Contents

Chapter 14
Simulation

Simulation refers to the general strategy of drawing from a simple distribution in order to sample from a more complicated distribution, or to approximate a function. In this chapter we show how simulation can be used to approximate integrals. Our leading example is the problem of computing integrals in Bayesian inference, but the techniques are widely applicable. The simulation methods we discuss include Monte Carlo integration, importance sampling, and Markov chain Monte Carlo (MCMC).

14.1 Introduction

Suppose that we wish to draw a random sample X from a distribution F. Since $F(X)$ is uniformly distributed over the interval $(0,1)$, a basic strategy is to sample $U \sim \text{Uniform}(0,1)$, and then output $X = F^{-1}(U)$. This is an example of *simulation*; we sample from a distribution that is easy to draw from, in this case $\text{Uniform}(0,1)$, and use it to sample from a more complicated distribution F. As another example, suppose that we wish to estimate the integral $\int_0^1 h(x)\,dx$ for some complicated function h. The basic simulation approach is to draw N samples $X_i \sim \text{Uniform}(0,1)$ and estimate the integral as

$$\int_0^1 h(x)\,dx \approx \frac{1}{N}\sum_{i=1}^{N} h(X_i). \tag{14.1}$$

This converges to the desired integral by the law of large numbers.

Simulation methods are especially useful in Bayesian inference, where complicated distributions and integrals are of the essence; let us briefly review the main ideas. Given a prior $\pi(\theta)$ and data $X^n = (X_1, \ldots, X_n)$ the posterior density is

$$\pi(\theta \,|\, X^n) = \frac{\mathcal{L}_n(\theta)\pi(\theta)}{c} \tag{14.2}$$

where $\mathcal{L}_n(\theta)$ is the likelihood function and

$$c = \int \mathcal{L}_n(\theta)\pi(\theta)\, d\theta \tag{14.3}$$

is the *normalizing constant*. The posterior mean is

$$\overline{\theta} = \int \theta\pi(\theta \mid X^n)d\theta = \frac{\int \theta\mathcal{L}_n(\theta)\pi(\theta)d\theta}{c}. \tag{14.4}$$

If $\theta = (\theta_1, \ldots, \theta_k)$ is multidimensional, then we might be interested in the posterior for one of the components, θ_1, say. This marginal posterior density is

$$\pi(\theta_1 \mid X^n) = \int\int \cdots \int \pi(\theta_1, \ldots, \theta_k \mid X^n)d\theta_2 \cdots d\theta_k \tag{14.5}$$

which involves high-dimensional integration. When θ is high-dimensional, it may not be feasible to calculate these integrals analytically. Simulation methods will often be helpful.

14.2 Basic Monte Carlo Integration

Suppose we want to evaluate the integral

$$I = \int_a^b h(x)\, dx \tag{14.6}$$

for some function h. If h is an "easy" function like a polynomial or trigonometric function, then we can do the integral in closed form. If h is complicated there may be no known closed form expression for I. There are many numerical techniques for evaluating I such as Simpson's rule, the trapezoidal rule and Gaussian quadrature. Monte Carlo integration is another approach for approximating I which is notable for its simplicity, generality and scalability.

Begin by writing

$$I = \int_a^b h(x)dx = \int_a^b w(x)f(x)dx \tag{14.7}$$

where $w(x) = h(x)(b-a)$ and $f(x) = 1/(b-a)$. Notice that f is the probability density for a uniform random variable over (a, b). Hence,

$$I = \mathbb{E}_f(w(X)) \tag{14.8}$$

where $X \sim \text{Uniform}(a, b)$. If we generate $X_1, \ldots, X_N \sim \text{Uniform}(a, b)$, then by the law of large numbers

$$\widehat{I} \equiv \frac{1}{N}\sum_{i=1}^{N} w(X_i) \xrightarrow{\text{P}} \mathbb{E}(w(X)) = I. \tag{14.9}$$

This is the basic *Monte Carlo integration* method. We can also compute the standard error of the estimate

$$\widehat{\text{se}} = \frac{s}{\sqrt{N}} \tag{14.10}$$

where

$$s^2 = \frac{\sum_{i=1}^N (Y_i - \widehat{I})^2}{N-1} \tag{14.11}$$

where $Y_i = w(X_i)$. A $1-\alpha$ confidence interval for I is $\widehat{I} \pm z_{\alpha/2}\widehat{\text{se}}$. We can take N as large as we want and hence make the length of the confidence interval very small.

14.12 Example. Let $h(x) = x^3$. Then, $I = \int_0^1 x^3 dx = 1/4$. Based on $N = 10,000$ observations from a Uniform$(0,1)$ we get $\widehat{I} = .248$ with a standard error of .0028. □

A generalization of the basic method is to consider integrals of the form

$$I = \int_a^b h(x) f(x) dx \tag{14.13}$$

where $f(x)$ is a probability density function. Taking f to be a Uniform(a,b) gives us the special case above. Now we draw $X_1, \ldots, X_N \sim f$ and take

$$\widehat{I} \equiv \frac{1}{N} \sum_{i=1}^N h(X_i) \tag{14.14}$$

as before.

14.15 Example. Let

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \tag{14.16}$$

be the standard normal pdf. Suppose we want to compute the cdf at some point x:

$$I = \int_{-\infty}^x f(s) ds = \Phi(x). \tag{14.17}$$

Write

$$I = \int h(s) f(s) ds \tag{14.18}$$

where

$$h(s) = \begin{cases} 1 & s < x \\ 0 & s \geq x. \end{cases} \tag{14.19}$$

Now we generate $X_1, \ldots, X_N \sim N(0,1)$ and set

$$\widehat{I} = \frac{1}{N} \sum_i h(X_i) = \frac{\text{number of observations } \leq x}{N}. \tag{14.20}$$

For example, with $x = 2$, the true answer is $\Phi(2) = .9772$ and the Monte Carlo estimate with $N = 10,000$ yields .9751. Using $N = 100,000$ we get .9771. □

14.21 Example (Bayesian inference for two binomials). Let $X \sim \text{Binomial}(n, p_1)$ and $Y \sim \text{Binomial}(m, p_2)$. We would like to estimate $\delta = p_2 - p_1$. The mle is $\widehat{\delta} = \widehat{p}_2 - \widehat{p}_1 = (Y/m) - (X/n)$. We can get the standard error $\widehat{\text{se}}$ using the delta method, which yields

$$\widehat{\text{se}} = \sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{m}} \tag{14.22}$$

and then construct a 95 percent confidence interval $\widehat{\delta} \pm 2\,\widehat{\text{se}}$. Now consider a Bayesian analysis. Suppose we use the prior $\pi(p_1, p_2) = \pi(p_1)\pi(p_2) = 1$, that is, a flat prior on (p_1, p_2). The posterior is

$$\pi(p_1, p_2 \,|\, X, Y) \propto p_1^X (1-p_1)^{n-X}\, p_2^Y (1-p_2)^{m-Y}. \tag{14.23}$$

The posterior mean of δ is

$$\overline{\delta} = \int_0^1 \int_0^1 \delta(p_1, p_2)\, \pi(p_1, p_2 \,|\, X, Y) = \int_0^1 \int_0^1 (p_2 - p_1)\, \pi(p_1, p_2 \,|\, X, Y). \tag{14.24}$$

If we want the posterior density of δ we can first get the posterior cdf

$$F(c \,|\, X, Y) = P(\delta \leq c \,|\, X, Y) = \int_A \pi(p_1, p_2 \,|\, X, Y) \tag{14.25}$$

where $A = \{(p_1, p_2) :\ p_2 - p_1 \leq c\}$, and then differentiate F. But this is complicated; to avoid all these integrals, let's use simulation.

Note that $\pi(p_1, p_2 \,|\, X, Y) = \pi(p_1 \,|\, X)\, \pi(p_2 \,|\, Y)$ which implies that p_1 and p_2 are independent under the posterior distribution. Also, we see that $p_1 \,|\, X \sim \text{Beta}(X+1, n-X+1)$ and $p_2 \,|\, Y \sim \text{Beta}(Y+1, m-Y+1)$. Hence, we can simulate $(P_1^{(1)}, P_2^{(1)}), \ldots, (P_1^{(N)}, P_2^{(N)})$ from the posterior by drawing

$$P_1^{(i)} \sim \text{Beta}(X+1, n-X+1) \tag{14.26}$$

$$P_2^{(i)} \sim \text{Beta}(Y+1, m-Y+1) \tag{14.27}$$

for $i = 1, \ldots, N$. Now let $\delta^{(i)} = P_2^{(i)} - P_1^{(i)}$. Then,

$$\overline{\delta} \approx \frac{1}{N} \sum_i \delta^{(i)}. \tag{14.28}$$

We can also get a 95 percent posterior interval for δ by sorting the simulated values, and finding the .025 and .975 quantile. The posterior density $f(\delta \,|\, X, Y)$ can be obtained by applying density estimation techniques to $\delta^{(1)}, \ldots, \delta^{(N)}$ or, simply by plotting a histogram. For example, suppose that $n = m = 10$, $X = 8$ and $Y = 6$. From a posterior sample of size 1000 we get a 95 percent posterior interval of $(-0.52, 0.20)$. The posterior density can be estimated from a histogram of the simulated values as shown in Figure 14.1. □

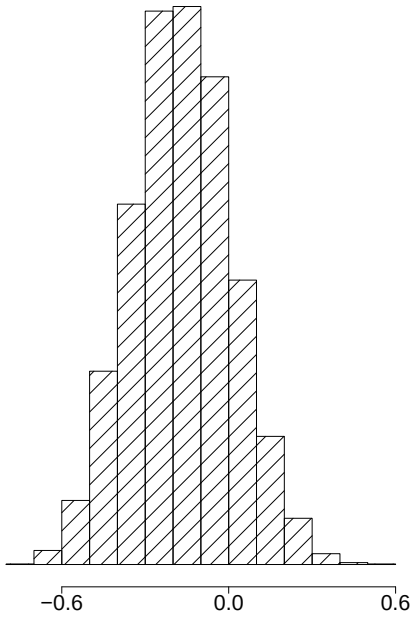

Figure 14.1. *Posterior of δ from simulation.*

14.29 Example (Bayesian inference for dose response). Suppose we conduct an experiment by giving rats one of ten possible doses of a drug, denoted by $x_1 < x_2 < \ldots < x_{10}$. For each dose level x_i we use n rats and we observe Y_i, the number that survive. Thus we have ten independent binomials $Y_i \sim \text{Binomial}(n, p_i)$. Suppose we know from biological considerations that higher doses should have higher probability of death; thus, $p_1 \leq p_2 \leq \cdots \leq p_{10}$. We want to estimate the dose at which the animals have a 50 percent chance of dying—this is called the *LD50*. Formally, $\delta = x_{j^*}$ where

$$j^* = \min\left\{j : \; p_j \geq \tfrac{1}{2}\right\}. \tag{14.30}$$

Notice that δ is implicitly just a complicated function of $p_1, \ldots, p_{10}$ so we can write $\delta = g(p_1, \ldots, p_{10})$ for some g. This just means that if we know $(p_1, \ldots, p_{10})$ then we can find δ. The posterior mean of δ is

$$\int\int \cdots \int_A g(p_1, \ldots, p_{10})\, \pi(p_1, \ldots, p_{10} \,|\, Y_1, \ldots, Y_{10})\, dp_1 dp_2 \ldots dp_{10}. \tag{14.31}$$

The integral is over the region

$$A = \{(p_1, \ldots, p_{10}) : \; p_1 \leq \cdots \leq p_{10}\}. \tag{14.32}$$

The posterior cdf of δ is

$$F(c \,|\, Y_1, \ldots, Y_{10}) = \mathbb{P}(\delta \leq c \,|\, Y_1, \ldots, Y_{10}) \tag{14.33}$$

$$= \int\int \cdots \int_B \pi(p_1, \ldots, p_{10} \,|\, Y_1, \ldots, Y_{10})\, dp_1 dp_2 \ldots dp_{10} \tag{14.34}$$

where

$$B = A \cap \left\{ (p_1, \ldots, p_{10}) : \ g(p_1, \ldots, p_{10}) \leq c \right\}. \tag{14.35}$$

The posterior mean involves a 10-dimensional integral over a restricted region A. We can approximate this integral using simulation.

Let us take a flat prior truncated over A. Except for the truncation, each P_i has once again a Beta distribution. To draw from the posterior we proceed as follows:

(1) Draw $P_i \sim \text{Beta}(Y_i + 1, n - Y_i + 1), i = 1, \ldots, 10$.

(2) If $P_1 \leq P_2 \leq \cdots \leq P_{10}$ keep this draw. Otherwise, throw it away and draw again until you get one you can keep.

(3) Let $\delta = x_{j^*}$ where

$$j^* = \min\{j : \ P_j > \tfrac{1}{2}\}. \tag{14.36}$$

We repeat this N times to get $\delta^{(1)}, \ldots, \delta^{(N)}$ and take

$$\mathbb{E}(\delta \,|\, Y_1, \ldots, Y_{10}) \approx \frac{1}{N} \sum_i \delta^{(i)}. \tag{14.37}$$

Note that δ is a discrete variable. We can estimate its probability mass function by

$$\mathbb{P}(\delta = x_j \,|\, Y_1, \ldots, Y_{10}) \approx \frac{1}{N} \sum_{i=1}^{N} I(\delta^{(i)} = x_j). \tag{14.38}$$

For example, consider the following data:

Dose	1	2	3	4	5	6	7	8	9	10
Number of animals n_i	15	15	15	15	15	15	15	15	15	15
Number of survivors Y_i	0	0	2	2	8	10	12	14	15	14

The posterior draws for $p_1, \ldots, p_{10}$ with $N = 500$ are shown in Figure 14.2. We find that $\overline{\delta} = 5.45$ with a 95 percent interval of (5,7). □

14.3 Importance Sampling

Consider again the integral $I = \int h(x) f(x) dx$ where f is a probability density. The basic Monte Carlo method involves sampling from f. However, there are cases where we may not know how to sample from f. For example, in Bayesian inference, the posterior density is is obtained by multiplying the likelihood $\mathcal{L}_n(\theta)$ times the prior $\pi(\theta)$, and there is generally no guarantee that $\pi(\theta \,|\, x)$ will be a known distribution like a normal or gamma.

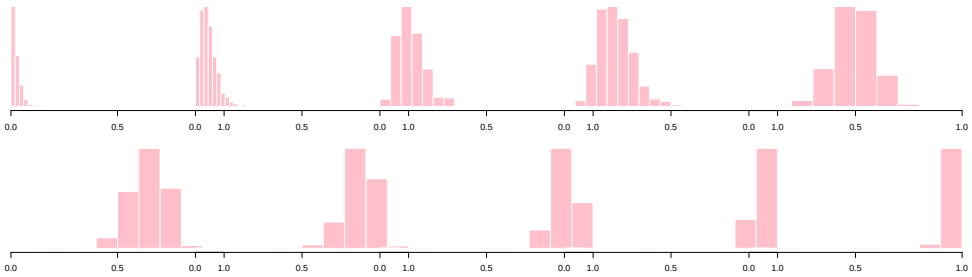

Figure 14.2. *Posterior distributions of the probabilities P_i, $i = 1, \ldots, 10$, for the dose response data of Example 14.29.*

Importance sampling is a generalization of basic Monte Carlo that addresses this problem. Let g be a probability density that we know how to sample from. Then

$$I = \int h(x)f(x)dx = \int \frac{h(x)f(x)}{g(x)}g(x)dx = \mathbb{E}_g(Y) \tag{14.39}$$

where $Y = h(X)f(X)/g(X)$ and the expectation $\mathbb{E}_g(Y)$ is with respect to g. We can simulate $X_1, \ldots, X_N \sim g$ and estimate I by the sample average

$$\widehat{I} = \frac{1}{N}\sum_i Y_i = \frac{1}{N}\sum_i \frac{h(X_i)f(X_i)}{g(X_i)}. \tag{14.40}$$

This is called *importance sampling*. By the law of large numbers, $\widehat{I} \xrightarrow{\text{P}} I$.

There's a catch, however. It's possible that $\widehat{I}$ might have an infinite standard error. To see why, recall that I is the mean of $w(x) = h(x)f(x)/g(x)$. The second moment of this quantity is

$$\mathbb{E}_g(w^2(X)) = \int \left(\frac{h(x)f(x)}{g(x)}\right)^2 g(x)dx = \int \frac{h^2(x)f^2(x)}{g(x)}dx. \tag{14.41}$$

If g has thinner tails than f, then this integral might be infinite. To avoid this, a basic rule in importance sampling is to sample from a density g with thicker tails than f. Also, suppose that $g(x)$ is small over some set A where $f(x)$ is large. Again, the ratio of f/g could be large leading to a large variance. This implies that we should choose g to be similar in shape to f. In summary, a good choice for an importance sampling density g should be similar to f but with thicker tails. In fact, we can say what the optimal choice of g is.

14.42 Theorem. *The choice of g that minimizes the variance of $\widehat{I}$ is*

$$g^*(x) = \frac{|h(x)|f(x)}{\int |h(s)|f(s)ds}. \tag{14.43}$$

Proof. The variance of $w = fh/g$ is

$$\mathbb{E}_g(w^2) - (\mathbb{E}(w^2))^2 = \int w^2(x)g(x)dx - \left(\int w(x)g(x)dx\right)^2 \tag{14.44}$$

$$= \int \frac{h^2(x)f^2(x)}{g^2(x)}g(x)dx - \left(\int \frac{h(x)f(x)}{g(x)}g(x)dx\right)^2 \tag{14.45}$$

$$= \int \frac{h^2(x)f^2(x)}{g^2(x)}g(x)dx - \left(\int h(x)f(x)dx\right)^2. \tag{14.46}$$

The second integral does not depend on g, so we only need to minimize the first integral. From Jensen's inequality (Theorem **??**) we have

$$\mathbb{E}_g(W^2) \geq (\mathbb{E}_g(|W|))^2 = \left(\int |h(x)|f(x)dx\right)^2. \tag{14.47}$$

This establishes a lower bound on $\mathbb{E}_g(W^2)$. However, $\mathbb{E}_{g^*}(W^2)$ equals this lower bound which proves the claim. □

This theorem is interesting but it is only of theoretical interest. If we did not know how to sample from f then it is unlikely that we could sample from $|h(x)|f(x)/\int |h(s)|f(s)ds$. In practice, we simply try to find a thick-tailed distribution g which is similar to $f|h|$.

14.48 Example (Tail probability). Let's estimate $I = \mathbb{P}(Z > 3) = .0013$ where $Z \sim N(0,1)$. Write $I = \int h(x)f(x)dx$ where $f(x)$ is the standard normal density and $h(x) = 1$ if $x > 3$, and 0 otherwise. The basic Monte Carlo estimator is $\widehat{I} = N^{-1}\sum_i h(X_i)$ where $X_1, \ldots, X_N \sim N(0,1)$. Using $N = 100$ we find (from simulating many times) that $\mathbb{E}(\widehat{I}) = .0015$ and $\text{Var}(\widehat{I}) = .0039$. Notice that most observations are wasted in the sense that most are not near the right tail. Now we will estimate this with importance sampling taking g to be a Normal(4,1) density. We draw values from g and the estimate is now $\widehat{I} = N^{-1}\sum_i f(X_i)h(X_i)/g(X_i)$. In this case we find that $\mathbb{E}(\widehat{I}) = .0011$ and $\text{Var}(\widehat{I}) = .0002$. We have reduced the standard deviation by a factor of 20. □

Many variants of the basic importance sampling scheme have been proposed and studied; see, for example Neal (1998) and Southey et al. (1999).

14.4 The Metropolis–Hastings Algorithm

Consider once more the problem of estimating the integral $I = \int h(x)f(x)dx$. Now we introduce Markov chain Monte Carlo (MCMC) methods. The idea is to construct a Markov chain $X_1, X_2, \ldots$, whose stationary distribution is f. Under certain conditions it will then follow that

$$\frac{1}{N}\sum_{i=1}^{N} h(X_i) \xrightarrow{\text{P}} \mathbb{E}_f(h(X)) = I. \tag{14.49}$$

This works because there is a law of large numbers for Markov chains; see Theorem 14.152.

The *Metropolis–Hastings algorithm* is a specific MCMC method that works as follows. Let $q(y \mid x)$ be an arbitrary, "friendly" distribution—that is, we know how to sample efficiently from $q(y \mid x)$. The conditional density $q(y \mid x)$ is called the *proposal distribution*. The Metropolis–Hastings algorithm creates a sequence of observations $X_0, X_1, \ldots$, as follows.

Metropolis–Hastings Algorithm

Choose X_0 arbitrarily.

Given $X_0, X_1, \ldots, X_i$, generate X_{i+1} as follows:

1. Generate a *proposal* or *candidate* value $Y \sim q(y \mid X_i)$.
2. Evaluate $r \equiv r(X_i, Y)$ where
$$r(x, y) = \min\left\{\frac{f(y)}{f(x)}\frac{q(x \mid y)}{q(y \mid x)},\ 1\right\}. \tag{14.50}$$
3. Set
$$X_{i+1} = \begin{cases} Y & \text{with probability } r \\ X_i & \text{with probability } 1 - r. \end{cases} \tag{14.51}$$

A simple way to execute step (3) is to generate $U \sim \text{Uniform}(0, 1)$. If $U < r$ set $X_{i+1} = Y$; otherwise set $X_{i+1} = X_i$. A common choice for $q(y \mid x)$ is $N(x, b^2)$ for some $b > 0$, so that the proposal is draw from a normal, centered at the current value. In this case, the proposal density q is symmetric, $q(y \mid x) = q(x \mid y)$, and r simplifies to

$$r = \min\left\{\frac{f(Y)}{f(X_i)},\ 1\right\}. \tag{14.52}$$

By construction, $X_0, X_1, \ldots$ is a Markov chain. But why does this Markov chain have f as its stationary distribution? Before we explain why, let us first do an example.

14.53 Example. The Cauchy distribution has density

$$f(x) = \frac{1}{\pi}\frac{1}{1 + x^2}. \tag{14.54}$$

Our goal is to simulate a Markov chain whose stationary distribution is f. As suggested in the remark above, we take $q(y \mid x)$ to be a $N(x, b^2)$. So in this case,

$$r(x, y) = \min\left\{\frac{f(y)}{f(x)},\ 1\right\} = \min\left\{\frac{1 + x^2}{1 + y^2},\ 1\right\}. \tag{14.55}$$

So the algorithm is to draw $Y \sim N(X_i, b^2)$ and set

$$X_{i+1} = \begin{cases} Y & \text{with probability } r(X_i, Y) \\ X_i & \text{with probability } 1 - r(X_i, Y). \end{cases} \tag{14.56}$$

The simulator requires a choice of b. Figure 14.3 shows three chains of length $N = 1,000$ using $b = .1$, $b = 1$ and $b = 10$. Setting $b = .1$ forces the chain to take small steps. As a result, the chain doesn't "explore" much of the sample space. The histogram from the sample does not approximate the true density very well. Setting $b = 10$ causes the proposals to often be far in the tails, making r small and hence we reject the proposal and keep the chain at its current position. The result is that the chain "gets stuck" at the same place quite often. Again, this means that the histogram from the sample does not approximate the true density very well. The middle choice avoids these extremes and results in a Markov chain sample that better represents the density sooner. In summary, there are tuning parameters and the efficiency of the chain depends on these parameters. We'll discuss this in more detail later. □

If the sample from the Markov chain starts to look like the target distribution f quickly, then we say that the chain is "mixing well." Constructing a chain that mixes well is somewhat of an art.

14.5 Why It Works

An understanding of why MCMC works requires elementary Markov chain theory, which is reviewed in an Appendix at the end of this chapter.

Recall that a distribution π satisfies *detailed balance* for a Markov chain if

$$p_{ij}\pi_i = p_{ji}\pi_j. \tag{14.57}$$

If π satisfies detailed balance, then it is a stationary distribution for the chain.

Because we are now dealing with continuous state Markov chains, we will change notation a little and write $p(x, y)$ for the probability of making a transition from x to y. Also, let's use $f(x)$ instead of π for a distribution. In this new notation, f is a stationary distribution if $f(x) = \int f(y)p(y, x)\, dy$ and detailed balance holds for f if

$$f(x)p(x, y) = f(y)p(y, x). \tag{14.58}$$

Detailed balance implies that f is a stationary distribution since, if detailed balance holds, then

$$\int f(y)p(y, x)\, dy = \int f(x)p(x, y)\, dy = f(x) \int p(x, y)\, dy = f(x) \tag{14.59}$$

which shows that $f(x) = \int f(y)p(y, x)\, dy$ as required. Our goal is to show that when $p(x, y)$ is the Markov chain defined by the Metropolis-Hastings algorithm, then f satisfies detailed balance, and therefore is a stationary distribution for the chain.

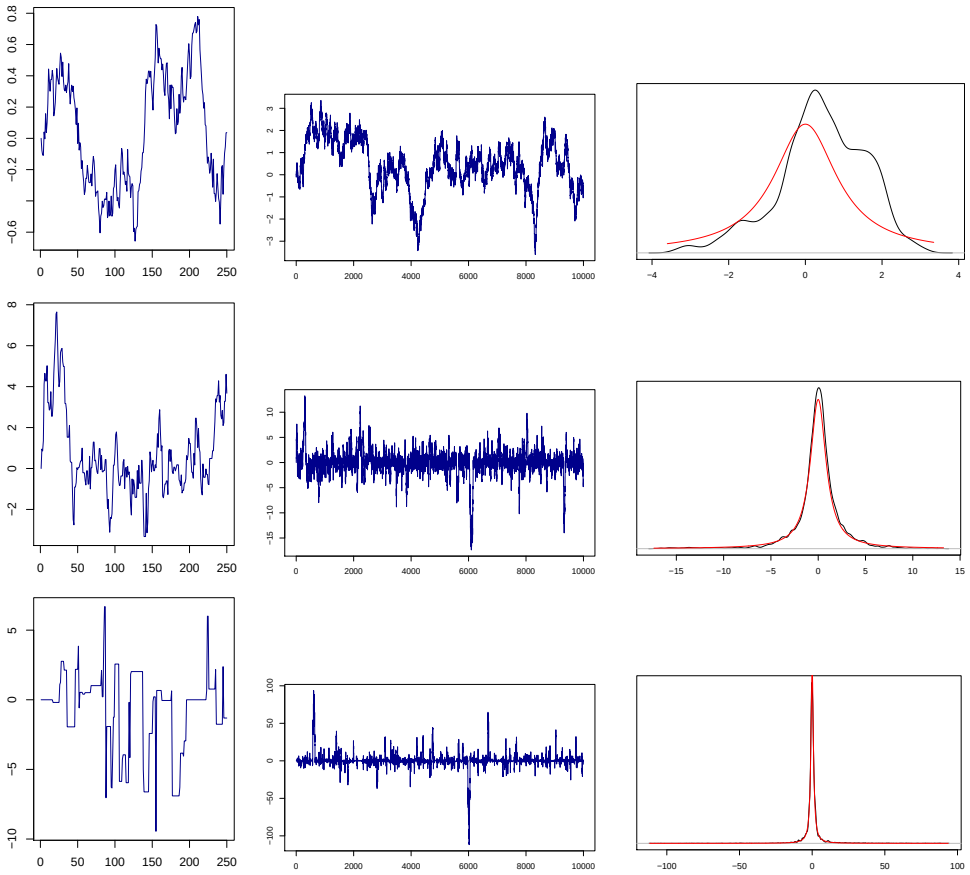

Figure 14.3. *Three Metropolis chains corresponding to* $b = .1$, $b = 1$, $b = 10$, *with acceptance rates 97%, 76%, and 27%, respectively.*

Consider two points x and y. Either

$$f(x)q(y \mid x) < f(y)q(x \mid y) \quad \text{or} \quad f(x)q(y \mid x) > f(y)q(x \mid y). \tag{14.60}$$

We will ignore ties (which occur with probability zero for continuous distributions). Without loss of generality, assume that $f(x)q(y \mid x) > f(y)q(x \mid y)$. This implies that

$$r(x, y) = \frac{f(y)}{f(x)} \frac{q(x \mid y)}{q(y \mid x)} < 1 \tag{14.61}$$

and that $r(y, x) = 1$. Now let $p(x, y)$ be the probability of jumping from x to y. This means that (i) the proposal distribution must generate y, and (ii) you must accept y. Thus,

$$p(x, y) = q(y \mid x)r(x, y) = q(y \mid x) \frac{f(y)}{f(x)} \frac{q(x \mid y)}{q(y \mid x)} = \frac{f(y)}{f(x)} q(x \mid y). \tag{14.62}$$

Therefore,

$$f(x)p(x,y) = f(y)q(x \mid y). \tag{14.63}$$

On the other hand, $p(y,x)$ is the probability of jumping from y to x. This requires two that (i) the proposal distribution must generate x, and (ii) you must accept x. This occurs with probability $p(y,x) = q(x \mid y)r(y,x) = q(x \mid y)$. Hence,

$$f(y)p(y,x) = f(y)q(x \mid y). \tag{14.64}$$

Comparing (14.63) and (14.64), we see that we have shown that detailed balance holds.

14.6 Different Flavors of MCMC

There are different types of MCMC algorithm. Here we will consider a few of the most popular versions.

RANDOM-WALK-METROPOLIS–HASTINGS. In the previous section we considered drawing a proposal Y of the form

$$Y = X_i + \epsilon_i \tag{14.65}$$

where ϵ_i comes from some distribution with density g. In other words, $q(y \mid x) = g(y - x)$. We saw that in this case,

$$r(x,y) = \min\left\{1, \frac{f(y)}{f(x)}\right\}. \tag{14.66}$$

This is called a *random-walk-Metropolis–Hastings* method. The reason for the name is that, if we did not do the accept–reject step, we would be simulating a random walk. The most common choice for g is a $N(0, b^2)$. The hard part is choosing b so that the chain mixes well. As mentioned earlier, a good rule of thumb is to choose b so that about 50 percent of the proposals are accepted.

Note that this method doesn't make sense unless X takes values on the whole real line. If X is restricted to some interval then it is best to transform X. For example, if $X \in (0, \infty)$ then you might take $Y = \log X$ and then simulate the distribution for Y instead of X.

INDEPENDENCE-METROPOLIS–HASTINGS. This is an importance-sampling version of MCMC. We draw the proposal from a fixed distribution g. Generally, g is chosen to be an approximation to f. The acceptance probability becomes

$$r(x,y) = \min\left\{1, \frac{f(y)}{f(x)}\frac{g(x)}{f(y)}\right\} = \min\left\{1, \frac{f(y)}{g(y)}\frac{g(x)}{f(x)}\right\}. \tag{14.67}$$

GIBBS SAMPLING. The two previous methods can be easily adapted, in principle, to work in higher dimensions. In practice, tuning the chains to make them mix well is hard. Gibbs sampling is a way to turn a high-dimensional problem into several one-dimensional problems.

Here's how it works for a bivariate problem. Suppose that (X,Y) has density $f_{X,Y}(x,y)$. First, suppose that it is possible to simulate from the conditional distributions $f_{X \mid Y}(x \mid y)$

and $f_{Y\,|\,X}(y\,|\,x)$. Let (X_0, Y_0) be starting values, and assume we have drawn $(X_0, Y_0), \ldots, (X_n, Y_n)$. Then the Gibbs sampling algorithm for getting (X_{n+1}, Y_{n+1}) is:

Gibbs Sampling

Iterate until convergence:

$$X_{n+1} \sim f_{X\,|\,Y}(x\,|\,Y_n) \tag{14.68}$$
$$Y_{n+1} \sim f_{Y\,|\,X}(y\,|\,X_{n+1}) \tag{14.69}$$

To see that this is a special case of the Metropolis-Hastings algorithm, suppose that the current state is (X_n, Y_n) and the proposal is (X_n, Y), with probability $f_{Y\,|\,X}(Y\,|\,X_n)$. Then the acceptance probability in the Metropolis-Hastings algorithm is

$$r((X_n, Y_n), (X_n, Y)) = \min\left\{1, \frac{f(X_n, Y)}{f(X_n, Y_n)} \frac{f_{Y\,|\,X}(Y_n\,|\,X_n)}{f_{Y\,|\,X}(Y\,|\,X_n)}\right\} \tag{14.70}$$
$$= \min\left\{1, \frac{f(X_n, Y)}{f(X_n, Y_n)} \frac{f(X_n, Y_n)}{f(X_n, Y)}\right\} = 1. \tag{14.71}$$

This generalizes in the obvious way to higher dimensions, where we cycle through the variables, sampling one of them at a time, conditioned on the others.

14.72 Example (Normal hierarchical model). Gibbs sampling is very useful for a class of models called *hierarchical models*. Here is a simple case. Suppose we have a sample of data from k cities. From each city we draw n_i people and observe how many people Y_i have a disease. Thus, $Y_i \sim \text{Binomial}(n_i, p_i)$, allowing for different disease rates in different cities. We can also think of the p_i's as random draws from some distribution F. We can write this model in the following way:

$$P_i \sim F \tag{14.73}$$
$$Y_i\,|\,P_i = p_i \sim \text{Binomial}(n_i, p_i). \tag{14.74}$$

We are interested in estimating the p_i's and the overall disease rate $\int p\, d\pi(p)$.

To proceed, it will simplify matters if we make some transformations that allow us to use some normal approximations. Let $\widehat{p}_i = Y_i/n_i$. Recall that $\widehat{p}_i \approx N(p_i, s_i)$ where $s_i = \sqrt{\widehat{p}_i(1-\widehat{p}_i)/n_i}$. Let $\psi_i = \log(p_i/(1-p_i))$ and define $Z_i \equiv \widehat{\psi}_i = \log(\widehat{p}_i/(1-\widehat{p}_i))$. By the delta method,

$$\widehat{\psi}_i \approx N(\psi_i, \sigma_i^2) \tag{14.75}$$

where $\sigma_i^2 = 1/(n\widehat{p}_i(1-\widehat{p}_i))$. Experience shows that the normal approximation for ψ is more accurate than the normal approximation for p so we shall work with ψ, treating

σ_i as known. Furthermore, we shall take the distribution of the ψ_i's to be normal. The hierarchical model is now

$$\psi_i \sim N(\mu, \tau^2) \tag{14.76}$$

$$Z_i \,|\, \psi_i \sim N(\psi_i, \sigma_i^2). \tag{14.77}$$

As yet another simplification we take $\tau = 1$. The unknown parameters are $\theta = (\mu, \psi_1, \ldots, \psi_k)$. The likelihood function is

$$\mathcal{L}_n(\theta) \propto \prod_i f(\psi_i \,|\, \mu) \prod_i f(Z_i \,|\, \psi_i) \tag{14.78}$$

$$\propto \prod_i \exp\left\{-\frac{1}{2}(\psi_i - \mu)^2\right\} \exp\left\{-\frac{1}{2\sigma_i^2}(Z_i - \psi_i)^2\right\}. \tag{14.79}$$

If we use the prior $f(\mu) \propto 1$ then the posterior is proportional to the likelihood. To use Gibbs sampling, we need to find the conditional distribution of each parameter conditional on all the others. Let us begin by finding $f(\mu \,|\, \text{rest})$ where "rest" refers to all the other variables. We can throw away any terms that don't involve μ. Thus,

$$f(\mu \,|\, \text{rest}) \propto \prod_i \exp\left\{-\frac{1}{2}(\psi_i - \mu)^2\right\} \tag{14.80}$$

$$\propto \exp\left\{-\frac{k}{2}(\mu - b)^2\right\} \tag{14.81}$$

where

$$b = \frac{1}{k}\sum_i \psi_i. \tag{14.82}$$

Hence we see that $\mu \,|\, \text{rest} \sim N(b, 1/k)$. Next we will find $f(\psi \,|\, \text{rest})$. Again, we can throw away any terms not involving ψ_i, leaving us with

$$f(\psi_i \,|\, \text{rest}) \propto \exp\left\{-\frac{1}{2}(\psi_i - \mu)^2\right\} \exp\left\{-\frac{1}{2\sigma_i^2}(Z_i - \psi_i)^2\right\} \tag{14.83}$$

$$\propto \exp\left\{-\frac{1}{2d_i^2}(\psi_i - e_i)^2\right\} \tag{14.84}$$

where

$$e_i = \frac{\dfrac{Z_i}{\sigma_i^2} + \mu}{1 + \dfrac{1}{\sigma_i^2}} \quad \text{and} \quad d_i^2 = \frac{1}{1 + \dfrac{1}{\sigma_i^2}} \tag{14.85}$$

and so $\psi_i \,|\, \text{rest} \sim N(e_i, d_i^2)$. The Gibbs sampling algorithm then involves iterating the following steps N times:

$$\text{draw } \mu \sim N(b, v^2) \tag{14.86}$$

$$\text{draw } \psi_1 \sim N(e_1, d_1^2) \tag{14.87}$$

$$\vdots \quad \vdots \tag{14.88}$$

$$\text{draw } \psi_k \sim N(e_k, d_k^2). \tag{14.89}$$

It is understood that at each step, the most recently drawn version of each variable is used.

We generated a numerical example with $k = 20$ cities and $n = 20$ people from each city. After running the chain, we can convert each ψ_i back into p_i by way of $p_i = e^{\psi_i}/(1 + e^{\psi_i})$. The raw proportions are shown in Figure 14.5. Figure 14.4 shows "trace plots" of the Markov chain for p_1 and μ. Figure 14.5 shows the posterior for μ based on the simulated values. The second panel of Figure 14.5 shows the raw proportions and the Bayes estimates. Note that the Bayes estimates are "shrunk" together. The parameter τ controls the amount of shrinkage. We set $\tau = 1$ but, in practice, we should treat τ as another unknown parameter and let the data determine how much shrinkage is needed. □

So far we assumed that we know how to draw samples from the conditionals $f_{X\,|\,Y}(x\,|\,y)$ and $f_{Y\,|\,X}(y\,|\,x)$. If we don't know how, we can still use the Gibbs sampling algorithm by drawing each observation using a Metropolis–Hastings step. Let q be a proposal distribution for x and let $\widetilde{q}$ be a proposal distribution for y. When we do a Metropolis step for X, we treat Y as fixed. Similarly, when we do a Metropolis step for Y, we treat X as fixed. Here are the steps:

Metropolis within Gibbs

(1a) Draw a proposal $Z \sim q(z\,|\,X_n)$.

(1b) Evaluate

$$r = \min\left\{\frac{f(Z, Y_n)}{f(X_n, Y_n)}\frac{q(X_n\,|\,Z)}{q(Z\,|\,X_n)},\ 1\right\}. \tag{14.90}$$

(1c) Set

$$X_{n+1} = \begin{cases} Z & \text{with probability } r \\ X_n & \text{with probability } 1 - r. \end{cases} \tag{14.91}$$

(2a) Draw a proposal $Z \sim \widetilde{q}(z\,|\,Y_n)$.

(2b) Evaluate

$$r = \min\left\{\frac{f(X_{n+1}, Z)}{f(X_{n+1}, Y_n)}\frac{\widetilde{q}(Y_n\,|\,Z)}{\widetilde{q}(Z\,|\,Y_n)},\ 1\right\}. \tag{14.92}$$

(2c) Set

$$Y_{n+1} = \begin{cases} Z & \text{with probability } r \\ Y_n & \text{with probability } 1 - r. \end{cases} \tag{14.93}$$

Note that in step (1) (and similarly for step (2)), with Y_n fixed, sampling from $f(Z\,|\,Y_n)$

is equivalent to sampling from $f(Z, Y_n)$, as the ratios are identical:

$$\frac{f(Z, Y_n)}{f(X_n, Y_n)} = \frac{f(Z \mid Y_n)}{f(X_n \mid Y_n)}. \tag{14.94}$$

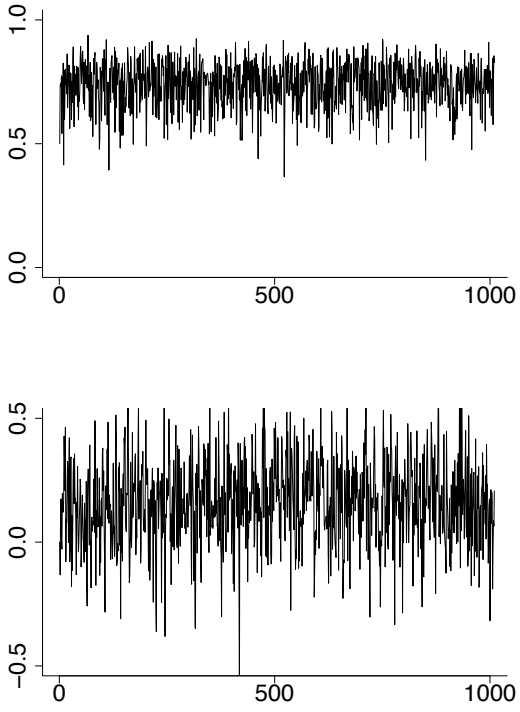

Figure 14.4. *Posterior simulation for Example 14.72. The top panel shows simulated values of p_1. The top panel shows simulated values of μ.*

14.7 Normalizing Constants

The beauty of MCMC is that we avoid having to compute the normalizing constant $c = \int \mathcal{L}_n(\theta)\pi(\theta)d\theta$. But suppose we do want to estimate c. For example, if $\mathcal{M}_1$ and $\mathcal{M}_2$ are two models then

$$\mathbb{P}(\mathcal{M}_1 \mid X_1, \ldots, X_n) = \frac{c_1 p}{c_1 p + c_2(1-p)} \tag{14.95}$$

where p is the prior probability of model 1 and c_1, c_2 are the normalizing constants for the two models. Thus, to do Bayesian model selection requires the normalizing constants.

In general, suppose that f is a probability density function and that

$$f(\theta) = cg(\theta) \tag{14.96}$$

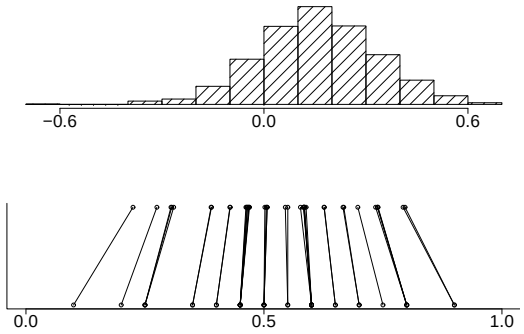

Figure 14.5. *Example 14.72. Top panel: posterior histogram of μ. Lower panel: raw proportions and the Bayes posterior estimates. The Bayes estimates have been shrunk closer together than the raw proportions.*

where $g(\theta) > 0$ is a known function and c is unknown; typically, $g(\theta) = \mathcal{L}_n(\theta)\pi(\theta)$. Let $\theta_1, \ldots, \theta_n$ be a sample from f. Let h be a known probability density function. Define

$$\widehat{c} = \frac{1}{n}\sum_{i=1}^{n} \frac{h(\theta_i)}{g(\theta_i)}. \tag{14.97}$$

Then

$$\mathbb{E}(\widehat{c}) = \int \frac{h(\theta)}{g(\theta)} f(\theta) d\theta = \int \frac{h(\theta)}{g(\theta)} c g(\theta) d\theta = c. \tag{14.98}$$

And if $\int h^2(\theta)/g(\theta)d\theta < \infty$, then $\widehat{c} - c = O_P(n^{-1/2})$.

14.8 Appendix: Basic Markov Chain Theory

A Markov chain is a stochastic process for which the distribution of X_n depends only on X_{n-1}. In this section we assume that the state space is discrete, either $\mathcal{X} = \{1, \ldots, N\}$ or $\mathcal{X} = \{1, 2, \ldots, \}$ and that the index set is $T = \{0, 1, 2, \ldots\}$.

14.99 Definition. *The process $\{X_n : n \in T\}$ is a Markov chain if*

$$\mathbb{P}(X_n = x \mid X_0, \ldots, X_{n-1}) = \mathbb{P}(X_n = x \mid X_{n-1}) \tag{14.100}$$

for all n and for all $x \in \mathcal{X}$.

For a Markov chain, the joint distribution of $X_1, \ldots, X_n$ can be written as

$$f(x_1, \ldots, x_n) = f(x_1) f(x_2 \,|\, x_1) f(x_3 \,|\, x_2) \cdots f(x_n \,|\, x_{n-1}). \tag{14.101}$$

A Markov chain can be represented by the following DAG:

$$X_0 \longrightarrow X_1 \longrightarrow X_2 \longrightarrow \cdots \longrightarrow X_n \longrightarrow \cdots$$

Each variable has a single parent, namely, the previous observation.

The theory of Markov chains is very rich and complex; we have to get through several definitions before we can do anything interesting. Our goal is to answer the following questions: When does a Markov chain "settle down" into some sort of equilibrium? How can we construct Markov chains that converge to a given equilibrium distribution.

To understand the first question, look at the two chains in Figure 14.6. The first chain oscillates all over the place; the second chain eventually settles into an equilibrium. If we constructed a histogram of the first process, it would keep changing as we got more and more observations. But a histogram from the second chain would eventually converge to some fixed distribution.

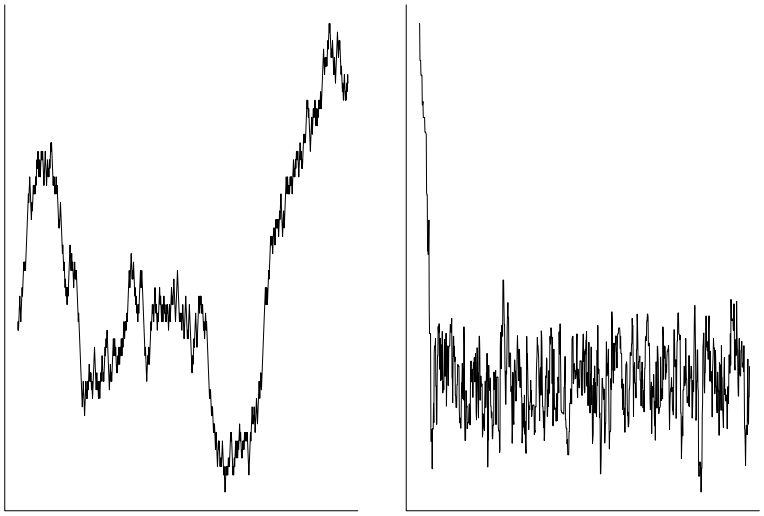

Figure 14.6. *Two Markov chains. The first chain does not settle down into an equilibrium. The second does.*

The key quantities of a Markov chain are the probabilities of jumping from one state into another state. A Markov chain is *homogeneous* if $\mathbb{P}(X_{n+1} = j \,|\, X_n = i)$ does not change with time. Thus, for a homogeneous Markov chain, $\mathbb{P}(X_{n+1} = j \,|\, X_n = i) = \mathbb{P}(X_1 = j \,|\, X_0 = i)$. We shall only deal with homogeneous Markov chains.

14.102 Definition. *We call*

$$p_{ij} \equiv \mathbb{P}(X_{n+1} = j \,|\, X_n = i) \tag{14.103}$$

the transition probabilities. The matrix $\mathbf{P}$ *whose* (i, j) *element is* p_{ij} *is called the transition matrix.*

Notice that $\mathbf{P}$ satisfies (i) $p_{ij} \geq 0$ and (ii) $\sum_i p_{ij} = 1$; thus each row can be regarded as a probability mass function.

14.104 Example (Random walk with absorbing barriers). Let $\mathcal{X} = \{1, \ldots, N\}$. Suppose you are standing at one of these points. Flip a coin with $\mathbb{P}(\text{heads}) = p$ and $\mathbb{P}(\text{tails}) = q = 1 - p$. If it is heads, take one step to the right. If it is tails, take one step to the left. If you hit one of the endpoints, stay there. The transition matrix is

$$\mathbf{P} = \begin{bmatrix} 1 & 0 & 0 & 0 & \cdots & 0 & 0 \\ q & 0 & p & 0 & \cdots & 0 & 0 \\ 0 & q & 0 & p & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 & q & 0 & p \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}. \tag{14.105}$$

□

14.106 Example. Suppose the state space is $\mathcal{X} = \{\text{sunny}, \text{cloudy}\}$. Then X_1, X_2, ... represents the weather for a sequence of days. The weather today clearly depends on yesterday's weather. It might also depend on the weather two days ago but as a first approximation we might assume that the dependence is only one day back. In that case the weather is a Markov chain and a typical transition matrix might be

	Sunny	Cloudy
Sunny	0.4	0.6
Cloudy	0.8	0.2

For example, if it is sunny today, there is a 60 per cent chance it will be cloudy tomorrow. □

Let

$$p_{ij}(n) = \mathbb{P}(X_{m+n} = j \,|\, X_m = i) \tag{14.107}$$

be the probability of of going from state i to state j in n steps. Let $\mathbf{P}_n$ be the matrix whose (i, j) element is $p_{ij}(n)$. These are called the *n-step transition probabilities*.

14.108 Theorem (The Chapman-Kolmogorov equations). *The n-step probabilities satisfy*

$$p_{ij}(m + n) = \sum_k p_{ik}(m) p_{kj}(n). \tag{14.109}$$

Proof. Recall that, in general,

$$\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x)\mathbb{P}(Y = y \,|\, X = x). \tag{14.110}$$

This fact is true in the more general form

$$\mathbb{P}(X = x, Y = y \,|\, Z = z) = \mathbb{P}(X = x \,|\, Z = z)\mathbb{P}(Y = y \,|\, X = x, Z = z). \tag{14.111}$$

Also, recall the law of total probability:

$$\mathbb{P}(X = x) = \sum_y \mathbb{P}(X = x, Y = y). \tag{14.112}$$

Using these facts and the Markov property we have

$$\begin{aligned}
p_{ij}(m+n) &= \mathbb{P}(X_{m+n} = j \,|\, X_0 = i) && (14.113)\\
&= \sum_k \mathbb{P}(X_{m+n} = j, X_m = k \,|\, X_0 = i) && (14.114)\\
&= \sum_k \mathbb{P}(X_{m+n} = j \,|\, X_m = k, X_0 = i)\mathbb{P}(X_m = k \,|\, X_0 = i) && (14.115)\\
&= \sum_k \mathbb{P}(X_{m+n} = j \,|\, X_m = k)\mathbb{P}(X_m = k \,|\, X_0 = i) && (14.116)\\
&= \sum_k p_{ik}(m)p_{kj}(n). && (14.117)
\end{aligned}$$

which gives (14.109). □

Equation (14.109) is nothing more than the equation for matrix multiplication. Hence we have shown that

$$\mathbf{P}_{m+n} = \mathbf{P}_m\mathbf{P}_n. \tag{14.118}$$

By definition, $\mathbf{P}_1 = \mathbf{P}$. Using the above theorem, $\mathbf{P}_2 = \mathbf{P}_{1+1} = \mathbf{P}_1\mathbf{P}_1 = \mathbf{P}\mathbf{P} = \mathbf{P}^2$. Continuing this way, we see that

$$\mathbf{P}_n = \mathbf{P}^n \equiv \underbrace{\mathbf{P} \times \mathbf{P} \times \cdots \times \mathbf{P}}_{\text{multiply the matrix n times}}. \tag{14.119}$$

Let $\mu_n = (\mu_n(1), \ldots, \mu_n(N))$ be a row vector where

$$\mu_n(i) = \mathbb{P}(X_n = i) \tag{14.120}$$

is the marginal probability that the chain is in state i at time n. and μ_0 is called the *initial distribution*. The following procedure is used to simulate a Markov chain:

Markov Chain Simulation

1. Draw $X_0 \sim \mu_0$. Thus, $\mathbb{P}(X_0 = i) = \mu_0(i)$.
2. Iterate:
   (a) Draw $X_{n+1} \sim \mathbf{P}_{X_n}$; thus $\mathbb{P}(X_{n+1} = j \,|\, X_n = i) = p_{ij}$.
   (b) $n \leftarrow n + 1$

To understand the meaning of μ_n, imagine simulating the chain many times, and collecting all the outcomes at time n from all the chains. This histogram would look approximately like μ_n. A consequence of theorem 14.108 is the following:

14.121 Lemma. *The marginal probabilities are given by*

$$\mu_n = \mu_0 \mathbf{P}^n. \tag{14.122}$$

Proof. We have $\mu_n(j) = \mathbb{P}(X_n = j) = \sum_i \mathbb{P}(X_n = j \,|\, X_0 = i) P(X_0 = i) = \sum_i \mu_0(i) p_{ij}(n) = \mu_0 \mathbf{P}^n$. □

Summary of Terminology

1. Transition matrix: $\mathbf{P}(i,j) = \mathbb{P}(X_{n+1} = j \,|\, X_n = i) = p_{ij}$.
2. n-step matrix: $\mathbf{P}_n(i,j) = \mathbb{P}(X_{n+m} = j \,|\, X_m = i)$.
3. $\mathbf{P}_n = \mathbf{P}^n$.
4. Marginal: $\mu_n(i) = \mathbb{P}(X_n = i)$.
5. $\mu_n = \mu_0 \mathbf{P}^n$.

The states of a Markov chain can be classified according to various properties.

14.123 Definition. *We say that i reaches j (or j is accessible from i) if $p_{ij}(n) > 0$ for some n, and we write $i \to j$. If $i \to j$ and $j \to i$ then we write $i \leftrightarrow j$ and we say that i and j communicate.*

14.124 Theorem. *The communication relation satisfies the following properties:*

1. *$i \leftrightarrow i$.*
2. *If $i \leftrightarrow j$ then $j \leftrightarrow i$.*
3. *If $i \leftrightarrow j$ and $j \leftrightarrow k$ then $i \leftrightarrow k$.*
4. *The set of states $\mathcal{X}$ can be written as a disjoint union of classes $\mathcal{X} = \mathcal{X}_1 \cup \mathcal{X}_2 \cup \cdots$ where two states i and j communicate with each other if and only if they are in the same class.*

If all states communicate with each other, then the chain is called *irreducible*. A set of states is *closed* if, once you enter that set of states you never leave. A closed set consisting of a single state is called an *absorbing state*.

14.125 Example. Let $\mathcal{X} = \{1, 2, 3, 4\}$ and

$$\mathbf{P} = \begin{pmatrix} \frac{1}{3} & \frac{2}{3} & 0 & 0 \\ \frac{2}{3} & \frac{1}{3} & 0 & 0 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{14.126}$$

The classes are $\{1, 2\}, \{3\}$ and $\{4\}$. State 4 is an absorbing state. □

Suppose we start a chain in state i. Will the chain ever return to state i? If so, that state is called persistent or recurrent.

14.127 Definition. *State i is recurrent or persistent if*

$$\mathbb{P}(X_n = i \text{ for some } n \geq 1 \mid X_0 = i) = 1. \tag{14.128}$$

Otherwise, state i is transient

14.129 Theorem. *A state i is recurrent if and only if*

$$\sum_n p_{ii}(n) = \infty. \tag{14.130}$$

A state i is transient if and only if

$$\sum_n p_{ii}(n) < \infty. \tag{14.131}$$

Proof. Define

$$I_n = \begin{cases} 1 & \text{if } X_n = i \\ 0 & \text{if } X_n \neq i. \end{cases} \tag{14.132}$$

The number of times that the chain is in state i is $Y = \sum_{n=0}^{\infty} I_n$. The mean of Y, given that the chain starts in state i, is

$$\mathbb{E}(Y \mid X_0 = i) = \sum_{n=0}^{\infty} \mathbb{E}(I_n \mid X_0 = i) = \sum_{n=0}^{\infty} \mathbb{P}(X_n = i \mid X_0 = i) = \sum_{n=0}^{\infty} p_{ii}(n). \tag{14.133}$$

Define $a_i = \mathbb{P}(X_n = i \text{ for some } n \geq 1 \mid X_0 = i)$. If i is recurrent, $a_i = 1$. Thus, the chain will eventually return to i. Once it does return to i, we argue again that since $a_i = 1$, the chain will return to state i again. By repeating this argument, we conclude that $\mathbb{E}(Y \mid X_0 = i) = \infty$. If i is transient, then $a_i < 1$. When the chain is in state i, there is a probability $1 - a_i > 0$ that it will never return to state i. Thus, the probability that the chain is in state i exactly n times is $a_i^{n-1}(1 - a_i)$. This is a geometric distribution which has finite mean. □

14.134 Theorem. *Recurrence satisfies the following properties.*

1. *If state i is recurrent and $i \leftrightarrow j$, then j is recurrent.*
2. *If state i is transient and $i \leftrightarrow j$, then j is transient.*
3. *A finite Markov chain must have at least one recurrent state.*
4. *The states of a finite, irreducible Markov chain are all recurrent.*

14.135 Theorem (Decomposition Theorem). *The state space $\mathcal{X}$ can be written as the disjoint union*

$$\mathcal{X} = \mathcal{X}_T \cup \mathcal{X}_1 \cup \mathcal{X}_2 \cdots \tag{14.136}$$

where $\mathcal{X}_T$ are the transient states and each $\mathcal{X}_i$ is a closed, irreducible set of recurrent states.

14.137 Example (Random walk). Let $\mathcal{X} = \{\ldots, -2, -1, 0, 1, 2, \ldots, \}$ and suppose that $p_{i,i+1} = p$, $p_{i,i-1} = q = 1 - p$. All states communicate, hence either all the states are recurrent or all are transient. To see which, suppose we start at $X_0 = 0$. Note that

$$p_{00}(2n) = \binom{2n}{n} p^n q^n \tag{14.138}$$

since the only way to get back to 0 is to have n heads (steps to the right) and n tails (steps to the left). We can approximate this expression using Stirling's formula which says that

$$n! \sim n^n \sqrt{n} e^{-n} \sqrt{2\pi}. \tag{14.139}$$

Inserting this approximation into (14.138) shows that

$$p_{00}(2n) \sim \frac{(4pq)^n}{\sqrt{n\pi}}. \tag{14.140}$$

It is easy to check that $\sum_n p_{00}(n) < \infty$ if and only if $\sum_n p_{00}(2n) < \infty$. Moreover, $\sum_n p_{00}(2n) = \infty$ if and only if $p = q = 1/2$. By Theorem (14.129), the chain is recurrent if $p = 1/2$ otherwise it is transient. □

Convergence of Markov Chains.

To discuss the convergence of chains, we need a few more definitions. Suppose that $X_0 = i$. Define the *recurrence time*

$$T_{ij} = \min\{n > 0 : \ X_n = j\} \tag{14.141}$$

assuming X_n ever returns to state i, otherwise define $T_{ij} = \infty$. The *mean recurrence time* of a recurrent state i is

$$m_i = \mathbb{E}(T_{ii}) = \sum_n n f_{ii}(n) \tag{14.142}$$

where

$$f_{ij}(n) = \mathbb{P}(X_1 \neq j, X_2 \neq j, \ldots, X_{n-1} \neq j, X_n = j \,|\, X_0 = i). \tag{14.143}$$

A recurrent state is *null recurrent* or *null* if $m_i = \infty$ otherwise it is called *non-null* or *positive*.

14.144 Lemma. *If a state is null and recurrent, then $p_{ii}^n \to 0$.*

14.145 Lemma. *In a finite state Markov chain, all recurrent states are positive.*

Consider a three-state chain with transition matrix

$$\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}. \tag{14.146}$$

Suppose we start the chain in state 1. Then we will be in state 3 at times 3, 6, 9, ... This is an example of a periodic chain. Formally, the *period* of state i is d if $p_{ii}(n) = 0$ whenever n is not divisible by d and d is the largest integer with this property. Thus, $d = \gcd\{n : \ p_{ii}(n) > 0\}$ where gcd means "greater common divisor." State i is *periodic* if $d(i) > 1$ and *aperiodic* if $d(i) = 1$. A state with period 1 is called *aperiodic*.

14.147 Lemma. *If state i has period d and $i \leftrightarrow j$ then j has period d.*

14.148 Definition. *A state is ergodic if it is recurrent, non-null and aperiodic. A chain is ergodic if all its states are ergodic.*

Let $\pi = (\pi_i : \ i \in \mathcal{X})$ be a vector of non-negative numbers that sum to one. Thus π can be thought of as a probability mass function.

14.149 Definition. *We say that π is a stationary (or invariant) distribution if $\pi = \pi\mathbf{P}$.*

Here is some intuition. Draw X_0 from π. Now draw X_1 according to the transition probability of the chain. The distribution of X_1 is then $\mu_1 = \mu_0\mathbf{P} = \pi\mathbf{P} = \pi$. Similarly, the distribution of X_2 is $\pi\mathbf{P}^2 = (\pi\mathbf{P})\mathbf{P} = \pi\mathbf{P} = \pi$. Continuing, we see that the distribution of X_n is $\pi\mathbf{P}^n = \pi$. In other words, If at any time the chain has distribution π, then it will continue to have distribution π forever.

14.150 Definition. *We say that a chain has limiting distribution π if*

$$\mathbf{P}^n \to \begin{bmatrix} \pi \\ \pi \\ \vdots \\ \pi \end{bmatrix} \tag{14.151}$$

for some π, that is, $\pi_j = \lim_{n\to\infty} \mathbf{P}_{ij}^n$ exists and is independent of i.

Here is the main theorem about convergence. The theorem says that an ergodic chain converges to its stationary distribution. Also, sample averages converge to their theoretical expectations under the stationary distribution.

14.152 Theorem. *An irreducible, ergodic Markov chain has a unique stationary distribution π. The limiting distribution exists and is equal to π. If g is any bounded function, then, with probability one,*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} g(X_n) \to \mathbb{E}_\pi(g) \equiv \sum_j \pi_j \, g(j). \tag{14.153}$$

We say that π satisfies *detailed balance* if

$$\pi_i p_{ij} = p_{ji} \pi_j. \tag{14.154}$$

Detailed balance guarantees that π is a stationary distribution.

14.155 Theorem. *If π satisfies detailed balance, then π is a stationary distribution.*

Proof. We need to show that $\pi\mathbf{P} = \pi$. The j^{th} element of $\pi\mathbf{P}$ is $\sum_i \pi_i p_{ij} = \sum_i \pi_j p_{ji} = \pi_j \sum_i p_{ji} = \pi_j$. □

Beware—just because a chain has a stationary distribution does not mean it converges.

14.156 Example. Let

$$\mathbf{P} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}. \tag{14.157}$$

Let $\pi = (1/3, 1/3, 1/3)$. Then $\pi P = \pi$ so π is a stationary distribution. If the chain is started with the distribution π it will stay in that distribution. Imagine simulating many chains and checking the marginal distribution at each time n. It will always be the uniform distribution π. But this chain does not have a limit. It continues to cycle around forever. □

14.158 Example. Let $\mathcal{X} = \{1, 2, 3, 4, 5, 6\}$. Let

$$\mathbf{P} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 \\ \frac{1}{4} & \frac{3}{4} & 0 & 0 & 0 & 0 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & 0 & 0 \\ \frac{1}{4} & 0 & \frac{1}{4} & \frac{1}{4} & 0 & \frac{1}{4} \\ 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix} \tag{14.159}$$

Then $C_1 = \{1, 2\}$ and $C_2 = \{5, 6\}$ are irreducible closed sets. States 3 and 4 are transient because of the path $3 \to 4 \to 6$ and once you hit state 6 you cannot return to 3 or 4. Since $p_{ii}(1) > 0$, all the states are aperiodic. In summary, 3 and 4 are transient while 1, 2, 5, and 6 are ergodic. □

14.160 Example (Hardy-Weinberg). Here is a famous example from genetics. Suppose a gene can be type A or type a. There are three types of people (called genotypes): AA, Aa, and aa. Let (p, q, r) denote the fraction of people of each genotype. We assume that everyone contributes one of their two copies of the gene at random to their children. We also assume that mates are selected at random. The latter is not realistic however, it is often reasonable to assume that you do not choose your mate based on whether they are AA, Aa, or aa. (This would be false if the gene was for eye color and if people chose mates based on eye color.) Imagine if we pooled everyone's genes together. The proportion of A genes is $P = p + (q/2)$ and the proportion of a genes is $Q = r + (q/2)$. A child is AA with probability P^2, aA with probability $2PQ$, and aa with probability Q^2. Thus, the fraction of A genes in this generation is

$$P^2 + PQ = \left(p + \frac{q}{2}\right)^2 + \left(p + \frac{q}{2}\right)\left(r + \frac{q}{2}\right). \tag{14.161}$$

However, $r = 1 - p - q$. Substitute this in the above equation and you get $P^2 + PQ = P$. A similar calculation shows that the fraction of "a" genes is Q. We have shown that the proportion of type A and type a is P and Q and this remains stable after the first generation. The proportion of people of type AA, Aa, aa is thus $(P^2, 2PQ, Q^2)$ from the second generation and on. This is called the Hardy-Weinberg law.

Assume everyone has exactly one child. Now consider a fixed person and let X_n be the genotype of their n^{th} descendant. This is a Markov chain with state space $\mathcal{X} = \{AA, Aa, aa\}$. Some basic calculations will show you that the transition matrix is

$$\begin{bmatrix} P & Q & 0 \\ \frac{P}{2} & \frac{P+Q}{2} & \frac{Q}{2} \\ 0 & P & Q \end{bmatrix}. \tag{14.162}$$

The stationary distribution is $\pi = (P^2, 2PQ, Q^2)$. □

INFERENCE FOR MARKOV CHAINS. Consider a chain with finite state space $\mathcal{X} = \{1, 2, \ldots, M\}$. Suppose we observe n observations $X_1, \ldots, X_n$ from this chain. The unknown parameters of a Markov chain are the initial probabilities $\mu_0 = (\mu_0(1), \mu_0(2), \ldots,)$ and the elements of the transition matrix $\mathbf{P}$. Each row of $\mathbf{P}$ is a multinomial distribution. So we are essentially estimating M distributions (plus the initial probabilities). Let n_{ij} be the observed number of transitions from state i to state j. The likelihood function is

$$\mathcal{L}(\mu_0, \mathbf{P}) = \mu_0(x_0) \prod_{r=1}^{n} p_{X_{r-1}, X_r} = \mu_0(x_0) \prod_{i=1}^{M} \prod_{j=1}^{M} p_{ij}^{n_{ij}}. \tag{14.163}$$

There is only one observation on μ_0 so we can't estimate that. Rather, we focus on estimating $\mathbf{P}$. The mle is obtained by maximizing $\mathcal{L}(\mu_0, \mathbf{P})$ subject to the constraint that the elements are non-negative and the rows sum to 1. The solution is

$$\widehat{p}_{ij} = \frac{n_{ij}}{n_i} \tag{14.164}$$

where $n_i = \sum_{j=1}^{M} n_{ij}$. Here we are assuming that $n_i > 0$. If not, then we set $\widehat{p}_{ij} = 0$ by convention.

14.165 Theorem (Consistency and asymptotic normality of the mle). *Assume that the chain is ergodic. Let $\widehat{p}_{ij}(n)$ denote the mle after n observations. Then $\widehat{p}_{ij}(n) \xrightarrow{\mathrm{P}} p_{ij}$. Also,*

$$\left[\sqrt{N_i(n)}(\widehat{p}_{ij} - p_{ij})\right] \rightsquigarrow N(0, \Sigma) \tag{14.166}$$

where the left-hand side is a matrix, $N_i(n) = \sum_{r=1}^{n} I(X_r = i)$ and

$$\Sigma_{ij,k\ell} = \begin{cases} p_{ij}(1 - p_{ij}) & (i,j) = (k,\ell) \\ -p_{ij}p_{i\ell} & i = k, j \neq \ell \\ 0 & \text{otherwise.} \end{cases} \tag{14.167}$$

14.9 Bibliographic Remarks

MCMC methods go back to the effort to build the atomic bomb in World War II. They were used in various places after that, especially in spatial statistics. There was a new surge of interest in the 1990s that still continues. The main reference for this chapter is Robert and Casella (1999). See also Gelman et al. (2003) and Gilks et al. (1998).

Exercises

14.1 Let

$$I = \int_1^2 \frac{e^{-x^2/2}}{\sqrt{2\pi}} dx. \tag{14.168}$$

(a) Estimate I using the basic Monte Carlo method. Use $N = 100,000$. Also, find the estimated standard error.
(b) Find an (analytical) expression for the standard error of your estimate in (a). Compare to the estimated standard error.
(c) Estimate I using importance sampling. Take g to be $N(1.5, v^2)$ with $v = .1$, $v = 1$ and $v = 10$. Compute the (true) standard errors in each case. Also, plot a histogram of the values you are averaging to see if there are any extreme values.
(d) Find the optimal importance sampling function g^*. What is the standard error using g^*?

14.2 Here is a way to use importance sampling to estimate a marginal density. Let $f_{X,Y}(x, y)$ be a bivariate density and let $(X_1, X_2), \ldots, (X_N, Y_N) \sim f_{X,Y}$.
(a) Let $w(x)$ be an arbitrary probability density function. Let

$$\widehat{f}_X(x) = \frac{1}{N} \sum_{i=1}^{N} \frac{f_{X,Y}(x, Y_i) w(X_i)}{f_{X,Y}(X_i, Y_i)}. \tag{14.169}$$

Show that, for each x,

$$\widehat{f}_X(x) \xrightarrow{p} f_X(x). \tag{14.170}$$

Find an expression for the variance of this estimator.
(b) Let $Y \sim N(0, 1)$ and $X \mid Y = y \sim N(y, 1 + y^2)$. Use the method in (a) to estimate $f_X(x)$.

14.3 Here is a method called *accept–reject sampling* for drawing observations from a distribution.
(a) Suppose that f is some probability density function. Let g be any other density and suppose that $f(x) \leq Mg(x)$ for all x, where M is a known constant. Consider the following algorithm:
(step 1): Draw $X \sim g$ and $U \sim \text{Uniform}(0, 1)$;
(step 2): If $U \leq f(X)/(Mg(X))$ set $Y = X$, otherwise go back to step 1. (Keep repeating until you finally get an observation.)
Show that the distribution of Y is f.
(b) Let f be a standard normal density and let $g(x) = 1/(1 + x^2)$ be the Cauchy density. Apply the method in (a) to draw 1,000 observations from the normal distribution. Draw a histogram of the sample to verify that the sample appears to be normal.

14.4 A random variable Z has a *inverse Gaussian distribution* if it has density

$$f(z) \propto z^{-3/2} \exp\left\{ -\theta_1 z - \frac{\theta_2}{z} + 2\sqrt{\theta_1 \theta_2} + \log\left(\sqrt{2\theta_2}\right) \right\}, \quad z > 0 \tag{14.171}$$

where $\theta_1 > 0$ and $\theta_2 > 0$ are parameters. It can be shown that

$$\mathbb{E}(Z) = \sqrt{\frac{\theta_2}{\theta_1}} \quad \text{and} \quad \mathbb{E}\left(\frac{1}{Z}\right) = \sqrt{\frac{\theta_1}{\theta_2}} + \frac{1}{2\theta_2}. \tag{14.172}$$

(a) Let $\theta_1 = 1.5$ and $\theta_2 = 2$. Draw a sample of size 1,000 using the independence-Metropolis–Hastings method. Use a Gamma distribution as the proposal density. To assess the accuracy, compare the mean of Z and $1/Z$ from the sample to the theoretical means Try different Gamma distributions to see if you can get an accurate sample.
(b) Draw a sample of size 1,000 using the random-walk-Metropolis–Hastings method. Since $z > 0$ we cannot just use a normal density. One strategy is this. Let $W = \log Z$. Find the density of W. Use the random-walk-Metropolis–Hastings method to get a sample $W_1, \ldots, W_N$ and let $Z_i = e^{W_i}$. Assess the accuracy of the simulation as in part (a).

14.5 Get the heart disease data from the book web site. Consider a Bayesian analysis of the logistic regression model

$$\mathbb{P}(Y = 1 \,|\, X = x) = \frac{e^{\beta_0 + \sum_{j=1}^k \beta_j x_j}}{1 + e^{\beta_0 + \sum_{j=1}^k \beta_j x_j}}. \tag{14.173}$$

Use the flat prior $f(\beta_0, \ldots, \beta_k) \propto 1$. Use the Gibbs–Metropolis algorithm to draw a sample of size 10,000 from the posterior $f(\beta_0, \beta_1 \,|\, \text{data})$. Plot histograms of the posteriors for the β_j's. Get the posterior mean and a 95 percent posterior interval for each β_j.
(b) Compare your analysis to a frequentist approach using maximum likelihood.

Bibliography

GELMAN, A., CARLIN, J. B., STERN, H. S. and RUBIN, D. B. (2003). *Bayesian Data Analysis*. Chapman & Hall, CRC Press. Second edition.

GILKS, W. R., RICHARDSON, S. and SPIEGELHALTER, D. J. (1998). *Markov chain Monte Carlo in practice*. Chapman & Hall Ltd.

NEAL, R. M. (1998). Annealed importance sampling. *Statistics and Computing* **11** 125–139.

ROBERT, C. P. and CASELLA, G. (1999). *Monte Carlo statistical methods*. Springer-Verlag Inc.

SOUTHEY, F., SCHUURMANS, D. and GHODSI, A. (1999). Regularized greedy importance sampling. In *Advances in Neural Information Processing Systems 12*. MIT Press.

Index