), you should find this easy.
- ▷ Exercise [5.5] is an example with a “dimensional analysis” flavor, showing a use of the Brownian scaling relationship to get substantial information about the form of a functional relationship with little effort.

Brownian motion is continually “restarting” in a probabilistic sense. The next proposition is one way of formulating this idea mathematically.

(5.5) PROPOSITION. *Suppose that W is a standard Brownian motion, and let $c > 0$. Define $X(t) = W(c + t) - W(c)$. Then $\{X(t) : t \geq 0\}$ is a standard Brownian motion that is independent of $\{W(t) : 0 \leq t \leq c\}$.*

▷ The proof is again an exercise; see Exercise [5.6].

The Proposition says that, at each time c , the Brownian motion “forgets” its past and continues to wiggle on just as if it were a new, independent Brownian motion. That is, suppose that we know that $W(c) = w$, say. Look at the graph of the path of W ; we are assuming the graph passes through the point (c, w) . Now imagine drawing a new set of coordinate axes, translating the origin to the point (c, w) . So the path now goes through the new origin. Exercise (5.5) says that if we look at the path past time c , relative to the new coordinate axes, we see the path of a new standard Brownian motion, independent of what happened before time c . Brownian motion is a *Markov process*: given the current state, future behavior does not depend on past behavior.

A standard Brownian motion has a constant mean of 0, so that it has no “drift,” and its variance increases at a rate of 1 per second. [For now take the unit of time to be 1 second.] A standard Brownian motion is a standardized version of a general Brownian motion, which need not have $W(0) = 0$, may have a nonzero “drift” μ , and has a “variance parameter” σ^2 that is not necessarily 1.

(5.6) DEFINITION. A process X is called a (μ, σ^2) **Brownian motion** if it can be written in the form

$$X(t) = X(0) + \mu t + \sigma W(t),$$

where W is a standard Brownian motion.

Notice that the mean and variance of a (μ, σ^2) Brownian motion increase at rate μ and σ^2 per second, respectively.

This situation here is analogous to that with normal distributions, where $Z \sim N(0, 1)$ is called a “standard” normal random variable, and general normal random variables are obtained by multiplying a standard normal random variable by something and adding something.

The following characterization of Brownian motion is sometimes useful.

(5.7) FACT. If a stochastic process X has continuous paths and stationary, independent increments, then X is a Brownian motion.

Thus, the assumptions of path continuity and stationary, independent increments is enough to give the normality of the increments “for free.” This is not surprising, from the Central Limit Theorem.

5.2 Visualizing Brownian motion

First, friends, it’s time for some frank talk about Brownian motion. Brownian motion can be very difficult to visualize; in fact, in various respects it’s impossible. Brownian motion has some “pathological” features that make it seem strange and somewhat intimidating. Personally, I remember that after having heard some weird things about Brownian motion, I felt rather suspicious and mistrustful of it, as if I could not use it with confidence or even speak of it without feeling apologetic somehow. We will not dwell unduly on the pathologies

here, but I do not want us to completely avert our eyes, either. Let's try to take enough of a peek so that we will not be forever saddled with the feeling that we have chickened out completely. Hopefully, such a peek will result in an improved level of confidence and familiarity in working with Brownian motion.

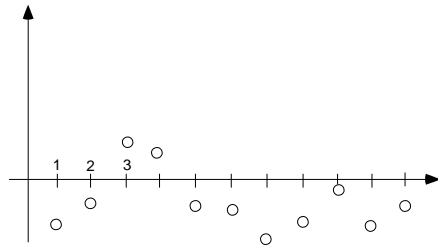
What is all this about “pathologies”? After all, Brownian motion has continuous sample paths, and continuous functions are quite nice already, aren't they? Here's one slightly strange feature, just to get started. Recall that $W(0) = 0$. It turns out that for almost all sample paths of Brownian motion [that is, for a set of sample paths having probability one], for all $\epsilon > 0$, the path has infinitely many zeros in the interval $(0, \epsilon)$. That is, the path changes sign infinitely many times, cutting through the horizontal axis infinitely many times, within the interval $(0, \epsilon)$; a Brownian motion does more before time ϵ than most of us do all day. Another rather mind-boggling property is that with probability 1, a sample path of Brownian motion does not have a derivative *at any time*! It's easy to imagine functions—like $f(t) = |t|$, for example—that fail to be differentiable at isolated points. But try to imagine a function that *everywhere* fails to be differentiable, so that there is not even one time point at which the function has a well-defined slope.

Such functions are not easy to imagine. In fact, before around the middle of the 19th century mathematicians generally believed that such functions did not exist, that is, they believed that every continuous function must be differentiable somewhere. Thus, it came as quite a shock around 1870 when Karl Weierstrass produced an example of a nowhere-differentiable function. Some in the mathematical establishment reacted negatively to this work, as if it represented an undesirable preoccupation with ugly, monstrous functions. Perhaps it was not unlike the way adults might look upon the ugly, noisy music of the next generation. It is interesting to reflect on the observation that, in a sense, the same sort of thing happened in mathematics much earlier in a different context with which we are all familiar. Pythagorus discovered that $\sqrt{2}$ —which he knew to be a perfectly legitimate number, being the length of the hypotenuse of a right triangle having legs of length one—is irrational. Such numbers were initially viewed with great distrust and embarrassment. They were to be shunned; notice how even the name “irrational” still carries a negative connotation. Apparently some Pythagoreans even tried to hide their regrettable discovery. Anyway, now we know that in a sense “almost all” numbers are of this “undesirable” type, in the sense that the natural measures that we like to put on the real numbers [like Lebesgue measure (ordinary length)] place all of their “mass” on the set of irrational numbers and no mass on the set of rational numbers. Thus, the proof of existence of irrational numbers by producing an example of a particular irrational number was dwarfed by the realization that if one chooses a real number at random under the most natural probability measures, the result will be an irrational number with probability 1. The same sort of turnabout has occurred in connection with these horrible nowhere-differentiable functions. Weierstrass constructed a particular function and showed that it was nowhere differentiable. The strange nature of this discovery was transformed in the same sense by Brownian motion, which puts probability 0 on “nice” functions and probability 1 on nowhere differentiable functions.

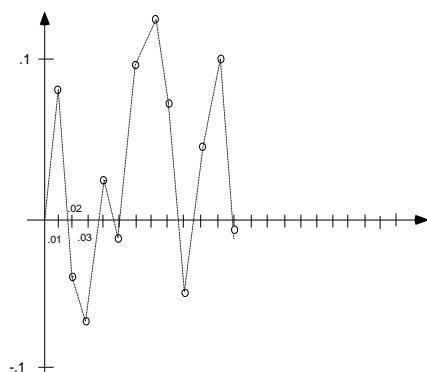
Having presented two scary pathologies, let us now argue that they are not really all that strange or unexpected. We'll start with the fact that a Brownian motion W has

infinitely many zeros in the interval $(0, \epsilon)$. Let b be an arbitrary positive number, perhaps very large. Do you believe that a Brownian motion will necessarily hit 0 infinitely many times in the time interval (b, ∞) ? This proposition seems to me to be quite believable and not at all scary [for example, by recurrence considerations we know that a simple random walk will behave this way]. Well, recall that $X(s) = sW(1/s)$ is a Brownian motion. So you believe that $X(s) = sW(1/s) = 0$ for infinitely many values of s in (b, ∞) . But this implies that $W(t) = 0$ for infinitely many values of t in the interval $(0, 1/b)$. Making the identification of $1/b$ with ϵ shows that the scary pathology and the believable proposition are the same. Now for the nondifferentiability of the Brownian paths. This should not be very surprising, by the assumption of independent increments. Indeed, for each t and each $\delta > 0$, the increment $W(t + \delta) - W(t)$ is independent of the increment $W(t) - W(t - \delta)$, so that it would just be the wildest stroke of luck if the increments on both sides of t “matched up” well enough for W to be differentiable at t !

Enough of that for a while. How does Brownian motion look and behave? We can get a good idea of the behavior on a rough scale by sampling the process at every integer, say. If we are looking at the process over a large time interval and are not concerned about little fluctuations over short time intervals, then this sort of view may be entirely adequate. It is also very easy to understand, since $W(0), W(1), W(2), \dots$ is just a random walk with *iid* standard normal increments. This a very familiar, non-mysterious sort of process.



What if we want to get a more detailed picture? Let's zoom in on the first tenth of a second, sampling the process in time intervals of length 0.01 instead of length 1. Then we might get a picture that looks like this.



We get another normal random walk, this time with the increments having variance 0.01 instead of variance 1. Notice that the standard deviation of the increments is 0.1, which is 10 times bigger than the time interval 0.01 over which the increments take place! That is, the random walk changes by amounts of order of magnitude 0.1 over intervals of length 0.01, so that we get a random walk that has “steep” increments having “slope” on the order of magnitude of 10.

We could continue to focus in on smaller and smaller time scales, until we are satisfied that we have seen enough detail. For example, if we sampled 10,000 times every second instead of 100 times, so that the sampling interval is .0001, the standard deviation of the increments would be $\sqrt{.0001} = .01$, so that the random walk would have even “steeper” increments whose slope is now measured in the hundreds rather than in tens. Notice again how we should not be surprised by Brownian motion’s catastrophic failure of differentiability.

It is reassuring to know that in a sense we can get as accurate and detailed a picture of Brownian motion as we like by sampling in this way, and that when we do so, we simply get a random walk with normally distributed increments.

5.3 A simple calculation with Brownian motion: the reflection principle

Let $\{W_t\}$ be a standard Brownian motion. For $b > 0$, define the first passage time

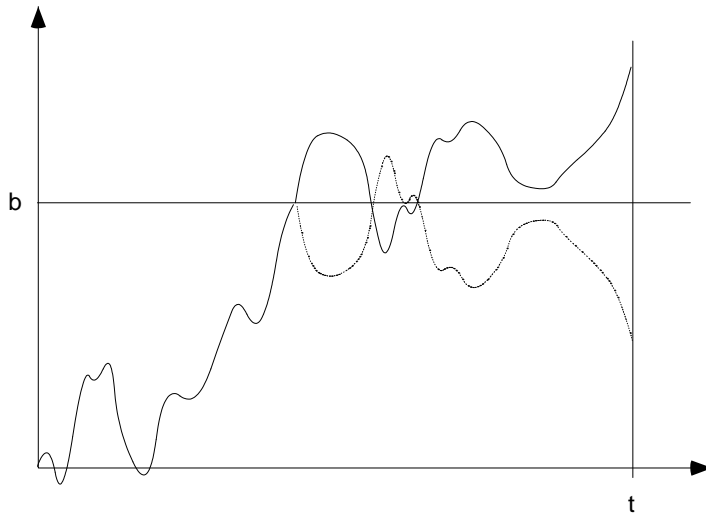
$$\tau_b = \inf\{t : W_t \geq b\};$$

by path continuity, an equivalent definition would be $\inf\{t : W_t = b\}$. Here is a problem: what is $P\{\tau_b \leq t\}$? Here is how to do it. First note that $\{\tau_b \leq t, W_t > b\} = \{W_t > b\}$, since by path continuity and the assumption that $W(0) = 0$, the statement that $W(t) > b$ implies that $\tau_b \leq t$. Using this,

$$\begin{aligned} P\{\tau_b \leq t\} &= P\{\tau_b \leq t, W_t < b\} + P\{\tau_b \leq t, W_t > b\} \\ &= P\{W_t < b \mid \tau_b \leq t\}P\{\tau_b \leq t\} + P\{W_t > b\}. \end{aligned}$$

The term $P\{W_t > b\}$ is easy: since $W_t \sim N(0, t)$, the probability is $1 - \Phi(\frac{b}{\sqrt{t}})$. Next, a little thought will convince you that $P\{W_t < b \mid \tau_b \leq t\} = \frac{1}{2}$. Indeed, path continuity guarantees

that $W_{\tau_b} = b$, so that, knowing that $\tau_b \leq t$, the process is equally likely to continue on to be above b or below b at time t . [A rigorous justification involves the “strong Markov property,” but let’s not get into that now. Also note that path continuity is important here. We could not make a statement like this about a discrete-time random walk having $N(0, 1)$ increments, for example, since there will always be an “overshoot” when such a process first jumps above b .]



Making the above substitutions and solving for $P\{\tau_b \leq t\}$ gives

$$P\{\tau_b \leq t\} = 2P\{W_t > b\} = 2 \left[1 - \Phi \left(\frac{b}{\sqrt{t}} \right) \right],$$

a nice explicit result! This is one reason why people like to use Brownian motion in models—it sometimes allows explicit, tidy results to be obtained.

Here’s another example of a nice, explicit formula. For now let’s just present it “for enrichment”; we’ll come back to derive it later. For Brownian motion $X_t = W_t + \mu t$ with drift μ , defining $\tau_b = \inf\{t : X_t \geq b\}$, we have

$$P_\mu\{\tau_b \leq t\} = 1 - \Phi \left(\frac{b - \mu t}{\sqrt{t}} \right) + e^{2\mu b} \Phi \left(\frac{-b - \mu t}{\sqrt{t}} \right).$$

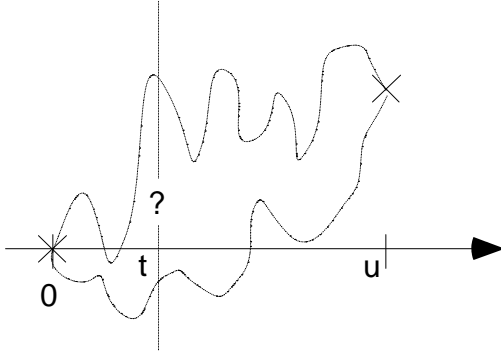
Isn’t that neat?

▷ Exercises [5.10], [5.11], [5.12], and [5.13] apply the reflection principle.

5.4 Conditional distributions for Brownian motion

What happens between sampled points of a Brownian motion? That is, if we are given the value of Brownian motion at two time points, what is the conditional distribution of the

process between those times?



We will examine this question in more detail when we discuss the Brownian bridge in Section 5.6. For now we'll just do what we need for the construction of Brownian motion in the next section. So here is a problem. Let W be a standard Brownian motion (so we know that $W_0 = 0$) and let $0 < t < u$. What is the conditional distribution of W_t given W_u ?

The next result gives a nice way of working with the relevant normal distributions.

(5.8) FACT. $W_t - (t/u)W_u$ is independent of W_u .

To verify this, just note that

$$\text{Cov}(W_t - (t/u)W_u, W_u) = (t \wedge u) - (t/u)(u \wedge u) = t - (t/u)u = 0,$$

and recall that jointly Gaussian variables are independent if they have zero covariance.

The simple fact (5.8) makes everything easy. To get the conditional mean $E(W_t | W_u)$, observe that

$$\begin{aligned} 0 &= E(W_t - (t/u)W_u) \\ &\stackrel{(a)}{=} E(W_t - (t/u)W_u | W_u) \\ &= E(W_t | W_u) - (t/u)W_u, \end{aligned}$$

where (a) follows from the claim. Thus,

$$E(W_t | W_u) = (t/u)W_u.$$

This relation makes sense; undoubtedly you would have guessed it from the picture! The conditional mean of Brownian motion is obtained by linearly interpolating between the points we are given.

For the conditional variance,

$$\begin{aligned}
 \text{Var}(W_t \mid W_u) &= E[(W_t - E(W_t \mid W_u))^2 \mid W_u] \\
 &= E[(W_t - (t/u)W_u)^2 \mid W_u] \\
 &\stackrel{(a)}{=} E[(W_t - (t/u)W_u)^2] \\
 &= (t \wedge t) - 2(t/u)(t \wedge u) + (t/u)^2(u \wedge u) \\
 &= t - 2(t^2/u) + (t^2/u) = t \left(1 - \frac{t}{u}\right) = \frac{t(u-t)}{u},
 \end{aligned}$$

where we have again used the claim at (a). The functional form $t(u-t)/u$ for the conditional variance makes some qualitative sense at least: notice that it approaches 0 as t approaches either of the two points 0 or u , which makes sense. Also, for fixed u , the conditional variance is maximized by taking t in the middle: $t = u/2$.

In summary, we have found:

$$(5.9) \quad \text{For } 0 < t < u, \quad \mathcal{L}(W_t \mid W_u) = N\left(\frac{t}{u}W_u, \frac{t(u-t)}{u}\right).$$

Observe that the conditional variance does not depend on W_u ! Does this surprise you? For example, to take an extreme example, we have found that the conditional distribution of $W(1/2)$ given that $W(1) = 10$ billion is normal with mean 5 billion and variance $1/4$! How do you feel about that? In this example, should we really claim that $W(1/2)$ is within 1.5 (that's 3 standard deviations) of 5 billion with probability 0.997? Well, that is what is implied by the Brownian motion model. Here is one way to conceptualize what is going on. It's *extremely* painful for a Brownian motion to get to 10 billion at time 1 (that's why it is so extremely rare). Among all of the painful ways the Brownian motion can do this, by far the least painful is for it to spread the pain equally over the two subintervals $[0, 1/2]$ and $[1/2, 1]$, making an increment of very nearly 5 billion over each. This last property of Brownian motion, which basically stems from the small tail of the normal distribution, could be viewed as a defect of the model. In real life, if one observes a value that seems outlandish according to our model, such as the value $W(1) = 10$ billion as discussed above, it does not seem sensible to be pig-headedly sure about the value of $W(1/2)$. In fact, an outlandish observation should be an occasion for healthy respect for the limitations of models in general and for skepticism about the suitability of this model in particular, which should lead to a humbly large amount of uncertainty about the value of $W(1/2)$.

▷ Exercises [5.14], [5.15], and [5.16] concern this sort of conditional distribution.

5.5 Existence and construction of Brownian motion (Or: Let's Play Connect-the-Dots)

I like the way David Freedman puts it: "One of the leading results on Brownian motion is that it exists." It is indeed comforting to know that we have not been talking about nothing.

We will show that Brownian motion exists by “constructing” it. This means that we will show how we can obtain Brownian motion by somehow putting together other simpler, more familiar things whose existence we are not worried about. Why do I want to do this? It is not because I think that a mathematical proof of the existence of Brownian motion is somehow legally necessary before we can do anything else (although I suppose it could be a bit embarrassing to be caught analyzing things that do not exist). Rather, it is because I think that seeing how a construction of Brownian motion works gives one a much better, more “familiar” feeling for Brownian motion. Personally, after having heard some weird things about Brownian motion, I felt much less queasy about it after seeing how it could be constructed rather simply from familiar objects. For gaining familiarity and understanding, there’s nothing like taking something apart and putting it together again.

We will construct Brownian motion on the time interval $[0,1]$; having done that, it will be easy to construct Brownian motion on $[0, \infty)$. We’ll do it by an intuitive connect-the-dots approach, in which at each stage of the construction we obtain a more and more detailed picture of a sample path. We know $W(0) = 0$. At the initial stage, we start with the modest goal of simulating the value of the Brownian motion at time 1. Since we know that $W(1) \sim N(0, 1)$, we can do this by going to our computer and generating a $N(0, 1)$ random variable Z_1 ; take Z_1 to be $W(1)$. Given just the information that the path passes through the two points $(0, 0)$ and $(1, Z_1)$, the conditional expectation is the linear interpolation $X^{(0)}$ shown in Figure 1, that is, $X^{(0)}(t) = Z_1 t$. This will be our first crude approximation to a sample path.

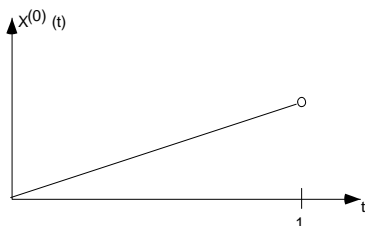


Figure 1

Next let’s simulate a value for $W(1/2)$. Given the values we have already generated for $W(0)$ and $W(1)$, we know that $W(1/2)$ is normally distributed with mean $Z_1/2$ and variance $(1/2)(1/2) = 1/4$. Since $X^{(0)}(1/2)$ is already $Z_1/2$, we need only add a normal random variable with mean 0 and variance $1/4$ to $X^{(0)}(1/2)$ to get the right distribution. Accordingly, generate another independent $N(0, 1)$ random variable Z_2 and take $W(1/2)$ to be $X^{(0)}(1/2) + (1/2)Z_2$. Having done this, define the approximation $X^{(1)}$ to be the piecewise linear path joining the three points $(0,0)$, $(1/2, W(1/2))$, and $(1, W(1))$ as in Figure 2.

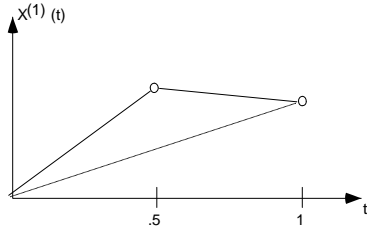


Figure 2

Now let's simulate $W(1/4)$ and $W(3/4)$. Notice how the correct conditional means are already given by the piecewise linear path $X^{(1)}$; that is, $E(W(t) \mid W(0), W(1/2), W(1)) = X^{(1)}(t)$; this holds for all t , and in particular for $t = 1/4$ and $t = 3/4$. The conditional variance of $W(1/4)$ given $W(0)$, $W(1/2)$, and $W(1)$ is $(1/4)(1/4)/(1/2) = 1/8$. Similarly, the conditional variance of $W(3/4)$ is $1/8$. Thus, to simulate these points we generate two more independent standard normal random variables Z_3 and Z_4 , and define

$$W(1/4) = X^{(1)}(1/4) + \frac{1}{\sqrt{8}}Z_3,$$

$$W(3/4) = X^{(1)}(3/4) + \frac{1}{\sqrt{8}}Z_4.$$

The approximation $X^{(2)}$ is then defined to be the piecewise linear interpolation of the simulated values we have obtained for the times 0, $1/4$, $1/2$, $3/4$, and 1, as in Figure 3.

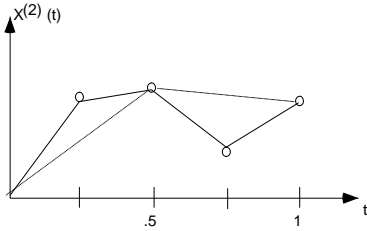


Figure 3

One more time: Given the values simulated so far, each of $W(1/8)$, $W(3/8)$, $W(5/8)$, and $W(7/8)$ has conditional variance $(1/8)(1/8)/(1/4) = 1/16$. So we can simulate $W(1/8)$, $W(3/8)$, $W(5/8)$, and $W(7/8)$ by multiplying some more standard normal random variables Z_5, Z_6, Z_7, Z_8 by $\sqrt{1/16} = 1/4$ and adding these to the values $X^{(2)}(1/8)$, $X^{(2)}(3/8)$, $X^{(2)}(5/8)$, and $X^{(2)}(7/8)$ given by the previous approximation. The piecewise linear interpolation gives $X^{(3)}$.

And so on. In general, to get from $X^{(n)}$ to $X^{(n+1)}$, we generate 2^n new standard normal random variables $Z_{2^n+1}, Z_{2^n+2}, \dots, Z_{2^{n+1}}$, multiply these by the appropriate conditional standard deviation $\sqrt{2^{-n-2}} = 2^{-(n/2)-1}$, and add to the values $X^{(n)}(1/2^{n+1}), X^{(n)}(3/2^{n+1}), \dots, X^{(n)}(1 - 1/2^{n+1})$ to get the new values $X^{(n+1)}(1/2^{n+1}), X^{(n+1)}(3/2^{n+1}), \dots, X^{(n+1)}(1 - 1/2^{n+1})$.

(5.10) CLAIM. *With probability 1, the sequence of functions $X^{(1)}, X^{(2)}, \dots$ converges uniformly over the interval $[0, 1]$.*

The importance of the uniformity of the convergence stems from the following fact from analysis:

The limit of a uniformly convergent sequence of continuous functions is a continuous function.

[[To appreciate the need for uniformity of convergence in order to be guaranteed that the limit function is continuous, recall the following standard example. For $n = 1, 2, \dots$ consider the function $t \mapsto t^n$ for $t \in [0, 1]$. Then as $n \rightarrow \infty$, this converges to 0 for all $t < 1$ whereas it converges to 1 for $t = 1$, so that the limit is not a continuous function.]] Since each of the functions $X^{(n)}$ is clearly continuous, the claim then implies that with probability 1, the sequence $X^{(1)}, X^{(2)}, \dots$ converges to a limit function X that is continuous.

PROOF: Define the maximum difference M_n between $X^{(n+1)}$ and $X^{(n)}$ by

$$M_n = \max_{t \in [0,1]} |X^{(n+1)}(t) - X^{(n)}(t)|.$$

Note that if $\sum M_n < \infty$, then the sequence of functions $X^{(1)}, X^{(2)}, \dots$ converges uniformly over $[0,1]$. Thus, it is sufficient to show that $P\{\sum M_n < \infty\} = 1$. Observe that

$$M_n = 2^{-(n/2)-1} \max\{|Z_{2^n+1}|, |Z_{2^n+2}|, \dots, |Z_{2^{n+1}}|\}.$$

We will use the following result about normal random variables.

(5.11) FACT. *Let G_1, G_2, \dots be iid standard normal random variables, and let c be a number greater than 2. Then*

$$P\{|G_n| \leq \sqrt{c \log n} \text{ for all sufficiently large } n\} = 1.$$

PROOF: Remember the tail probability bound

$$P\{G > x\} \leq \frac{\varphi(x)}{x} = \frac{1}{\sqrt{2\pi}} \frac{e^{-x^2/2}}{x}$$

for a standard normal random variable G and for $x > 0$. From this,

$$\sum_{n=1}^{\infty} P\{|G_n| > \sqrt{c \log n}\} \leq 2 \frac{1}{\sqrt{2\pi}} \sum_{n=1}^{\infty} \frac{e^{-(1/2)c \log n}}{\sqrt{c \log n}} = \sqrt{\frac{2}{\pi}} \sum_{n=1}^{\infty} \frac{n^{-(1/2)c}}{\sqrt{c \log n}},$$

which is finite for $c > 2$. Thus, by the Borel-Cantelli lemma,

$$P\{|G_n| > \sqrt{c \log n} \text{ infinitely often}\} = 0,$$

which is equivalent to the desired statement. □

Taking $c > 2$, the fact implies that with probability 1,

$$M_n \leq 2^{-(n/2)-1} \sqrt{c \log(2^{n+1})}$$

holds for all sufficiently large n . That is,

$$P\{M_n \leq 2^{-(n/2)-1} \sqrt{n+1} \sqrt{c \log 2} \text{ eventually}\} = 1.$$

Thus, since $\sum 2^{-(n/2)} \sqrt{n+1} < \infty$, we have $\sum M_n < \infty$ with probability 1, which completes the proof of Claim (5.10). \square

So we know that with probability 1, the limit $X = \lim_{n \rightarrow \infty} X^{(n)}$ is a well-defined, continuous function. That is, we have established the existence of a limit process X , and the process X has continuous paths. It remains to check that X satisfies the other defining properties of a standard Brownian motion, that is, $X(t) \sim N(0, t)$ and the process X has stationary independent increments.

To check that $X(t) \sim N(0, t)$, first note that for any *dyadic rational* r , that is, any number r of the form $r = k/2^m$, we already know that $X(r) \sim N(0, r)$. [Why? Because for all $n \geq m$, we have $X^{(n)}(k/2^m) = X^{(m)}(k/2^m)$; that is, in the construction process just described, once we assign a value to the process at a dyadic rational time $r = k/2^m$, we never change it. So $X(r) = X^{(m)}(r)$. But the process $X^{(m)}$ was constructed so that $X^{(m)}(k/2^m) \sim N(0, k/2^m)$.] Letting t be an arbitrary number in the interval $[0, 1]$, choose a sequence of dyadic rational numbers r_1, r_2, \dots such that $\lim_{n \rightarrow \infty} r_n = t$. Then $X(t) = \lim_{n \rightarrow \infty} X(r_n)$ with probability 1, by the path continuity of the process X . Thus, since $X(r_n) \sim N(0, r_n)$ and $\lim_{n \rightarrow \infty} r_n = t$, we must have $X(t) \sim N(0, t)$. [Just in case you have any doubts, think of it this way: we have shown that $X(r_n) \sim \sqrt{r_n} Z$, where $Z \sim N(0, 1)$. Take the limit of both sides as $n \rightarrow \infty$. $X(r_n) \rightarrow X(t)$ by path continuity, and $\sqrt{r_n} Z \rightarrow \sqrt{t} Z \sim N(0, t)$.]

Checking that X has stationary independent increments may be done by exactly the same idea. For example, to show that $X(u) - X(t) \sim N(0, u - t)$, $X(t) - X(s) \sim N(0, t - s)$, and $X(u) - X(t)$ is independent of $X(t) - X(s)$ for $s < t < u$, take three dyadic rational sequences $s_n \rightarrow s$, $t_n \rightarrow t$, and $u_n \rightarrow u$. Then note that by construction, $X(u_n) - X(t_n) \sim N(0, u_n - t_n)$ is independent of $X(t_n) - X(s_n) \sim N(0, u_n - t_n)$ for all n . That is,

$$(X(t_n) - X(s_n), X(u_n) - X(t_n)) \sim (\sqrt{t_n - s_n} Z_1, \sqrt{u_n - t_n} Z_2),$$

where Z_1 and Z_2 are *iid* $N(0, 1)$ random variables. Finally, take the limit of both sides and use the path continuity of X . This completes the construction of standard Brownian motion.

5.6 The Brownian bridge

A standard Brownian bridge over the interval $[0, 1]$ is a standard Brownian motion $W(\cdot)$ conditioned to have $W(1) = 0$. People say the Brownian motion is “tied down” at time 1 to have the value 0. By Exercise ([5.15]), we know that $E(W(t) \mid W(1) = 0) = 0$ and $\text{Cov}(W(s), W(t) \mid W(1) = 0) = s(1 - t)$ for $0 \leq s \leq t \leq 1$.

(5.12) DEFINITION. A standard Brownian bridge is a Gaussian process X with continuous paths, mean 0, and covariance function $\text{Cov}(X(s), X(t)) = s(1 - t)$ for $0 \leq s \leq t \leq 1$.

Here is an easy way to manufacture a Brownian bridge from a standard Brownian motion: define

$$(5.13) \quad X(t) = W(t) - tW(1) \quad \text{for } 0 \leq t \leq 1.$$

It is easy and pleasant to verify that the process X defined this way satisfies the definition of a Brownian bridge; I wouldn't dream of denying you the pleasure of checking it for yourself! Notice that, given the construction of standard Brownian motion W , now we do not have to worry about the existence or construction of the Brownian bridge. Another curious but sometimes useful fact is that the definition

$$(5.14) \quad Y(t) = (1-t)W\left(\frac{t}{1-t}\right) \quad \text{for } 0 \leq t < 1, \quad Y(1) = 0$$

also gives a Brownian bridge.

5.6.1 A boundary crossing probability

Earlier, using the reflection principle, we found the probability that a Brownian motion reaches a certain height by a certain time. What is this probability for a Brownian bridge? Letting W be a standard Brownian motion, recall the definition of the first hitting time of the positive level b :

$$\tau_b = \inf\{t : W(t) = b\}.$$

In the standard Brownian bridge, we considered the particular condition $W(1) = 0$. Instead of the particular time 1 and the particular value 0, let us consider the general condition where we tie the Brownian motion down at an arbitrary time t to an arbitrary value x , so that we are interested in the probability $P\{\tau_b \leq t \mid W(t) = x\}$. Clearly by path continuity the answer is 1 if $x \geq b$, so let us assume that $x < b$. Adopting rather informal notation, we have

$$(5.15) \quad P\{\tau_b \leq t \mid W(t) = x\} = \frac{P\{\tau_b \leq t, W(t) \in dx\}}{P\{W(t) \in dx\}}.$$

Heuristically, dx is a tiny interval around the point x . Or, somewhat more formally, think of the dx as shorthand for a limiting statement—the usual limiting idea of conditioning on a random variable taking on a particular value that has probability 0. We can calculate the right side of (5.15) explicitly as follows:

$$\begin{aligned} \text{numerator} &= P\{\tau_b \leq t\}P\{W(t) \in dx \mid \tau_b < t\} \\ &= P\{\tau_b \leq t\}P\{W(t) \in 2b - dx \mid \tau_b < t\} \\ &= P\{W(t) \in 2b - dx, \tau_b < t\} \\ &= P\{W(t) \in 2b - dx\} \\ &= \frac{1}{\sqrt{t}}\varphi\left(\frac{2b-x}{\sqrt{t}}\right) dx \end{aligned}$$

and of course

$$\text{denominator} = \frac{1}{\sqrt{t}}\varphi\left(\frac{x}{\sqrt{t}}\right) dx,$$

so that

$$\begin{aligned} P\{\tau_b < t \mid W(t) = x\} &= \frac{\varphi\left(\frac{2b-x}{\sqrt{t}}\right)}{\varphi\left(\frac{x}{\sqrt{t}}\right)} \\ &= \exp\left[-\frac{1}{2}\frac{(2b-x)^2}{t} + \frac{1}{2}\frac{x^2}{t}\right] \\ &= \exp\left[\frac{-2b(b-x)}{t}\right], \end{aligned}$$

a handy formula! Of course, it makes qualitative sense: the probability goes to 0 [very fast!] as $b \rightarrow \infty$, and the probability is nearly 1 if b is small or if $b - x$ is small or if t is large.

5.6.2 Application to testing for uniformity

Suppose U_1, \dots, U_n are *iid* having a distribution F on $[0,1]$, and we are interested in testing the hypothesis that F is the uniform distribution $F(t) = t$. The empirical distribution function F_n is defined by

$$F_n(t) = \frac{1}{n} \sum_{i=1}^n I\{U_i \leq t\} \quad \text{for } 0 \leq t \leq 1.$$

Thus, $F_n(t)$ is the fraction of the sample that falls in the interval $[0, t]$; this is a natural estimator of $F(t)$, the probability that one random observation from F falls in $[0, t]$. By the law of large numbers, $F_n(t) \rightarrow F(t)$ for all t as $n \rightarrow \infty$. So if F is $\text{Unif}[0,1]$, we have $F_n(t) \rightarrow t$. The idea of the **Kolmogorov-Smirnov test** is to look at the difference $F_n(t) - t$, and reject the uniformity hypothesis if the difference gets large enough at any $t \in [0, 1]$. The question is: how large is large enough? For example, we might want to find a rejection threshold that gives a probability of false rejection of .05; that is, find b so that $P\{\max(F_n(t) - t) : t \in [0, 1]\} = .05$.

Again, the Strong Law of Large Numbers says that for all t , the difference $F_n(t) - t$ approaches 0 as $n \rightarrow \infty$. A limit distribution is obtained by multiplying the difference by \sqrt{n} : since

$$\text{Var}(I\{U_1 \leq t\}) = P\{U_1 \leq t\} - (P\{U_1 \leq t\})^2 = t(1 - t)$$

the Central Limit Theorem tells us that

$$\sqrt{n}(F_n(t) - t) \xrightarrow{D} N(0, t(1 - t)).$$

So define $X_n(t) = \sqrt{n}(F_n(t) - t)$. Then, similarly to the above, since

$$\text{Cov}(1_{\{U_1 \leq s\}}, 1_{\{U_1 \leq t\}}) = P\{U_1 \leq s, U_1 \leq t\} - (P\{U_1 \leq s\})(P\{U_1 \leq t\}) = s - st = s(1 - t)$$

for $s \leq t$, the vector Central Limit Theorem tells us that

$$\begin{pmatrix} X_n(s) \\ X_n(t) \end{pmatrix} = \begin{pmatrix} \sqrt{n}(F_n(s) - s) \\ \sqrt{n}(F_n(t) - t) \end{pmatrix} \xrightarrow{D} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} s & s \\ s & t \end{pmatrix}\right) \sim \begin{pmatrix} X(s) \\ X(t) \end{pmatrix},$$

where X is a Brownian bridge, and, in general,

$$\begin{pmatrix} X_n(t_1) \\ \vdots \\ X_n(t_k) \end{pmatrix} \xrightarrow{\mathcal{D}} N \left(\begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} t_1 & \cdots & t_1 \\ \vdots & & \\ t_1 & & t_k \end{pmatrix} \right) \sim \begin{pmatrix} X(t_1) \\ \vdots \\ X(t_k) \end{pmatrix}.$$

Thus, as $n \rightarrow \infty$, the joint distribution of process X_n sampled at any finite number of time points converges to the joint distribution to the Brownian bridge X sampled at those same times. Therefore, for any finite collection of times $T = \{t_1, \dots, t_k\} \subset [0, 1]$,

$$\lim_{n \rightarrow \infty} P\{\max\{X_n(t) : t \in T\} \geq b\} = P\{\max\{X(t) : t \in T\} \geq b\}$$

This leads one to suspect that we should also have

$$\lim_{n \rightarrow \infty} P\{\max\{X_n(t) : t \in [0, 1]\} \geq b\} = P\{\max\{X(t) : t \in [0, 1]\} \geq b\}$$

In fact, this last convergence can be rigorously shown; the proof is a bit too involved for us to get into now. [For the general subject of weak convergence of stochastic processes see the books *Convergence of Stochastic Processes* by David Pollard and *Convergence of Probability Measures* by P. Billingsley.] Since we know the exact expression for the last probability, we can say that

$$\lim_{n \rightarrow \infty} P\{\max\{X_n(t) : t \in [0, 1]\} \geq b\} = e^{-2b^2}.$$

Thus, for example, since $e^{-2b^2} = 0.05$ for $b = 1.22$, then if n is large we have

$$P\{\max\{X_n(t) : t \in [0, 1]\} \geq 1.22\} \approx 0.05.$$

So we have found an approximate answer to our question of setting a rejection threshold in the test for uniformity.

5.7 Two Approaches to a Simple Boundary Crossing Problem

Let W be standard Brownian motion as usual, let $b > 0$, and define $\tau_b = \inf\{t : W_t = b\}$. Recall that the reflection principle gives

$$P\{\tau_b \leq t\} = 2P\{W_t > b\} = 2 \left[1 - \Phi \left(\frac{b}{\sqrt{t}} \right) \right],$$

from which, letting $t \rightarrow \infty$ with b fixed, it follows that $P\{\tau_b < \infty\} = 1$. That is, W is sure to cross the horizontal level b eventually.

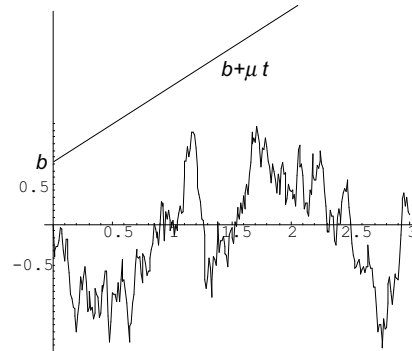
More generally, we could ask:

Problem A: *What is the probability*

$$P\{W(t) = b + \mu t \text{ for some } t \geq 0\}$$

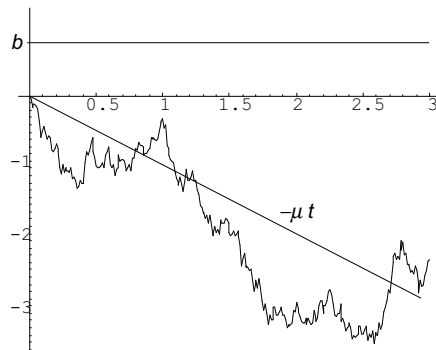
that the process W ever crosses the linear boundary $b + \mu t$?

Clearly the answer is 1 if $\mu < 0$, and we just observed in the previous paragraph that the answer is also 1 when $\mu = 0$. Accordingly, let's look at linear boundaries having positive slope $\mu > 0$.



Note that if we subtract μt from W , we get Brownian motion with drift $-\mu$, and if we subtract μt from the line $b + \mu t$, we get the horizontal level b . Thus, letting $X(t) := W(t) - \mu t$ and defining the stopping time τ_b by $\tau_b = \inf\{t : X_t = b\}$, it is clear that our problem is equivalent to the following problem.

Problem A': Find $P_{-\mu}\{\tau_b < \infty\}$. Here the subscript “ $-\mu$ ” is attached to the P in order to remind us that the Brownian motion that we are considering has drift $-\mu$.

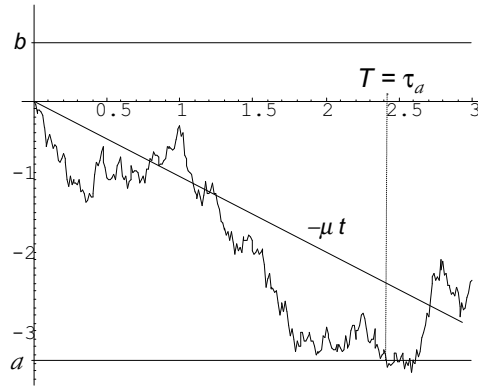


To solve problems A and A', we will in fact solve the following “better” problem.

Problem B: Let $a < 0 < b$ and μ be given, and define $T = \min\{\tau_a, \tau_b\}$. What is the probability

$$P_{-\mu}\{\tau_b < \tau_a\} = P_{-\mu}\{X_T = b\}$$

that the Brownian motion hits the level b before the level a ?



Solving Problem B will enable us to solve the other two, since

$$P_{-\mu}\{\tau_b < \infty\} = \lim_{a \rightarrow -\infty} P_{-\mu}\{\tau_b < \tau_a\}.$$

5.7.1 Differential Equations

We could formulate Problem B in terms of an absorbing Markov process. Consider a $(-\mu, 1)$ -Brownian motion $\{X_t\}$ having two absorbing states a and b . Then our problem is to find the probability that the process gets absorbed in b rather than a . This should have a familiar feel to it: we solved this sort of problem for finite-state Markov chains earlier. In fact, we did this by two different methods, both of which solved the given problem by simultaneously solving the whole family of analogous problems starting from all of the possible starting states of the chain. The first method used the “fundamental matrix.” The second method involved conditioning on what the chain did at the first step.

Here we will do the same sort of thing, using the continuous-time analog of the “conditioning on what happened at the first step” method. We won’t try to be rigorous here. Let P^x and E^x denote probability and expectation when the $(-\mu, 1)$ -Brownian motion $\{X_t\}$ starts in the state $X_0 = x$. Then, defining the function $u(x) = P^x\{X_T = b\}$, Problem B asks for the value of $u(0)$, which we will find by in fact solving for the whole function $u(x)$ for $x \in [a, b]$. Clearly $u(a) = 0$ and $u(b) = 1$, so let $x \in (a, b)$. In continuous time there is no “first step” of the process, but we can think of conditioning on the value of $X(h)$ where h is a tiny positive number. This gives

$$u(x) = E^x P^x\{X_T = b | X(h)\} = E^x u[X(h)] + o(h),$$

where in the last equality we have used the Markov “restarting” property to say that $P^x\{X_T = b | X(h), T > h\} = u[X(h)]$, and Exercise [5.11] to say that $P^x\{T \leq h\} = o(h)$.

Now since h is tiny, $X(h)$ will be very close to x with high probability under P^x . So $u[X(h)]$ can be closely approximated by the first few terms of its Taylor expansion

$$u[X(h)] = u(x) + u'(x)[X(h) - x] + (1/2)u''(x)[X(h) - x]^2 + \cdots.$$

Combining this with the previous equation gives

$$u(x) = u(x) + u'(x)E^x[X(h) - x] + (1/2)u''(x)E^x\{[X(h) - x]^2\} + \cdots$$

However, the $(-\mu, 1)$ -Brownian motion $\{X_t\}$ satisfies

$$\begin{aligned} E^x[X(h) - x] &= -\mu h + o(h), \\ E^x\{[X(h) - x]^2\} &= h + o(h), \end{aligned}$$

and

$$E^x\{[X(h) - x]^k\} = o(h) \text{ for } k > 2.$$

Thus,

$$0 = u'(x)(-\mu h) + (1/2)u''(x)h + o(h),$$

so that, dividing through by h and letting $h \downarrow 0$, we see that u satisfies the differential equation

$$(5.16) \quad (1/2)u''(x) - \mu u'(x) = 0.$$

The boundary conditions are $u(a) = 0$ and $u(b) = 1$. This differential equation is very easy to solve: the general solution is $u(x) = Ce^{2\mu x} + D$ where C and D are constants, so the solution satisfying the boundary conditions is

$$(5.17) \quad u(x) = \frac{e^{2\mu x} - e^{2\mu a}}{e^{2\mu b} - e^{2\mu a}}.$$

This is a handy, explicit result, from which the solution to our original problems follow easily. In particular, since we wanted our Brownian motion X to start at 0, we are interested in

$$P_{-\mu}\{\tau_b < \tau_a\} = u(0) = \frac{1 - e^{2\mu a}}{e^{2\mu b} - e^{2\mu a}}.$$

Since $\mu > 0$, by letting $a \rightarrow -\infty$ we obtain

$$P_{-\mu}\{\tau_b < \infty\} = \lim_{a \rightarrow -\infty} P_{-\mu}\{\tau_b < \tau_a\} = e^{-2\mu b}.$$

Let's pause to make sure that (5.17) is consistent with what we know about the cases where μ is not positive. If μ were negative, so that $-\mu$ were positive, then we would have obtained the limit $\lim_{a \rightarrow -\infty} u(0) = 1$, which makes sense, as we said before: Brownian motion with positive drift is sure to pass through the level b eventually. What if $\mu = 0$? In that case the solution (5.17) breaks down, reducing to the form "0/0". What happens is that the differential equation (5.16) becomes $u''(x) = 0$, whose solutions are linear functions of x . The solution satisfying the boundary conditions $u(a) = 0$ and $u(b) = 1$ is

$$u(x) = \frac{x - a}{b - a},$$

so that $u(0) = a/(a - b)$, which approaches 1 as $a \rightarrow -\infty$, again as expected.

It is interesting to contemplate our solution still a bit more. Let M denote the maximum height ever attained by the Brownian motion, that is, $M = \max\{X_t : t \geq 0\}$. Then what we have found is that $P_{-\mu}\{M \geq b\} = e^{-2\mu b}$, or, in other words, $M \sim \text{Exp}(2\mu)$. Now this result also makes a good deal of sense, at least qualitatively: we could have guessed that M should be exponentially distributed, since it seems intuitively clear that M should have the memoryless property. To see this, suppose I tell you that I observed M to be greater than b , and I ask you for your conditional probability that $M > b + y$. Then you should think to yourself: “He just told me that $\tau_b < \infty$. The portion of the BM path after time τ_b should look just like ordinary BM with drift $-\mu$, except started at the level b . [Then you should mutter something about the strong Markov property [see Exercise 5.8]]: τ_b is a stopping time.] So I should say $P_{-\mu}\{M > b + y \mid M > b\} = P_{-\mu}\{M > y\}$.” This is the memoryless property.

The fact that the parameter of the exponential distribution should be 2μ does not seem obvious, so our calculation has given us some substantial information. Again, you should check that 2μ at least makes some crude, qualitative sense; for example, it is monotone in the right direction and that sort of thing.

5.7.2 Martingales

The results we discussed in the previous chapter for martingales in discrete time have counterparts for continuous-time martingales. Here is a definition. Suppose that W is a standard Brownian motion.

(5.18) DEFINITION. *A stochastic process M is a martingale with respect to W if it satisfies the following two conditions:*

1. *M is adapted to W ; that is, for every t , $M(t)$ is some deterministic function of the portion $\langle W \rangle_0^t := \{W(s) : 0 \leq s \leq t\}$ of the path of W up to time t .*
2. *For all $0 < s < t$ we have $E\{M(t) \mid \langle W \rangle_0^s\} = M(s)$.*

The second condition is a “fair game” sort of requirement. If we are playing a fair game, then we expect to neither win nor lose money on the average. Given the history of our fortunes up to time s , our expected fortune $M(t)$ at a future time $t > s$ should just be the fortune $M(s)$ that we have at time s .

The important *optional sampling* (“conservation of fairness”) property of martingales extends to continuous time. Let M be a martingale with respect to W , and suppose that we know that $M(0)$ is just the constant m_0 . By the “fair game” property, $EM(t) = m_0$ for all times $t \geq 0$. [Exercise: check this!] That is, I can say “stop” at any predetermined time t , like $t = 8$, say, and my winnings will be “fair”: $EM(8) = m_0$. As before, the issue of optional sampling is this: If τ is a *random* time, that is, τ is a nonnegative random variable, does the equality $EM(\tau) = m_0$ still hold? As before, we can be assured that this holds if we rule out two sorts of obnoxious behaviors: “taking too long” and “taking back moves.” That is, optional sampling holds for bounded stopping times.

(5.19) DEFINITION. *We say that a nonnegative random variable τ is a stopping time (with*

respect to W) if for each t it is possible to determine whether or not $\tau \leq t$ just by looking at $\langle W \rangle_0^t$. That is, the indicator random variable $I\{\tau \leq t\}$ is a function of $\langle W \rangle_0^t$.

[[We could also perversely express this in the language introduced above by saying that τ is a stopping time if the process X defined by $X(t) = \{\tau \leq t\}$ is adapted to W .]] We say that a random time τ is bounded if there is a number c such that $P\{\tau \leq c\} = 1$.

(5.20) THEOREM [OPTIONAL SAMPLING THEOREM]. *Suppose that M is a martingale with respect to W starting at the value $M(0) = m_0$, and let τ be a bounded stopping time with respect to W . Then we have $EM(\tau) = m_0$.*

You may be disappointed by the boundedness restriction on the stopping time τ . However, often we can prove optional sampling for unbounded stopping times by combining the above optional sampling theorem with a result like the bounded convergence theorem, for example. We will see this in our application, to which I think it is high time we got back now. If in the force Yoda's so strong, construct a sentence with the words in the proper order then why can't he?

The next result introduces a martingale called "Wald's martingale" or "the exponential martingale." Any martingale with more than one name must be important!

(5.21) CLAIM. *For any real λ , $M(t) := \exp\{\lambda W(t) - \frac{1}{2}\lambda^2 t\}$ is a martingale.*

PROOF: Easy exercise. Use the fact that if $Z \sim N(0, 1)$, then $E\{e^{\theta Z}\} = e^{\theta^2/2}$ for real θ ; this is the moment generating function of the $N(0, 1)$ distribution. \square

We can use this to form a martingale out of our process $X(t) = W(t) - \mu t$, which has drift $-\mu$: since $X(t) + \mu t$ is a standard Brownian motion, for every λ

$$(5.22) \quad M(t) := \exp\{\lambda[X(t) + \mu t] - \frac{1}{2}\lambda^2 t\} = \exp\{\lambda X(t) + \lambda(\mu - \frac{1}{2}\lambda)t\}$$

is a martingale. The nice thing is that since this holds for every λ , we are free to choose any λ that we like. There is a clear choice here that appears to simplify things: if we take $\lambda = 2\mu$, we see that

$$M(t) := e^{2\mu X(t)} \text{ is a martingale.}$$

Retaining the notation $T = \min\{\tau_a, \tau_b\}$ from before, in accordance with the optional sampling ideas we discussed above, we would like to say that

$$E\{e^{2\mu X(T)}\} = EM(T) = M(0) = 1.$$

Is this right? Well, clearly T is a stopping time; that's good. However, T is not bounded; that might be bad. Here's the trick. For any number n , the random time $T \wedge n$ is a bounded

stopping time, so that we can say that $E\{e^{2\mu X(T\wedge n)}\} = 1$. So clearly we would like to take a limit as $n \rightarrow \infty$ as follows

$$\begin{aligned} E\{e^{2\mu X(T)}\} &= E\left\{\lim_{n \rightarrow \infty} e^{2\mu X(T\wedge n)}\right\} \\ &= \lim_{n \rightarrow \infty} E\left\{e^{2\mu X(T\wedge n)}\right\} \\ &= \lim_{n \rightarrow \infty} 1 = 1. \end{aligned}$$

This interchange of limit and expectation is permitted by the bounded convergence theorem in our case, since the fact that $X(T\wedge n)$ must be between a and b implies that the random variables $e^{2\mu X(T\wedge n)}$ are bounded \llbracket between $e^{2\mu a}$ and $e^{2\mu b}\rrbracket$.

▷ *Strictly speaking, in order to apply bounded convergence this way, we should show that T is finite with probability one. Can you do that? This is Exercise [\[5.20\]](#)*

We are almost done. Write

$$1 = E\{e^{2\mu X(T)}\} = e^{2\mu a} P\{X(T) = a\} + e^{2\mu b} P\{X(T) = b\}.$$

Noting that $P\{X(T) = a\} = 1 - P\{X(T) = b\}$, this becomes a simple equation for the desired probability $P\{X(T) = b\}$, and the answer is the same as before.

Let's end this section with more food for thought. If you really think about the optional sampling result, from a certain point of view it is exceedingly strange. Here is what I mean. We know that if T is any bounded stopping time and W is a standard Brownian motion, then $E[W(T)] = 0$. In gambling terms, you cannot stop a Brownian motion before time 1, for example, and make a profit — i.e. end up with a positive expected amount of money. However, we also know that with probability 1, a Brownian motion *must* become positive before time 1. In fact, in each interval of the form $(0, \epsilon)$, W hits 0 infinitely often and is positive infinitely often and is negative infinitely often. Not only that, but since W has continuous paths, whenever W is positive at any point in time, in fact there is a whole *interval* of times over which W is positive. Imagine yourself “riding along” on the bumpy graph of a Brownian motion. We have just noted that, with probability 1, before time 1 you will ride through many intervals of times on which the Brownian motion is positive, that is, the graph lies above the time axis. As you ride, can't you just look down, see the time axis below us, think “Good, I'm positive now,” and say “Stop”? It doesn't seem hard, does it? If you are a microscopic rider on the graph of a Brownian path, there will be all these stretches of time over which the Brownian motion is positive. You don't have to be greedy; just choose *any* of those times and say “Stop.” Isn't it clear that you can do this with probability 1? If so, your winnings will be positive with probability 1, and so obviously your expected winnings are positive. But the optional sampling theorem implies that there is no such stopping time!

5.8 Some confusing questions (or answers)

I suspect the effect of this section may be to toggle your state of confusion about some issues — if you were not already confused and searching for answers, this section may confuse

you, whereas if you have been wondering about these issues yourself, I hope this section will help you sort them out. So, have you ever wondered about questions of this sort?

- (1) WHAT IS THE Ω FOR A BROWNIAN MOTION?
- (2) WHAT DOES IT MEAN TO “FIX ω ” FOR A POLLEN PARTICLE IN WATER?

I’ve been asked questions like these a number of times by a number of students, and they are good questions (perhaps a little too good... grumble grumble...).

The answer to (1) is: the set Ω is unimaginably complicated. Or not. Actually, there are many ways to answer this sort of question. This shouldn’t be surprising. For example, there are many different random variables that have (or are commonly modeled as having) a $N(0, 1)$ distribution; for example, (women’s height – 65 inches)/3, (IQ – 100)/10, and so on. Just as there are many different $N(0, 1)$ random variables, there are many different Brownian motions.

Let’s start by reviewing these ideas in a context much simpler than Brownian motion: tossing a coin, just once. We want to define a random variable X that models a coin toss, according to the mathematical framework of probability. That is, we need a probability space (Ω, \mathcal{F}, P) and then we define X as a function $X : \Omega \rightarrow \mathbb{R}$. [Recall Ω is called the sample space, and P is a probability measure on Ω . \mathcal{F} is a collection of subsets of Ω called events.]

A standard description of the concept of a sample space Ω is that “ Ω is the set of all possible outcomes of the experiment under consideration.” Here that would be $\Omega = \{H, T\}$. So defining the probability measure P by $P\{H\} = P\{T\} = 1/2$ and the random variable X by $X(\omega) = \omega$ for $\omega \in \Omega$, we have a model. Notice that the random variable X here is rather trivial—the identity function. Given this generic choice for X , we have customized the probability P to model the phenomenon.

Here is another way to model a coin toss. Imagine simulating the toss using a uniformly distributed random number: take $\Omega = [0, 1]$, the unit interval, and P =Lebesgue measure (ordinary length) on Ω . Then we could define X by $X(\omega) = H$ if $\omega \leq 1/2$ and $X(\omega) = T$ if $\omega > 1/2$. Notice that this is an entirely different random variable than the X defined in the previous paragraph: they are different functions, with different domains! However, the two random variables have the same probability distribution: each satisfies $P\{\omega \in \Omega : X(\omega) = H\} = P\{\omega \in \Omega : X(\omega) = T\} = 1/2$. In this second setup we have used a generic sort of source of randomness: a uniformly distributed ω . So P was not tailored to our application here; it is the random variable X that was carefully defined to give the desired distribution. Notice the contrast with the last sentences of the previous paragraph.

Now for a more physical picture. Imagine a person actually tossing a coin and letting it fall to the floor. What is it about the randomness in this situation that we would need to specify in order to know the outcome? This motivates a description like $\Omega = \{\text{all possible initial conditions, i.e., all possible values for (initial position, initial velocity, initial angular velocity)}\}$. Here P would be the probability distribution over Ω that describes the way that our flipper will “choose” initial conditions. And X is a complicated function that you could in principle write down from the laws of physics [good luck] that tells us, for each possible initial condition, whether the toss will be heads or tails.

What does it mean to fix ω ? For each fixed ω , the outcome $X(\omega)$ is determined—recall that X is simply a function of ω . So having fixed ω , there is no randomness left in $X(\omega)$. The “randomness” is all in the choice of ω ; this applies to each of the three descriptions above.

Let’s stop modeling a coin toss now; you can think of other probability spaces and random variables that we could use to do it. Now to Brownian motion: each of the 3 ways of thinking described above has a natural analog here. The first approach had Ω as the set of all possible outcomes. We know that the “outcome” of Brownian motion is a continuous path. So we could take Ω to be the set of all continuous functions

$$\Omega = C[0, \infty) = \{\omega : \omega(\cdot) \text{ is a continuous function on } [0, \infty)\}.$$

Then P would be the probability measure on $C[0, \infty)$ that corresponds to Brownian motion; this is called *Wiener measure*. It is rather complicated conceptually; an example of a simple statement that one could make about Wiener measure P is

$$P\{\omega \in C[0, \infty) : \omega(1) < 2\} = \Phi(2)$$

where Φ is the standard normal cumulative distribution function. The definition of Brownian motion W in terms of this sample space is trivially simple, just as the definition of X was trivial in the first example above; just define $W(\omega) = \omega$. That is, for each t , we define $W_t(\omega) = \omega(t)$. So simple it looks like gibberish. [Remember, ω is a function in its own right; it is an element of $C[0, \infty)$.] The function W is trivial; the interesting part is the Wiener measure.

A second approach uses a simpler measure and a more complicated function. The question comes down to: how could you simulate a realization (that is, a path) of Brownian motion? We have seen (from our “construction” of Brownian motion) that it can be done from an independent sequence of uniformly distributed random variables. But in fact a whole sequence of such random variables can be produced from a single uniformly distributed random variable. [How?] So Brownian motion can be defined on the nice friendly probability space $[0, 1]$ with Lebesgue measure. The tradeoff for the simple probability space is that the function W must then be more complicated — we can’t get by with the trivial identity function any more! This function must perform the tasks of producing a sequence of *iid* uniforms, transforming them to a sequence of *iid* $N(0, 1)$ ’s, then combining them into the series of functions discussed in the construction of Brownian motion, all starting from a single uniformly distributed random variable.

Finally, one could imagine modeling an actual pollen particle in a drop of water. We could define Ω to be the set of all possible initial conditions of the positions and velocities of the pollen particle and of all of the molecules of water in the drop. Wow. Then P would be our probability distribution for such initial conditions. The function W would again be determined [in principle!] from the laws of physics, with $W_t(\omega)$ giving the position of the pollen at time t if the initial conditions were ω .

Here are some words. A function $X : \Omega \rightarrow \mathbb{R}$ is usually called a random variable. We can consider functions from Ω to a more general space \mathcal{X} , say. Such a function $X : \Omega \rightarrow \mathcal{X}$ would be called a *random variable taking values in \mathcal{X}* or a *random element of \mathcal{X}* . We can

consider a stochastic process such as W to be a random element of a function space such as $C[0, \infty)$, since, for each ω , we get a whole path (i.e. continuous function) $W(\omega)$.

I hope you are not put off by all these different modeling approaches. Thankfully, as we have seen, most of the time, we don't have to agonize over just what Ω we have in our minds and how complicated or trivial the corresponding random element is. This works because usually we are interested in probabilities that the random element is in various sets, and these depend only on the probability distribution of the random element. It is well to keep in mind that this is all a game going on inside our heads. Of course, the same statement applies to all probability and statistics, and indeed to all of mathematics (and perhaps much more...). There are no points or lines in the physical universe, yet the mental game of geometry can be useful.

5.9 Exercises

▷ In the exercises of this chapter, W will denote a standard Brownian motion process.

[5.1] [[A sojourn time problem]] Let $X(t) = \mu t + \sigma W(t)$, where W is a standard Brownian motion and $\mu > 0$.

(i) For $\delta > 0$, find the expected total amount of time that the process X spends in the interval $(0, \delta)$. That is, defining

$$T = \int_0^\infty I\{X(t) \in (0, \delta)\} dt,$$

what is $E(T)$? A rather involved calculus calculation should eventually arrive at a strikingly simple answer.

(ii) Can you give a convincing but calculation-free argument why the simple answer is correct?

[5.2] [[Another sojourn time problem]] As in the previous problem, let $X(t) = \mu t + \sigma W(t)$, where W is a standard Brownian motion and $\mu > 0$. What is the expected amount of time that the process X spends below 0? [[The calculus is easier in this problem than in the previous one.]]

[5.3] For $0 < a < b$, calculate the conditional probability $P\{W_b > 0 \mid W_a > 0\}$.

[5.4] Prove the Brownian scaling property, Theorem (5.4).

[5.5] Imagine that you do not already know the answer to Exercise [5.2]; you know only that the answer is some function of μ and σ^2 . Use Brownian scaling to argue without calculus that the desired function must be of the form $a\sigma^2/\mu^2$ for some number a .

[5.6] Prove Proposition (5.5).

[5.7] Let $\{W_t : t \in [0, 1]\}$ and $\{X_t : t \in [0, 1]\}$ be standard Brownian motions that are independent of each other. Show that with probability 1, there are infinitely many times t in $[0, 1]$ such that $W_t = X_t$.

[5.8] The *strong Markov property* is an extension of the restarting property of Proposition 5.5 from fixed times c to random *stopping times* γ : For a stopping time γ , the process x defined by $X(t) = W(\gamma + t) - W(\gamma)$ is a Brownian motion, independent of the path of W up to time γ . Explain the role of the stopping time requirement by explaining how the restarting property can fail for a random time that isn't a stopping time. For example, let $M = \max\{B_t : 0 \leq t \leq 1\}$ and let $\beta = \inf\{t : B_t = M\}$; this is the first time at which B achieves its maximum height over the time interval $[0, 1]$. Clearly β is not a stopping time, since we must look at the whole path $\{B_t : 0 \leq t \leq 1\}$ to determine when the maximum is attained. Argue that the restarted process $X(t) = W(\beta + t) - W(\beta)$ is not a standard Brownian motion.

[5.9] [[Ornstein-Uhlenbeck process]] Define a process X by

$$X(t) = e^{-t}W(e^{2t})$$

for $t \geq 0$, where W is a standard Brownian motion. X is called an *Ornstein-Uhlenbeck process*.

- (a) Find the covariance function of X .
- (b) Evaluate the functions μ and σ^2 , defined by

$$\begin{aligned}\mu(x, t) &= \lim_{h \downarrow 0} \frac{1}{h} E[X(t+h) - X(t) \mid X(t) = x] \\ \sigma^2(x, t) &= \lim_{h \downarrow 0} \frac{1}{h} \text{Var}[X(t+h) - X(t) \mid X(t) = x].\end{aligned}$$

[5.10] Let W be a standard Brownian motion.

- (i) Defining $\tau_b = \inf\{t : W(t) = b\}$ for $b > 0$ as above, show that τ_b has probability density function

$$f_{\tau_b}(t) = \frac{b}{\sqrt{2\pi}} t^{-3/2} e^{-b^2/(2t)}$$

for $t > 0$.

- (ii) Show that for $0 < t_0 < t_1$,

$$P\{W(t) = 0 \text{ for some } t \in (t_0, t_1)\} = \frac{2}{\pi} \tan^{-1} \left(\sqrt{\frac{t_1}{t_0} - 1} \right) = \frac{2}{\pi} \cos^{-1} \left(\sqrt{\frac{t_0}{t_1}} \right).$$

[[Hint: The last equality is simple trigonometry. For the previous equality, condition on the value of $W(t_0)$, use part (i), and Fubini (or perhaps integration by parts).]]

(iii) Let $L = \sup\{t \in [0, 1] : W_t = 0\}$ be the last zero of W on $[0, 1]$. Find and plot the probability density function of L . Rather peculiar, wouldn't you say?

[5.11] Let $X_t = W_t + \mu t$ be a Brownian motion with drift μ , and let $\epsilon > 0$. Show that

$$P\{\max_{0 \leq t \leq h} |X_t| > \epsilon\} = o(h) \text{ as } h \downarrow 0.$$

[[That is,

$$\frac{1}{h} P\{\max_{0 \leq t \leq h} |X_t| > \epsilon\} \rightarrow 0 \text{ as } h \downarrow 0.$$

Hint: You might want to do the special case $\mu = 0$ first. For the general case, you can transform the problem about X into an equivalent problem about W , and use the special case to do an easy bound. This shouldn't be a very involved calculation.]] This result is useful when we are calculating something and keeping terms up to order h , and we want to show that the probability of escaping from a strip can be neglected.

[5.12] Let $M(t) = \max\{W(s) : 0 \leq s \leq t\}$. Find the joint probability density function of $M(t)$ and $W(t)$.

[5.13] Let $(X(t), Y(t))$ be a two-dimensional standard Brownian motion; that is, let $\{X(t)\}$ and $\{Y(t)\}$ be standard Brownian motion processes that are independent of each other. Let $b > 0$, and define $\tau = \inf\{t : X(t) = b\}$. Find the probability density function of $Y(\tau)$. That is, find the probability density of the height at which the two-dimensional Brownian motion first hits the vertical line $x = b$.

[[Hint: The answer is a Cauchy distribution.]]

[5.14] Show that for $0 \leq s < t < u$,

$$\mathcal{L}(W_t \mid W_s, W_u) = N\left(W_s + \frac{t-s}{u-s}(W_u - W_s), \frac{(t-s)(u-t)}{u-s}\right).$$

[[Hint: This may be obtained by applying the result of (5.9) to the Brownian motion \tilde{W} defined by $\tilde{W}(v) = W(s+v) - W(s)$ for $v \geq 0$.]]

[5.15] Let $0 < s < t < u$.

(a) Show that $E(W_s W_t \mid W_u) = \frac{s}{t} E(W_t^2 \mid W_u)$.

(b) Find $E(W_t^2 \mid W_u)$ [[you know $\text{Var}(W_t \mid W_u)$ and $E(W_t \mid W_u)$!]] and use this to show that

$$\text{Cov}(W_s, W_t \mid W_u) = \frac{s(u-t)}{u}.$$

[5.16] What is the conditional distribution of the vector (W_5, W_9, W_{12}) given that $W_{10} = 3$? [[It is a joint normal distribution; you should give the (3×1) mean vector and (3×3) covariance matrix.]]

- [5.17] Verify that the definitions (5.13) and (5.14) give Brownian bridges.
- [5.18] We defined the Brownian bridge by conditioning a standard Brownian motion to be 0 at time 1. Show that we obtain the same Brownian bridge process if we start with a $(\mu, 1)$ Brownian motion and condition it to be 0 at time 1.
- [5.19] Let $X(t) = x_0 + \mu t + \sigma W(t)$ be a (μ, σ^2) Brownian motion starting from x_0 at time 0. What is the probability

$$P\{X(s) \geq b + cs \text{ for some } s \leq t \mid X(t) = x_t\}?$$

[[Your answer should be a function of μ , σ , x_0 , x_t , b , c , and t (maybe not depending on all of these). This should not require significant calculation, but rather a reduction to something we have done.]]

- [5.20] As in Section 5.7, let T denote $\min\{\tau_a, \tau_b\}$, where $a < 0 < b$. Show that $P\{T < \infty\} = 1$.
- [5.21] The exponential martingale [[see (5.21) and (5.22)]] is quite useful. For example, let $\mu > 0$ and $b > 0$ and consider $X(t)$ to be BM with drift μ . Find the moment generating function $E(e^{\theta \tau_b})$ of τ_b , as a function of θ . Note that the answer is finite for some values of θ and infinite for others. Can you solve the same problem for $T = \tau_a \wedge \tau_b$?

6. Diffusions and Stochastic Calculus

We have discussed Markov processes in discrete time with discrete state space—these are discrete time Markov chains. We have also studied Brownian motion, a special Markov process in continuous time that has continuous sample paths. In this chapter, we will study more general continuous-time Markov processes that have continuous sample paths. These are called *diffusions*.

There is a whole family of Brownian motions: for each μ and σ^2 there is a (μ, σ) Brownian motion. However, in a sense these are all the same—all are obtained from the standard Brownian motion by multiplying by a constant and adding on a deterministic linear function. This is analogous to the way that all normal random variables are obtained from a standard normal random variable simply by adding and multiplying by constants. Thus, the family of Brownian motions has limited flexibility in modeling and describing random phenomena. For example, unless your random process is expected to follow a linear trend, you will be frustrated trying to fit a (μ, σ^2) Brownian motion to it; you will not be able to get it to do what you want. We have seen some other sorts of processes built out of Brownian motion, such as the Brownian bridge and geometric Brownian motion, for example. These processes are simple examples of diffusions, which is a much more general class of processes that can exhibit a variety of more interesting behaviors, and are therefore flexible and useful tools in stochastic modeling and analysis.

Diffusions are built up, in a sense, from Brownian motions. They are built up from Brownian motions in the same way as in ordinary calculus, general differentiable functions are built up out of linear functions. Think about the functions of the form $x(t) = Ce^{-t}$, for example. These functions satisfy the differential equation $x'(t) = -x(t)$. What is this equation telling us? Suppose you wanted to graph the function that satisfies the equation $x'(t) = -x(t)$ and the initial condition $x(0) = 3$. Let's pretend that you never heard of exponential functions; all you know is that you want to graph the function—whatever it may be—that satisfies the given differential equation and initial condition. So you start with the point you know: $x(0) = 3$. Then the differential equation tells you that the slope of the solution curve at that point is $x'(0) = -x(0) = -3$. In other words, the equation is telling you that the desired curve is closely approximated by the straight line $3 - 3t$ over a short time interval $[0, \epsilon]$, say. Drawing that linear function $3 - 3t$ as an approximation over the interval $[0, \epsilon]$ brings us to the new point $(\epsilon, 3 - 3\epsilon)$. Note that this point is wrong; in fact, we are wrong as soon as we leave our initial point, since as $x(t)$ leaves the initial value 3, we should no longer be using the initial slope -3 ! However, since the slope change little over tiny intervals, we are nearly right. So we have gotten to the point $(\epsilon, 3 - 3\epsilon)$. Now we would use the differential equation to re-evaluate the slope we should use: the slope is now $-3 + 3\epsilon$. Then we can move along a line with the new slope over the time interval $[\epsilon, 2\epsilon]$. Then we would re-evaluate the slope using the differential equation, and so on.

If we use a small enough ϵ , we get an arbitrarily good approximation to the true solution

$x(t) = 3e^{-t}$ of the differential equation this way. Notice that the true solution curve has no straight lines in it, but it is arbitrarily well approximated by a piecewise linear function. In a sense, we might say that the function $x(t) = 3e^{-t}$ is *always* behaving according to a linear function; it is just continually adjusting its slope according to the recipe specified by the differential equation: slope = −current value. The differential equation is telling us, at each time, the slope of the line we should be following. The equation expresses the current slope as a function of the current state.

Diffusions are the stochastic analog of such solutions of differential equations. They are built up out of the family of Brownian motions in the same spirit as the way in which functions like $3e^{-t}$ are built up from the family of linear functions. In the same loose sense as above, a diffusion is always behaving according to a Brownian motion; it is just continually re-adjusting its drift μ and variance σ^2 parameters according to its current state. We specify a diffusion by giving a rule for determining what Brownian motion we should be following, as a function of the current state. That is, we specify two functions $\mu(\cdot)$ and $\sigma^2(\cdot)$; if the current state is X_t , we should be running a $(\mu(X_t), \sigma^2(X_t))$ -Brownian motion.

We can “solve,” or simulate, a stochastic differential equation on a computer using the same idea as in solving a deterministic differential equation on a computer. Of course, we can’t simulate a diffusion perfectly, just as we cannot simulate a Brownian motion perfectly. [Recall that the “construction” of Brownian motion on $[0,1]$ that we described would take an infinite amount of time to “finish”.] However, as with Brownian motion, we can get an arbitrarily good approximate simulation by working with a fine grid on the time axis, say $0, h, 2h, \dots$. Start by simulating X_0 from the initial distribution. Let Z_1, Z_2, \dots denote a sequence of *iid* $N(0, 1)$ random variables; we know how to simulate these. We can use these normal random variables to simulate the process X on our grid as follows: take

$$X_h = X_0 + \mu(X_0)h + \sigma(X_0)\sqrt{h}Z_1,$$

then

$$X_{2h} = X_h + \mu(X_h)h + \sigma(X_h)\sqrt{h}Z_2,$$

and so on. This is a natural idea: instead of doing the ideal of “continually readjusting the drift and variance functions at each instant,” we are content to adjust the drift and variance only at the times $0, h, 2h, \dots$, keeping their values constant between adjustment times. The approximate diffusion consists of piecing together the resulting little Brownian motions run over those tiny time intervals. Again, although the actual diffusion will generally have no time intervals on which it is truly a pure Brownian motion, since it is continually readjusting its drift and variance, if we take h sufficiently small the resulting simulation is a good approximation to the real thing. This is like the situation with deterministic differential equations: on a computer, we approximate a solution—which is a curve—by a piecewise linear function.

6.1 Specifying a diffusion

Let’s state a common definition of a diffusion process, at least so we can say we did it.

(6.1) DEFINITION. A stochastic process that has the strong Markov property and (almost surely) continuous sample paths is called a **diffusion process**.

I do not want to emphasize this definition, which, incidentally, is not universally agreed upon for various subtle reasons. We will ignore those subtleties. Just think of a diffusion as a continuous-time Markov process that has continuous paths; you won't go far wrong.

Let us take a more pragmatic or operational point of view. As we did with Markov chains, we'll start by saying how to specify a diffusion. That is, if I am thinking of a diffusion, how do I tell you which particular diffusion $\{X_t : t \geq 0\}$ I am thinking about? Just as with Markov chains, to specify a diffusion, we need to specify a state space, an initial distribution, and a probability transition structure. The state space will be an interval I of the form (l, r) , $[l, r]$, $[l, r)$, or $(l, r]$, where $-\infty \leq l < r \leq \infty$. The initial distribution, of course, is just a probability distribution on I . Now to describe the probability transition structure: this is not easy to spell out rigorously in a few words, but the intuitive idea is not hard. Again thinking in terms of simulation, suppose that we have already simulated the process X up to time t and we are currently at the position $X(t) = x$, where x is some point in the interior of the state space. How do we simulate the process over the next tiny time interval $[t, t + h]$? The answer is (approximately): we simulate a certain (μ, σ^2) -Brownian motion. Which μ and σ^2 we use are determined by the current position x ; that is, μ and σ^2 are functions of x .^{*} Thus, an intuitive description of the way we run the process is: at each time t , we check to see where we are, calling our current state X_t . Then we evaluate the two functions μ and σ^2 at X_t , and run a $(\mu(X_t), \sigma^2(X_t))$ -Brownian motion for a tiny amount of time. We are always continually checking where we are, and adjusting the drift and variance of our Brownian motion accordingly.

Thus, in the interior (l, r) of the state space, the probability transition structure of a time-homogeneous diffusion is specified by two functions $\mu = \mu(x)$ and $\sigma^2 = \sigma^2(x)$, which satisfy the relations

$$(6.2) \quad E[X(t+h) - X(t) \mid X(t) = x] = \mu(x)h + o(h)$$

and

$$(6.3) \quad \text{Var}[X(t+h) - X(t) \mid X(t) = x] = \sigma^2(x)h + o(h)$$

as $h \downarrow 0$.

(6.4) EXERCISE. Show that in the presence of (6.2), the condition (6.3) is equivalent to

$$E[(X(t+h) - X(t))^2 \mid X(t) = x] = \sigma^2(x)h + o(h).$$

Terms like “infinitesimal drift” or “infinitesimal mean function” are used for $\mu(\cdot)$, and $\sigma^2(\cdot)$ has names like “infinitesimal variance” and “diffusion function.”

Let us also assume that moments of the increments $X(t+h) - X(t)$ higher than the second moment are negligible when compared with the first and second moments:

$$E[|X(t+h) - X(t)|^p \mid X(t) = x] = o(h) \text{ as } h \rightarrow 0$$

^{*}This is for time-homogeneous diffusions. For general diffusions, μ and σ^2 may depend on t as well as x .

for all $p > 2$.

So far we have discussed the probability transition structure of our diffusion only in the interior (l, r) of I . The behavior at the boundary points must be specified separately. There are many possible types of boundary behaviors: absorbing boundaries, reflecting boundaries, “sticky” boundaries, and others. We might talk about this more later.

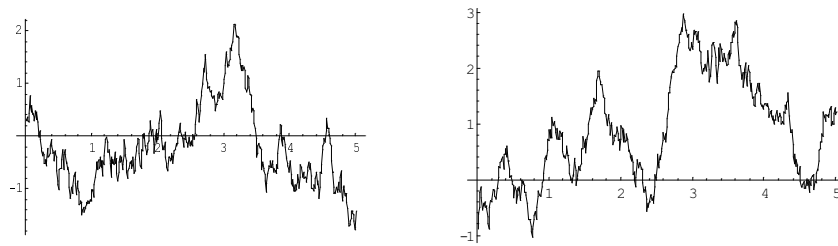
Going back to the infinitesimal mean and variance functions, another interesting point to appreciate is that it is not obvious that we have really specified a probability transition structure just by giving functions $\mu(\cdot)$, and $\sigma^2(\cdot)$ —these just describe the behaviors of the first and second moments of increments of the process, not the whole distribution. However, in fact it is generally true that these two functions are enough. We will see how the whole probability transition structure is determined by these two functions when we discuss Kolmogorov’s forward and backward equations.

(6.5) EXAMPLE [GEOMETRIC BROWNIAN MOTION]. Suppose $\{X_t\}$ is a (μ, σ^2) Brownian motion, and define $Y_t = e^{X_t}$. Then $\{Y_t\}$ is called a *geometric Brownian motion*. A typical application of this process is as a model of stock prices, where it seems reasonable to assume that “returns,” which are *ratios* like $Y(t_2)/Y(t_1)$ and $Y(t_4)/Y(t_3)$, are independent for nonoverlapping intervals (t_1, t_2) and (t_3, t_4) . \square

(6.6) EXAMPLE [ORNSTEIN-UHLENBECK PROCESS]. A *standard Ornstein-Uhlenbeck process* is a diffusion having $\mu(x) = -x$ and $\sigma^2(x) = 2$ on the state space $I = (-\infty, \infty)$. If the initial distribution is taken to be $N(0, 1)$, then the resulting process is in fact stationary. This is the reason for the modifier *standard*—the process has the standard normal distribution $N(0, 1)$ as its stationary distribution. Here the $\sigma^2(\cdot)$ is not very interesting; the process has constant wiggleness. But look at the drift function $\mu(x) = -x$. The process is always drifting toward 0: for example if $X_t = 1.5$ then the drift at time t is -1.5 , while if $X_t = -3$ the drift at time t is 3. If $X_t = 0$ the drift at time t is 0. Thus, the process is happy when it is at 0 (i.e. it just wanders around aimlessly not particularly going in either direction). If it is near to 0, then it has a slight tendency to move back toward 0, while if it is far away from 0 then there is a large drift pulling it strongly back toward 0. It is as if there is a “restoring force” always pulling the process back toward 0, with the force being stronger the further the process is away from 0.

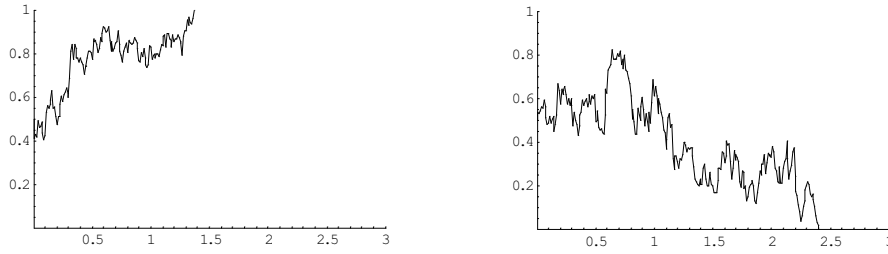
Thus, there is a continual struggle between the drift part (which wants to pull the process back toward 0) and the diffusion part (which wants the process to wander randomly). The result is a wiggly process that tends to hover near to 0. It will occasionally by chance make larger excursions away from 0, but then the drift will pull it quickly back to a vicinity of 0.

The Ornstein-Uhlenbeck process, sometimes called a *mean-reverting* process, is often used as a model for stochastic systems that have an equilibrium state and a spring-like force that tends to move the process back toward the equilibrium state. It approximates many systems of this type. A simple example is the Ehrenfest chain. Recall there that we had $2d$ balls, say, in two urns, and at each time we choose at ball at random and move it to the other urn. This process tends to move toward the equilibrium state of d balls in both urns. The tendency to move toward equilibrium is stronger the further the system is away from equilibrium.

(6.7) FIGURE. *Two simulated Ornstein-Uhlenbeck sample paths.*

(6.8) EXAMPLE [WRIGHT-FISHER PROCESS]. This is a diffusion on the state space $I = [0, 1]$ having drift and variance functions $\mu(x) = 0$ and $\sigma^2(x) = x(1-x)$. The states 0 and 1 are absorbing. There is no drift, so the process never has any systematic tendency to go up or down. The interesting part here is the variance function: $\sigma^2(x)$ is maximized for $x = 1/2$ and approaches 0 as $x \rightarrow 0$ or $x \rightarrow 1$. What this means is that, if we watch a sample path, then it is “more wiggly” when x is nearly $1/2$, and it is less wiggly as x approaches either endpoint of I .

Does the process actually get absorbed? If you consider this question, you may find yourself thinking Zeno paradox-like thoughts: In order to get to the boundary, the process has to get close to the boundary first. As the process gets nearer and nearer to the boundary, the form of the σ^2 function causes it to move around less and less. Maybe it never quite gets all the way to the boundary... In fact, we’ll see that the process does reach the boundary, and we’ll even find the expected time to absorption as a function of the starting state.



(6.9) FIGURE. Two simulated Wright-Fisher sample paths.



6.2 A calculation with diffusions

Several interesting quantities associated with diffusions can be calculated by solving a differential equation, and, in fact, this can often be done explicitly.

For example, here's a problem. Consider a diffusion on the state space I , having infinitesimal parameters $\mu = \mu(x)$ and $\sigma^2 = \sigma^2(x)$. Note that we are considering a time-homogeneous diffusion: the functions μ and σ^2 depend only on x , and not on t . Choose two points a and b in the interior of I . Consider starting the diffusion $\{X(t)\}$ off at some $x \in (a, b)$ at time 0 and letting the diffusion run along until the random time T at which it first attains the value a or the value b . Let $g(\cdot)$ be a “cost function” that gives the rate $g(x)$ at which “cost” accrues per unit time when the diffusion X is in the state x , so that the total cost of running the diffusion until time T is the random variable $\int_0^T g(X_t) dt$. The question we are asking is: what is the expected cost $E^x[\int_0^T g(X_t) dt]$? [Notation: here the superscript “ x ” on the “ E ” indicates that we are assuming that $X_0 = x$.] For example, a simple but important special case to consider is when the function g is the constant $g(x) \equiv 1$, in which case $\int_0^T g(X_t) dt = T$, so that our problem is to find $E^x(T)$.

Defining $w(x) = E^x[\int_0^T g(X_t) dt]$, our question is answered by the following result.

(6.10) CLAIM. The function w satisfies the differential equation

$$(6.11) \quad \mu(x)w'(x) + \frac{1}{2}\sigma^2(x)w''(x) = -g(x)$$

with the boundary conditions $w(a) = w(b) = 0$.

You may be getting the idea that the whole area of diffusion processes is intimately connected with differential equations; this is another example of that connection. We would expect the solution of the second-order equation (6.11) to involve two arbitrary constants; presumably the two boundary conditions $w(a) = 0$ and $w(b) = 0$ will allow us to specify the values of these constants.

Let's do a very simple example before justifying the claim. This is a problem that we have done before. Consider a standard Brownian motion [which we have agreed starts at

0], let $a < 0 < b$, and as above let $T = \inf\{t : W_t = a \text{ or } W_t = b\}$. Our question is: what is $E(T)$? The idea we use is familiar by now: to solve our particular given problem, we imbed it in a whole family of similar problems and solve them all at once. Define $w(x) = E^x(T) = E[T \mid W(0) = x]$. As mentioned above, the appropriate cost function here is simply $g(x) = 1$ for all x . Also, $\mu(x) \equiv 0$ and $\sigma^2(x) \equiv 1$. So (6.11) becomes $\frac{1}{2}w''(x) = -1$, which has general solution $w(x) = -x^2 + c_1x + c_2$. Thus, clearly the solution satisfying the boundary conditions $w(a) = 0 = w(b)$ is $w(x) = (x-a)(b-x)$. This is a nice, handy result. It's interesting to stick in a couple of numbers. For example, if $-a = b = 100$ then the Brownian motion takes a long time $E(T) = 10,000$ on the average to escape the interval (a, b) . On the other hand, if $-a = b = .01$ then the Brownian motion just shoots out in the tiny average time $E(T) = .0001$! This is another reflection of the fact that Brownian paths are very "steep" over short intervals of time.

Now in support of the claim, let us see where (6.11) comes from. [The boundary conditions seem pretty clear.] Here is a heuristic derivation. Begin with

$$(6.12) \quad \begin{aligned} w(x) &= E^x \left\{ \int_0^h g(X_t) dt + \int_h^T g(X_t) dt \right\} \\ &= hg(x) + E^x \left\{ \int_h^T g(X_t) dt \right\} + o(h). \end{aligned}$$

We are imagining that h is extremely tiny, so that we can be virtually certain that $T > h$. Using the basic property $EY = E\{E(Y \mid Z)\}$ of conditional expectation,

$$(6.13) \quad \begin{aligned} E^x \left\{ \int_h^T g(X_t) dt \right\} &= E^x \left[E^x \left\{ \int_h^T g(X_t) dt \mid X_h \right\} \right] \\ &= E^x \left[E^x \left\{ \int_h^T g(X_t) dt \mid X_h, T > h \right\} \right] + o(h) \\ &= E^x [w(X_h)] + o(h). \end{aligned}$$

The last equality follows from the Markov property: given that $X_h = x_h$ and $T > h$, the expected value of $\int_h^T g(X_t) dt$ is the same as the expected cost of running the diffusion until time T starting at time 0 in the state x_h , which is $w(x_h)$. However,

$$(6.14) \quad \begin{aligned} E^x [w(X_h)] &= E^x \left[w(x) + w'(x)(X_h - x) + \frac{1}{2}w''(x)(X_h - x)^2 \right] + o(h) \\ &= w(x) + w'(x)\mu(x)h + \frac{1}{2}w''(x)\sigma^2(x)h + o(h). \end{aligned}$$

Substituting (6.14) into (6.13), and then (6.13) into (6.12) gives

$$w(x) = hg(x) + w(x) + w'(x)\mu(x)h + \frac{1}{2}w''(x)\sigma^2(x)h + o(h),$$

or

$$\left[\mu(x)w'(x) + \frac{1}{2}\sigma^2(x)w''(x) + g(x) \right] h = o(h),$$

which implies that $\mu(x)w'(x) + \frac{1}{2}\sigma^2(x)w''(x) + g(x) = 0$, which is (6.11).

6.3 Infinitesimal parameters of a function of a diffusion

Here is an example of the kind of question we'll consider. Suppose $\{X_t\}$ is a Brownian motion having $\mu_X(x) \equiv \mu$ and $\sigma_X^2(x) \equiv \sigma^2$, say, and define the geometric Brownian motion $Y_t = e^{X_t}$. Our question is: What are the infinitesimal parameters $\mu_Y(\cdot)$ and $\sigma_Y^2(\cdot)$ of the Y process? This type of question is answered by the following result.

(6.15) PROPOSITION. *Suppose X is a $(\mu_X(\cdot), \sigma_X^2(\cdot))$ diffusion. Let f be a strictly monotone function that is “smooth enough” [e.g., twice continuously differentiable], and define $Y_t = f(X_t)$. Then Y is a diffusion having infinitesimal parameters*

$$\mu_Y(y) = \mu_X(x)f'(x) + \frac{1}{2}\sigma_X^2(x)f''(x)$$

and

$$\sigma_Y^2(y) = [f'(x)]^2\sigma_X^2(x),$$

where $x = f^{-1}(y)$.

Taylor expansions underlie just about everything we will do with diffusions. One example is the derivation of (6.11); see (6.14). Deriving the formulas in the above Proposition provides another example. In fact, we have

$$\begin{aligned} E[Y_{t+h} - Y_t \mid Y_t = y] &= E[f(X_{t+h}) - f(x) \mid X_t = x] \quad \text{recall } x = f^{-1}(y) \\ &= E[f'(x)(X_{t+h} - x) + \frac{1}{2}f''(x)(X_{t+h} - x)^2 \mid X_t = x] + o(h) \\ &= f'(x)\mu_X(x)h + \frac{1}{2}f''(x)\sigma_X^2(x)h + o(h), \end{aligned}$$

so that

$$\mu_Y(y) = \mu_X(x)f'(x) + \frac{1}{2}\sigma_X^2(x)f''(x).$$

Similarly,

$$\begin{aligned} E[(Y_{t+h} - Y_t)^2 \mid Y_t = y] &= E\{[f'(x)]^2[X_{t+h} - x]^2 \mid X_t = x\} + o(h) \\ &= [f'(x)]^2\sigma_X^2(x)h + o(h), \end{aligned}$$

so that

$$\sigma_Y^2(y) = [f'(x)]^2\sigma_X^2(x).$$

To return to the example of geometric Brownian motion, where $\mu_X(x) \equiv \mu$, $\sigma_X^2(x) \equiv \sigma^2$, and $y = e^x = f(x) = f'(x) = f''(x)$, we have

$$\mu_Y(y) = \mu y + \frac{1}{2}\sigma^2 y.$$

and

$$\sigma_Y^2(y) = y^2\sigma^2.$$

6.4 Kolmogorov's backward and forward equations

Our aim here is to investigate the probability transition structure of diffusions. Recall that for Markov chains, the probability transition rule was specified by giving a matrix P , whose job is to tell us how to calculate the distribution at time $n + 1$ in terms of the distribution at time n : just multiply $\pi(n + 1) = \pi(n)P$. This formula answers the question: given the distribution of the state now, how does one compute the distribution at the next time? In continuous time, there is no “next time” after t , but we can hope to say how to calculate the distribution of the process state a tiny infinitesimal time later, so to speak. That is, the probability transition rule will give the rate of change, per unit time, of the probability density function of the state. This is done by a partial differential equation. Why *partial*? An ordinary differential equation describes a function of time $f : \mathbb{R} \rightarrow \mathbb{R}$ by giving its rate of change $f'(t)$. The derivative says approximately how to compute the number $f(t + h)$ in terms of $f(t)$, if h is small. That is not what is happening in our case—here, at each time, we are concerned with a whole *function*, namely, the probability density function at time t . In other words, we are concerned with a function of two variables: $f(t, x)$ is the probability density at state x at time t . A partial differential equation describes such an f by giving the rate of change $(\partial/\partial t)f(t, x)$. Such an equation describes the time evolution of the function $f(t, \cdot)$ —for small h the time derivative $(\partial/\partial t)f(t, x)$ tells us approximately how to compute the function $f(t + h, \cdot)$ in terms of the function $f(t, \cdot)$. The partial differential equation gives this time derivative $(\partial/\partial t)f(t, x)$ in terms of the function $f(t, \cdot)$, and possibly its derivatives $(\partial/\partial x)f(t, x)$, $(\partial^2/\partial x^2)f(t, x)$, and so on.

Let X be a diffusion with infinitesimal mean and variance functions $\mu(\cdot)$ and $\sigma^2(\cdot)$. For the “backward” equation, fix a state y , and define the function $f(t, x)$ to be the density of X_t evaluated at y , given that $X_0 = x$; that is,

$$f(t, x) = f_{X_t}(y \mid X_0 = x).$$

Kolmogorov's backward equation says that

$$\partial_t f(t, x) = \mu(x)\partial_x f(t, x) + \frac{1}{2}\sigma^2(x)\partial_{xx} f(t, x).$$

For the “forward” equation, fix an initial probability density for X_0 , and define $g(t, y)$ to be the density of X_t evaluated at y . The forward equation describes the evolution of this density over time:

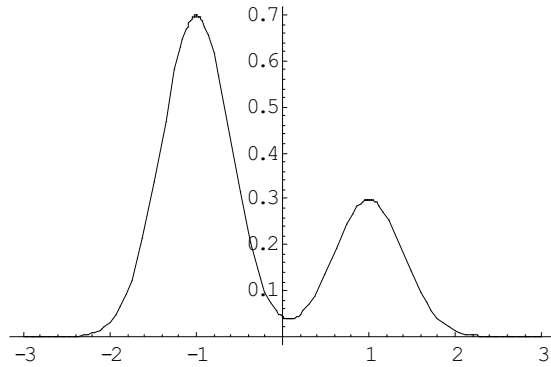
$$\partial_t g(t, y) = -\partial_y [\mu(y)g(t, y)] + \partial_{yy} \left[\frac{1}{2}\sigma^2(y)g(t, y) \right].$$

You can understand the names for these equations if you think of x as the “backward” variable and y as the “forward” variable. These names make sense since x describes the state of the process way back at time 0, while the variable y refers to the value of X_t .

As promised, these are partial differential equations, but don't let that scare you! For example, for driftless Brownian motion the forward equation becomes

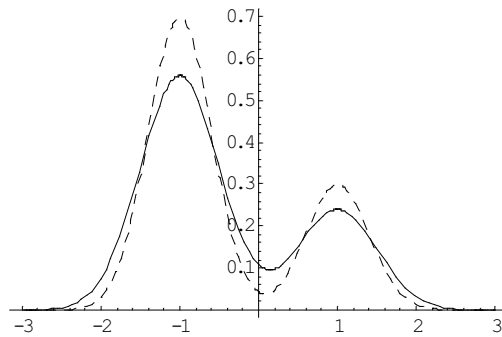
$$(6.16) \quad \partial_t g(t, y) = \frac{1}{2}\partial_{yy} g(t, y).$$

This equation gives a rather intuitive description of how the density evolves. For example, suppose that at time t the probability density $g(t, \cdot)$ happens to look like this:



(6.17) FIGURE. *Density at time t for a Brownian motion.*

[[Of course, this would not happen with standard Brownian motion started out at state 0 at time 0, but we could arrange it by starting the Brownian motion off with an initial density that is bimodal.]] What does the probability density $g(t+h, \cdot)$ look like a short time h later? Of course, if h is small we expect $g(t+h, \cdot)$ to look similar to $g(t, \cdot)$. But the density $g(t+h, \cdot)$ should be slightly more “spread out” than $g(t, \cdot)$. That is, over the time interval $[t, t+h]$, the peaks of the density $g(t, \cdot)$ should flatten out some by decreasing in height, while the valleys of $g(t, \cdot)$ should compensate by increasing in height. The peaks are characterized by a negative value of the second derivative $\partial_{yy}g(t, y)$. The equation (6.16) tells us that at such a point y , the time derivative $\partial_t g(t, y)$ will be negative, so that $g(t+h, y)$ will indeed be less than $g(t, y)$: the peaks decrease in height. On the other hand, for states y in a valley of $g(t, y)$, where the second derivative $\partial_{yy}g(t, y)$ is positive, (6.16) tells us that $\partial_t g(t, y)$ is positive, so that $g(t+h, y) > g(t, y)$. In the example pictured above, a short time later the density will have changed as shown in the next figure, in which $g(t, \cdot)$ and $g(t+h, \cdot)$ are plotted in dashed and solid lines, respectively.



(6.18) FIGURE. *Density at time $t+h$ for the same Brownian motion.*

Let us now derive the backward and forward equations. We'll begin with the backward equation, which will then be used in the derivation of the forward equation. The backward equation actually applies to more than just the transition density $f(x, t)$; it describes a more general class of functions that includes f . To work at this level of generality, let ρ be a real-valued function defined on the state space of the diffusion. Define $u(x, t) = E^x \rho(X_t)$;

for example, if ρ is interpreted as a “reward” function, then $u(x, t)$ is the conditional expected reward if we stop the diffusion at time t , given that it started in state x at time 0. We will derive a backward equation satisfied by u . We have

$$u(x, t + h) = E^x \rho(X_{t+h}) = E^x [E^x \rho(X_{t+h} | X_h)] = E^x u(X_h, t).$$

So

$$\begin{aligned} \partial_t u(x, t) &= \lim_{h \downarrow 0} \frac{1}{h} [E^x u(X_h, t) - u(x, t)] \\ &= \lim_{h \downarrow 0} \frac{1}{h} E^x \left[\partial_x u(x, t)(X_h - x) + \frac{1}{2} \partial_{xx} u(x, t)(X_h - x)^2 + o(h) \right] \\ &= \mu(x) \partial_x u(x, t) + \frac{1}{2} \sigma^2(x) \partial_{xx} u(x, t) \end{aligned}$$

This is the backward equation for u .

To specialize this to the transition density f , let us think of y as fixed but arbitrary for the moment, and choose the function ρ to be the indicator function $\rho(z) = 1\{z \leq y\}$. Then

$$u(x, t) = E^x \rho(X_t) = P^x \{X_t \leq y\} =: F(t, x, y),$$

so that we have shown that

$$\partial_t F(t, x, y) = \mu(x) \partial_x F(t, x, y) + \frac{1}{2} \sigma^2(x) \partial_{xx} F(t, x, y).$$

Taking ∂_y of both sides and noting that $\partial_y F(t, x, y)$ is the function $f(t, x)$ defined above gives the backward equation for f .

Next, we will use the backward equation in a derivation of the useful forward equation — “Life can only be understood backwards, but it has to be lived forwards,” according to Kierkegaard. Denote the density of X_t by $v(\cdot, t)$ for all $t \geq 0$. Our question is to find the equation governing evolution over time of the function v . Retain the definition of the function $u(y, s) = E^y \rho(X_s)$ from before, where ρ is some given function that we assume to be smooth [twice continuously differentiable say] with compact support [that is, $\rho(x) = 0$ for x outside some bounded interval]. Using the assumed time-homogeneity of the diffusion X , notice that

$$u(y, s) = E[\rho(X_{s+t}) | X_t = y].$$

So we have

$$\int u(y, s) v(y, t) dy = \int E[\rho(X_{s+t}) | X_t = y] P\{X_t \in dy\} = E\rho(X_{s+t}).$$

Thus, we have shown that $\int u(y, s) v(y, t) dy$ is a function of $(s + t)$, from which it follows that

$$\partial_s \left[\int u(y, s) v(y, t) dy \right] = \partial_t \left[\int u(y, s) v(y, t) dy \right],$$

or

$$(6.19) \quad \int [\partial_s u(y, s)] v(y, t) dy = \int u(y, s) [\partial_t v(y, t)] dy.$$

But we know what $\partial_s u(y, s)$ is, by the backward equation! So the left side of (6.19) is

$$(6.20) \quad \int \mu(y) [\partial_y u(y, s)] v(y, t) dy + \frac{1}{2} \int \sigma^2(y) [\partial_{yy} u(y, s)] v(y, t) dy.$$

Integration by parts in the first of the two integrals in (6.20) gives

$$\int \mu(y) [\partial_y u(y, s)] v(y, t) dy = - \int u(y, s) [\partial_y \mu(y) v(y, t)] dy;$$

we have not written the difference of the “boundary terms” $\lim_{y \rightarrow \pm\infty} u(y, s) \mu(y) v(y, t)$ since these two limits will both be zero under mild assumptions: in fact, $u(y, s)$ will approach zero because we have assumed ρ to have compact support, and we would also expect $v(y, s)$ to approach zero, as most good densities do. Similarly, integration by parts twice gives

$$\frac{1}{2} \int \sigma^2(y) [\partial_{yy} u(y, s)] v(y, t) dy = \frac{1}{2} \int u(y, s) [\partial_{yy} \sigma^2(y) v(y, t)] dy$$

for the second term in (6.20). Thus, by substituting the last two displays into (6.20) and then into (6.19),

$$\int u(y, s) \left\{ -\partial_y [\mu(y) v(y, t)] + \frac{1}{2} \partial_{yy} [\sigma^2(y) v(y, t)] \right\} dy = \int u(y, s) [\partial_t v(y, t)] dy,$$

which in turn becomes

$$\int \rho(y) \left\{ -\partial_y [\mu(y) v(y, t)] + \frac{1}{2} \partial_{yy} [\sigma^2(y) v(y, t)] \right\} dy = \int \rho(y) [\partial_t v(y, t)] dy$$

by letting $s \downarrow 0$. Finally, by observing that the last display holds for all ρ , where we have allowed ρ to be quite an arbitrary function, we obtain the forward equation for v

$$(6.21) \quad \partial_t v(y, t) = -\partial_y [\mu(y) v(y, t)] + \frac{1}{2} \partial_{yy} [\sigma^2(y) v(y, t)].$$

Of the two Kolmogorov equations, in a sense the forward equation gives the more natural description of the evolution of the process, being an equation for the density as a function of the “forward” variable y , and holding the initial distribution fixed. In contrast, the backward equation keeps the forward variable y fixed, and describes the density at time t as a function of the “backward” variable x that gives the initial state of the process at time 0. An advantage of the backward equation is that it requires weaker assumptions than the forward equation. You can get a hint of this just by looking at the form of the two equations: notice that the functions μ and σ appear inside derivatives in the forward equation, while they do not in the backward equation. So one might expect that the forward equation requires more smoothness assumptions on μ and σ .

6.5 Stationary distributions

The forward equation can be used to find stationary distributions. Suppose $\pi(\cdot)$ is a stationary density for the diffusion X . This means that if we start the diffusion in the density π , then it stays in π . Consequently, the function

$$v(t, y) = \pi(y)$$

should satisfy the forward equation (6.21), so that π satisfies the ordinary differential equation

$$(6.22) \quad 0 = -\frac{d}{dy}[\mu(y)\pi(y)] + \frac{1}{2} \frac{d^2}{dy^2}[\sigma^2(y)\pi(y)].$$

Thus, we can solve for stationary distributions by solving ordinary differential equations.

(6.23) EXAMPLE [DIFFUSION IN A POTENTIAL WELL]. Let $\psi : \mathbb{R} \rightarrow \mathbb{R}$ be a “potential function”; we could think of ψ as a function that we want to minimize. Assume that ψ has a derivative [almost everywhere at least] and $\psi(y) \rightarrow \infty$ as $|y| \rightarrow \infty$. A deterministic gradient descent method for minimizing ψ could be described as follows. Say we are at position X_t at time t . Then we evaluate the gradient [i.e. derivative] $\psi'(X_t)$ and take a small step to move to the position $X_t - \psi'(X_t) \Delta t$ at time $t + \Delta t$. Letting $\Delta t \rightarrow 0$, we obtain the continuous time model

$$\frac{dX_t}{dt} = -\psi'(X_t).$$

In the tiny time interval dt , the increment in the function X is $dX_t = -\psi'(X_t) dt$.

This was a model of a deterministic algorithm. Suppose that, in the spirit of simulated annealing, we add some randomness to the method in the hope that this will prevent our getting stuck in local minima. We'll model this by taking X to be a diffusion with infinitesimal drift function $\mu(x) = -\psi'(x)$ and constant infinitesimal variance $\sigma^2(x) \equiv \sigma^2$, say. Incidentally, we can write this symbolically in the form

$$dX_t = -\psi'(X_t) dt + \sigma dW_t ;$$

more about this later. Let us find the stationary distribution of the process X . The equation (6.22) becomes

$$0 = \frac{d}{dy}[\psi'(y)\pi(y)] + \frac{\sigma^2}{2} \frac{d^2}{dy^2}[\pi(y)],$$

or

$$\pi'(y) + \frac{2}{\sigma^2} \psi'(y)\pi(y) = B$$

for some constant B . Introducing the integrating factor $\exp[(2/\sigma^2)\psi(y)]$ gives

$$\left(e^{(2/\sigma^2)\psi(y)} \pi(y) \right)' = B e^{(2/\sigma^2)\psi(y)},$$

so that

$$\pi(y) = e^{-(2/\sigma^2)\psi(y)} \left[B \int_0^y e^{(2/\sigma^2)\psi(x)} dx + C \right].$$

[[Notice that the lower limit 0 in the integral could equally well be any number; a change in this limit would simply be compensated by a corresponding change in the value of C . This is why many people write “ $\int^y e^{(2/\sigma^2)\psi(x)} dx$ ” here—the lower limit does not matter.]] Our assumption that $\psi(y) \rightarrow \infty$ as $|y| \rightarrow \infty$ clearly implies that the last integral approaches ∞ as $y \rightarrow \infty$ and $-\infty$ as $y \rightarrow -\infty$. Therefore, since $\pi(y)$ must be nonnegative for all y , we must have $B = 0$. So we have found that

$$\pi(y) = C e^{-(2/\sigma^2)\psi(y)}$$

where C is the normalizing constant that makes $\int_{-\infty}^{\infty} \pi(y) dy = 1$.

For example,

1. If $\psi(y) = (1/2)y^2$ (so that $\mu(y) = -y$) and $\sigma^2 = 2$, then $\pi(y) = C e^{-y^2/2}$. This is the standard Ornstein-Uhlenbeck process, and we have found that it has a standard normal distribution as its stationary distribution.
2. If $\psi(y) = |y|$ (so that $\mu(y)$ is -1 when y is positive and $+1$ when y is negative) and $\sigma^2 = 2$, then $\pi(y) = C e^{-|y|}$.

□

(6.24) EXERCISE. *Can you find a diffusion that has a standard Cauchy distribution as its stationary distribution?*

6.6 Probability flux for diffusions

Think of the probability $P\{X_t < x, X_{t+h} > x\}$ as the flux from the set $(-\infty, x)$ to the set (x, ∞) over the time interval $[t, t+h]$; that is, it is the flux across x “from left to right”. Similarly, the probability $P\{X_t > x, X_{t+h} < x\}$ is the flux across x in the other direction, from right to left. We are interested in the net flux across x , defined to be the difference

$$P\{X_t < x, X_{t+h} > x\} - P\{X_t > x, X_{t+h} < x\}.$$

For example, this will be 0 for a stationary process. Let $v(t, \xi)$ denote the density of X_t evaluated at the state ξ , so that

$$P\{X_t \in d\xi\} = v(\xi, t) d\xi.$$

For a small positive h , let $\Delta(\xi, y)$ denote the conditional density of the increment $X_{t+h} - X_t$, given that $X_t = \xi$, evaluated at y ; that is,

$$P\{X_{t+h} - X_t \in dy \mid X_t = \xi\} = \Delta(\xi, y) dy.$$

Then

$$\begin{aligned} P\{X_t < x, X_{t+h} > x\} &= \int_{y=0}^{\infty} \int_{\xi=x-y}^x P\{X_{t+h} - X_t \in dy \mid X_t = \xi\} P\{X_t \in d\xi\} \\ &= \int_{y=0}^{\infty} \int_{\xi=x-y}^x v(t, \xi) \Delta(\xi, y) d\xi dy. \end{aligned}$$

Similarly,

$$P\{X_t > x, X_{t+h} < x\} = \int_{y=-\infty}^0 \int_{\xi=x}^{x-y} v(t, \xi) \Delta(\xi, y) d\xi dy.$$

Thus, the net flux is given by

$$\int_{y=-\infty}^{\infty} \int_{\xi=x-y}^x v(t, \xi) \Delta(\xi, y) d\xi dy.$$

Consider the integrand $v(t, \xi) \Delta(\xi, y)$ as a function of ξ . For small h , only values of ξ that are very near to x will contribute significantly. This motivates a Taylor expansion about $\xi = x$, which we will do without worrying about being rigorous. This gives

$$v(t, \xi) \Delta(\xi, y) = v(t, x) \Delta(x, y) + \{\partial_x v(t, x) \Delta(x, y)\} (\xi - x) + \cdots,$$

so that

$$\begin{aligned} &P\{X_t < x, X_{t+h} > x\} - P\{X_t > x, X_{t+h} < x\} \\ &= \int_{y=-\infty}^{\infty} \int_{\xi=x-y}^x \left[v(t, x) \Delta(x, y) + \{\partial_x v(t, x) \Delta(x, y)\} (\xi - x) + \cdots \right] d\xi dy \\ &= \int_{y=-\infty}^{\infty} \left[y v(t, x) \Delta(x, y) - \frac{y^2}{2} \{\partial_x v(t, x) \Delta(x, y)\} + \cdots \right] dy \\ &= v(t, x) \int_{-\infty}^{\infty} y \Delta(x, y) dy - \frac{1}{2} \partial_x \left\{ v(t, x) \int_{-\infty}^{\infty} y^2 \Delta(x, y) dy \right\} + \cdots \\ &= v(t, x) \mu(x) h - \frac{1}{2} \partial_x \{v(t, x) \sigma^2(x) h\} + o(h). \end{aligned}$$

To convert this into a *rate* of flux per unit time, divide by the time increment h . Then letting h tend to 0 gives an instantaneous rate at time t . Thus, the rate of net probability flux across x at time t is given by

$$v(t, x) \mu(x) - \frac{1}{2} \partial_x \{v(t, x) \sigma^2(x)\}.$$

You should think about the form of this expression to make sure it makes qualitative sense. For example, the first term is consistent with the obvious thought that increasing $\mu(x)$ should increase the net flux across x from left to right. To think about the second term, suppose $\mu(x)$ were 0. Then, for example, if $\sigma^2(\cdot)$ were constant, then the flux at x would be determined by the derivative $\partial_x v(t, x)$. For example, if the density were increasing through x , so that it is larger to the right of x than to the left, there would be a negative probability flux—a net flux to the left—which makes sense. The qualitative effect of the

σ^2 function also makes sense: for example, if $\sigma^2(\cdot)$ were increasing through x , so that is it bigger to the right than to the left of x , this would again contribute a net flux to the left.

This probability flux allows us to recover Kolmogorov's forward equation. In fact, the net flux has a simpler alternative expression, found by adding and subtracting the probability $P\{X_t < x, X_{t+h} < x\}$:

$$\begin{aligned} & P\{X_t < x, X_{t+h} > x\} - P\{X_t > x, X_{t+h} < x\} \\ &= P\{X_t < x, X_{t+h} > x\} + P\{X_t < x, X_{t+h} < x\} \\ &\quad - P\{X_t < x, X_{t+h} < x\} - P\{X_t > x, X_{t+h} < x\} \\ &= P\{X_t < x\} - P\{X_{t+h} < x\}. \end{aligned}$$

Thus, we have found that

$$P\{X_{t+h} < x\} - P\{X_t < x\} = -v(t, x)\mu(x)h + \frac{1}{2}\partial_x \{v(t, x)\sigma^2(x)h\} + o(h),$$

so that

$$\partial_t P\{X_t < x\} = -v(t, x)\mu(x) + \frac{1}{2}\partial_x \{v(t, x)\sigma^2(x)\}.$$

Finally, differentiating with respect to x gives the familiar forward equation

$$\partial_t v(t, x) = -\partial_x \{v(t, x)\mu(x)\} + \frac{1}{2}\partial_{xx} \{v(t, x)\sigma^2(x)\}.$$

6.7 Quadratic Variation of Brownian Motion

A very important property of the standard Brownian motion process W is that its “quadratic variation” over the interval $[0, t]$ is t , with probability 1. This property is fundamental to stochastic integration. Soon we will see that the formula

$$\int_0^t \{[dW(s)]^2\} = t$$

makes some sort of sense. Since we also have $\int_0^t ds = t$ for all t , it seems natural to write

$$[dW(s)]^2 = ds.$$

In fact, this is the essence of “Ito’s formula,” to be discussed below: Ito’s formula just involves doing a few terms of a Taylor series expansion and changing each $[dW(t)]^2$ that arises into a dt .

We will start with a definition of quadratic variation for an ordinary, nonrandom function f . Then we will look at the quadratic variation of a typical sample path of Brownian motion.

(6.25) DEFINITION. Let f be a real-valued function defined at least on the interval $[0, t]$. The **quadratic variation** $q_t(f)$ of f over $[0, t]$ is defined to be

$$q_t(f) = \lim_{n \rightarrow \infty} \sum_{k=1}^{2^n} \left[f\left(\frac{kt}{2^n}\right) - f\left(\frac{(k-1)t}{2^n}\right) \right]^2$$

if the limit exists (otherwise the quadratic variation is undefined).

The concept of quadratic variation is not very interesting for most of the nice, tame functions we think about every day, as the next result shows.

(6.26) FACT. *If f is continuous and of bounded variation, then $q_t(f) = 0$ for all t .*

PROOF: For any given t , observe that

$$\begin{aligned} & \sum_{k=1}^{2^n} \left[f\left(\frac{kt}{2^n}\right) - f\left(\frac{(k-1)t}{2^n}\right) \right]^2 \\ & \leq \left\{ \max_{1 \leq j \leq 2^n} \left| f\left(\frac{jt}{2^n}\right) - f\left(\frac{(j-1)t}{2^n}\right) \right| \right\} \sum_{k=1}^{2^n} \left| f\left(\frac{kt}{2^n}\right) - f\left(\frac{(k-1)t}{2^n}\right) \right|. \end{aligned}$$

However, the maximum over j approaches 0 as $n \rightarrow \infty$, since the continuity assumed of the function f on the closed interval $[0, t]$ implies that f is uniformly continuous there. The last sum over k is bounded as $n \rightarrow \infty$; this is the meaning of bounded variation. Thus, the limit in the definition (6.25) is 0. \square

Thus, for example, any continuous function f that we can draw has zero quadratic variation on $[0, t]$: the finite total length of the graph is greater than the sum of the increments in the function, which implies the function is of bounded variation.

The quadratic variation concept comes alive in the sample paths of Brownian motion and related processes. For example, sample paths of standard Brownian motion have the interesting property that, despite their continuity, with probability 1 they have quadratic variation t on the interval $[0, t]$. Thus, in particular, by the previous fact, of course the paths of Brownian motion must have infinite variation with probability 1.

To save writing, let $\Delta W_{k,n}$ denote the increment

$$\Delta W_{k,n} = W\left(\frac{kt}{2^n}\right) - W\left(\frac{(k-1)t}{2^n}\right)$$

and let $Q_n = \sum_{k=1}^{2^n} (\Delta W_{k,n})^2$. Then here is our main result about quadratic variation.

(6.27) THEOREM. *With probability 1, $Q_n \rightarrow t$ as $n \rightarrow \infty$. Also, $Q_n \rightarrow t$ in mean square, that is, $E[(Q_n - t)^2] \rightarrow 0$.*

PROOF: The proof is based on simple calculations of EQ_n and $\text{Var}Q_n$.

$$EQ_n = \sum_{k=1}^{2^n} E[(\Delta W_{k,n})^2] = \sum_{k=1}^{2^n} \text{Var}(\Delta W_{k,n}) = \sum_{k=1}^{2^n} \left(\frac{t}{2^n}\right) = t.$$

That's reassuring. Next, using the independence of the increments $\Delta W_{k,n}$ for different values of k ,

$$\begin{aligned} \text{Var}(Q_n) &= \sum_{k=1}^{2^n} \text{Var}[(\Delta W_{k,n})^2] = 2^n \text{Var}[N(0, t/2^n)^2] \\ &= 2^n \text{Var}\left[\left(\frac{t}{2^n}\right) N(0, 1)^2\right] = t^2 2^{-n} \underbrace{\text{Var}[N(0, 1)^2]}_2 \rightarrow 0 \quad \text{as } n \rightarrow \infty. \end{aligned}$$

This is already enough to show mean square convergence:

$$E[(Q_n - t)^2] = [E(Q_n - t)]^2 + \text{Var}(Q_n - t) \rightarrow 0 \text{ as } n \rightarrow \infty.$$

To prove almost sure convergence, let $\epsilon > 0$. By Chebyshev,

$$P\{|Q_n - t| > \epsilon\} \leq \frac{\text{Var} Q_n}{\epsilon^2} = \frac{t^2}{\epsilon^2 2^{n-1}},$$

which implies that $\sum_n P\{|Q_n - t| > \epsilon\} < \infty$, which, by the Borel-Cantelli Lemma implies that $P\{|Q_n - t| > \epsilon \text{ infinitely often}\} = 0$. Thus, $P\{|Q_n - t| > \epsilon \text{ just finitely often}\} = 1$, so that, since ϵ was arbitrary, $Q_n \rightarrow t$ almost surely. \square

There are two remarkable statements contained in this result. The first is that the quadratic variation of Brownian motion is positive, not 0. Second, all Brownian sample paths (up to a set of probability 0) have precisely the same quadratic variation over every interval. Although different Brownian paths obviously will look very different, they have exactly the same quadratic variation.

(6.28) EXERCISE. Let $X(t) = \mu t + \sigma W(t)$ be a (μ, σ^2) -Brownian motion. Show that, with probability 1, the quadratic variation of X on $[0, t]$ is $\sigma^2 t$.

This has an interesting implication in terms of statistical estimation. Suppose that we are observing a (μ, σ^2) -Brownian motion X and we want to estimate μ and σ^2 . If we observe the X process for a very long time T , we can estimate the drift μ well by its slope $X(T)/T$. This estimator gets better as T increases, since $X(t)/t \rightarrow \mu$ as $t \rightarrow \infty$. Of course there is always some error in this estimator; this error just gets smaller as the observation time T increases. However, in contrast, for estimating σ^2 , it is enough to observe the X process over *any* interval of time (arbitrarily short), and we can infer σ^2 *exactly*! The reason for this is: by the previous exercise the quadratic variation of X over an interval $[0, t]$ is $\sigma^2 t$ with probability 1. Therefore, since the quadratic variation over an interval is exactly determined by the path just over that interval, σ^2 is determined by the path over any interval. Thus, we can estimate the variance parameter of a Brownian motion exactly [or at least as closely as we desire, depending on how much calculation we are willing to do] from observing a sample path over any time interval, arbitrarily short.

6.8 The Idea of Stochastic differential equations

Since differential equations play such an important role in modelling and analyzing deterministic phenomena, it is natural to contemplate a stochastic analog of differential equations, which we would hope could do the same for phenomena in which randomness is important.

So what is a stochastic differential equation (SDE)? Let's start with a couple of thoughts about ordinary, deterministic, first-order differential equations, which look something like this:

$$\frac{dX_t}{dt} = f(X_t, t).$$

It is interesting to observe that such equations have a kind of “Markov-like” property: in generating a solution curve of a differential equation on a computer, for example, we just have to know the value of X at time t in order to perform the calculation $X(t + \Delta t) \doteq X(t) + f(X(t), t)\Delta t$. Given $X(t)$, all past values $X(s)$ for $s < t$ are superfluous for determining future behavior. The Markovishness of this last statement is apparent! On the other hand, second order equations do not have this property. In solving a second order differential equation numerically, one must use $X(t - \Delta t)$ and $X(t)$ to approximate $X(t + \Delta t)$. However, a familiar device in differential equations is to convert a second order equation for X into a first order differential equation for the vector $(X \ \dot{X})$. Thus, by expanding the “state description” to include the information $\dot{X}(t)$ as well as $X(t)$, we may convert the “non-Markov” process $X(t)$ into the “Markov process” $(X(t) \ \dot{X}(t))$. This is generally true of Markov (stochastic) processes: if we include enough information in the state, we can get the process to be Markov. However, such augmentation of the state space is seldom practical.

A simple example of a deterministic differential equation is $dX/dt = rX$, which might model exponential growth of a population. If we wanted to model a “noisy” growth rate, we might propose an equation of the form

$$\frac{dX(t)}{dt} = [r + N(t)]X(t),$$

where N is a noise process. Here $\{N(t), t \geq 0\}$ is a stochastic process, and hence so is $\{X(t), t \geq 0\}$. More generally, we can consider equations of the form

$$\frac{dX_t}{dt} = \mu(X_t, t) + \sigma(X_t, t)N_t.$$

What sorts of assumptions should we make about the noise process? Here are some that seem desirable in many applied contexts. They are the characteristics of so-called “Gaussian white noise.”

1. $\{N_t, t \geq 0\}$ is a stationary Gaussian process.
2. $EN_t = 0$.
3. N_s is independent of N_t for $s \neq t$.

There is nothing really to the second assumption; it’s purely for convenience. The third assumption is very strong, but seems reasonable as an idealization of a process in which knowing the value at any one time gives very little information concerning values at other times.

Rewriting our equation as

$$\begin{aligned} dX_t &= \mu(X_t, t) dt + \sigma(X_t, t)N_t dt \\ &=: \mu(X_t, t) dt + \sigma(X_t, t) dV_t \end{aligned}$$

where we have let $\{V_t, t \geq 0\}$ be a process having increments $dV_t = N_t dt$, we see that the process $\{V_t, t \geq 0\}$ must be a Gaussian process with stationary independent increments.

There is only one sort of process satisfying these requirements: $\{V_t, t \geq 0\}$ must be a Brownian motion. So let's rewrite our equation one more time:

$$dX_t = \mu(X_t, t) dt + \sigma(X_t, t) dW_t,$$

where as usual $\{W_t, t \geq 0\}$ is a standard Brownian motion.

There is one serious problem with all of this: the paths of $\{W_t, t \geq 0\}$ are not differentiable! So what could we mean by dW_t ? In fact, it is a nontrivial problem to make sense of this. The theory we will discuss later *defines* the previous equation to be just a convenient and suggestive way of writing the integrated form

$$X_t - X_0 = \int_0^t \mu(X_s, s) ds + \int_0^t \sigma(X_s, s) dW_s.$$

Now, do we know what this equation means? It turns out that the first integral presents no conceptual problems; it can be defined for each fixed ω as just an ordinary Riemann integral; that is, it is just the random variable (i.e., function of ω) $\omega \mapsto \int_0^t \mu(X_s(\omega), s) ds$. The second integral is a different story. It is natural first to try the idea that worked above. That is, consider defining the second integral to be the random variable $\omega \mapsto \int_0^t \sigma(X_s(\omega), s) dW_s(\omega)$. By this we would presumably mean that the value of the integral at ω is the ordinary, non-stochastic, Riemann-Stieltjes integral $\int_0^t \sigma(X_s(\omega), s) dW_s(\omega)$. However, as we will discuss in more detail later, the last integral is generally not well-defined for the sorts of stochastic processes $\{X_t, t \geq 0\}$ that we want to consider. This is the sticking point that causes the definition of the stochastic integral to be such a long story; stochastic integrals need to be given a special definition that is not an obvious extension of definitions that apply to deterministic functions.

The issues here are in a sense analogous to some issues concerning the δ function. For most practical purposes, one can think heuristically of the δ function $\delta(\cdot)$ as a “function” satisfying $\delta(x) = 0$ for all $x \neq 0$ and $\int_{-\infty}^{\infty} \delta(x) dx = 1$. Of course, no such function exists, but that does not stop people from using this idea to good advantage in many situations. There is even a “calculus” of the δ function. These rules for manipulating the δ function are convenient and useful for automating and expediting certain routine calculations — in fact, they help *make* such calculations routine. However, when the δ function manipulator inevitably begins to wonder whether the manipulations are justified or even what they really mean, or if any sort of nonstandard situation arises, then heuristic concepts and manipulation rules may not be enough. Fortunately, in such circumstances one can turn to a piece of mathematics that puts the δ function and δ calculus on a firm mathematical foundation. As it turns out, the theory says that the δ function has rigorous meaning only when it is inside an integral, and the manipulation rules of the δ calculus are really just convenient shorthand for longer statements involving integrals. Stochastic calculus puts “white noise” on a firm mathematical foundation in an analogous manner. White noise is a useful heuristic concept, but does not exist in the usual sense of a Gaussian stochastic process. The rules of stochastic calculus, which tell us how to manipulate “stochastic differentials” involving white noise, are really shorthand for statements involving stochastic integrals. There is a body of mathematical theory that gives stochastic integrals a rigorous meaning.

Stochastic calculus is quite a technical subject, and a rigorous mathematical treatment is not easy. The typical path for a student interested in learning stochastic calculus is this. First take a course in measure-theoretic probability. Then take a course in stochastic calculus. In such a course, typically a good fraction of the semester is required just to define a stochastic integral. By the time the definition is reached, the class is confused and exhausted, and has forgotten why they wanted to do any of this in the first place. By the end of the semester, only a fortunate few are not completely lost. So, still intent on finding out what stochastic calculus is about, the student takes the class again. And the process iterates.

Now, contrary to the dreary picture just presented, it is not the case that the purpose of stochastic calculus is to make your life miserable; in fact, ideally it should make your life easier. The original purpose of a calculus is to help you calculate. So let's discuss a bit of stochastic calculus at a basic level, without worrying about being mathematically rigorous.

6.9 Simple Examples of Stochastic Calculus and Ito's Formula

Let's start just by getting an idea of what some simple manipulations of stochastic calculus look like.

A diffusion $\{X(t), t \geq 0\}$ with infinitesimal parameters $\mu(x, t)$ and $\sigma^2(x, t)$ has stochastic differential

$$dX = \mu(X, t) dt + \sigma(X, t) dW.$$

Thus, turning this around, it is useful to know the stochastic differential of a diffusion, since from it we can simply read off the infinitesimal parameters of the diffusion.

A version of **Ito's formula**, one of the fundamental results of stochastic calculus, says that if X is a stochastic process having stochastic differential dX and f is a suitably nice function, then the process $Y := f(X)$ has stochastic differential

$$dY_t = f'(X_t)dX_t + \frac{1}{2}f''(X_t)(dX_t)^2,$$

where $(dX_t)^2$ is computed using the rules

1. $(dt)[d(\text{anything})] = 0$,
2. $(dW_t)^2 = dt$.

Note that the second rule makes some sense heuristically from what we showed about the quadratic variation of standard Brownian motion.

In more general situations, such as when Y may be of the form $Y = f(X, t)$, Ito's formula says to compute dY by first doing a Taylor expansion, keeping terms up to quadratic order in the various differentials, and then simplifying the result by using the rules above.

(6.29) EXAMPLE. Let's redo the geometric Brownian motion example we did earlier. That is, consider the Brownian motion with drift $X_t = \mu t + \sigma W_t$, which has stochastic differential

$dX_t = \mu dt + \sigma dW_t$. Define the geometric Brownian motion $Y_t = e^{X_t}$. What are the infinitesimal mean and variance functions of the diffusion Y ?

Here $Y = f(X) = e^X$, so that $f'(X) = f''(X) = e^X = Y$. Also,

$$\begin{aligned}(dX_t)^2 &= (\mu dt + \sigma dW_t)^2 \\ &= \mu^2(dt)^2 + 2\mu\sigma(dt)(dW_t) + \sigma^2(dW_t)^2 \\ &= 0 + 0 + \sigma^2 dt.\end{aligned}$$

Thus,

$$\begin{aligned}dY_t &= f'(X_t) dX_t + (1/2)f''(X_t) (dX_t)^2 \\ &= Y_t(\mu dt + \sigma dW_t) + (1/2)Y_t\sigma^2 dt \\ &= Y_t[\mu + (1/2)\sigma^2] dt + \sigma Y_t dW_t,\end{aligned}$$

so that Y is a diffusion process having infinitesimal parameters $\mu_Y(y) = [\mu + (1/2)\sigma^2]y$ and $\sigma_Y^2(y) = \sigma^2 y^2$. \square

The term *Ito process* refers to a stochastic process that has a stochastic differential of the form

$$dZ_t = X_t dW_t + Y_t dt$$

where the processes X and Y satisfy some conditions: roughly,

1. X and Y are adapted; that is, X_t and Y_t are determined by the portion of the Brownian path $W_0^t = \{W_s : s \leq t\}$ up to time t .
2. Conditions that assure that X and Y are “not too big”.

(6.30) EXAMPLE. There are simple Ito processes that are not diffusions. For example, suppose $Z_t = X_t^2$ where $X_t = \mu t + \sigma W_t$. Then by Ito's formula,

$$\begin{aligned}dZ &= 2X dX + (dX)^2 \\ &= 2X(\mu dt + \sigma dW) + \sigma^2 dt \\ &= (2\sigma X)dW + (2\mu X + \sigma^2)dt.\end{aligned}$$

Now we can see that Z is not a diffusion. In fact, if it were, then we could read off its infinitesimal mean and variance parameters from its stochastic differential. Let's try to do so and see where the trouble arises. The infinitesimal variance part looks OK: $\sigma_Z^2(z) = (2\sigma X)^2 = 4\sigma^2 z$. However, our candidate for an infinitesimal mean parameter, $2\mu x + \sigma^2$, cannot (unless $\mu = 0$) be expressed as a function of $z := x^2$, since x is not a function of z . In fact, Z is not even a Markov process. Given only that $Z_t = z$, say, of course we cannot tell whether $X_t = \sqrt{z}$ or $X_t = -\sqrt{z}$. For purposes of predicting the future of Z , it seems clear that we would like to know more about the past behavior of Z , because that could give us a hint about whether $X_t = \sqrt{z}$ or $X_t = -\sqrt{z}$. This reeks of nonmarkovianity. Note that if $\mu = 0$, then Z is a diffusion, with $\sigma_Z^2(z) = 4\sigma^2 z$ and $\mu_Z(z) = \sigma^2$. \square

6.10 The Black-Scholes Formula

In this section we will use some Ito calculus ideas to give a derivation of the celebrated Black-Scholes formula of finance. You financial insiders will notice that I have adhered to the unspoken rule of attaching the obligatory modifier “celebrated” to the hallowed formula. Yes, indeed it is celebrated. I’d give it a 98; it has a good beat, and you can hope to make a lot of money from it.

The problem addressed by the formula is determining how much an “option” should cost. We’ll talk about “call” options. A call option on a certain stock is the right to buy a share of the stock at a certain fixed price (the “strike price”) at a certain fixed time in the future (the “maturity date”). If I buy a call option from you, I am paying you a certain amount of money in return for the right to force you to sell me a share of the stock, if I want it, at the strike price on the maturity date. Our problem is, what is the “right” amount of money for me to pay for this right?

The meaning of the term “right” here relates to the economic term *arbitrage*. An arbitrage opportunity is the opportunity to make money instantly and without risk. That is, you get some money for sure, right now. Such free lunches are not supposed to exist, or at least should be rare and short-lived. The basic reason for believing this is that many people are looking for such opportunities to make money. If the price of commodity A were so low, for example, that some clever financial transaction involving buying commodity A and perhaps selling some others were guaranteed to make an instantaneous profit, then many eager arbitrage seekers would try to perform the transaction many times. The resulting increased demand for commodity A would cause its price to increase, thereby destroying the arbitrage opportunity.

Let the stock price at time t be $X(t)$. Let t_1 denote the maturity date, and k , the strike price of the option. A little thought shows that the value of the option at time t_1 is the random variable $(X(t_1) - k)_+$, since it makes sense for me to “exercise” the option if and only if $X(t_1) > k$. Let $Y(t)$ denote the magic, no-arbitrage price for the option that we are seeking. Assume that $Y(t)$ may be expressed as some function $f(X(t), t)$ of $X(t)$ and t ; our goal is to determine the function f . As a final assumption, let the “interest rate” or the “riskless rate of return” be r ; that is, \$1 in a riskless investment today becomes $\$e^{rt}$ at time t .

Stocks and options are risky investments; they are not like banks, because we do not know how much we will be able to get back at any future time. We assume a simple probabilistic model for the evolution of the stock price: suppose X is the geometric Brownian motion having stochastic differential

$$dX = \mu X dt + \sigma X dW.$$

Thus, X is the exponential of a Brownian motion with drift. It is comforting that this process will not become negative. Indeed, in accordance with our view of Brownian motions as the stochastic analog of linear functions, geometric Brownian motion seems like a sensible first model: riskless investments change as $\exp(\text{linear function})$, and stocks change as $\exp(\text{Brownian motion})$. What we are really assuming is that returns, that is, proportional changes in the stock price, are stationary and independent over different time intervals.

A key idea behind the Black-Scholes formula is that by trading both the stock and the option simultaneously, we can reduce our risk. If we just buy a call option, we will be happy if the stock price goes up and sad if it goes down. What can we do with the stock to restore ourselves to blissful equanimity, in which we are indifferent whether the stock price goes up or down? We need to do something that will make us sad if the stock goes up and happy if it goes down. Selling some stock will do the trick!

Thus, let us consider the portfolio: “buy one call, sell A shares of stock, and put the rest of our money in a bond.” The price of the bond is assumed to follow the riskless form e^{rt} . Our investing strategy consists of two stochastic processes: at time t we hold $A = A(t)$ shares of stock and $B = B(t)$ “shares” of bond, where A and B are assumed to be adapted stochastic processes. The value of this portfolio at time t is

$$(6.31) \quad V(t) = Y(t) - A(t)X(t) + B(t)e^{rt}.$$

We say that the strategy (A, B) is *self-financing* if $X(t) dA(t) = e^{rt} dB(t)$ for all t . The interpretation of this is: our trades of the stock and bond neither remove money from the portfolio nor require adding additional money from the outside. The changes $dA(t)$ and $dB(t)$ we make in our holdings just reallocate the money in the portfolio, without adding or removing money. This relationship between the processes A and B says that the proceeds $X(t) dA(t)$ from the sale of stock at time t are precisely offset by the cost $e^{rt} dB(t)$ of the stock purchases made at time t . [These changes could be positive or negative; if they are negative we are buying stock and selling bonds at time t , and we are using the money from the bond sales to buy the stock.] Any gains or losses in the value of the portfolio then must come from changes in the prices of the assets, not from our changing of our allocations $A(t)$ or $B(t)$.

From the expression for $V(t)$ we have

$$dV(t) = dY(t) - A(t) dX(t) - X(t) dA(t) + B(t)re^{rt} dt + e^{rt} dB(t).$$

The self-financing condition causes two of the terms to cancel, leaving

$$(6.32) \quad dV(t) = dY(t) - A(t) dX(t) + B(t)re^{rt} dt.$$

With our assumption that $Y(t) = f(X(t), t)$, Ito’s formula gives

$$dY = f_x(X, t)dX + f_t(X, t)dt + (1/2)f_{xx}\sigma^2 X^2 dt.$$

Therefore,

$$(6.33) \quad dV = [f_x(X, t) - A(t)]dX(t) + [f_t(X, t) + (1/2)f_{xx}\sigma^2 X^2 + B(t)re^{rt}]dt.$$

There is an obvious choice that looks interesting: Suppose we adopt the strategy $A(t) = f_x(X(t), t)$ —we can conceive of doing this, even though we do not happen to know what f is yet. Observe that the differential dV then would have no dX part, just a dt part, and, by (6.31),

$$B(t)re^{rt} = r[V(t) - Y(t) + A(t)X(t)] = rV(t) - rf(X, t) + rX(t)f_x(X(t), t).$$

Since the stochastic differential dV has only a deterministic “ dt ” part and no stochastic “ dX ” part, the portfolio V is riskless over a short time interval following time t . Now a little economic no-arbitrage reasoning gives that the instantaneous rate of return of this portfolio V must be the assumed riskless rate of return r , that is,

$$(6.34) \quad dV(t) = rV(t) dt.$$

Equating (6.33) and (6.34) gives the equation

$$\begin{aligned} rV(t) &= f_t(X, t) + (1/2)f_{xx}\sigma^2 X^2 + B(t)re^{rt} \\ &= f_t(X, t) + (1/2)f_{xx}\sigma^2 X^2 + \{rV(t) - rf(X, t) + rX(t)f_x(X(t), t)\}, \end{aligned}$$

or

$$f_t(X(t), t) = rf(X(t), t) - rX(t)f_x(X(t), t) - (\sigma^2/2)X^2 f_{xx}(X(t), t).$$

This will hold if f satisfies the partial differential equation

$$f_t = rf - rxf_x - \frac{1}{2}\sigma^2 x^2 f_{xx}.$$

Thus, solving this equation together with the obvious boundary condition

$$f(x, t_1) = (x - k)_+$$

should in principle give us the desired function f .

The following problem gives a nice way of using probability to solve the equation.

(6.35) EXERCISE [FEYNMANN-KAC FORMULA AND SOLUTION OF BLACK-SCHOLES PDE].

1. To put the PDE in a slightly more convenient form, change variables by replacing “ t ” by “ $t_1 - t$ ” [=“time to go”], and show that the new function $g(x, t) := f(x, t_1 - t)$ satisfies the PDE

$$g_t = -rg + rxg_x + \frac{1}{2}\sigma^2 x^2 g_{xx}$$

with the boundary condition $g(x, 0) = (x - k)_+$.

2. Show that if Z is a diffusion satisfying $dZ = rZ dt + \sigma Z dW$, then the function

$$g(x, t) = E\{e^{-rt}[Z(t) - k]_+ \mid Z(0) = x\}$$

satisfies the PDE and boundary condition of part 1.

3. Using the expression in part 2, compute the answer

$$g(x, t) = x\Phi\left(\frac{\ln(x/k) + [r + (1/2)\sigma^2]t}{\sigma\sqrt{t}}\right) - ke^{-rt}\Phi\left(\frac{\ln(x/k) + [r - (1/2)\sigma^2]t}{\sigma\sqrt{t}}\right).$$

It is interesting to observe that the solution does not involve the parameter μ (in fact, no μ appears in the PDE). This seems quite mysterious at first thought: for example, one would think that a call option would be more valuable for a stock that is expected to rise a great deal than for one that is expected to fall! This is a point well worth pondering, and we will discuss this further later.

This lack of dependence on μ is fortunate. To use the B-S formula, we do not need to know or even try to estimate the μ that appeared in our original stochastic model of the stock price behavior. The parameter σ does enter into the formula, however. But σ is easier to estimate than μ .

Some final B-S comments: the Black-Scholes formula is really a sort of test of consistency: the prices of the stock, the option, and the interest rate should be consistent with each other. In particular, the theory underlying the formula is not an equilibrium theory; we do not have to worry about people's utility functions and other imponderables. That may be a key reason why the formula is so successful in practice as compared with many other results in economics: it does not require us to pretend that we know many things that we cannot know. Briefly, the reason we can get away with this is that the option is in fact "redundant" given the stock and bond—a portfolio can be formed using just the stock and the bond that duplicates the cash flows from the option. Thus, if we observe the stock price process and the interest rate, then the option price is determined by a no-arbitrage condition.

6.11 A Little About Stochastic Integrals

It seems quite clear that Ito's formula is useful. In fact, I'd say it's useful enough to be worthwhile to prove. O.K., let's prove it. So, let's see, hmmm ... what is it exactly that we're trying to show? That's a real problem; in fact, we don't know what we're trying to prove because we haven't defined the ingredients of Ito's formula. But we did say something about Ito's formula actually just being a convenient form for expressing statements about stochastic integrals. To progress with the theory, we will finally have to say what we mean by a stochastic integral.

Let's start with a simple example of an Ito integral. To choose a first example, we can use Ito's formula to generate an example for which we know the answer. By Ito's formula, $d(W^2) = 2W dW + dt$. The meaning of this statement is given by the integrated form

$$W^2(t) - W^2(0) = 2 \int_0^t W dW + t.$$

Using $W(0) = 0$ and solving for the integral gives

$$\int_0^t W dW = \frac{1}{2}[W^2(t) - t].$$

This gives us an answer to shoot for. Notice that it is different from the answer "ordinary calculus" (Riemann-Stieltjes integral) would give, which is $\int_0^t W(s) d[W(s)] = W^2(t)/2$.

Stochastic integrals are defined in terms of *mean-square convergence* of random variables.

(6.36) DEFINITION. Let Z_1, Z_2, \dots and Z be random variables with $\mathbb{E}(Z_n^2) < \infty$ for all n and $\mathbb{E}(Z^2) < \infty$. We say $Z_n \rightarrow Z$ **in mean square** (or Z is the **mean-square limit** of $\{Z_n\}$) if $\lim_{n \rightarrow \infty} \mathbb{E}[(Z_n - Z)^2] = 0$.

Back to our example: The integral $\int_0^t W dW$ is defined to be the mean square limit of the sum

$$(6.37) \quad \sum_{k=0}^n W(t_k)[W(t_{k+1}) - W(t_k)]$$

as $n \rightarrow \infty$ and the partition $0 = t_0 < t_1 < \dots < t_n = t$ becomes more and more refined.

To calculate the limit in our example, we make the substitution

$$W(t_k) = \frac{1}{2}[W(t_{k+1}) + W(t_k)] - \frac{1}{2}[W(t_{k+1}) - W(t_k)],$$

for the first $W(t_k)$ in (6.37). Then the sum becomes

$$\frac{1}{2} \sum_{k=0}^n [W^2(t_{k+1}) - W^2(t_k)] - \frac{1}{2} \sum_{k=0}^n [W(t_{k+1}) - W(t_k)]^2.$$

But $\sum_{k=0}^n [W^2(t_{k+1}) - W^2(t_k)]$ telescopes to $W^2(t)$, independently of n . Also, by what we showed about the quadratic variation of Brownian motion, $\sum_{k=0}^n [W(t_{k+1}) - W(t_k)]^2$ approaches t in mean square. Thus, the integral $\int_0^t W dW$ comes out to be $(1/2)[W^2(t) - t]$, as we expected.

(6.38) EXERCISE. Show that if we had evaluated the integrand W at the “right-hand endpoint” t_{k+1} instead of the “left-hand endpoint” t_k in the definition of the stochastic integral — that is, if we had defined $\int_0^t W dW$ to be the mean-square limit of $\sum_{k=0}^n W(t_{k+1})[W(t_{k+1}) - W(t_k)]$ — we would have obtained $\int_0^t W dW = (1/2)[W^2(t) + t]$.

The sort of modification discussed in the previous exercise would not make any difference in the definition of a Riemann-Stieltjes integral; there the integrand may be evaluated at any point of the interval that contains it. Thus, the definition of the stochastic integral is a matter of some subtlety.

In fact, more than one definition of stochastic integral with respect to Brownian motion would be reasonable. The definition that has turned out to be the most useful is the Ito integral that we have described. Thinking about our little example $\int_0^t W_s dW_s = (1/2)(W_t^2 - t)$ from above, although at first you may find the extra term $-t/2$ disconcerting, it does have a nice property: it makes the answer a martingale. That is, although the natural guess $(1/2)W_t^2$ is not a martingale, as a process in t , the odd extra term is just what we need to add in order to make the answer into a martingale. That is not a coincidence, as we will see below.

The example we worked out shows the idea used to define the stochastic integral $\int_a^b X dW$ for a whole class of \mathcal{G} of “good” processes X . Of course, the class \mathcal{G} will contain our example $X = W$. The following definition is meant to give the idea of what “good” is; it is not rigorously spelled out.

(6.39) DEFINITION. *Let \mathcal{G} be a class of “good” stochastic processes, having certain nice properties that we will not enumerate in complete detail. One important requirement of a good process $\{X_t\}$ is that for each t , the random variable X_t is not allowed to depend on any future values of the Brownian motion—that is, X_t may depend only on the values $\{W_s : s \leq t\}$ and not on the values $\{W_u : u > t\}$. The other requirements on a good process are more technical. For example, there are measurability requirements. Another property says that the values of a good process $\{X_t\}$ are not too big, in the sense that $\int_0^t E(X_s^2) ds < \infty$ for all t . Finally, we assume a certain amount of regularity on the sample paths: we require them to be right continuous and have left limits.*

(6.40) DEFINITION. *For a good process $X \in \mathcal{G}$, define the stochastic integral $\int_0^t X dW$ as the mean-square limit of the sum*

$$(6.41) \quad \sum_{k=0}^n X(t_k)[W(t_{k+1}) - W(t_k)]$$

as $n \rightarrow \infty$ and the partition $0 = t_0 < t_1 < \cdots < t_n = t$ becomes more refined.

Back to the issue of evaluating X at the left endpoint of each interval versus the right endpoint or some other choice: here is the promised result stating the useful martingale property of Ito integrals.

(6.42) FACT. *If the process $\{X_t\}$ is in the class \mathcal{G} , then the process $\{Y_t\}$ defined by $Y_t = \int_0^t X_s dW_s$ is a martingale.*

The main idea here is that evaluating the integrand at the left endpoint produces a martingale. To understand this, look at a sum of the form $I(n) = \sum_{k=1}^n X(t_{k-1})[W(t_k) - W(t_{k-1})]$; we take the limit of such sums to define the Ito integral. It is easy to see that $\{I(n)\}$ is a martingale:

$$\begin{aligned} E(I(n+1) \mid \{W(s) : 0 \leq s \leq t_n\}) \\ &= I(n) + E(X(t_n)[W(t_{n+1}) - W(t_n)] \mid \{W(s) : 0 \leq s \leq t_n\}) \\ &= I(n) + X(t_n)E(W(t_{n+1}) - W(t_n) \mid \{W(s) : 0 \leq s \leq t_n\}) = I(n) \end{aligned}$$

It is easy to see intuitively the role of the left endpoint. In a gambling interpretation of the sums $\sum_{k=1}^n X(t_{k-1})[W(t_k) - W(t_{k-1})]$ defining the stochastic integral, the increments $W(t_k) - W(t_{k-1})$ in the Brownian motion can be thought of as the gains from a sequence of fair games. The sum is our fortune when we are allowed to multiply the stakes by varying amounts; in particular, we multiply the winnings $W(t_k) - W(t_{k-1})$ by $X(t_{k-1})$. Clearly it would be to our advantage if we were allowed to multiply by $W(t_k) - W(t_{k-1})$, for example, since then our total winnings would be $\sum_{k=1}^n [W(t_k) - W(t_{k-1})]^2$! These sort

of gains correspond to evaluating the integrand at the right endpoint, and are prevented if we evaluate at the left endpoint.

(6.43) EXERCISE. Let $\{W_t\}$ be a standard Brownian motion and define $X_t = (W_t)^3$.

1. Use Ito's formula to find the mean and variance functions $\mu_X(x)$ and $\sigma_X^2(x)$ of the diffusion X .
2. You should have found that $\mu_X(0) = 0$ and $\sigma_X^2(0) = 0$. Yet the process is perfectly free to pass through the state 0, taking both positive and negative values. What do you make of this?

(6.44) EXERCISE. Find a function $f = f(t)$ so that the process $X_t := W_t^3 - f(t)W_t$ is a martingale.

A PARTING THOUGHT...

Again, why bother with stochastic calculus?

We've mentioned two major roles for Ito calculus. Firstly, it provides the mathematical machinery needed to for a rigorous treatment of diffusions and diffusion-like processes. Secondly, it is useful for doing calculations. Although we have tried to glimpse a bit of the theory, in this course it is more appropriate to emphasize applications and the calculation aspect. In your first exposure to ordinary calculus, I hope you were not forced to agonize over what an area is and precisely what functions are integrable. You said, let's see how to calculate some areas, and volumes, and work, and ...

7. Likelihood Ratios

The idea of this chapter is known under various names: likelihood ratios, Radon-Nikodym derivatives, and change of measure. It is the basis of the technique of exponential tilting in the study of large deviations, importance sampling in simulation, and the Cameron-Martin-Girsanov transformation in stochastic differential equations.

In a sense, the idea is mathematically rather trivial: we multiply and divide by something inside a sum or integral, and observe that the answer does not change. But the probabilistic interpretation is powerful. When we are given a random variable whose expectation we wish to evaluate, we are not restricted to work only with the particular probability measure given in the problem. We are free to choose another probability measure—that is, pretend the random variable is generated by a different distribution—as long as we compensate for the change in probability measure by multiplying the random variable by a likelihood ratio before taking the expectation under the new probability measure.

7.1 The idea of likelihood ratios

Suppose X is a random variable on a probability space Ω that has two different probability measures P_1 and P_2 defined on it. For simplicity, suppose for now that the probability space Ω has only a finite number of elements. Then we may define the *likelihood ratio* L simply by

$$(7.1) \quad L(\omega) = \frac{P_1\{\omega\}}{P_2\{\omega\}}.$$

We need to worry about possible division by zero to make sure the quotient in (7.1) makes sense. In fact, the appropriate condition to require is this:

$$(7.2) \quad \text{For each } \omega \text{ satisfying } P_2\{\omega\} = 0 \text{ we also have } P_1\{\omega\} = 0.$$

That is, as we will see in a moment, it will be all right for the denominator to be zero as long as the numerator is also zero.

The likelihood ratio L is used to relate expectations (and hence also probabilities) taken under P_1 to those taken under P_2 . In the discrete situation that we are now considering,

$$(7.3) \quad \begin{aligned} E_1(X) &= \sum X(\omega)P_1\{\omega\} = \sum X(\omega)\frac{P_1\{\omega\}}{P_2\{\omega\}}P_2\{\omega\} \\ &= \sum X(\omega)L(\omega)P_2\{\omega\} = E_2(XL). \end{aligned}$$

Now you should be able to think through why we are not bothered by those ω that make $L(\omega)$ of the form “0/0”. In fact, the set of all such dubious ω values will have probability zero under P_2 , so it cannot affect the expectation $E_2(XL)$.

[[To write this down more carefully:

$$\begin{aligned}
 E_1(X) &= \sum_{P_1\{\omega\}>0} X(\omega)P_1\{\omega\} \\
 &= \sum_{P_1\{\omega\}>0} X(\omega)P_1\{\omega\} + \sum_{P_1\{\omega\}=0, P_2\{\omega\}>0} X(\omega)P_1\{\omega\} \\
 &= \sum_{P_2\{\omega\}>0} X(\omega)P_1\{\omega\} \\
 &= \sum_{P_2\{\omega\}>0} X(\omega) \frac{P_1\{\omega\}}{P_2\{\omega\}} P_2\{\omega\} \\
 &= \sum X(\omega)L(\omega)P_2\{\omega\} = E_2(XL).
 \end{aligned}
]$$

How does this go if Ω is not necessarily discrete? In that case, the set $B = \{\omega : P_2\{\omega\} = 0\}$ need not be negligible under P_2 ; in fact, even in many of the simplest cases of interest B will already be the whole space Ω . For example, if Ω is the real line, then any continuous distribution will put zero probability on each point in Ω . So we can no longer think of $L(\omega)$ as the quotient $P_1\{\omega\}/P_2\{\omega\}$. However, intuitively we can think of $L(\omega)$ as some sort of limit like

$$L(\omega) = \lim_{\Delta\omega \rightarrow 0} \frac{P_1(\Delta\omega)}{P_2(\Delta\omega)}.$$

Whatever that might mean, it certainly looks like some kind of derivative, and in fact the notation dP_1/dP_2 is often used for L . It would seem sensible to think of $\Delta\omega$ roughly as a small set that contains ω . This would allow us, in the limit as $\Delta\omega$ shrinks down to the point ω , to speak of the likelihood ratio *at the point* ω , while at each stage keeping a nondegenerate quotient of positive numbers $P_1(\Delta\omega)$ and $P_2(\Delta\omega)$. If Ω were the Euclidean space \mathbb{R}^d , say, and the probabilities P_1 and P_2 had densities f_1 and f_2 on \mathbb{R}^d , then L would be the ratio of densities

$$L(x_1, \dots, x_d) = \frac{f_1(x_1, \dots, x_d)}{f_2(x_1, \dots, x_d)}.$$

The rigorous foundation for such likelihood ratios in general probability spaces is found in measure theory, in the theory of *Radon-Nikodym derivatives*. In fact, in that context, the requirement that the relation

$$E_1(X) = E_2(XL)$$

hold for all bounded random variables X becomes the *definition* of the likelihood ratio L . The condition for the existence of the likelihood ratio that is analogous to condition (7.2) in the discrete case is that P_1 be *absolutely continuous* with respect to P_2 , that is,

For each event A satisfying $P_2A = 0$ we also have $P_1A = 0$.

7.2 The idea of importance sampling

This idea of changing probability measures using the likelihood ratio is the basis of the technique of *importance sampling* for the Monte Carlo evaluation of integrals and expectations.

Importance sampling is particularly useful in simulations of low probability events. In such cases it can sometimes increase the efficiency of the simulation by many thousands of times over straightforward “naive” sampling. Naive sampling for the given problem would sample from the given probability density, which gave the event of interest low probability. Our estimate of the probability would be the fraction of times the event of interest occurred during our simulations. Unfortunately, since the probability of this event is small, we may well sample many times and not even see one occurrence of the event of interest. This would give an estimated probability of 0, which is not very useful—although it confirms that the probability of interest is small, it gives very little idea of precisely how small the probability is. For example, if the event occurs 0 times in 1000 samples, the probability of interest might be .002 or 10^{-6} or 10^{-9} or whatever. With importance sampling, we sample from a different probability density of our choosing. The idea is to choose a density that concentrates more of its mass on the event of interest, that is, on the more “important” region of the space. Multiplication by the likelihood ratio compensates for the fact that we have changed from the given, desired probability measure to a different measure.

[[More on the problem of low probability events. Just compute the standard deviation of our estimator of p to see the difficulty. The standard deviation is $\sqrt{p(1-p)/n}$, so that, since $\sqrt{p} \gg p$, we need to take n large just to get the standard deviation down to around p , and still much larger in order to make the standard deviation acceptably small relative to p .]]

(7.4) EXAMPLE. Just to see the idea in a simple setting, let’s simulate a simple problem for which we already know the answer: letting $X \sim N(0,1)$, what is the probability that X is greater than 2.5? Let us write this desired probability as $P_0\{X > 2.5\}$, where the subscript 0 is there to remind us that X has mean 0.

From a table of the normal distribution, we see that the answer is 0.0062. Pretending that we do not know this, let us naively simulate the event of interest 100 times. Below is a little Mathematica experiment that generates 100 $N(0,1)$ random variables X_1, \dots, X_{100} and records a 1 for each i such that $X_i > 2.5$ and records a 0 otherwise.

```
list={}
nit=100;
For [it=1,it<=nit,it++,
  x=Random[NormalDistribution[0,1]];
  If [x > 2.5, list=Append[list,1], list=Append[list,0]];
];

list
{0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
  0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
  0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
  0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
  0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0}

mean=Apply[Plus,list]/nit
```

0.

Just as we feared: no successes, and an estimated probability of 0! The problem, of course, is that the $N(0, 1)$ distribution places very little probability on the even $\{X > 2.5\}$. Let us change to the $N(2.5, 1)$ distribution and see what happens. The likelihood ratio of the $N(0, 1)$ distribution with respect to the $N(2.5, 1)$ distribution is the ratio of densities

$$L(x) = \frac{f_0(x)}{f_{2.5}(x)} = \frac{(1/\sqrt{2\pi}) \exp[-x^2/2]}{(1/\sqrt{2\pi}) \exp[-(x-2.5)^2/2]} = \exp[-2.5x + (1/2)(2.5)^2],$$

so that the desired probability is

$$P_0\{X > 2.5\} = E_{2.5}(L(X)I\{X > 2.5\}).$$

To simulate the right-hand side, we generate 100 $N(2.5, 1)$ random variables X_1, \dots, X_{100} , recording $L(X_i)$ when $X_i > 2.5$ and recording 0 otherwise. The expectation in the right-hand side is estimated by the average of the recorded numbers.

```
L[x_] := Exp[-2.5 x + 0.5 (2.5)^2];
list = {};
nit = 100;
For [it = 1, it <= nit, it++,
  x = Random[NormalDistribution[2.5, 1]];
  If [x > 2.5, list = Append[list, L[x]], list = Append[list, 0]];
];
```

```
N[list, 3]
{0.00147, 0., 0.000688, 0., 0.00194, 0.0386, 0.0294, 0., 0.,
 0., 0.00216, 0.023, 0.00911, 0.00619, 0.0222, 0., 0., 0., 0.,
 0., 0.0419, 0., 0.00769, 0.00109, 0.000943, 0.0134, 0., 0.,
 0., 0., 0.0115, 0., 0.0334, 0.0191, 0.00523, 0., 0.00462, 0.,
 0.0065, 0.00294, 0.0319, 0., 0., 0.0282, 0., 0., 0.0241,
 0.0378, 0.00491, 0., 0., 0.0242, 0., 0., 0., 0., 0.00119,
 0., 0.0125, 0., 0.000842, 0.0396, 0.00299, 0.00627, 0.0165,
 0., 0.00115, 0., 0.021, 0.0361, 0., 0.0177, 0., 0., 0.0371,
 0., 0., 0.00164, 0., 0., 0., 0., 0., 0.0148, 0.00703, 0., 0.,
 0., 0.000764, 0., 0., 0.0251, 0.0324, 0., 0., 0., 0., 0.,
 0.00217}
```

```
mean = Apply[Plus, list] / nit
0.00711085
```

```
Sqrt[Apply[Plus, (list - mean)^2] / ((nit - 1) nit)]
0.00119063
```

Observe that we have found an estimate of 0.0071 for the probability, and this estimate has a standard error of 0.0012, which is respectably small for the modest sample size of 100 that we used. Thus, for example, an approximate 95% confidence interval for the unknown probability would be $0.0071 \pm 0.0024 = (0.0047, 0.0095)$. □

7.3 A gambler's ruin problem

Let's consider a gambler's ruin problem for a random walk having normally distributed increments: let X_1, X_2, \dots be *iid* and distributed as $N(\mu_1, 1)$ with $\mu_1 < 0$, and define the random walk $S_n = X_1 + \dots + X_n$. Our problem is to find $P_{\mu_1}\{\tau_b < \infty\}$, where as usual τ_b denotes the first passage time $\inf\{n : S_n > b\}$.

(7.5) EXERCISE. Show that $P_{\mu_1}\{\tau_b < \infty\} \leq e^{2\mu_1 b}$. *[Hint: Note that the right-hand side is a probability for Brownian motion.]*

For $x = (x_1, \dots, x_n) \in \mathbb{R}^n$, we have the likelihood ratio

$$\begin{aligned} \frac{f_{\mu_1}(x_1, \dots, x_n)}{f_{\mu_2}(x_1, \dots, x_n)} &= \frac{\exp[-\frac{1}{2} \sum_1^n (x_i - \mu_1)^2]}{\exp[-\frac{1}{2} \sum_1^n (x_i - \mu_2)^2]} \\ &= \frac{\exp[\mu_1 \sum_1^n x_i - \frac{n}{2} \mu_1^2]}{\exp[\mu_2 \sum_1^n x_i - \frac{n}{2} \mu_2^2]} \\ &= \exp\left[(\mu_1 - \mu_2) \sum_1^n x_i - \frac{n}{2} (\mu_1^2 - \mu_2^2)\right], \end{aligned}$$

so that for $A \subseteq \mathbb{R}^n$

$$\begin{aligned} P_{\mu_1}\{(X_1, \dots, X_n) \in A\} &= \int_{x \in A} f_{\mu_1}(x) dx \\ &= \int_{x \in A} \exp\left[(\mu_1 - \mu_2) \sum_1^n x_i - \frac{n}{2} (\mu_1^2 - \mu_2^2)\right] f_{\mu_2}(x) dx \\ &= E_{\mu_2}\left[e^{(\mu_1 - \mu_2)S_n - n(\mu_1^2 - \mu_2^2)/2} \{(X_1, \dots, X_n) \in A\}\right]. \end{aligned}$$

That is, for any event B that is determined by (X_1, \dots, X_n) , we have

$$(7.6) \quad P_{\mu_1}(B) = E_{\mu_2}\left[e^{(\mu_1 - \mu_2)S_n - n(\mu_1^2 - \mu_2^2)/2} B\right].$$

Things get really interesting when we apply this to stopping times like τ_b . For convenience, let's repeat the definition here.

(7.7) DEFINITION. A nonnegative integer-valued random variable τ is a **stopping time** with respect to the process X_1, X_2, \dots if for each n , the event $\{\tau = n\}$ is a function of the random variables X_1, \dots, X_n .

That is, whether or not the event $\{\tau = n\}$ occurs is determined by the values of the X process only up to time n . We do not have to look at any values of X_m for $m > n$ to determine whether $\tau = n$. Clearly the first passage time τ_b is a stopping time.

To remind ourselves that μ_1 is negative, let us take $\mu_1 = -\mu$, where $\mu > 0$. So our object is to find the probability $P_{-\mu}\{\tau_b < \infty\}$. Since the event $\{\tau_b = n\}$ is determined by (X_1, \dots, X_n) , (7.6) gives

$$\begin{aligned} P_{-\mu}\{\tau_b = n\} &= E_{\mu_2} \left[e^{(-\mu-\mu_2)S_n - n(\mu^2-\mu_2^2)/2} \{\tau_b = n\} \right] \\ &= E_{\mu_2} \left[e^{(-\mu-\mu_2)S_{\tau_b} - \tau_b(\mu^2-\mu_2^2)/2} \{\tau_b = n\} \right]. \end{aligned}$$

Thus,

$$\begin{aligned} P_{-\mu}\{\tau_b < \infty\} &= \sum_{n=1}^{\infty} P_{-\mu}\{\tau_b = n\} \\ &= \sum_{n=1}^{\infty} E_{\mu_2} \left[e^{(-\mu-\mu_2)S_{\tau_b} - \tau_b(\mu^2-\mu_2^2)/2} \{\tau_b = n\} \right] \\ &= E_{\mu_2} \left[e^{(-\mu-\mu_2)S_{\tau_b} - \tau_b(\mu^2-\mu_2^2)/2} \sum_{n=1}^{\infty} \{\tau_b = n\} \right] \\ &= E_{\mu_2} \left[e^{(-\mu-\mu_2)S_{\tau_b} - \tau_b(\mu^2-\mu_2^2)/2} \{\tau_b < \infty\} \right] \end{aligned}$$

We are free to choose whatever value of μ_2 might be convenient. Notice that a nice simplification in the likelihood ratio takes place when $\mu^2 - \mu_2^2 = 0$, in which case the coefficient of τ_b in the exponential vanishes. It does no good to take $\mu_2 = -\mu$, since in that case the last equation reduces to a triviality. Choosing $\mu_2 = \mu$ gives

$$P_{-\mu}\{\tau_b < \infty\} = E_{\mu} \left[e^{-2\mu S_{\tau_b}} \{\tau_b < \infty\} \right] = E_{\mu} \left[e^{-2\mu S_{\tau_b}} \right],$$

where we are allowed to omit the event $\{\tau_b < \infty\}$ in the last equality because the assumption that $\mu > 0$ implies that $P_{\mu}\{\tau_b < \infty\} = 1$. By definition of τ_b , we have $S_{\tau_b} > b$. The difference is the “overshoot” [or “excess over the boundary” or “residual”] $R_b := S_{\tau_b} - b$. That is, the residual R_b is the amount by which the random walk overshoots the level b when it first exceeds b . In terms of this overshoot, we have

$$(7.8) \quad P_{-\mu}\{\tau_b < \infty\} = e^{-2\mu b} E_{\mu} \left[e^{-2\mu R_b} \right].$$

By the way, you showed in Exercise (7.5) that the probability $P_{-\mu}\{\tau_b < \infty\}$ is bounded above by $e^{-2\mu b}$. Since the overshoot R_b is positive, the result (7.8) agrees with this bound, and gives some insight into how tight one might expect the bound to be.

Similarly, I’ll leave it to you to show that for $a < 0 < b$ and $T = \tau_a \wedge \tau_b$ defined as above, we have

$$(7.9) \quad P_{-\mu}\{S_T > b\} = e^{-2\mu b} E_{\mu} \left[e^{-2\mu R_b} \{S_T > b\} \right].$$

Let us give a more general formulation of the above pattern of reasoning. To do this, assume that X_1, X_2, \dots are random variables with X_1, \dots, X_n having joint density f_{1n} and f_{2n} under the two probability measures P_1 and P_2 , respectively. Suppose for simplicity, so that we don't have to worry about dividing by 0, that f_{1n} and f_{2n} are positive functions. Let \mathcal{F}_n denote the set of random variables that are functions of (X_1, \dots, X_n) . Then for $Y = h(X_1, \dots, X_n) \in \mathcal{F}_n$ we have, letting x denote (x_1, \dots, x_n) and dx denote $dx_1 \cdots dx_n$,

$$(7.10) \quad E_1(Y) = \int h(x) f_{1n}(x) dx = \int h(x) \frac{f_{1n}(x)}{f_{2n}(x)} f_{2n}(x) dx = E_2(Y L_n),$$

where

$$L_n = \frac{f_{1n}(X_1, \dots, X_n)}{f_{2n}(X_1, \dots, X_n)}.$$

As in the random walk problems discussed above, things really get interesting when we apply this to stopping times. In terms of the sets of random variables \mathcal{F}_n , we say that τ is a stopping time if for each n the event $\{\tau = n\}$ is in \mathcal{F}_n . Thus, $\{\tau = n\}$ is a function of (X_1, \dots, X_n) , so that whether or not we stop at time n is determined by the random variables X_1, \dots, X_n ; we do not have to look at any of the future values X_{n+1}, X_{n+2}, \dots . Next, we define \mathcal{F}_τ to be the set of random variables Y such that

$$(7.11) \quad Y\{\tau = n\} \in \mathcal{F}_n$$

for all n .

(7.12) EXERCISE. Show that τ is a stopping time if and only if $\{\tau \leq n\} \in \mathcal{F}_n$ for all n , and $Y \in \mathcal{F}_\tau$ if and only if

$$Y\{\tau \leq n\} \in \mathcal{F}_n$$

for all n .

(7.13) PROPOSITION [WALD'S LIKELIHOOD RATIO IDENTITY]. For bounded random variables $Y \in \mathcal{F}_\tau$,

$$E_1[Y\{\tau < \infty\}] = E_2[Y L_\tau\{\tau < \infty\}].$$

PROOF: We have assumed Y bounded just so we can invoke the Bounded Convergence Theorem to eliminate any worries about interchanging sums and expectations. We have

$$\begin{aligned} E_1[Y\{\tau < \infty\}] &= \sum_{n=1}^{\infty} E_1[Y\{\tau = n\}] \\ &= \sum_{n=1}^{\infty} E_2[Y L_n\{\tau = n\}] \\ &= \sum_{n=1}^{\infty} E_2[Y L_\tau\{\tau = n\}] \\ &= E_2[Y L_\tau\{\tau < \infty\}], \end{aligned}$$

where the second equality uses (7.10) together with the definition (7.11) of \mathcal{F}_τ . □

7.4 Importance sampling for the gambler's ruin

Suppose we want to simulate the probability $p = P_{-\mu}\{S_T > b\}$ from (7.9), where either the drift $-\mu$ is very negative or the height b is large or both, so that p is very small. The straightforward “naive” method of simulating this probability would be to simulate many realizations of a random walk with drift $-\mu$, for each realization stopping at the first time T at which the walk either goes below a or above b . Our estimate of p would then be the fraction of our simulated realizations that escaped above b rather than below a . The difficulty with this is that if p is very small, we would have to simulate a very long time before we see even one realization that escapes above b . You know the problem; a very large number of replications would be required to achieve acceptable precision.

However, (7.9) gives us a way to do much better than the naive method. Look at the right-hand side of (7.9). The main contributor to the “smallness” of p is the factor $e^{-2\mu b}$. The remaining expected value can be simulated quite efficiently: since we have changed to the measure P_μ , the simulation would involve realizations of random walks having positive drift μ . This positive drift guarantees that nearly all of the realizations would satisfy $S_T > b$, so that the random variable $e^{-2\mu R_b}\{S_T > b\}$ would be positive. The residual R_b is a nice, moderate quantity, so that in a relatively small number of realizations we would have a good estimate of the expectation $E_\mu[e^{-2\mu R_b}\{S_T > b\}]$.

In the Mathematica experiment below, I simulated the probability (7.9) with $a = -3$, $b = 3$, and a random walk having drift $-\mu = -1$. First I tried the “naive” method, simulating 400 realizations of the random walk and getting a total of zero realizations in which S_T was greater than b . This gives the very uninformative estimator 0 for p —getting no “successes” out of 400 iterations is consistent with having p as large as a few times $1/400$ or so, or anything smaller. The second program performs 100 iterations of the importance sampling method, and gets an estimate of 0.00089 for p , with a standard error of 0.00007. Thus, we have gotten quite a good idea of the size of p from only 100 iterations.

The naive method:

```
mu=1.0;a=-3.0;b=3.0;
nit=400;
nsucc=0;
For [it=1,it<=nit,it++,
  s=0;
  While[s >= a && s <= b,
    s=s+Random[NormalDistribution[-mu,1]]];
  If [s > b, nsucc=nsucc+1];
];

nsucc
0
```

Using importance sampling:


```

mu=1.0;a=-3.0;b=3.0;
list={}
nit=100;
For [it=1,it<=nit,it++,
    s=0;
    While[s >= a && s <= b,
        s=s+Random[NormalDistribution[mu,1]]];
    If [s > b, list=Append[list,Exp[-2.0 mu s]]];
];

N[list,3]
{0.00201, 0.000104, 0.00166, 0.0000459, 0.0000633, 0.00169,
  0.000887, 0.00129, 0.00182, 0.00107, 0.000625, 0.00047,
  0.00096, 0.00165, 0.00052, 0.0000113, 0.000879, 0.00144,
  0.000574, 0.000117, 0.000285, 0.0000792, 0.000217, 0.00212,
  0.00072, 0.00222, 0.00043, 0.000131, 0.000696, 0.000759,
  0.000925, 0.000354, 0.00059, 0.0000381, 0.00014, 0.00231,
  0.00169, 0.000273, 0.00239, 0.000733, 0.00119, 0.00214,
  0.000363, 0.00165, 0.0000509, 0.00235, 0.00128, 0.000355,
  0.00212, 0.00171, 0.00132, 0.000234, 0.000136, 0.000208,
  0.00046, 0.000443, 0.000101, 0.0000684, 0.00064, 0.000994,
  0.000681, 0.000138, 0.00159, 0.00219, 0.00101, 0.000231,
  0.000185, 0.0000257, 0.000591, 0.00146, 0.000864, 0.00193,
  0.00225, 0.00123, 0.00097, 0.000376, 0.00169, 0.00024,
  0.000294, 0.000718, 0.00204, 0.000912, 0.000896, 0.000203,
  0.00203, 0.00236, 0.00144, 0.0000242, 0.000374, 0.0000961,
  0.00016, 0.000254, 0.00105, 0.000102, 0.00131, 0.000327,
  0.00133, 0.00104, 0.00193, 0.0000586}

mean=Apply[Plus,list]/nit
0.000894185

Sqrt[Apply[Plus,(list-mean)^2]/((nit-1)nit)]
0.0000733288

```

7.5 Likelihood ratios for Brownian motion

Next let's see how to change measure with Brownian motion. First we need to see what the likelihood ratio looks like. Here \mathcal{F}_t will denote the set of random variables that are a function of the random variables $W_0^t = \{W_s : 0 \leq s \leq t\}$. Letting P_μ denote the probability distribution making $\{W_t\}$ into a $(\mu, 1)$ Brownian motion, we want to find the likelihood ratio of W_0^t under P_{μ_1} with respect to P_{μ_2} . The idea is to approximate the infinite set of random variables W_0^t by the finite set $(W(t_1), \dots, W(t_n))$ where $0 = t_0 < t_1 < t_2 < \dots < t_n = t$, and then to take the limit as $n \rightarrow \infty$ and the points t_1, \dots, t_n become dense in $(0, t)$. The

likelihood ratio of $(W(t_1), \dots, W(t_n))$ under P_{μ_1} with respect to P_{μ_2} is the ratio of densities

$$\begin{aligned}
 & \frac{P_{\mu_1}\{W(t_1) \in dw_1, \dots, W(t_n) \in dw_n\}}{P_{\mu_2}\{W(t_1) \in dw_1, \dots, W(t_n) \in dw_n\}} \\
 &= \frac{\exp\left\{-\frac{1}{2} \sum_{i=1}^n \frac{1}{t_i - t_{i-1}} (w_i - w_{i-1} - \mu_1(t_i - t_{i-1}))^2\right\}}{\exp\left\{-\frac{1}{2} \sum_{i=1}^n \frac{1}{t_i - t_{i-1}} (w_i - w_{i-1} - \mu_2(t_i - t_{i-1}))^2\right\}} \\
 &= \exp\left\{(\mu_1 - \mu_2) \sum_{i=1}^n (w_i - w_{i-1}) - \frac{1}{2}(\mu_1^2 - \mu_2^2) \sum_{i=1}^n (t_i - t_{i-1})\right\} \\
 &= \exp\left\{(\mu_1 - \mu_2)W(t) - \frac{t}{2}(\mu_1^2 - \mu_2^2)\right\}.
 \end{aligned}$$

Here the situation is particularly simple, since the previous likelihood ratio does not depend on n . Letting $n \rightarrow \infty$ we get that the likelihood ratio of W_0^t under P_{μ_1} with respect to P_{μ_2} is

$$(7.14) \quad L_t = \exp\left\{(\mu_1 - \mu_2)W(t) - \frac{t}{2}(\mu_1^2 - \mu_2^2)\right\}.$$

Analogously with the discrete time case, the nonnegative random variable τ is a stopping time if $\{\tau \leq t\} \in \mathcal{F}_t$ for all t . Also, \mathcal{F}_τ is defined to be the collection of random variables Y such that

$$Y\{\tau \leq t\} \in \mathcal{F}_t$$

for all $t \geq 0$. With these definitions, the following analog of Proposition (7.13) holds.

(7.15) PROPOSITION. For nonnegative $Y \in \mathcal{F}_\tau$,

$$E_{\mu_1}[Y\{\tau < \infty\}] = E_{\mu_2}[YL_\tau\{\tau < \infty\}],$$

where the process $\{L_t\}$ is given by (7.14).

(7.16) EXAMPLE. The previous likelihood ratio identity allows us to derive a beautiful fact about Brownian motion. For a particular example of this fact, consider a Brownian motion W with negative drift $-\mu$. Letting $b > 0$ be large, we know that $P_{-\mu}\{\tau_b < \infty\}$ is only $e^{-2\mu b}$, which is small. But suppose that we are told that the unusual happened: our sample path has $\tau_b < \infty$. What is the conditional distribution of τ_b ? We will see that the answer is that this conditional distribution is exactly the same as the unconditional distribution of τ_b for a Brownian motion having the positive drift $+\mu$!

More generally, for $Y \in \mathcal{F}_{\tau_b}$,

$$\begin{aligned}
 E_{-\mu}[Y\{\tau_b < \infty\}] &= E_\mu[YL_{\tau_b}\{\tau_b < \infty\}] \\
 &= E_\mu[Ye^{-2\mu W(\tau_b)}\{\tau_b < \infty\}] \\
 &= e^{-2\mu b}E_\mu[Y]. \quad [\text{Note } P_\mu\{\tau_b < \infty\} = 1]
 \end{aligned}$$

Notice that in the special case $Y = 1$, the previous equation becomes our old friend $P_{-\mu}\{\tau_b < \infty\} = e^{-2\mu b}$. Thus,

$$(7.17) \quad E_{-\mu}[Y \mid \tau_b < \infty] = \frac{E_{-\mu}[Y\{\tau_b < \infty\}]}{P_{-\mu}\{\tau_b < \infty\}} = \frac{e^{-2\mu b}E_{\mu}[Y]}{e^{-2\mu b}} = E_{\mu}[Y].$$

Now specializing to the choice $Y = \{\tau_b \leq t\}$, we get the promised result

$$P_{-\mu}\{\tau_b \leq t \mid \tau_b < \infty\} = P_{\mu}\{\tau_b \leq t\}.$$

That is, the conditional distribution under drift $-\mu$ of τ_b given that τ_b is finite is the same as the unconditional distribution of τ_b under drift $+\mu$.

In fact, given that τ_b is finite, the path of a Brownian motion with drift $-\mu$ up to time τ_b cannot be probabilistically distinguished from the path of a Brownian motion with drift μ up to time τ_b . This follows as a result of (7.17); what is important is that (7.17) holds *for all* random variables $Y \in \mathcal{F}_{\tau_b}$ that are determined by the path up to time τ_b . For example, we could consider Y to be an event of the form

$$Y = \{W(\tau_b/m) \in A_1, W(2\tau_b/m) \in A_2, \dots, W((m-1)\tau_b/m) \in A_{m-1}, W(\tau_b) \in A_m\},$$

where each A_i is a subset of the real numbers. By taking m large, we can get as detailed a picture as we like of the path up to time τ_b . The result (7.17) tells us that the probabilities of such events coincide for a Brownian motion with drift $-\mu$ that is lucky enough to reach the height b and for a Brownian motion with drift $+\mu$.

This sort of result is useful for large deviations. If b is large then the rare $(-\mu, 1)$ Brownian motion that hits the level b does so in a neighborhood of b/μ that is only of size of order \sqrt{b} , which is negligibly small compared to b . There is a sort of optimization problem that the Brownian motion is doing. Maybe I'll say more about this later... \square

(7.18) EXAMPLE [INVERSE GAUSSIAN DISTRIBUTION]. Here is another example of a useful and explicit Brownian motion formula. We know the distribution of τ_b when the drift $\mu = 0$; we did this by using the reflection principle. How about general μ ? The formula is

$$(7.19) \quad P_{-\mu}\{\tau_b \leq t\} = 1 - \Phi\left(\frac{b + \mu t}{\sqrt{t}}\right) + e^{-2\mu b}\Phi\left(\frac{-b + \mu t}{\sqrt{t}}\right),$$

where Φ is the standard normal distribution function. The distribution of τ_b is called an *Inverse Gaussian distribution*.

In typical calculus-intensive derivations of (7.19), the answer mysteriously appears as the dust settles at the end of the calculation. When I first saw this formula, I found it tantalizing: after all, it just involves two simple normal probabilities together with the very interpretable, familiar factor $e^{-2\mu b}$. Here is a derivation that uses the ideas that we have been discussing; it also is very much in the same spirit as the reflection principle. Start with

$$P_{-\mu}\{\tau_b \leq t\} = P_{-\mu}\{\tau_b \leq t, W_t > b\} + P_{-\mu}\{\tau_b \leq t, W_t \leq b\}.$$

The first term on the right side is simple:

$$P_{-\mu}\{\tau_b \leq t, W_t > b\} = P_{-\mu}\{W_t > b\} = 1 - \Phi\left(\frac{b + \mu t}{\sqrt{t}}\right).$$

Here is a way to get the second term: conditioning on τ_b gives

$$P_{-\mu}\{\tau_b \leq t, W_t \leq b\} = \int_0^t P_{-\mu}\{W_t \leq b \mid \tau_b = s\} P_{-\mu}\{\tau_b \in ds\}.$$

However,

$$\begin{aligned} P_{-\mu}\{W_t \leq b \mid \tau_b = s\} &= P_{-\mu}\{W_t \leq b \mid W_s = b\} \\ &= P\{N[-\mu(t-s), t-s] \leq 0\} \\ &= P\{N[\mu(t-s), t-s] \geq 0\} \\ &= P_{\mu}\{W_t \geq b \mid W_s = b\} \\ &= P_{\mu}\{W_t \geq b \mid \tau_b = s\}, \end{aligned}$$

where the first and last equalities follow from the Markov property. Notice that this is just the sort of symmetry we used in the special case $\mu = 0$ for the reflection principle. Furthermore, by the likelihood ratio identity,

$$P_{-\mu}\{\tau_b \in ds\} = e^{-2\mu b} P_{\mu}\{\tau_b \in ds\}.$$

Thus,

$$\begin{aligned} P_{-\mu}\{\tau_b \leq t, W_t \leq b\} &= e^{-2\mu b} \int_0^t P_{\mu}\{W_t \geq b \mid \tau_b = s\} P_{\mu}\{\tau_b \in ds\} \\ &= e^{-2\mu b} P_{\mu}\{W_t \geq b, \tau_b \leq t\} \\ &= e^{-2\mu b} P_{\mu}\{W_t \geq b\} \\ &= e^{-2\mu b} P_{-\mu}\{W_t \leq -b\} \\ &= e^{-2\mu b} \Phi\left(\frac{-b + \mu t}{\sqrt{t}}\right). \end{aligned}$$

There it is! □

7.6 The Sequential Probability Ratio Test

Let f_0 and f_1 be two given probability density functions. Suppose X_1, X_2, \dots are *iid* and distributed according to the density f , and we are trying to decide whether $f = f_0$ or $f = f_1$. That is, we want to test the null hypothesis $H_0 : f = f_0$ against the alternative hypothesis $H_1 : f = f_1$. Define the *likelihood ratio statistic*

$$L_n = \prod_{i=1}^n \frac{f_1(X_i)}{f_0(X_i)}$$

for each n . If we have decided to take a sample X_1, \dots, X_n of size n and then do a test, the best tests are the *likelihood ratio tests*, which take the following form: choose a number B , and reject H_0 if $L_n > B$, and accept H_0 otherwise. Such tests are optimal in that for each “significance level” α they maximize the power $P_1\{\text{reject } H_0\}$ over all tests that have error probability $P_0\{\text{reject } H_0\} \leq \alpha$. [Here we have introduced the notation P_i for probability when the true density of the observations is f_i .]

Now suppose that we are interested in making a decision between H_0 and H_1 as “quickly” as possible, that is, after looking at as few observations as possible. We might be doing a clinical trial comparing two treatments, and we want to decide as quickly as possible which treatment is better, so that we can give the inferior treatment to as few patients as possible. Another way to describe the problem is to imagine that we are observing the data sequentially, and we want to stop as soon as enough evidence has accumulated in order for us to make a reliable decision. The idea is that, rather than committing to a particular choice of a sample size n before looking at any data, we allow ourselves to look at the observations one by one, and after each observation we decide whether to stop sampling or to take another observation. The extra flexibility this gives us can lead to significant improvements in speed. If we are lucky enough to get some observations that are particularly informative, we can stop early; we don’t force ourselves to continue sampling after already accumulating sufficient evidence to make a reliable decision, merely in order to fulfill some previously specified fixed sample size.

For example, suppose f_0 and f_1 are the normal densities $N(-1/2, 1)$ and $N(1/2, 1)$, respectively. If our first two observations happened to be 4.1 and 5.2, say, we might like to stop sampling and conclude that f_1 is the true density, while if our first two observations were 0.3 and -0.2, we would probably want to take at least one more observation.

This field of study is a part of statistics called *sequential analysis*. See Wald (1947), Siegmund (1985), and Wetherill and Glazebrook (1986), for example. The major development of the field began in World War II, and in fact was apparently kept secret until after the war because it was considered to be so valuable in practice. My memory of the history is fuzzy as you can see, but I believe that Wald (1947) contains interesting discussions.

Hypothesis tests that make use of sequential observations, deciding after each observation whether or not to stop sampling, are called *sequential*. In the situation described above, where we are just trying to decide between two given densities, the optimal tests have the name *sequential probability ratio test*, or “SPRT”. A SPRT is performed by choosing two numbers A and B , sampling until the stopping time

$$T = \inf\{n : L_n < A \text{ or } L_n > B\},$$

rejecting H_0 if $L_T > B$ and accepting H_0 if $L_T < A$. Such tests have a strong optimality property; see Ferguson or Lehmann, for example.

What does this have to do with the gambler's ruin problem? Define $a = \log(A)$, $b = \log(B)$, and

$$S_n = \log(L_n) = \sum_{i=1}^n \log \left(\frac{f_1(X_i)}{f_0(X_i)} \right).$$

Then

$$T = \inf\{n : S_n < a \text{ or } S_n > b\},$$

and the significance level of the test, or probability of a type I error, is

$$(7.20) \quad P_0\{S_T > b\}.$$

Under the probability P_0 , note that S_n is the sum of n iid random variables, that is, the process $\{S_n\}$ is a random walk. Thus, the problem of finding the probability (7.20) is exactly in the form of a gambler's ruin problem.

As an example, let's look briefly at what happens in the normal case, in which case the test takes a particularly simple form. For example, if f_0 and f_1 are the normal densities $N(-1/2, 1)$ and $N(1/2, 1)$, then we have

$$\log \left(\frac{f_1(X_i)}{f_0(X_i)} \right) = \frac{-1}{2} \left(X_i - \frac{1}{2} \right)^2 + \frac{1}{2} \left(X_i + \frac{1}{2} \right)^2 = X_i.$$

Thus, in this case $S_n = \log(L_n)$ is simply the sum $S_n = X_1 + \cdots + X_n$.

Incidentally, the drift of the random walk in (7.20) is

$$E_0 \log \left(\frac{f_1(X_1)}{f_0(X_1)} \right) = -D(f_0 \| f_1) \leq 0,$$

where the *Kullback-Leibler* distance defined by

$$D(f \| g) = E_{X \sim f} \log \left(\frac{f(X)}{g(X)} \right)$$

is nonnegative [a simple consequence of Jensen's inequality; Exercise!]. The negativity of the drift of the random walk under P_0 is fortunate, since it allows us to make the error probability (7.20) small by taking b large. Analogous remarks apply to the other error probability $P_1\{S_T < a\}$.

(7.21) EXERCISE [OPTIMALITY OF THE SPRT]. ...

8. Extremes, Large Deviations, and the Poisson Clumping Heuristic

8.1 The Poisson clumping heuristic

This is an interesting and useful technique for deriving probability approximations. The wonderful reference for this topic is David Aldous' book *Probability Approximations via the Poisson Clumping Heuristic* by (Springer-Verlag, 1989). Aldous has over 100 substantial examples from all different areas of probability, often deriving in a heuristic and intuitive manner and with relatively little effort (a few paragraphs or so) an approximation whose rigorous justification requires a long, difficult paper. I refer you to Aldous' preface for more selling points on the method and for more on the philosophy of approximations.

Let us discuss the method through an example. Let $\{X_t : t \geq 0\}$ be a standard, stationary Ornstein-Uhlenbeck process: $\mu(x) = -x$, $\sigma^2(x) = 2$. What is the probability $P\{\max_{t \leq t_1} X_t \geq b\}$ that the process reaches the level b by time t_1 ? We will think of b as a high level, so that the desired probability will be small.

Consider the random set $\mathcal{S} := \{t : X_t \geq b\}$. In the picture above, \mathcal{S} is the set of highlighted points on the time axis. The first idea in the Poisson clumping heuristic is the key assertion that \mathcal{S} “looks like” or “behaves like” or “has almost the same distribution as” a *mosaic process*. What is a mosaic process? Here is how to make one. Start with an *iid* sequence of random *sets*. For example, the first three such sets might be as shown here.

Let's call the random sets C_1, C_2, C_3, \dots ; they are small, localized “clumps,” one might say. Next, let $0 < Y_1 < Y_2 < Y_3 < \dots$ be the ordered points in a Poisson process with rate λ ,

say. For each i , translate the set C_i by the amount Y_i , and write the resulting translated set as the sum $Y_i + C_i$. The union $\bigcup_{i=1}^{\infty} (Y_i + C_i)$ of the randomly translated random sets is a mosaic process. Let's call each set $Y_i + C_i$ a "clump," and call the point Y_i the "origin" of the clump $Y_i + C_i$. The rate λ is called the *clump rate* of the mosaic process.

Given the suggestive [but not accurately drawn, of course] picture above, I hope you get a feeling for what the clumps are in the Ornstein-Uhlenbeck example and why the key assertion above might make sense. When X first reaches the high level b , that's the origin of the first clump. Recall that for some time after that origin, X behaves locally like a $(-b, 2)$ Brownian motion. In particular, there is a strong negative drift pulling X back down. So it won't be long at all before X is back to being far below b , and probably back around 0, in fact. So, after that first origin, there will be a complicated but very localized set (clump!) of times at which $X_t \geq b$, followed a very much longer stretch of time over which X is below b . Eventually, after a long time, by chance X will reach the high level b again, and we'll get another clump of values t at which $X_t \geq b$. And so on. It makes sense that the time between clump origins should be approximately exponentially distributed with some large mean [and so a small rate]—that the time between clumps should be nearly memoryless makes intuitive sense. If we know that it has been 50 seconds since the last clump, then we suspect that we must be down around 0 somewhere just hovering around, and it will be a long time before we get back up to b . We would think the same way if we were told that it has been 100 seconds or 1000 seconds since the last clump. This is the memoryless property, which characterizes the exponential distribution.

Taking the key assertion to heart, let's heuristically assume that the set $\mathcal{S} = \{t : X_t \geq b\}$ actually *is* a mosaic process with rate λ , where λ is to be determined. Then we would have

$$P\{\max_{t \leq t_1} X_t \geq b\} = P\{\mathcal{S} \cap [0, t_1] \neq \emptyset\} = P\{Y_1 \leq t_1\} = 1 - e^{-\lambda t_1},$$

where Y_1 is the random origin of the first clump. Well, we don't really believe that \mathcal{S} is *exactly* a mosaic process, so we propose the approximation

$$P\{\max_{t \leq t_1} X_t \geq b\} \approx 1 - e^{-\lambda t_1}.$$

Observe at this point that our problem has reduced to determining the clump rate λ . To do this, recall that the X process is stationary, and let π denote the stationary distribution, which is $N(0, 1)$ in this case. The second major idea of the Poisson clumping heuristic is the fundamental relation

$$(8.1) \quad \pi[b, \infty) = \lambda E(C),$$

where $E(C)$ is the expected length of a clump. For example, if a clump happened to be a finite union of intervals, then C would be defined to be the sum of the lengths of those intervals. Why is (8.1) true? Think about what happens over a long stretch of time: we claim that as $t \rightarrow \infty$,

$$(8.2) \quad \pi[b, \infty)t \sim (\lambda t)E(C).$$

To see why this last claim should be true, interpret the stationary probability $\pi[b, \infty)$ as the long-run fraction of time that $X_s \in [b, \infty)$, so that the amount of time the X process

spends in $[b, \infty)$ up to some large time t should be about $\pi[b, \infty)t$. On the other hand, there should be about λt clumps in the time interval $[0, t]$, each of which has an average length of $E(C)$, so that the total length of $\mathcal{S} \cap [0, t]$ should be about $(\lambda t)E(C)$. This argument justifies (8.2), and therefore also (8.1).

Since $\pi[b, \infty) = 1 - \Phi(b)$ is known, by the fundamental relation (8.1), to find λ it is sufficient to find $E(C)$. However, by the local description of the Ornstein-Uhlenbeck process, we would expect to be able to approximate $E(C)$ well by the expected amount of time a $(-b, 2)$ Brownian motion W , started at $W_0 = b$, spends above the level b . In other words, $E(C) \approx ET_{[0, \infty)}$, where $T_{[0, \infty)}$ is the sojourn time in $[0, \infty)$ of a $(-b, 2)$ Brownian motion started at 0. Equivalently, $EC = ET_{(-\infty, 0]}$, where here $T_{(-\infty, 0]}$ is the sojourn time of a $(b/\sqrt{2}, 1)$ Brownian motion started at 0. Thus, by one of the infamous exercises on sojourn times,

$$E(C) \approx \frac{1}{2(b/\sqrt{2})^2} = \frac{1}{b^2}.$$

Now we can put the approximation together:

$$\lambda = \frac{\pi[b, \infty)}{EC} \approx \frac{1 - \Phi(b)}{(1/b^2)} = b^2(1 - \Phi(b)),$$

or $\lambda \approx b\varphi(b)$ if we like, since $1 - \Phi(b) \sim \varphi(b)/b$, which gives

$$(8.3) \quad P\{\max_{t \leq t_1} X_t \geq b\} \approx 1 - e^{-b\varphi(b)t_1}.$$

Here is another, slightly different, way to derive (8.3). Instead of taking $\mathcal{S} = \{t : X_t \geq b\}$ as we did in the preceding derivation, let's take $\mathcal{S} = \{t : X_t \in [b - \delta, b]\}$, where δ is a very small positive number. Again we assert that \mathcal{S} is nearly a mosaic process, with rate λ , say. Again we want to determine λ , since then we will be able to say just as above

$$P\{\max_{t \leq t_1} X_t \geq b\} \approx P\{\mathcal{S} \cap [0, t_1] \neq \emptyset\} \approx 1 - e^{-\lambda t_1}.$$

The “fundamental relation” (8.1) now takes the form

$$\pi[b - \delta, b] = \lambda EC.$$

But $\pi[b - \delta, b] \approx \varphi(b)\delta$, and by the first [even more infamous!] sojourn time homework problem, $EC = \delta/b$, since the drift of X when $X(t) = b$ is $-b$. Thus,

$$\lambda = \frac{\pi[b - \delta, b]}{EC} \approx \frac{\varphi(b)\delta}{(\delta/b)} = b\varphi(b),$$

which again gives (8.3)!

So, there's the Poisson clumping heuristic, applied to the Ornstein-Uhlenbeck process, in two slightly different ways. Now let's go back and look at the neat, one-paragraph description of the idea of the Poisson clumping heuristic that Aldous gives in his preface:

1. Problems about random extrema can often be translated into problems about sparse random sets in $d \geq 1$ dimensions.

2. Sparse random sets often resemble *iid* random clumps thrown down randomly (i.e. centered at points of a Poisson process).
3. The problem of interest reduces to estimating a mean clump size.
4. This mean clump size can be estimated by approximating the underlying random process locally by a simpler, known process for which explicit calculations are possible.

Get it? Here's how it applied to the Ornstein-Uhlenbeck example:

1. The “random extremum” was the maximum $\max_{t \leq t_1} X_t$, and the problem was to approximate the probability $P\{\max_{t \leq t_1} X_t \geq b\}$. The sparse random set was $\mathcal{S} = \{t : X_t \in [b - \delta, b]\}$, say. The “translation” was

$$P\{\max_{t \leq t_1} X_t \geq b\} \approx P\{\mathcal{S} \cap [0, t_1] \neq \emptyset\}.$$

2. This is the key heuristic assertion of the Poisson clumping heuristic: \mathcal{S} is approximately a mosaic process, with some rate λ .
3. Since $P\{\mathcal{S} \cap [0, t_1] \neq \emptyset\} \approx 1 - e^{-\lambda t_1}$, we want to know λ . But the “fundamental relation” gives $\pi[b - \delta, b] = \lambda EC$ and we know $\pi[b - \delta, b] \approx \varphi(b)\delta$, so the problem reduces to estimating the mean clump size EC .
4. The mean clump size EC is estimated by approximating the Ornstein-Uhlenbeck process locally by an appropriate Brownian motion. EC becomes an expected sojourn time. This is simple enough for Brownian motion so that “explicit calculations are possible.”

A. Monotone Convergence, Dominated Convergence, and All That

Suppose that $X_n \rightarrow X$ with probability 1 [that is, $P\{\omega : X_n(\omega) \rightarrow X(\omega)\} = 1$.] The question is: Can we say that $E(X_n) \rightarrow E(X)$? The following results give sufficient conditions.

(A.1) MONOTONE CONVERGENCE THEOREM. *If $X_n \rightarrow X$ with probability 1 and $0 \leq X_1 \leq X_2 \leq \dots$, then $E(X_n) \rightarrow E(X)$.*

Note that $E(X)$ might be infinity in the previous result!

(A.2) BOUNDED CONVERGENCE THEOREM. *If $X_n \rightarrow X$ with probability 1 and there is a finite number b such that $|X_n| \leq b$ for all n , then $E(X_n) \rightarrow E(X)$.*

Here is a generalization of the last result.

(A.3) DOMINATED CONVERGENCE THEOREM. *If $X_n \rightarrow X$ with probability 1 and there is a random variable Y such that $|X_n| \leq Y$ for all n and $E(Y) < \infty$, then $E(X_n) \rightarrow E(X)$.*

Still more generally, we say that a family of random variables $\{X_n\}$ is *uniformly ntegrable* if

$$\lim_{b \rightarrow \infty} \sup_n E[|X_n| \{ |X_n| > b \}] = 0.$$

(A.4) THEOREM. *If $X_n \rightarrow X$ with probability 1 and $\{X_n\}$ is uniformly integrable, then $E(X_n) \rightarrow E(X)$.*

B. Conditioning

B.1 Definitions

In the study of probability, the word “condition” is a verb as often as it a noun. In this appendix, I’d like us to get to the point where it is clear what is meant by a phrase like “condition on X ,” and an assertion like $\mathbb{E}(Y) = \mathbb{E}[\mathbb{E}(Y | X)]$.

Suppose we are interested in the probability of an event A , and X is some discrete random variable. To **condition on** X means to write the probability as

$$(B.1) \quad \mathbb{P}(A) = \sum_x \mathbb{P}(X = x) \mathbb{P}(A | X = x).$$

This is known as the Law of Total Probability, and is just a simple consequence of the definition of the conditional probability

$$(B.2) \quad \mathbb{P}(A | X = x) = \frac{\mathbb{P}[A \cap \{X = x\}]}{\mathbb{P}\{X = x\}}.$$

[[Note: the probability (B.2) is well-defined if $\mathbb{P}\{X = x\} > 0$. The sum in (B.1) can be taken over such values of x that have positive probability.]] We can do the same thing with conditional expectations: from the definition

$$\mathbb{E}(Y | X = x) = \frac{\mathbb{E}[Y I\{X = x\}]}{\mathbb{P}\{X = x\}},$$

it follows that

$$\mathbb{E}(Y) = \sum_x \mathbb{E}[Y I\{X = x\}] = \sum_x \mathbb{P}(X = x) \mathbb{E}(Y | X = x).$$

In the case where the random variable X has a continuous distribution with probability density function f , analogous results replace the sums by integrals:

$$\begin{aligned} \mathbb{P}(A) &= \int \mathbb{P}(A | X = x) f(x) dx, \\ \mathbb{E}(Y) &= \int \mathbb{E}(Y | X = x) f(x) dx. \end{aligned}$$

A general comment: Defining conditional probabilities and expectations of the form $\mathbb{P}(A | X = x)$ and $\mathbb{E}(Y | X = x)$ is a bit subtle when the event $\{X = x\}$ has probability 0. At an elementary level, it is common to treat the discrete case in detail, then state more general results by analogy. That is probably the best way to proceed here also.

The discrete and the continuous cases are special cases of the more general results

$$(B.3) \quad \mathbb{P}(A) = \mathbb{E}[\mathbb{P}(A | X)],$$

$$(B.4) \quad \mathbb{E}(Y) = \mathbb{E}[\mathbb{E}(Y | X)].$$

What do these mean?

Define a function h by $h(x) = \mathbb{E}(Y | X = x)$. For example $h(3) = \mathbb{E}(Y | X = 3)$ and $h(5) = \mathbb{E}(Y | X = 5)$. Now, since h is a perfectly well defined function, we could consider applying h to the random variable X getting a new random variable $h(X)$.

(B.5) DEFINITION. *Given two random variables X and Y , the notation $\mathbb{E}(Y | X)$ stands for the random variable $h(X)$, where the function $h : \mathbb{R} \rightarrow \mathbb{R}$ is defined by $h(x) = \mathbb{E}(Y | X = x)$. So, at the risk of creating complete confusion, I ask: What is $h(X)$? Beware of mechanically substituting X into the definition of h — you will get the incorrect suggestion $h(X) \stackrel{??}{=} \mathbb{E}(Y | X = X)$, which is just the constant number $\mathbb{E}(Y)$, not a random variable. The correct definition $h(X) = \mathbb{E}(Y | X)$ is truly a random variable; for example, it takes on the value $\mathbb{E}(Y | X = 3)$ with probability $\mathbb{P}\{X = 3\}$ and it takes on the value $\mathbb{E}(Y | X = 5)$ with probability $\mathbb{P}\{X = 5\}$. Note the distinction: while $\mathbb{E}(Y | X = x)$ is a number for each x value, $\mathbb{E}(Y | X)$ is a random variable—it is just the random variable $h(X)$ where the function h is as defined above.*

For example, if X and Y have $N(0, 1)$ distributions with correlation ρ , then $\mathbb{E}(Y | X = x) = \rho x$ for all x . So $\mathbb{E}(Y | X) = \rho X$. Also, in the same situation, $\mathbb{E}(Y^2 | X = x) = 1 - \rho^2 + \rho^2 x^2$. So $\mathbb{E}(Y^2 | X) = 1 - \rho^2 + \rho^2 X^2$.

Since $\mathbb{E}(Y | X)$ is a random variable, we can take its expectation: defining $h(x) = \mathbb{E}(Y | X = x)$ as above,

$$\begin{aligned} \mathbb{E}[\mathbb{E}(Y | X)] &= \mathbb{E}[h(X)] = \sum_x h(x) \mathbb{P}\{X = x\} \\ &= \sum_x \mathbb{E}(Y | X = x) \mathbb{P}\{X = x\} = \mathbb{E}(Y). \end{aligned}$$

This is the identity (B.4).

Given the concept of conditional expectation, there is nothing new with conditional probabilities. In fact, we can think of conditional probabilities as special cases of the notion of conditional expectation, where the random variable Y is the indicator of an event A ; that is, $\mathbb{P}(A | X) = \mathbb{E}(I(A) | X)$. But let's give a definition, just to make sure.

(B.6) DEFINITION. *The notation $\mathbb{P}(A | X)$ stands for the random variable $g(X)$, where the function $g : \mathbb{R} \rightarrow \mathbb{R}$ is defined by $g(x) = \mathbb{P}(A | X = x)$.*

The expected value of the random variable $\mathbb{P}(A | X)$ is given by

$$\begin{aligned} \mathbb{E}[\mathbb{P}(A | X)] &= \mathbb{E}[g(X)] = \sum_x g(x) \mathbb{P}\{X = x\} \\ &= \sum_x \mathbb{P}(A | X = x) \mathbb{P}\{X = x\} = \mathbb{P}(A). \end{aligned}$$

Thus, the cryptic-looking identity (B.3) is just a concise way of expressing the law of total probability.

(B.7) EXAMPLE [EXPECTATION OF THE GEOM(p) DISTRIBUTION, BY CONDITIONING]. Let X_1, X_2, \dots be *iid* Bernoulli(p) random variables, that is, $\mathbb{P}\{X_k = 1\} = p = 1 - \mathbb{P}\{X_k = 0\}$. Define $T = \inf\{k : X_k = 1\}$, that is, T is the time of the first “success” (where we are thinking of “1” as success here), and $T \sim \text{Geom}(p)$. Find $\mathbb{E}(T)$ by conditioning on X_1 to write an equation for $\mathbb{E}(T)$ in terms of itself.

SOLUTION. Conditioning on X_1 , we obtain

$$\begin{aligned}\mathbb{E}(T) &= \mathbb{E}[\mathbb{E}(T \mid X_1)] \\ &= \mathbb{E}(T \mid X_1 = 1)\mathbb{P}\{X_1 = 1\} + \mathbb{E}(T \mid X_1 = 0)\mathbb{P}\{X_1 = 0\} \\ &= \mathbb{E}(T \mid X_1 = 1)p + \mathbb{E}(T \mid X_1 = 0)(1 - p).\end{aligned}$$

But $X_1 = 1$ implies $T = 1$, so that $\mathbb{E}(T \mid X_1 = 1) = 1$. Also, if $X_1 = 0$, then it is as if we are restarting the process of waiting for the first success, having already used one trial, so that $\mathbb{E}(T \mid X_1 = 0) = 1 + \mathbb{E}(T)$. Making these substitutions gives

$$\mathbb{E}(T) = p + (1 + \mathbb{E}(T))(1 - p),$$

or $p\mathbb{E}(T) = 1$, or $\mathbb{E}(T) = 1/p$. □

B.2 Summary of some rules

1. If X and Y are independent, then $\mathbb{E}(Y \mid X) = \mathbb{E}(Y)$.

To see this, just observe that $\mathbb{E}(Y \mid X = x) = \mathbb{E}(Y)$ for all x .

2. $\mathbb{E}(g(X)Y \mid X) = g(X)\mathbb{E}(Y \mid X)$ holds for any deterministic function g .

This is intuitive: if we know that value of X , then we also know the value of a function $g(X)$ of X , which then acts like a constant, so that it can be pulled outside the expectation. More formally, $\mathbb{E}(g(X)Y \mid X = x) = \mathbb{E}(g(x)Y \mid X = x) = g(x)\mathbb{E}(Y \mid X = x)$.

3. $\mathbb{E}[\mathbb{E}(Y \mid X)] = \mathbb{E}(Y)$.

(B.8) EXAMPLE. For g a deterministic function, show that $\mathbb{E}[\mathbb{E}(Y \mid X)g(X)] = \mathbb{E}[Yg(X)]$.

SOLUTION:

$$\mathbb{E}[\mathbb{E}(Y \mid X)g(X)] \stackrel{(a)}{=} \mathbb{E}[\mathbb{E}(Yg(X) \mid X)] \stackrel{(b)}{=} \mathbb{E}[Yg(X)],$$

where (a) and (b) are consequences of rules 2 and 3, respectively. □

The result of example (B.8) is important; in some more advanced treatments it is taken as the basis of the definition of conditional expectation.

B.3 Conditional probabilities are probabilities, conditional expectations are expectations

A conditional probability is a probability measure in its own right. For example, if Ω denotes the sample space as usual, then $\mathbb{P}(\Omega \mid X = x) = 1$. Also, if A and B are disjoint events, then $\mathbb{P}(A \cup B \mid X = x) = \mathbb{P}(A \mid X = x) + \mathbb{P}(B \mid X = x)$. Thus, $\mathbb{P}(\Omega \mid X) = 1$, and $\mathbb{P}(A \cup B \mid X) = \mathbb{P}(A \mid X) + \mathbb{P}(B \mid X)$ for disjoint A and B . Formulas that hold for ordinary probabilities and expectations have analogs for conditional probabilities and expectations. For example, Jensen's inequality says that $f(\mathbb{E}(Y)) \leq \mathbb{E}(f(Y))$ if f is a convex function. Replacing each of the expectations “ \mathbb{E} ” in the previous formula by a conditional expectation “ $\mathbb{E}(\cdot \mid X)$ ”, say, we get the conditional form of Jensen's inequality $f(\mathbb{E}(Y \mid X)) \leq \mathbb{E}(f(Y) \mid X)$.

As another example, consider the relation $\mathbb{E}(Y) = \mathbb{E}[\mathbb{E}(Y \mid X)]$. For any random variable Z , we can replace each of the expectations “ \mathbb{E} ” in the previous formula by a conditional expectation “ $\mathbb{E}(\cdot \mid Z)$ ” to obtain the identity

$$(B.9) \quad \mathbb{E}(Y \mid Z) = \mathbb{E}[\mathbb{E}(Y \mid X, Z) \mid Z].$$

This is a conditional analog of rule 3.

Here is a conditional analog of rule 1: If X and Y are independent conditional on Z , then

$$(B.10) \quad \mathbb{E}(Y \mid X, Z) = \mathbb{E}(Y \mid Z).$$

To check this in the discrete case,

$$\begin{aligned} \mathbb{E}(Y \mid X = x, Z = z) &= \frac{\mathbb{E}[Y I\{X = x\} I\{Z = z\}]}{\mathbb{P}\{X = x, Z = z\}} \\ &= \frac{\mathbb{E}[Y I\{X = x\} \mid Z = z]}{\mathbb{P}\{X = x \mid Z = z\}} \\ &= \frac{\mathbb{E}[Y \mid Z = z] \mathbb{P}\{X = x \mid Z = z\}}{\mathbb{P}\{X = x \mid Z = z\}} \\ &= \mathbb{E}(Y \mid Z = z). \end{aligned}$$

(B.11) EXAMPLE. Letting W be a standard Brownian motion, show that $\mathbb{E}(W_s W_t \mid W_u) = \frac{s}{t} \mathbb{E}(W_t^2 \mid W_u)$ for $0 \leq s \leq t \leq u$.

SOLUTION.

$$\begin{aligned} \mathbb{E}(W_s W_t \mid W_u) &\stackrel{(a)}{=} \mathbb{E}[\mathbb{E}(W_s W_t \mid W_t, W_u) \mid W_u] \\ &\stackrel{(b)}{=} \mathbb{E}\left[W_t \mathbb{E}(W_s \mid W_t, W_u) \mid W_u\right] \\ &\stackrel{(c)}{=} \mathbb{E}\left[W_t \mathbb{E}(W_s \mid W_t) \mid W_u\right] \\ &= \mathbb{E}\left[W_t \left(\frac{s}{t} W_t\right) \mid W_u\right] \\ &= \frac{s}{t} \mathbb{E}\left[W_t^2 \mid W_u\right], \end{aligned}$$

where (a) is by (B.9), and (b) uses rule 2 of the list in the previous section to pull the W_t outside the inner conditional expectation. Equality (c) uses (B.10) together with the fact that W_s and W_u are independent conditional on W_t . \square