Markov Chain Monte Carlo Lecture Notes

Charles J. Geyer

Course notes originally used Spring Quarter 1998
Last changed: November 21, 2005
Last typeset: November 21, 2005

Contents

Chapter 1

Introduction

1.1 Monte Carlo

Monte Carlo is a cute name for learning about probability models by simulating them, Monte Carlo being the location of a famous gambling casino. A half century of use as a technical term in statistics, probability, and numerical analysis has drained the metaphor of its original cuteness. Everybody uses "Monte Carlo" as the only technical term describing this method.

Whenever we can simulate a random process, we can calculate probabilities and expectations by averaging over the simulations. This means we can handle any calculation we might want to. If we can't do pencil and paper calculations deriving closed-form expressions for the quantities we want, we can always use brute force computation. The Monte Carlo method may not be as elegant as the pencil and paper method, and it may not give as much insight into the problem, but it applies to any random process we can simulate, and we shall see that we can simulate almost any random process. Pencil and paper methods are nice when they work, but they only apply to a small set of simple, computationally convenient probability models. Monte Carlo brings a huge increase in the models we can handle.

Suppose X_1, X_2, ... are a sequence of independent, identically distributed (i. i. d.) simulations of some probability model. Let X denote a generic realization of the model, so all of the X_i have the same distribution as X. We want to calculate the expectation of some random variable $g(X)$. If we can do it by pencil and paper calculations, fine. If not, we use Monte Carlo. Write the expectation in question as $\mu = E\{g(X)\}$. The *Monte Carlo approximation* of μ is the sample average over the simulations

$$\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} g(X_i). \tag{1.1}$$

Since $\hat{\mu}_n$ is the sample mean of i. i. d. random variables $g(X_1)$, ..., $g(X_n)$ having expectation μ, the strong law of large numbers (SLLN) says $\hat{\mu}_n$ converges

almost surely to μ as the number of simulations goes to infinity,

$$\hat{\mu}_n \xrightarrow{\text{a. s.}} \mu, \qquad n \to \infty. \tag{1.2}$$

Furthermore, if $\operatorname{Var}\{g(X)\}$ is finite, say σ^2, then the central limit theorem (CLT) says $\hat{\mu}_n$ is asymptotically normal with mean μ and variance σ^2/n,

$$\sqrt{n}(\hat{\mu}_n - \mu) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2).$$

The nice thing for statisticians about Monte Carlo is that we already understand the theory. It is just elementary statistics.

All of this applies to calculating probabilities as well as expectations, because probabilities are expectations of indicator functions.

Example 1.1. Evaluating a Probability.
If X and Y are independent normal random variables with mean zero and the same variance, what is $P(Y < X^2)$? We could do this by numerical integration

$$\mu = \int \Phi(x^2)\phi(x)\, dx$$

where ϕ is the standard normal probability density function and Φ is the standard normal distribution function (Mathematica gives $\mu = 0.719015$), but we will pretend we can't and use Monte Carlo.

We generate a large number n of pairs (X_i, Y_i) of independent standard normal random variables. Then $\hat{\mu}_n$ is the proportion of pairs having $Y_i < X_i^2$. The first time I tried this with $n = 1000$, I got $\hat{\mu}_n = 0.700$.

We do not know σ^2 but can, as everywhere else in statistics, estimate it by the sample variance, which when we are estimating a probability has the binomial form $p(1-p)$ where $p = \hat{\mu}_n$. Thus for $\hat{\mu}_n = 0.700$ we get an estimate of $\sigma/\sqrt{n}$ of $\sqrt{0.7 \times 0.3/1000} = 0.0145$.

So we find that statistics works (no surprise). The standard error (SE) calculation says that our Monte Carlo approximation 0.700 is about one SE, 0.0145 more or less, from the exact answer, and it is just a little over one SE low.

In order to avoid confusion we call n the *Monte Carlo sample size* when it is necessary to distinguish it from some other "sample size" involved in the problem. Often in statistics, the random process X we are simulating is a model for data. If X is a vector of length m, the usual terminology of statistics says we have sample size m. Calling n the Monte Carlo sample size avoids confusion between m and n.

Similarly we call the standard error of the Monte Carlo approximation the *Monte Carlo standard error* (MCSE) in order to distinguish it from any other "standard error" involved in the problem. It my be that the very thing we are trying to calculate by Monte Carlo is the standard error of a parameter estimate or a related quantity like Fisher information. Then the term MCSE avoids confusion.

1.2 Problems with Ordinary Monte Carlo

The main problem with ordinary independent-sample Monte Carlo is that it is very hard to do for multivariate random quantities. A huge number of methods exist for simulating univariate random quantities. Devroye (1986) is the definitive source. Ripley (1987) is more introductory but is authoritative as far as it goes. Knuth (1998) is also authoritative, though oriented more toward computer science than statistics.

There are a few tricks for reducing multivariate problems to univariate problems. A general multivariate normal random vector $X \sim \mathcal{N}(\mu, \Sigma)$ can be simulated using the Cholesky decomposition of the variance matrix $\Sigma = LL^T$. Let Z be a $N(0, I)$ random vector (each component is standard normal and the components are independent). Then $X = \mu + LZ$ has the desired $N(\mu, \Sigma)$ distribution (Ripley 1987, p. 98). Wishart distributions can also be simulated (Ripley 1987, p. 99–100). There are a few other special cases in which independent simulations of a multivariate process are possible, but not many.

One general method that has occurred to many people is to use the laws of conditional probability. Simulate the first component using its marginal distribution, simulate the second component using its conditional distribution given the first, then simulate the third component using its conditional distribution given the first two, and so forth. The sad fact is that this is almost never useful, because the required marginal and conditional distributions are not known and cannot be used for simulation.

In summary, ordinary independent-sample Monte Carlo is not useful for most multivariate random quantities. Something better is needed.

1.3 Stochastic Processes

A *discrete-time stochastic process* is the same as what is called a *random sequence* in Fristedt and Gray (1997, Chapter 22). It is a sequence X_1, X_2, ... of random elements of some fixed set called the *state space* of the stochastic process. A specific familiar example is a sequence of i. i. d. random variables.

The point of calling this object a "random sequence" or "stochastic process" is to serve as a reminder that the entire sequence can be considered a random object. A familiar example of this is the SLLN (1.2), which can be rewritten

$$\Pr(\hat{\mu}_n \to \mu) = 1$$

where the probability refers to the whole infinite sequence. This is a measure-theoretic technicality that will play only a very minor role in our study of Markov chains. It is a theorem of measure-theoretic probability that the probability law of a "random sequence" contains no more information than the so-called "finite-dimensional distributions," meaning the distributions of finite segments of the sequence $X_1, \ldots, X_n$. The probability law of the infinite sequence, thought of as an infinite vector $(X_1, X_2, \ldots)$, determines the joint distribution of the vector $(X_1, \ldots, X_n)$ for each n, and vice versa: the finite-dimensional distributions

collectively determine the probability law of the infinite sequence. Thus mostly finite-dimensional distributions are enough.

A *continuous-time stochastic process* is a set of random variables X_t indexed by a continuous variable, say $t \in [0, 1]$. An example is Brownian motion (Fristedt and Gray 1997, Chapter 19). These have not played much role in Markov chain Monte Carlo, and we shall ignore them.

1.4 Markov Chains

In this course, the term *Markov chain* refers to a discrete-time stochastic process on a general state space that has the Markov property: the future is independent of the past given the present state. This follows one of the two conflicting standard usages of the term "Markov chain." Older Markov chain literature (Chung 1967) uses "Markov chain" to refer to a discrete-time *or continuous-time* stochastic process on a *countable* state space that satisfies the Markov property. The limitation to a countable state space would rule out most of the interesting applications. Thus much of the modern Markov chain literature (Nummelin 1984; Meyn and Tweedie 1993) and all of the Markov chain Monte Carlo (MCMC) literature follows the usage adopted here.

So to repeat our definition with more specificity, a *Markov chain* is a discrete-time stochastic process $X_1, X_2, \ldots$ taking values in an arbitrary state space and having the property that the conditional distribution of X_{n+1} given the past, $X_1, \ldots, X_n$, depends only on the present state X_n. Following Nummelin (1984) and Meyn and Tweedie (1993) and all of the MCMC literature, we will further restrict the term "Markov chain" to refer to a Markov chain *with stationary transition probabilities,* that is, the conditional distribution of X_{n+1} given X_n is the same for all n.

The specification of a Markov chain model has two pieces, the initial distribution and the transition probabilities. The *initial distribution* is the marginal distribution of X_1. The *transition probabilities* specify the conditional distribution of X_{n+1} given X_n. Since we always assume stationary transition probabilities, this is just one conditional distribution, the same for all n.

By mathematical induction, these two pieces determine the marginal distribution of $X_1, \ldots, X_n$ for any n. The base of the induction is obvious, the marginal distribution of X_1 is the initial distribution. Assuming the distribution of $X_1, \ldots, X_{n-1}$ is known, the distribution of $X_1, \ldots, X_n$ is determined by the usual

$$\text{joint} = \text{conditional} \times \text{marginal}$$

formula when densities exist, where "marginal" refers to the distribution of $X_1, \ldots, X_{n-1}$, "joint" refers to the distribution of $X_1, \ldots, X_n$, and "conditional" refers to the distribution of X_n given $X_1, \ldots, X_{n-1}$, which by the Markov property depends on X_{n-1} alone and is the specified transition probability. A more general proof that does not depend on the existence of densities will be given later after we have developed the required notation.

Example 1.2. AR(1) Time Series.
An *AR(1) time series* is a stochastic process X_1, X_2, ... with state space $\mathbb{R}$ defined recursively by

$$X_n = \rho X_{n-1} + e_n \tag{1.3}$$

where e_1, e_2, ... are i. i. d. $\mathcal{N}(0, \tau^2)$ and where ρ and τ^2 are real numbers that are parameters of the model. The distribution of X_1 may be specified arbitrarily.

It is easy to see that this stochastic process is a Markov chain (with stationary transition probabilities). The conditional distribution of X_n given X_1, ..., X_{n-1} is $\mathcal{N}(\rho X_{n-1}, \tau^2)$, which is the same as the conditional distribution conditioning on X_{n-1} only. Thus the process has the Markov property. Since the conditional distribution of X_n given X_{n-1} is the same for all n, the process has stationary transition probabilities. If the last point is not clear, perhaps different notation will help. The conditional distribution of X_n given $X_{n-1} = x$ is $\mathcal{N}(\rho x, \tau^2)$, and it is now clear that this does not depend on n.

For those who are curious about the name, "AR(1)" stands for autoregressive of order one. Equation (1.3) looks like the specification of a regression model except that the same variables occur on both sides of the equation at different times, thus the "auto-" to indicate this.

An AR(k) time series is defined by the recursion

$$X_n = \rho_1 X_{n-1} + \rho_2 X_{n-2} + \cdots + \rho_k X_{n-k} + e_n. \tag{1.4}$$

It is clear that this is *not* a Markov chain, because the conditional distribution of X_n given the past depends on X_{n-k}, ..., X_{n-1} rather than just on X_{n-1}.

An AR(k) time series can be turned into a Markov chain by redefining the state space. Consider the stochastic process Y_1, Y_2, ... with state space $\mathbb{R}^k$ defined by

$$Y_n = \begin{pmatrix} X_n \\ X_{n+1} \\ X_{n+2} \\ \vdots \\ X_{n+k-1} \end{pmatrix}$$

where the X_i form an AR(k) process. The new process is Markov, since the conditional distribution of Y_n given Y_1, ..., Y_{n-1} depends only on Y_{n-1}. It also obviously has stationary transition probabilities.

This is a special case of a vector-valued AR(1) time series with state space $\mathbb{R}^k$, defined by

$$Y_n = AY_{n-1} + e_n \tag{1.5}$$

where now Y_n and e_n are vectors in $\mathbb{R}^k$ with e_1, e_2, ... are i. i. d. $\mathcal{N}(0, M)$ and where and A is a linear transformation from $\mathbb{R}^k$ to $\mathbb{R}^k$, which can be represented by a $k \times k$ matrix, and M is also a $k \times k$ matrix, the covariance matrix of the vectors e_i. The initial distribution (the distribution of Y_1) can be specified arbitrarily. The scalar parameter ρ in (1.3) corresponds to the matrix A in (1.5), and the scalar parameter τ^2 in (1.3) corresponds to the covariance matrix M in (1.5).

In the example, we took a scalar-valued AR(k) time series, which is not Markov, and simply by changing what we thought of as the state space, it became a vector-valued AR(1) time series, which is Markov. This illustrates a very important general principle.

Whether a process is Markov depends on what you consider the state.

We will see many examples of the use of this principle. Adding more variables to the state, can make a process Markov that wasn't before. It can also turn a process that was Markov into a different Markov process with simpler properties.

1.5 Stationary Stochastic Processes

A discrete-time stochastic process X_1, X_2, ..., not necessarily Markov, is *stationary* if the joint distribution of the vector $(X_n, X_{n+1}, \ldots, X_{n+k})$ does not depend on n for each fixed k.

This definition simplifies considerably when applied to a Markov chain. The conditional distribution of $(X_n, X_{n+1}, \ldots, X_{n+k})$ given the entire past history is a function of X_n alone by the Markov property. Therefore a Markov chain is stationary if the distribution of X_n does not depend on n. Note well the distinction, a Markov chain having *stationary transition probabilities* is not necessarily *stationary*. The former is a property of the transition properties alone, the latter involves the initial distribution.

A probability distribution is *invariant* for a specification of transition probabilities if the Markov chain that results from using that distribution as the initial distribution is stationary.

A important problem in the theory of Markov chains is determining for a specification of transition probabilities whether an invariant distribution exists and is unique. For us the existence aspect of this problem will not be interesting, because in Markov chain Monte Carlo we always construct chains to have a specified invariant distribution. We will be interested in the uniqueness question.

Example 1.3. I. I. D. Sequences.
A trivial special case of Markov chains is an i. i. d. sequence X_1, X_2, ... Since the conditional distribution of X_n given any other variables is the same as its unconditional distribution by independence, the Markov property holds. The Markov chain is stationary because the X_n are identically distributed. The unique invariant distribution is the distribution of the X_n.

Example 1.4. Maximally Uninteresting Chains.
A very trivial special case of Markov chains is defined by the recursion $X_{n+1} = X_n$. This specifies a set of transition probabilities for which *any* probability distribution is invariant. Since $X_n = X_1$ for all n, of course the distribution of X_n is the same for all n. The reason this chain is "maximally uninteresting" is because it goes nowhere and does nothing. Observing the whole chain tells us nothing more than observing X_1.

Needless to say, we will not be very interested in "maximally uninteresting" chains. The only point in knowing about them at all is to provide simple examples. For example, they do tell us that the uniqueness question for invariant distributions is a real question. There do exist transition probabilities with more than one invariant distribution.

Example 1.5. AR(1) Time Series (Continued).
The fact that linear combinations of normal random variables are normal leads one to suspect that an AR(1) time series has an invariant distribution that is normal, say $\mathcal{N}(\mu, \sigma^2)$. We can determine μ and σ^2 by checking the first and second moments of (1.3).

$$\mu = E(X_n) = \rho E(X_{n-1}) + E(e_n) = \rho\mu \tag{1.6a}$$

and

$$\sigma^2 = \operatorname{Var}(X_n) = \rho^2 \operatorname{Var}(X_{n-1}) + \operatorname{Var}(e_n) = \rho^2\sigma^2 + \tau^2. \tag{1.6b}$$

From (1.6a) we see that we must have either $\rho = 1$ or $\mu = 0$. The choice $\rho = 1$ combined with (1.6b) requires $\tau^2 = 0$, which gives us the maximally uninteresting chain as a degenerate special case of the AR(1) model.

The choice $\mu = 0$ places no restriction on ρ, but we get other restrictions from (1.6b). Since σ^2 and τ^2 are both nonnegative, $\rho^2 \geq 1$ would require $\sigma^2 = \tau^2 = 0$, which again gives a degenerate model. Thus the only Gaussian invariant distributions for nondegenerate AR(1) models (i. e., $\tau^2 > 0$) have $\mu = 0$ and $\rho^2 < 1$ and

$$\sigma^2 = \frac{\tau^2}{1 - \rho^2}. \tag{1.7}$$

In fact, this is the unique invariant distribution (Exercise 1.1).

1.6 Asymptotics for Stationary Processes and Markov Chains

1.6.1 The Law of Large Numbers

The theorem for stationary stochastic processes that is analogous to the SLLN for i. i. d. sequences is often called the Birkhoff ergodic theorem (Fristedt and Gray 1997, Section 28.4). Under a certain technical condition called "ergodicity" it has exactly the same conclusion as the SLLN. If Y_1, Y_2, ... is a stationary real-valued stochastic process that is ergodic, and $E(Y_i) = \mu$, then

$$\overline{Y}_n \xrightarrow{\text{a. s.}} \mu, \qquad n \to \infty. \tag{1.8}$$

A stationary Markov chain X_1, X_2, ... is a stationary stochastic process, but it needn't be real-valued. If g is a real-valued function on the state space of the Markov chain, then $g(X_1)$, $g(X_2)$, ... is a stationary real-valued stochastic process. Note well that it is not necessarily a Markov chain, because conditioning on $g(X_n)$ as opposed to X_n may not give the Markov property. However,

the process $g(X_1)$, $g(X_2)$, ... does have many nice properties. It is called a "functional" of the original chain.

If the original Markov chain has a unique invariant distribution, then it is an ergodic process in the sense required for the Birkhoff ergodic theorem, and the SLLN holds for the functional of the chain if the functional has finite expectation, that is, if $Y_i = g(X_i)$ and $E(Y_i) = \mu$, then (1.8) holds, which is the same except for different notation as (1.2), which we used in analyzing ordinary independent-sample Monte Carlo.

It is not completely obvious from the statement we just gave, but the SLLN for Markov chains does not have anything to do with the initial distribution or stationarity. Because it involves almost sure convergence, the convergence happens from almost all starting points. Thus we could restate the result as follows. If for a fixed specification of transition probabilities there is a unique invariant distribution, then the SLLN holds for any initial distribution that is dominated by the invariant distribution (is absolutely continuous with respect to it).

One should not get too excited about this formulation of the SLLN. Later we will see that an even stronger version is typically true. Under a slightly stronger regularity condition than uniqueness of the invariant distribution, called Harris recurrence, the SLLN holds for any initial distribution whatsoever. This condition is too technical to go into now. We will look at it later.

1.6.2 The Central Limit Theorem

We have just seen that the SLLN is no more complicated for Markov chains than for i. i. d. random variables. This is not the case with the CLT. The reason the CLT is more complicated is that "the expectation of a sum is the sum of the expectations" holds for any random variables, dependent or not, but the analogous rule for variances, "the variance of a sum is the sum of the variances," only holds for independent random variables. The general rule is

$$\begin{aligned} \operatorname{Var}\left(\sum_{i=1}^{n} Y_i\right) &= \sum_{i=1}^{n} \sum_{j=1}^{n} \operatorname{Cov}(Y_i, Y_j) \\ &= \sum_{i=1}^{n} \operatorname{Var}(Y_i) + 2 \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \operatorname{Cov}(Y_i, Y_j) \end{aligned}$$

If the variables form a stationary stochastic process, then $\operatorname{Var}(Y_n)$ does not depend on n and $\operatorname{Cov}(Y_n, Y_{n+k})$ does not depend on n for fixed k. Hence

$$\operatorname{Var}\left(\sum_{i=1}^{n} Y_i\right) = n \operatorname{Var}(Y_j) + 2 \sum_{k=1}^{n-1} (n-k) \operatorname{Cov}(Y_j, Y_{j+k})$$

(where by stationarity, the right hand side does not depend on j). To simplify notation, we define for any real-valued stationary stochastic process Y_1, Y_2, ... the *lag k autocovariance* $\gamma_k = \operatorname{Cov}(Y_j, Y_{j+k})$ (which does not depend on j by

stationarity. Note that as a special case $\gamma_0 = \operatorname{Var}(Y_j)$. Using this notation, the variance of the sample mean $\overline{Y}_n$ becomes

$$n \operatorname{Var}\left(\overline{Y}_n\right) = \gamma_0 + 2 \sum_{k=1}^{n-1} \frac{n-k}{n} \gamma_k. \tag{1.9}$$

In the special case where the Y_i are i. i. d. with $\operatorname{Var}(Y_i) = \sigma^2$, this reduces to the familiar $n \operatorname{Var}\left(\overline{Y}_n\right) = \gamma_0 = \sigma^2$ because all the covariances are zero. When we have dependence (1.9) makes it clear that the variance in the CLT cannot be the same as with independence.

So far so good, but now things get very murky. If we look in the literature on central limit theorems for stationary processes, for example in Peligrad (1986), we find central limit theorems under many different conditions, but none of the conditions seem easy to verify, nothing like the very simple condition in the i. i. d. case (there is a CLT if the variance is finite). For now we will not worry about conditions that imply the CLT. Let us just assume the CLT holds and proceed.

If the CLT holds, we might expect the limiting variance to be the limit of (1.9) as $n \to \infty$, and if things are simple this limit will be

$$\sigma^2_{\text{clt}} = \gamma_0 + 2 \sum_{k=1}^{\infty} \gamma_k. \tag{1.10}$$

There are two issues here. First is the limit of the variances the variance of the limiting random variable? The answer is not necessarily, a condition implying that is uniform integrability (Fristedt and Gray 1997, p. 108 and Problem 26 of Chapter 14). The second issue is whether the limit of (1.9) as n goes to infinity is actually (1.10). The answer to that is also not necessarily. The limit

$$\lim_{n \to \infty} \sum_{k=1}^{n-1} \frac{n-k}{n} \gamma_k. \tag{1.11}$$

is what is called in real analysis the *Cesáro sum* of the γ_k. It is a theorem of real analysis (Stromberg 1981, Theorem 7.81) that the Cesáro sum is equal to the ordinary sum

$$\sum_{k=1}^{\infty} \gamma_k = \lim_{n \to \infty} \sum_{k=1}^{n} \gamma_k \tag{1.12}$$

if the series is *absolutely summable*, that is, if $\sum_{k=1}^{\infty} |\gamma_k| < \infty$. When the series is not absolutely summable, it may be the case that the Cesáro sum (1.11) exists, but the ordinary sum (1.12) does not exist. Neither of these points enters Markov chain theory in an important way. We have fussed about these two issues only so that is it is clear what you *cannot* say about the variance in the CLT for stationary processes.

In the special case where $Y_i = g(X_i)$ is a functional of a Markov chain, the situation remains murky. Theorems that are sharp have conditions that are hard

to verify. There is one condition that implies a CLT and which can be verified in at least some practical examples, that the Markov chain be *geometrically ergodic* and that $E\{g(X_i)^{2+\epsilon}\}$ exist for some $\epsilon > 0$ (Chan and Geyer 1994), but this condition is still too complicated to discuss now. Sorting out what we can say about the CLT for Markov chains will be a major topic of the course.

As was the case with the SLLN, the CLT for a Markov chain does not require stationarity. The same technical condition, Harris recurrence, that guarantees the SLLN holds for all initial distributions if it holds for the invariant distribution guarantees the same thing about the CLT: the CLT holds for all initial distributions if it holds for the invariant distribution.

Example 1.6. AR(1) Time Series (Continued).
For a stationary, scalar-valued AR(1) time series, autocovariances are easy to calculate using the recursion (1.3). Recall that $E(X_n) = 0$ and $\operatorname{Var}(X_n) = \sigma^2$, where σ^2 is given by (1.7). So

$$\operatorname{Cov}(X_n, X_{n+k}) = \operatorname{Cov}(X_n, \rho X_{n+k-1} + e_{n+k}) = \rho \operatorname{Cov}(X_n, X_{n+k-1}) \tag{1.13}$$

By mathematical induction we get

$$\operatorname{Cov}(X_n, X_{n+k}) = \rho^k \sigma^2. \tag{1.14}$$

The base of the induction, the case $k = 0$, is clear. Plugging (1.14) into (1.13) shows the induction step is correct.

Now we can find the asymptotic variance (1.10)

$$\begin{aligned} \sigma^2_{\text{clt}} &= \gamma_0 + 2 \sum_{k=1}^{\infty} \gamma_k. \\ &= \sigma^2 \left(1 + 2 \sum_{k=1}^{\infty} \rho^k \right) \\ &= \sigma^2 \left(1 + 2 \frac{\rho}{1 - \rho} \right) \\ &= \sigma^2 \frac{1 + \rho}{1 - \rho} \end{aligned} \tag{1.15}$$

the geometric series being summable because of the condition $|\rho| < 1$ which is required for stationarity. This model is so simple we can show both the SLLN and the CLT by direct methods (Exercise 1.2).

A Caution: The γ_k are the lagged autocovariances for the *stationary* Markov chain, started in the invariant distribution. Thus (1.9) is the variance of $\sqrt{n}\, \overline{Y}_n$ for the *stationary* Markov chain. We have seen that when the sequence of autocovariances is absolutely summable, this variance converges to the asymptotic variance (1.10).

A tempting error, that many people have fallen prey to, is the very similar statement that the variance of $\sqrt{n}\, \overline{Y}_n$ converges to the asymptotic variance

without requiring stationarity. The error is easily seen by considering the AR(1) process.

To simplify notation a bit, let us start with X_0 rather than X_1, then

$$\begin{aligned} X_1 &= \rho X_0 + e_1 \\ X_2 &= \rho X_1 + e_2 \\ &= \rho^2 X_0 + \rho e_1 + e_2 \\ &\vdots \\ X_n &= \rho^n X_0 + \rho^{n-1} e_1 + \rho^{n-2} e_2 + \cdots + e_n \end{aligned} \tag{1.16}$$

It is clear that if X_0 does not have finite variance, then neither does any X_n, nor does $\overline{X}_n$. Thus the variance of $\sqrt{n}\ \overline{X}_n$ (which is always infinite) does not converge to the asymptotic variance (1.10) even though the CLT holds (Exercise 1.2).

1.6.3 Estimating the Asymptotic Variance

It is not enough to have a CLT. We must also be able to estimate the variance in the CLT (1.10). There are many ways to do this, the simplest and the only one we will look at now is the method of *batch means*. It is based on the fact that if a Markov chain X_1, X_2, ... satisfies the CLT and we want to estimate the mean of a functional $Y_n = g(X_n)$ using the estimate $\overline{Y}_n$ and

$$n \operatorname{Var}(\overline{Y}_n) \to \sigma^2_{\text{clt}},$$

then the variance of the average over a segment of the chain of sufficiently long length will be a good estimate. Hence divide the chain into consecutive segments of length m. These are called *batches*. Write $\sigma^2_m = m \operatorname{Var}(\overline{Y}_m)$, which for sufficiently large m will be close to σ^2_{clt}, because σ^2_m is given by (1.9) with n replaced by m, and (assuming absolute summability of the autocovariance sequence) this converges to σ^2_{clt} as $m \to \infty$.

Now we use a trick like the one we used in converting an AR(k) process, which was not Markov, into a vector-valued AR(1) process, which was. Write

$$Z_n = \begin{pmatrix} X_{m(n-1)+1} \\ \vdots \\ X_{mn} \end{pmatrix}$$

Then the Z_n form a Markov chain, and the *batch means*

$$B_n = g(Z_n) = \frac{1}{m} \sum_{i=1}^{m} g(X_{m(n-1)+i})$$

are a functional of this Markov chain. Hence by the SLLN for Markov chains (a. k. a. the Birkhoff ergodic theorem),

$$\overline{B}_n = \frac{1}{n} \sum_{i=1}^{n} B_i \xrightarrow{\text{a. s.}} E(B_1) = \mu,$$

where $\mu = E(Y_i)$ for all i, assuming stationarity of the Markov chain, and

$$\frac{1}{n}\sum_{i=1}^{n}[B_i - \mu]^2 \xrightarrow{\text{a. s.}} \operatorname{Var}(B_1) = \operatorname{Var}(\overline{Y}_m) = \frac{\sigma_m^2}{m} \approx \frac{\sigma_{\text{clt}}^2}{m}$$

Combining these gives

$$s_{\text{batch}}^2 = \frac{1}{n}\sum_{i=1}^{n}[B_i - \overline{B}_n]^2 \approx \frac{\sigma_{\text{clt}}^2}{m}$$

Combining this with the CLT gives

$$\overline{Y}_n \approx \mathcal{N}\left(\mu, \frac{\sigma_{\text{clt}}^2}{n}\right) \approx \mathcal{N}\left(\mu, \frac{m}{n} s_{\text{batch}}^2\right)$$

All of this can be explained without formulas if you trust such an argument. The batch means B_i have approximately the same variance as Y_n except for a factor m/n that arises from the different lengths of the sequences. The variance of the batch means is estimated by their sample variance. End of argument.

So how large should the batch size be? To be useful, it should be large enough so that $\sigma_m^2 \approx \sigma_{\text{clt}}^2$. And how large is that? It depends on the details of the Markov chain problem. Since we rarely know anything about those details, we want a batch size as large as possible.

On the other hand we want the number of batches to be large so that s_{batch}^2 will be a good estimate of σ_{clt}^2/m. We want at least 20 batches, and 100 or more would be desirable.

This creates something of a conflict. We want the batch size to be large, very large. We also want the batch size to be small relative to the Monte Carlo sample size n. Unless n is very, very, very large, we may not be able to satisfy both wants. It is frustrating that we need a much larger Monte Carlo sample size to estimate the MCSE accurately than we need to estimate accurately the quantity of interest. However, we do not need a very accurate MCSE, one significant figure will do, whereas we want as much accuracy as possible, two or more significant figures, for the sample mean (our Monte Carlo approximation of the quantity of interest).

So there often is a batch size that works. The question is how to find it. One recommendation that has been made in the literature (Schmeiser 1982) is that the number of batches should be small, no more than thirty, since that will give a decent estimate of σ_m^2 and there is generally no telling how large m must be so that σ_m^2 is close to σ_{clt}^2/m.

A possible diagnostic of a batch size being too small is to check the lagged autocovariances of the batches. Since the batch means form a functional of a Markov chain, the variance in the CLT is given by a formula like (1.10), say

$$n \operatorname{Var}(\overline{B}_n) \to \gamma_{m,0} + 2\sum_{k=1}^{\infty} \gamma_{m,k}$$

where

$$\begin{aligned}\gamma_{m,k} &= m \operatorname{Cov}(B_i, B_{i+k}) \\ &= \frac{1}{m^2} \sum_{i=1}^{m} \sum_{j=mk+1}^{m(k+1)} \gamma_{j-i} \\ &= \sum_{l=-(m-1)}^{m-1} \frac{m - |l|}{m} \gamma_{mk+l}\end{aligned}$$

The lag zero autocovariance is $\gamma_{m,0} = \sigma_m^2$. The other lagged autocovariances $\gamma_{m,k}$ for $k \geq 1$ converge to zero as $m \to \infty$, because in order for the original autocovariance sequence to be absolutely summable we need $|\gamma_k| \to 0$ as $k \to \infty$.

Thus an equivalent way to think about the batch length m being large enough, is considering whether the batch means B_i are almost uncorrelated. If the $\gamma_{m,k}$ for $k \geq 1$ are not significantly different from zero, then m is large enough. We shall defer till later an explanation of how to test whether autocovariances are zero, but computer packages with time series capabilities may have such a test built in. In S-PLUS, for example, the `acf` function makes an autocorrelation plot with 95% confidence limits about zero. Autocorrelations within the confidence limits can be considered negligible.

1.7 Markov Chain Monte Carlo

We are finally ready to say something about Markov chain Monte Carlo. Specific algorithms for MCMC will be a major focus of the course. Here we will just mention one algorithm, not the best, nor the most useful, but the easiest to explain. This is the *Gibbs sampler*, thus named by Geman and Geman (1984), although special cases of the algorithm had been used by earlier authors, for example, Ripley (1979).

The general notion of MCMC is to estimate probabilities or expectations by simulating a Markov chain and averaging over the simulations. The probabilities or expectations calculated are those for functionals $g(X_i)$ of the *stationary* chain, hence they are probabilities or expectations with respect to the *invariant* distribution. Thus the first task in any MCMC application is to find a Markov chain having a specified invariant distribution.

The Gibbs sampler is a method that does this using almost no theory, no more than the definition of conditional probability. Before we can define it, though we need to look at even more basic concept: combining update mechanisms. Let us call any well-defined procedure that makes a random change in the state of a system according to a probability law that depends only on the current state a *Markov update mechanism*. A Markov chain results from iterating a Markov update mechanism. In the context of MCMC, we can think of a Markov update mechanism as a bit of computer code that makes a random[1] change in the state.

[1]Pedants will insist on "pseudo-random" rather than "random" here to indicate that com-

The point of isolating the notion of an update mechanism, is that we can use it to define new Markov chains. Let us say that an update mechanism *preserves* a specified probability distribution if that distribution is invariant for the Markov chain obtained by iterating the update mechanism. So another way to state the "first task in MCMC" is to find a Markov update mechanism that preserves a specified distribution.

1.7.1 Combining Update Mechanisms

There are several ways of combining update mechanisms that preserve a specified distribution to obtain a new update mechanism that also preserves the same distribution. The first is *composition*, which is following one update mechanism with another. It is clear that if an update mechanism U_1 preserves a specified distribution, and so does another update mechanism U_2, then so does U_1 followed by U_2, which we will denote U_1U_2. It is also clear that this can be applied to more than two update mechanisms that all preserve the same distribution: $U_1U_2 \ldots U_k$ preserves a distribution if each of the U_i does.

Another way of combining update mechanisms is *mixing*, which is making a random choice among update mechanisms. Suppose U_1, ..., U_k preserve the same distribution and p_1, ..., p_k is a fixed probability vector (the p_i are nonnegative and sum to one). Then the mechanism that updates the state by chosing U_i with probability p_i and then performing U_i is called the mixture of the U_i with mixing probabilities p_i. It is clear that this also preserves the specified distribution, because no matter which U_i is chosen the distribution is preserved. Later we will meet several more complicated ways of combining update mechanisms. These two will do for now.

The terms used here are not standard. Most of the literature uses the word "scan" in this context, the idea being that if you have several update mechanisms preserving the same distribution, you want to "scan" though them to use them all. What we call "composition" most MCMC authors call "fixed scan," and what we call "mixing" most MCMC authors call "random scan." There are two reasons for our new terminology. First, it is more comprehensive. As we will see, it covers many ways of combining update mechanisms that are not described by the terms "fixed scan" and "random scan." Second, it is more closely connected to Markov chain theory. As we will see, composition corresponds to composition of the Markov kernels representing the update mechanisms, and mixing corresponds to linear combinations.

1.7.2 The Gibbs Sampler

Now we can present the notion of a *Gibbs update* mechanism. At the beginning of an application of MCMC we don't have a Markov chain, just a specified distribution we want our Markov chain (when invented) to preserve. Let X be a random element of the state space having this distribution, and let $h(X)$

puters don't have really truely random numbers. We won't bother with this distinction.

be any function of X. A Gibbs update gives X a new value simulated from the conditional distribution of X given $h(X)$. That is this update preserves the specified distribution is a straighforward consequence of the definition of conditional probability. If $g(X)$ is any integrable function, then

$$E\{E[g(X)|h(X)]\} = E\{g(X)\}$$

(sometimes called the iterated expectation formula) shows that the expectation of $g(X)$ is unchanged by the update, hence, since g could be the indicator of any measurable set A, this shows that $\Pr(X \in A)$ is unchanged by the update for A.

This usage is also not standard. What we have described here includes what is usually called a Gibbs update as a special case, but it also includes many updates most MCMC authors would call "block Gibbs" or "generalized Gibbs" or perhaps not even recognize as updates closely related to what they think of as Gibbs. It seems foolish not to collect all updates based on the same extremely simple idea under one name, and your humble author dislikes terminology of the form "generalized blah de blah."

The usual notion of a Gibbs update is the following. The state X is a vector $X = (X_1, \ldots, X_k)$. (Warning: for the next few paragraphs, subscripts indicate components of the state vector, not the time index of a discrete-time stochastic process, as they have up to now.) There are k Gibbs update mechanisms. Each changes only one component X_i giving it a new value simulated from its conditional distribution given the rest of the variables. It is a very useful notational convenience when dealing with Gibbs sampling to have a notation for "the rest." A widely used notation is

$$X_{-i} = (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_k).$$

Thus a Gibbs update gives X_i a new value simulating from the conditional distribution of X_i given X_{-i}. These k conditional distributions of each X_i given X_{-i} are called the *full conditionals* of the distribution of X in the Gibbs sampling literature.

The very limited view of Gibbs updates just described is an obvious special case of the more general view. Taking $h(X) = X_{-i}$ gives the Gibbs update of X_i. In a very curious inversion, the "general view" is a special case of the "limited view" if looked at the right way, a way that starts with the question: what is a "variable" to be Gibbsed? The "limited view" starts with a fixed list X_1, ..., X_k of variables. It declares that these are the only mathematical objects that will be allowed to be called "variables" in discussion of the problem at hand. The "general view" says, why not some other list of "variables"? If we consider $h(X)$ and X to be the "variables," we get the "generalized Gibbs" update as a special case of the "limited Gibbs" update, which is absurd. Better to call them all just plain Gibbs, as we have recommended here.

Example 1.7. Bayesian Inference for the Two-Parameter Normal.
Suppose we observe data X_1, ..., X_n i. i. d. $\mathcal{N}(\mu, \lambda^{-1})$ and want to make

Bayesian inference about the parameters μ and λ. The distribution we want to know about here is the posterior distribution of μ and λ given the data $X_1, \ldots, X_n$. The posterior depends on the data and on our prior, which we will assume has a probability density function $g(\mu, \lambda)$.

As is well known (DeGroot 1970, Section 9.6), there is a closed-form solution to this problem, if (big if) we choose the prior for reasons of mathematical convenience to be of the form[2]

$$\lambda \sim \text{Gamma}(\alpha, \beta) \tag{1.17a}$$

$$\mu | \lambda \sim \mathcal{N}\big(\gamma, \delta^{-1}\lambda^{-1}\big) \tag{1.17b}$$

where α, β, γ, and δ are hyperparameters of the prior to be chosen to reflect subjective prior opinion (or objective, call it what you will, your humble author has no ax to grind here).

This is a so-called *conjugate family* of prior distributions (DeGroot 1970, Chapter 9), one that is *closed under sampling*, which means the posterior distribution is in the same family (with different values of the hyperparameters) as the prior for any sample size n and any values of the data.

Bayesians have always felt a need to justify the curious prior dependence between μ and λ. Why have $\text{Var}(\mu|\lambda) = \delta^{-1}\lambda^{-1}$? How does that represent anyone's prior opinion? The justification is that, because this is a conjugate family, the prior could have arisen as the posterior from some earlier data analysis, and even if a flat prior would have been used then, the resulting posterior, which is now our prior, would exhibit this "curious" dependence (DeGroot 1970, p. 170). There is something unsatisfactory about this explanation. Why this particular family of priors?

The family obtained by keeping (1.17a) and changing (1.17b) to

$$\mu | \lambda \sim \mathcal{N}\big(\gamma, \delta^{-1}\big), \tag{1.17c}$$

is no longer a conjugate family, but it is a reasonable family of priors, perhaps more reasonable. We cannot resolve the question of which family is better. A subjectivist Bayesian always resolves the issue by asking an "expert" or a "user" or whatever one wishes to call the person whose subjective opinion is to be used. The opinion of statisticians, especially those not involved in the particular application is irrelevant. Having no particular application in mind and hence no users to ask, we can have no opinion about which family of priors is better, or for that matter whether either family is any use at all. We can only proceed with the example to see how it turns out, leaving questions of relevance unanswered.

The conjugate family is "pencil-and-paper-friendly" (to coin a phrase by analogy with "user-friendly"). The family described by (1.17a) and (1.17c) is

[2]The notation $\text{Gamma}(\alpha, \beta)$ here indicates the distribution with density

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, \qquad x > 0$$

rather than the other convention which replaces β by $1/\beta$.

"Gibbs-friendly" because as we will see, there is no problem sampling it with the Gibbs sampler. The likelihood times the prior is proportional to

$$h(\mu, \lambda) = \lambda^{n/2} \exp\left\{ -\frac{n\lambda v_n}{2} - \frac{n\lambda}{2}(\bar{x}_n - \mu)^2 \right\} \lambda^{\alpha-1} e^{-\beta\lambda} \exp\left\{ -\frac{\delta}{2}(\mu - \gamma)^2 \right\}$$

where $\bar{x}_n$ is the sample mean and v_n is the sample variance with n rather than $n - 1$ in the formula, that is

$$\begin{aligned} \bar{x}_n &= \frac{1}{n} \sum_{i=1}^{n} x_i \\ v_n &= \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2 \end{aligned}$$

Staring a bit at the definition of $h(\mu, \lambda)$ we see that for fixed λ it is e to a quadratic function of μ, hence the "full conditional" for μ is normal, and that for fixed μ it is a power of λ times e to a constant times λ, hence the "full conditional" for λ is gamma. Specifically,

$$\lambda | \mu \sim \text{Gamma}\left(\alpha + \frac{n}{2}, \beta + \frac{n v_n}{2} + \frac{n}{2}(\bar{x}_n - \mu)^2 \right) \tag{1.18a}$$

$$\mu | \lambda \sim \mathcal{N}\left(\frac{n\lambda \bar{x}_n + \delta\gamma}{n\lambda + \delta}, \frac{1}{n\lambda + \delta} \right) \tag{1.18b}$$

(Exercise 1.7).

So here is the recipe for the Gibbs sampler for this problem. Start anywhere, say at the prior means $\mu_1 = \gamma$ and $\lambda_1 = \alpha/\beta$. Then alternate the update steps. Simulate λ_2 from the distribution (1.18a) with μ_1 plugged in for μ. Then simulate μ_2 from the distribution (1.18b) with λ_2 (the current value) plugged in for λ. And repeat.

- Simulate λ_n from the distribution (1.18a) with μ_{n-1} plugged in for μ.
- Simulate μ_n from the distribution (1.18b) with λ_n plugged in for λ.

This produces a Markov chain (λ_n, μ_n), $n = 1, 2, \ldots$ with state space $\mathbb{R}^2$.

There are several ways to look at the simulation output. One is to look at time-series plots of functionals of the chain. An example is Figure 1.1, which plots μ_n versus n. The time series plot shows very little autocorrelation. The reader should be warned that this example is very atypical. Most MCMC time-series plots show much more autocorrelation. This is a very easy Markov chain problem.

Another way to look at the simulation output is a scatter plot of two functionals of the chain. An example is Figure 1.2, which plots μ_n versus $\sigma_n = 1\sqrt{\lambda_n}$. In this figure we have lost the time-series aspect. It gives no indication that the sample is from a Markov chain or how much dependence there is in the Markov chain. There is no way to tell, just looking at the figure, whether this is an MCMC sample or an ordinary, independent-sampling sample. This is an important principle of MCMC.

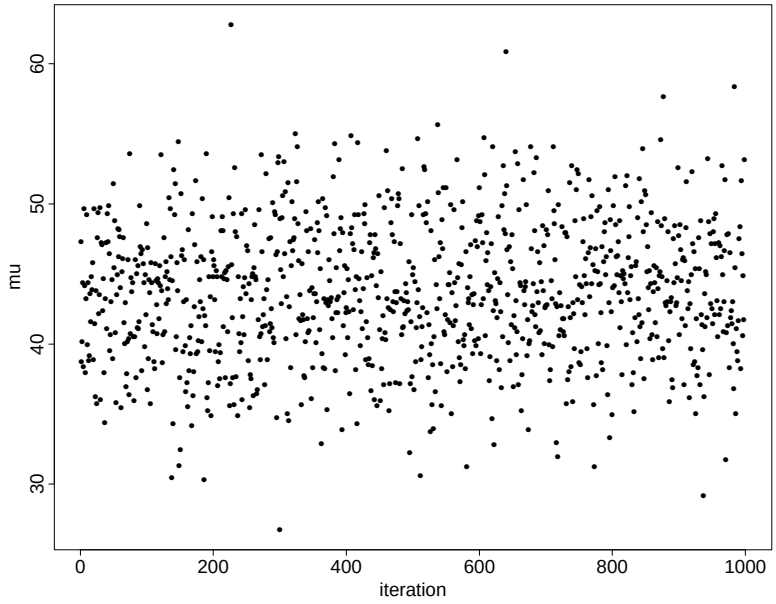


Figure 1.1: Time series plot of Gibbs sampler output for μ in the two-parameter normal model. Sufficient statistics for the data were $\bar{x}_n = 41.56876$, $v_n = 207.5945$, and $n = 10$. Hyperparameters of the prior were $\alpha = 1$, $\beta = 20^2$, $\gamma = 50$, and $\delta = 1/10^2$. The starting point was $\mu = \gamma$ and $\lambda = \alpha/\beta$.

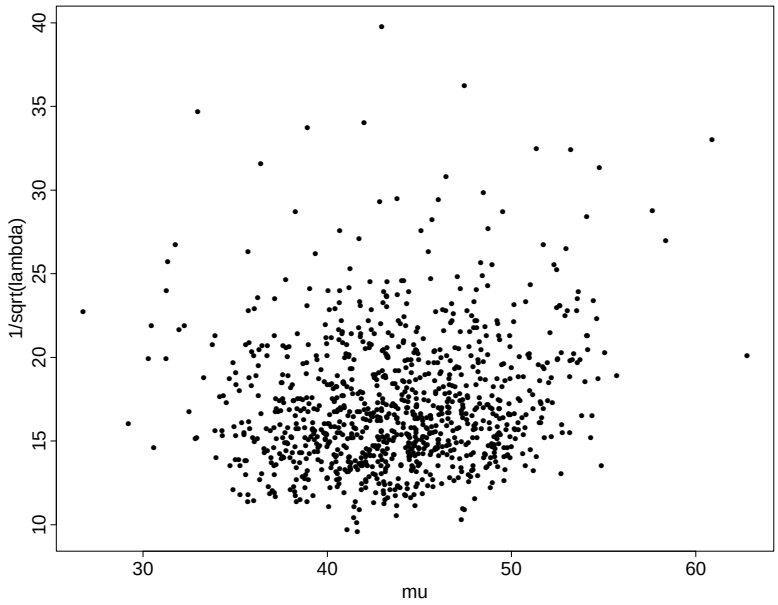


Figure 1.2: Scatter plot of Gibbs sampler output for μ and $\sigma = 1/\sqrt{\lambda}$ in the two-parameter normal model, the same run as shown in Figure 1.1.

An MCMC scatter plot approximates the distribution of interest, just like an OMC (ordinary Monte Carlo) scatter plot.

This follows from the SLLN. Suppose A is any event (some region in the figure). Then the SLLN says

$$\frac{1}{n} \sum_{i=1}^{n} 1_A(\lambda_n, \mu_n) \xrightarrow{\text{a. s.}} \Pr\{(\lambda, \mu) \in A | \text{data}\}$$

Without the symbols, this says the fraction of points in a region A in the figure approximates the posterior probability of that region.

Yet another way to look at the simulation output is a histogram of one functional of the chain. An example is Figure 1.3, which plots a histogram of the μ_n. By the SLLN again, this is the MCMC approximation of the marginal posterior distribution of μ (same argument as for scatter plots).

A clever method due to Wei and Tanner (1990) gives a much better estimate of the marginal posterior for μ. Curiously, it ignores the simulated values of μ and uses only the simulated values of λ. The distribution of μ given λ is a known normal distribution (1.18b). Denote its density by $f(\mu|\lambda, \text{data})$. Let $f_\lambda(\lambda|\text{data})$ denote the marginal posterior density of λ (which is not known). The marginal

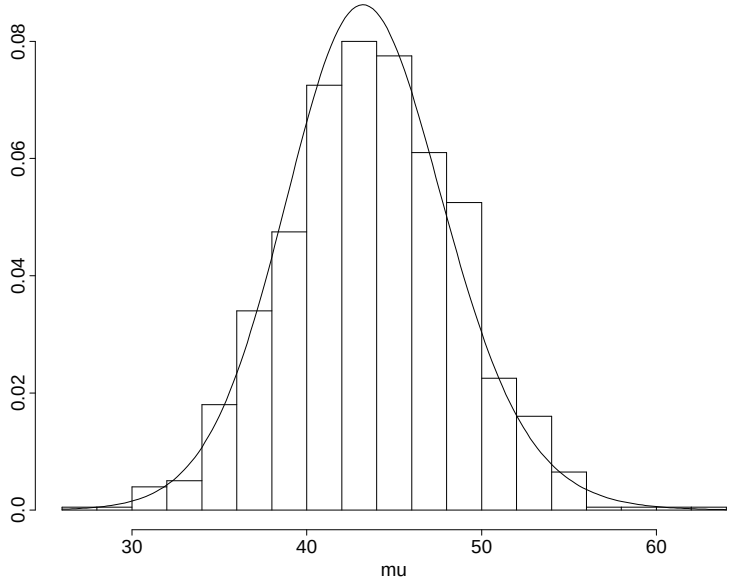


Figure 1.3: Histogram of Gibbs sampler output for μ in the two-parameter normal model, the same run as shown in Figure 1.1. The curve is the estimator of Wei and Tanner (1990) given by (1.19).

posterior for μ is then given by

$$f_\mu(\mu|\text{data}) = \int f(\mu|\lambda, \text{data}) f_\lambda(\lambda|\text{data})\, d\lambda.$$

The integrand is the joint posterior of (μ, λ) given the data, so integrating out λ gives the marginal for μ. We cannot easily do the integral analytically, but we can do it by Monte Carlo

$$f_{\mu,n}(\mu|\text{data}) = \frac{1}{n} \sum_{i=1}^{n} f(\mu|\lambda_i, \text{data}) \tag{1.19}$$

where the λ_i are the simulated values from the MCMC run. Note well that (1.19) is to be considered a function of μ. For fixed data and MCMC output λ_1, ..., λ_n, we vary μ obtaining the smooth curve in Figure 1.3. Clearly the smooth curve is a much better estimate of the marginal posterior than the histogram. It is also much better than the histogram smoothed using standard methods of density estimation, such as kernel smoothing.

We can also get a highest posterior density (HPD) region for μ. An HPD region is a level set of the posterior density, in this case a set of the form

$$A_c = \{\, \mu : f_\mu(\mu|\text{data}) \geq c \,\}$$

for some constant c, which is chosen to give a desired posterior coverage, e. g., a 95% HPD region choses c so that $P(\mu \in A_c|\text{data}) = 0.95$. For any event A, the SLLN says that this probability is approximated by

$$P(\mu \in A|\text{data}) \approx \frac{1}{n} \sum_{i=1}^{n} 1_A(\mu_i)$$

So a region A will have 95% coverage, as estimated by MCMC, if it contains 95% of the points μ_1, ..., μ_n. It will be a HPD region if it has the property that $f_\mu(\mu|\text{data})$ is larger for any $\mu \in A$ than for any $\mu \notin A$. Thus we estimate c by the 5-th percentile of the n numbers $f_{\mu,n}(\mu_i|\text{data})$, $i = 1$, ..., n, and estimate A_c by

$$A_{c,n} = \{\, \mu : f_{\mu,n}(\mu|\text{data}) \geq c \,\}$$

Then the MCMC estimate of $P(\mu \in A_{n,c}|\text{data})$ is 0.95 by construction, and $A_{c,n}$ approximates the HPD region A_c. For the run shown in Figure 1.3, the 5-th percentile is 0.086, giving a 95% HPD region $(33.76, 53.4)$.

1.7.3 The Moral of the Story

It's a bit hard to say exactly what lessons are to be drawn from this example, because it's a toy problem. From the *Jargon File* (Raymond 1996)

toy problem /n./ [AI] A deliberately oversimplified case of a challenging problem used to investigate, prototype, or test algorithms for a real problem. Sometimes used pejoratively.

In statistics, toy problems include analyses of real data that look at questions much simpler than the original questions the data were collected to shed light on. By this criterion most examples in textbooks and papers are toy problems. As the definition from the *Jargon File* says, the term is only sometimes used pejoratively. If a toy problem is a good illustration of some specific issues, then there's nothing wrong with it.

> *Toy problems are all right if you draw the right lessons from them.*
>
> *But it's hard to know what lessons to draw from a toy problem.*

The trouble is that toy problems lack realism. At best they have pseudo-realism, when they use real data for a toy purpose,

> Merely corroborative detail, intended to give artistic verisimilitude to an otherwise bald and unconvincing narrative.
>
> Pooh-Bah (Lord High Everything Else)
> in Gilbert and Sullivan's *Mikado*

And it's hard to know what in a toy problem is realistic and what is merely artistic verisimilitude.

You might draw the lesson that all MCMC problems are this easy, which is very wrong. You might draw the lesson that "pencil-and-paper-friendly" models are now obsolete, that "for reasons of mathematical convenience" is no longer a good excuse. I hope you got that lesson. It's an important one. You might go a little farther and draw the lesson that "Gibbs-friendly" models are an important new class of models we need to theorize about. That would be a wrong lesson. The Gibbs sampler is a very limited algorithm, but there are many other MCMC algorithms. One of them almost always does the job.

> *MCMC does anything. Hence there is never any excuse for doing the Wrong Thing.*

From the *Jargon File* (Raymond 1996)

Right Thing /n./ That which is *compellingly* the correct or appropriate thing to use, do, say, etc. Often capitalized, always emphasized in speech as though capitalized. Use of this term often implies that in fact reasonable people may disagree. "What's the right thing for LISP to do when it sees `(mod a 0)`? Should it return `a`, or give a divide-by-0 error?" Oppose *Wrong Thing*.

Wrong Thing /n./ A design, action, or decision that is clearly incorrect or inappropriate. Often capitalized; always emphasized in speech as if capitalized. The opposite of the Right Thing; more generally, anything that is not the Right Thing. In cases where 'the good is the enemy of the best', the merely good—although good—is nevertheless the Wrong Thing. "In C, the default is for module-level declarations to be visible everywhere, rather than just within the module. This is clearly the Wrong Thing."

As the definition says, "reasonable people may disagree." If you are a Bayesian, you think a Bayesian analysis is the Right Thing. If you are a frequentist, you may think a hypothesis test is the Right Thing. The same goes for finer details, if you are a subjective Bayesian you think the prior must be elicited from a user or an expert, and so forth. Whatever the philosophical analysis that leads you to conclude that a particular statistical procedure is the Right Thing, that is what you must do, because some form of MCMC will enable you to do it.

It follows that there is no excuse for "algorithm-friendly" analyses. Changes made to the *statistical model* or the *mode of statistical inference* for the sake of using a particular MCMC algorithm or a simpler MCMC algorithm, are the Wrong Thing. In particular, "Gibbs-friendly" is dumb.

Another lesson you might draw from the example is that MCMC has its own bag of tricks not taken from the rest of statistics, like the method of Wei and Tanner (1990) for HPD regions. This is also a good lesson to draw. We will see other tricks, that do more than just calculate a simple sample average.

Exercises

1.1. For the scalar-valued AR(1) time series with nondegenerate error distribution ($\tau^2 > 0$), show that

(a) When $\rho^2 < 1$, the invariant distribution found in Example 1.5 is the unique invariant distribution.

(b) When $\rho \geq 1$, an invariant probability distribution does not exist.

Hint: use characteristic functions (both parts).

1.2. For a stationary, scalar-valued AR(1) time series with nondegenerate error distribution ($|\rho| < 1$ and $\tau^2 > 0$), show that, for any initial distribution,

(a) the marginal distribution of X_n converges to the invariant distribution $\mathcal{N}(0, \sigma^2)$ with σ^2 given by (1.7),

(b) the CLT holds, that is

$$\sqrt{n}\, \overline{X}_n \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2_{\text{clt}})$$

where σ^2_{clt} is given by (1.15), and

(c) the SLLN holds, that is

$$\overline{X}_n \xrightarrow{\text{a. s.}} 0.$$

Hints: In (b) use the fact that the autocovariances are absolutely summable so (1.11) and (1.12) agree. For (c) the Borel-Cantelli lemma implies that SLLN holds if the sequence $\Pr(|\overline{X}_n| \geq \epsilon)$ is summable.

1.3. Implement a scalar-valued AR(1) sampler with $\rho = 0.95$, and $\sigma^2 = 1$. Use a run of the chain of length 10,000 to estimate $\mu = \Phi(-2) = 0.02275$ using as your Monte Carlo approximation the fraction of the run that has $X_n < -2$. Find the MCSE of your estimate using the method of batch means.

1.4. For the vector-valued AR(1) time series with nondegenerate error distribution (the error variance matrix M is strictly positive definite), show that an invariant distribution exists if and only if $A^n \to 0$ as $n \to \infty$.

1.5. Verify the formulas (1.18a) and (1.18b) for the full conditionals in Example 1.7.

1.6. Produce a marginal density plot for $\sigma = \lambda^{-1/2}$ and a 95% HPD region for σ using the method of Wei and Tanner (1990) as described in Example 1.7. Use the data and hyperparameter values given in the caption for Figure 1.3. Hint: Don't forget the Jacobian.

1.7. Find the "full conditionals" for mean-zero exchangeable bivariate normal distribution (exchangeable meaning both components have the same variance). What is the connection of the Gibbs sampler for this distribution with the scalar-valued AR(1) time series?

Chapter 2

Basic Markov Chain Theory

To repeat what we said in the Chapter 1, a *Markov chain* is a discrete-time stochastic process X_1, X_2, ... taking values in an arbitrary state space that has the Markov property and stationary transition probabilities:

- the conditional distribution of X_n given X_1, ..., X_{n-1} is the same as the conditional distribution of X_n given X_{n-1} only, and
- the conditional distribution of X_n given X_{n-1} does not depend on n.

The conditional distribution of X_n given X_{n-1} specifies the *transition probabilities* of the chain. In order to completely specify the probability law of the chain, we need also specify the *initial distribution*, the distribution of X_1.

2.1 Transition Probabilities

2.1.1 Discrete State Space

For a discrete state space S, the transition probabilities are specified by defining a matrix

$$P(x, y) = \Pr(X_n = y | X_{n-1} = x), \qquad x, y \in S \tag{2.1}$$

that gives the probability of moving from the point x at time $n-1$ to the point y at time n. Because of the assumption of stationary transition probabilities, the transition probability matrix $P(x, y)$ does not depend on the time n.

Some readers may object that we have not defined a "matrix." A *matrix* (I can hear such readers saying) is a rectangular array P of numbers p_{ij}, $i = 1$, ..., m, $j = 1$, ..., n, called the *entries* of P. Where is P? Well, enumerate the points in the state space $S = \{x_1, \ldots, x_d\}$, then

$$p_{ij} = \Pr\{X_n = x_j | X_{n-1} = x_i\}, \qquad i = 1, \ldots, d, \ j = 1, \ldots d.$$

I hope I can convince you this view of "matrix" is the Wrong Thing. There are two reasons.

First, the enumeration of the state space does no work. It is an irrelevancy that just makes for messier notation. The mathematically elegant definition of a matrix does not require that the index sets be $\{1, \ldots, m\}$ and $\{1, \ldots, n\}$ for some integers m and n. Any two finite sets will do as well. In this view, a *matrix* is a function on the Cartesian product of two finite sets. And in this view, the function P defined by (2.1), which is a function on $S \times S$, *is* a matrix.

Following the usual notation of set theory, the space of all real-valued functions on a set A is written $\mathbb{R}^A$. This is, of course, a d-dimensional vector space when A has d points. Those who prefer to write $\mathbb{R}^d$ instead of $\mathbb{R}^A$ may do so, but the notation $\mathbb{R}^A$ is more elegant and corresponds to our notion of A being the index set rather than $\{1, \ldots, d\}$. So our matrices P being functions on $S \times S$ are elements of the d^2-dimensional vector space $\mathbb{R}^{S \times S}$.

The second reason is that P is a conditional probability mass function. In most contexts, (2.1) would be written $p(y|x)$. For a variety of reasons, partly the influence of the matrix analogy, we write $P(x, y)$ instead of $p(y|x)$ in Markov chain theory. This is a bit confusing at first, but one gets used to it. It would be much harder to see the connection if we were to write p_{ij} instead of $P(x, y)$.

Thus, in general, we define a *transition probability matrix* to be a real-valued function P on $S \times S$ satisfying

$$P(x, y) \geq 0, \qquad x, y \in S \tag{2.2a}$$

and

$$\sum_{y \in S} P(x, y) = 1. \tag{2.2b}$$

The state space S must be countable for the definition to make sense. When S is not finite, we have an infinite matrix. Any matrix that satisfies (2.2a) and (2.2b) is said to be *Markov* or *stochastic*.

Example 2.1. Random Walk with Reflecting Boundaries.
Consider the symmetric random walk on the integers 1, ..., d with "reflecting boundaries." This means that at each step the chain moves one unit up or down with equal probabilities, $\frac{1}{2}$ each way, except at the end points. At 1, the lower end, the chain still moves up to 2 with probability $\frac{1}{2}$, but cannot move down, there being no points below to move to. Here when it wants to go down, which is does with probability $\frac{1}{2}$, it bounces off an imaginary reflecting barrier back to where it was. The behavior at the upper end is analogous. This gives a transition matrix

$$\begin{pmatrix}
\frac{1}{2} & \frac{1}{2} & 0 & 0 & \ldots & 0 & 0 & 0 \\
\frac{1}{2} & 0 & \frac{1}{2} & 0 & \ldots & 0 & 0 & 0 \\
0 & \frac{1}{2} & 0 & \frac{1}{2} & \ldots & 0 & 0 & 0 \\
0 & 0 & \frac{1}{2} & 0 & \ldots & 0 & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\
0 & 0 & 0 & 0 & \ldots & 0 & \frac{1}{2} & 0 \\
0 & 0 & 0 & 0 & \ldots & \frac{1}{2} & 0 & \frac{1}{2} \\
0 & 0 & 0 & 0 & \ldots & 0 & \frac{1}{2} & \frac{1}{2}
\end{pmatrix} \tag{2.3}$$

We could instead use functional notation

$$P(x, y) = \begin{cases} 1/2, & |x - y| = 1 \text{ or } x = y = 1 \text{ or } x = y = d \\ 0, & \text{otherwise} \end{cases}$$

Either works. We will use whichever is most convenient.

2.1.2 General State Space

For a general state space S the transition probabilities are specified by defining a *kernel*

$$P(x, B) = \Pr\{X_n \in B | X_{n-1} = x\}, \qquad x \in S, \; B \text{ a measurable set in } S,$$

satisfying

- for each fixed x the function $B \mapsto P(x, B)$ is a probability measure, and
- for each fixed B the function $x \mapsto P(x, B)$ is a measurable function.

In other words, the kernel is a *regular* conditional probability (Breiman 1968, Section 4.3).

Lest the reader worry that this definition signals an impending blizzard of measure theory, let me assure you that it does not. A little bit of measure theory is unavoidable in treating this subject, if only because the major reference works on Markov chains, such as Meyn and Tweedie (1993), are written at that level. But in practice measure theory is entirely dispensable in MCMC, because the computer has no sets of measure zero or other measure-theoretic paraphernalia. So if a Markov chain really exhibits measure-theoretic pathology, it can't be a good model for what the computer is doing.

In any case, we haven't hit serious measure theory yet. The main reason for introducing kernels here is purely notational. It makes unnecessary a lot of useless discussion of special cases. It allows us to write expressions like

$$E\{g(X_n) | X_{n-1} = x\} = \int P(x, dy) g(y) \tag{2.4}$$

using one notation for all cases. Avoiding measure-theoretic notation leads to excruciating contortions.

Sometimes the distribution of X_n given X_{n-1} is a continuous distribution on $\mathbb{R}^d$ with density $f(y|x)$. Then the kernel is defined by

$$P(x, B) = \int_B f(y|x) \, dy$$

and (2.4) becomes

$$E\{g(X_n) | X_{n-1} = x\} = \int g(y) f(y|x) \, dy.$$

Readers who like boldface for "vectors" can supply the appropriate boldface. Since both x and y here are elements of $\mathbb{R}^d$, every variable is boldfaced. I don't like the "vectors are boldface" convention. It is just one more bit of distinguishing trivial special cases that makes it much harder to see what is common to all cases.

Often the distribution of X_n given X_{n-1} is more complicated. A common situation in MCMC is that the distribution is continuous except for an atom at x. The chain stays at x with probability $r(x)$ and moves with probability $1 - r(x)$, and when it moves the distribution is given by a density $f(y|x)$. Then (2.4) becomes

$$E\{g(X_n)|X_{n-1} = x\} = r(x)g(x) + [1 - r(x)] \int g(y) f(y|x)\, dy.$$

The definition of the kernel in this case is something of a mess

$$P(x, B) = \begin{cases} r(x) + [1 - r(x)] \int_B f(y|x)\, dy, & x \in B \\ [1 - r(x)] \int_B f(y|x)\, dy, & \text{otherwise} \end{cases} \tag{2.5}$$

This can be simplified by introducing the *identity kernel* (yet more measure-theoretic notation) defined by

$$I(x, B) = \begin{cases} 1, & x \in B \\ 0, & x \notin B \end{cases} \tag{2.6}$$

which allows us to rewrite (2.5) as

$$P(x, B) = r(x)I(x, B) + [1 - r(x)] \int_B f(y|x)\, dy.$$

We will see why the identity kernel has that name a bit later.

Another very common case in MCMC has the distribution of X_n given X_{n-1} changing only one component of the state vector, say the i-th. The Gibbs update discussed in Chapter 1 is an example. The distribution of the i-th component has a density $f(y|x)$, but now x is an element of $\mathbb{R}^d$ and y is an element of $\mathbb{R}$ (not $\mathbb{R}^d$). Then (2.4) becomes

$$E\{g(X_n)|X_{n-1} = x\} = \int g(x_1, \ldots, x_{i-1}, y, x_{i+1}, \ldots, x_d) f(y|x)\, dy.$$

The notation for the kernel is even uglier unless we use "probability is a special case of expectation." To obtain the kernel just take the special case where g is the indicator function of the set B.

The virtue of the measure-theoretic notation (2.4) is that it allows us to refer to all of these special cases and many more without getting bogged down in a lot of details that are irrelevant to the point under discussion. I have often wondered why this measure-theoretic notation isn't introduced in lower

level courses. It would avoid tedious repetition, where first we woof about the discrete case, then the continuous case, even rarely the mixed case, thus obscuring what is common to all the cases. One can use the notation without knowing anything about measure-theoretic probability. Just take (2.4) as the definition of the notation. If you understand what expectations mean in the model at hand, then you can write out what the notation means in each case, as we have done above. Regardless of whether you think this would be a good idea in lower level courses, or not, I hope you are convinced that the notation is necessary in dealing with Markov chains. One would never see the forest for the trees without it.

2.1.3 Existence of Infinite Random Sequences

Transition probabilities do not by themselves define the probability law of the Markov chain, though they do define the law conditional on the initial position, that is, given the value of X_1. In order to specify the unconditional law of the Markov chain we need to specify the *initial distribution* of the chain, which is the marginal distribution of X_1.

If λ is the initial distribution and P is the transition kernel and g_1, ..., g_n are any real-valued functions, then

$$\begin{aligned} &E\{g_1(X_1) \ldots g_n(X_n)\} \\ &\qquad = \int \cdots \int \lambda(dx_1) P(x_1, dx_2) \cdots P(x_{n-1}, dx_n) g_1(x_1) \cdots g_n(x_n) \end{aligned}$$

provided the expectation exists. This determines the joint probability distribution of X_1, ..., X_n for any n. Just take the special case where the g_i are indicator functions.

Let Q_n denote the probability distribution of X_1, ..., X_n, a measure on the cartesian product S^n, where S is the state space. The Q_n are called the *finite-dimensional* distributions of the infinite random sequence X_1, X_2, ... The finite-dimensional distributions satisfy the obvious consistency property: $Q_n(A) = Q_{n+1}(A \times S)$. It is a theorem of measure-theoretic probability (Fristedt and Gray 1997, Theorem 3 of Chapter 22 and Definition 10 of Chapter 21) that for any consistent sequence of finite-dimensional distributions, there exists a unique probability measure Q_∞ for the infinite sequence such that Q_∞ agrees with the finite-dimensional distributions, that is, if A is a measurable set in S^n and

$$B = \{\, (x_1, x_2, \ldots) \in S^\infty : (x_1, \ldots, x_n) \in A \,\},$$

then $Q_n(A) = Q_\infty(B)$.

We will only rarely refer explicitly or even implicitly to Q_∞. One place where it cannot be avoided is the strong law of large numbers, which says that the set of infinite sequences $(X_1, X_2, \ldots)$ having the property that $\overline{X}_n \to \mu$ has probability one, the probability here referring to Q_∞, since it refers to probabilities on the space of infinite sequences. But mostly we deal only with

finite-dimensional distributions. The CLT, for example, is a statement about finite-dimensional distributions only.

Anyway, this issue of Q_∞ has nothing to do particularly with Markov chains. It is needed for the SLLN in the i. i. d. case too. If you are not bothered by the SLLN for i. i. d. random sequences, then the SLLN for Markov chains should not bother you either. The measure-theoretic technicalities are exactly the same in both cases.

2.2 Transition Probabilities as Operators

When the state space is finite, we have seen that the transition probabilities form a matrix, an $d \times d$ matrix if the state space has d points. From linear algebra, the reader should be familiar with the notion that a matrix represents a linear operator. This is true for Markov transition matrices as well. Actually, we will see it represents two different linear operators.

In the general state space case, transition probabilities also represent linear operators. In this case the vector spaces on which they operate are infinite-dimensional. We do not assume the reader should be familiar with these notions and so develop what we need of this theory to work with Markov chains.

2.2.1 Finite State Space

Right Multiplication

When the state space S is finite (2.4) becomes

$$E\{g(X_n)|X_{n-1} = x\} = \sum_{y \in S} P(x, y)g(y).$$

Although the notation is unusual, the right hand side corresponds to the matrix multiplication of the matrix P on the right by the "column vector" g. Using this notation we write the function defined by the right hand side as Pg. Hence we have

$$Pg(x) = E\{g(X_n)|X_{n-1} = x\}.$$

If we were fussy, we might write the left hand side as $(Pg)(x)$, but the extra parentheses are unnecessary, since the other interpretation of $Pg(x)$, that P operates on the real number $g(x)$, is undefined.

As mentioned above, the vector space of all real-valued functions on S is denoted $\mathbb{R}^S$. The operation of right multiplication defined above takes a function g in $\mathbb{R}^S$ to another function Pg in $\mathbb{R}^S$. This map $R_P : g \mapsto Pg$ is a linear operator on $\mathbb{R}^S$ represented by the matrix P. When we are fussy, we distinguish between the matrix P and the linear operator R_P it represents, as is common in introductory linear algebra books (Lang 1987, Chapter IV). But none of the Markov chain literature bothers with this distinction. So we will bother with making this distinction only for a little while. Later we will just write P instead of R_P as all the experts do, relying on context to make it clear whether P means

a matrix or a linear operator. We don't want the reader to think that making a clear distinction between the matrix P and the linear operator R_P is essential. Holding fast to that notational idiosyncrasy will just make it hard for you to read the literature.

Left Multiplication

A probability distribution on S is also determines a vector in $\mathbb{R}^S$. In this case the vector is the probability mass function $\lambda(x)$. If X_{n-1} has the distribution λ, then the distribution of X_n is given by

$$\Pr(X_n = y) = \sum_{x \in S} \lambda(x) P(x, y). \tag{2.7}$$

Again we can recognize a matrix multiplication, this time of the matrix P on the left by the "row vector" λ. Using this notation we write the probability distribution defined by the right hand side as λP. and hence have

$$\lambda P(y) = \Pr(X_n = y),$$

when X_{n-1} has the distribution λ. Again if we were fussy, we might write the left hand side as $(\lambda P)(y)$, but again the extra parentheses are unnecessary, since the other interpretation of $\lambda P(y)$, that $P(y)$ operates on λ, is undefined because $P(y)$ is undefined.

Equation (2.7) makes sense when λ is an arbitrary element of $\mathbb{R}^S$, in which case we say it represents a *signed measure* rather than a probability measure. Thus the matrix P also represents another linear operator on $\mathbb{R}^S$, the operator $L_P : \lambda \mapsto \lambda P$. Note that L_P and R_P are not the same operator, because P is not a symmetric matrix, so right and left multiplication produce different results.

When we are not being pedantic, we will usually write P instead of L_P or R_P. So how do we tell these two operators apart? In most contexts only one of the two is being used, so there is no problem. In contexts where both are in use, the notational distinction between Pf and λP helps distinguish them.

Invariant Distributions

Recall from Section 1.5 that a probability distribution π is an *invariant* distribution for a specified transition probability matrix P if the Markov chain that results from using π as the initial distribution is stationary. (An invariant distribution is also called a *stationary* or an *equilibrium* distribution.) Because the transition probabilities are assumed stationary, as we always do, it is enough to check that $X_{n-1} \sim \pi$ implies $X_n \sim \pi$. But we have just learned that $X_{n-1} \sim \lambda$ implies $X_n \sim \lambda P$. Hence we can use our new notation to write the characterization of invariant distributions very simply: a probability distribution π is invariant for a transition probability matrix P if and only if $\pi = \pi P$.

Recall from Section 1.7 that the "first task in MCMC" is to find a Markov update mechanism that preserves a specified distribution. Now we can state

that in notation. We are given a distribution π. The "first task" is to find one transition probability matrix P such that $\pi = \pi P$. Often, we want to find several such matrices or kernels, intending to combine them by composition or mixing.

Matrix Multiplication (Composition of Operators)

The distribution of X_{n+2} given X_n is given by

$$\Pr(X_{n+2} = z | X_n = x) = \sum_{y \in S} P(x, y) P(y, z).$$

Now we recognize a matrix multiplication. The right hand side is the (x, z) entry of the matrix P^2, which we write $P^2(x, z)$. Carrying the process further we see that

$$\Pr(X_{n+k} = z | X_n = x) = P^k(x, z),$$

where $P^k(x, z)$ denotes the (x, z) entry of the matrix P^k.

We can use these operations together. $P^k g$ is the conditional expectation of $g(X_{n+k})$ given X_n, and λP^k is the marginal distribution of X_{n+k} when X_n has marginal distribution λ.

We also want to use this operation when the transition probability matrices are different. Say $P(x, y)$ and $Q(x, y)$ are two transition probability matrices, their product is defined in the obvious way

$$(PQ)(x, z) = \sum_{y \in S} P(x, y) Q(y, z).$$

We met this object in Chapter 1 under the name of the composition of P and Q, which we wrote as PQ, anticipating that it would turn out to be a matrix multiplication. The reason for calling it "composition" is that it is functional composition when we think of P and Q as linear operators. Obviously, $(PQ)g = P(Qg)$. This translates to

$$R_{PQ} = R_P \circ R_Q \tag{2.8a}$$

when we use the notation R_P for the linear operator $f \mapsto Pf$. It translates to

$$L_{PQ} = L_Q \circ L_P \tag{2.8b}$$

when we use the notation L_P for the linear operator $\lambda \mapsto \lambda P$. In both cases matrix multiplication represents functional composition, but note that P and Q appear in opposite orders on the right hand sides of (2.8a) and (2.8b), the reason being the difference between right and left multiplication.

Convex Combinations of Matrices (Mixing)

Besides multiplication of matrices, linear algebra also defines the operations of matrix addition and multiplication of a matrix by a scalar. Neither of these

operations turns a Markov matrix into a Markov matrix, because matrix addition loses property (2.2b) and multiplication by a negative scalar loses property (2.2a).

If we use both operations together, we can get an operation that preserves Markovness. Transition probability matrices are elements of the vector space $\mathbb{R}^{S\times S}$, a d^2-dimensional vector space if the state space S has d elements. Addition of matrices is just vector addition in this vector space. Multiplication of a matrix by a scalar is just scalar multiplication in this vector space. If P_1, ..., P_k are elements of any vector space, and a_1, ..., a_k are scalars, then

$$P = a_1 P_1 + \cdots + a_k P_k \tag{2.9}$$

is called a *linear combination* of the P_i. If the a_i also satisfy $\sum_i a_i = 1$, a linear combination is called an *affine combination*. If the a_i also satisfy $a_i \geq 0$ for each i, an affine combination is called a *convex combination*.

For Markov matrices P_1, ..., P_k,

- if P in (2.9) is Markov, then linear combination is affine,
- conversely, if the linear combination is convex, then P is Markov.

(Exercise 2.2).

Convex combinations correspond exactly to the operation of mixing of update mechanisms (also called "random scan") described in Section 1.7. if there are k update mechanisms, the i-th mechanism described by transition probability matrix P_i, and we choose to execute the i-the mechanism with probability a_i, then the transition probability matrix for the combined update mechanism is given by (2.9). In order to be probabilities the a_i must be nonnegative and sum to one, which is exactly the same as the requirement for (2.9) to be a convex combination. We would have called this notion "convex combination" rather than "mixture," but that seemed too long for everyday use.

2.2.2 General State Space

Now we turn to general state spaces, and kernels replace matrices. The objects on which the kernels operate on the left and right now are very different, a function on the state space (an object for right multiplication) is not at all like a measure on the state space (and object for left multiplication).

Signed Measures

In the discrete case we wanted to talk about measures that were not probability measures. We need a similar notion for general state spaces. A real-valued measure on a measurable space[1] $(S, \mathcal{B})$ is a function $\mu : \mathcal{B} \to \mathbb{R}$ that is countably additive.

[1] A *measurable space* is a pair $(S, \mathcal{B})$ consisting of a set S, in this case the state space, and a σ-field of subsets of S. The elements of $\mathcal{B}$ are called the *measurable sets* or, when we are talking about probabilities, *events*. So $\mathcal{B}$ is just the set of all possible events.

Although not part of the definition, it is a theorem of real analysis that μ is actually a bounded function (Rudin 1987, Theorem 6.4), that is, there are constants a and b such that $a \le \mu(B) \le b$ for all $B \in \mathcal{B}$. If $\mu(B) \ge 0$ for all measurable sets B, then we say μ is a *positive* measure. The general case, in which $\mu(B)$ takes values of both signs, is sometimes called a real *signed* measure, although strictly speaking the "signed" is redundant.

Another theorem (Rudin 1987, Theorem 6.14) says that there exists a partition[2] of the state space into two measurable sets A_1 and A_2 such that

$$\begin{aligned} \mu(B) &\le 0, \qquad B \subset A_1 \\ \mu(B) &\ge 0, \qquad B \subset A_2 \end{aligned}$$

This is called the *Hahn decomposition* of the state space S. Then the measures μ^+ and μ^- defined by

$$\begin{aligned} \mu^-(B) &= -\mu(B \cap A_1), && B \in \mathcal{B} \\ \mu^+(B) &= \mu(B \cap A_2), && B \in \mathcal{B} \end{aligned}$$

are both positive measures on S and they are mutually singular. Note that $\mu = \mu^+ - \mu^-$, which is called the *Jordan decomposition* of μ. It is entirely analogous to the decomposition $f = f^+ - f^-$ of a function into its positive and negative parts. The measure $|\mu| = \mu^+ + \mu^-$ is called the *total variation* of μ. And $\|\mu\| = |\mu|(S)$ is called the *total variation norm* of μ.

Let $\mathcal{M}(S)$ denote the set of all real signed measures on S. From the Jordan decomposition, we see that every element of $\mathcal{M}(S)$ is a difference of positive finite measures, hence a linear combination of probability measures. Thus $\mathcal{M}(S)$ is the vector space spanned by the probability measures. Hence it is the proper replacement for $\mathbb{R}^S$ in our discussion of left multiplication in the discrete case.

Norms and Operator Norm

For any vector space V, a function $x \mapsto \|x\|$ from V to $[0, \infty)$ is called a *norm* on V if it satisfies the following axioms (Rudin 1987, p. 95)

(a) $\|x + y\| \le \|x\| + \|y\|$ for all $x, y \in V$,

(b) $\|ax\| = |a| \cdot \|x\|$ for all $a \in \mathbb{R}$ and $x \in V$, and

(c) $\|x\| = 0$ implies $x = 0$.

Axiom (a) is called the *triangle inequality*. The pair $(V, \|\cdot\|)$ is called a *normed vector space* or a *normed linear space*.

Total variation norm makes $\mathcal{M}(S)$ a normed vector space. We do need to verify that total variation norm does satisfy the axioms for a norm (Exercise 2.3).

Denote the set of all linear operators on a vector space V by $L(V)$. Then $L(V)$ is itself a vector space if we define vector addition by

$$(S + T)(x) = S(x) + T(x), \qquad S, T \in L(V),\ x \in V \tag{2.10a}$$

[2]Partition means $A_1 \cap A_2 = \varnothing$ and $A_1 \cup A_2 = S$

and scalar multiplication by

$$(aT)(x) = aT(x), \qquad a \in \mathbb{R},\ T \in L(V),\ x \in V. \tag{2.10b}$$

These definitions are the obvious ones, arrived at almost without thinking. How else would you define the sum of two functions S and T except as the sum (2.10a)?

When V is normed, there is a natural corresponding norm for $L(V)$ defined by

$$\|T\| = \sup_{\substack{x \in V \\ x \neq 0}} \frac{\|Tx\|}{\|x\|} \tag{2.11}$$

Or, more precisely, we should say that (2.11) defines a norm for the subset of $L(V)$ consisting of T such that (2.11) is finite. We denote that subset $B(V)$, and call its elements the *bounded operators* on $L(V)$. The bounded operators are the well behaved ones.

A normed linear space is also a metric space, the metric being defined by $d(x, y) = \|x - y\|$. Hence we can discuss topological notions like continuity and convergence of sequences. A sequence $\{x_n\}$ in V converges to a point x if $\|x_n - x\| \to 0$. An operator $T \in L(V)$ is continuous at a point x if $Tx_n \to Tx$ (meaning $\|Tx_n - Tx\| \to 0$) for every sequence $\{x_n\}$ converging to x. Since $Tx_n - Tx = T(x_n - x)$ by linearity, a linear operator T is continuous at x if and only if it is continuous at zero. Thus linear operators are either everywhere continuous or nowhere continuous. A linear operator T is continuous if and only if it is bounded (Rudin 1991, Theorem 1.32). Thus the unbounded operators are nowhere continuous, a fairly obnoxious property. If V is finite-dimensional, then every operator in $L(V)$ is bounded (Halmos 1958, p. 177). But if V is infinite-dimensional, there are lots of unbounded operators.

Let's check that operator norm satisfies the norm axioms. Essentially it satisfies the axioms because vector norm does. For the triangle inequality

$$\begin{aligned}
\|S + T\| &= \sup_{\substack{x \in V \\ x \neq 0}} \frac{\|Sx + Tx\|}{\|x\|} \\
&\leq \sup_{\substack{x \in V \\ x \neq 0}} \frac{\|Sx\| + \|Tx\|}{\|x\|} \\
&\leq \sup_{\substack{x \in V \\ x \neq 0}} \frac{\|Sx\|}{\|x\|} + \sup_{\substack{y \in V \\ y \neq 0}} \frac{\|Ty\|}{\|y\|} \\
&= \|S\| + \|T\|
\end{aligned}$$

The first inequality is the triangle inequality for the vector norm. The second inequality is subadditivity of the supremum operation. For any functions f and g on any set S

$$f(x) + g(x) \leq f(x) + \sup_{y \in S} g(y),$$

so taking the sup over x gives

$$\sup_{x \in S} [f(x) + g(x)] \le \sup_{x \in S} f(x) + \sup_{y \in S} g(y).$$

For axiom (b),

$$\|aT\| = \sup_{\substack{x \in V \\ x \neq 0}} \frac{\|aTx\|}{\|x\|} = \sup_{\substack{x \in V \\ x \neq 0}} \frac{|a| \cdot \|Tx\|}{\|x\|} = a\|T\|.$$

Finally, for axiom (c), $\|T\| = 0$ only if $\|Tx\| = 0$ for all $x \in V$, but axiom (c) for vector norm implies $\|Tx\| = 0$ if and only if $Tx = 0$. Thus $\|T\| = 0$ implies that T is the operator that maps every x to 0. And this operator is indeed the zero of the vector space $L(V)$, because then

$$(S + T)(x) = S(x) + T(x) = S(x) + 0 = S(x), \qquad x \in V$$

so $S + T = S$ for all $S \in L(V)$, and this is the property that makes T the zero of the vector space $L(V)$.

Operator norm satisfies two important inequalities. The first

$$\|Tx\| \le \|T\| \cdot \|x\| \tag{2.12}$$

follows immediately from the definition (2.11).

The second involves the notion of operator "multiplication," which is defined as composition of functions: ST is shorthand for $S \circ T$. As we saw above, this agrees with our usual notation in the finite-dimensional case: matrix multiplication corresponds to functional composition of the corresponding operators. With this notion of multiplication $B(V)$ becomes an *operator algebra*. A *vector algebra*, also called *linear algebra*, is a vector space in which a multiplication is defined. The reason the subject "linear algebra" is so called is because matrices form a vector algebra.

The second important inequality is

$$\|ST\| \le \|S\| \cdot \|T\|. \tag{2.13}$$

I call (2.13) the *Banach algebra inequality* because it is one of the defining properties of a Banach algebra. Since we will have no need of Banach algebras in this course, it is a really horrible name. Maybe we should call it the *mumble mumble* inequality. Whatever we call it, the proof is a trivial consequence of operator "multiplication" actually being functional composition.

$$\|ST\| = \sup_{\substack{x \in V \\ x \neq 0}} \frac{\|S(Tx)\|}{\|x\|} \le \sup_{\substack{x \in V \\ x \neq 0}} \frac{\|S\| \cdot \|Tx\|}{\|x\|} = \|S\| \cdot \|T\|$$

where the inequality is just (2.12).

Left Multiplication

If λ is a probability measure on the state space, and X_{n-1} has distribution λ, then the distribution of X_n is given by

$$\lambda P(A) = \int \lambda(dx) P(x, A). \tag{2.14}$$

This is no longer a matrix multiplication, but it does define a linear operator, because integration is a linear operation. Using the Jordan decomposition, we see that (2.14) makes sense for any $\lambda \in \mathcal{M}(S)$. Hence (2.14) defines a linear operator on $\mathcal{M}(S)$.

The next question to answer is whether it is a well-behaved operator, that is, whether it is bounded. In fact, it is. For any Markov kernel P, let L_P denote the linear operator on $\mathcal{M}(S)$ defined by $\lambda \mapsto \lambda P$. Then $\|L_P\| = 1$ (Exercise 2.5).

As was the case for discrete state spaces, a probability measure π is invariant for a transition probability kernel if and only if $\pi = \pi P$. This is an integral equation

$$\pi(B) = \int \pi(dx) P(x, B), \qquad B \in \mathcal{B}$$

but we do not usually attempt to find a P that satisfies this equation by direct means. Usually we exploit some trick (if this is mysterious, it will all become clear in the next chapter).

Function Spaces

Before we can define the analog to right matrix multiplication, we must decide what space the linear operator $f \mapsto Pf$ is to act upon. There are a number of possibilities. The ones we will consider are the so-called $L^p(\pi)$ spaces, where $1 \le p \le \infty$ and π is a probability measure.

The $L^p(\pi)$ norm of a real-valued measurable function f on the probability space $(S, \mathcal{B}, \pi)$ is defined by

$$\|f\|_p = \left(\int |f(x)|^p \pi(dx) \right)^{1/p}$$

when $1 \le p < \infty$. The vector space $L^p(\pi)$ is the set of all measurable functions f on $(S, \mathcal{B})$ such that $\|f\|_p < \infty$. It is easy to see that the $L^p(\pi)$ norm satisfies axiom (b) for norms. That it satisfies axiom (a) is a well-known inequality called *Minkowski's inequality* (Rudin 1987, Theorem 3.5). It is also easy to see that the $L^p(\pi)$ norm fails to satisfy axiom (c), since $\|f\|_p = 0$ only implies $\pi\{|f(X)| > 0\} = 0$. If S is not discrete, there must be nonempty sets of probability zero, and any function f that is zero except on a set of probability zero has $\|f\|_p = 0$.

In order to make $L^p(\pi)$ a normed vector space, we need to work around this problem by redefining equality in $L^p(\pi)$ to mean equal except on a set of probability zero. Then axiom (c) is satisfied too, and $L^p(\pi)$ is a legitimate normed vector space.

We also redefine what we mean by inequalities as well. The statement $f \le g$ only means $f(x) \le g(x)$ except on a set of probability zero, and similarly for the other inequality relations. The space $L^\infty(\pi)$ consists of the bounded elements of $L^p(\pi)$, that is $|f| \le c$ for some real number c. Following the conventions for L^p spaces, this only means $|f(x)| \le c$ except on a set of probability zero. The $L^\infty(\pi)$ norm is the smallest c that will work

$$\|f\|_\infty = \inf\{\, c > 0 : \pi\{|f(X)| > c\} = 0 \,\}$$

This is also now easily seen to satisfy the axioms for norms, axiom (c) holding because we consider $f = 0$ if it is zero except on a set of probability zero. Thus all the $L^p(\pi)$ spaces for $1 \le p \le \infty$ are normed vector spaces[3].

An useful fact about $L^p(\pi)$ spaces is that $1 \le p \le q \le \infty$ implies $L^p(\pi) \supset L^q(\pi)$ (Exercise 2.12). (Warning: this uses the fact that π is a bounded measure. It is not true otherwise. However, we will be interested only in the case where π is a probability measure.)

Right Multiplication

We are finally ready to define "multiplication" of a kernel on the right by a function. If f is any nonnegative measurable function on $(S, \mathcal{B})$,

$$Pf(x) = \int P(x, dy) f(y) \tag{2.15}$$

is well-defined, though possibly $+\infty$. So we have no trouble defining "right multiplication" for nonnegative functions.

General functions are a bit more tricky. The issue is whether we can even define Pf for f that are both positive and negative. The trouble is that we want f to be integrable with respect to an infinite collection of probability measures, $P(x, \cdot)$, $x \in S$.

It turns out that we get everything we need, if π is an invariant probability measure for a transition probability kernel P and we use integrability with respect to π as our criterion. For $f \in L^1(\pi)$, define

$$g(x) = \int P(x, dy) |f(y)|.$$

Then

$$\begin{aligned} \int \pi(dx) g(x) &= \iint \pi(dx) P(x, dy) |f(y)| \\ &= \int \pi(dx) |f(y)| \\ &= \|f\|_1 \end{aligned} \tag{2.16}$$

[3]Actually they are Banach spaces, a *Banach space* being a complete normed vector space, where *complete* means every Cauchy sequence converges. But that will not play any role in the theory used in this course.

because $\pi = \pi P$. The interchange of the order of integration going from line 2 to line 3 is the conditional Fubini theorem (Fristedt and Gray 1997, Theorem 2 of Chapter 22). Hence the set

$$B = \{\, x \in S : g(x) < \infty \,\}.$$

satisfies $\pi(B^c) = 0$, because if g were infinite on a set of positive probability, the integral (2.16) would be infinite. This means we can define $Pf(x)$ by (2.15) for $x \in B$ and arbitrarily (say $Pf(x) = 0$) for $x \in B^c$ and have a function well defined in the $L^p(\pi)$ sense. Since $L^p(\pi) \subset L^1(\pi)$ for any $p > 1$, this makes the map $f \mapsto Pf$ well-defined on $L^p(\pi)$ for $1 \le p \le \infty$.

Now we want to show that the linear transformation $R_P : f \mapsto Pf$ actually maps $L^p(\pi)$ into $L^p(\pi)$. For $x \in B$ and $1 \le p < \infty$, Jensen's inequality gives

$$\begin{aligned} |Pf(x)|^p &= \left| \int P(x, dy) f(y) \right|^p \\ &\le \int P(x, dy) |f(y)|^p \end{aligned}$$

When we integrate both sides with respect to π, the fact that the left hand side is not defined for $x \in B^c$ does not matter because $\pi(B^c) = 0$. Hence

$$\begin{aligned} \|Pf\|_p^p &= \int \pi(dx) |Pf(x)|^p \\ &\le \iint \pi(dx) P(x, dy) |f(y)|^p \\ &= \int \pi(dy) |f(y)|^p \\ &= \|f\|_p^p \end{aligned}$$

Again $\pi = \pi P$ and the conditional Fubini theorem were used in going from line 2 to line 3.

The case $p = \infty$ is even simpler, for $x \in B$

$$\begin{aligned} |Pf(x)| &= \left| \int P(x, dy) f(y) \right| \\ &\le \int P(x, dy) |f(y)| \\ &\le \|f\|_\infty \int P(x, dy) \\ &= \|f\|_\infty \end{aligned}$$

Integrating with respect to π gives $\|Pf\|_\infty \le \|f\|_\infty$.

Thus we see that for $1 \le p \le \infty$ the linear transformation $R_P : f \mapsto Pf$ maps $L^p(\pi)$ into $L^p(\pi)$ and the corresponding operator norm satisfies

$$\|R_P\|_p = \sup_{\substack{f \in L^p(\pi) \\ f \neq 0}} \frac{\|R_P f\|_p}{\|f\|_p} \le 1. \tag{2.17}$$

In fact $\|R_P\|_p = 1$ because for $f \equiv 1$,

$$Pf(x) = \int P(x, dy) = 1 = f(x)$$

so $\|Pf\|_p = \|f\|_p$ for constant functions and the supremum in (2.17) is actually equal to one.

This has been an important section, so we summarize our results. If f is a measurable function from the state space to $[0, \infty]$, then $Pf(x)$ is well defined, though it may have the value $+\infty$. Since the set of functions on which this operation is defined is not a vector space, we cannot call P a linear operator here, but this notion is useful in various places in the theory of Markov chains.

If a kernel P has an invariant distribution π and $f \in L^p(\pi)$ for some $p \geq 1$, then Pf is a well defined element of $L^p(\pi)$. The linear operator $R_P : f \mapsto Pf$ is a bounded operator on $L^p(\pi)$ having operator norm equal to one.

General Kernels

In discrete state spaces, we wanted to discuss matrices that were not necessarily Markov. We need the analogous definitions for kernels. If $(S, \mathcal{B})$ is a measurable space, then a map K from $S \times \mathcal{B}$ to $\mathbb{R}$ is a *kernel* if

- for each fixed x the function $B \mapsto K(x, B)$ is a real signed measure, and
- for each fixed B the function $x \mapsto K(x, B)$ is a measurable function.

Multiplication of Kernels

The operation on kernels that is analogous to matrix multiplication is defined by

$$(K_1 K_2)(x, A) = \int K_1(x, dy) K_2(y, A).$$

Kernel multiplication is associative,

$$(K_1 K_2) K_3 = K_1 (K_2 K_3) \tag{2.18}$$

for any kernels K_1, K_2, and K_3, by the conditional Fubini theorem (Fristedt and Gray 1997, Theorem 2 of Chapter 22).

Kernel multiplication is not, in general, commutative: $K_1 K_2 = K_2 K_1$ may be false.

All of the results for composition and mixing of transition operators that we described in the discrete case carry over unchanged to the general case. In particular, multiplication of kernels corresponds to composition of operators (also called "fixed scan") in just the same way as we saw in (2.8a) and (2.8b). And a convex combination of Markov operators again produces a Markov operator and still corresponds to the operation of choosing an update mechanism at random and applying it (also called "random scan").

The Identity Kernel

The identity element any of the kernel operations is indeed the identity kernel defined back in (2.6). The identity kernel has connections with other notations widely used in probability. For fixed x, the measure $I(x, \cdot)$ is the probability measure concentrated at x, sometimes written δ_x, sometimes called the *Dirac measure*. For fixed A, the function $I(\cdot, A)$ is the indicator of the set A, more commonly written 1_A.

The identity kernel is the identity for kernel multiplication because

$$(IK)(x, A) = \int I(x, dy)K(y, A) = \int \delta_x(dy)K(y, A) = K(x, A),$$

and

$$(KI)(x, A) = \int K(x, dy)I(y, A) = \int K(x, dy)1_A(y) = \int_A K(x, dy) = K(x, A).$$

For this reason, we define $K^0 = I$ for any kernel K. Then the so-called Chapman-Kolmogorov equation

$$K^n = K^m K^{n-m}$$

holds whenever $0 \le m \le n$ as a direct consequence of the associative law (2.18).

The identity kernel is the identity for left multiplication of a kernel by a signed measure because

$$(\lambda I)(A) = \int \lambda(dx)I(x, A) = \int \lambda(dx)1_A(x) = \int_A \lambda(dx) = \lambda(A)$$

It is the identity for right multiplication of a kernel by a function because

$$(If)(x) = \int I(x, dy)f(y) = \int \delta_x(dy)f(y) = f(x).$$

Needless to say, the operators $L_P : \lambda \mapsto \lambda P$ and $R_P : f \mapsto Pf$ are the identity operators on the relevant vector spaces when P is the identity kernel.

The identity kernel is Markov, because, as we have seen $I(x, \cdot)$ is a probability measure, δ_x, for each x. If $X_n \sim \delta_x$, then $X_{n+1} \sim \delta_x$, because $\delta_x I = \delta_x$. Hence the chain never moves. Thus the identity kernel is the transition probability for the "maximally uninteresting chain" described in Example 1.4.

2.2.3 Hilbert Space Theory

Inner Product Spaces

An *inner product* on a complex vector space V is a map from $V \times V$ to $\mathbb{C}$, the value for the ordered pair of vectors x and y being written (x, y), that satisfies the following axioms (Halmos 1958, p. 121)

(a) $(x, y) = \overline{(y, x)}$,

(b) $(ax + by, z) = a(x, z) + b(y, z)$, for $a, b \in \mathbb{C}$,

(c) $(x, x) \geq 0$, and

(d) $(x, x) = 0$ implies $x = 0$.

where the overline in (a) denotes complex conjugation. An *inner product space* is a vector space equipped with an inner product.

For the most part, we will only be interested in real inner product spaces, in which case the complex conjugation in (a) does nothing and the scalars in (b) must be real. Since in applications we have no complex numbers, why should the theory involve them? The answer is eigenvalues and eigenvectors. Transition probability matrices are nonsymmetric and hence may have complex eigenvalues even though all their entries are real. So we will not be able to avoid mentioning complex inner product spaces. However, we will see they play a very minor role in Markov chain theory.

An inner product space is also a normed vector space with the norm defined by $\|x\| = \sqrt{(x, x)}$. It is easily verified that the norm axioms are implied by the inner product axioms (Exercise 2.6), the only bit of the proof that is nontrivial being the triangle inequality, which follows directly from

$$|(x, y)| \leq \|x\| \cdot \|y\|,$$

which is known to statisticians as the *Cauchy-Schwarz* inequality. It, of course, is proved exactly the same way as one proves that correlations are between -1 and 1.

Hilbert Spaces

A *Hilbert space* is a complete inner product space, where *complete* means every Cauchy sequence converges, a sequence $\{x_n\}$ being *Cauchy* if $\|x_m - x_n\| \to 0$ as $\min(m, n) \to \infty$. We will not develop any of the consequences of this definition, since they are well beyond the level of real analysis taken by most statistics graduate students, but we will steal a few results here and there from Hilbert space theory, explaining what they mean but blithely ignoring proofs.

One important fact about Hilbert space theory is the existence of the adjoint of an operator, which is analogous to the transpose of a matrix. If T is a bounded operator on a Hilbert space H. Then there is a unique bounded operator T^* on H that satisfies

$$(x, Ty) = (T^*x, y), \qquad x, y \in H$$

(Rudin 1991, Section 12.9). T^* is called the *adjoint* of T. If $T^* = T$, then T is said to be *self-adjoint*.

To see the connection between adjoints and transposes, equip the vector space $\mathbb{R}^S$ for some finite set S with the usual inner product

$$(f, g) = \sum_{x \in S} f(x)g(x). \tag{2.19}$$

A linear operator on $\mathbb{R}^S$ is represented by a matrix $M(x, y)$, the linear operator being $T_M : f \mapsto Mf$ (the same as the right multiplication we studied in Section 2.1.1 but with M not necessarily a transition probability matrix). Then

$$(f, T_M g) = \sum_{x \in S} \sum_{y \in S} f(x) M(x, y) g(y)$$

and

$$(T_M^* f, g) = \sum_{x \in S} \sum_{y \in S} g(x) M^*(x, y) f(y)$$

where M^* is the matrix that represents T_M^*. Clearly, M and M^* are transposes of each other.

For Markov chain theory, there are only two important Hilbert spaces. The first we have already met: $L^2(\pi)$ is a Hilbert space when the inner product is defined by

$$(f, g) = \int f(x) \overline{g(x)} \pi(dx). \tag{2.20}$$

That this defines an inner product (with the usual proviso that equality means only equality with probability one) is obvious. The completeness comes from the fact that every $L^p(\pi)$ is a complete metric space (Rudin 1987, Theorem 3.11). Usually we consider $L^p(\pi)$ a real Hilbert space, in which case the complex conjugate in (2.20) does nothing.

The reason why $L^2(\pi)$ is so important is that (2.20) is $\operatorname{Cov}\{f(X), g(X)\}$ in the special case when both variables have mean zero. In order to cater to this special case of interest to statisticians, we introduce the subspace of $L^2(\pi)$ that consists of mean-zero functions

$$L_0^2(\pi) = \left\{ f \in L^2(\pi) : \int f(x) \pi(dx) = 0 \right\}$$

Another characterization of $L_0^2(\pi)$ uses the notion of orthogonality. Vectors x and y in a Hilbert space are *orthogonal* if $(x, y) = 0$. If 1 represents the constant function equal to 1 almost surely, then we can also write

$$L_0^2(\pi) = \big\{ f \in L^2(\pi) : (f, 1) = 0 \big\}$$

Thus $L_0^2(\pi)$ is the subspace of $L^2(\pi)$ orthogonal to the constant functions. Since the linear function $f \mapsto (f, 1)$ is continuous, $L_0^2(\pi)$ is a topologically closed subspace of $L^2(\pi)$ and hence is also a Hilbert space.

Warning: The characterization of the adjoint as the transpose is incorrect for $L^2(\pi)$ even in the finite state space case. The reason is that (2.19) is not the inner product on $L^2(\pi)$. The inner product is defined by (2.20). The same formula applies to finite state spaces as for general state spaces (general includes finite). Exercise 2.9 derives the correct formula for the adjoint.

In the preceding section, we saw that the operator norm for the linear operator $f \mapsto Pf$ is exactly equal to one, no matter which $L^p(\pi)$ we have the operator act on. The Hilbert space $L^2(\pi)$ is no exception, but $L_0^2(\pi)$ is different. Reducing the domain of the operator cannot increase the norm, but may decrease it, the supremum in (2.17) being over a smaller set. The proof that the norm is exactly one no longer applies, because it used the fact that $Pf = f$ for constant functions f, and those functions are no longer in the domain. Thus when we consider $R_P : f \mapsto Pf$ an operator on $L_0^2(\pi)$ we have $\|R_P\|_2 \le 1$ with strict inequality now a possibility.

2.2.4 Time-Reversed Markov Chains

The measure-theoretic construction of infinite sequences of random variables discussed in Section 2.1.3, says that specification of the probability distribution of an infinite sequence is equivalent to specifying a consistent set of finite-dimensional distributions. This allows us to specify a stationary Markov chain as a doubly infinite sequence $\ldots$, X_{-2}, X_{-1}, X_0, X_1, X_2, $\ldots$. Specifying the distribution of the doubly infinite sequence is the same as specifying the joint distribution of X_n, X_{n+1}, $\ldots$, X_{n+k} for any $k > 0$. Stationarity implies that this joint distribution does not depend on n.

Two questions naturally arise about the time-reversed sequence. First, is it Markov? Second, what is its kernel? That the time-reversed sequence has the Markov property is a trivial consequence of conditional independence being a symmetric property, that is, the following three statements are equivalent.

- The future is independent of the past given the present.
- The past is independent of the future given the present.
- The past and future are independent given the present.

If this isn't mathy enough for you, here are some equations. What is to be shown is that

$$\begin{aligned} &E\{f(X_{n+1}, X_{n+2}, \ldots) g(X_{n-1}, X_{n-2}, \ldots) | X_n\} \\ &\qquad = E\{f(X_{n+1}, X_{n+2}, \ldots) | X_n\} E\{g(X_{n-1}, X_{n-2}, \ldots) | X_n\} \end{aligned} \tag{2.21}$$

for any functions f and g such that both sides are well defined. This says the σ-field generated by X_{n+1}, X_{n+2}, $\ldots$ (the future) and the σ-field generated by X_{n-1}, X_{n-2}, $\ldots$ (the past) are conditionally independent given the σ-field generated by X_n (the present) (Fristedt and Gray 1997, Definition 23 of Chapter 21).

The proof is

$$\begin{aligned}
&E\{f(X_{n+1}, X_{n+2}, \ldots) g(X_{n-1}, X_{n-2}, \ldots) | X_n\} \\
&\quad = E\{E[f(X_{n+1}, X_{n+2}, \ldots) g(X_{n-1}, X_{n-2}, \ldots) | X_n, X_{n-1}, X_{n-2}, \ldots] | X_n\} \\
&\quad = E\{g(X_{n-1}, X_{n-2}, \ldots) E[f(X_{n+1}, X_{n+2}, \ldots) | X_n, X_{n-1}, X_{n-2}, \ldots] | X_n\} \\
&\qquad = E\{g(X_{n-1}, X_{n-2}, \ldots) E[f(X_{n+1}, X_{n+2}, \ldots) | X_n] | X_n\} \\
&\qquad\quad = E\{f(X_{n+1}, X_{n+2}, \ldots) | X_n\} E\{g(X_{n-1}, X_{n-2}, \ldots) | X_n\}
\end{aligned}$$

The equality between lines 3 and 4 is the Markov property of the original chain running forwards in time. The other equalities are standard properties of conditional expectation. The equalities between lines 2 and 3 and between lines 4 and 5 are the property that functions of the conditioning variables can be taken outside a conditional expectation (Fristedt and Gray 1997, Problem 27 of Chapter 23). The equality between lines 1 and 2 is the general iterated conditional expectation formula (Fristedt and Gray 1997, Proposition 6 of Chapter 23).

By Propositions 25 and 27 of Chapter 23 in Fristedt and Gray (1997) (2.21) implies the Markov property for the time-reversed chain

$$E\{1_A(X_{n-1}) | X_n, X_{n+1}, X_{n+2}, \ldots\} = E\{1_A(X_{n-1}) | X_n\}.$$

Clearly, the time-reversed chain is also stationary, in particular, it has stationary transition probabilities. As to whether these transition probabilities are representable by a kernel, the answer is not necessarily, but usually. The issue is whether there exists a kernel P^* satisfying

$$\int_A \pi(dx) P^*(x, B) = \int_B \pi(dx) P(x, A), \qquad A, B \in \mathcal{B}, \tag{2.22}$$

(where $\mathcal{B}$ is the σ-field of the state space), that is, whether P^* exists as a *regular* conditional probability. Conditional probabilities always exist, but *regular* ones do not. The key is whether the state space is "nice" enough. If the state space is a so-called *Borel space*, then regular conditional probabilities (a. k. a. kernels) exist (Fristedt and Gray 1997, Theorem 19 of Chapter 21). Euclidean spaces $\mathbb{R}^d$ are Borel spaces, as are most (all?) other state spaces that arise in practical examples. So we may take it for granted that P^* exists. It is not, however, uniquely defined. $P^*(x, \cdot)$ can be defined arbitrarily for x in a set of π-probability zero without effecting (2.22). Thus there are many kernels P^*, all of which give the same probability law for the time-reversed chain.

Now that we have a kernel P^* for the time-reversed chain, we know that P^* and the marginal distribution π of X_n, which is invariant for both P and P^*, determine the probability distribution of the infinite sequence. We can also look at P^* as an operator. In particular, (2.22) is equivalent to

$$\int \pi(dx) P^*(x, dy) f(x) g(y) = \int \pi(dx) P(x, dy) g(x) f(y), \qquad f, g \in L^2(\pi) \tag{2.23}$$

by linearity of expectation and monotone convergence. In Hilbert space notation (2.23) is

$$(f, P^*g) = (Pf, g)$$

so now we see why the choice of P^* for the kernel of the time-reversed chain. It is the adjoint operator on $L^2(\pi)$.

2.2.5 Reversibility

A stationary Markov chain is *reversible* (also called *time-reversible*) if the doubly infinite sequence has the same probability distribution when time is reversed. We also say a kernel P is reversible with respect to π if (2.22) holds with $P^* = P$, that is,

$$\int_A \pi(dx)P(x, B) = \int_B \pi(dx)P(x, A), \qquad A, B \in \mathcal{B}. \tag{2.24}$$

Taking the case where A is the whole state space in (2.24) gives

$$\int \pi(dx)P(x, B) = \int_B \pi(dx) = \pi(B), \qquad B \in \mathcal{B},$$

which says $\pi P = \pi$. Thus (2.24) implies that π is invariant for P.

This is a very important principle.

> *If P is reversible with respect to π, then P preserves π.*

This will turn out to be our main method for accomplishing the "first task" of MCMC. Given a distribution π, how do we find Markov update mechanisms that preserve π? Answer: show they are reversible with respect to π.

If (2.24) holds, then so does (2.23) with $P^* = P$, that is,

$$\iint f(x)g(y)\pi(dx)P(x, dy) = \iint g(x)f(y)\pi(dx)P(x, dy), \qquad f, g \in L^2(\pi). \tag{2.25}$$

Hence P is self-adjoint.

> *P is reversible with respect to π, if and only if P is a self-adjoint operator on $L^2(\pi)$.*

We can rewrite (2.24) as

$$\Pr(X_n \in A \,\&\, X_{n+1} \in B) = \Pr(X_n \in B \,\&\, X_{n+1} \in A) \tag{2.26}$$

This gives yet another slogan.

> *A stationary Markov chain is reversible, if and only if X_n and X_{n+1} are exchangeable.*

For a discrete state space, transition probability matrix P and invariant distribution π, and state space S, the reversibility property is

$$\Pr(X_n = x \,\&\, X_{n+1} = y) = \Pr(X_n = y \,\&\, X_{t+1} = x),$$

or stated in terms of π and P

$$\pi(x)P(x,y) = \pi(y)P(y,x), \qquad x, y \in S, \tag{2.27}$$

a condition that is referred to as *detailed balance*. Our main tool for establishing that a particular transition probability P has a specified invariant distribution π will be verification of the detailed balance condition (2.27) and its counterparts for general state spaces. This is generally much easier than verifying $\pi P = \pi$ directly.

The analogue of (2.27) for general state spaces (2.26) involves probabilities of sets rather than points, and so does not lead to an analog of the detailed balance condition. You will sometimes see

$$\pi(dx)P(x,dy) = \pi(dy)P(y,dx)$$

called "detailed balance for general state spaces," but strictly speaking this is merely a shorthand for (2.24) or (2.25).

Exercises

2.1. Find an invariant distribution and show that it is unique for

(a) The random walk with reflecting barriers, Example 2.1.

(b) The modification of random walk with reflecting barriers, so that the first row of the transition probability matrix is $0, 1, 0, \ldots$ and the last row is modified similarly to $\ldots, 0, 1, 0$, the rest of the rows remaining as in (2.3).

2.2.

(a) Show that a linear combination of Markov transition operators is Markov if and only if the linear combination is an affine combination.

(b) Provide a counterexample that shows an affine combination of Markov transition operators that is not a convex combination but is still Markov.

2.3. Show that total variation norm satisfies the norm axioms.

2.4. Show that the map $L_P : \lambda \mapsto \lambda P$ is a linear operator on $\mathcal{M}(S)$ when P is a Markov kernel. There are two things to show, first that L_P is a linear transformation

$$L_P(a\lambda + b\mu) = aL_P(\lambda) + bL_P(\mu), \qquad a, b \in \mathbb{R},\ \lambda, \mu \in \mathcal{M}(S),$$

and second that L_P maps $\mathcal{M}(S)$ to $\mathcal{M}(S)$ (that is, λP is a countably additive set function).

2.5. Show that the map $L_P : \lambda \mapsto \lambda P$ satisfies $\|L_P\| = 1$ when P is a Markov kernel.

2.6. Show that $\|x\| = \sqrt{(x, x)}$ defines a norm, when (x, y) is an inner product. Include a proof of the Cauchy-Schwarz inequality for inner product spaces.

2.7. Show that the stationary scalar-valued AR(1) time series discussed in Examples 1.2 and 1.5 is reversible.

2.8.

(a) Show that the random walk with reflecting barriers of Example 2.1 is reversible.

(b) Show that the modified random walk of Problem 2.1 (b) is reversible.

(c) Show that the "maximally uninteresting chain" having the identity kernel as its kernel is reversible for any invariant distribution π.

2.9. Suppose P is a transition probability matrix on a finite state space S having invariant distribution π considered as a vector $\pi \in \mathbb{R}^S$. Find the formula for the adjoint of $R_P : f \to Pf$ considered as an operator on $L^2(\pi)$.

2.10. Find a Markov chain transition probability kernel that is *not* reversible.

2.11. Show that the Gibbs update described in Section 1.7 is reversible.

2.12. If π is a probability measure, show that $1 \leq p \leq q \leq \infty$ implies $L^p(\pi) \supset L^q(\pi)$.

Chapter 3

Basic Algorithms

This chapter describes the two basic "algorithms" for Markov chain Monte Carlo. The word "algorithms" is in quotation marks because what will actually be described are elementary update steps, bits of algorithm that change the state variable of the Markov chain in such a way so as to preserve a specified invariant distribution. These updates can be combined as described in Section 1.7.1 to make more complicated Markov transition mechanisms preserving the same invariant distribution. Repeating an update mechanism, basic or combined, again and again simulates a Markov chain. The two types of basic update step are the Gibbs update described in Section 1.7.2, the basic component of the "Gibbs sampler," and the Metropolis-Hastings-Green update, the basic component of the so-called "Metropolis-Hastings-Green algorithm."

3.1 Combining Update Mechanisms

3.1.1 Simple Composition and Mixing

We already met "composition" and "mixing" of elementary update mechanisms in Section 1.7.1 (commonly called "fixed scan" and "random scan" in the MCMC literature). Then in Chapter 2 we learned that composition corresponded to operator multiplication and mixing to a convex combination of operators.

The composition of update mechanisms that correspond to Markov transition kernels P_1, ..., P_d is the kernel $P_1 \cdots P_d$. The proof that if P_1, ..., P_d each preserves a distribution π, then so does the composition $P_1 \cdots P_d$ is trivial, just the fact that kernel multiplication is associative (2.18), so

$$\pi P_1 P_2 \cdots P_d = \pi P_2 \cdots P_d = \cdots = \pi P_d = \pi.$$

The mixture of update mechanisms that correspond to Markov transition kernels P_1, ..., P_d and uses the mixing distribution with probabilities a_1, ..., a_d is $\sum_i a_i P_i$. The proof that if P_1, ..., P_d each preserves a distribution π, then

so does the composition $\sum_i a_i P_i$ is just as trivial

$$\pi \left(\sum_{i=1}^{d} a_i P_i \right) = \sum_{i=1}^{d} a_i \pi P_i = \left(\sum_{i=1}^{d} a_i \right) \pi = \pi$$

No good theoretical reasons are known for choosing any particular mixing distribution, but the most common choice is the discrete uniform distribution $a_i = 1/d$, perhaps because of lack of imagination and spirit of adventure in MCMC practitioners.

3.1.2 Non-Finite Mixtures

Mixtures can use any mixing distribution, discrete or continuous. We need an argument that this is o. k. when the mixture is not finite.

Theorem 3.1. *Suppose μ is a probability distribution and for each z in the domain of μ there is a Markov kernel P_z satisfying $\pi = \pi P_z$, and suppose that the map $(z, x) \mapsto P_z(x, A)$ is jointly measurable for each A. Then*

$$Q(x, A) = \int \mu(dz) P_z(x, A)$$

defines a kernel Q that is Markov and satisfies $\pi = \pi Q$.

Proof. First we need to show that Q is a kernel. The double integral

$$\iint (\pi \times \mu)(dx, dz) P_z(x, A)$$

exists because $(x, z) \mapsto P_z(x, A)$ is jointly measurable and bounded. Hence $x \mapsto Q(x, A)$ is measurable (one of the conclusions of the Fubini theorem). To check that $A \mapsto Q(x, A)$ is a measure, we need only check countable additivity. If $A_n \uparrow A$, then

$$\lim_{n \to \infty} Q(x, A_n) = Q(x, A)$$

by the monotone convergence theorem.

The Markovness of Q is obvious. That π is invariant for Q is just the Fubini theorem

$$\begin{aligned} \int \pi(dx) Q(x, A) &= \int \pi(dx) \int \mu(dz) P_z(x, A) \\ &= \int \mu(dz) \int \pi(dx) P_z(x, A) \\ &= \int \mu(dz) \pi(A) \\ &= \pi(A) \end{aligned}$$

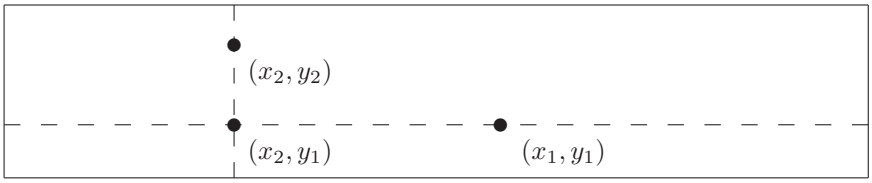
□

3.1.3 The Hit-and-Run Algorithm

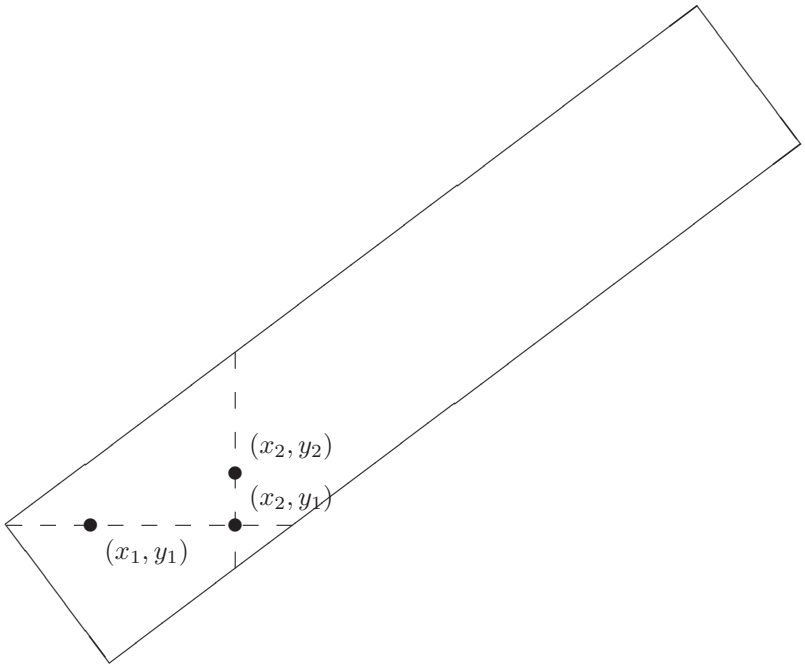
An example of a non-finite mixing distribution is the so-called "hit-and-run" algorithm (Bélisle, Romeijn, and Smith 1993; Chen and Schmeiser 1993). In its simplest form this algorithm is just a mixture of Gibbs updates that condition on a direction in the state space.

Example 3.1. Gibbs Sampling a Uniform Distribution.
Consider a bounded set A in $\mathbb{R}^d$. A conventional Gibbs sampler uses d updates, one for each coordinate. The i-th update updates the i-coordinate, giving it a new value simulated from its conditional distribution given the rest of the coordinates, which is uniform on some set.

If the region A is a rectangle parallel to the coordinate axes, the sampler produces i. i. d. samples. Starting at the point (x_1, y_1) in the figure, it simulates a new x value uniformly distributed over its possible range thereby moving to a position uniformly distributed along the horizontal dashed line, say to (x_2, y_1). Then it simulates a new y value uniformly distributed over its possible range thereby moving to a position uniformly distributed along the vertical dashed line, say to (x_2, y_2). This clearly produces a point uniformly distributed in the rectangle and uncorrelated with the previous point.



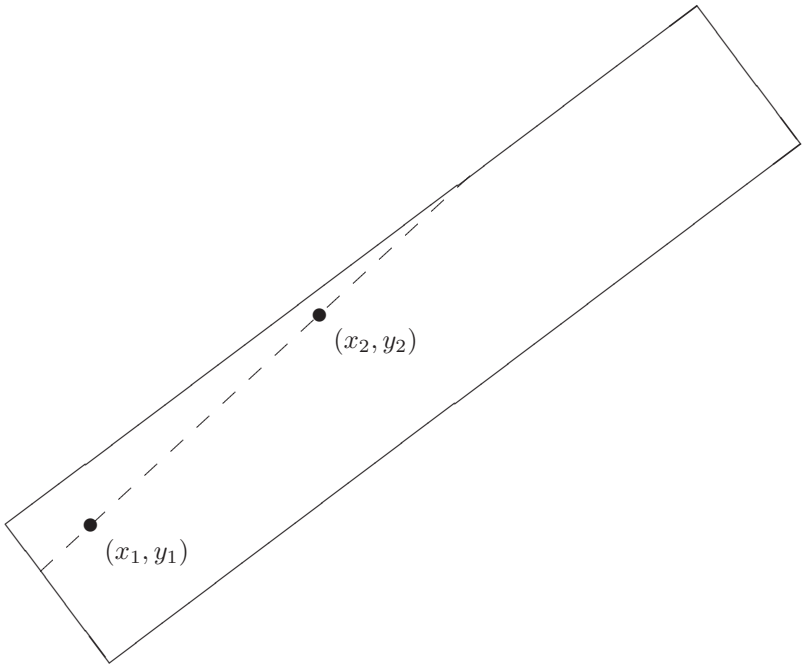
If the region A is not a rectangle parallel to the coordinate axes, then the Gibbs sampler has autocorrelation.



The update moves are still parallel to the coordinate axes. The possible range of values for each update is the intersection of a horizontal or vertical line, as the case may be, with A. Clearly, starting from the point (x_1, y_1) shown in the figure, it would take several moves to get into the upper half of the rectangle. Conclusion: the Gibbs sampler for the second rectangle is less efficient.

This example is an important toy problem. What it lacks in realism, it makes up for in simplicity. It is very easy to visualize this Gibbs sampler. Moreover, it does share some of the characteristics of realistic problems.

Example 3.2. Hit-and-Run Sampler for a Uniform Distribution. The hit-and-run sampler is almost the same as the Gibbs sampler, except that it moves in an arbitrary direction. A hit-and-run step simulates a random angle θ uniformly distributed between 0 and 2π. Then it simulates a new point uniformly distributed along the intersection of A and the line through the current point making angle θ.



It is obvious from the figure that some hit-and-run update steps move farther than Gibbs update steps. Some hit-and-run steps, not many, only those in a fairly small range of angles, can go from one end of the rectangle to the other. No Gibbs update step can do that.

Tentative conclusion: the hit-and-run sampler is more efficient than the Gibbs sampler. Is that right? When we think about the the comparison a bit more deeply we see that it is not at all obvious that hit-and-run is better. If we really want to know, we will have to do some simulation experiments and see.

3.1.4 Random Sequence Scans

Composition and mixing are the only ways to combine kernels, since multiplication and convex combination are the only operations that combine kernels to make other kernels, but we can mix a set of kernels that are themselves products of other kernels.

The best known example of combining composition and mixing is the so-called "random sequence scan." If there are d elementary update mechanisms having kernels P_1, ..., P_d, a random sequence scan chooses a random permutation $(k_1, k_2, \ldots, k_d)$ of the integers 1, 2, ..., d and then applies the updates in that order. We may use any distribution for the mixing distribution. If we let $\mathcal{P}$ denote the set of all $d!$ permutations, then a mixing distribution is given by real numbers a_k, $k \in \mathcal{P}$ that are nonnegative and sum to one. The random sequence scan update can then be described as follows.

1. Choose a random permutation $k = (k_1, \ldots, k_d) \in \mathcal{P}$, choosing k with

probability a_k.

2. Update the state using the composite update mechanism with kernel $P_{k_1} \ldots P_{k_d}$.

The composite update mechanism referred to in step 2 first does the update with kernel P_{k_1}, next the update with kernel P_{k_2}, and so forth. The whole random sequence scan update has kernel

$$P = \sum_{(k_1,\ldots,k_d)\in\mathcal{P}} a_k P_{k_1} \cdots P_{k_d}. \tag{3.1}$$

This is clearly a mixture, the mixing distribution being the uniform distribution on $\mathcal{P}$, and the kernels being mixed having the form $P_{k_1} \cdots P_{k_d}$.

When $a_k = 1/d!$ for all k, we say we are using a *uniform random sequence scan*, but the "uniform" is often dropped. As with the simple random scan, the uniform mixing distribution seems to be the default. An efficient procedure for producing uniform random permutations is given by Knuth (1998, p. 145). It uses computer memory and time proportional to d to generate the random permutation. Since it also takes time proportional to d to execute the scan, this is a minor issue, but there is some reason to consider random sequence scans that don't require additional memory proportional to d.

For example, we could choose uniformly at random from among the $2(d-1)$ permutations that cycle through the integers in normal or reversed order. With four variables these permutations are

1234	2341	3412	4123
4321	3214	2143	1432

This random sequence scan uses only two random variates per iteration, one to decide whether to cycle forward or backward and one to decide which update to start with. The uniform random sequence scan needs $d - 1$ random variates to generate a random permutation.

3.1.5 Auxiliary Variable Random Sequence Scans

Random scan and random sequence scan have an odd property when used with Gibbs updates. Gibbs updates are idempotent, that is, they satisfy $P^2 = P$ (Exercise 3.1). Thus whenever a scan starts with the same update that ended the preceding scan, no progress is made, but we cannot just omit the useless update, because then we would not have a Markov chain. For example if there are two updates with kernels P_1 and P_2 and we are using simple random scan and the first 10 updates are $P_1P_2P_1P_1P_2P_1P_1P_2P_1P_1$, then the distribution of X_{10} given X_0 is

$$P_1P_2P_1P_1P_2P_1P_1P_2P_1P_1 = P_1P_2P_1P_2P_1P_2P_1$$

But we cannot use the kernel on the right hand side, because we must do 10 elementary updates and output the state X_n after each one.

Of course, this problem only occurs in $1/d$ scans on average, so is not serious when d is large. Even when d is small, it does not affect correctness only efficiency. Still there is some reason to see whether we can find a random sequence scan that never repeats an update consecutively.

To accomplish this we need a new idea: let the random sequence we choose depend on the preceding one. If this is not to destroy the Markov property, we must enlarge the state space to include the scan sequence and verify that we still have a Markov chain with the desired invariant distribution. This trick of enlarging the state space is widely used in MCMC under the name "auxiliary variable methods." We will see it again and again.

Suppose we try choosing a scan sequence uniformly at random from all possible scans that do not begin with same elementary update that was the end of the preceding scan, so there are no repeats of elementary updates. Then the scan chosen depends on the index of the last elementary update of the preceding scan. In order to continue using Markov chain theory, we must add that index to the state space.

> *You can do anything in MCMC, but everything the update depends on must be part of the state.*

If the original state space was S, then the enlarged state space is $D \times S$, where $D = \{1, \ldots, d\}$ is the index set of the updates. The Markov chain we simulate will have the form (I_n, X_n), $i = 1, 2, \ldots$, where $I_n \in D$ and $X_n \in S$. The "auxiliary variable random sequence scan" update can now be described as follows.

- Choose a scan sequence $k_1, \ldots, k_d$ uniformly from the permutations of $(1, \ldots, d)$ not beginning with I. Set $I = k_d$.
- Update X using the update mechanism with kernel $P_{k_1} \ldots P_{k_d}$.

In a Markov chain problem, the "given" is a probability distribution π on S that we want to study. If we had not enlarged the state space, π would have been the invariant distribution of our Markov chain. Now, however, the invariant distribution of the chain (assuming it has one) will be a distribution on $D \times S$, since that is now the state space. In order for the new Markov chain to be of any use in learning about π, we need the X_n to still have marginal distribution π. Thus the marginal for X of the invariant distribution should be π. Since the update mechanism for X preserves π regardless of the index I and all of the index values are treated the same, it stands to reason that the invariant distribution if $\mu \times \pi$ where $\mu(i) = 1/d$.

We must, of course, check that this guess is correct. The kernel of the update on the enlarged state space can be written

$$P\big((i, x), \{j\} \times A\big) = \frac{1}{(d-1) \cdot (d-1)!} \sum_{\substack{(k_1, \ldots, k_d) \in \mathcal{P} \\ k_1 \neq i \\ k_d = j}} P_{k_1} \ldots P_{k_d}(x, A),$$

where $\mathcal{P}$ is the set of all permutations of indices as in Section 3.1.4, the factor $(d-1)\cdot(d-1)! = d! - (d-1)!$ being the number of permutations that do not start with i. If we left multiply by $\mu \times \pi$, we get

$$\begin{aligned}
&\frac{1}{d}\sum_{i=1}^{d}\int \pi(dx)P\big((i,x),\{j\}\times A\big) \\
&\qquad = \frac{1}{(d-1)\cdot d!}\sum_{i=1}^{d}\sum_{\substack{(k_1,\ldots,k_d)\in\mathcal{P}\\ k_1\neq i\\ k_d=j}}\int \pi(dx)P_{k_1}\ldots P_{k_d}(x,A) \qquad (3.2)\\
&\qquad = \pi(A)\frac{1}{d}
\end{aligned}$$

because each P_k preserves π so the integral is $\pi(A)$ and the result must integrate to one with respect to $\mu\times\pi$. Thus we have proved that this update does preserve $\mu \times \pi$.

There is something strange about the case $d = 2$. There is no longer any randomness in the scan orders. If we start with $I = 2$, then we must use the scan P_1P_2 and have $I = 2$ at the end of the scan. So every scan uses the same order and $I_n = 2$ for all n. Similarly, if we start with $I = 1$. Thus the method is essentially fixed scan. We choose one scan order at the beginning and use it ever after.

3.1.6 Subsampling a Markov Chain

Powers are a special case of kernel multiplication (composition). If P is a Markov kernel, so is P^n, and if P preserves π, so does P^n. Doing one P^n update is the same as doing the P update n times. Hence the algorithm that does n update steps between each "sample" that is used in subsequent calculations has kernel P^n. In effect we run the chain with kernel P, but only use X_n, X_{2n}, X_{3n}, ... This is called *subsampling* the chain.

If we take a mixture of powers, we get a randomly subsampled chain. Consider a sampling distribution on the nonnegative integers giving probability a_n to n. Then the kernel of the mixture is

$$P_a = \sum_{n=0}^{\infty} a_n P^n \qquad (3.3)$$

(recall that $P^0 = I$). We are assured by our theorem about mixtures that this kernel preserves π.

What simulation has P_a as its kernel? Just follow the instructions for a random mixture.

- Generate a random nonnegative integer N_i with distribution a, i. e., $P(N_i = n) = a_n$.

- Run the chain having kernel P for N_i steps. (Running for zero steps means doing nothing.)
- Output the current state as Y_i. (This means $Y_i = Y_{i-1}$ if $N_i = 0$.)
- Set $i = i + 1$ and repeat.

If X_1, X_2, ... are a Markov chain with transition probability kernel P, then Y_1, Y_2, ..., where

$$Y_k = X_{N_1 + \cdots + N_k}$$

is a Markov chain with transition probability kernel P_a.

Curiously the notion of subsampling a chain at a fixed interval, using the kernel P^n, is very widely used, probably overused, in MCMC. But random subsampling, using the kernel P_a, is almost never used. This is surprising because random subsampling, using the kernel P_a is a major tool of Markov chain theory, used again and again in (Meyn and Tweedie 1993, Section 5.5). They call the notion "sampled chains" rather than our "subsampled," but the concept is the same.

3.1.7 Preserving Reversibility

Reversibility of a Markov chain is not necessary for MCMC and much of the literature ignores reversibility. However, reversibility does have some theoretical and practical consequences (Besag and Clifford 1989; Geyer 1992), and most elementary update mechanisms that have been proposed for MCMC are reversible, because the easiest way to show that an update mechanism preserves a specified distribution is to show that it is reversible with respect to that distribution. Hence the only way anyone makes a Markov chain for Monte Carlo that is nonreversible is to combine reversible elementary update steps in a nonreversible way. This is all right if one doesn't care whether the sampler is reversible, but one should know how to obtain a reversible sampler.

Suppose that we have d elementary update mechanisms with kernels P_i that are reversible with respect to the same distribution π. Let us see whether composition and mixing preserve reversibility.

Composition

When we combine by composition, we immediately see that reversibility is *not*, in general, preserved. Since P_1 and P_2 are self-adjoint operators on $L^2(\pi)$,

$$(f, P_1 P_2 g) = (P_1 f, P_2 g) = (P_2 P_1 f, g), \qquad f, g \in L^2(\pi),$$

and this says the adjoint of $P_1 P_2$ is $P_2 P_1$. Thus the composition is self-adjoint if and only if $P_1 P_2 = P_2 P_1$, that is, if P_1 and P_2 are *commuting* operators on $L^2(\pi)$. In general the elementary update operators do not commute and hence the composition is not self-adjoint and reversibility is not preserved. Similarly, for d operators, the adjoint of $P_1 \ldots P_d$ is $P_d \ldots P_1$, and reversibility is not preserved.

Some special forms of composition do, however, preserve reversibility. Consider the "scan" $P_1P_2P_2P_1$. Its adjoint has the operators multiplied together in reversed order, but that gives us the same thing again. Hence it is self-adjoint.

Let us say a composition of operators is *palindromic* if it reads the same forwards and backwards.[1] Then it is obvious that any palindromic composition of self-adjoint operators is self-adjoint and preserves reversibility.

Mixing

What happens when we combine by mixing? Now it is obvious that reversibility is preserved. Since P_1 and P_2 are self-adjoint operators on $L^2(\pi)$,

$$\begin{aligned}(f, [aP_1 + bP_2]g) &= a(f, P_1g) + b(f, P_2g) \\ &= a(P_1f, g) + b(P_2f, g) \\ &= ([aP_1 + bP_2]f, g), \qquad f, g \in L^2(\pi),\end{aligned}$$

and this says $aP_1 + bP_2$ is self-adjoint for any real scalars a and b. This obviously extends to arbitrary linear combinations, even to arbitrary non-finite mixtures (Exercise 3.2).

Random Sequence Scans

The the kernel (3.1) is self-adjoint if

$$\sum_{(k_1,\ldots,k_d)\in\mathcal{P}} a_k P_{k_1} \cdots P_{k_d} = \sum_{(k_1,\ldots,k_d)\in\mathcal{P}} a_k P_{k_d} \cdots P_{k_1}. \tag{3.4}$$

If we define an operator r (for reverse) on $\mathcal{P}$ by $r\big((k_1, \ldots, k_d)\big) = (k_d, \ldots, k_1)$, then (3.4) holds if $a_{r(k)} = a_k$ for all k. In words, a random sequence scan is reversible if each scan sequence has the same probability as its reverse sequence. Both of the specific methods discussed in Section 3.1.4 have this property.

3.1.8 State-Dependent Mixing

Green (1995) proposed an algorithm that involves *state-dependent mixing* having mixing probabilities that depend on the current state. Even in the case of finite mixtures, the theory developed so far does not work. Consider a mixing distribution with probabilities $a_i(x)$ that depend on the current state x. That is, we propose to use the kernel

$$P(x, A) = \sum_{i=1}^{d} a_i(x) P_i(x, A)$$

[1] A *palindrome* is a phrase that reads the same forwards and backwards, such as "Able was I ere I saw Elba."

Now $\pi P = \pi$ is

$$\sum_{i=1}^{d} \int \pi(dx) a_i(x) P_i(x, A) = \pi(A),$$

and this equation is no longer easy to verify. It is not implied by $\pi P_i = \pi$ for each i. The problem is that multiplication of a kernel by $a_i(x)$ is not multiplication of the operator P_i by a scalar. In fact, this operation is another kernel multiplication. Define the kernel

$$M_i(x, B) = a_i(x) I(x, B)$$

and check that

$$(M_i P_i)(x, B) = \int a_i(x) I(x, dy) P_i(y, B) = a_i(x) P_i(x, B).$$

Now we see that in operator notation

$$P = \sum_{i=1}^{d} M_i P_i$$

There is no reason why P should preserve π whenever all the P_i do, because M_i does not preserve π.

Green's ingenious notion was to use reversibility directly. Define $K_i = M_i P_i$, written out in full

$$K_i(x, A) = a_i(x) P_i(x, A). \tag{3.5}$$

Suppose each K_i is reversible with respect to π, that is, satisfies (2.24) with P replaced by K_i. Then clearly P is also reversible with respect to π. If P is Markov, then it does everything we want.

Thus we are lead to treating the K_i rather than the P_i as the primary objects. Let us see what the relation between the two is. Since $a_i(x)$ is a probability, it is between zero and one. Hence

$$K_i(x, B) \ge 0, \qquad B \in \mathcal{B} \tag{3.6a}$$

$$K_i(x, S) \le 1, \tag{3.6b}$$

where $(S, \mathcal{B})$ is, as usual, the state space. A kernel having these properties is called *substochastic*. Using

$$a_i(x) = K_i(x, S) \tag{3.7}$$

we see that

$$P_i(x, A) = \frac{K_i(x, A)}{K_i(x, S)} \tag{3.8}$$

So (3.5) and the pair of equations (3.7) and (3.8) can be used to go back and forth between K's and P's, and we may consider that we have been given the K_i to specify the algorithm.

As in Theorem 3.1 we can consider arbitrary mixtures. For those we change the index from i to z.

Theorem 3.2. *Suppose μ is a σ-finite positive measure and for each z in the domain of μ there is a substochastic kernel K_z that is reversible with respect to π, suppose that the map $(z, x) \mapsto K_z(x, A)$ is jointly measurable for each A, and suppose*

$$\int \mu(dz) K_z(x, A) \le 1, \qquad x \in S, \ A \in \mathcal{B}. \tag{3.9}$$

Then

$$Q(x, A) = \int \mu(dz) K_z(x, A) \tag{3.10}$$

defines a substochastic kernel Q that is reversible with respect to π.

Proof. The proof that Q is a kernel is exactly like the proof in Theorem 3.1. That Q is substochastic is again obvious. By the Fubini theorem

$$\int_A \pi(dx) Q(x, B) = \int_A \pi(dx) \int \mu(dz) K_z(x, B) = \int \mu(dz) \int_A \pi(dx) K_z(x, B). \tag{3.11}$$

Reversibility of Q with respect to π is the property that the left hand side of (3.11) is unchanged by swapping A and B, which is true because swapping A and B in the right hand side leaves it unchanged by the reversibility of each K_z. □

This theorem is often used in the case where μ is counting measure, so for ease of reference we state that as a corollary.

Corollary 3.3. *Suppose $\{ K_i : i \in I \}$ is a family of substochastic kernels, each reversible with respect to π, and suppose*

$$\sum_{i \in I} K_i(x, A) \le 1, \qquad x \in S, \ A \in \mathcal{B}.$$

Then

$$Q(x, A) = \sum_{i \in I} K_i(x, A)$$

defines a substochastic kernel Q that is reversible with respect to π.

Remark. If the index set I is finite or countable, the meaning of the sums is obvious. If I is uncountable, the sum means integration with respect to counting measure on I, that is,

$$\sup_{\substack{F \subset I \\ F \text{ finite}}} \sum_{i \in F} K_i(x, A)$$

The kernel Q defined in the corollary will be stochastic (Markov) if and only if the mixing probabilities $a_i(x) = K_i(x, S)$ sum to one for each x. Sometimes this is hard to verify (more precisely, it is hard to invent K_i having this property). Then a simple trick allows us to use the corollary anyway. Define the defect

$$d(x) = 1 - \sum_{i \in I} K_i(x, S), \qquad x \in S \tag{3.12}$$

and a new kernel

$$\widetilde{K}(x, A) = d(x)I(x, A). \tag{3.13}$$

Then $\widetilde{K}$ is reversible with respect to any distribution π since

$$\iint f(x)g(y)\pi(dx)\widetilde{K}(x, dy) = \iint f(x)g(y)d(x)\pi(dx)$$

is trivially symmetric under the interchange of f and g. If we add $\widetilde{K}$ to our set of kernels, then the sum is stochastic.

Thus we have the following formulation of state-dependent mixing. Suppose we are given a family of substochastic kernels as described in the corollary. Then the combined update described as follows

1. Choose a random index $i \in I$, choosing index i with probability $p_i(x)$ defined by (3.7). With probability (3.12) skip step 2 and stay at the current position.

2. Simulate a new value of x from the probability distribution $P_i(x, \cdot)$ defined by (3.8).

has the stochastic transition kernel $\widetilde{K} + \sum_i K_i$ and is reversible with respect to π if each of the K_i is reversible with respect to π.

In the general case described by the theorem, the algorithm is a bit more complicated to describe, partly because the notation is a bit confusing. Now the probability of using the kernel K_z is denoted $a_z(x) = K_z(x, S)$, and we need to think of this as a subprobability density with respect to μ, but in that role z is the variable, x being fixed. So let us write $f_x(z) = a_z(x)$. Then

$$\int f_x(z)\mu(dz) \le 1$$

by (3.9) so f_x is indeed a subprobability density. The defect of f_x is

$$d(x) = 1 - \int f_x(z)\mu(dz), \tag{3.14}$$

and we define $\widetilde{K}$ by (3.13) exactly as before except that the defect is defined by (3.14) rather than (3.12).

In order to carry out the combined update described by the theorem, we need to be able to simulate a random variate Z having this density with respect to μ. The update is described as follows.

1. Simulate a random variate z having probability density function f_x with respect to μ. With probability (3.14) skip step 2 and stay at the current position.

2. Simulate a new value of x from the probability distribution $P_z(x, \cdot)$ defined by (3.8) with i replaced by z.

3.2 The Metropolis-Hastings Algorithm

In one form (Metropolis, Rosenbluth, Rosenbluth, Teller, and Teller 1953), this is the oldest MCMC algorithm, dating to the dawn of the computer age when the only place something like this could have been done was Los Alamos. In its modern form (Green 1995), it is the newest MCMC algorithm, which solves many problems MCMC researchers have stumbled over in the past. In between, a key improvement was made by Hastings (1970), which in a curious episode in the sociology of science was not really understood for 20 years. The paper was published in a prestigious journal (*Biometrika*) and was cited by some MCMC authors (Ripley 1987), but many problems that now seem trivial ("just use Metropolis-Hastings") were stumbled over because the importance of Hastings' improvement was not understood.

3.2.1 Unnormalized Probability Densities

The section heading refers to a concept that is familiar, being a standard problem in introductory probability courses, but usually is not given a name. Here we do give it a name so we can use it better. A function h is an *unnormalized probability density* with respect to a positive measure μ if h is nonnegative and has a finite, nonzero integral. Then the integral $c = \int h(x)\mu(dx)$ is called the *normalizing constant* for h, and the function f defined by $f(x) = h(x)/c$ is called the *normalized density* corresponding to h.

As we said, this is concept is very familiar from introductory probability problems like: What constant k makes kx^2 a probability density for $0 < x < 1$? But lack of a name for this concept keeps people from noticing that it plays a key role in several areas of statistics.

It is part of the definition, but it needs to be emphasized that calling h and unnormalized density asserts

- it is nonnegative,
- it does not integrate to zero (i. e., is strictly positive on some set having positive μ-measure), and
- it does not integrate to infinity.

Checking the first two items is trivial. Checking the third is nontrivial, but it must be done. Arguments about "unnormalized densities" that integrate to infinity are mathematical nonsense.

Bayesian Inference

The computational problems that make Bayesian inference difficult all involve unnormalized densities, the reason being

$$\textit{likelihood} \times \textit{prior} = \textit{unnormalized posterior}$$

If a Bayesian has a data model $f(x|\theta)$ and a prior $g(\theta)$, the problem is to calculate properties of the posterior

$$h(\theta|x) = \frac{f(x|\theta)g(\theta)}{\int f(x|\theta)g(\theta)\,d\theta} \tag{3.15}$$

Because f and g appear in both the numerator and the denominator, both may be unnormalized, considered as functions of θ. Unnormalized versions of $f(x|\theta)$ are a concept with a name. A function $L_x(\theta)$ is a *likelihood* for the problem if

$$L_x(\theta) = a(x)f(x|\theta)$$

for an arbitrary strictly positive function $a(x)$. If we plug this into (3.15) we get

$$h(\theta|x) = \frac{L_x(\theta)g(\theta)}{\int L_x(\theta)g(\theta)\,d\theta} \tag{3.16}$$

(the $a(x)$ terms in the numerator and denominator cancel). It is also clear that we could plug in $cg(\theta)$ for $g(\theta)$ for an arbitrary positive constant c and the $c's$ would cancel, leaving the result unchanged.

Equation (3.16) even makes sense when g is not an unnormalized density. It can be any nonnegative function on the parameter space, so long as the numerator $L_x(\theta)g(\theta)$ is an unnormalized density. When $g(\theta)$ does not integrate, we say that it is an *improper prior*.

When the prior is proper, there is no need to show that the likelihood times the prior is integrable. It is automatically integrable by the laws of probability. The integral of the numerator in (3.15) is the marginal density for x, which is finite. When the prior is improper, a proof that the likelihood times the prior is integrable is a required part of the problem. Omitting the proof risks committing nonsense.[2]

Conditioning and Unnormalized Densities

Not surprising, Bayes rule being just a rearrangement of the definition of conditional probability, the relationship between unnormalized densities and conditioning we saw in Bayesian inference is a general phenomenon

> *A joint density is an unnormalized conditional density. The marginal is its normalizing constant.*

[2]It happened once to your humble author (Geyer 1992, see the "Note added in proof"). Don't let it happen to you. There is some MCMC literature on what happens when you try to simulate an "improper posterior" (you omitted the proof of integrability, and there isn't a proof, and you are in the realm of mathematical nonsense), but a short digest of that literature is that there is nothing to be said, no one has a clue about what will happen. Moreover, the whole notion of "improper posterior" seems to have no theoretical foundation. Even if you could simulate it in some sense, no Bayesian theoretician I've talked to thinks it has any meaning.

What this means is the following. Say $f(x, y)$ is a joint density considered as a function of two variables x and y. Considered as a function of one variable, say x, it is an unnormalized density defined by

$$h_y(x) = f(x, y).$$

The normalizing constant for h_y is the marginal of y

$$p(y) = \int h_y(x)\, dx = \int f(x, y)\, dx.$$

Really we should call $p(y)$ a "normalizing function" rather than "normalizing constant" because it is a function of y. Dividing by the normalizing function gives the conditional density

$$f(x|y) = \frac{f(x, y)}{p(x)}.$$

The same phenomenon holds when the joint distribution is unnormalized, but we have to be a bit careful with our terminology. Suppose we now have the unnormalized density $h(x, y) = cf(x, y)$, where c is an unknown constant. Again, we write

$$h_y(x) = h(x, y),$$

but now the normalizing function is not the marginal, though it is proportional to the marginal

$$c(y) = \int h_y(x)\, dx = \int cf(x, y)\, dx = cp(y).$$

But still, normalizing h_y gives the conditional density

$$\frac{h_y(x)}{c(y)} = \frac{h(x, y)}{c(y)} = \frac{f(x, y)}{p(x)} = f(x|y)$$

Models Specified by Unnormalized Densities

If for each θ in a parameter space Θ we have a function h_θ that is an unnormalized probability density with respect to μ, we say that the family of unnormalized densities $\{ h_\theta : \theta \in \Theta \}$ is a *family of unnormalized densities*. Again the normalizing constants

$$c(\theta) = \int h_\theta(x)\mu(dx), \qquad \theta \in \Theta$$

define a function $c : \Theta \to (0, \infty)$ called the *normalizing function* of the family. As always, the use of the term "unnormalized densities" implies that $0 < c(\theta) < \infty$ for all θ. The normalized densities of the family are defined by

$$f_\theta(x) = \frac{1}{c(\theta)} h_\theta(x), \qquad x \in S \tag{3.17}$$

(where S is, as usual, the sample space).

This notion may seem unfamiliar, but it is a widely used technique for specifying models for complicated phenomena. It may be very difficult to specify a model for which the normalizing constant is known for complicated data. As we will see, it is not necessary to have a closed-form expression for the normalizing constant in order to use the family as a statistical model. We will always be able to simulate data from the model by MCMC, and

when we can simulate, we can do inference.

This assertion may be a bit hard to swallow until some examples have been seen, but we will see them in due course.

3.2.2 The Metropolis-Hastings Update

The Metropolis-Hastings update preserves any distribution π specified by an unnormalized density h with respect to a measure μ. There is no restriction on $h(x)$ other than that it actually be an unnormalized density (its normalizing constant is nonzero and finite) and that it can be evaluated, that is, for each x we can calculate $h(x)$. There is no requirement that we be able to do any integrals or know the value of the normalizing constant. In particular, unlike the Gibbs sampler, we do not need to know anything about any conditional distributions of π.

The Metropolis-Hastings update uses an auxiliary transition probability specified by a density $q(x, y)$ called the *proposal distribution*. For every point x in the state space, $q(x, \cdot\,)$ is a (normalized) probability density with respect to μ having two properties: for each x we can simulate a random variate y having the density $q(x, \cdot\,)$ and for each x and y we can evaluate the $q(x, y)$. To summarize, this is what we need

1. For each x we can evaluate $h(x)$.
2. For each x and y we can evaluate $q(x, y)$.
3. For each x we can simulate a random variate with density $q(x, \cdot\,)$ with respect to μ.

There is no necessary connection between the auxiliary density $q(x, y)$ and the density $h(x)$ of the stationary distribution. We can choose any density that we know how to simulate. For example, if the state space is d-dimensional Euclidean space $\mathbb{R}^d$ we could use a multivariate normal proposal density with mean x and variance a constant times the identity. If ϕ denotes a Normal$(0, \sigma^2 I)$ density, then we have $q(x, y) = \phi(y - x)$. We can easily simulate multivariate normal variates and evaluate the density.

The Metropolis-Hastings update then works as follows. The current position is x, and the update changes x to its value at the next iteration.

1. Simulate a random variate y having the density $q(x, \cdot\,)$.

2. Calculate the "Hastings ratio"

$$R = \frac{h(y)q(y,x)}{h(x)q(x,y)}. \tag{3.18}$$

3. Do "Metropolis rejection:" with probability $\min(1, R)$ set $x = y$.

Later in this section we will prove that this update always preserves π.

We often say we "accept" the "proposal" y if we set the value $x = y$ in step 3. Otherwise we say we "reject" the proposal. When we reject, the value of the state of the Markov chain remains the same for two consecutive iterations.

Warning: Those familiar with so-called rejection sampling in ordinary Monte Carlo note that Metropolis rejection is completely different. In ordinary rejection sampling, proposals are made over and over until one is accepted. The first proposal accepted is the next sample. In Metropolis rejection only one proposal is made, if it is not accepted, then the Markov chain doesn't move and X_{n+1} is equal to X_n. If Metropolis rejection were done like ordinary rejection, the resulting Markov chain *would not preserve* π.

Note also that the denominator of the Hastings ratio (3.18) can never be zero if the chain starts at a point where $h(x)$ is nonzero. A proposal y such that $q(x, y) = 0$ occurs with probability zero, and a proposal y such that $h(y) = 0$ is accepted with probability zero. Thus there is probability zero that denominator of the Hastings ratio is ever zero during an entire run of the Markov chain so long as $h(X_1) > 0$. If we do not start in the support of the stationary distribution we have the problem of defining how the chain should behave when $h(x) = h(y) = 0$, that is, how the chain should move when both the current position and the proposal are outside the support of the stationary distribution. The Metropolis-Hastings algorithm says nothing about this. It is a problem that is best avoided by starting at a point where $h(x)$ is positive.

Also note specifically that there is no problem if the proposal is outside the support of the stationary distribution. If $h(y) = 0$, then $R = 0$ and the proposal is always rejected, but this causes no difficulties.

3.2.3 The Metropolis Update

The special case when we use a proposal density satisfying $q(x, y) = q(y, x)$ is called the Metropolis update. In this case the Hastings ratio (3.18) reduces to the odds ratio

$$R = \frac{h(y)}{h(x)}$$

and there is no need to be able to evaluate $q(x, y)$ only to be able to simulate it. Thus the requirements for Metropolis are a bit different from those for Metropolis-Hastings

1. For each x we can evaluate $h(x)$.

2. $q(x, y) = q(y, x)$ for each x and y.

3. For each x we can simulate a random variate with density $q(x, \cdot)$ with respect to μ.

(the first and third requirements are unchanged, only the second is different).

Metropolis proposals save the trouble of evaluating $q(x, y)$ in calculating the Hastings ratio. Evaluating $q(x, y)$ is usually not that much work, so avoiding it is not worth much additional trouble in making proposals.

> *Gibbs and Metropolis are all right when they are easy and effective. Otherwise they are part of the problem, not part of the solution.*

Always keep the general method in mind (for now "general" means Metropolis-Hastings, later it will mean Metropolis-Hastings-Green).[3]

3.2.4 A Good Default MCMC Sampler

The objective of this section is to outline a good "default" MCMC sampler. One way to think of what we are looking for is a method that will give reasonably good answers with a minimum of trouble.[4]

The normal proposal mentioned above is a Metropolis proposal. By the symmetry of the multivariate normal distribution, $q(x, y) = \phi(y - x)$ is equal to $q(y, x) = \phi(x - y)$, where ϕ is any non-degenerate multivariate normal density, that is, the proposal is $y \sim \text{Normal}(x, \Sigma)$, where Σ is any positive-definite matrix and x is the current position.

Although there are good reasons for using this method with general Σ, a method that asks the user to specify an arbitrary covariance matrix having the dimension of the state space has to many parameters to be considered easy to use. So we will restrict Σ to be diagonal. If the coordinate variables of the state vector have approximately the same variance under the distribution π we want to simulate, then we can use an even simpler proposal with $\Sigma = \sigma^2 I$. Now there is only one parameter (σ) that must be adjusted by the user. We can't do any

[3] If I had a nickel for every time I've been asked for help with an MCMC problem and answered, "Why are you using a Gibbs update there? Metropolis-Hastings would be easy and fix the problem," I'd be rich.

[4] Another way to think of what we are looking for is a default setting for the worlds most obnoxious seminar question. A statistician who shall remain nameless often asks seminar questions of the following form: "The most simple minded approach to this problem I can think of is blah. Can you explain why your method works any better than that?" Here "blah" stands for any really simple method, preferably one that can be explained in one sentence and took about fifteen seconds to think up. The reason the question is so obnoxious is that many people do write papers and give talks about very complicated methods that can be proved to have various properties, but cannot be proved to be better than the "most simple minded approach" I can think of. If the speaker understands the question, he is left with nothing to say. If the speaker doesn't get the point, and blathers on without addressing the issue of whether is method is good for anything, he seems a fool. In MCMC the method of this section is a good "most simple minded approach." I can't tell you how many MCMC talks I've heard or papers I've read that gave no reason to believe the methods proposed were better than this default.

better than that. If σ is chosen ridiculously small, say 10^{-10}, the chain can't get anywhere in any reasonable number of iterations. If σ is chosen ridiculously large, say 10^{10}, all of the proposals will be so far out in the tail that none will be accepted in any reasonable number of iterations. In either case, the chain will not produce a representative sample from its invariant distribution in the amount of time anyone is willing to wait. So we have a "Goldilocks problem." We don't want the porridge too cold or too hot. Of course we could choose $\sigma = 1$ and hope that will be about right for most problems, but that seems a too much to hope for.

How do we choose σ? Gelman, Roberts, and Gilks (1996) considering the performance of this algorithm in simulating multivariate normal distributions showed that adjusting σ so that about 20% of proposals are accepted gives the best performance (if you are simulating a multivariate normal). This came as a shock to many MCMC practitioners whose naive intuition told them that high acceptance rates like 90% would be right. So even though the recommendation was not exactly right for any non-toy problem it had a huge effect on practice, because what everyone was doing was grossly wrong. Geyer and Thompson (1995) came to a similar conclusion, that a 20% acceptance rate is about right, in a very different situation. They also warned that a 20% acceptance rate could be very wrong and produced an example where a 20% acceptance rate was impossible and attempting to reduce the acceptance rate below 70% would keep the sampler from ever visiting part of the state space. So the 20% magic number must be considered like other rules of thumb we toss around in statistics: $n > 30$ means the z-test is o. k. and more than 5 expected in each cell of a contingency table means the chi-square test is o. k. We know these rules of thumb can fail. There are many examples in the literature where they do fail. We keep repeating them because we want something simple to tell beginners, and they are all right for many problems.

> *The rule of thumb says 20% but* your mileage may vary.

From the *Jargon File* (Raymond 1996)

> **Your mileage may vary** (YMMV) /caveat/ [from the standard disclaimer attached to EPA mileage ratings by American car manufacturers] 1. A ritual warning often found in Unix freeware distributions. Translates roughly as "Hey, I tried to write this portably, but who knows what'll happen on your system?" 2. More generally, a qualifier attached to advice. "I find that sending flowers works well, but your mileage may vary."

Example 3.3. Bayesian Logistic Regression.
Here we do Bayesian logistic regression with a flat prior on the kyphosis data that comes with S-PLUS (Chambers and Hastie 1993, pp. 200 ff.). The problem has three predictor variables plus an intercept, so the log likelihood is

$$L(\beta) = \prod_{i=1}^{n} p(\theta_i)^{y_i} q(\theta_i)^{1-y_i}$$

where

$$p(\theta) = \frac{e^\theta}{e^\theta + 1} \qquad q(\theta) = 1 - p(\theta) = \frac{1}{e^\theta + 1}$$

and

$$\theta_i = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + x_{i3}\beta_3.$$

The responses y_i are all zero or one. The covariates x_{ij} are arbitrary real numbers. Here we use a flat prior $g(\beta) \equiv 1$.

A few short runs, the first four lines of the following table, establish that $\sigma = 0.2$ is about right.

sample size	subsample spacing	σ	acceptance rate (%)	computer time (sec)
10000	1	1.00	0.0	2.3
10000	1	0.10	2.1	2.1
10000	1	0.01	35.5	2.0
10000	1	0.03	11.9	2.1
10000	1	0.02	18.1	2.1
10000	10	0.02	17.9	18.7
10000	100	0.02	17.9	187.3

Figure 3.1 shows a time series plot for β_0. Of the four parameters, this one has the worst plot. The series hardly looks stationary. We need a longer run, because we don't want to fill up the disk, we use a wider spacing. The last line of the table shows a run of 10^6 iterations, subsampled at every 100 iterations, so we only write out 10^4 samples. We can't plot more than that anyway. Figure 3.2 is better than Figure 3.1 but not by much. The chain appears more or less stationary, but has so much autocorrelation that any estimates based on it will have low precision. Since this run only took three minutes we could increase the spacing by a factor of 100 again if we were willing to wait several hours for the results, but we could also think a little bit.

A little though about regression (not about MCMC) comes to the idea that the problem may be ill conditioned because of correlation among the predictor variables (a. k. a. collinearity). This leads to high correlation among the regression coefficients. When we check for that, we see that β_0 and β_3 are fairly highly correlated (Figure 3.3). This leads to the further idea that if we used orthogonal predictors, we might get a better behaved sampler. In fact, since the constant predictor is one of the ones causing trouble, we might just orthogonalize the other predictors to it, i. e., subtract off their means. This is equivalent to a change of parameters. Call the new parameters β_i'. Then we have

$$\beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + x_{i3}\beta_3 = \beta_0' + (x_{i1} - \bar{x}_1)\beta_1' + (x_{i2} - \bar{x}_2)\beta_2' + (x_{i3} - \bar{x}_3)\beta_3'$$

from which we see

$$\begin{aligned} \beta_i &= \beta_i', \qquad i = 1, 2, 3 \\ \beta_0 &= \beta_0' - \bar{x}_1\beta_1' - \bar{x}_2\beta_2' - \bar{x}_3\beta_3' \end{aligned} \tag{3.19}$$

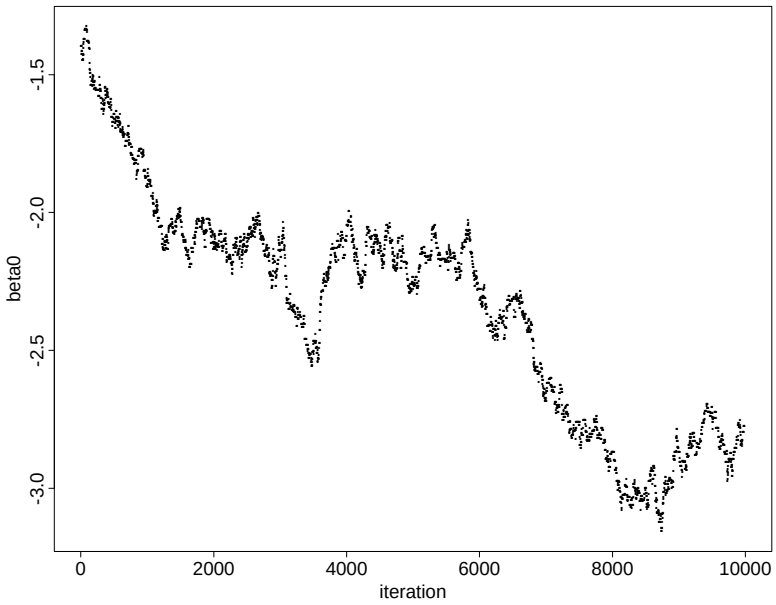


Figure 3.1: Time series plot of Metropolis sampler output for β_0 in the four-parameter logistic regression for the kyphosis data (Chambers and Hastie 1993). The sampler is the "default" Metropolis with $\sigma = 0.02$.

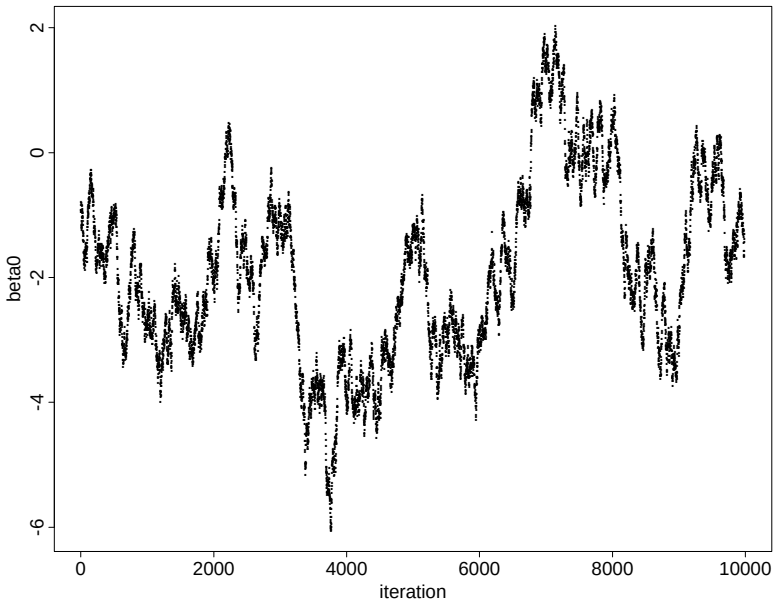


Figure 3.2: Time series plot of Metropolis sampler output for β_0 in the same model as in Figure 3.1. The only difference is the chain is subsampled with spacing 100 and runs 100 times as long.

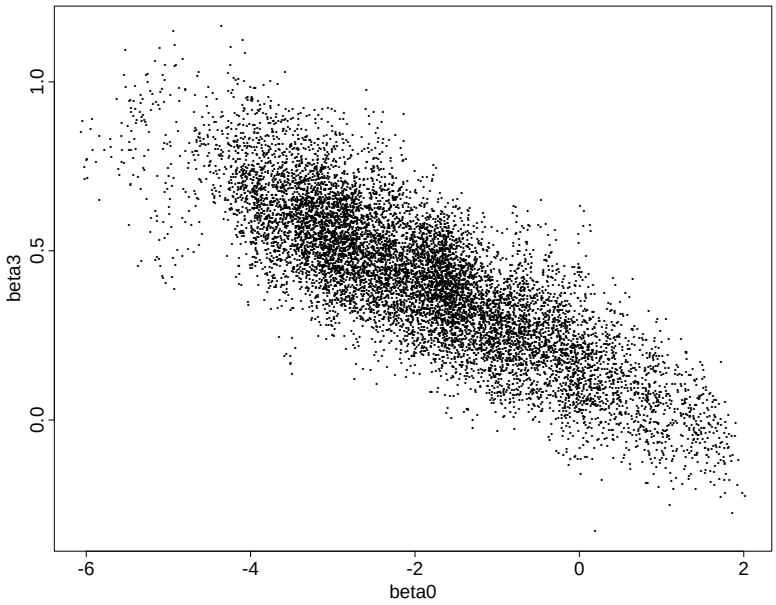


Figure 3.3: Scatter plot of β_0 versus β_3 for the Metropolis sampler output for the same run as in Figure 3.2.

Thus we can easily convert back to the original parameters if we so desire. We may not bother. The coefficient for the constant predictor has to be in the model, but we are not interested in its actual value.

With this change of parameters, things go much better.

sample size	subsample spacing	σ	acceptance rate (%)	computer time (sec)
10000	1	0.020	33.4	2.1
10000	1	0.030	22.6	2.1
10000	1	0.040	16.2	2.1
10000	1	0.035	19.0	2.1
10000	10	0.035	18.9	18.6
10000	100	0.035	19.4	185.6

One indication we are doing better is that we get higher acceptance rates for the same σ or, what is the same thing, can take bigger steps with the same acceptance rate. Figure 3.4 is the analog of Figure 3.1 for the run in line four of this table. Figure 3.4 looks much better than Figure 3.1. We continue making longer runs (the last two lines of the table) and then look at the analog of Figure 3.3. In order to do this we have to transform back to the original parameterization using (3.19).

It is clear that the transformation has turned a moderately hard problem

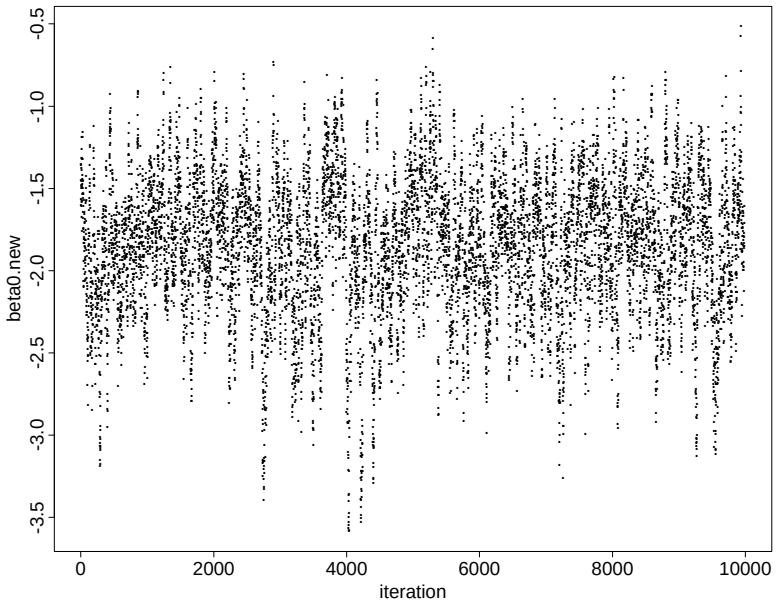


Figure 3.4: Time series plot of Metropolis sampler output for β_0 the same logistic regression data as in 3.1 but using the parameterization (3.19). The length of run and spacing of samples is the same as in 3.1.

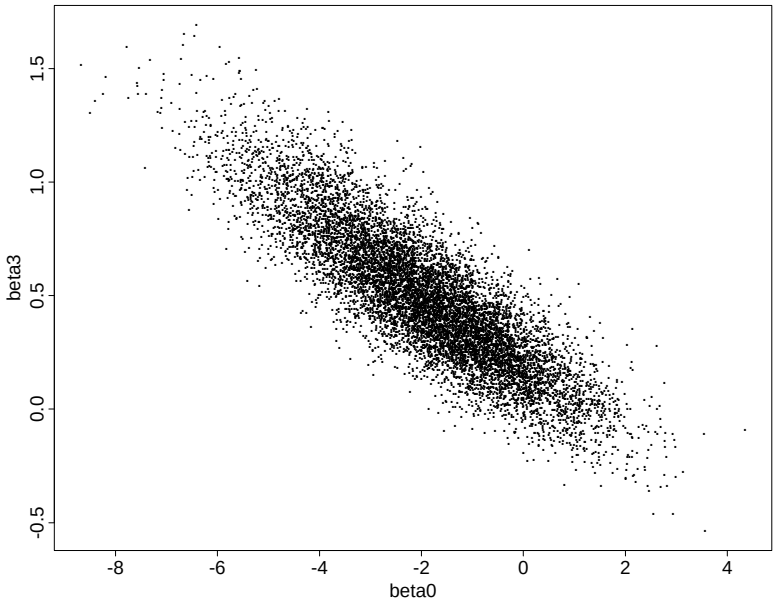


Figure 3.5: Scatter plot of β_0 versus β_3 for the Metropolis sampler output for the same data, same Monte Carlo sample size and same spacing of subsamples as in Figure 3.3. The only difference is that the parameters β_i' were used and then translated back to the original parameterization.

into an easy one. We do not continue with the example, because we have already seen what was to be learned. That we needed a simple trick should not be surprising, nothing in statistics works "right out of the box." Why would MCMC be an exception?

There are no idiot-proof MCMC methods, not even the "default."

Example 3.4. The Dumbbell Distribution.
This is a toy problem that shows the 20% rule failing.

A few quick runs show us that $\sigma = 1.3$ is about right according to the 20% rule. But what σ is really optimal?

sample size	subsample spacing	σ	acceptance rate (%)	computer time (sec)
10000	1	1.0	30.4	0.2
10000	1	2.0	11.1	0.2
10000	1	1.3	21.6	0.2

Suppose we are trying to estimate the mean of x (the horizontal coordinate). Of course, we know this is the center of symmetry in this toy problem, but you have to imagine we don't know the mean and must estimate it. What σ gives the most accuracy in estimating the mean?

We look at some more runs, this time also estimating the variance in the central limit theorem σ^2_{clt} (1.10) by the method of batch means (Section 1.6.3) with 100 batches.

sample size	subsample spacing	σ	acceptance rate (%)	computer time (sec)	σ^2_{clt}
10^5	100	1.3	21.3	130.1	407.42
10^5	100	2.0	11.4	131.7	170.30
10^5	100	3.0	6.2	132.8	93.66
10^5	100	4.0	4.0	132.7	55.04
10^5	100	5.0	2.9	136.5	47.40
10^5	100	6.0	2.3	133.2	39.76
10^5	100	7.0	1.8	133.1	38.89
10^5	100	8.0	1.5	133.0	44.79

It is clear that $\sigma = 1.3$ in not optimal and in fact $\sigma = 7$ is more like it and the optimal acceptance rate is more like 2% than 20%.

I imagine some reader will now protest that most problems are not like the "dumbell distribution" so what is the point? I reply by saying that asking the question like that misses the point. Unlike criminal defendents, math is guilty

until proven innocent. You are not entitled to assume that "most problems" are not "like" the dumbell distribution until you have a precise definition of the class of problems you are talking about and a proof that 20% acceptance rate (or whatever) is optimal for all problems in the class. As it stands now, we have a counterexample that disproves the conjecture that 20% is optimal for all problems. Until someone comes up with a better conjecture, that's the end of the story.

I imagine that some readers are still not satisfied. They would be happy to leave math and rely on practical experience. To them I would say that practical experience with complicated problems shows they do have bottlenecks like this toy problem. It is easy for the sampler to move around some parts of the state space, but hard for the sampler to get from one part of the state space to another (through a "bottleneck"). Real problems with bottlenecks tend to be so hard that the kind of experimentation we did here would take a very long time. But there is every reason to suspect that real problems do exhibit phenomena similar to the dumbell distribution.

3.2.5 Reversibility of Metropolis-Hastings

We can now write down the transition probability kernel for the Metropolis-Hastings update. The transition probability has two terms. For accepted proposals, we propose y and then accept it, which happens with probability density

$$p(x, y) = q(x, y)a(x, y),$$

where $a(x, y) = \min(R, 1)$ is the acceptance probability. Hence for any set A

$$\int_A q(x, y)a(x, y)\mu(dy)$$

is the part of $P(x, A)$ that results from accepted proposals. If the integral on the right hand side is taken over the whole state space, it gives the total probability that the proposal will be accepted. Thus the probability that the proposal is rejected is

$$r(x) = 1 - \int q(x, y)a(x, y)\mu(dy).$$

If the proposal is rejected we stay at x. Hence

$$P(x, A) = r(x)I(x, A) + \int_A q(x, y)a(x, y)\mu(dy), \tag{3.20}$$

Where $I(x, A)$ is the identity kernel, which we now recognize as the Markov kernel that corresponds to "doing nothing."

We now want to verify that the Metropolis-Hastings update is reversible with respect to π.

Lemma 3.4. *Suppose the transition probability kernel of a Markov chain has the following form*

$$P(x, A) = r(x)I(x, A) + \int_A p(x, y)\mu(dy), \tag{3.21}$$

where $p(x, \cdot)$ is a subprobability density for each x and

$$r(x) = 1 - \int p(x, y)\mu(dy).$$

Suppose $h(x)$ is an unnormalized density with respect to μ and

$$h(x)p(x, y) = h(y)p(y, x), \qquad \text{for all } x \text{ and } y. \tag{3.22}$$

Then this Markov chain is reversible with respect to the distribution π having unnormalized density h with respect to μ.

Proof. What is to be shown is that

$$\begin{aligned}\iint f(x)g(y)\pi(dx)P(x, dy) \\ = \int f(x)g(x)r(x)\pi(dx) + \iint f(x)g(y)\pi(dx)p(x, y)\mu(dy).\end{aligned}$$

is unchanged when we interchange f and g (2.25).

The first term is obviously unchanged by interchanging f and g. So we work on the second term, which multiplied by the normalizing constant for $h(x)$ is

$$\begin{aligned}\iint f(x)g(y)h(x)p(x, y)\mu(dx)\mu(dy) &= \iint f(x)g(y)h(y)p(y, x)\mu(dx)\mu(dy) \\ &= \iint f(y)g(x)h(x)p(x, y)\mu(dy)\mu(dx)\end{aligned}$$

where (3.22) gives the first equality, and interchanging the dummy variables x and y gives the second. Now, except for the order of integration, the second line is just the left hand side of the first with f and g interchanged. Reversal of the order of integration is justified by the Fubini theorem. □

Corollary 3.5. *The Metropolis-Hastings update is reversible with respect to the distribution π having unnormalized density h with respect to μ.*

Proof. The Metropolis-Hastings kernel (3.20) has the form (3.21) with $p(x, y) = q(x, y)a(x, y)$. Thus we need only verify (3.22).

The probability that a proposal is accepted is

$$a(x, y) = \min(1, R) = \min\left(1, \frac{h(y)q(y, x)}{h(x)q(x, y)}\right).$$

Note that if $R \le 1$ then

$$a(x, y) = \frac{h(y)q(y, x)}{h(x)q(x, y)} \qquad \text{and} \qquad a(y, x) = 1$$

and if $R \ge 1$ then

$$a(x, y) = 1 \qquad \text{and} \qquad a(y, x) = \frac{h(x)q(x, y)}{h(y)q(y, x)}$$

In either case

$$a(x, y)h(x)q(x, y) = a(y, x)h(y)q(y, x),$$

which is (3.22). □

3.2.6 One-Variable-at-a-Time Metropolis-Hastings

When the state X is a vector $X = (X_1, \ldots, X_d)$, the Metropolis-Hastings update can be done one variable at a time, just like the Gibbs update. The algorithm is essentially the same as before, although some changes in notation are required because the proposal only changes a single variable and hence the proposal density $q(x, y)$ is not a density with respect to the measure μ on the whole space. (Warning: for the rest of the section, subscripts indicate components of the state vector, not the time index of a Markov chain.)

Suppose μ is a product measure $\mu_1 \times \cdots \times \mu_d$. For a Metropolis-Hastings update of the i-th variable, we need a proposal density $q_i(x, \cdot)$ with respect to μ_i. The update then works as follows. The current position is x, and the update changes x to its value at the next iteration.

1. Simulate a random variate y having the density $q_i(x, \cdot)$. Note that y has the dimension of x_i not x. Let x_y denote the state with x_i replaced by y

$$x_y = (x_1, \ldots, x_{i-1}, y, x_{i+1} \ldots x_d).$$

2. Evaluate the Hastings ratio

$$R = \frac{h(x_y)q_i(x_y, x_i)}{h(x)q_i(x, y)}.$$

3. Do Metropolis rejection: with probability $\min(1, R)$ set $x = x_y$.

Note that, as with the original Metropolis-Hastings update, this update also stays in feasible states if started in a feasible state.

It is easy enough to go through the statements and proofs of Lemma 3.4 and Corollary 3.5 making the necessary notational changes to obtain the analogous results for one-variable-at-a-time Metropolis-Hastings. But we won't bother, since variable-at-a-time Metropolis is a special case of the Metropolis-Hastings-Green algorithm, and we will give proofs for that.

3.2.7 Why Gibbs is a Special Case of Metropolis-Hastings

Gibbs updates a variable x_i from its conditional distribution given the rest. The unnormalized joint density of all the variables is $h(x) = h(x_1, \ldots, x_d)$. We know from our slogan about conditioning and unnormalized densities that this is also an unnormalized conditional density of x_i given x_{-i}.

A Gibbs update is a Metropolis-Hastings update in which the proposal density is $x_i \mapsto h(x_1, \ldots, x_d)$. Thus

$$q_i(x, y) = h(x_1, \ldots, x_{i-1}, y, x_{i+1}, \ldots, x_d)/c$$

where c is the unknown normalizing constant that makes h a proper conditional probability density. Then using the notation of the preceding section, the Hastings ratio is

$$\frac{h(x_y)q_i(x_y, x_i)}{h(x)q_i(x, y)} = \frac{h(x_y)h(x)}{h(x)h(x_1, \ldots, x_{i-1}, y_i, x_{i+1}, \ldots, x_d)} = 1.$$

Thus this Metropolis-Hastings simulates a new value of x_i from its conditional given the rest and always accepts the proposal. Hence it does exactly the same thing as a Gibbs update.

3.3 The Metropolis-Hastings-Green Algorithm

Metropolis-Hastings-Green is just like Metropolis-Hastings except that measures replace densities. Why would we want something like that? One reason is one-variable-at-a-time Metropolis-Hastings in which the whole state space is $\mathbb{R}^d$, but the proposal lies in a one-dimensional subset

$$A_{i,x} = \{\, (x_1, \ldots, x_{i-1}, y, x_{i+1} \ldots x_d) : y \in \mathbb{R} \,\}.$$

Since the support $A_{i,x}$ of the proposal depends on the current position x, the proposal distribution cannot have a density with respect to one single measure, that is, it cannot have a density $q_i(x, \cdot\,)$ with respect to μ we used in the general Metropolis-Hastings algorithm. That's why we were forced to use different notation for one-variable-at-a-time Metropolis-Hastings (and would have needed a different proof of reversibility had we attempted one).

But, as we shall see, there are many other situations in which we want to make proposals in subsets of the state space that depend on the current position. In order to describe all of these using the same theory, we need a more general theory.

3.3.1 Metropolis-Hastings-Green, the Dominated Case

The Metropolis-Hastings-Green (MHG) update (Green 1995) is best described as Metropolis-Hastings with measures replacing densities.

- The unnormalized density h is replaced by an unnormalized measure η.

- The proposal density $q(x, y)$ is replaced by a proposal kernel $Q(x, A)$.
- The Hastings ratio (3.18) is replaced by "Green's ratio"

$$R(x, y) = \frac{\eta(dy)Q(y, dx)}{\eta(dx)Q(x, dy)} \tag{3.23}$$

Before we can make sense of this we have to clarify what each of these means.

By an "unnormalized measure" we mean a positive real measure. Here we want an unnormalized measure η that is proportional to the desired invariant distribution π, that is, $\eta = c\pi$ or, written out in more detail, $\eta(B) = c\pi(B)$ for all measurable sets B. Since π is a probability measure, $c = \eta(S)$, where S is the state space. Allowing the measure to be unnormalized doesn't affect the characterization of reversibility. We say the kernel P is reversible with respect to the positive measure η if (2.24) holds when π is replaced by η. Clearly, a kernel is reversible with respect to both η and π or neither.

The proposal kernel Q needs almost no explanation. When x is the current position, $Q(x, \cdot\,)$ is a probability measure used to make the proposal.

Strictly speaking (3.23) is meaningless nonsense. It is shorthand for a Radon-Nikodym derivative. We will later give precise definitions, for now we adopt the temporary definition[5] that (3.23) means

$$\iint g(x, y)R(x, y)\eta(dx)Q(x, dy) = \iint g(x, y)\eta(dy)Q(y, dx) \tag{3.24}$$

holds for every function g for which the integrals are defined, in particular for every indicator function.

There is ambiguity in defining R by (3.24), since R can be arbitrarily redefined on a set of measure zero without affecting the values of the integrals. In many interesting examples the point (x, y) will have measure zero. If we are allowed to redefine R before each use, the value $R(x, y)$ will be arbitrary whenever we use it. That's won't do at all! In order to have an algorithm we need to settle on one version of R, that is, one function that satisfies (3.24), and use that same function always. It doesn't matter which version we choose, so long as we stick with our choice ever after.

Now the obvious changes of notation transform Metropolis-Hastings into the more general MHG update. The current position is x, and the update changes x to its value at the next iteration.

1. Simulate a random variate y having the probability distribution $Q(x, \cdot\,)$.
2. Calculate "Green's ratio" $R(x, y)$.
3. Do "Metropolis rejection:" with probability $\min[1, R(x, y)]$ set $x = y$.

We see that the conditions we need are

1. For each x we can simulate a random variate with distribution $Q(x, \cdot\,)$.
2. For each x and y we can evaluate $R(x, y)$.

[5]The meaning of (3.23) will later be generalized to cases in which (3.24) does not hold.

Green's Algorithm

The MHG update really gains power when combined with state-dependent mixing. The algorithm proposed in Green (1995) used both ideas. There are a finite or infinite set of proposal kernels $Q_i(x, A)$, $i \in I$, which are permitted to be *substochastic*. The requirements on the proposal kernels are

- $Q_i(x, S)$ is known for all i.
- $$\sum_{i \in I} Q_i(x, S) \le 1, \qquad \forall x \in S$$
- For all $i \in I$
$$R_i(x, y) = \frac{\pi(dy) Q_i(y, dx)}{\pi(dx) Q_i(x, dy)} \tag{3.25}$$
is known[6] and it is possible to evaluate $R_i(x, y)$ for all x and y.
- for each x and i, it is possible to simulate realizations from the distribution having the normalized proposal distribution
$$P_i(x, \cdot) = \frac{Q_i(x, \cdot)}{Q_i(x, S)} \tag{3.26}$$

Then one step of Green's algorithm, starting from current position x goes as follows.

1. Simulate a random index i, choosing $i \in I$ with probability $Q_i(x, S)$. With probability $1 - \sum_{i \in I} Q_i(x, S)$, skip the remaining steps and stay at x.
2. Simulate $y \sim P_i(x, \cdot)$ defined by (3.26).
3. Calculate Green's ratio $R_i(x, y)$.
4. Accept y with probability $\min[1, R_i(x, y)]$.

All of this is just the MHG update described in preceding section combined with the idea of state-dependent mixing (Section 3.1.8).

3.3.2 Spatial Point Processes

Poisson Processes

A *spatial point process* is a random process having values that are point patterns in a region of $\mathbb{R}^d$. Both the number of points and their positions within the region are random. A point process is *simple* if the locations of points never

[6] We take the Radon-Nikodym derivative here to have the same meaning here as in the preceding section, i. e., (3.24) holds with Q and R replaced by Q_i and R_i. Also we must fix one version of R_i to be used throughout. As promised for the simple MHG update, we will later generalize to cases in which (3.24) does not hold.

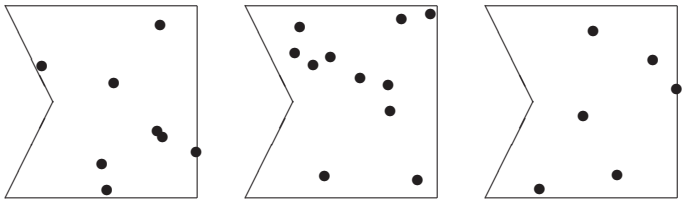


Figure 3.6: Three realizations of the same spatial point process.

coincide, that is, with probability one the location of every point is different. A point process if *finite* if the number of points is finite with probability one. We will only be interested in finite simple point processes.

The process illustrated in Figure 3.6 is the simplest of all spatial point processes, the *homogeneous Poisson process*, which is simulated as follows.

- Simulate a Poisson random variate N.
- Simulate N i. i. d. points uniformly distributed in the region.

For the patterns in Figure 3.6, the expected number of points was 8.75 (the actual numbers are 8, 11, and 6). Any nonnegative number of points is possible, including zero (the empty pattern) though this may be very rare (probability 1.6×10^{-4} in this example). The notch in the side of the region is only to avoid being square. The region can be any shape.

For any point process on a region A and any measurable subset B of A, let N_B denote the number of points in B. This is a random variable, because it is a function of the random point pattern. Define $\lambda(B) = E(N_B)$. Then λ is a positive measure on A, called the *parameter measure* of the process. When the process is simple, the only case of interest to us, λ is also called the *intensity measure* of the process.

Any finite, nonatomic[7] measure λ on a region A determines an *inhomogeneous Poisson process* with intensity measure λ, which is simulated as follows.

- Simulate a Poisson random variate N with expectation $\lambda(A)$.
- Simulate N i. i. d. points with distribution ν defined by

$$\nu(B) = \lambda(B)/\lambda(A) \tag{3.27}$$

It is a remarkable fact about the Poisson process that it has two characterizations that have no obvious connection with each other.

Theorem 3.6. *In order that a simple, finite point process be Poisson, it is necessary and sufficient that there be a finite nonatomic measure λ such that $E(N_B) = \lambda(B)$ for each measurable set B.*

[7] A measure is *nonatomic* if every one-point set has measure zero. A positive measure λ is *finite* if $\lambda(A) < \infty$.

This combines Theorems 2.4.II and 2.4.III in Daley and Vere-Jones (1988).

Theorem 3.7. *In order that a simple, finite point process be Poisson, it is necessary and sufficient that for any measurable partition B_1, B_2, ..., B_k of the domain, the random variables N_{B_1}, N_{B_2}, ..., N_{B_k} are independent.*

This is Theorem 2.4.VII in Daley and Vere-Jones (1988). That the simulation method described above satisfies the characterizations in the theorems is left as an exercise (Exercise 3.5).

Non-Poisson Processes

So far we have gotten away with not precisely specifying the probability measure for the Poisson process, or even the sample space. This turns out to be slightly tricky, the issue being whether we consider the points of the pattern to be ordered or not. Notationally, the easiest to work with is to consider ordered patterns of points. Then conditional on $N_A = n$, the n points of the pattern are an element of A^n. This is not the Right Thing because we really want to consider the points as unordered, in which case the ordered view overcounts by distinguishing the $n!$ permutations of n points. However, the Wrong Thing can be made to work as long as we choose probability models that are symmetric under permutations of the points in a pattern. Then both views will produce the same answers to all questions that do not explicitly mention the ordering. For more on this issue, see Daley and Vere-Jones (1988, Section 5.3).

In the "ordered view," the state space of a finite simple point process in a region A can be taken to be

$$S = \bigcup_{n=0}^{\infty} A^n.$$

When there are n points, the state is a vector of a points in A, hence an element of A^n. A^0 is the singleton set $\{\varnothing\}$. This agrees with the definition of A^0 in abstract set theory, where 0 is defined to be the empty set, so $A^0 = A^{\varnothing}$, which is the set of all functions from the empty set to A and there is one such function, the empty function. This notation is felicitous, the empty set being an appropriate notation to represent the empty point pattern having zero points. If $\mathcal{A}$ is the σ-field for A, then the product σ-field for A^k is denoted $\mathcal{A}^k$, and the natural σ-field for S, call it $\mathcal{B}$, is the family of sets $B \subset S$ such that $B \cap A^k$ is an element of $\mathcal{A}^k$.

Now we can write down the probability measure of the Poisson process with intensity measure λ. It is a measure P on $(S, \mathcal{B})$ defined by

$$P(B) = \sum_{n=0}^{\infty} \frac{\lambda^n(B \cap A^n)}{n!} e^{-\lambda(A)}, \qquad B \in \mathcal{B}.$$

We see that this is the right formula because

$$\Pr(N_A = n) = P(A^n) = \frac{\lambda^n(A^n)}{n!} e^{-\lambda(A)} = \frac{\lambda(A)^n}{n!} e^{-\lambda(A)}$$

which is the right formula for N_A to be Poisson with mean $\lambda(A)$, and

$$\Pr(X \in B | N_A = n) = \frac{P(B \cap A^n)}{P(A^n)} = \frac{\lambda^n(B \cap A^n)}{\lambda^n(A^n)}$$

is just λ^n renormalized to be a probability measure, which is also the right thing (the n points are i. i. d. because λ^n is product measure). It saves a little bit of ink in formulas if we also define the unnormalized measure μ for the Poisson process that throws away the constant $e^{-\lambda(A)}$, giving

$$\mu(B) = \sum_{n=0}^{\infty} \frac{\lambda^n(B \cap A^n)}{n!}, \qquad B \in \mathcal{B}. \tag{3.28}$$

We now want to consider families of probability distributions for point processes defined by families of unnormalized densities $\{ h_\theta : \theta \in \Theta \}$ with respect to μ. The Poisson process is symmetric under permutation of the points in the point patterns. We want the same property for our new models. Write $x \equiv y$ if $x, y \in S$ are patterns having the same number of points and the same locations of the points only a different ordering. Then we need to require that our unnormalized densities satisfy the symmetry requirement

$$h_\theta(x) = h_\theta(y), \qquad \text{whenever } x \equiv y. \tag{3.29}$$

Recall that h_θ on S is an unnormalized density if it is nonnegative, not almost everywhere zero, and integrable. The first two are easy to check. The last is not trivial. The normalizing function for the family is given by

$$c(\theta) = \int h_\theta(x) \mu(dx) = \sum_{n=0}^{\infty} \frac{1}{n!} \int_{A^n} h_\theta(x) \lambda^n(dx)$$

if the integral is finite (that's what we have to check). The normalized density f_θ corresponding to h_θ is given, as usual, by (3.17), from which we see that the probability of a measurable set B in S is

$$\frac{1}{c(\theta)} \int_B h_\theta(x) \mu(dx) = \frac{1}{c(\theta)} \sum_{n=0}^{\infty} \frac{1}{n!} \int_{B \cap A^n} h_\theta(x) \lambda^n(dx) \tag{3.30}$$

It turns out that for a variety of reasons we will only be interested in processes that satisfy the following stability condition

Condition 3.8. *A process with unnormalized density h with respect to μ is* stable *if there exists a real number M such that*

$$h(x \cup \xi) \le M h(x), \qquad \text{for all } x \in S \text{ and } \xi \in S. \tag{3.31}$$

This condition will have other uses later on. For now, it implies that the normalizing function is finite on Θ. First we see that if x has n points, then by using (3.31) n times, we obtain

$$h_\theta(x) \le M^n h(\varnothing)$$

and this implies

$$c(\theta) \le h(\varnothing) \sum_{n=0}^{\infty} \frac{M^n}{n!} \int_{A^n} \lambda^n(dx) = h(\varnothing) \sum_{n=0}^{\infty} \frac{M^n \lambda(A)^n}{n!} = h(\varnothing) e^{M\lambda(A)}$$

which is finite. For more on this subject see the chapter by Geyer and the chapter by Baddeley in (Kendall, Barndorff-Nielsen, and van Lieshout 1998).

Simulating Spatial Point Processes

This section actually discusses a "prequel" of the Metropolis-Hastings-Green algorithm, a method for simulating spatial point processes due to Geyer and Møller (1994) that, although a special case of Metropolis-Hastings-Green, was invented prior to it. This is typical of the way theories develop, special cases first, general theories later.

It is a truism that textbooks and research papers make for bad history, bad psychology, and bad sociology of science. Textbooks and papers never tell it like it was and hence are useless for learning about science was done or should be done. Authors start with a half-baked idea, often a wrong idea. They work it over, modify it to make proofs easier (or possible!) or interpretations simpler. Sometimes they make the treatment more abstract and mathematically sophisticated. By the time an article appears in print, there may be no trace of the train of thought that lead the authors to their discovery. Result: you can't learn about how to do science by reading science (or math). Textbooks are worse. The start with the distortions of the original authors and add more of their own. One of the best services the author of a textbook can perform is to really clean up a subject, eliminating all the blind alleys and presenting a clear path through the material. But that really distorts the history. It requires presenting material out of historical sequence and selecting material to present on the basis of importance to the textbook author's take on the subject rather than historical importance. This book is no different, but for once, I'll present a subject as it really developed.

One way to think of the state of a point process is as a random integer N and a random N-vector $X = (X_1, \ldots, X_N)$. Before Green (1995) there was no general method for simulating such a thing, no way to "jump dimensions". But if we could put every state on a space of the same dimension, we could use ordinary Metropolis-Hastings. No finite dimensional space will do, so let's pad out the space to $\mathbb{R}^\infty$. Now the state of the point process is a random nonnegative integer N and a random sequence $X = (X_1, X_2, \ldots) \in \mathbb{R}^\infty$. The observable state of the point process is $(X_1, \ldots, X_N)$. The rest of the variables are junk added to help us apply the Metropolis algorithm. They can be defined any way we like. A simple definition that turns out to be useful is to define them to be i. i. d. on the region containing the process.

Starting with a model having unnormalized density h_θ with respect to the measure μ defined by (3.28), which is proportional to the probability measure for a Poisson process with intensity measure λ, we want to define a new model

as one having unnormalized density $\tilde{h}_\theta(x, n)$ with respect to some measure $\tilde{\mu}$ on $\mathbb{R}^\infty \times \mathbb{N}$. We take $\tilde{\mu}$ to be the measure on $\mathbb{R}^\infty \times \mathbb{N}$ that is ν^∞ times counting measure on $\mathbb{N}$, where ν is the measure defined by (3.27), that is, λ normalized to be a probability measure. Then we define $\tilde{h}_\theta$ by

$$\tilde{h}_\theta(x, n) = \frac{h_\theta\big((x_1, \ldots, x_n)\big)\lambda(A)^n}{n!} \tag{3.32}$$

Since (3.32) does not involve x_{n+1} x_{n+2}, ..., it says that conditional on $N = n$ the variable X_{n+i} is independent of all other X_k and has the distribution ν, which was one property we wanted. It is also clear that for any measurable set B in A^n that

$$\Pr\big((x_1, \ldots, x_n) \in B \,\&\, N = n\big) = \frac{1}{n!} \int_B h_\theta(x) \lambda^n(dx)$$

Comparing with (3.30) we see that this model does capture the same probability structure as the other.

Now consider a Metropolis-Hastings update of N. The simplest is to propose to increase N by one with probability $\frac{1}{2}$ and decrease it by one with probability $\frac{1}{2}$ (unless $N = 0$ already, in which case increase N by one with probability $\frac{1}{2}$ and do nothing with probability $\frac{1}{2}$). This is a Metropolis proposal: between each two numbers n and $n+1$ there is the same probability of a proposal going up and a proposal going down (i. e., $\frac{1}{2}$). The odds ratio for a move from n to $n + 1$ is

$$R = \frac{h_\theta\big((x_1, \ldots, x_{n+1})\big)}{h_\theta\big((x_1, \ldots, x_n)\big)} \cdot \frac{\lambda(A)}{n+1} \tag{3.33}$$

and the odds ratio for a move the other way, from $n + 1$ to n is the reciprocal of (3.33), but we usually think of a move from n to $n - 1$ (the current position being n). That gives

$$R = \frac{h_\theta\big((x_1, \ldots, x_{n-1})\big)}{h_\theta\big((x_1, \ldots, x_n)\big)} \cdot \frac{n}{\lambda(A)} \tag{3.34}$$

One problem with this description of the algorithm is that it seems to require an infinite state. We can't allow that! But since the infinite tail past N is independent of the part of the state we are interested in, we can ignore it and simulate as needed. When we move from n to $n + 1$ we get a new X_{n+1}, but it is independent of the other X_i and has distribution ν. We can simulated it when needed in the proposal part of the update.

One update step, starting from current position $(x_1, \ldots, x_n)$ goes as follows.

1. Flip a coin. On heads try to move from n to $n + 1$. On tails, try to move from n to $n - 1$, unless $n = 0$, in which case skip the remaining steps (doing nothing).

2. If going up simulate x_{n+1} independent of the current state and having distribution ν given by (3.27).

3. Evaluate the odds ratio, (3.33) if going up or (3.34) if going down.

4. Accept the move with probability $\min(1, R)$.

There's no question algorithm has the correct invariant distribution. It's just Metropolis. There's nothing fancy about it except for the somewhat mysterious and ghostly infinite sequence of random variables that are only used in woofing about the algorithm, playing no role in the simulation. It seems likely that most examples of the Metropolis-Hastings-Green algorithm could be treated something like this, thereby eliminating any need to know what Radon-Nikodym derivatives are, but then the algorithm would lose its generality and every doable example would require a special story with its own special ghostly variables. Better to suffer the measure theory.

So let's translate our algorithm into Metropolis-Hastings-Green terminology. We know what the proposal is, going down, we will delete x_n, and going up we will add a new x_{n+1}, which will have distribution ν given by (3.27). The way Green's algorithm works is that one kernel, call it Q_n describes both a move and its "reverse move". If Q_n describes a move up from A^n to A^{n+1}, it should also describe the reverse move down from A^{n+1} to A^n. To keep things simple, we should leave it at that. Then there will be a different Q_n for every $n \ge 0$.

The next task is to figure out what $\pi(dx)Q_n(x, dy)$ is in each case, going up or down. Going up the current state x can be any element of A^n, but the proposal y must agree with x in the first n coordinates, so the pair (x, y) is concentrated on the set

$$D_n = \{\, (x, y) \in S^2 : x \in A^n,\ y \in A^{n+1},\ x_i = y_i,\ i = 1, \ldots, n \,\}.$$

The unnormalized joint distribution of (x, y) is

$$\begin{aligned} \eta(dx)Q_n(x, dy) &= h_\theta(x)\mu(dx)I(x, A^n)\frac{\lambda(dy_{n+1})}{\lambda(A)} \\ &= h_\theta(x)\frac{\lambda^n(dx)}{n!}\frac{\lambda(dy_{n+1})}{\lambda(A)} \\ &= h_\theta(x)\frac{\lambda^{n+1}(dy)}{n!\lambda(A)} \end{aligned} \tag{3.35}$$

Going down the current state x can be any element of A^{n+1} and the proposal y is deterministic, being the element of A^n that agrees with x in the first n coordinates, so the pair (x, y) is concentrated on the set $\varphi(D_n)$ where φ is the function that swaps coordinates in S^2, that is, $\varphi : (x, y) \mapsto (y, x)$. The unnormalized joint distribution of (x, y) is

$$\begin{aligned} \eta(dx)Q_n(x, dy) &= h_\theta(x)\mu(dx)I(x, A^{n+1}) \\ &= h_\theta(x)\frac{\lambda^{n+1}(dx)}{(n+1)!} \end{aligned} \tag{3.36}$$

Thus going up Green's ratio is (3.36) with x and y interchanged divided by (3.35)

$$R(x, y) = \frac{h_\theta(y)}{h_\theta(x)} \cdot \frac{\lambda(A)}{n + 1}$$

which is just the expression we had before, (3.33) in slightly different notation. Similarly going down Green's ratio is (3.35) with x and y interchanged divided by (3.36)

$$R(x, y) = \frac{h_\theta(y)}{h_\theta(x)} \cdot \frac{n + 1}{\lambda(A)}$$

which agrees with (3.34) when we recall that in (3.34) we had changed n to $n - 1$.

In calculating Green's ratio we just "cancelled" the $\lambda^{n+1}(dy)$ terms in the numerator and denominator. To be very careful, we should have checked that (3.24) holds, but it obviously does.

A minor blemish on this algorithm is the way it treats the points in the pattern asymmetrically. Recall that we really consider the points unordered. We insist that the model have a symmetric density, so that the probability of a pattern does not depend on the ordering of the points. But the MHG algorithm described above doesn't treat the points symmetrically. It always adds or deletes the last point in the ordering. We can cure this blemish by composing our MHG update with another basic update, which simply reorders the n points of the pattern, choosing among the $n!$ orders with equal probability. This clearly preserves the distribution with unnormalized density h_θ because we have required h_θ to be symmetric. We do not even have to actually permute the points. The only effect this random permutation has on the MHG updates is that in steps down a random point rather than the n-th is deleted. This gives us an algorithm that reflects the symmetry of the model.

As usual, we describe one basic update step starting at a pattern x with n points

1. Flip a coin. On heads try to add a point. On tails, try to delete one (or if $n = 0$ so there are no points to delete, do nothing, skip the remaining steps).

2. If going up simulate ξ independent of the current state and having distribution ν given by (3.27).

3. Evaluate the odds ratio, (3.33) if going up or (3.34) if going down.

4. Accept the move with probability $\min(1, R)$.

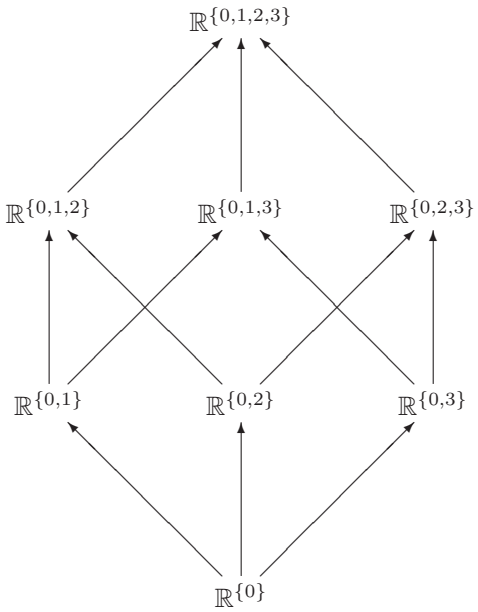
3.3.3 Bayesian Model Selection

The Bayesian competitor to frequentist model selection procedures (like all-subsets regression) involves computing Bayes factors for the various models under consideration. For a concrete example, consider again Bayesian logistic

regression (Example 3.3). In that model there were three predictors. There are $2^3 = 8$ different models that can be formed by including or excluding any of these predictors. One, the *full model*, which has all three predictors and four regression coefficients including the intercept, is the one we already analyzed in Example 3.3. Another, the *null model* has no predictors and just one regression coefficient, the intercept, and just fits a Bernoulli model to the data (i. e. the data Y_i are i. i. d. Bernoulli(p) with p the single unknown parameter). Between these are three models with one predictor and another three with two predictors. The *model selection problem* is to select the single model that that best fits the observed data. The *model comparison problem* is a bit more vague. It only asks for comparison of the models, leaving a decision to the user. The Bayesian solution to either involves Bayes factors.

The parameter spaces for different submodels typically have different dimensions. For our logistic regression example, the parameter spaces have dimensions between one (for the null model) and four (for the full model). The parameter spaces for the models have the form $\mathbb{R}^I$, where I is a subset of $\{0, 1, 2, 3\}$ that contains 0, and are shown in the diagram below.[8] The parameter spaces of the logistic regression model selection problem are partially ordered by embedding, the arrows in the diagram denoting the natural embeddings, which set certain coordinates to zero, for example, the the arrow going from $\mathbb{R}^{\{0,1,2\}}$ to $\mathbb{R}^{\{0,2\}}$ represents the embedding $(\beta_0, 0, \beta_2) \mapsto (\beta_0, \beta_2)$.

[8]Recall that $\mathbb{R}^S$ means the set of all functions from S to $\mathbb{R}$, hence an element $\beta \in \mathbb{R}^{\{0,1,3\}}$ is a function from $\{0, 1, 3\}$ to $\mathbb{R}$, which can be specified by giving its values $\beta(0)$, $\beta(1)$ and $\beta(3)$ at the points of the domain. If we write β_i instead of $\beta(i)$ we get the more familiar notation for vectors. An element $\beta \in \mathbb{R}^{\{0,1,3\}}$ represents a 3-vector $(\beta_0, \beta_1, \beta_3)$. Notice the value of the notation. The parameter spaces $\mathbb{R}^{\{0,1,3\}}$ and $\mathbb{R}^{\{0,2,3\}}$ are different. They index different models. If we denoted both of them by $\mathbb{R}^3$, we would not be able to distinguish them.



We now need an abstract framework that describes any model selection problem. Let $\mathcal{M}$ be an index set for the models. Corresponding to a model $M \in \mathcal{M}$, there is a parameter space Θ_M. In the logistic regression problem the Θ_M are the spaces $\mathbb{R}^I$ in the diagram. Assume the Θ_M are disjoint. Then the parameter space for the entire problem is the union[9]

$$\Theta = \bigcup_{M \in \mathcal{M}} \Theta_M.$$

For each $\theta \in \Theta$ there is a data model $f(x|\theta)$, and there is also a prior, which is a probability measure γ on Θ. In model comparison, proper priors are *de rigeur*. See Bernardo and Smith (1994, pp. 421–424) for the reasons why, and read all of Chapter 6 in Bernardo and Smith (1994) if you really want to understand Bayesian model comparison.

The object of Bayesian analysis is, as always, to calculate the posterior. In the model comparison problem, we are not interested in the posterior distribution of the parameter values θ, but only in the posterior probabilities of the models

$$p(M|x) = \frac{\int_{\Theta_M} f(x|\theta)\gamma(d\theta)}{\int_{\Theta} f(x|\theta)\gamma(d\theta)}$$

We do not need the denominator, since we are only interested in the relative probabilities of the models

$$p(M|x) \propto \int_{\Theta_M} f(x|\theta)\gamma(d\theta)$$

[9]If the Θ_M were not disjoint, then we would have to use the notion of *disjoint union* (Jänich 1984, p. 10), which treats the sets as if they were disjoint.

and not even in them, exactly. The prior γ can be divided into two parts: the marginal for the models $\gamma(\Theta_M)$ and the conditional distribution for θ given M

$$\gamma(A|M) = \frac{\gamma(A \cap \Theta_M)}{\gamma(\Theta_M)}$$

If you and I agree about the conditional of θ given M, but disagree about the marginals, then our posterior probabilities will be proportional to our prior probabilities

$$p(M|x) \propto \gamma(\Theta_M) \int_{\Theta_M} f(x|\theta)\gamma(d\theta|M)$$

One way to take out part of the subjectivity involved in this inference is to divide by the prior odds $\gamma(\Theta_M)$. This gives the *Bayes factor*, which is the ratio of posterior to prior odds

$$B(M) = \frac{p(M|x)}{\gamma(\Theta_M)} \propto \int_{\Theta_M} f(x|\theta)\gamma(d\theta|M).$$

The integral defines the Bayes factors up to an overall constant of proportionality. Call it the *unnormalized* Bayes factors

$$B_u(M) = \int_{\Theta_M} f(x|\theta)\gamma(d\theta|M).$$

To use the Bayes factors to compare models, you multiply $B_u(M)$ by your (or your client's) personal prior probabilities $\gamma(\Theta_M)$ to obtain your own posterior model probabilities $p(M|x)$ up to a constant of proportionality. The constant usually does not matter. For example, the solution to the model selection problem is to select the model with the highest $p(M|x)$ and this is the same as the model with the highest $\gamma(\Theta_M)B_u(M)$ because multiplying by a constant does not change which model is highest. If you need actual probabilities, simply normalize the unnormalized Bayes factors by dividing by their sum

$$p(M|x) = \frac{\gamma(\Theta_M)B_u(M)}{\sum_{M \in \mathcal{M}} \gamma(\Theta_M)B_u(M)}$$

To return to our logistic regression model, the data model is the same as before (Example 3.3). The only difference is that for the submodels we set some of the regression coefficients β_i to zero. So far we haven't specified the set $\mathcal{M}$ except to say that it indexes the models. To be specific now, let $\mathcal{M}$ be the set of exponents in the diagram, the subsets of $\{0, 1, 2, 3\}$ that contain 0. Then $\Theta_M = \mathbb{R}^M$. The prior must be a probability measure on $\Theta = \bigcup_M \Theta_M$. Only measure theory gives us a simple notation for something like that. We might, for example, choose a normal distribution for the restriction of γ to the parameter space $\mathbb{R}^{\{0,1,2,3\}}$ of the full model and obtain the all the restrictions of γ to the parameter spaces of the submodels by conditioning the normal distribution for the full model to lie in the the parameter spaces of the submodels.[10]

[10]To be continued. The code for an MHG sampler for this model is yet to be written.

3.3.4 Metropolis-Hastings-Green, the General Case

The description of the MHG update given in the preceding section is usable for many problems, but in some respects it is a step backward. It doesn't include some ordinary Metropolis updates, such as the one for the dumbell distribution.

Radon-Nikodym Derivatives and Lebesgue Decomposition

This section briefly sketches three important measure-theoretic notions: absolute continuity, Lebesgue decomposition, and Radon-Nikodym derivatives.

If μ and ν are two positive measures on the same measurable space $(S, \mathcal{B})$, we say μ is *absolutely continuous* with respect to μ if $\nu(B) = 0$ implies $\mu(B) = 0$. An alternative terminology is that ν *dominates* μ. A notation indicating this condition is $\mu \ll \nu$.

If $\mu \ll \nu$ and $\nu \ll \mu$, we say that μ is *equivalent* to ν and write $\mu \sim \nu$. Note that this says only that μ and ν have the same null sets. It is easy to see that this is an equivalence relation on the class of all positive real measures.

A function f on S is said to be a *density* of μ with respect ν if

$$\mu(B) = \int_B f(x)\nu(dx), \qquad B \in \mathcal{B}, \tag{3.37}$$

which implies

$$\int g(x)\mu(dx) = \int g(x)f(x)\nu(dx)$$

for any integrable function g. This is a generalization of the usual notion of a probability density function. When ν is Lebesgue measure dx and μ is a probability measure, f is just the familiar p. d. f. of μ.

The Radon-Nikodym theorem (Rudin 1987, Theorem 6.10) says that $\mu \ll \nu$ implies that μ has a density with respect to μ. The converse assertion is also true: if (3.37) holds, then $\mu \ll \nu$.

The Radon-Nikodym theorem implies the existence of a density, but is it unique? Since integrals over sets of measure zero are zero, a density can be redefined arbitrarily on a set of measure zero and still be a density. But an elementary theorem of measure theory (Rudin 1987, Theorem 1.39(b)) says that is the only arbitrariness allowed: two densities of μ with respect to ν must be equal except on a set of ν measure zero. Another way to say this is that if f is a density of μ with respect to ν, then f is unique considered as an element of $L^1(\nu)$.

Because a density f of μ with respect to ν is unique (in the L^1 sense), it makes sense to give it a name and notation as something determined by μ and ν. When (3.37) holds, we say that f is the *Radon-Nikodym derivative* of μ with respect to ν and write

$$f = \frac{d\mu}{d\nu}$$

This is just another terminology for (3.37). We are not defining a new operation.

So now we see where (3.24) comes from. If the measure in the numerator of (3.23) is absolutely continuous with respect to the measure in the denominator, then the condition that $R(x, y)$ be a density of one with respect to the other is (3.24). We now want to generalize to the situation when absolute continuity is not present.

Measures μ and ν are *mutually singular* if there exists a measurable set B such that $\mu(B) = 0$ and $\nu(B^c) = 0$ (hence μ is concentrated on B^c and ν is concentrated on B). A notation indicating this condition is $\mu \perp \nu$. In a sense mutual singularity is the opposite of absolute continuity.

The Lebesgue decomposition theorem (Rudin 1987, Theorem 6.10) says that if μ and ν are arbitrary positive real measures on the same state space, then μ can be decomposed as the sum $\mu = \mu_a + \mu_s$, where $\mu_a \ll \nu$ and $\mu_s \perp \nu$. The pair (μ_a, μ_s) is called the *Lebesgue decomposition* of μ relative to ν.

Now we can give the most general notion of a Radon-Nikodym derivative. If μ and ν are arbitrary postive real measures on the same state space, and $\mu = \mu_a + \mu_s$ is the Lebesgue decomposition of μ relative to ν, then we often say that $f = d\mu_a/d\nu$ is the Radon-Nikodym derivative of μ with respect to ν. Of course, f is now the density of μ_a (not μ) with respect to ν, but that is the best we can do. The mutually singular part μ_s has no relation to ν whatsoever.

With these preliminaries out of the way, let us return to considering what (3.23) means. We said it was a Radon-Nikodym derivative, but of what measures? It is obvious that the intention is that $\eta(dx)Q(x, dy)$ indicate the unnormalized joint distribution of the current state x and the proposal y. To be mathematically precise we must define this as a measure μ on $(S^2, \mathcal{B}^2)$ by

$$\mu(B) = \iint 1_B(x, y)\eta(dx)Q(x, dy), \qquad B \in \mathcal{B}^2. \tag{3.38}$$

The numerator in (3.23) is the denominator with x and y reversed, but μ is a function of one argument (the set B) rather than two, so we can't obtain the measure in the numerator by swapping arguments. Instead we have to proceed a bit differently, first defining the function $\varphi : (x, y) \mapsto (y, x)$ that switches coordinates in S^2. Then the measure in the numerator is $\mu \circ \varphi$, defined by

$$(\mu \circ \varphi)(B) = \mu[\varphi(B)]. \tag{3.39}$$

So we finally have a rigorous general definition of Green's ratio

$$R = \frac{d(\mu \circ \varphi)}{d\mu} \tag{3.40}$$

where μ is defined by (3.38).

The following lemmas give some useful properties of Radon-Nikodym derivatives that are helpful in calculations.

Lemma 3.9 (Chain Rule)**.** *If $\lambda \ll \mu \ll \nu$, then*

$$\frac{d\lambda}{d\nu} = \frac{d\lambda}{d\mu} \cdot \frac{d\mu}{d\nu}$$

holds ν almost everywhere.

Corollary 3.10 (Reciprocal Rule)**.** *If $\mu \sim \nu$, then*

$$\frac{d\mu}{d\nu} = \left(\frac{d\nu}{d\mu} \right)^{-1} \tag{3.41}$$

holds μ almost everywhere.

Remark. "μ almost everywhere" here is the same as "ν almost everywhere" because μ and ν have the same null sets. The set on which the right hand side is undefined because $\frac{d\nu}{d\mu} = 0$ is a set of ν measure zero. Hence we may define the right hand side arbitrarily for such points so long as we produce a measurable function (for example, we could set it equal to an arbitrary constant).

Lemma 3.9 and Corollary 3.10 are Problems 32 and 33 of Chapter 8 in Fristedt and Gray (1997).

Corollary 3.11 (Ratio Rule)**.** *If $\mu \ll \xi$ and $\nu \ll \xi$, then*

$$\frac{d\mu}{d\nu} = \frac{\frac{d\mu}{d\xi}}{\frac{d\nu}{d\xi}} \tag{3.42}$$

holds ν almost everywhere.

Remark. The set on which the right hand side is undefined because $\frac{d\nu}{d\xi} = 0$ is a set of ν measure zero. Hence we may define the right hand side arbitrarily for such points so long as we produce a measurable function.

Proof. Let (μ_a, μ_s) be the Lebesgue decomposition of μ with respect to ν. Then $\mu_a \ll \nu \ll \xi$, so by the chain rule

$$\frac{d\mu_a}{d\xi} = \frac{d\mu_a}{d\nu} \frac{d\nu}{d\xi}.$$

Also

$$\frac{d\mu_s}{d\xi} \frac{d\nu}{d\xi} = 0, \qquad \xi \text{ almost everywhere,}$$

because otherwise we would have $\mu_s(B) > 0$ and $\nu(B) > 0$ for some set B, which contradicts $\mu_s \perp \xi$. By the remark, we need only prove (3.42) when $\frac{d\nu}{d\xi} > 0$, which implies $\frac{d\mu_s}{d\xi} = 0$ and

$$\frac{d\mu}{d\nu} = \frac{d\mu_a}{d\nu} = \frac{\frac{d\mu_a}{d\xi}}{\frac{d\nu}{d\xi}} = \frac{\frac{d\mu}{d\xi}}{\frac{d\nu}{d\xi}}$$

and we are done. □

Lemma 3.12. *If (μ_a, μ_s) is the Lebesgue decomposition of μ relative to ν and (ν_a, ν_s) is the Lebesgue decomposition of ν relative to μ, then $\mu_a \perp \nu_s$, $\mu_a \sim \nu_a$, and*

$$\frac{d\mu}{d\nu} = \frac{d\mu_a}{d\nu_a}$$

Proof. First, $\mu \perp \nu_s$ implies $\mu_a \perp \nu_s$. Together with $\mu_a \ll \nu$, this implies $\mu_a \ll \nu_a$. Suppose D is a set such that $\nu_s(D) = 0$ and $\nu_a(D^c) = 0$, the existence of such a set being guaranteed by the Lebesgue decomposition theorem. Then if $f = d\mu/d\nu = d\mu_a/d\nu$

$$\mu_a(B) = \int_{B\cap D} f(x)\nu_a(dx) + \int_{B\cap D^c} f(x)\nu_s(dx) \tag{3.43}$$

Taking $B = D^c$, we get $\mu_a(D^c) \leq \nu_a(D^c) = 0$, so we must have $f(x) = 0$, $x \in D^c$. Thus the second term on the right hand side of (3.43) is always zero and f is also a density of μ_a with respect to ν_a. □

Lemma 3.13. *If φ is a function on the domain of μ satisfying $\varphi = \varphi^{-1}$. If $\nu = \mu \circ \varphi$, and if μ_a, μ_s, ν_a, ν_s are as in Lemma 3.12, then*

$$\nu_a = \mu_a \circ \varphi \quad \textit{and} \quad \nu_s = \mu_s \circ \varphi.$$

Proof. First we note that

$$\mu_a \circ \varphi + \mu_s \circ \varphi = (\mu_a + \mu_s) \circ \varphi = \nu$$

is a decomposition of ν, so what we need to show is

$$\mu_a \circ \varphi \ll \mu \tag{3.44a}$$

$$\mu_s \circ \varphi \perp \mu \tag{3.44b}$$

What we are given to work with is

$$\mu_a \ll \mu \circ \varphi \tag{3.44c}$$

$$\mu_s \perp \mu \circ \varphi \tag{3.44d}$$

(3.44a) is shown by

$$\mu(B) = 0 \Longleftrightarrow (\mu \circ \varphi)(\varphi[B]) = 0 \implies \mu_a(\varphi[B]) = 0 \Longleftrightarrow (\mu_a \circ \varphi)(B) = 0,$$

the middle implication being (3.44c) and the other implications being $\varphi = \varphi^{-1}$ and the definition of functional composition.

Now (3.44d) implies the existence of a set B such that $\mu_s(B) = (\mu\circ\varphi)(B^c) = 0$. Hence

$$(\mu_s \circ \varphi)(\varphi[B]) = \mu_s(B) = 0$$

and

$$\mu(\varphi[B]^c) = \mu(\varphi[B^c]) = (\mu \circ \varphi)(B^c) = 0$$

and this proves (3.44b). □

Corollary 3.14. *Suppose μ, μ_a and φ are as in the lemma, and ξ satisfies $\xi \circ \varphi = \xi$ and $\mu \ll \xi$. Then*

$$\frac{d(\mu \circ \varphi)}{d\mu} = \frac{\frac{d\mu}{d\xi} \circ \varphi}{\frac{d\mu}{d\xi}}$$

What does all this tell about MHG calculations? Taking (3.40) as our official definition of Green's ratio,

Metropolis-Hastings-Green is Reversible

We can now write down the transition probability kernel for the Metropolis-Hastings-Green update. As we saw with Metropolis-Hastings, the transition probability has two terms. For accepted proposals, we propose y and then accept it, which happens with probability density $a(x, \cdot\,)$ with respect to $Q(x, \cdot\,)$ where $a(x, y)$ is again the acceptance probability

$$a(x, y) = \min[1, R(x, y)].$$

Hence for any set A

$$\int_A Q(x, dy)a(x, y)$$

is the part of $P(x, A)$ that results from accepted proposals. If the integral on the right hand side is taken over the whole state space, it gives the total probability that the proposal will be accepted. Thus the probability that the proposal is rejected is

$$r(x) = 1 - \int Q(x, dy)a(x, y).$$

If the proposal is rejected we stay at x. Hence

$$P(x, A) = r(x)I(x, A) + \int_A Q(x, dy)a(x, y). \tag{3.45}$$

We now want to verify reversiblity of the MHG update, but first we collect some simple facts about Radon-Nikodym derivatives.

Lemma 3.15. *If μ and ν are positive real measures, (μ_a, μ_s) is the Lebesgue decomposition of μ relative to ν, (ν_a, ν_s) is the Lebesgue decomposition of ν relative to μ, then $\mu_a \ll \nu_a$ and $\nu_a \ll \mu_a$,*

$$\frac{d\mu}{d\nu} = \frac{d\mu_a}{d\nu_a} \qquad \textit{and} \qquad \frac{d\nu}{d\mu} = \frac{d\nu_a}{d\mu_a}.$$

Moreover,

$$\frac{d\mu_a}{d\nu_a} = \left(\frac{d\nu_a}{d\mu_a}\right)^{-1}$$

μ_a (or ν_a) almost everywhere.

Proof. Since $\mu_a \ll \nu$ and $\mu \perp \nu_s$, we must have $\mu_a \ll \nu_a$, and similarly with μ and ν reversed. If ν_a is concentrated on B and ν_s on B^c, then $\mu(B^c) = 0$, and if $f = d\mu/d\nu$, then

$$\mu(B^c) = \int_{B^c} f(x)\nu_s(dx) = 0$$

Hence $f = 0$, almost everywhere (ν_s), and

$$\mu(A) = \int_A f(x)\nu(dx) = \int_A f(x)\nu_a(dx)$$

which shows that $f = d\mu/d\nu_a = d\mu_a/d\nu_a$.

Finally, if $f = d\mu_a/d\nu_a$ and $g = d\nu_a/d\mu_a$, then

$$\mu_a(B) = \int_B f(x)\nu_a(dx) = \int_B f(x)g(x)\mu_a(dx)$$

holds for all measurable B, which implies $fg = 1$ almost everywhere (μ_a) (Rudin 1987, Theorem 1.39(b)). This is the same as almost everywhere (ν_a) because μ_a and ν_a have the same sets of measure zero. □

Corollary 3.16. *If R is defined by* (3.40)*, then $R(x, y) = 1/R(y, x)$ almost everywhere μ.*

Proof. Let μ_a denote the part of μ that is absolutely continuous with respect to $\mu \circ \varphi^{-1}$, and apply the lemma, yielding the conclusion that

$$R = \frac{d(\mu_a \circ \varphi^{-1})}{d\mu_a} \qquad \text{and} \qquad S = \frac{d\mu_a}{d(\mu_a \circ \varphi^{-1})}$$

$RS = 1$ almost everywhere μ_a, hence almost everywhere μ. Also

$$\int_B R\, d\mu_a = (\mu_a \circ \varphi^{-1})(B) = \int_{\varphi^{-1}(B)} d\mu_a = \int_{\varphi^{-1}(B)} S\, d(\mu_a \circ \varphi^{-1}) = \int_B (S \circ \varphi)\, d\mu_a,$$

the first equality being the definition of R, the second the definition of $\mu_a \circ \varphi^{-1}$, the third the definition of S, and the fourth the change of variable theorem for abstract integration (Billingsley 1979, Theorem 16.12). Since this holds for all B, we conclude $R = S \circ \varphi$. □

Theorem 3.17. *The Metropolis-Hastings-Green update is reversible with respect to η.*

Proof. What is to be shown is that

$$\begin{aligned} &\iint f(x)g(y)\eta(dx)P(x, dy) \\ &\qquad = \int f(x)g(x)r(x)\eta(dx) + \iint f(x)g(y)\eta(dx)Q(x, dy)a(x, y) \end{aligned}$$

is unchanged when we interchange f and g, as in the proof of Lemma 3.4. Again, the first term is obviously unchanged by interchanging f and g. So we work on the second term.

$$\begin{aligned} \iint f(x)g(y)a(x, y)\eta(dx)Q(x, dy) &= \iint f(y)g(x)a(y, x)\eta(dy)Q(y, dx) \\ &= \iint f(y)g(x)a(y, x)R(x, y)\eta(dx)Q(x, dy) \end{aligned}$$

the first equality from interchanging the dummy variables x and y and the second being (3.23). In order to finish the proof we only need to show that

$$a(x, y) = a(y, x)R(x, y), \qquad x, y \in S, \tag{3.46}$$

which is the "detailed balance for densities" condition analogous to (3.22) that we need here.

The proof is just like the proof of Corollary 3.5. In the case $R(x, y) \geq 1$ we have

$$a(x, y) = 1 \qquad \text{and} \qquad a(y, x) = R(y, x) \tag{3.47}$$

which implies (3.46), and in the case (3.23) less than or equal to one we have (3.47) with x and y interchanged, which also implies (3.46). Now

$$\begin{aligned} \int g(x)h(y)a(x, y)f(x, y)\xi(dx, dy) &= \int g(x)h(y)a(y, x)f(y, x)\xi(dx, dy) \\ &= \int g(y)h(x)a(x, y)f(x, y)\xi(dy, dx) \\ &= \int g(y)h(x)a(x, y)f(x, y)\xi(dx, dy) \end{aligned} \tag{3.48}$$

where (3.46) gives the first equality, interchanging the dummy variables x and y gives the second, and the symmetry of ξ gives the third. We do not need Fubini here, because there are no iterated integrals.[11] □

Exercises

3.1. Prove that Gibbs updates are idempotent (satisfy $P^2 = P$).

3.2. Prove that if each kernel P_z in Theorem 3.1 is reversible with respect to π, then so is the kernel Q.

3.3. Verify directly that lines 2 and 3 of (3.2) are equal, that is, count the number of terms in the double sum, divide by $d! \cdot (d - 1)$ and get d.

[11] We do need something, because, strictly speaking, the notation $\xi(dx, dy)$ is meaningless, ξ being a measure on S^2. What we need is the general change of variable formula for integration, for any function w, any measure ξ, and any measurable transformation φ

$$\int (w \circ \varphi)\, d\xi = \int w\, d(\xi \circ \varphi^{-1})$$

(Billingsley 1979, Theorem 16.12).

A formally correct argument now goes as follows. Let

$$w(x, y) = g(y)h(x)a(y, x)f(y, x)$$

[the last integrand in (3.48)]. Then we can rewrite the second and third equalities in (3.48) as

$$\int (w \circ \varphi)\, d\xi = \int w\, d(\xi \circ \varphi^{-1}) = \int w\, d\xi$$

the first equality being the change-of-variable formula and the second being the symmetry of ξ.

3.4. Explain why μ was not required to be a σ-finite measure in the definition of "unnormalized probability density" at the beginning of Section 3.2.1. Show that if h is an unnormalized density with respect to μ and h is strictly positive, then μ is automatically σ-finite, it need not be part of the definition. Then show that even if h is not strictly positive, the restriction of μ to the support of h (i. e., the set $\{ x : h(x) > 0 \}$ is σ-finite.

3.5. Show that the simulation method described for the Poisson process does indeed satisfy the characterizations in Theorems 3.6 and 3.7.

3.6. Redo the logistic regression example using the kyphosis data set that comes with S-PLUS. Calculate posterior means and variances with Monte Carlo standard errors. The info on the computing info web page may help.

If you are feeling adventurous, do probit instead of logit regression (the C library functions `erf` and `erfc` may help with the probit calculation).

3.7. Show that one-variable-at-a-time Metropolis-Hastings is a special case of Metropolis-Hastings-Green.

3.8. Formulate the Metropolis-Hastings analog of the hit and run algorithm of Section 3.1.3. Show that your algorithm is a special case of Metropolis-Hastings-Green with general state-dependent mixing and hence is valid with no further proofs. This is not new, see Chen and Schmeiser (1993), but don't look up the reference. Reinvent the wheel.

Chapter 4

Stochastic Stability

This chapter discusses asymptotics of Markov chains, or, as Meyn and Tweedie (1993) call it, the "stochastic stability" of Markov chains. We shall see that in many respects Markov chains are no so different from independent samples, and hence Markov chain Monte Carlo is not so different from ordinary independent-sample Monte Carlo.

In particular, the law of large numbers and the central limit theorem still hold for many Markov chains, although the conditions that must be verified in order to know whether they hold are more complicated than in the case of independent sampling. Whatever one does in independent-sample Monte Carlo can also be done in MCMC.

The difference between Markov chains and independent sampling is that with independent sampling there is a tight connection between the size of errors that can occur and the probability of the relevant events. To take the simplest possible example, suppose the distribution of interest is π and we are interested in the probability of a set A with $0 < \pi(A) < 1$. We are to estimate $\pi(A)$ by ordinary Monte Carlo using independent simulations X_1, X_2, ... from π. Consider the the probability that all n samples completely miss A giving us a Monte Carlo estimate of zero for the probability of A. Although the absolute error is small if $\pi(A)$ is small, the relative error is not. The probability of this error is

$$[1 - \pi(A)]^n$$

which goes to zero exponentially fast, and what is more important, at a rate which is determined by $\pi(A)$.

If we use MCMC, so X_1, X_2, ... is a Markov chain with invariant distribution π, the situation is qualitatively the same, but may be very different quantitatively. We usually have exponential convergence to zero of the probability that an n-sample entirely misses A. For so-called geometrically ergodic chains, for π-almost any starting point x the number of iterations s_A that the chain takes to hit A has a moment generating function, that is, for some $r > 1$ the expectation of r^{s_A} is finite (Nummelin 1984, Proposition 5.19). Thus by Markov's

inequality, there exists a constant $M < \infty$ such that

$$\Pr(s_A \geq n) \leq Mr^{-n}$$

which says the same thing as in the independent case except that we usually have no sharp bounds for M and r. With independence we know that $M = 1$ and $r = 1/[1 - \pi(A)]$ will do. For a Markov chain we only know that some $M < \infty$ and $r > 1$ will do.

This is not of merely theoretical concern. In practical situations, it may take a very large number of iterations to get a sample that is reasonably representative of the invariant distribution, and there is usually no simple calculation that tells us how many iterations are required.

4.1 Irreducibility

The weakest form of stochastic stability is irreducibility. Among other things, if a Markov chain has an invariant distribution and is irreducible, then the invariant distribution is unique. Irreducibility also implies that the law of large numbers holds. It has many other important consequences. One should never use a chain that is not irreducible for Monte Carlo. Irreducibility is generally easy to demonstrate. When one cannot demonstrate irreducibility for a sampling scheme, one should find a different sampling scheme for which one can demonstrate irreducibility. This is always possible, since there are so many ways to construct samplers with a specified invariant distribution.

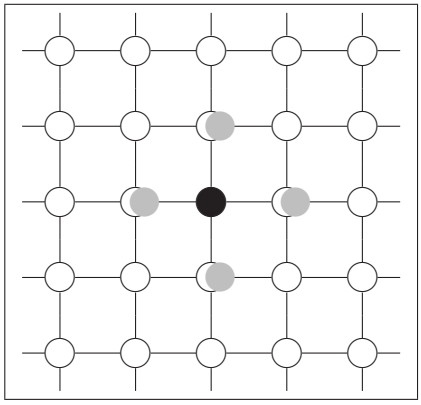
4.1.1 Countable State Spaces

Irreducibility is the one notion that has a different definition for discrete and continuous state spaces. Since both definitions are widely used, one should know both. Recall from Sections 2.1.1 and 2.2.1 that for a countable state space the transition probabilities are described by a matrix P and that the n-step transition probabilities are given by P^n. A Markov chain on a countable state space is *irreducible* if for any points x and y in the state space there exists an integer n such that $P^n(x, y) > 0$, that is, if for some n there is positive probability that the chain can move from x to y in n steps. The colloquial version of this is that the chain can get "from anywhere to anywhere" (not necessarily in one step).

In order to see how this definition works we need an example with a discrete state space.

4.1.2 The Ising Model

The Ising model is a spatial lattice process. The state is a vector $x = \{ x_i : i \in W \}$ where W is a subset of vertices of the infinite rectangular lattice $\mathbb{Z}^2$, the set of all pairs of points in the two-dimensional plane $\mathbb{R}^2$ having integer coordinates.



In the figure, the circles represent the *vertices* of the lattice. Associated with each node i there is a random variable x_i, and together these random variables form the state x of the spatial lattice process. Vertices joined by lines are called neighbors. The relation of being neighbors is denoted by $\sim$, if vertices i and j are neighbors we write $i \sim j$. In the figure, the vertices colored gray are the neighbors of the vertex colored black. In the infinite lattice, every vertex has four neighbors. When we look at a finite region W, some vertices have neighbors outside of W.

The random variables x_i making up the state of the Ising model have two possible values. These are often coded as zero and one, but for reasons of symmetry -1 and $+1$ is a better choice. When we illustrate realizations of an Ising model, we will just show a black and white image each pixel representing a variable x_i.

The probability model for the vector x is a two-parameter exponential family with unnormalized density

$$h_\theta(x) = e^{\theta_1 t_1(x) + \theta_2 t_2(x)} \tag{4.1}$$

where the canonical statistics are defined by

$$t_1(x) = \sum_{i \in W} x_i$$

and

$$t_2(x) = \sum_{\substack{i,j \\ i \sim j}} x_i x_j. \tag{4.2}$$

When the x_i take values in $\{-1, +1\}$, the first canonical statistic is the number of black pixels minus the number of white pixels, and the second canonical statistic is the number of concordant neighbor pairs (same color) minus the number of discordant neighbor pairs (different color). When the x_i take values in $\{0, 1\}$, and we use the same definitions of the canonical statistics, the same family of stochastic models are defined but the parameterization is different.

The notation in (4.2) is deliberately ambiguous about what happens at the boundary of the region W. There are three different ways in which the boundary is commonly treated.

The first is to condition on the boundary. The sums in (4.2) extend over all pairs i and j such that one of i or j is in W and the other is either in W or just outside. The variables x_j for $j \notin W$ are not part of the state of the Markov chain, they are fixed and can be thought of as another parameter of the model.

The second way is to sum only over pairs i and j that are neighbors and both in W. Then vertices at the edge of the region have fewer neighbors than the rest. This method is referred to as "free boundary conditions."

The third way is to eliminate the boundary altogether by gluing the edges of the region W together to form a torus. Then the set W is no longer a subset of the infinite lattice, but each vertex has four neighbors and there is no need to specify data on a boundary. Using a toroidal lattice is also referred to as imposing "periodic boundary conditions" because we can think of extending our finite region to the whole infinite lattice by periodic repetition. All three kinds of boundary conditions are artificial in one way or another. We will say more about dealing with boundary conditions presently.

A Gibbs or Metropolis sampler updating one vertex at a time is very simple. The Gibbs update chooses a new value for x_i from its conditional distribution given the rest, which is proportional to $h_\theta(x)$. The only terms that matter are those containing x_i, hence this conditional has the unnormalized density

$$h_\theta(x_i|x_{-i}) = e^{\theta_1 x_i + \theta_2 x_i \sum_{j\sim i} x_j}$$

The only sum required in calculating the unnormalized conditional density is the sum of the four neighbors of x_i, and the only sum required in calculating the normalized conditional distribution is over the two possible states of x_i

$$p(x_i|x_{-i}) = \frac{h_\theta(x_i|x_{-i})}{h_\theta(x_i = 0|x_{-i}) + h_\theta(x_i = 1|x_{-i})}$$

The Metropolis update is simpler still. The proposal y has the sign of x_i reversed and all the rest of the x_j unchanged. The odds ratio is

$$R = \frac{h_\theta(y)}{h_\theta(x)} = e^{-2\theta_1 x_i - 2\theta_2 x_i \sum_{j\sim i} x_j} \tag{4.3}$$

This is a symmetric proposal so the proposal is accepted with probability $\min(1, R)$.

4.1.3 Coding Sets

The elementary update steps are combined in any of the usual ways, usually by fixed scan, random scan, or random sequence scan. A fixed scan can be either a "raster scan" in which one scans along rows, and the rows follow one another in order. A better way is a scan by "coding sets" (Besag 1974; Besag, Green,

Higdon, and Mengersen 1995). If we color the lattice like a checkerboard, the red squares are one coding set and the black squares the other. The colors here are not the random variables, they are just a way of describing sets of vertices of the lattice. The random variables in the red coding set are conditionally independent given those in the black coding set and vice versa, since no vertex in the red coding set is a neighbor of any in the black coding set. For i and j not neighbors we have

$$h_\theta(x) = e^{\theta_1 x_i + \theta_2 x_i \sum_{k \sim i} x_k} e^{\theta_1 x_j + \theta_2 x_j \sum_{l \sim j} x_l} \times \text{term not containing } x_i \text{ or } x_j$$

Hence these variables are conditionally independent given the rest by the factorization criterion. If i and j are neighbors, the density contains a term $e^{\theta_2 x_i x_j}$ and these variables are not conditionally independent.

If a fixed scan updates all of the variables in one coding set and then all the variables in the other coding set, the order of updating within coding sets does not matter. While updating the red coding set, no update changes any neighbor of a red vertex, since no neighbors are red. Thus when a red vertex is updated it makes no difference how many other red vertices have been updated since neither the Gibbs nor the Metropolis update rule depends on any variables except the one being updated and its neighbors. If we had a computer that could do parallel computations, we could update a whole coding set simultaneously. Thus when scanning by coding sets there are really only two block variables (the two coding sets).

4.1.4 Irreducibility of Ising Model Samplers

Irreducibility is simplest for the Gibbs sampler, because anything is possible. When we update a variable x_i, it can receive either of the two possible values. One of the probabilities may be small, but size of probabilities does not matter when discussing irreducibility, only whether they are zero or nonzero.

A fixed scan Gibbs sampler can go from any state x to any other state y in one scan. It is possible (not very likely but the probability is nonzero) that each i where $x_i \neq y_i$ will be changed and each i where $x_i = y_i$ will be left unchanged. The same logic applies to any scan chosen by a random sequence scan. A random scan cannot go from any x to any y in one step, because each step of the chain only changes one vertex. But if x and y differ at n vertices, then a random scan could choose to update those n vertices in n iterations, each update changing the variable. Again, this is not very likely, but all that matters is whether the probability is nonzero. Thus any Gibbs sampler for an Ising model is irreducible.

The logic here applies to many samplers besides Gibbs samplers for Ising models. We say a Markov chain transition probability satisfies a *positivity condition* if $P(x, y) > 0$ for all x and y, that is if the chain can go from any state to any other in one step. Clearly, positivity implies irreducibility, since it says that $P^n(x, y) > 0$ for the special case $n = 1$. Just as clearly, positivity is not a necessary condition, and the implication that positivity implies irreducibility

is rather trivial. However one often hears that a chain is irreducible "because the positivity condition holds" so one has to know what positivity means in this context.

Metropolis samplers are a bit more complicated. The problem is that positivity does not hold for elementary updates and whether it holds for a scan depends on the scan. When the odds ratio (4.3) is greater than one, the proposal is always accepted, so the variable being updated cannot remain the same. For a random scan, this is no problem. The same argument we used for the Gibbs sampler, says that if x and y differ at n vertices, the random scan could choose to update those n vertices in n iterations, each update changing the variable, thus moving from x to y in n steps.

Suppose we have a symmetric Ising model ($\theta_1 = 0$) and periodic boundary conditions. Suppose the lattice size is even, and consider the state composed of vertical stripes of alternating colors. Each site has two black neighbors and two white neighbors and $\sum_{j \sim i} x_j = 0$. Hence $R = 1$ and and a Metropolis update is always accepted. If we do a scan by coding sets, we will go through a whole coding set and change every vertex in the coding set. This changes the pattern of vertical stripes of alternating colors to horizontal stripes of alternating colors. The state of the system is just a 90° rotation of the original state. Hence the scan through the other coding set does the same thing and changes the pattern back to vertical stripes. The state is not the same as the original; every vertex has changed color. But one more complete scan does take us back to the original state. Although there are 2^d possible states if there are 2^d vertices, the Metropolis sampler using a fixed scan by coding sets only visits two states, if started with alternating stripes. It is not irreducible.

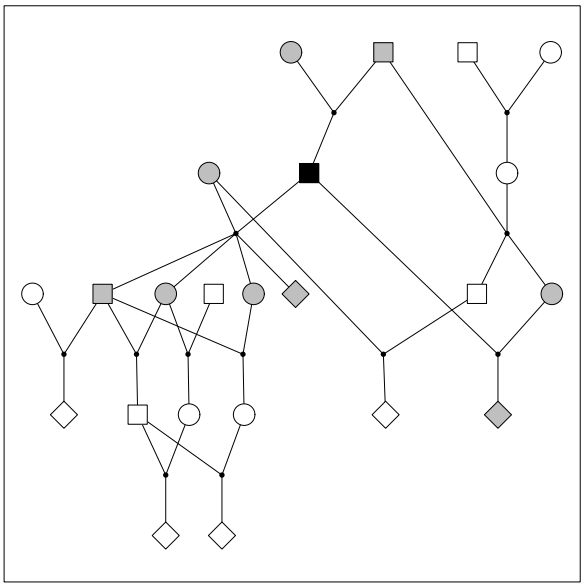
A symmetric Ising model with periodic boundary conditions can also fail to be irreducible when a raster scan is used. For that we need a lattice size that is odd and a checkerboard pattern.

It seems that fixed scan, Metropolis updates, and discrete state spaces do not mix well. If one uses Metropolis updates, perhaps it is best to use a random scan.

4.1.5 Mendelian Genetics

Another stochastic process with a discrete state space is Mendelian genetics. Consider a *pedigree* or *genealogy* of individuals such as that shown in the figure. The large squares, circles, and diamonds represent individuals (male, female, and unspecified, respectively). The small dots represent marriages. From each marriage node lines go up to the parents and down to the children.

Everyone has two copies of genes that are not on sex chromosomes, one copy inherited from their father and one from their mother. These copies are not necessarily identical. A number of variants of a gene called *alleles* are usually found in any large population. A gene passed from a parent to a child is equally likely to be either of the two copies of the gene in that parent, the one inherited from the grandfather or the one from the grandmother. This specifies the probability distribution of all the genes in the pedigree except for



the individuals at the top of the pedigree, called *founders*, whose parents are not recorded. The usual assumption made about the genes of founders is that their genes are randomly drawn from the population gene pool. This requires that the *population allele frequencies* be specified. Then the probability model for genes in the pedigree is completely specified.

The random variables of this probability model are usually taken to be the *genotypes* of the individuals, which say which alleles an individual has, but not which parent they were inherited from. Denote the alleles by $a_1, \ldots, a_m$. Then there are m possible genotypes $a_i a_i$ where both alleles are the same and $m(m-1)/2$ possible genotypes $a_i a_j$ where $i \neq j$. Denote the population allele frequencies by $p_1, \ldots, p_m$. Then the founder genes have a multinomial distribution. The probability of genotype $a_i a_i$ is p_i^2 and the probability of $a_i a_j$ is $2p_i p_j$.

Conditional on parental genotypes, the probability distribution genotypes of children is easy to work out. There are four possible states for the child, each having probability 1/4. These four possible states are not necessarily distinguishable depending on the genotypes of the parents. If both parents have the same genotype $a_1 a_2$, then the child is $a_1 a_1$ or $a_2 a_2$ with probability 1/4 and $a_1 a_2$ with probability 1/2. If one parent is $a_1 a_1$ and the other is $a_2 a_2$, then the child is $a_1 a_2$ with probability one. Other cases can be worked out similarly.

If we denote the probabilities of founders by $p(g)$ and the conditional probabilities of children given parents by $p(g_i | g_{f(i)}, g_{m(i)})$ where $f(i)$ and $m(i)$ are the father and mother of i. Then the probability of a vector of genotypes

$g = (g_1, \ldots, g_m)$ is given by

$$\prod_{\text{children } i} p(g_i | g_{f(i)}, g_{m(i)}) \prod_{\text{founders } i} p(g_i)$$

It is easy to draw independent samples from this distribution. Draw founders first with the specified probabilities. Then draw every child whose parents have already been drawn with the specified probabilities, and repeat this step until everyone has been drawn. A much harder problem is to simulate the conditional distribution of genotypes given observed on some of the individuals in the pedigree.

We often cannot see genotypes. A standard example is a *recessive* genetic disease like cystic fibrosis or phenylketonuria. There are two alleles, conventionally denoted A and a, the normal allele and the disease allele, respectively. The possible genotypes are then AA, Aa, and aa. A recessive disease is one in which one normal gene is enough for normal function, so it is impossible to distinguish the AA and Aa genotypes from the observable characteristics of the individual, which are called the *phenotype*. Individuals with the disease phenotype are known to have genotype aa, but individuals with the normal phenotype can have genotype AA or Aa. Denote these probabilities by $p(\text{data}|g_i)$. Then the joint distribution of phenotypes (data) and genotypes is given by

$$h(g) = \prod_{\text{all individuals } i} p(\text{data}|g_i) \prod_{\text{children } i} p(g_i | g_{f(i)}, g_{m(i)}) \prod_{\text{founders } i} p(g_i) \tag{4.4}$$

The genetics that requires MCMC is to simulate the conditional distribution of genotypes given data. The unnormalized density is given by (4.4). Probability models like this with discrete phenotypes and genotypes are called Mendelian, after Gregor Mendel who formulated the laws of genetics in 1865, to distinguish them from probability models for continuous traits like height and weight, the study of which is called *quantitative genetics*.

A Gibbs sampler for a Mendelian genetics problem is a bit more complicated than one for the Ising model, but not much. The conditional distribution of one individual given the rest only depends on that individuals neighbors in the graph, which are that individuals parents, children, and spouses. In the figure, the neighbors of the individual colored black are colored gray. As always we obtain the conditional for one variable given the rest by keeping only the terms involving that variable.

$$h(g_i | g_{-i}) = p(\text{data}|g_i) p(g_i | g_{f(i)}, g_{m(i)}) \prod_{\substack{\text{children } j \\ \text{of individual } i}} p(g_j | g_{f(j)}, g_{m(j)})$$

if individual i is not a founder and

$$h(g_i | g_{-i}) = p(\text{data}|g_i) p(g_i) \prod_{\substack{\text{children } j \\ \text{of individual } i}} p(g_j | g_{f(j)}, g_{m(j)})$$

if individual i is a founder. A Gibbs update of individual i calculates the unnormalized density $h(g_i|g_{-i})$, normalizes it to add to one when summed over the possible genotypes, and gives g_i a new value from this normalized conditional distribution. If we start in a possible state, one in which all individuals have genes that could have come from their parents, the Gibbs update is well defined and always results in another possible state.

4.1.6 Irreducibility of Mendelian Genetics Samplers

Sheehan and Thomas (1993) give the following proof of the irreducibility of of the Gibbs sampler for a recessive genetic trait. Individuals with the disease phenotype are known to have genotype aa. We can consider them fixed. The Gibbs sampler need only update the individuals with normal phenotype. The positivity condition does not hold. Suppose the sampler uses a fixed scan in which individual i is updated before his parents. Consider going from the genotype in which i and his parents are AA to a genotype in which i is Aa. When i is updated, his parents have not yet been updated, they are still AA which implies that i must also be AA, so he cannot change. After his parents have changed, then he can change, but this takes more than one step of the Markov chain. It would not help if all individuals were updated after their parents. It would still take more than one scan to change from any state to any other, though it is a bit less obvious.

Sheehan and Thomas's proof uses a path from any state to any other that goes through the state in which all individuals with the normal phenotype are Aa. If we start in any possible state, the Gibbs update has two properties (1) any individual can remain unchanged with positive probability and (2) any individual whose parents are both Aa has positive probability of being changed to Aa regardless of the genotypes of any children or spouses. The latter occurs because an Aa individual could have resulted from a marriage of Aa parents and can pass either allele to any child. Thus in one scan all founders can be changed to Aa. In the next scan all children of founders can be changed to Aa. Succeeding scans can change to Aa any individual whose parents have been changed to Aa in a previous scan, while leaving everyone else unchanged. After some number of scans less that the total number of individuals, every individual is Aa. This shows that any possible state can be taken to this special state with positive probability. By reversing the path, the chain can go from the special state to any other possible state.

The Gibbs sampler need not be irreducible for other models. This proof applies only to models having only two alleles. The ABO blood group has three alleles A, B, and O. The gene makes red cell surface antigens, proteins that stick out of the cell membrane of red blood cells and are recognized by the immune system. The A and B alleles make slightly different proteins and the O allele is nonfunctional and makes no protein. There are six genotypes AA, BB, OO, AB, AO, and BO, but only four distinguishable phenotypes AB, A, B, and O, respectively, both A and B antigens on red cells, only A, only B, and neither. Consider now the very simple pedigree with two parents and two children. The

children have blood types AB and O and hence have known genotypes AB and OO. The blood types of the parents are not known, but each must have passed an O allele to the OO child and each must have passed an A or a B to the AB child. Thus the parents are AO and BO, but we don't know which is which. The two possibilities are equally likely.

The Gibbs sampler for this problem is not irreducible. The only two individuals we need to sample are the parents, since the children's genotypes are known. When we update the AO parent, the genotype cannot change. The AB child must get an A allele from some parent, and the other parent, currently BO does not have one. The same goes for the other parent. A Gibbs sampler updating one individual at a time cannot work. A different sampler is required.

4.1.7 General State Spaces

Irreducibility for general state spaces is more complicated in theory but simpler in practice. The theory must deal with the problem that one cannot "get to" any state if the distribution is continuous. Points have probability zero and so are never hit. On the other hand, all real applications of MCMC on general state spaces are irreducible. The practical problems with irreducibility only arise on discrete state spaces.

As always in general state spaces, we talk about probability of hitting sets rather than points. If φ is a nonzero measure on the state space, a Markov chain is called *φ-irreducible* if for any point x and any measurable set A such that $\varphi(A) > 0$ there exists an integer n such that $P^n(x, A) > 0$.

There are equivalent ways to state this condition that use some different kernels. The kernel

$$U(x, A) = \sum_{n=1}^{\infty} P^n(x, A) \tag{4.5}$$

is the expected number of times the chain visits the set A in an infinite run. The chain is φ-irreducible if $U(x, A) > 0$ for all x and all φ-positive sets A. The kernel $L(x, A)$ is defined as the probability that the chain started at x ever hits the set A. A formula for $L(x, A)$ is rather complicated (Meyn and Tweedie 1993, p. 72) and not of immediate interest. What is important is that the chain is φ-irreducible if $L(x, A) > 0$ for all x and all φ-positive sets A.

The reason why an arbitrary measure φ is used in the definition, rather than the invariant distribution π is that the definition is formulated so as to apply to arbitrary Markov chains, including those that do not have an invariant probability distribution. If the chain has an invariant distribution π, then it is π-irreducible if it is φ-irreducible for any φ. So for MCMC where we always construct chains to have a specified invariant distribution π we could always check π-irreducibility, if we so desired, but we do not have to use π if that is inconvenient.

If a chain is φ-irreducible for any φ then there is a *maximal irreducibility measure* ψ having the following properties (Meyn and Tweedie 1993, Proposition 4.4.2)

(i) The chain is ψ-irreducible.

(ii) A measure φ' is an irreducibility measure if and only if it is dominated by ψ, that is, $\psi(A) = 0$ implies $\varphi'(A) = 0$.

(iii) If $\psi(A) = 0$ then $B = \{ x : L(x, A) > 0 \}$ also has ψ-measure zero.

The point of the irreducibility measure φ is to define a class of null sets which the chain does not need to hit. The maximal irreducibility measure ψ is the irreducibility measure having the smallest class of null sets. The measure itself is not unique, but the class of null sets of the maximal irreducibility measure is unique. If the chain has an invariant distribution π and is φ-irreducible, then the chain is recurrent (Meyn and Tweedie 1993, Proposition 10.1.1), the invariant distribution is unique (Proposition 10.4.4), and the invariant distribution is a maximal irreducibility measure (Proposition 10.4.9). Any other maximal irreducibility measure ψ has the same null sets, $\psi(A) = 0 \Leftrightarrow \pi(A) = 0$. We can always use π as the irreducibility measure, but there will be fewer sets to check if we use another measure φ dominated by π, and this may be more convenient.

Before continuing with general state spaces, let us stop and compare with the definition for countable state spaces. The definition for countable state spaces is essentially π-irreducibility in the case where every point has positive π-probability. All points of π-probability zero must be excluded from the state space, since if $\pi(\{y\}) = 0$, then by (iii) above, the set $B = \{ x : L(x, y) > 0 \}$ satisfies $\pi(B) = 0$. But by the definition of irreducibility for countable spaces B is the whole state space, which is impossible. Hence we must have $\pi(\{y\}) > 0$ for all y.

If we apply φ-irreducibility to countable state spaces, can use a measure φ concentrated at a single point y. Thus it is enough to show that that the chain can go from any point x to one single point y. It is not necessary to show that the chain can get to any other point, that follows from (iii) above. In the Mendelian genetics example, it was enough to show that the sampler could get from any state to the special state in which every individual with normal phenotype has genotype Aa. The proof could have stopped there.

4.1.8 Verifying ψ-Irreducibility

For most problems on continuous state spaces ψ-irreducibility is easy to verify. First consider a sampler that satisfies a very simple positivity condition, a Metropolis sampler that updates all variables at once with a proposal density $q(x, \cdot)$ and invariant density $h(x)$ that are everywhere positive. Then

$$P(x, A) \geq \int_A q(x, y) a(x, y) \mu(dy)$$

so if $\mu(A) > 0$ then $P(x, A) > 0$ because the integrand is strictly positive. Hence the chain is μ-irreducible.

Next consider a sampler that updates one variable at a time, but still has everywhere positive proposals and acceptance probabilities. If there are d variables

we prove irreducibility by induction on d. The induction hypothesis assumes that starting at $x = (x_1, \ldots, x_d)$ updating x_1, …, x_{d-1} has positive probability of hitting any set B of positive Lebesgue measure in $\mathbb{R}^{d-1}$. Write $Q_1(x, B)$ for this probability. The base of the induction, the case $d = 1$, was proved in the preceding paragraph. For any set A of nonzero Lebesgue measure in $\mathbb{R}^d$ and for any $x \in \mathbb{R}^d$ write $x = (x_{-d}, x_d)$ and

$$A_{x_{-d}} = \{\, x_d \in \mathbb{R} : (x_{-d}, x_d) \in A \,\}$$

for the "sections" of A, the possible values of x_d when the other x_{-d} is held fixed. It is a standard fact of measure theory that the sections are measurable sets and if A has positive measure then so does $A_{x_{-d}}$ for x_{-d} in a set of positive Lebesgue measure. Write $Q_2(x_{-d}, C)$ for the probability that $x_d \in C$ given x_{-d}. Then the preceding sentence says $Q_2(x_{-d}, A_{x_{-d}}) > 0$ for x_{-d} in a set of positive Lebesgue measure. Since

$$P(x, A) = \int Q_1(x, dx_{-d}) Q_2(x_{-d}, A_{x_{-d}})$$

is the integral of a function $Q_2(x_{-d}, A_{x_{-d}})$ that is not zero almost everywhere with respect to a measure $Q_1(x, \cdot\,)$, which is nonzero by the induction hypothesis, we have $P(x, A) > 0$. That proves φ-irreducibility where here φ is Lebesgue measure on $\mathbb{R}^d$.

Those unfamiliar with measure theory should take my word for it that these calculations involve only the elementary bits of measure theory that justify replacing integrals with respect to area or volume by iterated univariate integrals. They are only mystifying to the uninitiated.

These calculations have the drawback that they require positivity, something which we do not want to have to satisfy in general. For example, the first MCMC simulation ever (Metropolis, Rosenbluth, Rosenbluth, Teller, and Teller 1953) used the Metropolis algorithm for a point process with a fixed number of points and the proposal was to move the point to a position uniformly distributed in a ball around the current position. We would like to be able to show that simulation to be irreducible as well.

Theorem 4.1. *Suppose*

(a) *The state space of the chain is a second countable topological space.*

(b) *The state space is topologically connected.*

(c) *Every nonempty open set is φ-positive.*

(d) *Every point has a φ-communicating neighborhood.*

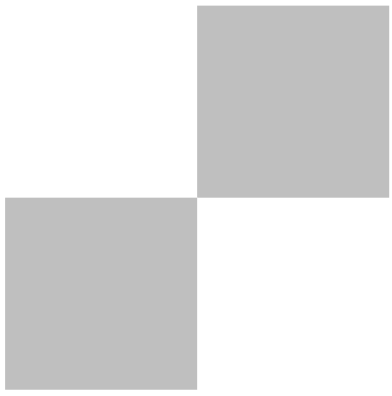
Then the chain is φ-irreducible. If all of the conditions hold except (b)*, then every connected component is φ-communicating.*

Some of these terms need explanation. A topological space is second countable if there is a countable family of open sets $\mathcal{U}$ such that every open set is a union of sets in $\mathcal{U}$. Every separable metric space, in particular any subset of a Euclidean space $\mathbb{R}^d$, has this property. A topological space is connected if it is not the union of disjoint open sets. A set B is φ-communicating if for every φ-positive subset C of B and every point x in B, there is an n such that $P^n(x, C) > 0$. This is the same as the definition of φ-irreducibility, except that it is applied to a subset rather than the whole space.

Before proving the theorem, let us see how it works. Consider a Metropolis sampler for the uniform distribution on any connected open set S in R^d that makes a proposal that is uniform in the ball $B(x, \varepsilon)$ of radius ε centered at the current point x. Because the uniform density is constant, the odds ratio is always zero or one. Every proposal that falls in S is accepted, and every proposal that falls outside is rejected. Checking the conditions of the theorem, (a) holds because the state space is a subset of $\mathbb{R}^d$, (b) holds by assumption, (c) holds if we take S to be the state space, and (d) holds by a variation of the argument using the positivity condition. For any point $x \in S$ there is a ball $B(x, \delta)$ contained in S, with $0 < \delta < \varepsilon/2$. Then for any $y \in B(x, \delta)$ we have $B(x, \delta) \subset B(y, \varepsilon)$. So for any y in $B(x, \delta)$ and any φ-positive $C \subset B(x, \delta)$, we also have $C \subset B(y, \varepsilon)$, so the proposal hits C with positive probability. This says that $B(x, \delta)$ is a φ-communicating neighborhood of x. Thus the theorem says this sampler is irreducible.

If the state space is not connected, then φ-irreducibility may not hold. Suppose the state space consists of two open sets S_1 and S_2 separated by a distance greater than ε. Then the sampler just described is not irreducible. It can never move from S_1 to S_2 or vice versa.

The interaction of conditions (b) and (d) is delicate. Consider a Gibbs sampler for the uniform distribution for the open set in $\mathbb{R}^2$ shown in the figure. The coordinate axes are horizontal and vertical. The update of the first variable



moves to a position uniform on the intersection of the horizontal line through the current point with the gray region, and similarly for the update of the second variable except the line is vertical. Neither update can ever move from one

square to the other and the chain is not irreducible. If the state space is taken to be the open set that is the gray region in the figure, it is not connected. So condition (b) doesn't hold, since the squares are disjoint and open. We can make the space connected by adding the point where the squares touch, but then condition (d) doesn't hold, since this new point does not have a φ-communicating neighborhood. Every neighborhood intersects both squares and the chain never moves from one square to another.

Proof. If A and B are any φ-communicating sets such that $\varphi(A \cap B) > 0$, then $A \cup B$ is φ-communicating. The reason is that for any $x \in A$, the chain must eventually hit $A \cap B$, and from there it must hit any φ-positive $C \subset B$. Formally

$$U(x, C) \geq \int_{A \cap B} P^m(x, dy) U(y, C),$$

where $U(x, A)$ is defined by (4.5). For some m, $P^m(x, A) > 0$, because A is φ-communicating, and $U(y, C) > 0$ because B is φ-communicating. By symmetry, the same holds if $x \in B$ and $C \subset A$. Hence $A \cup B$ is φ-communicating.

Now choose for each point $x \in S$ a φ-communicating neighborhood W_x that is an element of $\mathcal{U}$. This is possible because every neighborhood of x contains another neighborhood of x that is an element of $\mathcal{U}$ and subsets of φ-communicating sets are φ-communicating. Let $\mathcal{W} = \bigcup_{x \in S} W_x$. Then $\mathcal{W}$ is countable because $\mathcal{U}$ is countable.

Consider two sequences of sets $\{V_k\}$ and $\{D_k\}$ defined recursively as follows. First, V_1 is an arbitrary element of $\mathcal{W}$. Then, assuming $V_1, \ldots, V_{k+1}$ have been defined, we define

$$D_k = \bigcup_{i=1}^{k} V_i$$

and let V_{k+1} be any element of $\mathcal{W}$ satisfying

$$V_{k+1} \cap D_k \neq \varnothing$$

and

$$V_{k+1} \not\subset D_k$$

If no element of $\mathcal{W}$ satisfies the condition, let $V_{k+1} = \varnothing$.

By induction D_{k+1} is φ-communicating for each k, because the intersection of V_{k+1} and D_k is nonempty and open and hence φ-positive by (c). Hence the argument above shows their union is φ-communicating.

Let $D = \bigcup_{k=1}^{\infty} D_k$. Then D is φ-communicating, because any $x \in D$ and φ-positive $A \subset D$ there is a k such that $x \in D_k$ and $\varphi(A \cap D_k) > 0$. Hence it is possible to get from x to A because D_k is φ-communicating.

Now there are two logical possibilities. $D = S$ in which case the chain is φ-irreducible or D and $S \setminus D$ are disjoint open sets and (b) is violated. Then D is a φ-communicating connected component and the same construction shows that each connected component is φ-communicating. □

If this theorem can't be used to prove ψ-irreducibility, then we are really in the discrete case in disguise. Consider Gibbs samplers for the uniform distributions on the regions on each side of the figure. The one on the left is irreducible

the one on the right is not. The theorem doesn't apply to either one, because neither has a connected state space. The theorem says that each of the squares is φ-communicating, but topology is no help with the question of whether the chain can move from one square to another. No general argument is likely to help. As in with discrete state spaces, a special argument is needed for each problem.

4.1.9 Harris recurrence

If a chain is ψ-irreducible and has an invariant distribution π then there exists a set N with $\pi(N) = 0$ such that $L(x, A) = 1$ for all $x \notin N$ and all ψ-positive A and $P(x, N) = 0$ for all $x \notin N$ (Meyn and Tweedie 1993, Proposition 9.0.1). Note that the definition of ψ-irreducibility only requires $L(x, A) > 0$, but requires it for all x. Something even stronger is true, not only is any ψ-positive set A hit with probability one, it is hit infinitely often with probability one (Meyn and Tweedie 1993, Proposition 9.1.1) when started at any $x \notin N$. This null set N of starting points from which bad things happen is a nuisance. The point of Harris recurrence is to eliminate it. A ψ-irreducible chain is *Harris recurrent* if $L(x, A) = 1$ for all x and all ψ-positive A. Any ψ-irreducible chain can be made into a Harris chain by removing the null set N from the state space. This does no harm since the chain can never hit N from outside N.

Harris recurrence essentially banishes measure theoretic pathology. It would be very strange if a Markov chain that is an idealization of a computer simulation would be ψ-irreducible but not Harris recurrent. If null sets matter when the computer's real numbers are replaced by those of real analysis, then the simulation cannot be well described by the theory.

Note that any irreducible chain on a countable state space is always Harris recurrent. Irreducibility requires that we eliminate from the state space all points of π-measure zero. That having been done, the only remaining π-null set is empty, and irreducibility trivially implies Harris recurrence. The difference between ψ-irreducibility and Harris recurrence is only an issue in general state spaces.

Fortunately, an irreducible Gibbs or Metropolis sampler is always Harris recurrent under very weak conditions. Tierney (1994) gives the following two simple propositions. If a Gibbs sampler is ψ-irreducible and $P(x, \cdot)$ is absolutely continuous with respect to π, then it is Harris recurrent (Corollary 1). A ψ-irreducible chain that iterates one Metropolis-Hastings elementary update is always Harris recurrent (Corollary 2). The condition on the Gibbs sampler merely says that the chain cannot hit π-null sets. $\pi(A) = 0$ implies $P(x, A) = 0$.

The situation is only a bit more complicated for Metropolis-Hastings samplers that update one variable at a time. Chan and Geyer (1994) give the following (Theorem 1). Suppose the invariant distribution π has an unnormalized density $h(x)$ with respect to Lebesgue measure on $\mathbb{R}^d$, each proposal distribution has a density with respect to Lebesgue measure on $\mathbb{R}$, and all of the unnormalized conditional densities make sense, that is, $h(x)$ considered as a function of some of the variables, the rest held fixed, is (1) not identically zero and (2) integrable with respect to Lebesgue measure on the subspace spanned by those variables. If the Metropolis-Hastings sampler for each conditional distribution obtained by updating only a subset of variables is ψ-irreducible, then Metropolis-Hastings sampler for the unconditional distribution is Harris recurrent. This sounds complicated, but the conditions are necessary. Assuming each elementary update is "nice" with no measure theoretic pathology, the only way a variable-at-a-time Metropolis-Hastings sampler can fail to be Harris recurrent is if for some starting position x some variable x_i has a positive probability of never being updated in an infinite run of the chain. This cannot happen if the chain that starts at x and keeps x_i fixed is ψ-irreducible, and we need to verify this for each starting position x and every subset of variables held fixed.

No theorem has been found that establishes Harris recurrence for general Metropolis-Hastings-Green samplers, but there is a general method involving a "drift condition" that can be used for any Markov chain. This method will be explained in Section 4.7.5.

4.2 The Law of Large Numbers

We now return to the law of large numbers mentioned in Section 1.6.1 and give a precise statement. Suppose we have a Markov chain with invariant distribution π and g is a π-integrable function so the integral

$$\mu = E_\pi g(X) = \int g(x)\pi(dx)$$

exists. Let

$$\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} g(X_i)$$

denote the sample average of $g(X)$ over a run of the Markov chain. We then have the following two results.

Theorem 4.2. *For a φ-irreducible chain with invariant distribution π, conditional on the starting point x, the sample mean $\hat{\mu}_n$ converges almost surely to μ, for π-almost all x.*

When φ-irreducibility is strengthened to Harris recurrence, the bad null set of starting points for which convergence fails disappears.

Theorem 4.3. *For a Harris recurrent chain with invariant distribution π, the sample mean $\hat{\mu}_n$ converges almost surely to μ regardless of the initial distribution of the chain.*

The latter follows from Theorems 17.0.1 and 17.1.6 in Meyn and Tweedie (1993). The former follows from Birkhoff's ergodic theorem (Breiman 1968, Theorem 6.21) together with the condition for a Markov chain to be ergodic given in Theorem 7.16 in Breiman (1968), which uses the criterion of indecomposability, which in turn is implied by π-irreducibility (Nummelin 1984, Proposition 2.3).

Again ψ-irreducibility leaves us with a bad null set of starting points for which convergence fails. From now on we shall always require the stronger Harris property and no longer need to mention these null sets.

In the presence of Harris recurrence the law of large numbers says exactly the same thing for Markov chains as it does for independent sampling. If the function $g(X)$ is integrable, then the strong law of large numbers holds. There is almost sure convergence of the sample mean to its expected value with respect to the invariant distribution.

4.3 Convergence of the Empirical Measure

The *empirical measure* for a sample $X_1, \ldots, X_n$ is the probability measure

$$\pi_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}$$

that puts mass $1/n$ at each of the sample points, where, as always, $\delta_x = I(x, \cdot)$ denotes the "Dirac measure" concentrated at x. Since it depends on the sample, π_n is a random probability measure. Probabilities and expectations are calculated just as with any other probability measure

$$\pi_n(B) = \int_B \pi_n(dx) = \frac{1}{n} \sum_{i=1}^{n} 1_B(X_i)$$

and

$$E_{\pi_n} g(X) = \int g(x) \pi_n(dx) = \frac{1}{n} \sum_{i=1}^{n} g(X_i) \tag{4.6}$$

Thus we see that the "empirical expectation" (4.6) is just a fancy way of rewriting a familiar concept, the sample average of a functional $g(X)$ of the Markov chain.

Now we want to consider what it means to say the empirical measure π_n converges in distribution to π. By the "portmanteau theorem" (Fristedt and Gray 1997, Theorem 6 of Chapter 18) there are several equivalent ways of saying this, including

$$\int g(x)\pi_n(dx) \to \int g(x)\pi(dx) \tag{4.7a}$$

holds for every bounded continuous function g and

$$\liminf_{n\to\infty} \pi_n(O) \geq \pi(O) \tag{4.7b}$$

holds for every open set O. Now we want to prove a theorem that says π_n converges in distribution to π almost surely. Because there are two types of convergence involved, this is confusing. More precisely, the statement is

$$\Pr\left(\pi_n \xrightarrow{\mathcal{D}} \pi\right) = 1$$

or for almost all sample paths of the Markov chain $\pi_n \xrightarrow{\mathcal{D}} \pi$.

Note that the law of large numbers implies (4.7a) for just one function g or (4.7b) for just one open set O. The issue is whether there is simultaneous convergence for all bounded continous functions in (4.7a) and open sets in (4.7b).

Theorem 4.4. *Suppose the state space of the Markov chain is a separable metric space and the chain is Harris recurrent, then π_n converges in distribution to π with probability one.*

Let $\mathcal{B}$ denote the countable family of sets consisting of open balls with centers at the points of some countable dense set and rational radii and all finite intersections of such balls. Then, for almost all sample paths of the Markov chain,

$$\pi_n(B) = \frac{1}{n}\sum_{i=1}^{n} 1_B(X_i) \to \pi(B), \qquad \text{for all } B \in \mathcal{B} \tag{4.8}$$

By Corollary 1 of Theorem 2.2 in Billingsley (1968), (4.8) implies π_n converges in distribution to π. A similar result under different regularity conditions is proved by Meyn and Tweedie (1993, Theorem 18.5.1).

This theorem is not very deep, being a straightforward consequence of the law of large numbers, but gives us an important way to think about MCMC. An n-sample X_1, ..., X_n obtained from a single run of the Markov chain approximates the invariant distribution π in the sense described by the theorem. The empirical distribution for this cloud of points gets closer and closer to π as n goes to infinity.

If X_1, X_2, ... are a Markov chain with invariant distribution π. Then we often call X_1, ..., X_n an MCMC sample from π. This bothers many people because they are so used to the notion of i. i. d. samples that thinking about any other kind makes their head hurt. It is true X_1, X_2, ... are not independent. Nor are they identically distributed unless the initial distribution is π, and it

never is because if we knew how to produce even one realization from π we wouldn't be using MCMC. So if they aren't independent, and aren't identically distributed, and none of them have the distribution π, how dare we call them samples from π? The theorem says. Just like the i. i. d. case we have $\pi_n \xrightarrow{\mathcal{D}} \pi$ almost surely. That's what's important.

4.4 Aperiodicity

A very different sort of convergence involves the marginal distribution of X_n. It is usually true that $\mathcal{L}(X_n) \to \pi$ (read "the law of X_n converges to π"). Such statements are not important in themselves for MCMC. Since MCMC estimates are sample averages, the important kinds of convergence are the LLN and the CLT. Convergence of marginals is a side issue.

But it is an important side issue for a number of reasons. First a large part of Markov chain theory involves questions about convergence of marginals, and much of this has been imported into the MCMC literature and colors discussions despite its questionable relevance. Second, Markov chain theory about convergence of marginals is intimately connected with theory about the CLT. The easiest way to prove the CLT holds is to show "geometric ergodicity," which is a form of convergence of marginals. Hence what seems like a detour is actually taking us toward our goal.

The law of large numbers can hold for a Markov chain even though marginal distributions do not converge. The simplest example is the deterministic Markov chain on a two-point state space that alternates between the points. Call the points 0 and 1 then

$$X_n = n \mod 2$$

if we start at $X_1 = 1$ and

$$X_n = (n + 1) \mod 2$$

if we start at $X_1 = 0$. The chain is clearly irreducible since it can go from 0 to 1 in one step and from 1 to 1 in two steps. The invariant distribution puts probability 1/2 at each point by symmetry, or we can check $\pi P = P$ directly, which written out in matrix notation is

$$(\tfrac{1}{2}, \tfrac{1}{2}) \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = (\tfrac{1}{2}, \tfrac{1}{2})$$

Hence the law of large numbers applies, as can also be checked by direct calculation. But the marginal distribution of X_n does not converge to π. It is always concentrated at one point, either 0 or 1 depending on whether n is odd or even and what the starting point was.

It is worth pointing out that this is a Metropolis sampler where the proposal is to go to the other point. The proposal is always accepted because the odds ratio is always one.

This example illustrates a general phenomenon. The state space of any ψ-irreducible Markov chain can partitioned into sets D_0, D_1, ..., D_{d-1} and N such that

(i) $P(x, D_i) = 1$, when $x \in D_j$ and $j = i - 1 \mod d$.

(ii) $\psi(N) = 0$.

This partition is unique up to null sets if d is chosen as small as possible (Meyn and Tweedie 1993, Theorem 5.4.4). The chain is said to be *aperiodic* if $d = 1$ and *periodic* if $d > 1$. In the periodic case the marginals cannot converge, since if we start with X_1 in D_1 then we have $\Pr(X_n \in D_i) = 1$ for $i = n \mod d$. Since the distributions of X_m and X_n have disjoint supports for $m \neq m \mod d$, convergence is impossible.

Fortunately we have the following theorems.

Theorem 4.5. *Any ψ-irreducible sampler that has $P(x, \{x\}) > 0$ for $x \in A$ where $\psi(A) > 0$ is aperiodic.*

Proof. Assume to get a contradiction that the sampler is periodic. Then we must have $\psi(A \cap D_i) > 0$ for one of the D_i in the cyclic decomposition of the state space. But then for $x \in A \cap D_i$ we have $P(x, D_i) \geq P(x, \{x\}) > 0$. But the cyclic decomposition requires $P(x, D_i) = 0$ for $x \in D_i$. The contradiction proves the sampler must be aperiodic. □

The theorem wouldn't be true without any conditions on the sampler, since our deterministic two-point sampler is Metropolis and not aperiodic.

Theorem 4.6. *Any ψ-irreducible Gibbs sampler is aperiodic.*

Proof. The argument is taken from Liu, Wong, and Kong (1995, Lemma 3.2). It uses the point of view that the transition probabilities define an operator on $L^2(\pi)$. When working with nonreversible samplers, we need $L^2(\pi)$ to be a complex Hilbert space. A complex function u is an eigenvector of the transition operator P associated with the eigenvalue λ if $Pu = \lambda u$. A periodic chain always has an eigenvector u associated with the eigenvalue $\omega = e^{2\pi i/d}$, the d-th root of unity, given by

$$u(x) = \sum_{k=0}^{d-1} \omega^k 1_{D_k}(x) \tag{4.9}$$

since

$$(Pu)(x) = \sum_{k=0}^{d-1} \omega^k P(x, D_k) = \sum_{k=0}^{d-1} \omega^k 1_{D_{k-1 \mod d}}(x) = \sum_{k=0}^{d-1} \omega^{k+1} 1_{D_k}(x) = \omega u(x)$$

For a fixed scan Gibbs sampler, the transition operator is a product of operators for elementary updates $P = P_1 \cdots P_d$. The P_i for a Gibbs sampler have the special property of being projections, that is they are self-adjoint and idempotent. We have shown that Gibbs updates are reversible and that this is equivalent to

the operator being self-adjoint. Idempotent means $P_i^2 = P_i$, something we have also noted: repeating a Gibbs elementary update twice is the same as doing it once. Thus by the analog of the Pythagorean theorem for Hilbert spaces

$$\|u\|^2 = \|P_i u\|^2 + \|(I - P_i)u\|^2$$

holds for any function $u \in L^2(\pi)$. Hence either $\|P_i u\| < \|u\|$ or $\|(I - P_i)u\| = 0$. The latter implies that $P_i u = u$ so u is an eigenvector associated with the eigenvalue 1. If the latter is true for all i, then $Pu = u$, which is false for the particular u given by (4.9). Hence we must have $\|P_i u\| < \|u\|$ for at least one i, say an i such that $P_j u = u$ for $j > i$. But then

$$\|Pu\| \le \|P_1\| \cdots \|P_{i-1}\| \cdot \|P_i u\| < 1$$

since $\|P_i\| < 1$ for all i. But this contradicts

$$\|Pu\| = \|\omega u\| = |\omega| \, \|u\| = 1$$

So a fixed scan Gibbs sampler cannot be periodic.

Neither can a random scan or a random sequence scan sampler be periodic, by slight variants of the same argument. □

4.5 The Total Variation Norm

A bounded signed measure is a real-valued countably additive set function defined on a σ-field. Any signed measure μ has a decomposition $\mu = \mu^+ - \mu^-$ as the difference of two positive measures with disjoint supports. The total variation norm of μ is

$$\|\mu\| = \mu^+(\mathcal{X}) + \mu^-(\mathcal{X})$$

where $\mathcal{X}$ is the whole space. An equivalent definition is

$$\|\mu\| = \sup_{|f| \le 1} \int f \, d\mu. \tag{4.10}$$

where the supremum is taken over all measurable functions f such that $|f(x)| \le 1$ for all x.

The total variation norm gives bounds for the measure of sets

$$\sup_A |\mu(A)| \le \|\mu\| \le 2 \sup_A |\mu(A)|$$

where the sup runs over all measurable sets.

4.6 Convergence of Marginals

Theorem 4.7. *For an aperiodic Harris recurrent chain with invariant distribution π and any initial distribution λ*

$$\|\lambda P^n - \pi\| = \left\| \int \lambda(dx) P^n(x, \cdot) - \pi \right\| \to 0, \qquad \text{as } n \to \infty \tag{4.11}$$

Moreover, the left hand side is nonincreasing in n.

This is Theorem 13.3.3 and 13.3.2 in Meyn and Tweedie (1993).

If X_0 has the distribution λ, then λP^n is the marginal distribution of X_n. The theorem says this marginal distribution converges to π in total variation. A trivial corollary is that this marginal converges in distribution to π, since convergence in total variation implies convergence in distribution.

In the special case where λ is the measure concentrated at the point x, (4.11) reduces to

$$\|P^n(x, \cdot) - \pi\| \to 0, \qquad \text{as } n \to \infty \tag{4.12}$$

4.7 Geometric and Uniform Ergodicity

4.7.1 Geometric Ergodicity

A Markov chain is said to be *geometrically ergodic* when the convergence in (4.12) occurs at a geometric rate, that is when there is a constant $\rho < 1$ and a nonnegative function $M(x)$ such that

$$\|P^n(x, \cdot) - \pi\| \le M(x)\rho^n \qquad \text{for all } n. \tag{4.13}$$

When this happens, something a bit stronger is actually true, and Meyn and Tweedie (1993) take this as the definition. A Harris recurrent Markov chain with invariant distribution π is *geometrically ergodic* if there exists a constant $r > 1$ such that

$$\sum_{n=1}^{\infty} r^n \|P^n(x, \cdot) - \pi\| < \infty, \qquad \text{for all } x. \tag{4.14}$$

Note that for this series to be summable, each term must go to zero, which implies (4.13) holds with $\rho = 1/r$.

The total variation convergence in (4.13) implies that

$$|P^n(x, C) - \pi(C)| \le M(x)\rho^n$$

holds for any set C. In fact, something stronger is true, but we need some preliminary definitions before we can state it.

4.7.2 Small and Petite Sets

A set C is *small* if there is an integer m, a real number $\delta > 0$, and a probability measure Q on the state space such that

$$P^m(x, A) \ge \delta Q(A), \qquad x \in C \text{ and } A \text{ a measurable set.} \tag{4.15}$$

If $Q(C) = 1$, this is referred to as a "minorization condition" for for the m-step transition kernel P^m. It is a deep theorem of Jain and Jamison (1967) that any ψ-irreducible chain has ψ-positive small sets.

Small sets are not a convenient notion if the chain is periodic, since any small set must be contained in one of the D_i in the partition defining the periodic behavior. So Meyn and Tweedie (1993) define a closely related concept of "petite set." If $a(n)$, $n = 0, 1, \ldots$ defines a probability distribution on the nonnegative integers, then

$$K_a(x, A) = \sum_{n=0}^{\infty} a(n) P^n(x, A) \tag{4.16}$$

is the kernel of the Markov chain having the following update mechanism: generate a random integer N with distribution a, run the original chain N steps. This gives a random subsample of the original chain. The sample is "with replacement" if $a(0) > 0$ so that $N = 0$ is possible. A set C is *petite* if there is a sampling distribution a, a $\delta > 0$, and a probability measure Q on the state space such that

$$K_a(x, A) \geq \delta Q(A), \qquad x \in C \text{ and } A \text{ a measurable set.} \tag{4.17}$$

Every small set is petite (use the sampling distribution concentrated at m) and if the chain is aperiodic and irreducible every petite set is small (Meyn and Tweedie 1993, Theorem 5.5.7). The only difference between the concepts is when the chain is periodic. In MCMC we have little interest in periodic chains, but it does no harm to use the more general term, following Meyn and Tweedie.

Petite sets can be rather large. For any ψ-irreducible chain, there is an increasing sequence $C_1 \subset C_2 \subset \cdots$ of petite sets that covers the state space. So $\pi(C_i)$ increases to 1 as $i \to \infty$.

4.7.3 Feller chains and T-chains

A Markov chain on a topological state space is called a Feller chain if $P(\,\cdot\,, O)$ is a lower semicontinuous function for every open set O. The requirement that the kernel P be lower semicontinuous can be expressed as

$$\liminf_n P(x_n, O) \geq P(x, O), \qquad \text{whenever } x_n \to x.$$

Meyn and Tweedie (1993) call a Markov chain a "T-chain" if the following conditions hold

(i) There exists a sampling distribution a and a kernel $T(x, A)$ such that $T(\,\cdot\,, A)$ is a lower semicontinuous function for any measurable set A.

(ii) For each x, the measure $T(x, \,\cdot\,)$ is nonzero.

The point of the concept is the following (Meyn and Tweedie 1993, Theorem 6.0.1) if every compact set is petite then the chain is a T-chain and conversely if the chain is a T-chain then every compact set is petite. So if we can verify that a chain is a T-chain, we immediately have a wealth of petite sets.

Verifying that a chain is a T-chain usually a simple application of Fatou's lemma. Consider a Gibbs sampler. Say x is the current state and y is the

state after one fixed scan, and suppose that all of the elementary updates have densities, then the density of y given x has the form

$$p_3(y_3|y_2, y_1)p_2(y_2|x_3, y_1)p_1(y_1|x_3, x_2)$$

when there are three variables, and similarly for other numbers of variables. Suppose for each fixed value of y the integrand is a lower semicontinuous function of x, which in this case happens when $x_3 \mapsto p_2(y_2|x_3, y_1)$ is lower semicontinuous and $(x_3, x_2) \mapsto p_1(y_1|x_3, x_2)$ is lower semicontinuous. Then by Fatou's lemma

$$\begin{aligned}
&\liminf_n P(x_n, A) \\
&= \liminf_n \iiint_A p_3(y_3|y_2, y_1)p_2(y_2|x_{n,3}, y_1)p_1(y_1|x_{n,3}, x_{n,2})\, dy_1\, dy_2\, dy_3 \\
&\geq \iiint_A \liminf_n \bigl[p_3(y_3|y_2, y_1)p_2(y_2|x_{n,3}, y_1)p_1(y_1|x_{n,3}, x_{n,2})\bigr]\, dy_1\, dy_2\, dy_3 \\
&= \iiint_A p_3(y_3|y_2, y_1)p_2(y_2|x_3, y_1)p_1(y_1|x_3, x_2)\, dy_1\, dy_2\, dy_3 \\
&= P(x, A)
\end{aligned}$$

So the kernel itself is lower semicontinuous, and the chain is actually Feller as well as being a T-chain.

Now consider Metropolis-Hastings algorithm, this time with only two variables to keep the equations shorter. Here we throw away the rejection part of the kernel, since it need not be lower semicontinuous. Let $T(x, A)$ be the probability that the chain moves from x to A and every proposal in the scan is accepted. Then $P(x, A) \geq T(x, A)$ and

$$\begin{aligned}
\liminf_n T(x_n, A) &\geq \liminf_n \iint_A p_2(y_2|x_{n,2}, y_1)p_1(y_1|x_{n,2}, x_{n,1})\, dy_1\, dy_2 \\
&\geq \iint_A \liminf_n \bigl[p_2(y_2|x_{n,2}, y_1)p_1(y_1|x_{n,2}, x_{n,1})\bigr]\, dy_1\, dy_2 \\
&= \iint_A p_2(y_2|x_2, y_1)p_1(y_1|x_2, x_1)\, dy_1\, dy_2 \\
&= T(x, A)
\end{aligned}$$

and $T(x, A)$ is lower semicontinuous if the p_i are lower semicontinuous functions of their x arguments, just as with the Gibbs sampler. Now the p_i have the Metropolis form (3.2.5). These will be lower semicontinuous if both the proposal and acceptance densities are lower semicontinuous functions of their x arguments. Since x appears in both the numerator and denominator of the Hastings ratio, the only simple condition that assures this is that unnormalized density $h(x)$ is actually a continuous function of x and that the proposal density $q(x, y)$ is separately continuous in x and y. We also have to verify part (ii) of the definition of T-chain, which held trivially for the Gibbs sampler. $T(x, \cdot\,)$ will be a positive measure for each x if every possible elementary update has positive probability of being accepted.

Verifying that a Metropolis-Hastings-Green sampler is a T-chain is more difficult. The fact that the proposals are discontinuous with respect to Lebesgue measure means that we have to consider more than a single elementary update step. That was also the case with Gibbs and Metropolis, but what constitutes a "full scan" in a Metropolis-Hastings-Green sampler is unclear.

4.7.4 Absorbing and Full Sets

A set S is said to be *absorbing* if $P(x, S) = 1$ for all $x \in S$. A set S is said to be full if $\psi(S^c) = 0$, where ψ is a maximal irreducibility measure. When the chain has an invariant distribution π, a set S is full if $\pi(S) = 1$. Every absorbing set is full if the chain is ψ-irreducible (Meyn and Tweedie 1993, Proposition 4.2.3).

If the chain is started in an absorbing set S it never leaves. Thus it makes sense to talk about the chain restricted to S. Restriction to an absorbing set does not change the kernel except to restrict the domain.

If the chain is ψ-irreducible and started outside of S, the law of large numbers says that almost all sample paths hit S and never leave. Moreover since $\pi(S) = 1$, the part of the state space outside S is uninteresting from the standpoint of Markov chain Monte Carlo. We don't want any samples from a set of π-measure zero.

4.7.5 Drift Conditions

How do we verify geometric ergodicity? The basic tool is a so-called "drift condition." We say a Markov chain satisfies the *geometric drift condition* if there exists a measurable function $V(x) \geq 1$, possibly taking the value $+\infty$ but finite at some x, a petite set C, and constants $\lambda < 1$ and $b < \infty$ such that

$$PV(x) \leq \lambda V(x) + b1_C(x), \qquad \text{for all } x \tag{4.18}$$

where

$$PV(x) = \int P(x, dy)V(y) = E[V(X_t)|X_{t-1} = x].$$

If $V(x) = \infty$ the drift condition is satisfied vacuously for that x.

A weaker drift condition is useful in establishing Harris recurrence. A Markov chain satisfies the *positive drift condition* if there exists a measurable function $V(x) \geq 1$, possibly taking the value $+\infty$ but finite at some x, a petite set C, and a constant $b < \infty$ such that

$$PV(x) \leq V(x) - 1 + b1_C(x), \qquad \text{for all } x \tag{4.19}$$

If the chain is ψ-irreducible, any solution $V(x)$ of the geometric drift condition satisfies

(i) The set $S = \{\, x : V(x) < \infty \,\}$ is absorbing and full.

(ii) V is unbounded off petite sets.

(iii) $\int V \, d\pi < \infty$.

by Lemma 15.2.2 and Theorem 14.3.7 in Meyn and Tweedie (1993), and any solution $V(x)$ of the positive drift condition satisfies (i) and (ii) by Lemmas 11.3.6 and 11.3.7 in Meyn and Tweedie.

Condition (ii) means that every sublevel set $\{ x : V(x) \le r \}$ is petite, for any $r \in \mathbb{R}$. Combining that with the fact that there is an increasing sequence of petite sets C_i whose union is the whole space, we see that $V(x)$ goes to infinity at infinity where "infinity" means away from petite sets.

Condition (i) means that the set S satisfies $\pi(S) = 1$, so although $V(x)$ is allowed to take the value ∞, it can only do so on a π-null set, and we can restrict the chain to the absorbing set S.

Since condition (ii) must hold for any solution of the drift condition, it does no harm to impose it as a requirement. This gives a simpler equivalent formulation (Meyn and Tweedie 1993, Lemma 15.2.8). A Markov chain satisfies the *geometric drift condition* if there exists a measurable function $V(x) \ge 1$ unbounded off petite sets, possibly taking the value $+\infty$ but finite at some x, a petite set C, and constants $\lambda < 1$ and $L < \infty$ such that

$$PV(x) \le \lambda V(x) + L. \qquad \text{for all } x \tag{4.20}$$

For any function $V \ge 1$ define the V-norm by

$$\|\mu\|_V = \sup_{|f| \le V} \int f \, d\mu. \tag{4.21}$$

Note the resemblance to the alternative definition (4.10) of the total variation norm. The only difference is that here the supremum is over all functions f dominated by V. The total variation norm is the special case $V \equiv 1$.

The geometric drift condition implies (Meyn and Tweedie 1993, Theorem 15.0.1) that there are constants $r > 1$ and $R < \infty$ such that

$$\sum_{n=1}^{\infty} r^n \| P^n(x, \cdot) - \pi \|_V \le RV(x) \qquad \text{for all } x. \tag{4.22}$$

holds for all x. This, of course, says nothing about x such that $V(x) = \infty$.

Comparison with the definition of geometric ergodicity (4.14) shows that (4.22) is stronger except that geometric ergodicity requires that the right hand side be finite for all x, which is not so in (4.22) when $V(x) = \infty$. But if we restrict the chain to the absorbing full set $S = \{ x : V(x) < \infty \}$, the geometric drift condition implies that the chain restricted to S is geometrically ergodic.

If the chain is ψ-irreducible and there is an everywhere finite solution to the positive drift condition, then the chain is Harris recurrent (Meyn and Tweedie, Theorem 11.3.4). The geometric drift condition implies the positive drift condition, so an everywhere finite solution to the geometric drift condition also implies Harris recurrence.

Thus in practice the nuisance of V being infinite at some points does not arise. One verifies the geometric drift condition using a V that is everywhere

finite. Why then allow for the possibility $V(x) = \infty$? For every geometrically ergodic chain, there is a V satisfying the geometric drift condition (Meyn and Tweedie 1993, Theorems 15.4.2 and 15.0.1), but the solution may take the value $+\infty$ at some points. Thus not only can one establish geometric ergodicity by verifying the geometric drift condition, but one loses nothing by taking this approach. If the chain is geometrically ergodic, then there is a function V that makes the geometric drift condition hold. Similarly, for every Harris recurrent chain, there is a V satisfying the positive drift condition (Meyn and Tweedie 1993, Theorem 11.0.1). Whether one can actually find such a function is another question, of course.

Further comparison shows that (4.22) is much stronger than (4.14) when V is everywhere finite, because of the appearance of the V-norm rather than the total variation norm in (4.22) and also because of the explicit formula for the dependence of the right hand side on x. Thus verifying the geometric drift condition implies something stronger than mere geometric ergodicity. One might call this V-geometric ergodicity, but Meyn and Tweedie apply that name to the situation where the left hand side of (4.22) is only known to be finite for all x. The still stronger (4.22) is called *V-uniform ergodicity.*

4.7.6 Verifying Geometric Drift

Bivariate Normal Gibbs

Verifying geometric drift ranges from the easy to the extremely difficult. To start, let us consider the Gibbs sampler for a bivariate normal distribution. Of course, one doesn't need MCMC to sample this distribution. This is a toy problem that makes a useful simple example for demonstrating a variety of techniques.

We may as well consider a symmetric normal distribution in which the two variables have the same variance σ^2 and mean zero. Their correlation is ρ. Then the conditional distribution of Y given X is normal with mean ρX and variance $\tau^2 = \sigma^2(1 - \rho^2)$, and vice versa. Since both updates use the same distribution, this Gibbs sampler is essentially an AR(1) time series, which is defined by $Z_n = \rho Z_{n-1} + e$ where e Normal$(0, \tau^2)$. The bivariate state of a fixed-scan Gibbs sampler for the bivariate normal is formed by taking consecutive pairs (Z_n, Z_{n+1}) from the univariate AR(1) time series.

Thus we can find out many things about this Gibbs sampler by looking in the time series literature. In particular, it is well known that this sampler is not only geometrically ergodic but satisfies much stronger properties. But let us, work through establishing the drift condition.

Since second moments are easy to calculate, we first try $V(x, y) = 1 + ax^2 + by^2$ for some positive constants a and b. This is clearly unbounded off compact sets, and compact sets are petite because this is a Gibbs sampler with continuous update densities. Suppose we update y last in the scan, so in order to take a conditional expectation PV for the whole scan, we first take the conditional expectation given x which gives a function of x alone and then take

a conditional expectation given y, where this y is the value in the preceding scan. The first conditional expectation gives

$$E(V|X) = 1 + ax^2 + b(\rho^2 x^2 + \tau^2) = (a + b\rho^2)x^2 + \text{constant}$$

From (4.20) we see there is no need to keep track of constants. Then the second conditional expectation gives

$$PV(x, y) = (a + b\rho^2)\rho^2 y^2 + \text{constant}$$

Thus we have geometric drift if we can choose a and b so that

$$(a + b\rho^2)\rho^2 < b,$$

which happens if

$$a < b(\rho^{-2} - \rho^2)$$

For example, if $\rho = .99$ then $b = 1$ and $a = .04$ will do.

A Theorem of Roberts and Tweedie

Roberts and Tweedie (1996) give a general theorem on geometric ergodicity of Metropolis samplers on $\mathbb{R}^d$ that iterate a single elementary update with a "random walk" proposal of the form $q(x, y) = f(y - x)$ where f is any density satisfying $f(x) = f(-x)$. They use a drift function of the form $V(x) = h(x)^{-1/2}$, where $h(x)$ is the unnormalized density of the invariant distribution. The conditions under which a drift function of this form can be used to establish geometric ergodicity can be roughly stated as $h(x)$ must have exponentially decreasing tails and asymptotically round contours. These conditions are violated by many models of practical interest, but the paper does show how the technical issues involved in proving geometric ergodicity using drift conditions are attacked. Presumably similar methods can be used with drift functions specifically tailored to the problem to establish geometric ergodicity for problems for which this specific choice does not work.

4.7.7 A Theorem of Rosenthal

Establishing the geometric drift condition tells us that a chain is geometrically ergodic (even V-uniformly ergodic) but doesn't tell us anything about the constants r and R in (4.22). By combining the geometric drift condition with a minorization condition like (4.15) we can say something about these constants.

Theorem 4.8. *Suppose $V(x) \geq 0$ is an everywhere finite function and satisfies a geometric drift condition*

$$PV(x) \leq \lambda V + L, \qquad \text{for all } x. \tag{4.23}$$

for some $\lambda < 1$ and some $L < \infty$. Suppose that the minorization condition

$$P(x, \cdot\,) \geq \delta Q(\cdot), \qquad \text{for all } x \text{ with } V(x) \leq d \tag{4.24}$$

holds for some $\delta > 0$, some probability measure Q, and some d satisfying

$$d > \frac{2L}{1 - \lambda}. \tag{4.25}$$

Then for $0 < r < 1$ and any initial distribution ν of the Markov chain

$$\|\nu P^k - \pi\| \le (1 - \delta)^{rk} + \left(\alpha^{-(1-r)} A^r\right)^k \left(1 + \frac{L}{1 - \lambda} + E_\nu V(X)\right)$$

where

$$\alpha^{-1} = \frac{1 + 2L + \lambda d}{1 + d} \qquad \text{and} \qquad A = 1 + 2(\lambda d + L)$$

This is Theorem 12 in Rosenthal (1995a, 1995b). The drift condition (4.23) is slightly different from the ones previously described, but if V satisfies (4.23) then $1 + V$ satisfies (4.18) with $C = \{ x : V(x) \le d \}$ which is petite because of the minorization condition (4.24) and a slightly larger λ. Note that (4.25) implies that $\alpha^{-1} < 1$, but A is always greater than one and may be very much larger. Thus it may be necessary to choose r very close to zero in order that $\alpha^{-(1-r)} A^r$ be less than one and the right hand side go to zero as $k \to \infty$.

Bivariate Normal Gibbs Again

Let us see how this works with the Gibbs sampler for the bivariate normal. First we must redo the drift condition calculation Section 4.7.6 keeping track of the constants to obtain L. But consideration of the minorization condition shows us that we can use a different drift function.

Since the conditional distribution of (X, Y) at time t only depends on the distribution of Y at time $t - 1$ (using a fixed scan that updates x and then y), the minorization condition will hold for all x if it holds for any x hence sets of the form $\mathbb{R} \times A$ are petite and we may as well use a function of y alone. Let us use $V(x, y) = by^2$.

Then

$$PV(x, y) = b[\tau^2 + \rho^2(\tau^2 + \rho^2 y^2)]$$

Hence $PV \le \lambda V + L$ with

$$\lambda = \rho^4$$

and

$$L = b\tau^2(1 + \rho^2).$$

Thus we must choose d satisfying

$$d > \frac{2b\tau^2(1 + \rho^2)}{1 - \rho^4} = \frac{2b\tau^2}{1 - \rho^2} = 2b\sigma^2$$

The small set on which the minorization condition needs to hold is

$$C = \{ (x, y) : V(x, y) \le d \},$$

which is of the form $\mathbb{R} \times A$ with

$$A = \{\, y : |y| \le \sqrt{d/b}\,\}.$$

The conditional distribution of X and Y at time $t+1$ given $Y_t = y_0$ has the density

$$\frac{1}{\sqrt{2\pi}\tau} \exp\left(-\frac{(y-\rho x)^2}{2\tau^2}\right) \frac{1}{\sqrt{2\pi}\tau} \exp\left(-\frac{(x-\rho y_0)^2}{2\tau^2}\right)$$

Taking the inf over all y_0 such that $|y_0| \le d/b$ gives

$$\frac{1}{\sqrt{2\pi}\tau} \exp\left(-\frac{(y-\rho x)^2}{2\tau^2}\right) \frac{1}{\sqrt{2\pi}\tau} \exp\left(-\frac{(|x|+\rho d/b)^2}{2\tau^2}\right) \tag{4.26}$$

Integrating with respect to y gives

$$\frac{1}{\sqrt{2\pi}\tau} \exp\left(-\frac{(|x|+\rho d/b)^2}{2\tau^2}\right)$$

and then integrating with respect to x gives

$$\delta = 2\Phi\left(-\frac{\rho}{\tau}\sqrt{\frac{d}{b}}\right) < 2\Phi\left(-\rho\sqrt{\frac{2}{1-\rho^2}}\right), \tag{4.27}$$

where Φ is the standard normal cumulative distribution function, that is, (4.26) is a proper probability distribution times δ.

Note that if ρ is very close to one, then (4.27) is extremely small. If $\rho = .99$, then $\delta < 3.28 \times 10^{-23}$. On the other hand, if $\rho = .9$, then $\delta < 0.0035$, which is not so bad. The parameters to be chosen are b, d, and r which together determine the bound. Some experimentation seemed to show that $b = 1$ and $d = 12.4$, just a little above its lower bound $2b/(1-\rho^2) = 10.526$, were about optimal. This makes $\alpha^{-1} = 0.9518$ and $A = 20.900$. If we now choose r so the two rate constants $(1-\delta)^r$ and $\alpha^{-(1-r)}A^r$ are about equal, we get $r = 0.0160$ making $(1-\delta)^r = \alpha^{-(1-r)} * A^r = 0.999976$. Hence

$$\|\nu P^k - \pi\| \le (0.999976)^k \left(2 + \frac{L}{1-\lambda} + E_\nu V(X)\right) = 7.263158(0.999976)^k$$

if we start at any point where $V(X) = bY^2 = 0$.

Thus when $\rho = .9$ we get a useful bound. It does say that to reduce the total variation norm to .01 we need 270,000 iterations, which is rather conservative, but is doable.

If $\rho = .99$ the bound is completely useless. It gives on the order of 10^{-23} iterations to reduce the bound much below one, and that is completely beyond any foreseeable available computer power. It is also ridiculously conservative. It is possible to use a minorization condition on the n-step kernel P^n rather than on P, which would give a better bound. But this would draw the wrong lesson from this toy problem. In problems of real practical interest, it is rarely, if ever, possible to say anything useful about n-step transition probabilities. Hence the appropriate lesson here seems to be that this theorem can be used to prove fast convergence, but that when convergence is moderately slow the bound becomes so conservative as to be useless.

4.7.8 Uniform Ergodicity

When the bound in the definition of geometric ergodicity is uniform, that is when there is a constant $R < \infty$ such that

$$\sum_{n=1}^{\infty} r^n \|P^n(x, \cdot) - \pi\| < R, \qquad \text{for all } x. \tag{4.28}$$

we say the chain is *uniformly ergodic*. This implies

$$\sup_{\text{all } x} \|P^n(x, \cdot) - \pi\| \to 0, \qquad \text{as } n \to \infty, \tag{4.29}$$

which Meyn and Tweedie take as the definition of uniform ergodicity. This makes sense because (4.29) also implies (4.28) by Theorems 16.2.1 and 15.0.1 in Meyn and Tweedie (1993).

Uniform ergodicity is implied by the geometric drift condition if the drift function V is bounded. Since any solution V of the geometric drift condition is unbounded off petite sets, boundedness of V implies that the whole state space is petite. Conversely, if a chain is uniformly ergodic, then the whole state space is petite and there exists a bounded solution of the geometric drift condition (Meyn and Tweedie 1993, Theorem 16.2.1).

Thus we obtain a very simple criterion for uniform ergodicity, that the whole state space be petite. In particular, if the chain is a T-chain and the state space is compact, then the chain is uniformly ergodic. No drift condition actually need be verified. For example, any Markov chain on a finite state space is uniformly ergodic. The chain is trivially a T-chain because $x \mapsto P(x, A)$ is trivially continuous for each A, since any function on a discrete space is continuous. The entire space is compact because any finite set is trivially compact. But this criterion also applies to more complicated examples. The Gibbs or Metropolis samplers for the Strauss process with a fixed number of points n are T-chains by the Fatou's lemma argument of Section 4.7.3. The state space is compact, since it is a closed and bounded subset of $\mathbb{R}^{2n}$ (or in the case of periodic boundary conditions a compact manifold of dimension $2n$). It is also easy to show the minorization condition directly: $0 \le s(x) \le n(n-1)/2$ implies that $h(x)$ is bounded and bounded away from zero and that this in turn implies that there is a $\delta > 0$ such that $P(x, A) \ge \delta\mu(A)$ for all points x and all measurable sets A, where $\mu(A)$ is the Lebesgue measure of A.

It is possible that a chain can be uniformly ergodic when the whole state space is not compact. A trivial example is independent sampling. A sequence $X_1, X_2, \ldots$ of independent, identically distributed random variables with distribution π is trivially a Markov chain with invariant distribution π and transition probability kernel $P(x, A) = \pi(A)$, for all x, and this is trivially a minorization condition for the whole space.

A nontrivial example of this phenomenon is a hierarchical Poisson model for data on pump failures at a nuclear power plant used by Gaver and O' Muircheartaigh (1987) who used empirical Bayes calculations that did not involve MCMC. Gelfand and Smith (1990) used this as an example where a fully Bayes

analysis could be done using the Gibbs sampler. Tierney (1994) showed that this Gibbs sampler is uniformly ergodic, even though the state space is an unbounded region of $\mathbb{R}^d$ and hence noncompact.

In general, however, one has no right to expect a Markov chain on a noncompact state space to be uniformly ergodic. For example, any sampler for the unconditional Strauss process that adds or deletes at most one point per iteration cannot be uniformly ergodic. Write S^m as before for the set of all realizations with exactly m points. Then for any $n > 0$ and any $x \in S^{m+n+1}$

$$\|P^n(x, \cdot) - \pi\| \geq |P^n(x, S^m) - \pi(S^m)| = \pi(S^m)$$

Since the chain cannot get from S^{m+n+1} to S^m in only n steps. Hence

$$\sup_{\text{all } x} \|P^n(x, \cdot) - \pi\| \geq \pi(S^m)$$

for all n, the left hand side cannot converge to zero, and the chain is not uniformly ergodic.

Another simple example is the Gibbs sampler for the bivariate normal. From the standard theory of AR(1) time series we know that the conditional distribution of Y_n given $Y_0 = y$ is normal with mean $\rho^{2n}y$. The unconditional variance of Y_n is σ^2 and the conditional variance given $Y_0 = y$ must be less since conditioning reduces variance. Hence for $y > 0$

$$\Pr(Y_n \leq 0 | Y_0 = y) \leq \Phi(\rho^{2n} y / \sigma) \tag{4.30}$$

In order for the chain to be uniformly ergodic this must be bounded uniformly in y, more precisely, for any $\epsilon > 0$ there is a n_ϵ such that $|\Phi(\rho^{2n}y/\sigma) - \pi(Y \leq 0)| \leq \epsilon$ whenever $n \geq n_\epsilon$ for all y. Clearly, this can't hold since $\pi(Y \leq 0) = \frac{1}{2}$ and (4.30) converges to 1 as $y \to \infty$.

4.8 The Central Limit Theorem

The assertion of the Markov chain central limit theorem (leaving aside momentarily the question of whether it is ever true) is the following. As when we were discussing the law of large numbers, define for any function $g(X)$

$$\mu = E_\pi g(X)$$

and

$$\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} g(X_i).$$

Then the law of large numbers says that $\hat{\mu}_n$ converges almost surely to μ, and we know this holds for any initial distribution for any Harris recurrent chain with invariant distribution π. The Monte Carlo error $\hat{\mu}_n - \mu$, how far a Monte Carlo estimate of μ based on a run of the chain of length n is from the true

value, converges to zero as the run length n goes to infinity. The central limit theorem asserts

$$\sqrt{n}\,(\hat{\mu}_n - \mu) \xrightarrow{\mathcal{D}} N(0, \sigma^2). \tag{4.31}$$

Root n times the Monte Carlo error converges in distribution to a normal distribution with mean zero and some variance σ^2, so $\hat{\mu}_n \pm 1.96\sigma/\sqrt{n}$ is an approximate 95% confidence interval for the unknown true value μ. In real problems there is never any way to calculate σ^2, but it can be estimated from the same run of the chain that produced the estimate $\hat{\mu}_n$. This is a familiar situation. Even with independent, identically distributed samples we rarely know the true variance, use the sample standard deviation s in place of σ in calculating the confidence interval.

One simple result about the central limit theorem is that if the chain is Harris recurrent, then if (4.31) holds for any initial distribution then it holds for every initial distribution (Meyn and Tweedie 1993, Theorem 17.1.6). Since the initial distribution does not effect the asymptotics, there is no harm in pretending that the initial distribution is the invariant distribution π, which allows us to make connections with the theory of stationary stochastic processes.

A stochastic process X_1, X_2, ... is *stationary* if for any positive integers n and k

$$(X_1, \ldots, X_k) \stackrel{\mathcal{D}}{=} (X_{n+1}, \ldots, X_{n+k})$$

meaning that the left hand side is equal in distribution to the right hand side. Any consecutive block of variables of length k has the same distribution. A Markov chain is a stationary stochastic process if X_1 has the invariant distribution π. Thus we can obtain a Markov chain central limit theorem from limit theorems for general stationary processes, including theorems about stationary time series.

4.8.1 The Asymptotic Variance

The variance σ^2 in the limiting distribution in the central limit theorem cannot simply be $\operatorname{Var}_\pi g(X)$ as it would be for independent sampling. The variance of the left hand side in (4.31) is

$$\sigma_n^2 = n \operatorname{Var}(\hat{\mu}_n) = \frac{1}{n} \sum_{i=1}^{n} \operatorname{Var}(g(X_i)) + \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \operatorname{Cov}(g(X_i), g(X_j))$$

Since the initial distribution makes no difference to the asymptotics, we may assume stationarity, in which case

$$\gamma_0 = \operatorname{Var}(g(X_i))$$

is the same for all i and

$$\gamma_k = \operatorname{Cov}(g(X_i), g(X_{i+k})) \tag{4.32}$$

is the same for all k. (4.32) is called the lag k autocovariance of the stationary time series $g(X_1)$, $g(X_2)$, ... Thus stationarity implies

$$\sigma_n^2 = \gamma_0 + 2 \sum_{k=1}^{n-1} \frac{n-k}{n} \gamma_k. \tag{4.33}$$

and σ_n^2 converges to

$$\sigma^2 = \gamma_0 + 2 \sum_{k=1}^{\infty} \gamma_k \tag{4.34}$$

as $n \to \infty$ if the series on the right hand side is summable. We can expect (4.34) to be the asymptotic variance if everything is well behaved.

4.8.2 Geometrically Ergodic Chains

The necessary conditions for such theorems involve so-called "mixing coefficients." There are several varieties of which we will look at three, so-called β-mixing, ρ-mixing, and ϕ-mixing. The reader should be warned that the definitions given here apply only to Markov chains and that the definition for a general stationary process is slightly different, for which see Bradley (1986).

β-Mixing

The mixing coefficient $\beta(n)$ is defined for a Markov chain by

$$\beta(n) = \tfrac{1}{2} \sup \sum_{i=1}^{I} \sum_{j=1}^{J} |\Pr(X_0 \in A_i \,\&\, X_n \in B_j) - \pi(A_i)\pi(B_j)|$$

where the supremum is taken over all partitions A_1, ..., A_I and B_1, ..., B_J of the state space by measurable sets.

This mixing coefficient is related to the total variation norm as follows. An alternative definition of the total variation norm of a signed measure μ is

$$\|\mu\| = \sup \sum_{j=1}^{J} |\mu(B_j)|$$

where again the supremum is over all measurable partitions of the state space. Thus

$$\sum_{j=1}^{J} |P^n(x, B_j) - \pi(B_j)| \le \|P^n(x, \cdot) - \pi\|,$$

for all measurable partitions $B_1, \ldots, B_J$ and

$$\begin{aligned}\sum_{j=1}^{J} |P^n(A_i, B_j) - \pi(A_i)\pi(B_j)| &= \sum_{j=1}^{J} \left| \int_{A_i} [P^n(x, B_j) - \pi(B_j)]\pi(dx) \right| \\ &\le \sum_{j=1}^{J} \int_{A_i} |P^n(x, B_j) - \pi(B_j)|\pi(dx) \\ &\le \int_{A_i} \|P^n(x, \cdot\,) - \pi\|\pi(dx)\end{aligned}$$

so

$$\begin{aligned}\beta(n) &= \tfrac{1}{2} \sum_{i=1}^{I} \sum_{j=1}^{J} |P^n(A_i, B_j) - \pi(A_i)\pi(B_j)| \\ &\le \tfrac{1}{2} \sum_{i=1}^{I} \int_{A_i} \|P^n(x, \cdot\,) - \pi\|\pi(dx) \\ &= \tfrac{1}{2} \int \|P^n(x, \cdot\,) - \pi\|\pi(dx)\end{aligned}$$

If the Markov chain is geometrically ergodic then (4.22) and $\int V \, d\pi < \infty$ imply there is an $r > 1$ such that

$$\sum_{n=1}^{\infty} r^n \beta(n) < \infty.$$

so $\beta(n)$ goes to zero exponentially fast. This implies a central limit theorem. A chain is said to be β-mixing if $\beta(n) \to 0$ and β-mixing exponentially fast if $\beta(n) \le A\rho^n$ for some $A < \infty$ and $\rho < 1$.

Theorem 4.9. *If a Markov chain is geometrically ergodic, then it is β-mixing exponentially fast. For any function g such that $\int |g|^{2+\epsilon} \, d\pi < \infty$ for some $\epsilon > 0$ the central limit theorem (4.31) holds for the stationary chain, and the asymptotic variance is given by (4.34). If the chain is Harris recurrent the central limit theorem holds for any initial distribution.*

This follows from a well-known stationary process central limit theorem (Ibragimov and Linnik 1971, Theorem 18.5.3). This connection between geometric ergodicity and mixing conditions was noted by Chan and Geyer (1994). Chan and Geyer only showed that geometric ergodicity implies a weaker form of mixing called α-mixing, but the proof of the stronger β-mixing is essentially the same, and β-mixing is need for some forms of empirical process central limit theorems (Arcones and Yu 1994; Doukhan, Massart, and Rio 1994).

It is possible to have $\sigma^2 = 0$, in which case the interpretation is that $\sqrt{n}(\hat{\mu}_n - \mu)$ converges in distribution to the degenerate distribution concentrated at the origin, which is the same thing as convergence in probability to zero. An example

of such behavior is the periodic chain on two states mentioned in Section 4.4. The average over a full period is the same as the average over the stationary distribution. Thus $\hat{\mu}_n$ is exactly μ for even n and off by at most $\frac{1}{n}\max(g(0), g(1))$ for odd n. So $\hat{\mu}_n - \mu = O(1/n)$ and $\sqrt{n}(\hat{\mu}_n - \mu)$ converges to zero.

The Liapunov condition $\int |g|^{2+\epsilon}\, d\pi < \infty$ can be suppressed, by considering the actual function V used in the geometric drift condition.

Theorem 4.10. *If a Markov chain is V-uniformly ergodic, then for any function g such that $g^2 \le V$ the central limit theorem (4.31) holds for the stationary chain, and the asymptotic variance is given by (4.34). If the chain is Harris recurrent the central limit theorem holds for any initial distribution.*

This is Theorem (17.5.4) in Meyn and Tweedie (1993). A very similar result is given by Chan (1993).

Which of the two theorems one uses depends on what what one knows. If it is not known whether g has $2 + \epsilon$ moments, then Theorem 4.10 or the similar theorem in Chan (1993) must be used. If one wants central limit theorems for many functions, all of which are known to satisfy the Liapunov condition, then Theorem 4.9 will be more useful, since there is no need to find a different drift condition for each function g.

ρ-Mixing

A stronger mixing condition is ρ-mixing. The mixing coefficient $\rho(n)$ is defined for a Markov chain by

$$\begin{aligned} \rho(n) &= \sup_{u,v \in L^2(\pi)} \operatorname{Cor}\bigl(u(X_i), v(X_{i+n})\bigr) \\ &= \sup_{u \in L^2(\pi)} \sqrt{\frac{\operatorname{Var}\bigl(E\{u(X_{i+n}) | X_i\}\bigr)}{\operatorname{Var}\bigl(u(X_i)\bigr)}} \end{aligned} \tag{4.35}$$

A chain is ρ-mixing if $\rho(n) \to 0$, as $n \to \infty$.

Thinking of P as an operator on the Hilbert space $L_0^2(\pi)$ as in Section 2.2.3 we have

$$\rho(n) = \sup_{u \in L_0^2(\pi)} \frac{\|P^n u\|}{\|u\|} = \|P^n\|.$$

The nth ρ-mixing coefficient is just the norm of P^n. Because $\|P\| \le 1$ (shown in Section 2.2.2) if $\|P^m\| < 1$ for any m

$$\|P^{mn+k}\| \le \|P^m\|^n$$

and so if a chain is ρ-mixing, then it is ρ-mixing exponentially fast.

In (4.35) it is usual to consider only real functions u and v, so $L^2(\pi)$ is considered a real Hilbert space. In defining the spectrum it is necessary to consider it a complex Hilbert space, but this makes no difference since P takes real functions to real functions, which implies $\|P(u + iv)\|^2 = \|Pu\|^2 - \|Pv\|^2$,

so the supremum over real functions is the same as the supremum over complex functions.

For any bounded operator T on a Hilbert space, the *spectrum* of T is the set of complex numbers λ such that $T - \lambda I$ is not invertible. If the state space is finite, so P is a matrix, then the spectrum of P is the set of right eigenvalues of P, the set of λ such that $Pu = \lambda u$ for some vector u. We have already seen that complex numbers are needed in the proof of theorem 4.6. If a chain is periodic with period d, then $e^{2\pi i/d}$ is an eigenvalue, and this is complex if $d > 2$. If the chain is reversible, so P is self-adjoint, then the spectrum is real.

If the state space is not finite, the notion of eigenvalues and eigenvectors may be insufficient to describe the spectrum. A function can fail to be invertible for two reasons, either it is not one-to-one or it is not onto. For a linear operator on a finite-dimensional vector space, these two collapse into one, but in general λ can be in the spectrum of P because $P - \lambda I$ is not one-to-one, which means that $(P - \lambda I)u = 0$ has a nonzero solution u and u is an eigenvector of P (also called *eigenfunction* to emphasize that u is a function on the state space) or $P - \lambda I$ is not onto, which means that there is a v that is not of the form $(P - \lambda I)u$ for any u in $L_0^2(\pi)$.

The spectrum of a bounded operator T is always a compact subset of the complex plane. The supremum of $|\lambda|$ for all λ in the spectrum is called the *spectral radius* $r(T)$. It is always true that $r(T) \le \|T\|$, so for a transition probability operator P which has $\|P\| \le 1$, the spectrum is a closed subset of the unit circle in general and a closed subset of the interval $[-1, +1]$ for self-adjoint P. A more precise bound is given by the spectral radius formula

$$r(P) = \lim_{n \to \infty} \|P^n\|^{1/n}.$$

If a chain is not ρ-mixing, then $\|P^n\| = 1$ for all n and $r(P) = 1$. If the chain is ρ-mixing, then there are constants $A < \infty$ and $b < 1$ such that $\rho(n) \le Ab^n$ and

$$r(P) \le \lim_{n \to \infty} A^{1/n} b = b < 1.$$

So a chain is ρ-mixing if and only if the spectral radius of P considered to be a operator on $L_0^2(\pi)$ is strictly less than one.

A method of demonstrating ρ-mixing has been devised by Schervish and Carlin (1992) and Liu, Wong, and Kong (1995). The connection between these methods and ρ-mixing was pointed out by Chan and Geyer (1994). These methods can only be applied to Gibbs samplers or other Metropolis-Hastings schemes in which all proposals are accepted for reasons explained by Chan and Geyer (1994).

The condition that a Markov chain be ρ-mixing is overly strong for obtaining a central limit theorem. What is important is that the spectrum not contain the point 1, that is, that the operator $I - P$, called the *Laplacian operator* of the chain be invertible. Clearly ρ-mixing implies this ($r(P) < 1$ implies that 1 is not in the spectrum).

Theorem 4.11. *If a Markov chain has an invertible Laplacian operator, then the central limit theorem (4.31) holds for the stationary chain, and the asymptotic variance is given by (4.34). If the chain is Harris recurrent the central limit theorem holds for any initial distribution.*

This is a simple corollary of a theorem of Gordin and Lifšic (1978) as is pointed out by Chan and Geyer (1994).

ϕ-Mixing

A stronger mixing condition is known as ϕ-mixing. For a Markov chain this is equivalent to a condition known a Doeblin's condition (Bradley 1986, p. 175) which is equivalent to uniform ergodicity (Meyn and Tweedie 1993, p. 384). Thus another method of establishing ρ-mixing is to establish uniform ergodicity. If the chain is uniformly ergodic, then the central limit holds for all functions in $L^2(\pi)$.

4.9 Estimating the Asymptotic Variance

A central limit theorem is not much use without a method of estimating the asymptotic variance σ^2. Three methods are presented in this section and a fourth method in the next section.

4.9.1 Batch Means

Given a Markov chain $X_1, X_2, \ldots$ and a function g for which there is a central limit theorem (4.31), fix an integer m, let l be the smallest integer greater than or equal to m/n and define the *batch means*

$$\hat{\mu}_{n,k} = \frac{1}{l} \sum_{i=(k-1)*l+1}^{kl} g(X_i), \qquad k = 1, \ldots, m-1$$

$$\hat{\mu}_{n,m} = \frac{1}{n - l(m-1)} \sum_{i=(m-1)*l+1}^{n} g(X_i).$$

It follows from the functional central limit theorem (Meyn and Tweedie 1993, Section 17.4) that the m batch means $\hat{\mu}_{n,k}$ are asymptotically independent and identically distributed Normal(μ, σ^2). Hence large sample confidence intervals for μ can be constructed using Student's t distribution. If $\bar{x}$ and s^2 are the sample mean and standard deviation of the batch means then $\bar{x} \pm t_{\alpha/2} s/\sqrt{m}$ is a $100(1-\alpha)\%$ confidence interval for μ, where $t_{\alpha/2}$ is the appropriate t critical value for $m-1$ degrees of freedom.

How does one choose the batch length l? A good recommendation (Schmeiser 1982) is that the number of batches should be small, no more than thirty. Using t rather than normal critical values correctly adjusts for a small number of batches, but nothing adjusts for batches that are too small. So the batches

should be as large as possible. One might use as few as ten batches if one were worried about the batches being too small.

4.9.2 Overlapping Batch Means

Although the theory of batch means is very simple, it is inefficient compared to a simple modification called *overlapping batch means* (Meketon and Schmeiser 1984; Pedrosa and W. 1993). For any batch length l, define

$$\hat{\mu}_{n,l,j} = \frac{1}{l} \sum_{i=j}^{j+l-1} g(X_i), \qquad j = 1, \ldots, n - l + 1$$

and

$$\hat{\sigma}^2_{n,l} = \frac{l}{n-l+1} \sum_{j=1}^{n-l+1} (\hat{\mu}_{n,l,j} - \hat{\mu}_n)^2 \tag{4.36}$$

It follows from the central limit theorem for $\hat{\mu}_n$ and uniform integrability, which always holds under exponentially fast β-mixing that $\hat{\sigma}^2_{n,l}$ converges to σ^2 in probability as $n \to \infty$ and $l/n \to 0$. Hence $\hat{\mu}_n \pm 1.96\hat{\sigma}_{n,l}/\sqrt{n}$ is an asymptotic 95% confidence interval for μ.

How does one chose the batch length for overlapping batch means. Now the choice is more difficult. In order for $\hat{\sigma}^2_{n,l}$ to be a consistent estimator l must be "large" and l/n must be "small." There seem to be no good criteria for choosing l unless n is very large, in which case a wide range of choices should be good enough. If n is "small" then no choice of l will be good.

4.9.3 Examples

Bivariate Normal Gibbs

One nice property of the Gibbs sampler for the bivariate normal distribution is that we can calculate its asymptotic variance exactly. Suppose we want to calculate the expectation of $g(X, Y) = Y$. For the stationary chain, the Y_n have variance σ^2 (not the variance in the central limit theorem but the marginal variance of Y) and correlation $\operatorname{Cor}(Y_i, Y_{i+k}) = \rho^{2k}$, thus the variance in the central limit theorem is

$$\begin{aligned} \operatorname{Var}(Y_i) + 2\sum_{k=1}^{\infty} \operatorname{Cov}(Y_i, Y_{i+k}) &= \sigma^2 \left(1 + 2\sum_{i=1}^{\infty} \rho^{2k} \right) \\ &= \sigma^2 \left(1 + 2\frac{\rho^2}{1-\rho^2} \right) \\ &= \sigma^2 \left(\frac{1+\rho^2}{1-\rho^2} \right) \end{aligned}$$

Figure 4.1 shows a run of length 10,000 of a Gibbs sampler for the bivariate

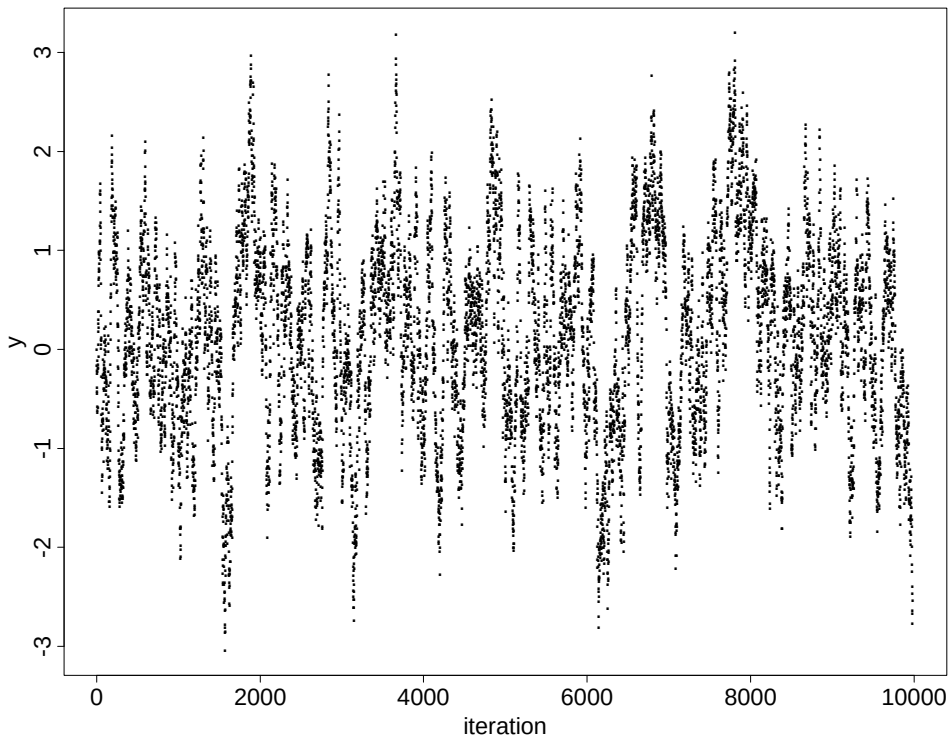


Figure 4.1: Output of the Gibbs sampler for the bivariate normal distribution with mean zero, variance one, and correlation $\rho = .99$. The starting position was $(0, 0)$ and the run length 10,000. The statistic plotted is the second component of the state vector.

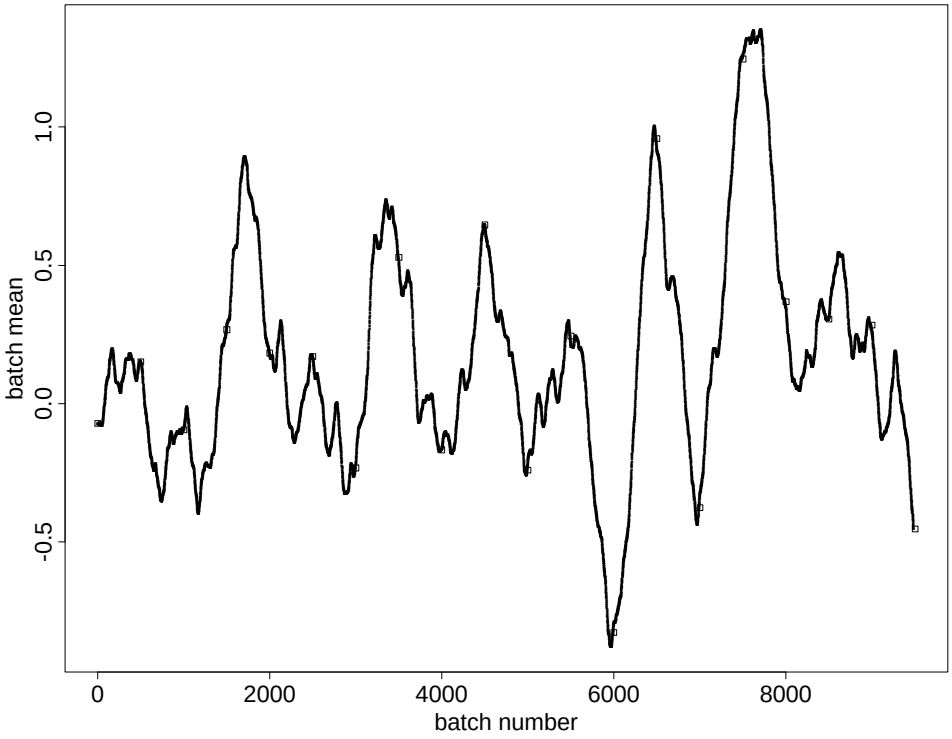


Figure 4.2: Overlapping batch means for the output shown in Figure 4.1. 9501 batches of length 500. Squares mark the 20 nonoverlapping batch means used in the ordinary batch means analysis.

normal distribution with a rather high correlation $\rho = 0.99$. The second variable Y of the state (X, Y) of the Markov chain is plotted.

Recall that in Section 4.7.6 we were able to show that this sampler is geometrically ergodic, hence a central limit theorem exists for any function satisfying a Liapunov condition and for Y in particular, but we were unable to get a tight bound on the convergence rate of the sampler in Section 4.7.7. A glance at Figure 4.1 shows that a run length of 10,000 is not long enough for the sampler to make many excursions to the extremes. The sample does have 0.0267 of its points above $+2$ and 0.0154 below -2 as compared to 0.025 for the invariant distribution π (which is standard normal), but only seven excursions above 1.96 make an appreciable contribution to the empirical expectation 0.0267 and only four excursions below -1.96 make an appreciable contribution to the empirical expectation 0.0154. So this Markov chain sample behaves something like an independent sample of size smaller than ten.

Figure 4.2 shows the batch means for batches of length 500. The ordinary batch means method uses the means of the twenty nonoverlapping batches

marked by squares in the figure. The mean and sample standard deviation are 0.145 and 0.484 giving a 95% confidence interval for the true mean $\mu = 0$ of $0.145 \pm 2.093 \cdot 0.484/\sqrt{20} = (-0.082, 0.371)$.

The estimated variance from the overlapping batch means is 81.27, which gives a confidence interval $0.145 \pm 1.96 \cdot \sqrt{81.27/10000} = (-0.032, 0.321)$. The correct theoretical value of the asymptotic variance is $(1+\rho^2)/(1-\rho^2) = 99.50$. Much of the underestimation of variance by the overlapping batch means estimator results from $\hat{\mu}_n$ not being μ. If μ were used (4.36) in place of $\hat{\mu}_n$ the estimate would be 95.14. There is, however, no way to correct for this, no way to widen the interval to account for something like degrees of freedom.

Conditional Strauss Process

Figure 4.3 shows a run of length 100,000 of a Metropolis sampler for a Strauss process with a fixed number of points. The distribution is bimodal with one mode near $s(x) = 175$ and another near $s(x) = 825$. Realizations in the low mode look much like those of a Poisson process. The points are almost independent. Realizations in the high mode have one cluster containing most of the points and a few scattered points outside. The Strauss process is not a very interesting model for clustering. It only serves as an interesting simple example of a spatial point process.

For this run, the mean of the canonical statistic $s(x)$ is 523.5 and the method of overlapping batch means with batch lengths of 2,000 estimates $\sigma^2 = 38981764$ giving a confidence interval of 523.5 ± 38.7 for the true expectation of $s(x)$.

4.9.4 Time Series Methods

A family of methods that are more complicated than batch means but also provide more information estimate the lagged autocovariances γ_k in (4.34) directly using the obvious estimator

$$\hat{\gamma}_k = \frac{1}{n} \sum_{i=1}^{n-k} [g(X_i) - \hat{\mu}_n][g(X_{i+k}) - \hat{\mu}_n]$$

This estimate is biased downwards, and one might think that dividing by $n - k$ rather than n would give a better estimate, but as we shall presently see, the estimates for large k are already too noisy and must be downweighted still further. Priestley (1981, pp. 323-324) discusses this in more detail. A naive estimate of σ^2 would be (4.34) with $\hat{\gamma}_k$ plugged in for γ_k, but it has long been known that this estimator is not even consistent (Priestley 1981, p. 432). For large k the variance of $\hat{\gamma}_k$ is approximately

$$\operatorname{Var}(\hat{\gamma}_k) \approx \frac{1}{n} \left(\gamma_0^2 + 2 \sum_{m=1}^{\infty} \gamma_m^2 \right) \tag{4.37}$$

(Bartlett 1946), assuming $\int g^4 \, d\pi < \infty$ and sufficiently fast mixing (ρ-mixing suffices). Figure 4.4 shows the estimated autocovariance function, $\hat{\gamma}_k$ as a func-

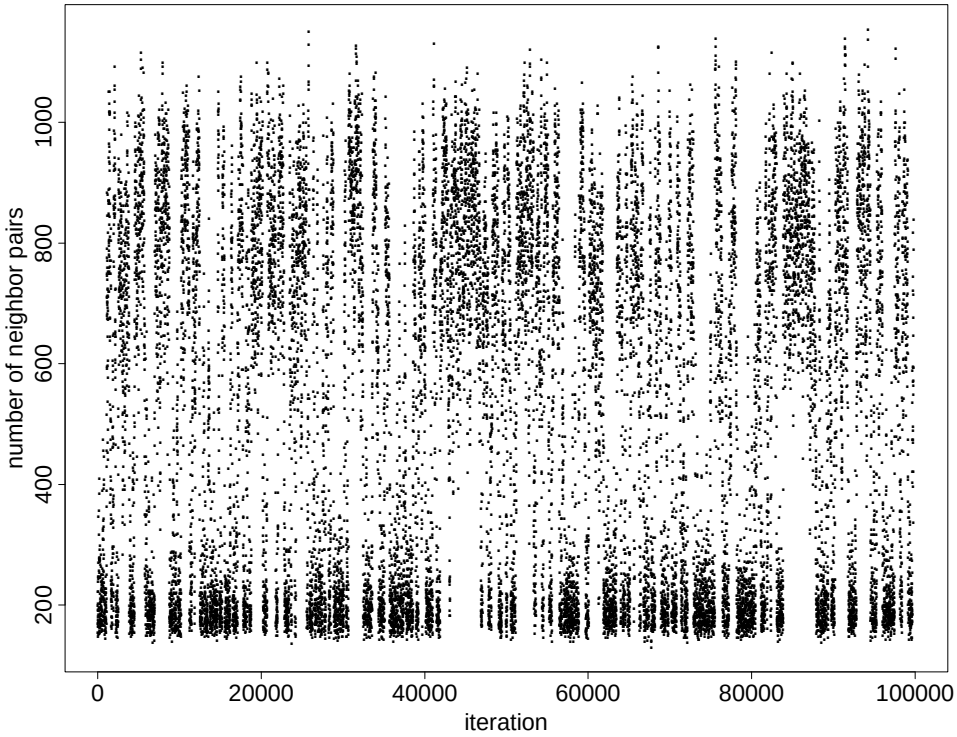


Figure 4.3: Metropolis sampler for the Strauss process with fixed number of points $n(x) = 50$ defined by (**??**) with canonical parameter $\beta = .126$. The vertical coordinate is the canonical statistic $s(x)$ which is the number of neighbor pairs. The run of length 100,000 was started at a realization of the Poisson process ($\beta = 0$). The plot only shows every fifth point, though all points were used in analyses.

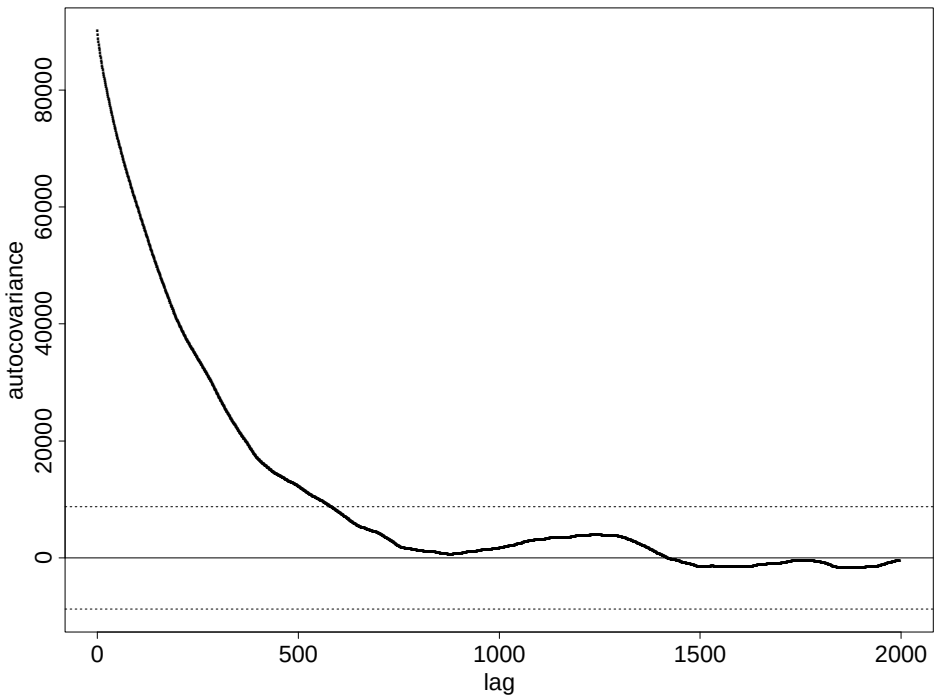


Figure 4.4: Empirical autocovariance function for the Metropolis sampler in Figure 4.3. The dotted lines are ± 1.96 times the asymptotic standard deviation of γ_k given by (4.37).

tion of k, with "large k confidence intervals calculated from (4.37) for the run shown in Figure 4.3.

In order to get an estimator of σ^2 that is even consistent, it is necessary to downweight the $\hat{\gamma}_k$ for large k.

$$\hat{\sigma}^2 = \hat{\gamma}_0 + 2 \sum_{k=1}^{\infty} w(k) \hat{\gamma}_k \tag{4.38}$$

where w is some weight function, called a *lag window*, satisfying $0 \le w \le 1$. Many weight functions have been proposed in the time-series literature. See Priestley (1981, p. 437 ff. and p. 563 ff.) for a discussion of choosing a lag window.

Typically one expects the autocovariance function to decline smoothly to zero and to be positive for all k, so it would seem that one could just truncate the sequence $\hat{\gamma}_k$ where it goes negative, but autocovariances can be negative, and usually nothing is known about the true autocovariance function of a sampler, so this approach is less than rigorous, except in one special case, when the chain

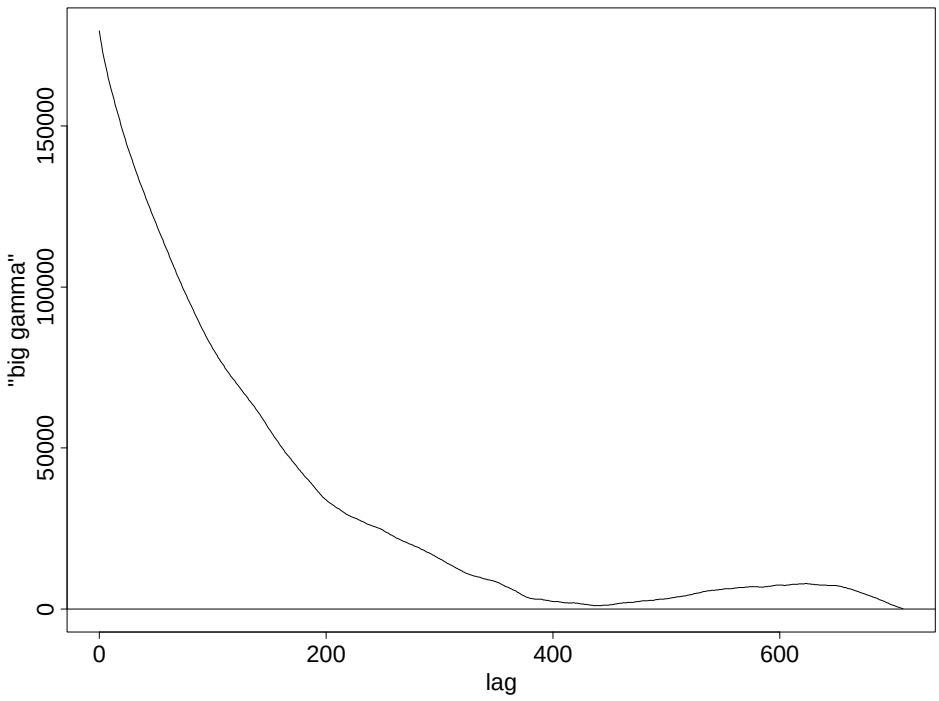


Figure 4.5: Plot of $\gamma_{2k} + \gamma_{2k+1}$ versus k for the Metropolis sampler in Figure 4.3.

is reversible. Geyer (1992) noted that the function $\Gamma_k = \gamma_{2k} + \gamma_{2k+1}$ is a strictly positive, strictly decreasing, and strictly convex function of k if the chain is reversible.

Thus for reversible chains it is rigorously correct to use any of the following three estimators based on using one of the three known properties of the "big gamma" function. The *initial positive sequence estimator* is the sum

$$\hat{\sigma^2} = \hat{\gamma}_0 + 2\hat{\gamma}_1 + \sum_{k=2}^{M} \hat{\Gamma}_k \tag{4.39}$$

where M is the largest integer such that the $\hat{\Gamma}_k$ are strictly positive for $k = 2$, $\ldots$, M.

The bulge in the figure above lag 450 is not like the behavior of a true "big gamma" function, so it makes sense to further to reduce the estimated $\hat{\Gamma}_k$ so that they are nondecreasing

$$\hat{\Gamma}_k^{(\text{mon})} = \min\left(\hat{\Gamma}_1, \ldots, \hat{\Gamma}_k\right)$$

and then replace $\hat{\Gamma}_k$ by $\hat{\Gamma}_k^{(\text{mon})}$ in (4.39). This gives the *initial monotone sequence estimator*.

The smaller bulges that make Figure 4.5 nonconvex can also be eliminated by taking the function $k \mapsto \hat{\Gamma}_k^{(\text{con})}$ to be the greatest convex minorant of $\hat{\Gamma}_1, \ldots, \hat{\Gamma}_M, 0$, and replacing $\hat{\Gamma}_k$ by $\hat{\Gamma}_k^{(\text{con})}$ in (4.39). This gives the *initial convex sequence estimator*. For any function g, the greatest convex minorant is supremum of all convex function $h \le g$. It can be constructed by the pool adjacent violators algorithm (Robertson, Wright, and Dykstra 1988, pp. 8–11).

For the run shown in Figure 4.3, the initial positive sequence estimator is 44.97×10^6, the initial monotone sequence estimator is 42.91×10^6, and the initial convex sequence estimator is 42.47×10^6. Recall that the overlapping batch means estimator was 38.98×10^6, which now seems too small. Increasing the batch length from 2,000 to 10,000 makes the overlapping batch means estimator 47.53×10^6. The choice of batch size can make a large difference in the estimator.

So which should one use, batch means, overlapping batch means, a lag window estimator using a window from the time series literature, or one of the initial sequence estimators? Ordinary batch means is the simplest and performs reasonably well. Overlapping batch means is better (Meketon and Schmeiser 1984). Unfortunately there is no good way to choose the batch length, one just chooses it to be reasonably long and hopes that is good enough. Any attempt to make a good choice by some adaptive procedure makes batch means more complicated than time series methods. The initial sequence methods provide a reasonable default lag window estimator, but do require that one use a reversible chain.

The choice of method is not as important as the choice to use *some* method. Variance calculations are still a rarity in the MCMC literature. Some have argued that because the do not diagnose "nonconvergence" there is no point in using them, that is, when $\hat{\mu}$ is very badly estimated because the run is far too short, then the estimate of σ^2 will be a gross underestimate. The same argument could be applied to all uses of confidence intervals—since they don't tell you when they fail to cover the true parameter value there is no point in using them—which is obvious nonsense. The right way to think about variance calculations is that they are the only way to say anything quantitative about the accuracy of an MCMC sampler or about the relative accuracy of two MCMC samplers. The following quotation from Geyer (1992) is still good advice.

> It would enforce a salutary discipline if the gold standard for comparison of Markov chain Monte Carlo schemes were asymptotic variance (asymptotic relative efficiency) for well-chosen examples that provide a good test of the methods. Experience shows that it is easier to invent methods than to understand exactly what their strengths and weaknesses are and what class of problems they solve especially well. Variance calculations seem to be the only sufficiently stringent standard for such investigations.

4.10 Regeneration

A very different method for estimating Monte Carlo error uses regeneration. A set α in the state space is said to be an *atom* if

$$P(x, \cdot) = P(y, \cdot), \qquad \text{for all } x, y \in \alpha. \tag{4.40}$$

This says the transition probabilities are the same from every point in the atom. Let $\tau_0, \tau_1, \ldots$ denote the times of visits to the atom, that is $X_j \in \alpha$ if and only if $j = \tau_i$ for some i. The τ_i are called *regeneration times* because the past history of the chain is forgotten. Because of (4.40) the future paths started from any two states in the atom have the same probability laws. In particular, segments of the sample path between regeneration times

$$X_{\tau_i+1}, \ldots, X_{\tau_{i+1}},$$

which are called *tours*, are independent and identically distributed.

If we are interested in calculating the expectation of a function g, the sums

$$Z_i = \sum_{k=\tau_{i-1}+1}^{\tau_i} g(X_k), \qquad i = 1, 2, \ldots$$

over the tours are independent and identically distributed random variables, as are the tour lengths

$$N_i = \tau_i - \tau_{i-1}, \qquad i = 1, 2, \ldots.$$

If the chain is Harris recurrent and the atom has positive probability under the invariant distribution, the atom is said to be *accessible*. An accessible atom is visited infinitely often with probability one, and there is an infinite sequence of regenerations. By the renewal theorem

$$E(N_i) = \frac{1}{\pi(\alpha)},$$

and by an analog of Wald's lemma in sequential sampling

$$E(Z_i) = E(N_i)\mu \tag{4.41}$$

where $\mu = E_\pi(g(X))$ (Nummelin 1984, pp. 76 and 81).

Another way to see this uses the identity

$$\frac{1}{n} \sum_{i=1}^{n} 1_\alpha(X_i) = \frac{k+1}{\tau_0 + N_1 + \cdots + N_k}$$

By the law of large numbers for Markov chains, the left hand side converges to $\pi(\alpha)$. By Harris recurrence, τ_0 is almost surely finite. Hence by the law of

large numbers for independent random variables, the right hand side converges to $1/E(N_i)$. Then

$$\frac{1}{n}\sum_{i=1}^{n} g(X_i) = \frac{1}{n}\sum_{i=1}^{\tau_0} g(X_i) + \frac{Z_1 + \cdots + Z_k}{\tau_0 + N_1 + \cdots + N_k}$$

and the same argument shows that the left hand side converges to μ and the right hand side converges to $E(Z_i)/E(N_i)$. It is not clear that this argument can be made noncircular, since the usual proofs of the law of large numbers and facts about Harris recurrence use regeneration, but it does help understand the phenomenon.

If $Z_i - \mu N_i$ has finite variance τ^2, then there will be a central limit theorem for

$$\hat{\mu}_k = \frac{\bar{z}_k}{\bar{n}_k} = \frac{Z_1 + \cdots + Z_k}{N_1 + \cdots + N_k}. \tag{4.42}$$

Write $\nu = E(N_i)$. Then

$$\sqrt{k}(\hat{\mu}_k - \mu) = \frac{\sqrt{k}(\bar{z}_k - \mu\bar{n}_k)}{\bar{n}_k} \xrightarrow{\mathcal{D}} \text{Normal}\left(0, \frac{\tau^2}{\nu^2}\right)$$

by Slutsky's theorem. The condition that $Z_i - \mu N_i$ have finite variance is a necessary and sufficient condition for the central limit theorem for $\sqrt{k}(\bar{z}_k - \mu\bar{n}_k)$ and hence is the weakest possible condition for a Markov chain central limit theorem. Being a necessary condition, it holds whenever there is a central limit theorem, such as when the chain is geometrically ergodic and g satisfies a Liapunov condition, but there seem to be no tools for verifying the condition other than those that apply in the absence of regeneration. When the geometric drift condition has been established with a drift function V that is bounded on the atom α and satisfies $g^2 \le V$, then both Z_i and N_i have finite variance by Theorem 14.2.3 in Meyn and Tweedie (1993).

If we average over a fixed number of complete tours, the numerator and denominator in (4.42) have the correct expectations by (4.41). The estimator $\hat{\mu}$ has a slight bias because the expectation of a ratio is not the ratio of the expectations, but the bias is asymptotically negligible and usually small in practice if the number of tours is large.

This property of the numerator and denominator have the correct expectations is preserved if we take a random number K of complete tours, so long as K is a *stopping time*, that is, the decision to stop at time k is made using only information available at time k, in particular it does not make use of (Z_i, N_i) for $i > k$. Then if Z_i and N_i have finite variance

$$E\left(\sum_{i=1}^{K} Z_i\right) = \mu E\left(\sum_{i=1}^{K} N_i\right) \tag{4.43}$$

$$\text{Var}\left(\sum_{i=1}^{K} (Z_i - \mu N_i)\right) = \tau^2 E(K) \tag{4.44}$$

(4.43) is the analog of Wald's lemma with random stopping, and (4.44) says that the natural estimate of τ^2 would have an unbiased numerator and denominator if the true value of μ were used the deviations. These follow from

$$\begin{aligned}
E\left(\sum_{i=1}^{K} Z_i\right) &= \mu\nu E(K) \\
E\left(\sum_{i=1}^{K} N_i\right) &= \nu E(K) \\
\mathrm{Var}\left(\sum_{i=1}^{K} Z_i - K\mu\nu\right) &= \mathrm{Var}(Z_i)E(K) \\
\mathrm{Var}\left(\sum_{i=1}^{K} N_i - K\nu\right) &= \mathrm{Var}(N_i)E(K) \\
\mathrm{Cov}\left(\sum_{i=1}^{K} Z_i - K\mu\nu, \sum_{i=1}^{K} N_i - K\nu\right) &= \mathrm{Cov}(Z_i, N_i)E(K)
\end{aligned}$$

which in turn follow from Theorem 5.3 and Remark 5.7 in Chapter I of Gut (1988).

The law of large numbers and the central limit theorem continue to hold for random stopping. If $K(t)$, $t \geq 0$ is a family of positive-integer-valued random variables such that $K(t) \to +\infty$ almost surely as $t \to \infty$ (not necessarily stopping times), then

$$\hat{\mu}_{K(t)} \xrightarrow{\text{a. s.}} \mu, \qquad t \to \infty.$$

This follows from Theorem 4.1 in Chapter I of Gut (1988). If Z_i and N_i have finite variance then

$$\sqrt{K(t)}\left(\hat{\mu}_{K(t)} - \mu\right) \xrightarrow{\mathcal{D}} \text{Normal}\left(0, \frac{\tau^2}{\nu^2}\right)$$

follows from Theorem 3.1 in Chapter I of Gut (1988) and the delta method.

4.10.1 Estimating the Asymptotic Variance

From (4.44)

$$\hat{\tau}_K^2 = \frac{1}{K}\sum_{i=1}^{K} \left(Z_i - N_i\hat{\mu}_K\right)^2 \tag{4.45}$$

is an approximately unbiased estimate of τ^2, only approximately unbiased because we have plugged in $\hat{\mu}_K$ for μ and because the expectation of a ratio is not equal to the ratio of the expectations when K is random. A consistent estimator of ν is, of course

$$\hat{\nu}_K = \frac{1}{K}\sum_{i=1}^{K} N_i.$$

Then $\hat{\sigma}_K^2 = \hat{\tau}_K^2/\hat{\nu}_K^2$ estimates the variance in the central limit theorem. This simple estimate has fairly good properties. It is analogous to the ratio estimator in finite population sampling.

Another possibility, discussed by Ripley (1987, pp. 160–161) is to jackknife the estimator μ_K. This will generally produce similar answers to the simple ratio estimator, leading to the conclusion that the biases are unimportant. See Section 4.10.7 for an example.

4.10.2 Splitting Markov Chains

Any Markov chain on a discrete state space has accessible atoms. Any point with positive probability is one since (4.40) is satisfied trivially when α only contains one point. But that is not much help unless the atom has fairly large probability so the regeneration rate $\pi(\alpha)$ is fairly large. And how does one find atoms for a chain with a continuous state space?

Nummelin (1978) and Athreya and Ney (1978) independently invented a method for constructing atoms for Markov chains on general state spaces. The method is used throughout the modern theory of Markov chains on general state spaces, which is laid out in the books by Nummelin (1984) and Meyn and Tweedie (1993). Mykland, Tierney, and Yu (1995) apply the technique to Markov chain Monte Carlo. The construction below follows Mykland, Tierney, and Yu (1995) who followed Nummelin (1984). The terminology has been changed to follow Meyn and Tweedie.

Suppose that we have a Harris recurrent Markov chain satisfying the following minorization condition: for some nonnegative measurable function s and some probability measure ν such that $\int s\,d\pi > 0$

$$P(x, A) \geq s(x)\nu(A) \qquad \text{for all points } x \text{ and measurable sets } A. \tag{4.46}$$

This is similar to the minorization conditions (4.15) used in the definition of small sets and (4.24) used in Rosenthal's theorem, but it is more general in replacing a constant δ with a function $s(x)$. It is also less general than (4.15) in that one must minorize the kernel P rather than an iterated kernel P^m.

Condition (4.46) allows the following construction of a chain on an enlarged sample space, called the *split chain*, that has an atom and that is related to the original chain by marginalization. We add to the state space a $\{0, 1\}$-valued variable S, that is the indicator of the atom. Thus the state of the split chain is the pair (X, S) where X takes values in the original state space.

The transition law of the split chain is described as follows. Note that if E is whole state space $1 = P(x, E) \geq s(x)\nu(E) = s(x)$, so $0 \leq s \leq 1$. At time t the state of the split chain is (X_t, S_t). If $S_t = 1$ then X_{t+1} is generated from the distribution ν, otherwise X_{t+1} is generated from the distribution

$$\frac{P(X_t, \cdot) - s(X_t)\nu(\cdot)}{1 - s(X_t)} \tag{4.47}$$

which is a normalized probability distribution because of the minorization condition (4.46). Then generate a Uniform$(0, 1)$ random variable U and set $S_{t+1} = 1$

if $U < s(X_{t+1})$ and otherwise set $S_{t+1} = 0$. It is clear that the distribution of (X_{t+1}, S_{t+1}) does not depend on the value of X_t when $S_t = 1$. Thus the set of points $\alpha = \{ (X, S) : S = 1 \}$ is an atom of the split chain.

Moreover, the sequence X_1, X_2, ... is a Markov chain with kernel P, since

$$\begin{aligned}
&\Pr(X_{t+1} \in A | X_t = x) \\
&= \Pr(S_t = 1 | X_t = x)\nu(A) + \Pr(S_t = 0 | X_t = x)\frac{P(x, A) - s(x)\nu(A)}{1 - s(x)} \\
&= s(x)\nu(A) + (1 - s(x))\frac{P(x, A) - s(x)\nu(A)}{1 - s(x)} \\
&= P(x, A)
\end{aligned}$$

So we have not disturbed the distribution of the X component of the state (X, S). The split chain has an invariant distribution in which X has the marginal distribution π and the conditional distribution of S given X has the density $s(x)$ with respect to π. The probability of the atom is thus $\int s \, d\pi$ and the atom is accessible.

Because of the Markov property, the S's are conditionally independent given the X's and the conditional distribution of S_t given all the X's depends only on X_t and X_{t+1} (Nummelin 1984, p. 62)

$$\begin{aligned}
r(x, y) &= \Pr(S_t = 1 | X_t = x, X_{t+1} = y) \\
&= \frac{s(x)\nu(dy)}{P(x, dy)},
\end{aligned}$$

where the last term is a Radon-Nikodym derivative. For every x such that $s(x) > 0$, the measure $P(x, \cdot)$ dominates ν and hence ν has a density f_x with respect to $P(x, \cdot)$. Then $r(x, y) = s(x) f_x(y)$.

We could thus simulate the split chain by first simulating X_1, X_2, ... using the original transition mechanism, and then go back later and simulate S_t as independent Bernoulli random variates with success probability $r(X_t, X_{t+1})$.

4.10.3 Independence Chains

Tierney (1994) proposed a simple special case of the Metropolis-Hastings algorithm called "independence" chains, something of a misnomer, because the proposals are independent, not the samples. The method proposes a new state y from a density $q(y)$ that does not depend on the current state x. Thus the Hastings ratio (3.18) becomes

$$R = \frac{h(y)q(x)}{h(x)q(y)}, \tag{4.48}$$

where $h(x)$ is an unnormalized density of the invariant distribution, both h and q being densities with respect to the same measure μ.

It is not clear that this idea is interesting used by itself. It should be compared to importance sampling using $q(x)$ as an importance distribution, which

will be explained in Section **??**. But no comparison seems to have been done, and it is not clear that independence chains have any advantage over importance sampling. Roberts and Tweedie (submitted) show that an independence chain is geometrically ergodic if and only if $h(x)/q(x)$ is bounded, in which case importance sampling is guaranteed to work well too.

4.10.4 Splitting Independence Chains

Mykland, Tierney and Yu (to appear) give the following simple recipe for splitting independence chains. Let c be an arbitrary positive constant. Define

$$\begin{aligned} w(x) &= \frac{h(x)}{q(x)}, \\ s(x) &= K \min\left\{\frac{c}{w(x)}, 1\right\}, \\ \nu(dy) &= \frac{1}{K} \min\left\{\frac{w(y)}{c}, 1\right\} q(y)\mu(dy) \end{aligned}$$

where K is chosen to make ν a probability measure. Without knowing K it is impossible to simulate the split chain by simulating S_t from its conditional distribution given X_t and X_{t+1} from its conditional distribution given X_t and S_t. Thus Mykland, Tierney and Yu (to appear) propose a method of simulating S_t from its conditional distribution given X_t and X_{t+1}, which differs a bit from the general scheme described in Section 4.10.2 in that we only set $S_t = 1$ when the Metropolis update from X_t to X_{t+1} is not a rejection. It uses the function

$$r_A(x, y) = \begin{cases} \max\left\{\frac{c}{w(x)}, \frac{c}{w(y)}\right\}, & w(x) > c \text{ and } w(y) > c, \\ \max\left\{\frac{w(x)}{c}, \frac{w(y)}{c}\right\}, & w(x) < c \text{ and } w(y) < c, \\ 1, & \text{otherwise.} \end{cases} \tag{4.49}$$

The overall update then goes as follows. Given $X_t = x$, propose a y with density q and accept the proposal with probability $\min(R, 1)$ where R is given by (4.48), that is $X_{t+1} = y$ if the proposal is accepted and $X_{t+1} = x$ otherwise. If the proposal is not accepted, set $S_t = 0$. If the proposal is accepted, set $S_t = 1$ with probability $r_A(x, y)$ given by (4.49) and $S_t = 0$ otherwise. Note that S_t is generated after X_{t+1}, which can be confusing if one is not careful.

Since this scheme does not refer to the normalizing constant K, it can be carried out. Although it works for any positive c, Mykland, Tierney and Yu (to appear) claim that it will be more efficient if c is chosen to be near the center of the distribution of the weights $w(X)$ when X has the invariant distribution. This does not appear to be correct. See Section 4.10.6.

The chain can be started with an arbitrary value for X_1 or it can be started at the regeneration point by setting $S_0 = 1$ and sampling X_1 from ν. This can be done without knowing the normalizing constant K by rejection sampling. Repeatedly simulate a y with density q and a Uniform$(0, 1)$ random variate u until $u < \min\left\{\frac{w(y)}{c}, 1\right\}$. Then y has the distribution ν. Set $X_1 = y$.

4.10.5 Metropolis-rejected Restarts

The independence proposal idea does have interesting application to restarting Markov chains (Tierney 1994). Restarting a Markov chain is an old idea of questionable validity that will be discussed further in Section **??**. If a Markov chain is very slowly mixing, then it seems to make sense to "restart" the Markov chain at some other point of the state space rather than wait for it to get there by itself. But this changes from an algorithm that converges, however slowly, to a known invariant distribution to an algorithm with unknown and generally unknowable properties. One thing is clear from Theorem 4.7, restarting always increases the distance from the marginal distribution of X_t to the invariant distribution π.

If, however, one wants to do something with restarts, it is not clear that they should ever be accepted without Metropolis rejection. If one attempts a restart y, then doing a Metropolis rejection with the Hastings ratio (4.48) preserves the invariant distribution and, if done at the beginning or end of each scan, preserves the Markov chain structure as well. We call this method Metropolis-rejected restarts. It is merely the composition of the original update mechanism with Tierney's "independence chain" update. It gives at least some of the benefits of restarting with none of the drawbacks.

4.10.6 Splitting Metropolis-rejected Restarts

Let Q denote the kernel for the split independence chain update described in Section 4.10.4. It updates the state (X, S). Let P denote any other kernel that preserves the same invariant distribution for X, which we trivially extend to an update rule for (X, S) by leaving S alone. Then the composite kernel QP preserves the invariant distribution of the split chain, and the times t when $S_t = 1$ are regenerations, because then the update of X by the Q kernel does not depend on the value of X_t.

Formally Q moves from (X_t, S_t) to an intermediate state (X', S'), and P moves from (X', S') to (X_{t+1}, S_{t+1}). Since P doesn't change S, we have $S' = S_{t+1}$. In practice, though, our mechanism for the split independence chain update does not produce (X', S_{t+1}) given (X_t, S_t). Instead it produces X' and S_t given X_t. We cannot produce S_t until we have produced the X' for the next iteration. Thus the algorithm goes as follows.

Set $S_0 = 1$
Generate x' from ν by rejection sampling
for $t = 1, 2, \ldots$ **do**
  Simulate x from $P(x', \cdot)$.
  Simulate y from q
  Simulate u Uniform(0,1)
  Calculate R given by (4.48)
  if $(u < R)$ **then**
    $x' = y$
    Simulate u Uniform(0,1)

```
    Calculate r_A(x, y) given by (4.49)
    if (u < r_A(x, y)) then
      s = 1
    else
      s = 0
    end if
  else
    x' = x
    s = 0
  end if
  Set X_t = x and S_t = s.
end do
```

The looping is a bit confusing if not explained. P is done at the top of the loop, though it is supposed to follow Q. The reason it that the loop begins in the middle of the iteration. At the top of the loop we have $X_{t-1} = x$ and $X' = x'$ and $S_{t-1} = s$. The loop begins by using P to generate $X_t = x$. Then it generates the x' for the next iteration so it can generate the $s = S_t$ for this iteration. At the bottom of the loop we output (X_t, S_t). The only state used in the following iteration is x'.

The code starts at the regeneration point. $S_0 = 1$. The value of X_0 is irrelevant, since the conditional distribution of X following a regeneration is independent of the previous value. In order to do this the first value of X' cannot be generated by the same code as used in the loop, we must generate a sample from ν using rejection sampling as described at the end of Section 4.10.4. This gives the x' value needed at the top of the loop.

4.10.7 Splitting the Strauss Process

The scheme of the preceding section is implemented for the Strauss process with a fixed number of points in the program `regen.c` described in Appendix **??**. The restart distribution is the binomial process (all points independently and uniformly distributed). Thus the density q is constant and the Hastings ratio for the Metropolis rejected restarts is simply

$$R = \frac{h(y)}{h(x)} = \exp\{\beta[t(y) - t(x)]\}$$

where we are now using $t(x)$ to denote the canonical statistic, number of neighbor pairs to avoid confusion with the splitting function $s(x)$. (4.49) can also be simplified to

$$r_A(x, y) = \begin{cases} \exp\{-\beta \min[t(x) - c', t(y) - c']\}, & t(x) > c' \text{ and } t(y) > c', \\ \exp\{-\beta \min[c' - t(x), c' - t(y)]\}, & t(x) < c' \text{ and } t(y) < c', \\ 1, & \text{otherwise.} \end{cases} \tag{4.50}$$

where $c' = (logc)/beta$. To start off the simulation we need one realization from ν which is sampled by repeatedly simulating realizations x from the binomial process and uniform random variates u until

$$u < \exp\{\beta[t(x) - c']\}.$$

The same process with $\beta = .126$ and $n(x) = 50$ as in Figure 4.3 was used. Since realizations from the binomial process only resemble realizations in the low mode of the Strauss process with $t(x)$ around 175, the first run of the sampler was done with $c' = 175$. About 45% of accepted restarts were regenerations, but the overall regeneration was only 2.9% because few restarts were accepted.

During this run, both the state x at the time of the attempted restart, the proposed restart y, and an indicator of whether the restart was accepted were written out. This permitted estimation of the expected regeneration by averaging $r_A(x, y)$ over iterations in which a restart was accepted. Figure 4.6 The figure shows that using $c' = 162$ should increase the regeneration rate to 66.2% of accepted restarts. Note that this is nowhere near the center of the distribution of $t(x)$ under the invariant distribution, which is about 480. If c' were set there, the sampler would not regenerate at all. The prediction from this calculation was borne out by another run with $c' = 162$ in which 66.8% of accepted restarts were regenerations for an overall regeneration rate of 4.6%.

This run proceeded to the first regeneration point after 100,000 iterations which was iteration 100,488 during which there were 4,628 tours, giving a mean tour length 21.7 (standard error 1.27). Taking μ to be the expectation of the canonical statistic $t(x)$, the estimator was $\hat{\mu} = 448.36$. The estimator (4.45) was $\hat{\tau}^2 = 6.67 \times 10^8$ giving an estimator $\hat{\sigma}^2 = 6.67 \times 10^8/21.7^2 = 1.42 \times 10^6$ for the variance in the central limit theorem and $\sqrt{\hat{\sigma}^2/4,628} = 17.49$ for the standard error of $\hat{\mu}$.

For comparison we computed the time-series estimators using the same run, which gave 18.01 for the standard error of $\hat{\mu}$ using the initial positive sequence and monotone sequence estimators and 17.98 using the convex sequence estimator.

Another comparison used the jackknife. This procedure makes a bias correction to $\hat{\mu}$ giving 449.33 for the estimator of μ. The estimated standard error is 17.66. The bias correction made by the jackknife is only 0.2the same as that calculated by the simple ratio estimate.

To see how well the estimation did we ran the sampler about nine times longer giving a total of 41,488 tours, including the run already used for estimation. This gave a new estimate $\hat{\mu} = 479.12$ with standard error 6.34. The difference between the two estimates is 30.76, which is about 1.7 estimated standard errors. So the Estimation of standard errors seems to have worked well.

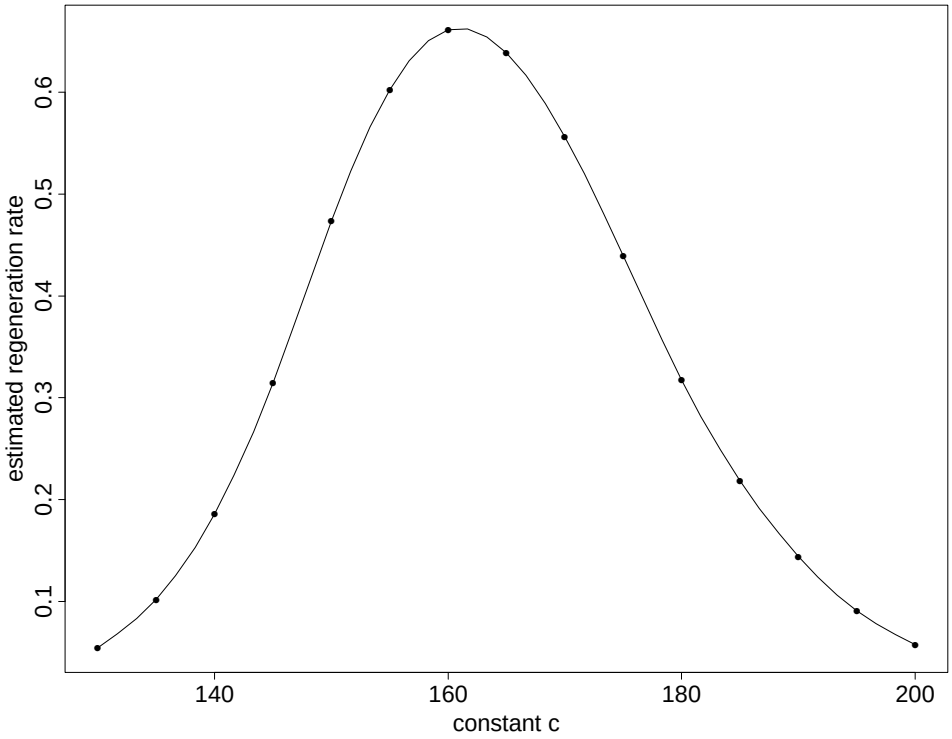


Figure 4.6: Expected regeneration rate versus the constant c' (4.50) for the Metropolis sampler with split Metropolis-rejected restarts for the Strauss process with 50 points $\beta = .126$. The horizontal coordinate is c' and the vertical coordinate is the estimated fraction of accepted restarts that will be regenerations.

Appendix A

Measure-theoretic Probability

A.1 Discrete, Continuous, and Other

A.1.1 Discrete

A *discrete probability space* consists of a finite or countable set S, called the *sample space*, and a nonnegative function p on S satisfying

$$\sum_{x \in S} p(x) = 1,$$

called the *probability mass function*. An *event* is a subset of S. For any event A the probability of A, written $P(A)$ is defined by

$$P(A) = \sum_{x \in A} p(x).$$

The map $A \mapsto P(A)$ is called the *probability measure* defined by p.

If g is a real-valued function on the sample space, then

$$E\{g(X)\} = \sum_{x \in S} g(x)p(x)$$

is called the *expectation* of the random variable $g(X)$, provided (in the case that S is not finite) that the summand on the right hand side is absolutely summable, so the order of summation does not matter.

Note that

$$P(A) = E\{1_A(X)\} \tag{A.1}$$

where 1_A denotes the so-called *indicator function* of the event A, defined by

$$1_A(x) = \begin{cases} 1, & x \in A \\ 0, & x \notin A \end{cases}$$

Hence the slogan

Probability is a trivial special case of expectation.

The set of all functions f for which expectations exist is denoted

A.1.2 Continuous

A *continuous probability space* consists of a nonnegative function f on some Euclidean space $\mathbb{R}^d$ satisfying

$$\int f(x)\,dx = 1,$$

called the *probability density function*. If g is a real-valued function on $\mathbb{R}^d$, then

$$E\{g(X)\} = \int g(x)p(x)\,dx$$

is called the *expectation* of the random variable $g(X)$, provided (in the case that S is not finite) that the integrand on the right hand side is absolutely integrable.

A.2 Measurable Spaces

Probability theory is a special case of a subject called measure theory. Both theories start with a set S. In probability theory, S is called the *sample space* or *state space*, the term we use when talking about Markov chains. In measure theory, S has no special name.

In elementary probability theory, subsets of S are called events, and probabilities are defined for all events, $P(B)$ is the probability of the event B. Thus a probability is a "set-function" $B \mapsto P(B)$ that takes subsets of S to real numbers. In general probability theory, more or less the same definition is used, but there is a problem, called the "problem of measure." The first issue any mathematical theory must deal with is whether the mathematical objects it woofs about (in this case probabilities) exist. Lacking such an existence theorem, there is no guarantee that the entire theory is not literally much ado about nothing, an elaborate discorse about elements of the empty set.

The set of all subsets of S is called the *power set* of S and is denoted $\mathcal{P}(S)$. With this notation, a "set-function" becomes an ordinary function. A probability is a map from $\mathcal{P}(S)$ to $\mathbb{R}$ that satisfies certain axioms, which will be met presently, as soon as we get past the problem of measure, which can now be stated: do there exist any maps $P : \mathcal{P}(S) \to \mathbb{R}$ that satisfy the axioms of probability? It is philosophically interesting that whether the existence problem has a solution depends on one's views on the foundations of mathematics. If one sticks to elementary set theory based on the so-called Zermelo-Frankel axioms, the problem of measure is a famous unsolved problem. If one adds to the elementary axioms the so-called *axiom of choice* then problem can be solved,

but the resolution is negative: there do not exist any probabilities on $\mathbb{R}$ or $\mathbb{R}^d$. But the axiom of choice has itself been the subject of vigorous debate for 100 years with no resolution of the argument about whether it should be included in the axioms. If instead of the axiom of choice one instead adds as an axiom *Cantor's continuum hypothesis*, one arrives at the same conclusion by another route, that probabilities on $\mathbb{R}$ or $\mathbb{R}^d$ do not exist. Thus we are left with the very unsettling conclusion that we aren't sure whether probabilities exist or not, but we certainly can't assert their existence.

The way to avoid the existence problem was found by Lebesgue, who proposed that instead of defining probabilities for all sets, we only define them for a family $\mathcal{B}$ of subsets of S. Thus a probability is a map $P : \mathcal{B} \to \mathbb{R}$ satisfying the axioms of probability. We need $\mathcal{B}$ to have two properties.

- $\mathcal{B}$ is large enough to be useful. It contains all of the events B for which we want to define $P(B)$.
- $\mathcal{B}$ is small enough to avoid the problem of measure. We need to be able to prove that probabilities exist.

It turns out that the right definition of $\mathcal{B}$ is the following. A family $\mathcal{B}$ of a set S is a *σ-field* if it satisfies the following axioms

Axiom 1. $S \in \mathcal{B}$.

Axiom 2. $B \in \mathcal{B}$ implies $B^c \in \mathcal{B}$.

Axiom 3. If B_1, B_2, ... are in $\mathcal{B}$ and $B_i \cap B_j = \varnothing$ when $i \neq j$, then $\bigcup_{i=1}^{\infty} B_i$ is in $\mathcal{B}$.

In words $\mathcal{B}$ contains the whole space S and is closed under complements and countable unions. Axioms 1 and 2 together imply $\varnothing \in \mathcal{B}$. Axioms 2 and 3 together with DeMorgan's laws

$$\left(\bigcap_{i=1}^{\infty} B_i\right) = \bigcup_{i=1}^{\infty} B_i^c$$

imply that $\mathcal{B}$ is also closed under countable intersections.

A *measurable space* is a pair $(S, \mathcal{B})$, where $\mathcal{B}$ is a σ-field for S. This definition is, of course, completely redundant because S is the largest element of $\mathcal{B}$, so knowing $\mathcal{B}$ tells you S. Thus the phrase "let $(S, \mathcal{B})$ be a measurable space" merely establishes notation. The point of the pairing is to establish both notations, S and $\mathcal{B}$ at once.

If $(S, \mathcal{B})$ is a measurable space, a *probability measure* on S is a map $P : \mathcal{B} \to \mathbb{R}$, satisfying

Axiom 1. $P(B) \geq 0$, for all $B \in \mathcal{B}$.

Axiom 2. $P(S) = 1$.

Axiom 3. If B_1, B_2, ... are in $\mathcal{B}$ and $B_i \cap B_j = \varnothing$ whenever $i \neq j$, then $P\left(\bigcup_{i=1}^{\infty} B_i\right) = \sum_{i=1}^{\infty} P(B_i)$.

Bibliography

Arcones, M. A. and B. Yu (1994). Central limit theorems for empirical and U-processes of stationary mixing sequences. *J. Theoret. Probab. 7*, 47–71.

Athreya, K. B. and P. Ney (1978). A new approach to the limit theory of recurrent Markov chains. *Trans. Amer. Math. Soc. 245*, 493–501.

Bartlett, M. S. (1946). On the theoretical specification and sampling properties of autocorrelated time-series. *Suppl J. Roy. Statist. Soc. 8*, 27–41.

Bélisle, C. J. P., H. E. Romeijn, and R. L. Smith (1993). Hit-and-run algorithms for generating multivariate distributions. *Math. Oper. Res. 18*, 255–266.

Bernardo, J. M. and A. F. M. Smith (1994). *Bayesian Theory*. New York: Wiley.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *J. Roy. Statist. Soc. Ser. B 36*, 192–236.

Besag, J. and P. Clifford (1989). Generalized Monte Carlo significance tests. *Biometrika 76*, 633–642.

Besag, J., P. Green, D. Higdon, and K. Mengersen (1995). Bayesian computation and stochastic systems (with discussion). *Statistical Science 10*, 3–66.

Billingsley, P. (1968). *Convergence of Probability Measures*. New York: Wiley.

Billingsley, P. (1979). *Probability and Measure*. New York: Wiley.

Bradley, R. C. (1986). Basic properties of strong mixing conditions. In E. Eberlein and M. S. Taqqu (Eds.), *Dependence in Probability and Statistics: A Survey of Recent Results (Oberwolfach, 1985)*, Boston, pp. 165–192. Birkhäuser.

Breiman, L. (1968). *Probability*. Reading, MA: Addison-Wesley.

Chambers, J. M. and T. J. Hastie (Eds.) (1993). *Statistical Models in S*. London: Chapman & Hall.

Chan, K. S. (1993). On the central limit theorem for an ergodic Markov chain. *Stochastic Process. Appl. 47*, 113–117.

Chan, K. S. and C. J. Geyer (1994). Discussion of the paper by Tierney. *Ann. Statist. 22*, 1747–1758.

Chen, M.-H. and B. Schmeiser (1993). Performance of the Gibbs, hit-and-run, and Metropolis samplers. *J. Comput. Graph. Statist. 2*, 251–272.

Chung, K. L. (1967). *Markov Chains with Stationary Transition Probabilities* (second ed.). Berlin: Springer-Verlag.

Daley, D. J. and D. Vere-Jones (1988). *An Introduction to the Theory of Point Processes*. New York: Springer-Verlag.

DeGroot, M. H. (1970). *Optimal Statistical Decisions*. New York: McGraw-Hill.

Devroye, L. (1986). *Non-Uniform Random Variate Generation*. New York: Springer-Verlag.

Doukhan, P., P. Massart, and E. Rio (1994). The functional central limit theorem for strongly mixing processes. *Ann. Inst. H. Poincaré Probab. Statist. 30*, 63–82.

Fristedt, B. and L. Gray (1997). *A Modern Approach to Probability Theory*. Boston: Birkhäuser.

Gaver, D. P. and I. G. O' Muircheartaigh (1987). Robust empirical Bayes analyses of event rates. *Technometrics 29*, 1–15.

Gelfand, A. E. and A. F. M. Smith (1990). Sampling-based approaches to calculating marginal densities. *J. Am. Statist. Assoc. 85*, 398–409.

Gelman, A., G. O. Roberts, and W. R. Gilks (1996). Efficient Metropolis jumping rules. In *Bayesian statistics, 5 (Alicante, 1994)*, pp. 599–607. New York: Oxford Univ. Press.

Geman, S. and D. Geman (1984). Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Machine Intell. 6*, 721–741.

Geyer, C. J. (1992). Practical Markov chain Monte Carlo (with discussion). *Statistical Science 7*, 473–511.

Geyer, C. J. and J. Møller (1994). Simulation and likelihood inference for spatial point processes. *Scand. J. Statist. 21*, 359–373.

Geyer, C. J. and E. A. Thompson (1995). Annealing Markov chain Monte Carlo with applications to ancestral inference. *J. Am. Statist. Assoc. 90*, 909–920.

Gordin, M. I. and B. A. Lifšic (1978). Central limit theorem for stationary Markov processes. *Soviet Math. Dokl. 19*, 392–394. English translation.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika 82*, 711–732.

Gut, A. (1988). *Stopped Random Walks*, Volume 5. New York: Springer-Verlag.

Halmos, P. R. (1958). *Finite-Dimensional Vector Spaces*. New York: Springer-Verlag.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika 57*, 97–109.

Ibragimov, I. A. and Y. V. Linnik (1971). *Independent and stationary sequences of random variables*. Groningen: Wolters-Noordhoff. English translation.

Jain, N. and B. Jamison (1967). Contributions to Doeblin's theory of Markov processes. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete 8*, 19–40.

Jänich, K. (1984). *Topology*. New York: Springer-Verlag.

Kendall, W. S., O. Barndorff-Nielsen, and M. C. van Lieshout (Eds.) (1998). *Trends in Stochastic Geometry*, London. Chapman & Hall.

Knuth, D. E. (1998). *Seminumerical Algorithms* (Third ed.), Volume 2 of *The Art of Computer Programming*. Reading, MA: Addison-Wesley.

Lang, S. (1987). *Linear Algebra* (Third ed.). New York: Springer-Verlag.

Liu, J., W. H. Wong, and A. Kong (1995). Covariance structure and convergence rate of the Gibbs sampler with various scans. *J. Roy. Statist. Soc. Ser. B 57*, 157–169.

Meketon, M. S. and B. W. Schmeiser (1984). Overlapping batch means: Something for nothing? In S. Sheppard, U. Pooch, and D. Pegden (Eds.), *Proceedings of the 1984 Winter Simulation Conference*, pp. 227–230.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller (1953). Equation of state calculations by fast computing machines. *J. Chem. Phys. 21*, 1087–1092.

Meyn, S. P. and R. L. Tweedie (1993). *Markov Chains and Stochastic Stability*. London: Springer-Verlag.

Mykland, P., L. Tierney, and B. Yu (1995). Regeneration in Markov chain samplers. *J. Amer. Statist. Assoc. 90*(429), 233–241.

Nummelin, E. (1978). A splitting technique for Harris recurrent Markov chains. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete 43*, 309–318.

Nummelin, E. (1984). *General Irreducible Markov Chains and Non-Negative Operators*. Cambridge: Cambridge University Press.

Pedrosa, A. C. and S. B. W. (1993). Asymptotic and finite-sample correlations between OBM estimators. In G. W. Evans, M. Mollaghasemi, E. C. Russell, and W. E. Biles (Eds.), *Proceedings of the 1993 Winter Simulation Conference*, pp. 481–488.

Peligrad, M. (1986). Recent advances in the central limit theorem and its weak invariance principle for mixing sequences of random variables (a survey). In E. Eberlein and M. S. Taqqu (Eds.), *Dependence in Probability and Statistics: A Survey of Recent Results (Oberwolfach, 1985)*, Boston, pp. 193–223. Birkhäuser.

Priestley, M. B. (1981). *Spectral Analysis and Time Series*. London: Academic Press.

Raymond, E. S. (Ed.) (1996). *The New Hacker's Dictionary* (Third ed.). Cambridge, MA: MIT Press. Also available on the World Wide Web as the *Jargon File* at `http://www.catb.org/~esr/jargon/`.

Ripley, B. D. (1979). Simulating spatial patterns: Dependent samples from a multivariate density. *Applied Statistics 28*, 109–112.

Ripley, B. D. (1987). *Stochastic Simulation*. New York: Wiley.

Roberts, G. O. and R. L. Tweedie (1996). Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika 83*, 95–110.

Robertson, T., F. T. Wright, and R. L. Dykstra (1988). *Order restricted statistical inference*. Chichester: John Wiley.

Rudin, W. (1987). *Real and Complex Analysis* (Third ed.). New York: McGraw-Hill.

Rudin, W. (1991). *Functional Analysis* (Second ed.). New York: McGraw-Hill.

Schervish, M. J. and B. P. Carlin (1992). On the convergence of successive substitution sampling. *J. Comput. Graph. Statist. 1*, 111–127.

Schmeiser, B. (1982). Batch size effects in the analysis of simulation output. *Oper. Res. 30*, 556–568.

Sheehan, N. and A. Thomas (1993). On the irreducibility of a Markov chain defined on a space of genotype configurations by a sampling scheme. *Biometrics 49*, 163–175.

Stromberg, K. R. (1981). *An Introduction to Classical Real Analysis*. Belmont, CA: Wadsworth.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *Ann. Statist. 22*, 1701–1762.

Wei, G. C. G. and M. A. Tanner (1990). A Monte Carlo implementation of the EM algorithm and poor man's data augmentation. *J. Am. Statist. Assoc. 85*, 699–704.