Markov Chain Monte Carlo Lecture Notes

Charles J. Geyer

Copyright 1995, 2005 by Charles J. Geyer

Course notes originally used Spring Quarter 1995
Last changed: November 21, 2005
Last typeset: November 21, 2005

Contents

Chapter 1

Introduction

1.1 Monte Carlo

Classical Monte Carlo involves the notion of learning about a probability distribution by simulating independent identically distributed realizations from it. Suppose we are interested in a probability distribution π and cannot do any pencil and paper calculations about it but can simulate in a computer a sequence X_1, X_2, ... of independent identically distributed realizations having this distribution. Then we can estimate expectations

$$E_\pi g(X) = \int g(x)\pi(dx) \tag{1.1}$$

of interesting functions g. The notation on the right hand side in (1.1) is not intended to intimidate anyone or warn of an impending onset of measure theory. It is just used as the appropriate shorthand sums or integrals or a combination of the two as the case may be. If X is a discrete random variable,

$$E_\pi g(X) = \int g(x)\pi(dx) = \sum_{\substack{\text{all possible} \\ x \text{ values}}} g(x)p(x),$$

where $p(x)$ is the probability mass function of the distribution π. If X is a continuous random variable,

$$E_\pi g(X) = \int g(x)\pi(dx) = \int g(x)p(x)\,dx,$$

where $p(x)$ is the probability density function of the distribution π. If X is multivariate these will be multiple sums or multiple integrals or perhaps sums and integrals if X has some discrete components and some continuous components.

In any case, although we can write down (1.1) we cannot actually evaluate the sums or integrals and so cannot actually calculate the expectation. Introduce the Greek letter $\mu = E_\pi g(X)$ for the quantity we want to calculate. Then

μ is the mean of the random variable $Y = g(X)$. If we can simulate a sequence X_1, X_2, ..., of independent realizations of the random process X having distribution π, then the random variables $Y_i = g(X_i)$ are independent and identically distributed with distribution π. Thus

$$\hat{\mu}_n = \bar{Y}_n = \frac{1}{n} \sum_{i=1}^{n} Y_i = \frac{1}{n} \sum_{i=1}^{n} g(X_i),$$

the sample mean of the Y_i, is an unbiased estimator of μ and $\hat{\mu}_n$ converges almost surely to μ as $n \to \infty$ by the strong law of large numbers. Moreover, if the variance of the random variable $Y = g(X)$ is finite, then the central limit theorem applies

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

where $\xrightarrow{\mathcal{D}}$ indicates convergence in distribution. Our Monte Carlo approximation $\hat{\mu}_n$ is asymptotically normally distributed with mean μ, the quantity to be calculated and standard deviation $\sigma/\sqrt{n}$, where σ^2 is the variance of the random variable Y

$$\sigma^2 = \int [g(x) - \mu]^2 \pi(dx).$$

Of course we usually do not know σ, but it can be consistently estimated from the simulations themselves by the sample standard deviation of the Y_i

$$s = \frac{1}{n-1} \sum_{i=1}^{n} [Y_i - \hat{\mu}_n]^2 = \frac{1}{n-1} \sum_{i=1}^{n} [g(X_i) - \hat{\mu}_n]^2$$

The Monte Carlo approximation $\hat{\mu}_n$ is a point estimate of of the quantity μ, which is the expectation to be calculated. The approximate error in the Monte Carlo approximation is $s/\sqrt{n}$. The nice thing for statisticians about ordinary Monte Carlo is that we already understand the theory. It is just elementary statistics.

1.2 Problems with Ordinary Monte Carlo

The main problem with ordinary Monte Carlo is that it is very hard to do for multivariate stochastic processes. A huge number of methods exist for simulating univariate random variables. Devroye (1986) is the definitive source. Ripley (1987) is more introductory but is authoritative as far as it goes. Knuth (1973) is also authoritative, though a bit dated.

There are a few tricks for reducing multivariate problems to univariate problems. A general multivariate normal random vector $X \sim N(\mu, \Sigma)$ can be simulated using the Cholesky decomposition of the dispersion 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 X_1 from its marginal distribution, simulate the second component X_2 from its conditional distribution given X_1, then simulate X_3 from its conditional distribution given X_1 and X_2, 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 stochastic processes. Something better is needed.

1.3 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." Some Markov chain literature (Chung, 1967, for example) uses "Markov chain" to refer to a discrete time *or continuous time* stochastic process on a *countable* state space. Much of the modern literature (Nummelin, 1984 or Meyn and Tweedie, 1993, for example) as well as all of the Markov chain Monte Carlo (MCMC) literature follows the usage adopted here.

A Markov chain is a discrete time stochastic process X_1, X_2, ... taking values in an arbitrary state space having the property that the conditional distribution of X_{t+1} given the past X_1, ..., X_t depends only on the present state X_t. 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_{t+1} given X_t is the same for all t.

1.3.1 Markov Chain Monte Carlo

There are stochastic processes more general than Markov chains that one might think would be useful for Monte Carlo, but this is not so because any computer program used for simulation is a Markov chain if one defines the state space properly. Consider a program

initialize x
for $i = 1, \ldots, n$ **do**
  update x
  output x
end

where x denotes the vector containing all of the variables the computer program uses except the loop index i and the step "update x" does not refer to the loop

index i. The program starts by defining some initial values for all the variables. Then each time through the loop the program makes some possibly random change in some or all of the variables and prints out the current values. The output is a Markov chain because the probability distribution of x when $i = t$ only depends on the value of x in the preceding iteration when $i = t-1$ because earlier values of x have been overwritten and are no longer available. The transition probabilities are stationary because the same computer code is used in the step "update x" each time through the loop and this code does not refer to the "time" i.

The program would not necessarily be a Markov chain if the program did not print out the values of all variables known to the program except the loop index i, but there is no reason not to include all of the variables in the computer program in the state space of the Markov chain. The transition probabilities would not be stationary if the step "update x" depended on i, but this conflicts with the notion of trying to simulate realizations of a fixed probability distribution π.

As we shall see, both the law of large numbers and the central limit theorem also hold for Markov chain under certain conditions. Thus the same principles used in ordinary Monte Carlo can be applied to Markov chain Monte Carlo.

1.3.2 Transition Probabilities

A Markov chain is defined by defining its *transition probabilities*. For a discrete state space S, these are specified by defining a matrix

$$p(x, y) = \Pr\{X_{t+1} = y | X_t = x\}, \qquad x, y \in S$$

that gives the probability of moving from any point x at time t to any other point y at time $t+1$. Because of the assumption of stationary transition probabilities, the transition probability matrix $p(x, y)$ does not depend on the time t. For a general state space S the transition probabilities are specified by defining a *kernel*

$$P(x, B) = \Pr\{X_{t+1} \in B | X_t = x\}, \qquad x \in S,\ B \text{ a measurable set in } S.$$

The kernel is defined so that 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 on S. In other words, the kernel is a *regular* conditional probability.

The 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 marginal distribution of X_1, which is called the *initial distribution* of the chain.

For those who like to keep track of measure-theoretic niceties, there is one technical condition always imposed on a general state space, that it be *countably generated*, meaning the σ-field is generated by a countable family of sets. This is required for some of the results of the modern theory of Markov chains that

make chains on general state spaces little more difficult than those on countable state spaces. In practice this is not restrictive. Any Euclidean space R^d or, more generally, any separable metric space is countably generated.

1.3.3 Stationary Distributions

We say that a probability distribution π is a *stationary* distribution or an *invariant* distribution for the Markov chain if it is "preserved" by the transition probability, that is if the initial distribution is π, then the marginal distribution of X_2 is also π. Hence so is the marginal distribution of X_3 and all of the rest the chain.

In the discrete case π is specified by a vector $\pi(x)$, and the stationary property is

$$\pi(y) = \sum_{x \in S} \pi(x) p(x, y). \tag{1.2}$$

For those who like to think of the transition probabilities as a matrix P with entries $p(x, y)$, (1.2) can be rewritten $\pi = \pi P$, since the right hand side of (1.2) is the multiplication of the matrix P on the left by the row vector π. This association with matrix multiplication does not buy much, because it does not extend to general state spaces.

For general state spaces the stationary property is

$$\pi(B) = \int \pi(dx) P(x, B). \tag{1.3}$$

(1.2) and (1.3) are the same except that a sum over a discrete state space has been replaced by an integral over a general state space. We sometimes use the same notation $\pi = \pi P$ to refer to (1.3) as well as (1.2) but it no longer refers to matrix multiplication. To further the analogy, we define multiplication on the left of a kernel P by an arbitrary positive measure ν by

$$(\nu P)(B) = \int \nu(dx) P(x, B).$$

This makes rigorous our calling the right hand side of (1.3) πP.

1.3.4 Law of Large Numbers

In MCMC we always use a Markov chain that is constructed to have a specified stationary distribution π, so there is never any question as to whether a stationary distribution exists—it does so by construction. There may be a question about whether the stationary distribution is unique, about whether (1.2) or (1.3) have any other solutions π other than the distribution used in the construction. If the stationary distribution is not unique, the chain is useless for Monte Carlo. Later on we will learn that a condition guaranteeing the uniqueness of the stationary distribution is *irreducibility* for chains on discrete state spaces and so-called *φ-irreducibility* for chains on general state spaces.

Under the same conditions, irreducibility or φ-irreducibility that guarantee uniqueness of the stationary distribution, there is a law of large numbers for the Markov chain. Define

$$\mu = E_\pi g(X)$$

and

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

just as we did in ordinary Monte Carlo. For a φ-irreducible Markov chain, conditional on the starting position $X_1 = x$, $\hat{\mu}_n$ converges almost surely to μ for π-almost all x. Under further regularity conditions the "π-almost all x" can be strengthened to "all x". This will be discussed further in Section 3.2. For now the point is that, having constructed a Markov chain with a unique stationary distribution π, averages over a simulation of the chain can be used to approximate expectations with respect to π, just as in ordinary Monte Carlo.

1.3.5 Operating on Functions

Although we don't need the notation immediately, it seems worth mentioning here the other way a transition probability can "multiply" a vector. If g is any non-negative function, the conditional expectation of $g(X_t)$ given $X_{t-1} = x$ is another non-negative function, which we denote Pg. If the state space is discrete

$$(Pg)(x) = E\{g(X_t) | X_{t-1} = x\} = \sum_{y \in S} p(x, y) g(y),$$

and, as the notation suggests, the vector Pg is obtained by multiplying the matrix P on the right by the column vector g.

For a general state space, the sum becomes an integral

$$(Pg)(x) = E\{g(X_t) | X_{t-1} = x\} = \int P(x, dy) g(y). \tag{1.4}$$

For discrete state spaces, both of these operations are matrix multiplication, the only distinction is that we multiply P on the left by a row vector μ and on right by a column vector g. Only the probabilistic interpretation of these operations tells us that we should consider μ a measure on the state space and g a function on the state space.

With a general state space, the distinction is clear. Because a kernel $P(x, A)$ has different kinds of arguments, x a point and A a set, we must have μ a measure and g a function, and these two kinds of mathematical objects cannot be confused as the "matrix multiplication" point of view invites.

We defined the operation Pg above only when g was a non-negative function. Because both the kernel and the function are non-negative, the result is always well-defined, although we may have $(Pg)(x)$ equal to infinity for some (or even all x).

We can also define P to operate on a different class of functions, the Hilbert space $L^2(\pi)$ where π is a stationary distribution for the kernel P. For complex-valued measurable functions u and v on the state space, define the inner product

$$(u, v) = \int u(x)\bar{v}(y)\pi(dx)$$

where $\bar{v}$ denotes the complex conjugate of v, and the norm $\|u\|$ by

$$\|u\|^2 = (u, u) = \int |u(x)|^2\pi(dx). \tag{1.5}$$

Then $L^2(\pi)$ is the space of function u for which (1.5) is finite. The Cauchy-Schwarz inequality $|(u, v)| \le \|u\| \, \|v\|$ guarantees the finiteness of the inner product.

If u is an element of $L^2(\pi)$, then so is Pu by Jensen's inequality

$$\begin{aligned}
\|Pu\|^2 &= \int \left| \int P(x, dy)u(y) \right|^2 \pi(dx) \\
&\le \iint \pi(dx)P(x, dy)u(y)^2 \\
&= \int \pi(dy)u(y)^2 \\
&= \|u\|^2
\end{aligned}$$

This allows us to interpret Pu when u is an arbitrary non-negative function or when u is not non-negative but is an element of $L^2(\pi)$.

The norm of a linear operator P on $L^2(\pi)$ is defined by

$$\|P\| = \sup_{u \in L^2(\pi)} \frac{\|Pu\|}{\|u\|}$$

The calculation above says $\|P\| \le 1$. Since $Pu = u$ for constant functions, we have $\|P\| = 1$. We are not interested in P operating on constant functions, because this is trivial, so we often restrict the domain of definition to the space of functions with mean zero

$$L_0^2(\pi) = \left\{ u \in L^2(\pi) : \textstyle\int u \, d\pi \right\}.$$

The norm of P considered to be an operator on the subspace $L_0^2(\pi)$ is still less than or equal to one. Because $L_0^2(\pi)$ does not contain any constant functions, the norm may be strictly less than one.

Chapter 2

Basic Algorithms

This section 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 (and the corresponding) code that change the state variable of the Markov chain in such a way so as to preserve the stationary distribution. Some of these updates can be used to simulate a Markov chain by repeating the update again and again. Some cannot be used that way, because the resulting Markov chain would not be irreducible, but these elementary update steps can be combined in various ways to make combined update steps that are useful. The two types of basic update step are the Gibbs update, the basic component of the so-called "Gibbs sampler," and the Metropolis-Hastings update, the basic component of the so-called "Metropolis-Hastings algorithm."

The Gibbs update is actually a special case of the Metropolis-Hastings update, so the "Gibbs sampler" is actually a special case of the "Metropolis-Hastings algorithm," but because it has gotten so much attention we will start with the Gibbs update.

2.1 The Gibbs Update

The rationale of the Gibbs update is very simple. The state variable of the system is a vector $x = (x_1, \ldots, x_d)$. In this section, subscripts indicate components of the state vector rather than time. An elementary Gibbs update changes only one component of the state vector, say x_i. This component is given a new value which is a realization from its conditional distribution given the rest $\pi(x_i|x_{-i})$ where $x_{-i} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_d)$ is conventional shorthand for the rest of the components of x besides x_i and the conditional distribution is the one derived from the desired stationary distribution π of the Markov chain. These one-dimensional conditional distributions of one component given the rest are often called "full conditionals" in the Gibbs sampling literature and "local characteristics" in the spatial statistics literature.

It is clear that this preserves the stationary distribution. If the current state x is a realization from π, then x_{-i} is distributed according to its marginal $\pi(x_{-i})$ derived from π and the state after the update will have the distribution

$$\pi(x_i|x_{-i})\pi(x_{-i}) \tag{2.1}$$

which is $\pi(x)$ by definition of conditional probability: joint equals conditional times marginal.

2.2 Combining update mechanisms

2.2.1 Composition

In order to get a useful Markov chain sampler, one must often combine elementary update mechanisms. There are two ways to combine update mechanisms: composition and mixing. Composition follows one update mechanism by another; the computer code for the first is executed, then the computer code for the second. If each step preserves the stationary distribution, then so does the combination. One reason why we are calling this composition of update mechanisms is that the procedure of following one bit of computer code by another is also composition of functions in an abstract sense. If one bit of code produces output $f(x)$ given input x and another bit produces output $g(y)$ given input y, then the first bit followed by the second produces output $g(f(x))$ given input x.

Hence if one has d different elementary update mechanisms U_1, ..., U_d, which we may think of as different bits of computer code or perhaps one bit of computer code that does d slightly different operations for the d different values of some index variable, we may combine them in a composite update $U_1 \cdots U_d$ that executes them in sequence, either by simply placing one bit of code after another in the program or by putting one bit of code in a loop

$$\textbf{for}(i \in 1, \ldots, d)\, \textbf{do}\, U_i$$

This procedure is also composition of kernels when we think of the kernels as operators on a function space. Suppose the corresponding transition probability kernels for these update steps are P_1, ..., P_d then the kernel for the combined update is $P_1 \cdots P_d$ where as usual in operator theory this means composition of operators. For any function $g \in L^2(\pi)$

$$(P_1 \cdots P_d)g = P_1(P_2 \cdots (P_d g) \cdots),$$

that is, the result of applying P_d to g yielding the $L^2(\pi)$ function $P_d g$, then applying P_{d-1} to that function, and so forth.

2.2.2 Multiplication of Kernels

In order to see why composition preserves stationary distributions, we need to look at the general definition of "multiplication of kernels." Suppose X has the

distribution μ, the conditional distribution of Y given X is given by the kernel P and the conditional distribution of Z given Y is given by the kernel Q. Then the joint distribution of X, Y, and Z is given by

$$\Pr(X \in A, Y \in B, Z \in C) = \int_A \int_B \mu(dx) P(x, dy) Q(y, C)$$

If we specialize to the case where $\mu = \delta_x$ the distribution concentrated at the point x and let B be the whole state space we get

$$\Pr(X = x, Z \in C) = \int P(x, dy) Q(y, C)$$

and this is another kernel, which is denoted by the notation PQ

$$(PQ)(x, A) = \int P(x, dy) Q(y, A).$$

An important special case arises when $P = Q$. For a Markov chain with stationary transition probabilities, the same kernel is used at each step and

$$\Pr(X_{t+n} \in A | X_t = x) = P^n(x, A)$$

the n-fold product of the transition probability kernel P. These are called the n-step transition probabilities. Because we can take marginals in any order, we have the so-called Chapman-Kolmogorov equations

$$P^{m+n}(x, A) = \int P^m(x, dy) P^n(y, A).$$

These hold even for m or n equal to zero if we define

$$P^0(x, A) = I(x, A) = \begin{cases} 1, & x \in A \\ 0, & x \notin A \end{cases} \tag{2.2}$$

to be the so-called identity kernel.

In the case of a discrete state space, when P is a matrix with entries $p(x, y)$, the composition PQ is matrix multiplication. PQ is the matrix with entries

$$\sum_y p(x, y) q(y, z).$$

The identity kernel I becomes the identity matrix. The n-step transition probabilities are the elements of the matrix P^n.

So far, our only examples of different update mechanisms are the d different Gibbs update steps for a d-dimensional state vector. If the kernel P_i denotes updating x_i from its conditional distribution given the rest, then the composite kernel $P_1 \cdots P_d$ denotes updating x_1, x_2, ..., x_d in that order. There always is a kernel corresponding to a Gibbs update because of the assumption of a countably generated state space, which assures the existence of regular conditional probabilities.

Now it is easy to see that composition of updates preserves a desired stationary distribution. If P_1, P_2, and P_3 all have the same stationary distribution π, then

$$\pi(P_1P_2P_3) = ((\pi P_1)P_2)P_3 = (\pi P_2)P_3 = \pi P_3 = \pi,$$

and similarly for the general case of d kernels.

2.2.3 Mixing

Mixing chooses at random amongst update mechanisms. If *runif*() generates uniform random variates strictly between zero and one and $U_1, \ldots, U_d$ are update mechanisms, possibly composite update mechanisms, then the following code combines the updates by mixing.

$$\begin{array}{l} u = \mathit{runif}() \\ i = \lceil d * u \rceil \\ U_i \end{array}$$

where $\lceil x \rceil$ denotes the smallest integer greater than or equal to x. Note that the code fails if u can be zero. If the random number generator can return zero, it is necessary to replace the first statement by

$$\textbf{repeat}\, u = \mathit{runif}()\, \textbf{until}(u > 0)$$

The joint probability distribution of the index I of the update chosen and the result Y of the update is $\frac{1}{d}P_i(x, dy)$, hence the kernel of the mixture update is obtained by marginalizing out I

$$\Pr(X_{t+1} \in A | X_t = x) = \frac{1}{d}\sum_{i=1}^{d} P_i(x, A)$$

or, more concisely, the kernel of the mixture update is the convex combination of elementary kernels

$$P = \frac{1}{d}\sum_{i=1}^{d} P_i \tag{2.3}$$

Many authors have noted that there is no reason to choose the elementary updates with equal probabilities. If update P_i was chosen with probability q_i then the mixture kernel would be

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

a general convex combination. However, little work has been done on choosing the q_i and the gains in efficiency resulting from using unequal mixing probabilities are likely to be small. So this is rarely used in actual practice.

2.2.4 Combining Composition and Mixing

Composition and mixing are the only ways to combine kernels, since multiplication and convex combination are the only operations that make 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 this is the so-called *random sequence scan* that combines d elementary update mechanisms by choosing a random permutation $(i_1, i_2, \ldots, i_d)$ of the integers 1, 2, ..., d and then applying the updates P_{i_j}, $j = 1, \ldots, d$ in that order. An efficient procedure for producing a random permutation is given by Knuth (1973, p. 139). So if $\mathcal{P}$ denotes the set of all $d!$ permutations, the kernel of this scan is

$$P = \frac{1}{d!} \sum_{(i_1,\ldots,i_d)\in\mathcal{P}} P_{i_1} \cdots P_{i_d}. \tag{2.4}$$

2.2.5 Reversibility

The transition probability kernel P of a Markov chain having a stationary distribution π (or the Markov chain itself) is said to be *reversible* if when X_t has the distribution π then X_t and X_{t+1} are exchangeable random variables, that is the pair (X_t, X_{t+1}) has the same joint distribution as the pair (X_{t+1}, X_t). Of course, the fact that π is a stationary distribution implies that X_t and X_{t+1} have the same marginal distribution (namely, π), but reversibility is a much stronger property that imposes severe restrictions on the joint distribution of X_t and X_{t+1} (and hence on the kernel P).

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 every elementary update mechanism that has so far been proposed for MCMC is reversible. The reason for this is the only simple way to show that an update mechanism has a specified stationary distribution is to show that it is reversible with respect to that stationary distribution. Hence the only way that 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.

Let us confirm that the Gibbs update is reversible. In order to avoid confusion about subscripts, let X denote the state at time t and Y the state at time $t + 1$ and subscripts again denote components. To prove reversibility we need to show that if X has the distribution π, then X and Y are exchangeable. Suppose we are updating the ith component of the state vector, then $X_{-i} = Y_{-i}$ since the rest of the components are not changed and both have the marginal $\pi(x_{-i})$ under the stationary distribution. Conditional on these variables X_i has the conditional distribution $\pi(x_i|x_{-i})$ and so does Y_i and these variables are conditionally independent given the rest. Since this description is obviously symmetric in X and Y these variables are exchangeable and the Gibbs update is reversible.

Another way to think about reversibility is that the Markov chain looks the same running forwards or backwards if started in the stationary distribution. The kernel for the "time reversed" Markov chain that relabels the variables so time runs backward is that same as the kernel of the original chain. So a way to verify reversibility is to find the kernel of the time reversed chain and see if it is the same as the original.

Now suppose that we have d elementary update mechanisms with kernels P_i that are reversible but not necessarily Gibbs and see whether composition and mixing preserve reversibility. Again for composition, let us do the special case of composing three kernels, the general case being similar. If X_t is the current state and X_{t+1} is the result of applying the composite update $P_1P_2P_3$, let Y and Z be the intermediate states, Y resulting from applying P_1 to X_t and Z resulting from applying P_2 to Y and X_{t+1} resulting from applying P_3 to Z, then the joint distribution of all four variables is

$$\begin{aligned}\Pr(X_t \in A, Y \in B, Z \in C, X_{t+1} \in D)& \\ = \int_A \int_B \int_C \pi(dx_t) P_1(x_t, dy) P_2(y, dz) P_3(z, D),& \quad (2.5)\end{aligned}$$

recalling that in checking reversibility we assume that X_t has the stationary distribution π and hence so do the other three variables. By reversibility of P_3 we know that the conditional distribution of Z given X_{t+1} is also given by the kernel P_3. Similarly the conditional distribution of Y given Z has the kernel P_2 and that of X_t given Y has P_1. So we can also write (2.5) as

$$\int_D \int_C \int_B \pi(dx_{t+1}) P_3(x_{t+1}, dz) P_2(z, dy) P_1(y, A),$$

from which we see that

$$P(X_t \in A | X_{t+1} = x) = \iint P_3(x, dz) P_2(z, dy) P_1(y, A)$$

and that the kernel for the time reversed chain is $P_3P_2P_1$. Since there is no reason in general why $P_1P_2P_3$ should be equal to $P_3P_2P_1$, combining updates by composition does not in general yield a reversible chain.

What happens when we combine by mixing? Now we need to consider the joint distribution of X_t, X_{t+1}, and the index I of the update applied. We have

$$\Pr(X_t \in A, X_{t+1} \in B, I = i) = \frac{1}{d} \int_A \pi(dx) P_i(x, B).$$

By reversibility of P_i this can also be written

$$\frac{1}{d} \int_B \pi(dx) P_i(x, A).$$

Summing over the d possible values of i, we see that both the original and the time reversed kernels have the same form (2.3).

Thus mixing maintains reversibility, but composition in general does not. What about the combination of mixing and composition discussed in the preceding section? The time reversal of the kernel (2.4) is the same because for each term $P_{i_1} \cdots P_{i_d}$ that appears in the sum the reversal $P_{i_d} \cdots P_{i_1}$ also appears. Although the individual terms are not reversible, the sum is reversible.

2.2.6 More on Combining Composition and Mixing

The three ways of combining elementary update mechanisms have a number of drawbacks if one wants to preserve reversibility. Pure composition does not preserve reversibility unless one uses a composite mechanism that performs some elementary updates more than once. $P_1P_2P_3P_3P_2P_1$ is reversible because it reads the same forwards or backwards, but this doubles the amount of work per step of the Markov chain, and it also has a very curious property when the elementary updates are Gibbs. For Gibbs updates $P_3P_3 = P_3$. Updating the same variable twice in a row is like doing it only once, because the distribution of the update only depends on the rest of the variables, which remain unchanged. Thus this update is the same as $P_1P_2P_3P_2P_1$, and variable 3 is effectively updated only half as frequently as variable 2. Worse, effectively the same thing happens with variable 1. If we look at several steps we have, for example

$$P^3 = P_1P_2P_3P_2P_1P_2P_3P_2P_1P_2P_3P_2P_1$$

so in effect variable 1 is also only updated half as often as variable 2. But we cannot dispense with either of the updates of variable 1 and maintain reversibility. One of the updates of variable 1 is essentially wasted, but cannot be avoided.

The "random sequence scan" has the same problem whenever the last elementary update performed in the preceding scan is the same as the first elementary update performed in the next, which happens with probability $1/d$ when there are d elementary updates. The simple mixture has the same problem with the same probability $1/d$, and it also has the drawback that it does not perform each elementary update once in d iterations.

Thus for maintaining reversibility, none of the methods of combining elementary updates considered so far are satisfactory. In order to do better we need a new idea. We can let the choice of the next update depend on which update has just been done. We could choose in each iteration to do the update with kernel $P_{i_1} \cdots P_{i_d}$ where $(i_1, \ldots, i_d)$ is a permutation chosen uniformly at random from among the $(d-1) \times (d-1)!$ permutations such that i_1 is not the same as the i_d used in the immediately preceding iteration. Even more simply we could choose in each iteration choose the permutation $(i_1, \ldots, i_d)$ uniformly at random from among the $2(d-1)$ permutations that cycle through the integers in normal or reversed order and do not start with the same update just done. With four variables these permutations are

1234	2341	3412	4123
4321	3214	2143	1432

If the last update in the preceding iteration was P_3, we use one of the six permutations that do not start with 3. This second method has the advantage of using fewer random variates to decide which permutation to use, only two per iteration, one to decide whether to cycle forward or backward and one to decide which update to start with. The first method would need $d - 1$ random variates to generate a random permutation.

Neither of these two methods makes a Markov chain unless we augment the state space of the Markov chain. Since the transition mechanism now depends on the index I of the last elementary update done in the preceding iteration, that must be part of the state. Hence the state variable of the Markov chain is now the pair (X, I) where X takes values in the original state space. And what is the stationary distribution? It can no longer be π because π lives on the original state space, not this augmented one. It turns out that the stationary distribution has X and I independent with X having the distribution π and I being uniform on the integers 1 to d. Why? Clearly for either of the two schemes, if I is uniformly distributed before an update, it is also uniformly distributed after the update, because both schemes treat all the indices the same. Also no matter which permutation is performed X_{t+1} has the distribution π if X_t has the distribution π.

In the case $d = 2$ this method behaves very counterintuitively. There are two possible fixed scan orders P_1P_2 and P_2P_1. In general these will not be equal and the fixed scan sampler is not reversible. If we use either of the schemes just described, we choose one of these two scan orders for the first iteration by flipping a coin. Then in each succeeding iteration, there is no freedom left. If we used P_1P_2 for the first iteration, then we must also use P_1P_2 the second to avoid updating variable 2 twice in succession, and for the same reason we must use P_1P_2 in every iteration. Thus this form of random scan has no randomness except for one coin flip at the beginning. In effect use one fixed scan sampler or the other and decide which one to use by a coin flip. Neither fixed scan sampler is reversible, but a random choice among them is reversible.

What is counterintuitive about this situation is that whether a chain is reversible or not does not depend on the initial distribution. Since the result of the coin flip needs to be part of the state of the Markov chain, we can use an initial distribution concentrated on the coin flip being heads. Thus the chain is not reversible if we decide to use the fixed scan order P_1P_2, but is reversible if we imagine a coin flip that comes up heads and then use the fixed scan order P_1P_2. Whether the chain is reversible or not depends on an entirely imaginary coin flip. The actual computer simulations are the same in both cases. This paradox tells us that a fixed scan chain that combines two elementary update steps is essentially reversible, even though it does not appear so at first sight.

These schemes illustrate a general principle that will be used again and again in MCMC. It is often useful to let the update mechanism in an MCMC simulation depend on additional variables other than those in the original statement of the problem. This is fine so long as one adds these additional variables to the state space of the Markov chain and is careful to consider the stationary distribution of the Markov chain on this augmented state space. As long as

the distribution of interest on the original state space is easily derived from the stationary distribution on the augmented space, a marginal or conditional, for example, the chain will still be useful for simulation.

Other schemes that let the method of combining updates depend on the current state are possible, but a general analysis of how to do that will have to wait until we have studied the Metropolis-Hastings algorithm.

2.3 The Gibbs Sampler

The term "Gibbs sampler" refers to a MCMC scheme in which all of the elementary update steps are Gibbs updates. The elementary updates are usually combined by composition $P_1 \cdots P_d$ making a so-called "fixed scan" Gibbs sampler, "scan" referring to running through the variables updating them in order. The simple mixture method of combination is usually called a "random scan" or "simple random scan" Gibbs sampler. As mentioned above, the combination using all $d!$ random permutations has been called a "random sequence scan."

The name "Gibbs sampler" was coined by Geman and Geman (1984). The algorithm itself is a special case of the Metropolis-Hastings algorithm specifically noted by Hastings (1970) and had been used even earlier in the physics literature, where it is called the "heat bath" algorithm. Another use in statistics prior to its naming was Ripley (1979). After several years of use in spatial statistics, particularly for Bayesian image reconstruction, the subject of Geman and Geman (1984), the use of the Gibbs sampler for general Bayesian inference was popularized by Gelfand and Smith (1990).

2.4 Bayesian Analysis of a Variance Components Model

The following example comes from Gelfand and Smith (1990), a Bayesian analysis of a simple variance component model. Suppose data y_{ij} are observed and are assumed to have distribution

$$y_{ij} \sim \text{Normal}\left(\theta_i, \frac{1}{\lambda_e}\right), \qquad i = 1, \ldots, K, \quad j = 1, \ldots, J,$$

and the group means θ_i are assumed to have distribution

$$\theta_i \sim \text{Normal}\left(\mu, \frac{1}{\lambda_\theta}\right), \qquad i = 1, \ldots, K.$$

A frequentist would take the parameters μ, λ_θ, and λ_e to be unknown constants, a Bayesian treats them as random quantities with prior distributions. In order for Gibbs sampling to be possible it is necessary that the priors have simple forms so that the one-dimensional conditionals of the posterior be known. Here

we take conjugate priors

$$\begin{aligned}\mu &\sim \text{Normal}\left(\mu_0, \frac{1}{\lambda_0}\right)\\ \lambda_\theta &\sim \text{Gamma}\,(a_1, b_1)\\ \lambda_e &\sim \text{Gamma}\,(a_2, b_2)\end{aligned}$$

The six hyperparameters μ_0, λ_0, a_1, b_1, a_2, and b_2 are assumed to be known, chosen so that the prior distributions represent one's a priori opinion about the parameters. The problem to be solved by Gibbs sampling is to obtain samples from the posterior distribution of the parameters given the data, which it does by simulating the joint distribution of the parameters μ, λ_θ, and λ_e and the random effects θ_i given the data.

The joint distribution of the data, the random effects and the parameters can then be written down. The unnormalized density (ignoring multiplicative constants) is

$$\begin{aligned}h(\theta_1, \ldots, \theta_K, \mu, \lambda_\theta, \lambda_e) = \lambda_e^{JK/2} e^{-\frac{\lambda_e}{2}\sum_{ij}(y_{ij}-\theta_i)^2} \lambda_\theta^{K/2} e^{-\frac{\lambda_\theta}{2}\sum_i(\theta_i-\mu)^2}\\ \times e^{-\frac{\lambda_0}{2}(\mu-\mu_0)^2} \lambda_\theta^{a_1-1} e^{-b_1\lambda_\theta} \lambda_e^{a_2-1} e^{-b_2\lambda_e} \quad (2.6)\end{aligned}$$

Not only is this the unnormalized joint density of the data, random effects, and parameters, it is also an unnormalized conditional density of any set of variables given the others. If we want the conditional density of λ_θ given the rest of the variables, it is given, up to a multiplicative constant, by (2.6) considered as a function of λ_θ, the rest of the variables being held fixed. Looking only at the factors involving λ_θ and collecting terms, the unnormalized density simplifies to

$$\lambda_\theta^{a_1+K/2-1} e^{-\left[b_1+\frac{1}{2}\sum_i(\theta_i-\mu)^2\right]\lambda_\theta}$$

and we see that considered as a function of λ_θ, this is of the form $\lambda_\theta^{a-1}e^{-b\lambda_\theta}$ with $a = a_1 + K/2$ and $b = b_1 + \frac{1}{2}\sum_i(\theta_i - \mu)^2$. Since this is the functional form of a gamma density with parameters a and b, the conditional distribution of λ_θ given the rest of the variables is Gamma(a, b), which we abbreviate to

$$\lambda_\theta|\text{rest} \sim \text{Gamma}\left(a_1 + K/2, b_1 + \frac{1}{2}\sum_i(\theta_i - \mu)^2\right)$$

Similarly

$$\lambda_e|\text{rest} \sim \text{Gamma}\left(a_2 + JK/2, b_2 + \frac{1}{2}\sum_{ij}(y_{ij} - \theta_i)^2\right)$$

To obtain the conditional density of μ given the rest, we look at factors containing μ and use the identity

$$\sum_i(\theta_i - \mu)^2 = \sum_i(\theta_i - \bar{\theta})^2 + K(\bar{\theta} - \mu)^2$$

(where $\bar{\theta}$ is the mean of the θ_i) giving

$$e^{-\frac{K\lambda_\theta}{2}(\bar{\theta}-\mu)^2-\frac{\lambda_0}{2}(\mu-\mu_0)^2}.$$

This is of the form $e^{-\frac{\lambda}{2}(\mu-m)^2}$ times a constant with $\lambda = \lambda_0 + K\lambda_\theta$ and $m = (\lambda_0\mu_0 + K\lambda_\theta\bar{\theta})/\lambda$, and so the distribution of μ given the rest is

$$\mu|\text{rest} \sim \text{Normal}\left(\frac{\lambda_0\mu_0 + K\lambda_\theta\bar{\theta}}{\lambda_0 + K\lambda_\theta}, \frac{1}{\lambda_0 + K\lambda_\theta}\right)$$

A similar calculation gives

$$\theta_i|\text{rest} \sim \text{Normal}\left(\frac{\lambda_\theta\mu + J\lambda_e\bar{y}_i}{\lambda_\theta + J\lambda_e}, \frac{1}{\lambda_\theta + J\lambda_e}\right)$$

(where $\bar{y}_i$ is the mean of the y_{ij}).

Note that the conditional distribution of one θ_i given the rest of the variables does not depend on any of the other θ_i. Hence the components of the vector $\theta = (\theta_1, \ldots, \theta_K)$ are conditionally independent given μ, λ_θ, and λ_e. Thus in principle the updates of the θ_i could be done simultaneously if we had a computer capable of parallel processing. Whether the updating is done simultaneously or not the effect is the same, so long as we update all of the θ_i consecutively. In effect, there are only four variables, the three scalar variables μ, λ_θ, and λ_e and the vector variable θ.

2.5 The Block Gibbs Sampler

This illustrates a general point about Gibbs sampling. The "variables" used need not be scalar. So long as each "variable" is updated using its conditional distribution given the rest derived from the desired stationary distribution, the argument establishing the validity of the Gibbs update works. This procedure is sometimes called "block Gibbs" because one updates a "block" of variables from their joint conditional distribution given the rest. But it is really just the ordinary Gibbs sampler. Nothing in the definition of the Gibbs update or proof that it is reversible with the specified stationary distribution required that the "variables" be one-dimensional.

The example in the preceding section is trivial, since one does the same thing whether or θ is considered one variable or K variables, so long as one adopts a scan order that updates all of the θ_i in a block. A nontrivial example of "block" Gibbs is obtained by considering the variance components model with only three "variables," the scalars λ_θ, and λ_e and the vector $\zeta = (\theta_1, \ldots, \theta_K, \mu)$, since μ and the θ_i are not conditionally independent given the λs, this gives a sampling scheme that is considerably different, and actually much better.

The conditional distribution of ζ given the rest is normal with precision matrix

$$V^{-1} = \begin{pmatrix} (\lambda_\theta + J\lambda_e)I & -\lambda_\theta 1 \\ -\lambda_\theta 1' & \lambda_0 + K\lambda_\theta \end{pmatrix}$$

that is, the upper left corner is a $K \times K$ matrix with $\lambda_\theta + J\lambda_e$ down the diagonal and zeros elsewhere, the upper right corner is a $K \times 1$ column vector with all elements $-\lambda_\theta$, the lower left corner is a $1 \times K$ row vector with all elements $-\lambda_\theta$, and the lower right corner is the scalar $\lambda_0 + K\lambda_\theta$. The mean vector is difficult to specify explicitly, but is is the solution ζ_0 of the system of linear equations

$$V^{-1}\zeta_0 = \begin{pmatrix} J\lambda_e \bar{y}_1 \\ \vdots \\ J\lambda_e \bar{y}_K \\ \lambda_0 \mu_0 \end{pmatrix}$$

In order to simulate from this distribution, we need a Cholesky factorization of $V^{-1} = LL^T$. Then the new value of ζ is

$$\zeta_0 + (L^{-1})^T z$$

where z is a random vector with independent standard normal components. The Cholesky factorization can also be used in solving for ζ_0. The Cholesky factorization can be done using standard numerical linear algebra routines, such as those found in LINPACK and LAPACK. Here it can also be done by hand. There are explicit, though rather complicated, formulas for L and L^{-1}.

2.6 Problems with the Gibbs Sampler

The problem with Gibbs is the requirement that one be able to sample from the conditional distribution of x_i given x_{-i} for each i. The beauty of Gibbs sampling is that one can sample the joint distribution knowing only the full conditionals, but one does need to know the full conditionals, and there is no reason why they should known in general, though in nice Bayesian problems with familiar sampling distributions and conjugate priors, they often turn out to be familiar, normal, gamma, beta, and the like.

If one does not know the full conditionals or is not able to sample from them efficiently, then Gibbs is either impossible or not competitive with other MCMC methods. Gibbs should only be used when it is easy and does the problem that one actually wants to do. At the first bit of difficulty, Gibbs should be abandoned and other MCMC methods adopted. An addiction to Gibbs tends to limit one to problems for which Gibbs works well, keeping to conjugate priors for example.

In the excitement in the Bayesian community following the publication of Gelfand and Smith (1990) there was a great emphasis on the Gibbs sampler. Only a few years later at a Royal Statistical Society one-day meeting on MCMC at which three papers were read (Smith and Roberts, 1993; Besag and Green, 1993; Gilks et al., 1993), the proposer of the vote of thanks (Clifford, 1993) said "Surely it has to be recognized that the Gibbs sampler attained prominence by accident. Currently, there are many statisticians trying to reverse out of this historical *cul-de-sac*." There was no disagreement in the replies. Nowadays this is well known to experts, but may still be only slowly seeping out of the primary

literature into the general knowledge of all statisticians. So it bears emphasizing here. The Gibbs sampler was overhyped. Reliance on the Gibbs sampler to the exclusion of other MCMC methods was mostly a result of ignorance of those other methods. There is no reason to prefer the Gibbs sampler to other MCMC methods, to which we now turn.

2.7 The Metropolis-Hastings Update

Unlike the Gibbs update, the Metropolis-Hastings (Metropolis et al., 1953; Hastings, 1970) update permits updating of all variables instead of just one, though it does not require this. It works much the same way whether we update all the variables or only a subset but the notation is a bit more complicated, so for simplicity we first consider an update of all the variables.

Also unlike the Gibbs update, the Metropolis-Hastings update works for any distribution π specified by an unnormalized density $h(x)$ with respect to some measure μ on the state space, usually counting measure for discrete state spaces or Lebesgue measure (dx) for Euclidean spaces. In any Bayesian problem, the unnormalized posterior density is the likelihood times the prior and so can always be written down whenever there is agreement on the model and the prior. Many distributions in spatial statistics are also specified by unnormalized densities. Other examples will be seen later. There is no restriction on $h(x)$ other than that it actually be an unnormalized density, that is $h(x) \ge 0$, for all x and

$$c = \int h(x)\mu(dx) < \infty, \tag{2.7}$$

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. Even for (2.7) we do not need to know the value of the normalizing constant c. We only need to know that the integral is finite. In particular, we do not need to know anything about any conditional distributions of π.

The normalized density of π is, of course,

$$f(x) = \frac{1}{c} h(x)$$

but this plays no role in the Metropolis-Hastings update. Unnormalized densities like $h(x)$ occur throughout MCMC and every student of the subject should become accustomed to thinking about them.

The Metropolis-Hastings update uses an auxiliary transition probability specified by a density $q(x, y)$ sometimes called the "proposal distribution" or the "candidate generating 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 $q(x, y)$. 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"

$$R = \frac{h(y)q(y, x)}{h(x)q(x, y)}. \tag{2.8}$$

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

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. Those familiar with so-called rejection sampling in ordinary Monte Carlo should 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_{t+1} is equal to X_t.

Note that the denominator of the Hastings ratio (2.8) 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.

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 (2.8) 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. The normal proposal mentioned above is a Metropolis proposal. By the symmetry $q(x, y) = \phi(y - x)$ is equal to $q(y, x) = \phi(x - y)$.

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

$$p(x, y) = q(x, y)a(x, y), \tag{2.9}$$

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 some proposal will be accepted, including the possibility that the proposal y is equal to x. Thus the probability that the proposal is rejected is

$$r(x) = 1 - \int q(x, y)a(x, y)\mu(dy), \tag{2.10}$$

If the proposal is rejected we stay at x. Hence

$$P(x, A) = r(x)I(x, A) + \int_A q(x, y)a(x, y)\mu(dy), \tag{2.11}$$

Where $I(x, A)$ was defined in (2.2). The first term is zero if $x \notin A$ and otherwise is $r(x)$.

2.7.1 Reversibility and Detailed Balance

We now want to verify that the Metropolis-Hastings update is reversible, and do this by verifying a condition called "detailed balance." Suppose the transition probability kernel of a Markov chain has the following form

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

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

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

As we just saw, the Metropolis-Hastings update has this form with $p(x, y) = q(x, y)a(x, y)$. Suppose $h(x)$ is an unnormalized density with respect to μ and

$$h(x)p(x, y) = h(y)p(y, x), \qquad \text{for all } x \text{ and } y, \tag{2.12}$$

which is called detailed balance. Then this Markov chain is reversible and $h(x)$ is the unnormalized density of a stationary distribution.

To prove this we need to verify the exchangeability of X_t and X_{t+1}, but we shall actually do a little bit more. We have already noted in Section 1.3.5

that the transition probability P can be thought of as an operator on $L^2(\pi)$. We now prove that it is a self-adjoint operator if detailed balance is satisfied. Self-adjoint means for any functions u and v in $L^2(\pi)$ that $(u, Pv) = (Pu, v)$, where (u, v) denotes the inner product on $L^2(\pi)$

$$(u, v) = \int u(x)v(x)\pi(dx).$$

Thus self-adjoint means

$$\iint u(x)v(y)\pi(dx)P(x, dy) = \iint u(y)v(x)\pi(dx)P(x, dy) \tag{2.13}$$

All that is required for reversibility is the special case where u and v are indicators of sets

$$\begin{aligned}\Pr(X_t \in A, X_{t+1} \in B) &= \iint 1_A(x)1_B(y)\pi(dx)P(x, dy) \\ &= \iint 1_B(x)1_A(y)\pi(dx)P(x, dy) = \Pr(X_t \in B, X_{t+1} \in A)\end{aligned} \tag{2.14}$$

It is not completely obvious that if (2.14) holds for all sets A and B then (2.13) holds for all functions u and v, but this is standard measure theory. Extend to simple functions by linearity, to nonnegative measurable functions by dominated convergence, and to all $L^2(\pi)$ functions by linearity. Thus "self-adjoint transition operator" is equivalent to "reversible Markov chain."

In (2.13) the two sides are the same except that u and v have been interchanged. Hence we need to show that one side is invariant under the interchange of u and v.

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

The first term is obviously unchanged by interchanging u and v. So we work on the second term, except for the normalizing constant for $h(x)$ this is

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

where detailed balance is used to get the first equality and the dummy variables x and y have been interchanged to get the second. Now, except for the order of integration, the second line is just the left hand side of the first with u and v interchanged. Reversal of the order of integration is justified by the Fubini theorem.

2.7.2 Reversibility of the Metropolis-Hastings Update

We still need to prove that the Metropolis-Hastings update satisfies the detailed balance condition (2.12).

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 $R \leq 1$ then

$$a(x, y) = \frac{h(y)q(y, x)}{h(x)q(x, y)} \qquad \text{and} \qquad a(y, x) = 1 \tag{2.15}$$

and if $R \geq 1$ then

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

Since x and y are dummy variables in the detailed balance formula (2.12) both cases reduce to the same thing, since they differ only by interchange of x and y. Thus assume without loss of generality that $R \leq 1$ so (2.15) holds. Then the left hand side of (2.12) is

$$h(x)p(x, y) = h(x)q(x, y)a(x, y) = h(x)q(x, y)\frac{h(y)q(y, x)}{h(x)q(x, y)} = h(y)q(y, x),$$

and the right hand side of (2.12) is

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

So both sides are the same, detailed balance holds, and any Metropolis-Hastings update is reversible.

2.7.3 Updating a Subset of Variables

The Metropolis-Hastings update can also be done on a subset of variables, including as a special case updating only one variable, like the Gibbs sampler. The algorithm is essentially the same. Some changes in notation are required because the proposal only changes a subset of the variables and hence the proposal density $q(x, y)$ is not a density with respect to the measure μ on the whole space. It must be a density with respect to a measure ν on the subspace spanned by the variables being updated.

For a particular example, suppose that we want to do the variance components example by the Metropolis algorithm rather than the Gibbs sampler, but we still want to update the variables one at a time. Suppose that we use a normal proposal centered at the current value for the θ_i and μ and a log normal proposal centered at the current value for λ_θ and λ_e, that is we propose a new value of λ_θ by generating an normal random variate z with mean zero and some

variance (an adjustable parameter of the Metropolis update) and taking $\lambda_\theta e^z$ as the proposal. The reason for a log normal proposal is to keep the λs positive. These are not symmetric proposals, so the Metropolis rejection is based on the Hastings ratio. Consider the update of μ. Denoting the proposal by μ', it has the density $\phi((\mu' - \mu)/\sigma_\mu)$ where ϕ is the univariate normal density and σ_μ is an adjustable parameter. This is a density with respect to Lebesgue measure on $\mathbb{R}$.

In order to keep the same notation as we used when all variables were updated simultaneously, let us maintain the convention that $q(x, y)$ is a function in which both x and y take values in the state space. When the proposal only updates some of the variables, then we will only evaluate $q(x, y)$ when the components not being changed are equal. To be more specific, suppose we are only updating x_i, leaving the rest of the variables x_{-i} unchanged. Then the proposal y always satisfies $y_{-i} = x_{-i}$. We write $q(x, y)$ with y having full dimension, but when we write the proposal density $q(x, \cdot)$ only y_i is variable, y_{-i} is fixed at x_{-i}. Although the notation looks like a d-dimensional density, it is really a 1-dimensional density of the variable x_i.

With this convention, the description of the Metropolis-Hastings update is unchanged, but the proof of its reversibility needs notational changes. In (2.10) and (2.11) $\mu(dy)$ must be replaced by $\nu(dy)$, and similar changes made throughout the proof. We shall not go through the details, because this is a special case of a more general Metropolis-like algorithm due to Green (submitted), and we shall go through the proof Metropolis-Hastings-Green.

2.7.4 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 $h(x) = h(x_1, \ldots, x_d)$. This is also the unnormalized conditional density of x_i given x_{-i} (or of an block of variables given the rest) because of conditional = joint/marginal the marginal of x_{-i} is a constant when we are considering the conditional of x_i. So this only changes the normalizing constant.

A Gibbs update is a Metropolis-Hastings update in which the proposal density is $\pi(x_i|x_{-i})$. Thus

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

where $y_j = x_j$ for $i \neq j$ and c is the unknown normalizing constant that makes h a proper conditional probability. Then the Hastings ratio is

$$\frac{h(y)q(y, x)}{h(x)q(x, y)} = \frac{h(y)h(y_1, \ldots, y_{i-1}, x_i, y_{i+1}, \ldots, y_d)}{h(x)h(x_1, \ldots, x_{i-1}, y_i, x_{i+1}, \ldots, x_d)} = \frac{h(y)h(x)}{h(x)h(y)} = 1$$

since the normalizing constant c cancels. 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.

2.8 The Strauss Process

A spatial point process is a stochastic process taking values that are point patterns in a region of the plane or a higher dimensional space. The simplest is a Poisson process. A Poisson process in the unit square in two dimensions is generated by simulating a random integer N having a Poisson distribution with mean λ and then placing N points uniformly distributed in the square. The $2N$ coordinates of the N points are all independent and uniformly distributed in the interval $(0, 1)$. The Poisson process is a model of complete randomness. The parts of the point pattern in disjoint regions are statistically independent. The parameter λ is the expected number of points per unit area, which is the same for all regions.

The Strauss process is perhaps the simplest non-Poisson spatial point process. It is an exponential family of distributions having unnormalized densities of the form

$$h(x) = e^{\alpha n(x) + \beta s(x)} \tag{2.16}$$

with respect to a Poisson process. This is called the "natural" or "canonical" parametrization of the exponential family, with canonical parameters α and β and canonical statistics $n(x)$ and $s(x)$. The first canonical statistic $n(x)$ is the number of points in the point pattern and the second canonical statistic $s(x)$ is the number of "neighbor pairs" where a pair of points are defined to be neighbors if they are within a distance r of each other. We can think of r as another parameter if we like, but if we consider r a parameter rather than a known constant, we no longer have an exponential family. Also $s(x)$ is a discontinuous function of r, so the likelihood is not continuous. The family is much less well behaved if r is taken to be a parameter.

The first task when dealing with any unnormalized density like (2.16) is to verify that it is one. Clearly (2.16) is nonnegative, but we must verify that it has a finite integral. To integrate with respect to the Poisson process we need to sum over all possible numbers of points and then integrate over all possible positions of points

$$c(\alpha, \beta) = \sum_{n=0}^{\infty} \frac{\lambda^n}{n!} e^{-\lambda} \int \cdots \int e^{\alpha n + \beta s(x)} \, dx_1 \cdots dx_n.$$

The integral is over the $2n$ coordinates of the n points. It cannot be calculated analytically because $s(x)$ is a very complicated function of the positions of the points.

We need to consider separately the cases $\beta \le 0$ and $\beta > 0$. If $\beta \le 0$, we increase the integral by setting β to zero, so

$$c(\alpha, \beta) \le \sum_{n=0}^{\infty} \frac{\lambda^n}{n!} e^{-\lambda} e^{\alpha n} = e^{-\lambda + \lambda e^{\alpha}}$$

So we do have a well-defined model specified by (2.16) when $\beta \le 0$. If $\beta > 0$, we only decrease the integral by integrating over smaller regions. Suppose we

integrate over the regions such that each point has both coordinates in $(0, d)$ where $d < \sqrt{r}$. In this region, every point is a neighbor of every other point, so $s(x) = n(n-1)/2$. Thus

$$c(\alpha, \beta) \geq \sum_{n=0}^{\infty} \frac{\lambda^n}{n!} e^{-\lambda} e^{\alpha n + \beta n(n-1)/2} d^{2n}$$

From Stirling's approximation $n! \leq \sqrt{2\pi} n^{n+1/2} e^{-n}$ so

$$\log \left(\frac{\lambda^n}{n!} e^{\alpha n + \beta n(n-1)/2} d^{2n} \right) \geq n \log \lambda - \tfrac{1}{2} \log(2\pi) - (n + \tfrac{1}{2}) \log n + n + \alpha n + \beta n(n-1)/2 + 2n \log d$$

The leading positive term $\beta n^2/2$ grows faster than the leading negative term $-n \log n$. Hence the terms of the series go to infinity and the sum is infinite.

We can also consider the Strauss process with a fixed number of points n. Then we have a one-parameter exponential family with unnormalized density

$$h(x) = e^{\beta s(x)}$$

with respect to the so-called binomial process, the Poisson process conditioned on $n(x) = n$, that is, there are n points uniformly distributed in the region. This process exists regardless of the sign of β since $h(x)$ is bounded above by $e^{\beta n(n-1)/2}$, and we are integrating over a set of finite measure. When $\beta \leq 0$, the Strauss process with a fixed number of points is obtained from the Strauss process with a random number of points by conditioning on the event $n(x) = n$. Hence the process with a fixed number of points is sometimes called the "conditional Strauss process" although this is a misnomer when $\beta > 0$ since there exist no unconditional Strauss processes with $\beta > 0$.

The Strauss process with fixed number of points was defined by Strauss (1975) and proposed as a model for clustering in point patterns. The Strauss process with a random number of points was defined by Kelly and Ripley (1976).

The Poisson process is the special case of the Strauss process obtained by setting $\beta = 0$. Letting β tend to $\pm\infty$ gives two other interesting stochastic processes. A general exponential family has unnormalized density

$$h_\theta(x) = e^{\langle t(x), \theta \rangle}$$

the notation $\langle t(x), \theta \rangle$ denoting the "inner product" $t_1(x)\theta_1 + \cdots + t_d(x)\theta_d$ of the d-dimensional canonical statistic $t(x)$ and canonical parameter θ. For any direction ϕ in the parameter space, the limit of the distributions defined by the unnormalized densities $h_{\theta + s\phi}$ as $s \to \infty$ is the distribution defined by h_θ conditioned on the set H_ϕ defined by

$$H_\phi = \{ \, x : \langle t(x), \phi \rangle = \operatorname{ess\,sup} \langle t(X), \phi \rangle \, \}$$

the set of points where $\langle t(x), \phi \rangle$ achieves its maximum value.

Sending β to $-\infty$ in the conditional Strauss process, is a special case of this where the parameter space is one-dimensional and the direction ϕ is -1. The result is conditioning the Strauss process on the set where $-s(x)$ achieves its maximum value, that is, $s(x)$ achieves its minimum value, which is zero. We get the same result regardless of which Strauss process we condition, so we may as well condition the binomial process. This process obtained by conditioning the binomial process on the event $s(x) = 0$, that is, the distance from every point to every other point is at least r, is called the hard core process. A similar result is obtained by sending the two dimensional parameter of the unconditional to infinity in the direction $(0, -1)$. Now we get a Poisson process conditioned on the event $s(x) = 0$.

Sending β to $+\infty$ in the Strauss process with a fixed number of points results in a binomial process conditioned on the event $s(x) = n(n-1)/2$, that is, every point is a neighbor of every other point, so all of the points are covered by a disk of radius r centered at any of the points.

2.9 Simulating the Strauss Process

The first method proposed for simulating a Strauss process with a fixed number of points was the Gibbs sampler (Ripley 1979). The "variables" Ripley considered were two-dimensional, the positions of points. Each elementary update step moved one point giving it a realization from its conditional distribution given the rest. Ripley used a random scan in his published computer code but noted that a fixed scan could also be used.

The conditional distribution for one point given the rest is analytically intractable, so rejection sampling from the uniform distribution was used.

repeat
  Simulate a new value of x_i uniformly distributed in the region
  Calculate the number m of points in x_{-i} that are neighbors of x_i
  if $\beta > 0$ **then**
    Set $p = e^{\beta(m-n+1)}$
  else
    Set $p = e^{\beta m}$
  end if
  Generate a uniform (0,1) random variate u
until $u < p$

This works because the proposed value of x_i is accepted with probability proportional to $e^{\beta m}$, which is proportional to the conditional density of x_i given x_{-i}.

This Gibbs sampler is very inefficient when the number of points n is large or when β is large and positive, because the rejection sampling is very inefficient. A Metropolis algorithm is much simpler and at least as efficient. A simple Metropolis update proposes to move one point, giving it a uniform distribution. This is clearly a symmetric proposal. Hence the Metropolis rejection is based on

the odds ratio $h(y)/h(x) = e^{\beta[s(y)-s(x)]}$. As with the Gibbs sampler, the change $s(y) - s(x)$ in the canonical statistic only involves the neighbors of the point being moved. It is the number of neighbors the point has at its new position minus the number of neighbors it had at the old position.

This is only the simplest proposal. The proposal could depend on x For example, if $\beta > 0$ we could increase the probability that the new point has many neighbors in the hope that this will make the simulation more efficient. We could propose a new position y_i for x_i that is uniform in the region with probability $1-\epsilon$ and with probability $\epsilon/(n-1)$ uniform in the disk D_j of radius r centered at the point x_j for each of the other $n-1$ points. Then the proposal probability is

$$q(x, y) = (1 - \epsilon) + \frac{\epsilon}{n-1} \sum_{j \neq i} 1_{D_j}(y) \tag{2.17}$$

Let $s_i(x)$ denote the number of neighbors of x_i

$$s_i(x) = \sum_{j \neq i} 1_{\|x_i - x_j\| < r}$$

So $s_i(y)$ is the number of neighbors of x_i in the proposed new position. Then $q(x, y) = (1 - \epsilon) + \epsilon s_i(y)/(n-1)$ and Then $q(y, x) = (1 - \epsilon) + \epsilon s_i(x)/(n-1)$, so this proposal is not symmetric, and we would have to use the Hastings ratio rather than the odds ratio in the Metropolis rejection.

This proposal sometimes proposes points that lie outside the region containing the process. For such points $h(y) = 0$ and the proposal is always rejected. We could alter the proposal so it only proposes points in the region, but it is not clear that it is worth the bother.

2.10 The Metropolis-Hastings-Green Update

In order to simulate an unconditional Strauss process we need a generalization of the Metropolis-Hastings algorithm described by Geyer and Møller (1994), which is a special case of a much more general algorithm due to Green (submitted). The problem with simulating a spatial point process is that the dimension of the problem changes as the number of points changes, and neither the Gibbs sampler or the Metropolis-Hastings algorithm handles that.

There are many situations where one also wants to deal with a state space that is a union of sets of different dimension. Another example is Bayesian model selection. In the variance components example, suppose we want to test whether $1/\lambda_\theta$ is zero. Then all the θ_i are equal to μ; the groups have the same mean. Then the model reduces to

$$y_{ij} \sim \text{Normal}\left(\mu, \frac{1}{\lambda_e}\right), \qquad i = 1, \ldots, K, \quad j = 1, \ldots, J.$$

Suppose we keep the same priors for μ and λ_e (θ and λ_θ no longer appears in the model). We also need a prior on models, say probability c_1 on the

large model and $c_0 = 1 - c_1$ on the small model. The posterior distribution of the parameters is still well-defined, but exists on a union of sets of different dimension. With some probability the large model is correct and the parameter vector has dimension $K + 3$. With one minus that probability the small model holds and the parameter vector has dimension 2. If we knew these probabilities we could sample each separately, but we can only find them by sampling both models together.

The Metropolis-Hastings-Green update makes proposals of varying dimension. Hence they cannot easily be described by a density. Thus we have a proposal kernel $Q(x, A)$ that satisfies the following condition. $\pi(dx)Q(x, dy)$ has a density $f(x, y)$ with respect to some symmetric measure ξ on the Cartesian product of the state space with itself. As before, we need to be able simulate random variates with the distribution $Q(x, \cdot\,)$ and be able to evaluate the density $f(x, y)$.

The update then works as follows. As with the Metropolis-Hastings 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 distribution $Q(x, \cdot\,)$.
2. Calculate "Green's ratio"
$$R = \frac{f(y, x)}{f(x, y)}.$$
3. Do "Metropolis rejection:" with probability $\min(1, R)$ set $x = y$.

Although the notation does not make it explicit, it is clear that we can use an unnormalized specification of the stationary distribution π in defining $f(x, y)$. This will only multiply $f(x, y)$ by a constant, which cancels in the ratio R.

2.10.1 Simulating the Unconditional Strauss Process

The notion of the density $f(x, y)$ and the measure ξ are best explained by example. Let us consider the following elementary update for the unconditional Strauss process. Suppose the points are ordered $x_1, \ldots, x_n$, where $n = n(x)$ is the number of points. If there are $n(x) = m$ points, then attempt to add a new point x_{m+1} uniformly distributed in the region, and if there are $n(x) = m + 1$ attempt to delete point x_{m+1}.

A single elementary update only changes the number of points from m to $m+1$ or from $m+1$ to m. Thus we have an infinite number of possible elementary updates, one for each nonnegative integer m. Let S denote the region in which the points are located, the unit square in $\mathbb{R}^2$. First consider the part of the proposal that attempts to add a point. Then x is in S^m, and on this set π has the unnormalized density
$$\frac{\lambda^m}{m!} e^{-\lambda + m\alpha + s(x)\beta}$$
with respect to Lebesgue measure on S^m. The new point y_{m+1} is uniformly distributed on S, and the rest of the points are not moved, $x_i = y_i$ for $i \le m$.

Thus the joint distribution of the pair (x, y) is concentrated on a set of dimension $2(m + 1)$ defined by

$$D_m^+ = \{\, (x, y) \in S^m \times S^{m+1} : x_i = y_i, i \le m \,\}.$$

We take part of the measure ξ to be Lebesgue measure on this set. In order that ξ be a symmetric measure it must have another part that is Lebesgue measure on

$$D_m^- = \{\, (x, y) \in S^{m+1} \times S^m : x_i = y_i, i \le m \,\},$$

which is the same except x and y are exchanged, x now having the larger dimension. Then the density $f(x, y)$ is given by

$$f(x, y) = \frac{\lambda^m}{m!} e^{-\lambda + m\alpha + s(x)\beta}$$

for (x, y) in D_m^+ and zero for (x, y) in D_m^-. Now consider the part of the proposal that deletes a point, going from $m+1$ points to m points. Then $x \in S^{m+1}$, and on this set π has the unnormalized density

$$\frac{\lambda^{m+1}}{(m + 1)!} e^{-\lambda + (m+1)\alpha + s(x)\beta}$$

with respect to Lebesgue measure on S^{m+1}. The proposal is not random; it attempts to delete point x_{m+1} with probability one. The pair (x, y) now lies in the set D_m^- and the density $f(x, y)$ is

$$f(x, y) = \frac{\lambda^{m+1}}{(m + 1)!} e^{-\lambda + (m+1)\alpha + s(x)\beta}$$

for (x, y) in D_m^- and zero for (x, y) in D_m^+.

When we are adding a point so (x, y) is in D_m^+ and (y, x) is in D_m^-

$$R = \frac{f(y, x)}{f(x, y)} = \frac{\lambda^{m+1} m! e^{-\lambda + (m+1)\alpha + s(y)\beta}}{\lambda^m (m + 1)! e^{-\lambda + m\alpha + s(x)\beta}} = \frac{\lambda}{m + 1} e^{\alpha + [s(y) - s(x)]\beta} \tag{2.18}$$

When we are deleting a point so (x, y) is in D_m^- and (y, x) is in D_m^+ we have

$$R = \frac{f(y, x)}{f(x, y)} = \frac{m + 1}{\lambda} e^{-\alpha + [s(y) - s(x)]\beta} \tag{2.19}$$

which is just the reciprocal with x and y interchanged.

These elementary updates are combined by mixing or composition. The mixing proposal is a bit tricky, so we describe it in detail. Let $Q_m(x, A)$ denote the proposal kernel just described, and $a_m(x, y)$ the acceptance probability. There are an infinite sequence of proposal kernels for $m = 0, 1, \ldots$. We combine the elementary updates by mixing over all of them

$$P(x, A) = r(x) I(x, A) + \frac{1}{2} \sum_{m=0}^{\infty} \int_A Q_m(x, dy) a_m(x, y)$$

where as before $r(x)$ is the rejection probability

$$r(x) = 1 - \frac{1}{2} \sum_{m=0}^{\infty} \int Q_m(x, dy) a_m(x, y)$$

Why divide by 2 when there is an infinite sum? For any x except the empty realization, which has no points, there are only two nonzero kernels. If $n(x) = m > 0$, then $Q_m(x, S^{m+1}) > 0$ and $Q_m(x, S^{m-1}) > 0$. The rest of the terms are zero. So dividing by 2 does yield a proper kernel.

Having verified all of this algebra, let us make sure we are clear what the algorithm does. If there are $m = n(x)$ points, we flip a coin, and, if it comes up heads we attempt to add a new point y_{m+1} uniformly distributed in the unit square. Both coordinates of y_{m+1} are Uniform$(0, 1)$. We accept the proposal with probability $\min(1, R)$ where R is given by (2.18). If the coin comes up tails, we attempt to delete point x_m if $m > 0$, and if $m = 0$ we make no proposal. We accept the proposal with probability $\min(1, R)$ where R is given by (2.19).

We can combine this update by composition with other updates. One useful update simply permutes the labels of the n points. Clearly this does not change the distribution, since the density does not depend on the labels of the points. Hence this proposal is always accepted. If we look at the effect, we see that this is equivalent to always deleting a random point rather than x_n, because we have just permuted the labels of the points. The point of this trick is to simplify the argument above.

Another possibility is to combine this dimension-changing update with the dimension-maintaining update used for simulating the conditional Strauss process that moves a point rather than adding or deleting one. Geyer and Møller, however, found that was not helpful. A chain with updates that only add and delete worked better.

Still another possibility is to use the nonuniform density (2.17) to locate the point being added. This would add a factor $q(x, y)$ to the denominator of (2.18) and the numerator of (2.19).

2.10.2 Reversibility of the Metropolis-Hastings-Green Update

To show that this algorithm is valid, we need to show that is satisfies a detailed balance condition, which can be written with a bit of abuse of notation as

$$\pi(dx) Q(x, dy) a(x, y) = \pi(dy) Q(y, dx) a(y, x),$$

the meaning being that integrating with respect to either side produces the same results, that is

$$\iint u(x) v(y) \pi(dx) Q(x, dy) a(x, y) \tag{2.20}$$

is the same if x and y are interchanged (or if u and v are interchanged) for all $L^2(\pi)$ functions u and v.

Substituting the definition of the function f in (2.20) we get

$$\iint u(x)v(y)f(x,y)\xi(dx,dy)a(x,y) \tag{2.21}$$

As with the proof for the Metropolis-Hastings update, we may assume without loss of generality that $R \le 1$ so that

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

Then (2.21) becomes

$$\begin{aligned}
\iint u(x)v(y)f(x,y)\xi(dx,dy)\frac{f(y,x)}{f(x,y)} &= \iint u(x)v(y)f(y,x)\xi(dx,dy) \\
&= \iint u(x)v(y)f(y,x)\xi(dx,dy)a(y,x) \\
&= \iint u(y)v(x)f(x,y)\xi(dy,dx)a(x,y)
\end{aligned}$$

where the third equality is the interchange of dummy variables. Now this is equal to (2.21) with u and v interchanged except for the order of integration, and this interchange is justified by the requirement that ξ be a symmetric measure.

It remains to check that detailed balance implies reversibility. The kernel for an elementary update is

$$P(x,A) = r(x)I(x,A) + \int_A Q(x,dy)a(x,y)$$

where

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

Calculating the inner product (u, Pv) gives

$$\int u(x)v(x)r(x)\pi(dx) + \iint u(x)v(y)\pi(dx)Q(x,dy)a(x,y).$$

which we need to show is unchanged by interchange of u and v. As before, the first term is obviously unchanged, and now that the second term is unchanged is trivial since the second term is (2.20). We have already verified that.

2.10.3 Bayesian Model Selection

We can use Green's algorithm to do model selection for the variance components model. Because there are two models, we now must be more careful about constants. We can have an unknown normalizing constant for the whole

posterior, but the relative likelihoods of the models must be known. Including all constants, the likelihood times the prior for the big model becomes

$$\begin{aligned}
&(2\pi)^{-JK/2} \lambda_e^{JK/2} e^{-\frac{\lambda_e}{2} \sum_{ij} (y_{ij} - \theta_i)^2} (2\pi)^{-K/2} \lambda_\theta^{K/2} e^{-\frac{\lambda_\theta}{2} \sum_i (\theta_i - \mu)^2} \\
&\quad \times (2\pi)^{-1/2} \lambda_0^{-1/2} e^{-\frac{\lambda_0}{2} (\mu - \mu_0)^2} \frac{1}{\Gamma(a_1)} b_1^{a_1} \lambda_\theta^{a_1 - 1} e^{-b_1 \lambda_\theta} \frac{1}{\Gamma(a_2)} b_2^{a_2} \lambda_e^{a_2 - 1} e^{-b_2 \lambda_e} c_1
\end{aligned} \tag{2.22}$$

and the likelihood times the prior for the small model is

$$\begin{aligned}
&(2\pi)^{-JK/2} \lambda_e^{JK/2} e^{-\frac{\lambda_e}{2} \sum_{ij} (y_{ij} - \mu)^2} \\
&\qquad \times (2\pi)^{-1/2} \lambda_0^{-1/2} e^{-\frac{\lambda_0}{2} (\mu - \mu_0)^2} \frac{1}{\Gamma(a_2)} b_2^{a_2} \lambda_e^{a_2 - 1} e^{-b_2 \lambda_e} c_0
\end{aligned} \tag{2.23}$$

The Gibbs update for the small model uses the one-dimensional conditionals

$$\begin{aligned}
\mu | \text{rest} &\sim \text{Normal}\left(\frac{\lambda_0 \mu_0 + JK \lambda_e \bar{\bar{y}}}{\lambda_0 + JK \lambda_e}, \frac{1}{\lambda_0 + JK \lambda_e} \right) \\
\lambda_e | \text{rest} &\sim \text{Gamma}\left(a_2 + JK/2, b_1 + \frac{1}{2} \sum_{ij} (y_{ij} - \mu)^2 \right)
\end{aligned}$$

These are the only two variables in the model, so "the rest" here just refers to the other variable.

Next we want a Metropolis-Hastings-Green update that jumps between models. A possible proposal is the following. Going down from the big model to the small model, set $\theta_i = \mu$ for all i and $\lambda_\theta = \infty$ while leaving μ and λ_e unchanged. Going up from the small model to the big, propose

$$\begin{aligned}
\lambda_\theta | \mu, \lambda_e &\sim \text{Gamma}\left(a_1 + K/2, b_1 \right) \\
\theta_i | \mu, \lambda_e, \lambda_\theta &\sim \text{Normal}\left(\frac{\lambda_\theta \mu + J \lambda_e \bar{y_i}}{\lambda_\theta + J \lambda_e}, \frac{1}{\lambda_\theta + J \lambda_e} \right)
\end{aligned}$$

The latter is just the Gibbs update in the big model. So is the former for the special case $\theta_i = \mu$ for all i. Of course, this proposal if always accepted would not produce the correct distribution. We need to do Metropolis a rejection. Going down the proposal is deterministic, so $f(x, y)$ is just (2.22), where the current position x is in the parameter space of the big model, y in the parameter space of the small model, and μ and λ_e have the same values in both. Going up $f(x, y)$ is (2.23) times

$$\begin{aligned}
&\frac{1}{\Gamma(a_1 + K/2)} b_1^{a_1 + K/2} \lambda_\theta^{a_1 + K/2 - 1} e^{-b_1 \lambda_\theta} \\
&\times (2\pi)^{-K/2} (\lambda_\theta + J \lambda_e)^{K/2} \exp\left(- \frac{\lambda_\theta + J \lambda_e}{2} \sum_i \left[\theta_i - \frac{\lambda_\theta \mu + J \lambda_e \bar{y_i}}{\lambda_\theta + J \lambda_e} \right]^2 \right)
\end{aligned} \tag{2.24}$$

The ratio of the two, (2.22) divided by the product of (2.23) and (2.24), is the ratio R used for the Metropolis rejection in Green's update.

$$R = \frac{c_1}{c_0}\frac{\Gamma(a_1 + K/2)}{\Gamma(a_1)}\frac{1}{b_1^{K/2}}\frac{1}{(\lambda_\theta + J\lambda_e)^{K/2}} \frac{\exp\left(-\frac{J\lambda_e}{2}\sum_i(\bar{y}_i - \theta_i)^2 - \frac{\lambda_\theta}{2}\sum_i(\theta_i - \mu)^2\right)}{\exp\left(-\frac{J\lambda_e}{2}\sum_i(\bar{y}_i - \mu)^2 - \frac{\lambda_\theta + J\lambda_e}{2}\sum_i\left[\theta_i - \frac{\lambda_\theta\mu + J\lambda_e\bar{y}_i}{\lambda_\theta + J\lambda_e}\right]^2\right)}$$

This is Green's ratio for a step up from little to big, and the reciprocal of the ratio for a step down from big to little.

This sampler seems to work well, at least for some data. Illustrative code for the Gibbs sampler, the block Gibbs sampler, and the Gibbs-Green sampler for model selection are described in Appendix A.

Chapter 3

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 most 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 we are interested in the probability of a set A and have independent simulations X_1, X_2, ... from the distribution of interest π. Consider the question of what is the probability that n samples will completely miss the set A thus giving us a Monte Carlo estimate of zero for the true probability $\pi(A)$, which we assume to be nonzero. The absolute error may be small if $\pi(A)$ is small, but the relative error is not. This probability is

$$[1 - \pi(A)]^n$$

which goes to zero exponentially fast, and what is more important, at a rate which is determined by $\pi(A)$. For Markov chains we usually have the exponential convergence to zero. 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

$$\Pr(s_A \geq n) \leq M r^{-n}$$

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 stationary distribution.

3.1 Irreducibility

The weakest form of stochastic stability is irreducibility. Among other things, if a Markov chain has a stationary distribution and is irreducible, then the stationary 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. It 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. There are many ways to construct samplers with a specified stationary distribution.

3.1.1 Countable State Spaces

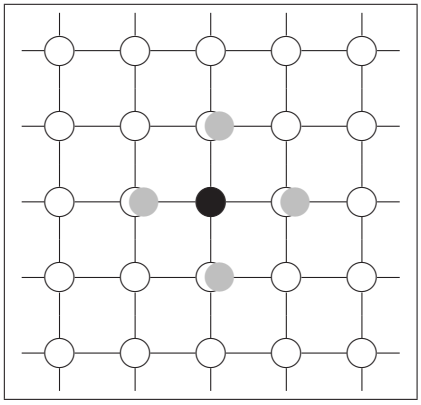
Irreducibility is the one notion that has a different definition for discrete and continuous state spaces. Since both definitions are widely used, one should know both. Recall from Sections 1.3.2 and 2.2.2 that for a countable state space the transition probabilities are described by a matrix P with entries $p(x, y)$ and that the n-step transition probabilities are given by 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."

In order to see how this definition works we need an example with a discrete state space.

3.1.2 The Ising Model

The Ising model is a spatial lattice process. The state is a vector $x = \{ x_i : i \in S \}$ where S 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 S, some vertices have neighbors outside of S.

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, like the Strauss process, a two-parameter exponential family with unnormalized density

$$h(x) = e^{\alpha t_1(x) + \beta t_2(x)} \tag{3.1}$$

where the canonical statistics are defined by

$$t_1(x) = \sum_{i \in S} x_i$$

and

$$t_2(x) = \sum_{i \in S} \sum_{j \sim i} x_i x_j. \tag{3.2}$$

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 (or vice versa depending on whether black is chosen to code for $+1$ or -1), and the second canonical statistic is the number of concordant neighbor pairs (same color) minus the number of discordant neighbor pairs. When the x_i take values in $\{0, 1\}$, and use the same definitions of the canonical statistics, the same family of stochastic models are defined but the parameterization is different.

The notation in (3.2) is deliberately ambiguous about what happens at the boundary of the region S. There are three different ways in which the boundary is commonly treated. The first is to condition on the boundary. The sums in (3.2) extend over all pairs i and j such that i is in S and j is a neighbor of i so that when i is at the edge of the region j may lie just outside the region.

The variables x_j for $j \notin S$ are not part of the state of the stochastic process, 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 S. 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 S together to form a torus. Then the set S 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 $h(x)$. The only terms that matter are those containing x_i, hence this conditional has the unnormalized density

$$h(x_i|x_{-i}) = e^{\alpha x_i + \beta x_i \sum_{j \sim i} x_j}$$

The only sum required in calculating the unnormalized 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

$$p(x_i|x_{-i}) = \frac{h(x_i|x_{-i})}{h(x_i = 0|x_{-i}) + h(x_i = 1|x_{-i})}$$

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

$$R = \frac{h(y)}{h(x)} = e^{-2\alpha x_i - 2\beta x_i \sum_{j \sim i} x_j} \tag{3.3}$$

This is a symmetric proposal so the proposal is accepted with probability $\min(1, R)$.

3.1.3 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, et al., 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

$$h(x) = e^{\alpha x_i + \beta x_i \sum_{k \sim i} x_k} e^{\alpha x_j + \beta x_j \sum_{l \sim j} x_l} \times \text{term not containing } x_i \text{ or } x_j$$

Hence these variables are conditionally independent given the rest by the factorization criterion. If i and j are neighbors, the density contains a term $e^{\beta x_i x_j}$ and these variables are not conditionally independent by the same criterion.

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 even update a whole coding set simultaneously. Thus when scanning by coding sets there are really only two block variables (the two coding sets). So the sampler is effectively reversible, as with any fixed scan with only two variables.

3.1.4 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 that does not matter when discussing irreducibility. It only matters that both probabilities are 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 it only matters that the probability be 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 de-

pends on the scan. When the odds ratio (3.3) 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 ($\alpha = 0$) 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.

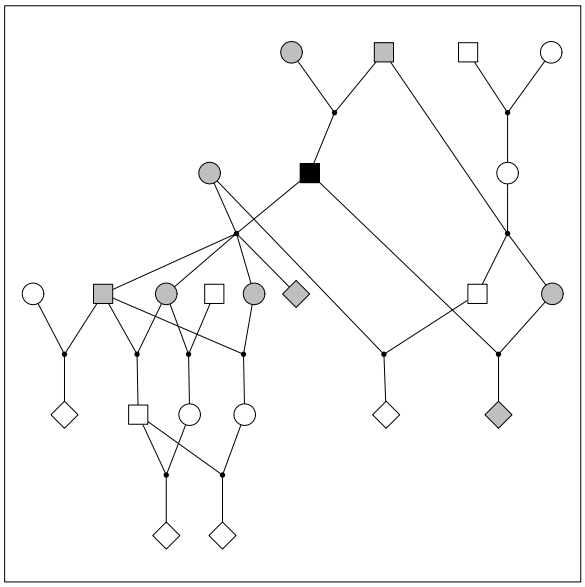
A symmetric Ising model with periodic boundary conditions can also fail to be irreducible when a raster scan is used. For that we need a lattice size that is odd and a checkerboard pattern.

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.

3.1.5 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, ..., 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, ..., 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

$$h(g) = \prod_{\text{all individuals } i} p(\text{data}|g_i) \prod_{\text{children } i} p(g_i|g_{f(i)}, g_{m(i)}) \prod_{\text{founders } i} p(g_i) \tag{3.4}$$

The genetics that requires MCMC is to simulate the conditional distribution of genotypes given data. The unnormalized density is given by (3.4). 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.

3.1.6 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 in one scan. 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 use 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 is not always irreducible. The proof applies only to problems in which there are 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 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.

3.1.7 Contingency Tables

3.1.8 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

$$U(x, A) = \sum_{n=1}^{\infty} P^n(x, A) \tag{3.5}$$

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 rather than the stationary distribution π is that the definition applies to arbitrary Markov chains, including those that do not have a stationary probability distribution. If the chain has a stationary 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 stationary 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 a stationary distribution π and is φ-irreducible, then the chain is recurrent (Meyn and Tweedie, 1993, Proposition 10.1.1), the stationary distribution is unique (Proposition 10.4.4), and the stationary 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.

3.1.9 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 stationary 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 a function that is not zero almost everywhere $Q_2(x_{-d}, A_{x_{-d}})$ 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 et al., 1953) used the Metropolis algorithm for a point process 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 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.*

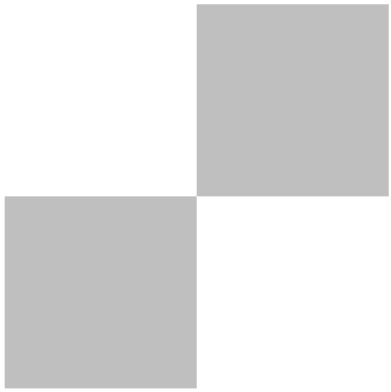
Some of these terms need explanation. A topological space is second countable if there is a countable family of open sets $\mathcal{U}$ such that every open set is a union of sets in $\mathcal{U}$. Every separable metric space, in particular any subset of a Euclidean space $\mathbb{R}^d$, has this property. 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.

Having seen how the theorem works, let us see why it is true. 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

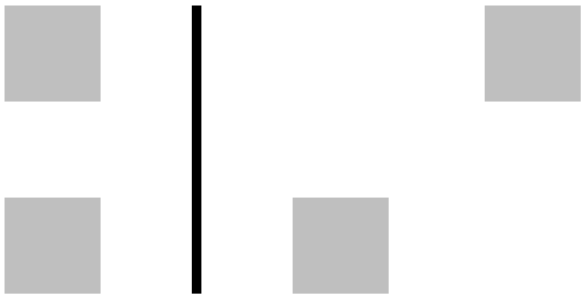
$$U(x, C) \geq \int_{A \cap B} P^m(x, dy) U(y, C),$$

where $U(x, A)$ is defined by (3.5). 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 in S a φ-communicating neighborhood W that is an element of $\mathcal{U}$, which is possible because every neighborhood of a point x contains a neighborhood of x that is an element of $\mathcal{U}$ and subsets of φ-communicating sets are φ-communicating. Let the set of such W be called $\mathcal{W}$. Then $\mathcal{W}$ is countable because $\mathcal{U}$ is countable. Consider the sequence V_1, V_2, ... of sets defined as follows. V_1 is any element of $\mathcal{W}$. Then for each k define V_{k+1} to be the union of V_k and any $W_k \in \mathcal{W}$ that intersects V_k but is not wholly contained in it. If for some k there is no such W, let $V_n = V_k$ for all $n > k$. Each V_k is φ-communicating by induction. The intersection of V_k and W_k is nonempty and open and hence φ-positive by (c). Hence the argument above shows their union is φ-communicating. Let $V = \bigcup_{k=1}^{\infty} V_k$. Then V is φ-communicating. because any $x \in V$ lies in some V_k, and any φ-positive $A \subset V$ intersects some V_k.

Now there are two logical possibilities. $V = S$ in which case the chain is irreducible or V and $S \setminus V$ are disjoint open sets and (b) is violated. Then V 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.

3.1.10 Harris recurrence

If a chain is ψ-irreducible and has a stationary 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 stationary 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 3.7.5.

3.2 The Law of Large Numbers

We now return to the law of large numbers mentioned in Section 1.3.4 and give a precise statement. Suppose we have a Markov chain with stationary 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 2 *For a φ-irreducible chain with stationary 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 3 *For a Harris recurrent chain with stationary 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 stationary distribution.

3.3 Convergence of the Empirical Measure

The law of large numbers tells us that for each fixed function g the sample mean of $g(X)$ converges to its expectation with respect to π for almost all sample paths of the Markov chain. Since a countable union of null sets is a null set, it also says that for any countable family of functions $\{g_i\}$ all of the sample means of the $g_i(X)$ converge to their expectations simultaneously with probability one. To be precise about null sets one last time, there is a null set N of sample paths of the Markov chain and for all sample paths not in N all of the sample means of the $g_i(X)$ converge.

By using continuity, we can go from a countable family to an uncountable one. We want to show that

$$\frac{1}{n} \sum_{i=1}^{n} g(X_i) \to E_\pi g(X) \tag{3.6}$$

simultaneously for all bounded continuous functions g with probability one (that is, for almost all sample paths of the Markov chain). Another way to say this is that the empirical distribution $\mathbb{P}_n$ that puts probability $1/n$ at each of the points of an n-sample converges in distribution to π. The left hand side of (3.6) is integration with respect to this empirical distribution

$$\int g(x) \mathbb{P}_n(dx) = \frac{1}{n} \sum_{i=1}^{n} g(X_i).$$

If for a particular sample path of the Markov chain

$$\int g(x) \mathbb{P}_n(dx) \to \int g(x) \pi(dx)$$

for all bounded continuous functions, this is by definition convergence in distribution of $\mathbb{P}_n$ to π.

Theorem 4 *Suppose the state space of the Markov chain is a separable metric space and the chain is Harris recurrent, then $\mathbb{P}_n$ 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,

$$\mathbb{P}_n(B) = \frac{1}{n} \sum_{i=1}^{n} 1_B(X_i) \to \pi(B), \qquad \text{for all } B \in \mathcal{B} \tag{3.7}$$

By Corollary 1 of Theorem 2.2 in Billingsley (1968), (3.7) implies $\mathbb{P}_n$ 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 stationary 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. In this way we can think of the n-sample as being "from" π even though the X_i are not independent, nor are they identically distributed, although as n goes to infinity the distribution of X_n converges to π, which is the next form of convergence we shall discuss.

3.4 Aperiodicity

The law of large numbers can hold for a Markov chain even though the marginal distributions do not. 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 stationary 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 5 *Any ψ-irreducible sampler that has $P(x, \{x\}) > 0$ for $x \in A$ where $\psi(A) > 0$ is aperiodic.*

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 6 *Any ψ-irreducible Gibbs sampler is aperiodic.*

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

$$u(x) = \sum_{k=0}^{d-1} \omega^k 1_{D_k}(x) \tag{3.8}$$

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 $P = P_1 \cdots P_k$. 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

$$\|u\|^2 = \|P_i u\|^2 + \|(I - P_i)u\|^2$$

for any function u. 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 (3.8). Hence we must have $\|P_i u\| < \|u\|$ for at least one i, which implies

$$\|P\| \le \|P_1\| \cdots \|P_k\| < 1$$

But then

$$\|Pu\| \le \|P\| \, \|u\| < 1$$

since $\|P\| < 1$ and u is a unit vector. 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.

3.5 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

$$\|\mu\| = \sup_{|f| \le 1} \int f \, d\mu. \tag{3.9}$$

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.

3.6 Convergence of Marginals

Theorem 7 *For an aperiodic Harris recurrent chain with stationary distribution π and any initial distribution λ*

$$\|\lambda P^n - \pi\| = \left\| \int \lambda(dx) P^n(x, \cdot) - \pi \right\| \to 0, \qquad \text{as } n \to \infty \tag{3.10}$$

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, (3.10) reduces to

$$\|P^n(x, \cdot) - \pi\| \to 0, \qquad \text{as } n \to \infty \tag{3.11}$$

3.7 Geometric and Uniform Ergodicity

3.7.1 Geometric Ergodicity

A Markov chain is said to be *geometrically ergodic* when the convergence in (3.11) occurs at a geometric rate, that is when there is a constant $\rho < 1$ and a nonnegative function $M(x)$ such that

$$\|P^n(x, \cdot) - \pi\| \le M(x)\rho^n \qquad \text{for all } n. \tag{3.12}$$

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 stationary distribution π is *geometrically ergodic* if there exists a constant $r > 1$ such that

$$\sum_{n=1}^{\infty} r^n \|P^n(x, \cdot) - \pi\| < \infty, \qquad \text{for all } x. \tag{3.13}$$

Note that for this series to be summable, each term must go to zero, which implies (3.12) holds with $\rho = 1/r$.

The total variation convergence in (3.12) 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.

3.7.2 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

$$P^m(x, A) \ge \delta Q(A), \qquad x \in C \text{ and } A \text{ a measurable set.} \tag{3.14}$$

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

$$K_a(x, A) = \sum_{n=0}^{\infty} a(n) P^n(x, A) \tag{3.15}$$

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

$$K_a(x, A) \geq \delta Q(A), \qquad x \in C \text{ and } A \text{ a measurable set.} \tag{3.16}$$

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 $i \to \infty$.

3.7.3 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 (2.9). 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.

Consider the unconditional Strauss process. An indirect proof that this sampler is a T-chain is given by Geyer and Møller (1994). The set $C_m = \{\, x : n(x) \le m \,\}$ is petite, because from (2.19) we see the probability of a step down is bounded below by $\frac{1}{2}e^{-\alpha}/\lambda$ (assuming $e^{-\alpha}/\lambda < 1$, otherwise it is bounded below by $\frac{1}{2}$) because removing a point always decreases the number of neighbor pairs so $[s(y) - s(x)]\beta > 0$. If $x = \varnothing$, the realization with no points, then the probability it stays there is at least $\frac{1}{2}$. Thus

$$\begin{aligned} P^{k+n}(x, \{\varnothing\}) &\ge P^k(x, \{\varnothing\}) P^n(\varnothing, \{\varnothing\}) \\ &\ge \left[\frac{1}{2}\frac{e^{-\alpha}}{\lambda}\right]^k \left[\frac{1}{2}\right]^n \\ &\ge \left[\frac{1}{2}\frac{e^{-\alpha}}{\lambda}\right]^{k+n} \end{aligned} \tag{3.17}$$

Hence for any $x \in C_m$

$$P^m(x, A) \ge \delta Q$$

where δ is given by (3.17) and Q is the probability measure concentrated at the empty realization $\varnothing$. This verifies directly that C_m is small, hence petite. The compact sets of the state space are the closed bounded sets, that is closed subsets of some C_m. Since every (measurable) subset of a petite set is petite, this proves that every compact set is petite. Hence the sampler is a T-chain.

The only reason we care about T-chains is for what they tell us about petite sets, so having directly proved that compact sets are petite, we no longer care about the T-chain property. It does hint, however, that a direct proof of the T-chain property should be possible.

3.7.4 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 a stationary 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.

3.7.5 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

$$PV(x) \leq \lambda V(x) + b1_C(x), \qquad \text{for all } x \tag{3.18}$$

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

$$PV(x) \leq V(x) - 1 + b1_C(x), \qquad \text{for all } x \tag{3.19}$$

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 $\{ x : V(x) \leq r \}$ 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 $V(x) \geq 1$ 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

$$PV(x) \leq \lambda V(x) + L. \qquad \text{for all } x \tag{3.20}$$

For any function $V \geq 1$ define the V-norm by

$$\|\mu\|_V = \sup_{|f| \leq V} \int f \, d\mu. \tag{3.21}$$

Note the resemblance to the alternative definition (3.9) 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

$$\sum_{n=1}^{\infty} r^n \|P^n(x, \cdot) - \pi\|_V \leq RV(x) \qquad \text{for all } x. \tag{3.22}$$

holds for all x. This, of course, says nothing about x such that $V(x) = \infty$.

Comparison with the definition of geometric ergodicity (3.13) shows that (3.22) is stronger except that geometric ergodicity requires that the right hand side be finite for all x, which is not so in (3.22) 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 (3.22) is much stronger than (3.13) when V is everywhere finite, because of the appearance of the V-norm rather than the total variation norm in (3.22) 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 (3.22) is only known to be finite for all x. The still stronger (3.22) is called *V-uniform ergodicity.*

3.7.6 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 (3.20) 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.

Unconditional Strauss Process

Now consider the unconditional Strauss process. As we have already seen, sets of the form $C_m = \{ x : n(x) \le m \}$ are petite. In order that $V(x)$ be unbounded

off petite sets it only need be a function of $n(x)$ that goes to infinity as $n(x)$ goes to infinity, say $V(x) = H(n(x))$. Since $n(x)$ can only change by one each step, we need

$$H(n(x) - 1) \le \lambda H(n(x))$$

for all large enough x. This suggests we take $V(x) = e^{cn(x)}$ for some fixed $c > 0$ to be chosen later.

If $n(x)$ is large enough, steps down will have a Hastings ratio greater than one and be accepted with probability one and steps up will have a Hastings ratio less than ε chosen as small as we please. Then for such x we have

$$\begin{aligned} PV(x) &\le \tfrac{1}{2} e^{c[n(x)-1]} + \tfrac{1}{2} \varepsilon e^{c[n(x)+1]} \\ &= \tfrac{1}{2} \left(e^{-c} + \varepsilon e^{c} \right) V(x) \end{aligned}$$

and this will be less than $\lambda V(x)$ for some $\lambda < 1$ and all large enough x if we choose ε small enough so that $e^{-c} + \varepsilon e^{c} \le 2\lambda$. For example, we could use $c = 1$, $\lambda = 1/2$, and $\varepsilon = .04$. From (2.19) we see the acceptance probability of a step down is greater than one if

$$R \ge \frac{n(x) + 1}{\lambda} e^{-\alpha} \ge 1,$$

and from (2.18) we see the acceptance probability of a step up is less than ε if

$$R \le \frac{\lambda}{n(x)} e^{\alpha} \le \varepsilon.$$

Both will hold whenever $n(x) \ge \lambda e^{\alpha} / \varepsilon$. For $n(x) < \lambda e^{\alpha} / \varepsilon$ we have

$$PV(x) \le e^{c[n(x)+1]} \le e^{c[\lambda e^{\alpha}/\varepsilon + 1]} = L$$

Putting the two bounds together we have (3.20).

A Theorem of Roberts and Tweedie

Roberts and Tweedie (submitted) 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 stationary 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.

3.7.7 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 (3.22). By combining the geometric drift condition with a minorization condition like (3.14) we can say something about these constants.

Theorem 8 *Suppose $V(x) \geq 0$ is an everywhere finite function and satisfies a geometric drift condition*

$$PV(x) \leq \lambda V + L, \qquad \text{for all } x. \tag{3.23}$$

for some $\lambda < 1$ and some $L < \infty$. Suppose that the minorization condition

$$P(x, \cdot\,) \geq \delta Q(\cdot), \qquad \text{for all } x \text{ with } V(x) \leq d \tag{3.24}$$

holds for some $\delta > 0$, some probability measure Q, and some d satisfying

$$d > \frac{2L}{1 - \lambda}. \tag{3.25}$$

Then for $0 < r < 1$ and any initial distribution ν of the Markov chain

$$\|\nu P^k - \pi\| \leq (1 - \delta)^{rk} + \left(\alpha^{-(1-r)} A^r\right)^k \left(1 + \frac{L}{1 - \lambda} + E_\nu V(X)\right)$$

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 (to appear). The drift condition (3.23) is slightly different from the ones previously described, but if V satisfies (3.23) then $1 + V$ satisfies (3.18) with $C = \{\, x : V(x) \leq d \,\}$ which is petite because of the minorization condition (3.24) and a slightly larger λ. Note that (3.25) 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 3.7.6 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

$$\frac{1}{\sqrt{2\pi}\tau} \exp\left(-\frac{(y-\rho x)^2}{2\tau^2}\right) \frac{1}{\sqrt{2\pi}\tau} \exp\left(-\frac{(|x|+\rho d/b)^2}{2\tau^2}\right) \tag{3.26}$$

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

$$\delta = 2\Phi\left(-\frac{\rho}{\tau}\sqrt{\frac{d}{b}}\right) < 2\Phi\left(-\rho\sqrt{\frac{2}{1-\rho^2}}\right), \tag{3.27}$$

where Φ is the standard normal cumulative distribution function, that is, (3.26) is a proper probability distribution times δ.

Note that if ρ is very close to one, then (3.27) 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.

3.7.8 Uniform Ergodicity

When the bound in the definition of geometric ergodicity is uniform, that is when there is a constant $R < \infty$ such that

$$\sum_{n=1}^{\infty} r^n \|P^n(x, \cdot) - \pi\| < R, \qquad \text{for all } x. \tag{3.28}$$

we say the chain is *uniformly ergodic*. This implies

$$\sup_{\text{all } x} \|P^n(x, \cdot) - \pi\| \to 0, \qquad \text{as } n \to \infty, \tag{3.29}$$

which Meyn and Tweedie take as the definition of uniform ergodicity. This makes sense because (3.29) also implies (3.28) 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 3.7.3. 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 stationary 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 $x \in S^{m+n+1}$

$$\|P^n(x, \cdot) - \pi\| \ge |P^n(x, S^m) - \pi(S^m)| = \pi(S^m)$$

Since the chain cannot get from S^{m+n+1} to S^m in only n steps. Hence

$$\sup_{\text{all } x} \|P^n(x, \cdot) - \pi\| \ge \pi(S^m)$$

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 $y > 0$

$$\Pr(Y_n \le 0 | Y_0 = y) \le \Phi(\rho^{2n} y/\sigma) \tag{3.30}$$

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 $|\Phi(\rho^{2n}y/\sigma) - \pi(Y \le 0)| \le \epsilon$

whenever $n \geq n_\epsilon$ for all y. Clearly, this can't hold since $\pi(Y \leq 0) = \frac{1}{2}$ and (3.30) converges to 1 as $y \to \infty$.

3.8 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 stationary 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

$$\sqrt{n}\,(\hat{\mu}_n - \mu) \xrightarrow{\mathcal{D}} N(0, \sigma^2). \tag{3.31}$$

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 (3.31) 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 stationary 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 stationary 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.

3.8.1 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 (3.31) 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

$$\gamma_k = \operatorname{Cov}(g(X_i), g(X_{i+k})) \tag{3.32}$$

is the same for all k. (3.32) is called the lag k autocovariance of the stationary time series $g(X_1)$, $g(X_2)$, ... Thus stationarity implies

$$\sigma_n^2 = \gamma_0 + 2 \sum_{k=1}^{n-1} \frac{n-k}{n} \gamma_k. \tag{3.33}$$

and σ_n^2 converges to

$$\sigma^2 = \gamma_0 + 2 \sum_{k=1}^{\infty} \gamma_k \tag{3.34}$$

as $n \to \infty$ if the series on the right hand side is summable. We can expect (3.34) to be the asymptotic variance if everything is well behaved.

3.8.2 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

$$\beta(n) = \tfrac{1}{2} \sup \sum_{i=1}^{I} \sum_{j=1}^{J} |\Pr(X_0 \in A_i \text{ and } X_n \in B_j) - \pi(A_i)\pi(B_j)|$$

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 (3.22) 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 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 (3.31) holds for the stationary chain, and the asymptotic variance is given by (3.34). 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 3.4. 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 10 *If a Markov chain is V-uniformly ergodic, then for any function g such that $g^2 \le V$ the central limit theorem (3.31) holds for the stationary chain, and the asymptotic variance is given by (3.34). 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 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 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

$$\begin{aligned} \rho(n) &= \sup_{u,v \in L^2(\pi)} \operatorname{Cor}\bigl(u(X_i), v(X_{i+n})\bigr) \\ &= \sup_{u \in L^2(\pi)} \sqrt{\frac{\operatorname{Var}\bigl(E\{u(X_{i+n}) | X_i\}\bigr)}{\operatorname{Var}\bigl(u(X_i)\bigr)}} \end{aligned} \tag{3.35}$$

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 1.3.5 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 1.3.5) 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 (3.35) 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 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 11 *If a Markov chain has an invertible Laplacian operator, then the central limit theorem (3.31) holds for the stationary chain, and the asymptotic variance is given by (3.34). 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 $L^2(\pi)$.

3.9 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.

3.9.1 Batch Means

Given a Markov chain X_1, X_2, ... and a function g for which there is a central limit theorem (3.31), 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.

3.9.2 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 Schmeiser, 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

$$\hat{\sigma}^2_{n,l} = \frac{l}{n-l+1} \sum_{j=1}^{n-l+1} (\hat{\mu}_{n,l,j} - \hat{\mu}_n)^2 \tag{3.36}$$

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.

3.9.3 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 $\text{Cor}(Y_i, Y_{i+k}) = \rho^{2k}$, thus the variance in the

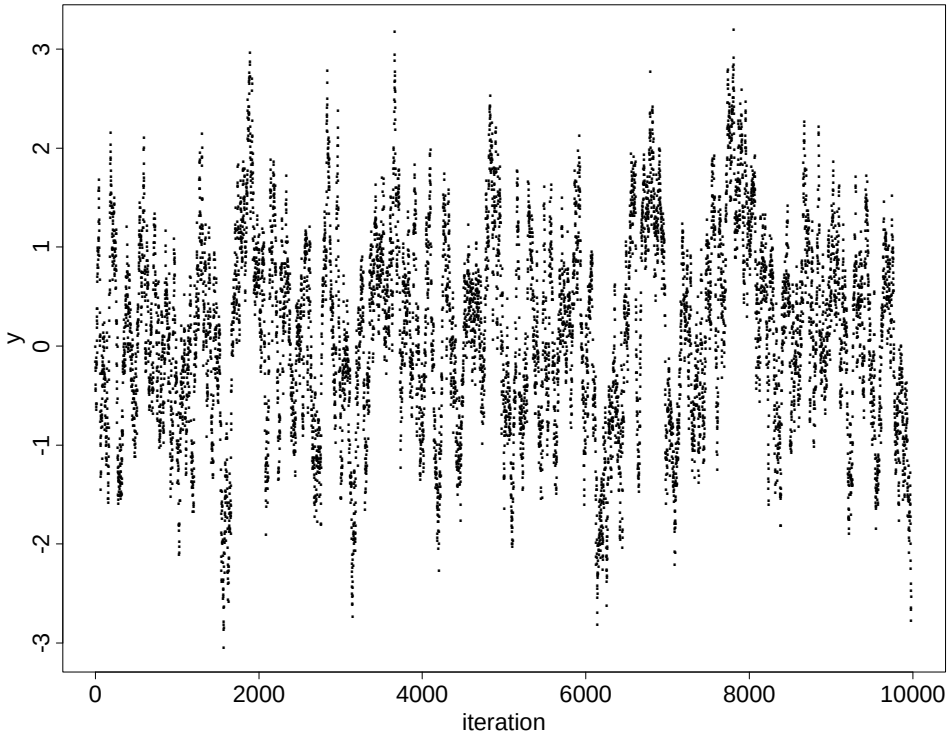


Figure 3.1: Output of the Gibbs sampler for the bivariate normal distribution with mean zero, variance one, and correlation $\rho = .99$. The starting position was $(0, 0)$ and the run length 10,000. The statistic plotted is the second component of the state vector.

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 3.1 shows a run of length 10,000 of a Gibbs sampler for the bivariate 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 3.7.6 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

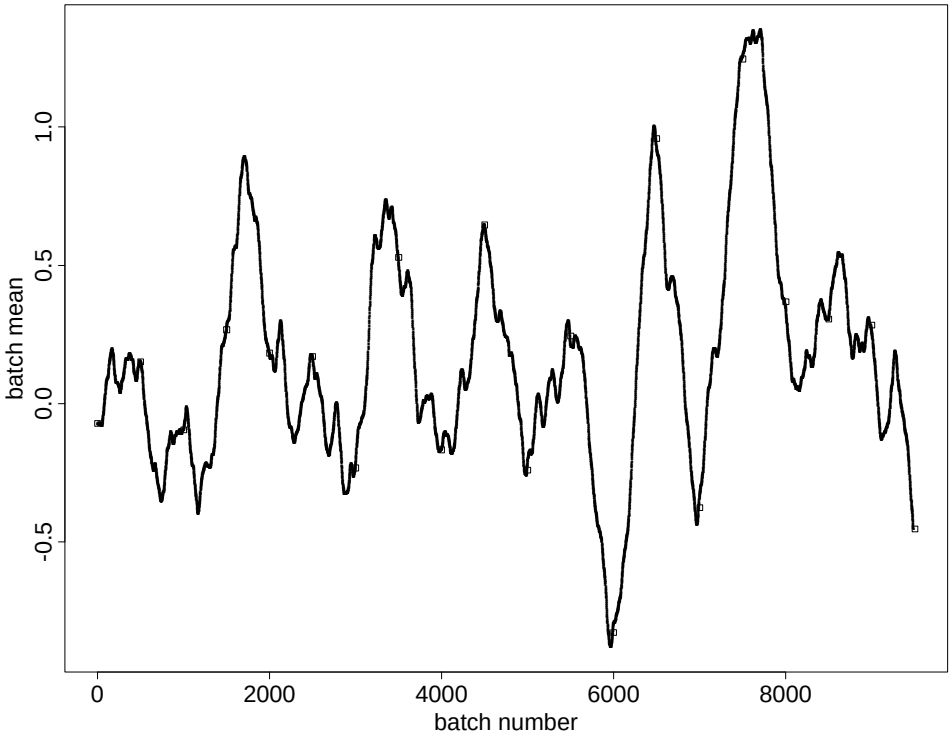


Figure 3.2: Overlapping batch means for the output shown in Figure 3.1. 9501 batches of length 500. Squares mark the 20 nonoverlapping batch means used in the ordinary batch means analysis.

bound on the convergence rate of the sampler in Section 3.7.7. A glance at Figure 3.1 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 stationary 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 3.2 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

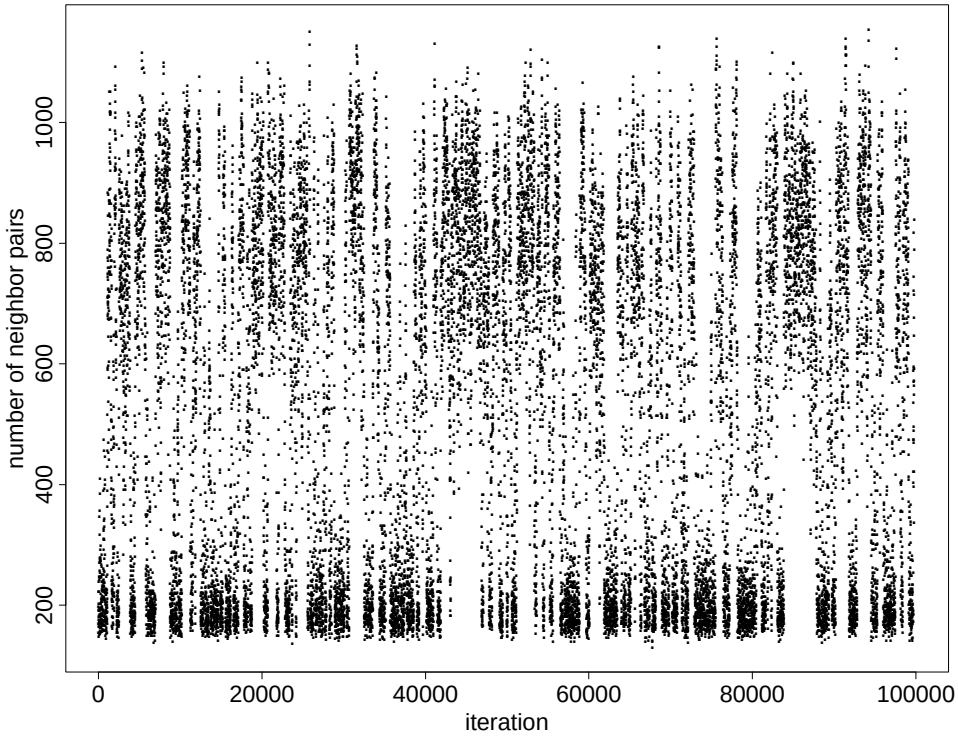


Figure 3.3: Metropolis sampler for the Strauss process with fixed number of points $n(x) = 50$ defined by (2.16) with canonical parameter $\beta = .126$. The vertical coordinate is the canonical statistic $s(x)$ which is the number of neighbor pairs. The run of length 100,000 was started at a realization of the Poisson process ($\beta = 0$). The plot only shows every fifth point, though all points were used in analyses.

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 (3.36) 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 3.3 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 $s(x)$.

3.9.4 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 (3.34) 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 (3.34) with $\hat{\gamma}_k$ plugged in for γ_k, but it has long been known that this estimator is not even consistent (see Priestley, 1981, p. 432). For large k the variance of $\hat{\gamma}_k$ is approximately

$$\operatorname{Var}(\hat{\gamma}_k) \approx \frac{1}{n} \left(\gamma_0^2 + 2 \sum_{m=1}^{\infty} \gamma_m^2 \right) \tag{3.37}$$

(Bartlett, 1946), assuming $\int g^4 \, d\pi < \infty$ and sufficiently fast mixing (ρ-mixing suffices). Figure 3.4 shows the estimated autocovariance function, $\hat{\gamma_k}$ as a function of k, with "large k confidence intervals calculated from (3.37) for the run shown in Figure 3.3.

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.

$$\hat{\sigma}^2 = \hat{\gamma}_0 + 2 \sum_{k=1}^{\infty} w(k) \hat{\gamma}_k \tag{3.38}$$

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

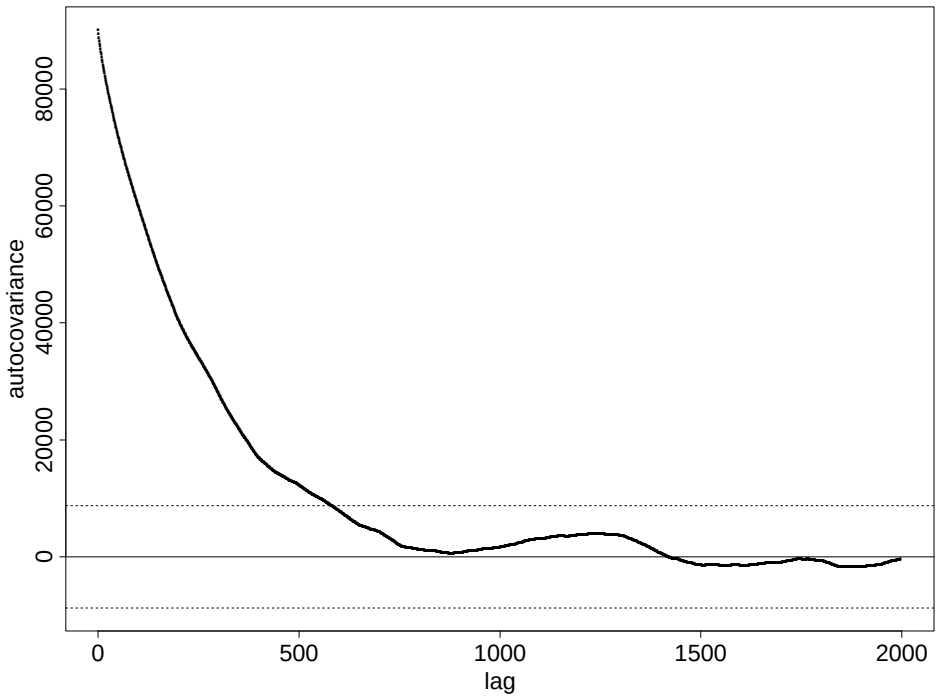


Figure 3.4: Empirical autocovariance function for the Metropolis sampler in Figure 3.3. The dotted lines are ± 1.96 times the asymptotic standard deviation of γ_k given by (3.37).

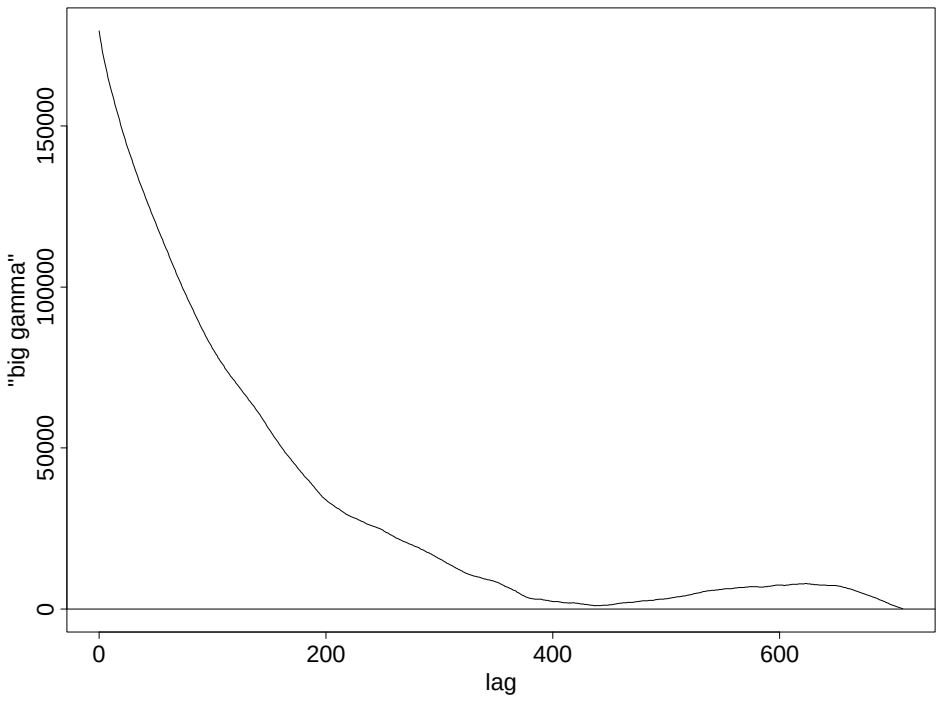


Figure 3.5: Plot of $\gamma_{2k} + \gamma_{2k+1}$ versus k for the Metropolis sampler in Figure 3.3.

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

$$\hat{\sigma^2} = \hat{\gamma}_0 + 2\hat{\gamma}_1 + \sum_{k=2}^{M} \hat{\Gamma}_k \tag{3.39}$$

where M is the largest integer such that the $\hat{\Gamma}_k$ are strictly positive for $k = 2$, ..., 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 $\hat{\Gamma}_k^{(\mathrm{mon})}$ in (3.39). This gives the *initial monotone sequence estimator.*

The smaller bulges that make Figure 3.5 nonconvex can also be eliminated by taking the function $k \mapsto \hat{\Gamma}_k^{(\mathrm{con})}$ to be the greatest convex minorant of $\hat{\Gamma}_1, \ldots, \hat{\Gamma}_M, 0$, and replacing $\hat{\Gamma}_k$ by $\hat{\Gamma}_k^{(\mathrm{con})}$ in (3.39). 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 3.3, 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.

3.10 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

$$P(x, \cdot\,) = P(y, \cdot\,), \qquad \text{for all } x, y \in \alpha. \tag{3.40}$$

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 (3.40) 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 stationary 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

$$E(Z_i) = E(N_i)\mu \tag{3.41}$$

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

$$\hat{\mu}_k = \frac{\bar{z}_k}{\bar{n}_k} = \frac{Z_1 + \cdots + Z_k}{N_1 + \cdots + N_k}. \tag{3.42}$$

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 (3.42) have the correct expectations by (3.41). 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

$$E\left(\sum_{i=1}^{K} Z_i\right) = \mu E\left(\sum_{i=1}^{K} N_i\right) \tag{3.43}$$

$$\text{Var}\left(\sum_{i=1}^{K} (Z_i - \mu N_i)\right) = \tau^2 E(K) \tag{3.44}$$

(3.43) is the analog of Wald's lemma with random stopping, and (3.44) 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

$$\begin{aligned}
E\left(\sum_{i=1}^{K} Z_i\right) &= \mu\nu E(K) \\
E\left(\sum_{i=1}^{K} N_i\right) &= \nu E(K) \\
\operatorname{Var}\left(\sum_{i=1}^{K} Z_i - K\mu\nu\right) &= \operatorname{Var}(Z_i)E(K) \\
\operatorname{Var}\left(\sum_{i=1}^{K} N_i - K\nu\right) &= \operatorname{Var}(N_i)E(K) \\
\operatorname{Cov}\left(\sum_{i=1}^{K} Z_i - K\mu\nu, \sum_{i=1}^{K} N_i - K\nu\right) &= \operatorname{Cov}(Z_i, N_i)E(K)
\end{aligned}$$

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.

3.10.1 Estimating the Asymptotic Variance

From (3.44)

$$\hat{\tau}_K^2 = \frac{1}{K}\sum_{i=1}^{K} \left(Z_i - N_i\hat{\mu}_K\right)^2 \tag{3.45}$$

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 3.10.7 for an example.

3.10.2 Splitting Markov Chains

Any Markov chain on a discrete state space has accessible atoms. Any point with positive probability is one since (3.40) 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 (to appear) apply the technique to Markov chain Monte Carlo. The construction below follows Mykland, Tierney and Yu (to appear) 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 $\int s\,d\pi > 0$

$$P(x, A) \geq s(x)\nu(A) \qquad \text{for all points } x \text{ and measurable sets } A. \tag{3.46}$$

This is similar to the minorization conditions (3.14) used in the definition of small sets and (3.24) 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 (3.14) in that one must minorize the kernel P rather than an iterated kernel P^m.

Condition (3.46) 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

$$\frac{P(X_t, \cdot\,) - s(X_t)\nu(\,\cdot\,)}{1 - s(X_t)} \tag{3.47}$$

which is a normalized probability distribution because of the minorization condition (3.46). 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 a stationary 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 $r(X_t, X_{t+1})$.

3.10.3 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 (2.8) becomes

$$R = \frac{h(y)q(x)}{h(x)q(y)}, \tag{3.48}$$

where $h(x)$ is an unnormalized density of the stationary 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 5.4. 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.

3.10.4 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 3.10.2 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

$$r_A(x,y) = \begin{cases} \max\left\{ \frac{c}{w(x)}, \frac{c}{w(y)} \right\}, & w(x) > c \text{ and } w(y) > c, \\ \max\left\{ \frac{w(x)}{c}, \frac{w(y)}{c} \right\}, & w(x) < c \text{ and } w(y) < c, \\ 1, & \text{otherwise.} \end{cases} \tag{3.49}$$

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 (3.48), 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 (3.49) 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 stationary distribution. This does not appear to be correct. See Section 3.10.6.

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 $X_1 = y$.

3.10.5 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 4.5. 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 stationary distribution to an algorithm with unknown and generally unknowable properties. One thing is clear from Theorem 7, restarting always increases the distance from the marginal distribution of X_t to the stationary 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 (3.48) preserves the stationary 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.

3.10.6 Splitting Metropolis-rejected Restarts

Let Q denote the kernel for the split independence chain update described in Section 3.10.4. It updates the state (X, S). Let P denote any other kernel that preserves the same stationary 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 stationary 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, ... do
  Simulate x from P(x', ·).
  Simulate y from q
  Simulate u Uniform(0,1)
  Calculate R given by (3.48)
  if (u < R) then
    x' = y
    Simulate u Uniform(0,1)
```

```
    Calculate r_A(x, y) given by (3.49)
    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 3.10.4. This gives the x' value needed at the top of the loop.

3.10.7 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 A. 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 $s(x)$. (3.49) can also be simplified to

$$r_A(x, y) = \begin{cases} \exp\{-\beta \min[t(x) - c', t(y) - c']\}, & t(x) > c' \text{ and } t(y) > c', \\ \exp\{-\beta \min[c' - t(x), c' - t(y)]\}, & t(x) < c' \text{ and } t(y) < c', \\ 1, & \text{otherwise.} \end{cases} \tag{3.50}$$

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 3.3 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 3.6 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 stationary 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 (3.45) 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.

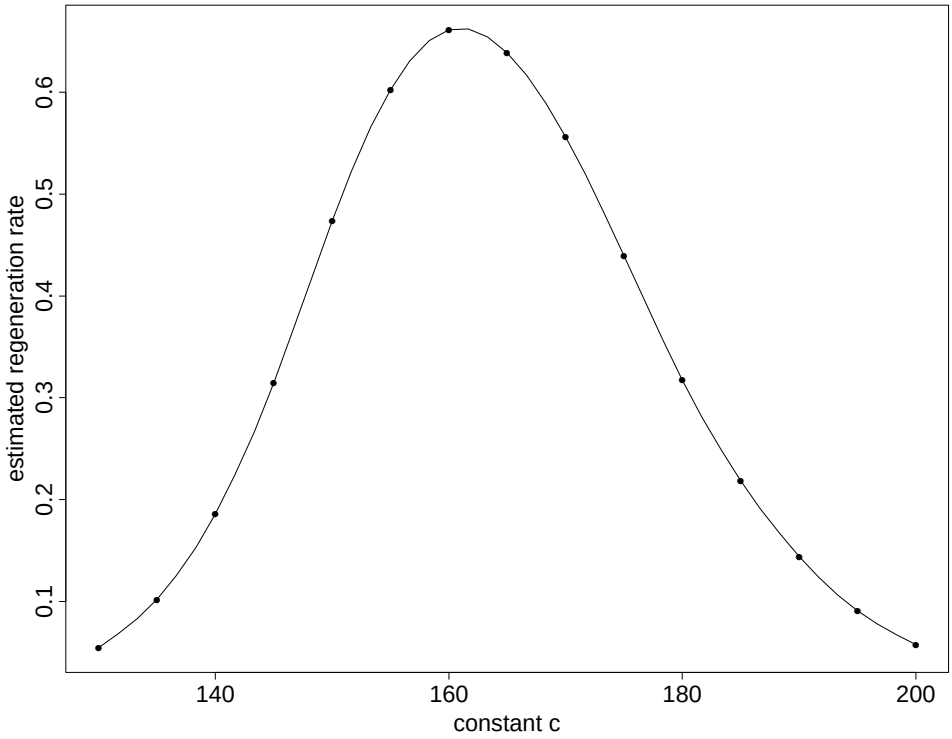


Figure 3.6: Expected regeneration rate versus the constant c' (3.50) for the Metropolis sampler with split Metropolis-rejected restarts for the Strauss process with 50 points $\beta = .126$. The horizontal coordinate is c' and the vertical coordinate is the estimated fraction of accepted restarts that will be regenerations.

Chapter 4

Running Markov Chains

> **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."
>
> —The Jargon File

This chapter is about the practice of Markov chain Monte Carlo, about the Right Things to do and to say about what one has done. "Right Thing" is meant in the hacker's sense explained in the epigraph. The practice of Markov chain Monte Carlo has been controversial (see Gelman and Rubin, 1992; Geyer, 1992; and the accompanying discussion), and it is likely that this chapter will not be the last word on the subject. The use of "Right Thing" and "Wrong Thing" concedes, as the definition says, that "reasonable people may disagree." Still it is important here, as everywhere in statistics, that we think about what is the Right Thing.

4.1 Many Short Runs

An old way of thinking about Markov chain Monte Carlo is what I have dubbed "many short runs." It does the following, start the chain with a sample from

some starting distribution μ_0. Run the chain m steps, throwing away all but the last iterate. This produces one sample from the distribution $\mu_0 P^m$. We know from the convergence of marginals that if the chain is Harris recurrent and aperiodic that $\mu_0 P^m$ can be made as close to the stationary distribution π as we please if we only take m large enough. Repeat this procedure n times producing n independent identically distributed samples from $\mu_0 P^m$. Compute estimates by averaging over these samples.

This is generally an absolutely horrible procedure, so bad that no one can now be found to defend it, although it was often recommended in papers written several years ago. The problem is that if m is taken very large, say 100,000, the procedure is so inefficient that is essentially precludes being able to do complicated problems. But if m is taken to be small, say 100, there will be no reason to believe that $\mu_0 P^m$ is close to the stationary distribution π. Even if we have bounds on $\|\mu_0 P^m - \pi\|$ from Theorem 8, they will generally be very conservative except in the easiest of problems and will not give a tight bound for m as small as 100. Moreover the procedure is still inefficient. Both wrong and inefficient, many short runs manages to do Markov chain Monte Carlo in a way that depends critically on information that is usually unknown and perhaps unknowable. It is clearly the Wrong Thing.

In hindsight it seems that the main reason why many short runs appealed to people is that it seemed to permit avoidance of Markov chain theory or indeed any theory of dependence in stochastic processes. The samples are independent (though from the wrong distribution $\mu_0 P^m$) and so the ordinary asymptotics of i. i. d. sampling seems to apply. A little more thought shows the situation is more complicated than that. The Monte Carlo estimate only converges to the correct answer if m and n both go to infinity. So this theoretical justification is also clearly the Wrong Thing; Markov chains should involve Markov chain theory.

4.2 One Long Run

Diametrically opposed to "many short runs" is "one long run." Start a chain at some point X_0 run for N steps and average over the samples. The starting position can be any point x_0 in the state space or can be a realization X_0 from any starting distribution μ_0. Similarly the length of the run can be any random integer N, though for the reasons given in Section 3.10 it seems best that N be a stopping time.

One long run is the Right Thing in MCMC. All of Markov chain theory can be directly applied to it. One can even say that anything else is not *Markov chain* Monte Carlo. A method that uses Markov chain updates but doesn't let the chain run isn't using Markov chains, it's fighting the Markov chain.

There are a variety of schemes intermediate between many short runs and one long run. They use some combination of three different ideas: subsampling, burn-in, and restarting. Subsampling and burn-in are special cases of one long run. Subsampling changes the Markov chain, but one still does one long run of

the new chain produced by subsampling. Burn-in is just one particular method of choosing a starting point for the run. Restarting is a special case of one long run only if the "restarting" is produced by regeneration. Any other form of restarting is not one long run.

4.3 Subsampling Markov Chains

A subsampled Markov chain with fixed spacing m, is the chain X_0, X_m, X_{2m}, ... obtained by taking every mth sample from the original chain. The subsampled chain is also a Markov chain. Its kernel is P^m, where P is the original kernel.

We can also have a chain with random spacing. If M is a random integer with distribution $P(M = m) = a(m)$, and M_1, M_2, ... are i. i. d. samples from this distribution, then X_0, X_{M_1}, $X_{M_1+M_2}$, ... obtained by taking samples with spacing M_1, M_2, ... is again a Markov chain with kernel (3.15).

Subsampling with random spacing has not been discussed in the applied literature, although it is used as a basic theoretical tool by Meyn and Tweedie (1993). It would be important if the chain were close to periodicity. Then subsampling with fixed spacing that is a multiple of the approximate period could result in a horrible loss of efficiency. Subsampling with a random spacing that is not periodic cannot interact badly with a periodic or almost periodic chain.

4.4 Starting Methods and "Burn-in"

Asymptotically, the starting position does not matter if the chain is Harris recurrent. The law of large numbers and the central limit theorem have the same limits regardless of the starting position, and if the chain is aperiodic the total variation convergence of marginals also does not depend on the starting position. If the starting position is a random variable X_0 from an arbitrary starting distribution μ_0, this also does not affect the asymptotics.

For a finite sample, the starting position does matter. Figure 4.1 illustrates the problem.

We know the stationary distribution for the variable plotted is standard normal. The section of the path before iteration 300 does not matter asymptotically, but unless the sample size is huge this "initial transient" causes large errors. 114 points have $y > 10$. The probability of this event under the stationary distribution is 1.3×10^{-2174}. The asymptotics will eventually make the initial transient irrelevant, but "eventually" is a very long time, more than 10^{2000} iterations.

To get good answers with practical sample sizes we must throw away the initial transient. Just be safe we might as well throw away a bit more, say everything before iteration 500. On my workstation this code takes 24 microseconds per iteration plus 2300 microseconds in initialization and wrap-up. So throwing away even 10,000 iterations would only lose a few seconds.

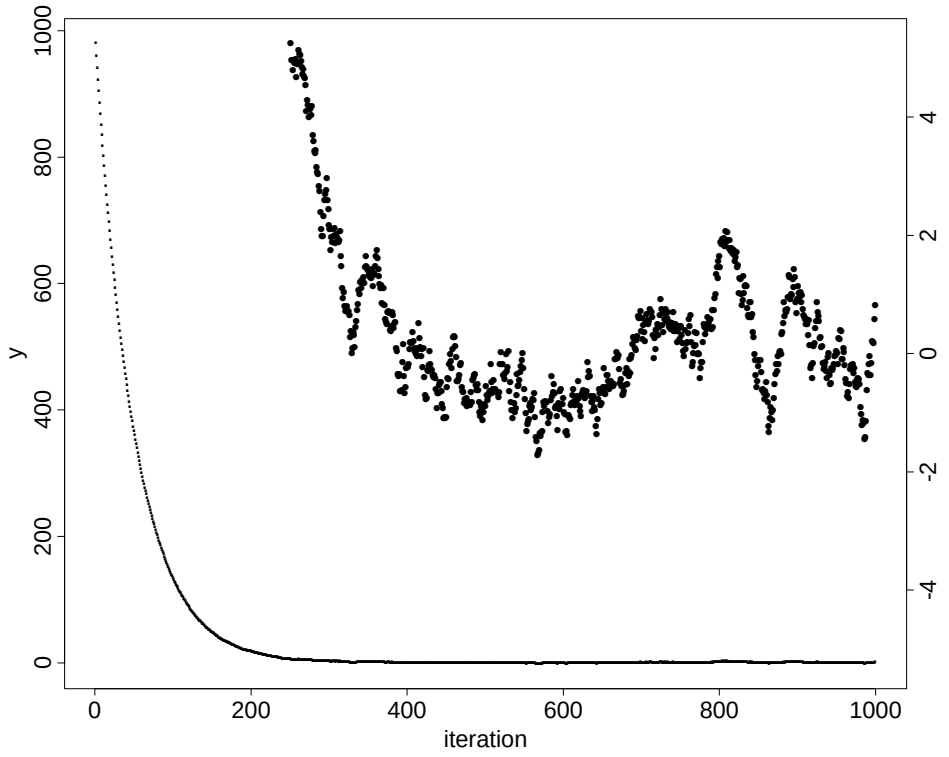


Figure 4.1: A run of the Gibbs sampler for the bivariate normal with the same stationary distribution as in Figure 3.1 but started at $(1000, 1000)$ instead of $(0, 0)$. The second component y of the state vector (x, y) is plotted. Small dots show the whole run (axis at left). Large dots show the portion of the run from iteration 251 to the end (axis at right).

So the question arises, how long do we have to run to avoid an initial transient. This period for which the sampler is allowed to run without generating output has been called "burn-in" or "warm-up." In that terminology, the question becomes "How much burn-in do we need?"

This is not the usual formulation of the burn-in question. What is usually asked is "how long must the chain be run until it reaches equilibrium?" Of course it never does reach equilibrium. The marginal distribution of X_n is never exactly π unless the chain was started with a realization from π. But we could ask how long the chain must be run to reduce $\|\mu_0 P^n = \pi\|$ to less than some prespecified tolerance. The people who formulate the burn-in question this way never say why one would want to start the run with an X_n whose marginal distribution is near π. With many short runs it is critical that the distribution of X_n be as near π as possible, but with one long run it does not matter. All that is required is that the starting position not be so far out in the tail that there is an "initial transient" that will require an extremely long run to wash out. If we start the run at some point that would occur with reasonable probability in the sample of the size we intend to collect, that is enough. There is no bonus for "starting when the chain has reached equilibrium."

It is important to realize that the burn-in question as posed usually has no answer. Unless the Markov chain is uniformly ergodic, there is no fixed amount of burn-in that will do the job for all starting points. This could be taken as the definition of uniform ergodicity. By Theorem 16.2.1 in Meyn and Tweedie (1993) a chain is uniformly ergodic if and only if there exists a petite set C and an $M < \infty$ such that the expected time for the chain to hit C started at x is less than M for all x. For the Gibbs sampler for the bivariate normal we know the expectation of Y_n given Y_0 is $\rho^{2n} Y_0$. So even if we decide on a burn-in of 10^9 iterations. This will not be enough if the chain starts at $Y_0 = 10^{10^{10}}$.

Even if the chain is uniformly ergodic, so the "burn-in question" does have a finite solution, the answer may be entirely useless. Consider the Gibbs sampler for the Ising model, an archetype of a slowly mixing Markov chain.

The symmetric Ising model with $\alpha = 0$ has what the physicists call a *phase transition.* The probability distribution induced by the model can be radically different for parameter values that are only slightly different. More precisely, there is a value β_c for the second canonical parameter such that for $\beta < \beta_c$, the distribution of $t_1(X)$ for a square lattice of size n with free or periodic boundary conditions converges to a unimodal distribution as $n \to \infty$. Conversely, if $\beta > \beta_c$, then the distribution of $t_1(X)$ converges to a bimodal distribution. Kindermann and Snell (1980) give a very readable introduction to these issues. The critical value of β is $\beta_c = \frac{1}{2} \sinh^{-1}(1) = \frac{1}{2} \log(1 + \sqrt{2}) = 0.4406868$.

If we keep the lattice size fixed and let β vary, the distribution of $t_1(X)$ makes a smooth transition from unimodal for $\beta < \beta_c$ to bimodal for $\beta > \beta_c$. The sharp transition only occurs in the limit as $n \to \infty$. For lattices of even moderate size, any Markov chain Monte Carlo scheme that only updates a single site per elementary update step will be very slowly mixing. Figure FigIsing illustrates the problem. In a run of 10,000 iterations, the sampler only makes one crossing from one mode to the other. For larger lattice sizes the problem

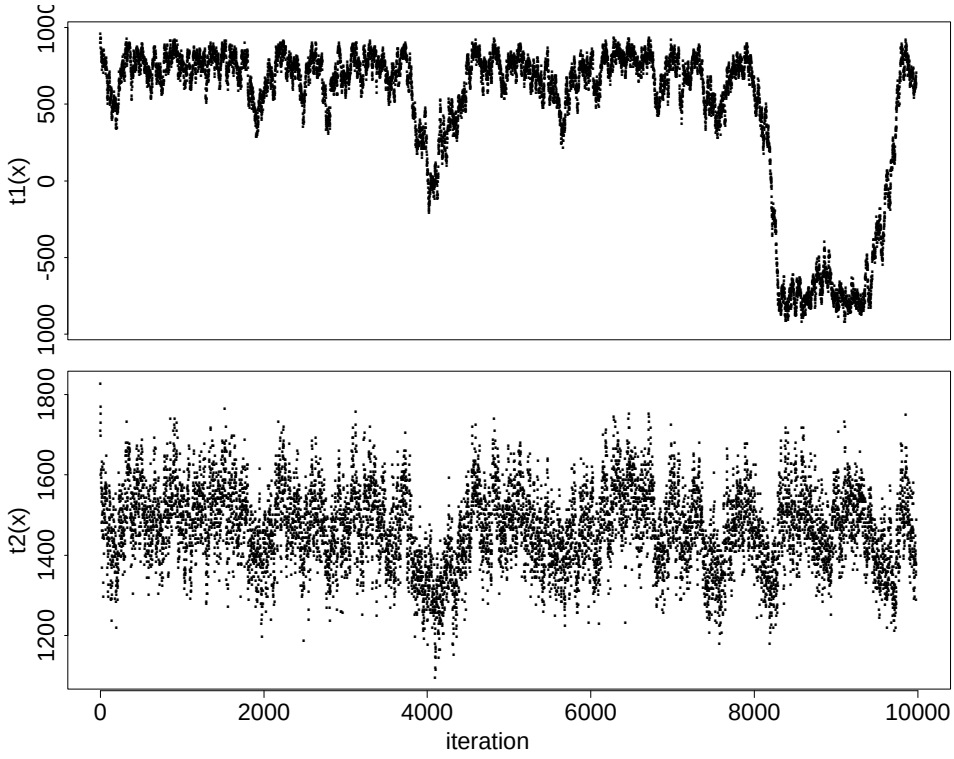


Figure 4.2: A run of the Gibbs sampler for the Ising model on a 32×32 toroidal lattice with critical parameter values, $\alpha = 0$, $\beta = .44$, started at the all-black image. The top panel is a time series plot of values of the first canonical statistic $t_1(X)$, black minus white pixels. The bottom panel is a time series plot of the second canonical statistic $t_2(X)$, concordant minus discordant neighbor pairs.

Table 4.1: Hitting times for the Gibbs sampler for the Ising model for various lattice sizes. Time for the Gibbs sampler for the Ising model with $\alpha = 0$ and $\beta = 2\beta_c = .88$ started at a random realization from the Ising model with $\alpha = \beta = 0$ to reach the set $t_1(x) \geq \mu$, where mu is the approximate expectation of $t_1(X)$ calculated using a run of 10,000 iterations of the Swendsen-Wang algorithm (Section 5.2). The first column gives lattice size ($n \times n$ lattice), the other columns give the minimum, median, mean, and maximum of 50 runs.

n	min	median	mean	max
16	9	36.0	109.02	746
32	44	135.5	725.88	5105
64	172	545.5	7432.58	77361
128	640	2634.0	76846.58	525720

gets much worse. Any sampler that only updates a single pixel at a time is useless for large lattices.

There is a simple fix that eliminates the worst of the problem. If $\alpha = 0$ and the distribution is symmetric, we can use the symmetry taking

$$\hat{\mu}_n = \frac{1}{2n} \sum_{i=1}^{n} g(X_i) + g(-X_i)$$

to be our estimate of $E_\pi g(X)$, where $-x$ is the image obtained from x by reversing the colors of all pixels. Even when $\alpha \neq 0$ a trick called "mode jumping" can be used. At the end of each scan propose to move from x to $-x$ and use Metropolis rejection to maintain the correct stationary distribution, the odds ratio being

$$R = \frac{h(-x)}{h(x)} = e^{-2\alpha t_1(x)}$$

because $t_1(-x) = -t_1(x)$ and $t_2(-x) = t_2(x)$. It is not clear whether this mode jumping idea has wide applicability. There seem to be other applications beside the Ising model. Usually there is no simple proposal like x to $-x$ for the Ising model that jumps between modes.

Even for the Ising model this trick of symmetrizing or mode jumping only fixes the worst problems. Though the sampler works much better, it is still bad. In particular, the burn-in problem is intractable.

It is clear that the notion of finding a fixed amount of burn-in that will suffice for all runs is hopeless. The distribution of the times to hit the part of the state space that has high probability under the model has a very long tail. Since the median remains small, it is clear that a moderate amount of burn-in, say 1,000 iterations will work a fair amount of the time, but it is clear that it takes an enormous amount of burn-in, more than 1,000,000 iterations to be sure of convergence most of the time for the 128×128 lattice. The situation rapidly becomes wore with larger lattices.

The burn-in problem goes away if we start from a better starting point. From the all-black image, the high-probability region of the state space was hit in less than 10 iterations for 50 runs for each of these lattice sizes.

The moral of the story is that if we start from a good starting point almost no burn-in is necessary, but if we start from a bad starting point there is no telling how much burn-in is required. This suggests that the "burn-in" question is a problem that has been studied from the wrong perspective. If a sampler is not uniformly ergodic, there is no fixed amount of burn-in that is sufficient. Even for samplers that are uniformly ergodic, so there is in principle some fixed amount of burn-in that is sufficient, the amount may be so astronomically large that it is useless in practice. Thus the emphasis should be placed on finding a good starting point.

It is fortunately not necessary that one be able to easily construct a good starting point. It is only neccessary that one be able to recognize a good starting point when the sampler reaches it. Suppose there is a subset C of the state space that we think consists of useful starting points. If we start the chain a some point x_0 and let it run until the first time τ_C that it hits the set C. We can then take the position X_{τ_C} where the chain enters C as the starting point of the run. There is no reason why this random starting position X_{τ_C} is worse than X_n for some fixed burn-in time n. In practice it is much better. If one starts in C one is not "out in the tail" of π and there is no initial transient.

Nothing in this analysis tells one how to find good starting points. The only point I am trying to make is that burn-in does not relieve one of the necessity of determining good starting starting points.

4.5 Restarting Markov Chains

Once we have started questioning "many short runs" the question arises as to why the Markov chain is restarted at all. Theorem 7 says that restarting always increases the distance from the marginal distribution of X_t to the stationary distribution π, and this agrees with the intuition that the chain has gotten approximate stationarity it is senseless to destroy the stationarity by restarting.

There is a research tradition that poses the following question, in which restarting seems to make sense. Suppose an update mechanism corresponding to a kernel P, a starting distribution μ_0, and a number of iterations n are fixed. How does one make choices of numbers b, d, and m so that if we restart m times, throw away b samples at the beginning of each run, and use every dth iteration, we will have optimal performance according to some loss function? Say we choose total variation norm, so the question is what choices of b, d, and m make

$$\left\| \sum_{i=1}^{\lceil n/m \rceil} \sum_{j} \mu_0 P^{b+mj} \right\|$$

It is often said that restarting may give better answers. In the worst case, suppose the chain does not mix at all, just stays put wherever it is started. This

chain is, of course, not irreducible, but there are irreducible chains arbitrarily close to this worst case in behavior. Then the estimate of a function $g(x)$ from one run of the chain is $g(X_0)$, the function evaluated at the starting position. The estimate obtained from a chain restarted n times using a restarting distribution Q is $g(X_0) + \sum_{i=1}^{n} g(X_i)$ where the X_i are independent samples from Q (the restarts). Clearly this can be a better answer than $g(X_0)$, but unless $Q = \pi$ this "better" answer is wrong.

In general, the analysis of when restarting might produce a better answer is very difficult (Kelton and Law, 1984; Fishman, 1991) and in any case misses the point in practical applications of Markov chain Monte Carlo. If a sampler is so slowly mixing that one run of reasonable length does not give satisfactory answers, then the solution is not restarts, which will also give unsatisfactory answers, but a different sampler that mixes better. The Metropolis-Hastings algorithm gives enormous scope for inventing samplers. There are many more ideas to try when the first one fails.

Despite some heated discussion in the literature of the validity of restarting (Gelman and Rubin, 1992; Geyer, 1992; and the accompanying discussion), there seems to be no theory that could be applied to practical examples that justifies restarting Markov chains, there seems to be no example where pure restarts have been shown to be superior to Metropolis-rejected restarts, and it is hard to imagine an example in which restarts were superior to methods such as simulated tempering (Section 5.3) using the distribution $q(x)$ as a "hot" distribution rather than for restarts.

Chapter 5

Tricks and Swindles

"Swindle" is used in classical Monte Carlo to refer to any cute trick that seems to get something for nothing. It hasn't been used much in the Markov chain Monte Carlo literature. A more colorless term is "variance reduction" but that is so mild as to be misleading in Markov chain Monte Carlo. In a situation in which a swindle is needed in MCMC the sampler that has been implemented may not converge or at least there is worry that it doesn't converge. Being irreducible, it does, of course, converge eventually. What is meant is that there is very low probability of obtaining a representative sample of the stationary distribution in the number of iterations we are willing to contemplate. So we are trying to get an improvement in convergence time from perhaps 10^20 to 10^3 iterations.

Reduction in variance doesn't measure the kind of efficiency gains that are needed here. Suppose we are trying to calculate the probability of a set. Then since probabilities are between zero and one, the variance of our estimator can never be greater than one, no matter how far the chain is from convergence. We are trying to get the variance down to something reasonably small, say 10^{-4}. So we need a factor of 10^4 in variance reduction, but we need a factor of 10^{17} improvement in running time to get it. This just says that it doesn't make much sense to talk about variance before asymptotics kicks in and the square root law becomes valid.

Is it possible to get improvements by enormous factors like 10^{17}? There are two algorithms that are know do give that kind of improvement in some problems. No doubt there are other algorithms still to be invented with similar properties. The first of these, the Swendsen-Wang algorithm deals with a generalization of the Ising model called the Potts model.

5.1 The Potts Model

The Potts model (Potts, 1952) is a generalization of the Ising model that allows more than two values for the random variables at the lattice sites and also for

interations between lattice sites that are not nearest neighbors. At each lattice site there is a variable x_i taking values in a finite set. We will call the elements of this set "colors" thinking of the state vector $x = \{ x_i : i \in S \}$ as an image with pixels of different colors. A Potts model has an unnormalized density of the form

$$h(x) = \exp\left\{ \sum_i \alpha_i(x_i) + \sum_{i<j} \beta_{ij} 1_{[x_i = x_j]} \right\} \tag{5.1}$$

where for each lattice site i there is a real-valued function $alpha_i$ on the set of possible colors and for each pair of lattice sites i and j there is a real number β_{ij}.

Note that there is no mathematical structure imposed on the set of colors. The model treats all the colors symmetrically, and the interaction term only depends on whether two colors are the same or different. There is no notion of some colors being more similar than others. Thus any finite set can be substituted for the colors. Moreover, the model does not require any particular structure of the lattice. The lattice sites can also be an arbitrary set. All that is required is the specification of a β_{ij} for each pair of lattice sites.

Allowing α_i to depend on i is important in applications in image processing, where the task is to reconstruct the true image given a corrupted version. Then α_i incorporates the information about the ith pixel obtained from the corrupted image.

To derive the Ising model as a special case of the Potts model, let x_i take values in $\{-1, 1\}$, let $\alpha_i(x_i) = \alpha x_i$, and let $\beta_{ij} = 0$ unless $i \sim j$, in which case $\beta_{ij} = \beta$. Then we get

$$h(x) = \exp\left\{ \alpha \sum_i x_i + \beta \sum_{i \sim j} 1_{[x_i = x_j]} \right\}$$

Since

$$\begin{aligned} \sum_{i \sim j} x_i x_j &= \sum_{i \sim j} 1_{[x_i = x_j]} - \sum_{i \sim j} 1_{[x_i \neq x_j]} \\ &= 2 \sum_{i \sim j} 1_{[x_i = x_j]} - N \end{aligned}$$

where N is the number of lattice sites. Comparing with (3.1) we see that this special case of the Potts model is an Ising model. The only difference is that the parameter value β for the Ising model with parametrization (3.1) corresponds to 2β with the Potts model parametrization.

Like the Ising model the Potts model exhibits phase transitions. For a symmetric model $\alpha_i(x) \equiv 0$ with $\beta_{ij} = \beta$ for first nearest neighbor interactions and $\beta_{ij} = 0$ otherwise and r colors, the critical parameter value is $\beta_c = \log(1 + \sqrt{r})$.

5.2 The Swendsen-Wang Algorithm

The Swendsen-Wang algorithm (Swendsen and Wang 1987) was the first of a family of algorithms now called "cluster" algorithms by statistical physicists (Wang and Swendsen, 1990). They can be applied to any Potts model in which the dependence is attractive, that is, all the β_{ij} are nonnegative.

The clever trick in the Swendsen-Wang algorithm is the addition of a large number of variables called "bonds" to the state space, one bond for each pair of variables. For each pair i and j there is a random variable y_{ij} taking values in $\{0, 1\}$. The value $y_{ij} = 1$ indicates there is a "bond" between pixels i and j.

In order for this addition of variables to be useful for MCMC, there must be a stationary distribution π defined on the whole state space (pixels and bonds) such that we can easily learn about the distribution of interest, the Potts model for the pixels, by sampling π. For the Swendsen-Wang algorithm the joint distribution of the pixels and bonds is specified by specifying the marginal distribution of the pixels to be the Potts model (5.1) and then specifying the conditional distribution of the bonds given the pixels. This gives the pixels the marginal distribution of interest.

The conditional distribution of the bonds given the pixels is particularly simple. The bonds are conditionally independent given the pixels, the distribution of a single bond being

$$P(y_{ij} = 1|x) = \begin{cases} \gamma_{ij}, & \beta_{ij} > 0 \text{ and } x_i = x_j \\ 0, & \text{otherwise} \end{cases}$$

where the γ_{ij} are constants to be determined later. The Swendsen-Wang algorithm proceeds by "block" Gibbs sampling the two sets of variables pixels and bonds. It first samples from the conditional distribution of bonds given pixels just described and then samples from the conditional distribution of pixels given bonds, which we now have to figure out.

First note that any two bonded pixels must be the same color, hence so must any two pixels connected by a chain of bonds. Thus if we think of the set of lattice sites as the vertices of a graph and the bonds as edges, any maximal connected component of the graph must have all pixels the same color. A maximal connected component is a set of vertices that cannot be divided into two subsets with no edges connecting them and that is not a subset of any larger set with the same property. Call the maximal connected components "patches." These can be found by standard computer science algorithms. The function `sw.c` described in Appendix A is an example. Conditional on the bonds a patch must have all its pixels the same color, so the only randomness in the conditional distribution of pixels given bonds is the colors of the patches.

The joint distribution of pixels and bonds is

$$P(x, y) \propto \prod_i e^{\alpha_i(x_i)} \prod_{i<j} e^{\beta_{ij} 1_{[x_i = x_j]}} 1_{[(\beta_{ij} > 0 \text{and} x_i = x_j) \text{or} y_{ij} = 0]} \left(\gamma_{ij}^{y_{ij}} (1 - \gamma_{ij})^{1 - y_{ij}} \right)^{1_{[\beta_{ij} > 0 \text{and} x_i = x_j]}}$$

This is also the conditional distribution of pixels given bonds when considered as a function of x for fixed y. Let $\mathcal{A}$ denote the set of patches and x_A the color

of patch A. Note that $P(x|y) = 0$ unless for each i and j either $y_{ij} = 0$ (the pixels are not bonded) or $\beta_{ij} > 0$ and $x_i = x_j$ (the pixels are bonded hence they must be the same color). For possible y, those satisfying this condition,

$$\begin{aligned} P(x|y) &\propto \left\{ \prod_{A \in \mathcal{A}} \exp\left[\sum_{i \in A} \alpha_i(x_A) \right] \prod_{\substack{i,j \in A \\ i<j}} e^{\beta_{ij}} \left(\gamma_{ij}^{y_{ij}} (1 - \gamma_{ij})^{1 - y_{ij}} \right)^{1_{[\beta_{ij} > 0]}} \right\} \\ &\quad \times \left\{ \prod_{\substack{A,B \in \mathcal{A} \\ A<B}} \prod_{\substack{i \in A \\ j \in B}} e^{\beta_{ij} 1_{[x_i = x_j]}} (1 - \gamma_{ij})^{1_{[\beta_{ij} > 0 \text{and} x_i = x_j]}} \right\} \\ &\propto \left\{ \prod_{A \in \mathcal{A}} \exp\left[\sum_{i \in A} \alpha_i(x_A) \right] \right\} \left\{ \prod_{\substack{A,B \in \mathcal{A} \\ A<B}} \prod_{\substack{i \in A \\ j \in B}} e^{\beta_{ij} 1_{[x_i = x_j]}} (1 - \gamma_{ij})^{1_{[\beta_{ij} > 0 \text{and} x_i = x_j]}} \right\} \end{aligned}$$

because $x_i = x_j$ whenever i and j are in the same patch and $y_{ij} = 0$ whenever i and j are not in the same patch, and because the terms dropped in going from the second line to the third are constant (do not depend on x). The notation $A < B$ means that each pair of patches only enters once.

We now choose γ_{ij} to make the last term cancel

$$1 - \gamma_{ij} = e^{-\beta_{ij}}$$

which gives

$$P(x|y) \propto \prod_{A \in \mathcal{A}} \exp\left[\sum_{i \in A} \alpha_i(x_A) \right]$$

This says that with these particular choices of the γ_{ij} the colors of the patches are conditionally independent given the bonds, and the probability $x_A = x$ is proportional to $\exp\left[\sum_{i \in A} \alpha_i(x)\right]$.

Thus the Swendsen-Wang algorithm performs block Gibbs updates in which the bonds are conditionally independent given the pixels and the colors of patches are conditionally independent given the bonds. It is thus easy to carry out given code for finding maximal connected components of a graph. Even this in not necessary. There is a variant called Wolff's algorithm (see Wang and Swendsen, 1990) that updates only a single patch in each elementary update step.

Despite its simplicity, the Swendsen-Wang algorithm is certainly nontrivial. Statistical physicists worked on simulation of Potts models and similar spatial lattice processes for three decades before its discovery.

5.3 Simulated Tempering

Simulated tempering

5.4 Importance Sampling

Chapter 6

Likelihood Inference and Optimization

This chapter deals using MCMC to do likelihood inference in problems where the likelihood cannot be evaluated analytically and must be approximated by Monte Carlo. It also deals with other optimization problems in which the objective function (the function to be minimized or maximized) cannot be evaluated analytically and must be approximated by Monte Carlo.

6.1 Likelihood in Normalized Families

The kind of problem to which Monte Carlo likelihood inference was first applied is that of Gibbs distributions or Markov random fields. Examples are the Ising and Potts models and the Strauss process. If we restrict ourselves to finite-volume models that can be simulated on a computer these are just exponential families with untractable normalizing constants. They have unnormalized densities with respect to some measure λ on the state space of the form

$$h_\theta(x) = e^{\langle t(x), \theta \rangle},$$

where $t(x) = (t_1(x), \ldots, t_d(x))$ is a d-dimensional statistic and $\theta = (\theta_1, \ldots, \theta_d)$ is a d-dimensional parameter, called the *canonical* or *natural* statistic and parameter. The notation $\langle t(x), \theta \rangle$ denotes the standard inner product on $\mathbb{R}^d$

$$\langle t(x), \theta \rangle = \sum_{i=1}^{d} t_i(x) \theta_i.$$

The normalized densities of the family are

$$f_\theta(x) = \frac{1}{c(\theta)} h_\theta(x), \tag{6.1}$$

where the normalizing constant $c(\theta)$ is defined by

$$c(\theta) = \int h_\theta(x)\mu(dx). \tag{6.2}$$

What makes likelihood inference for these models hard is that the integral (6.2) is analytically intractable, so the log likelihood

$$l(\theta) = \log f_\theta(x) = \langle t(x), \theta \rangle - \log c(\theta)$$

is also analytically intractable. Neither the log likelihood or its derivatives can be calculated exactly, and this makes ordinary methods of likelihood inference impossible. Monte Carlo methods for such problems have been given by Ogata and Tanemura (1981, 1984, 1989), by Penttinen (1984), and by Geyer and Thompson (1992).

It turns out that much of theory of Monte Carlo likelihood does not use any properties of exponential families, so Geyer (1994) defines the following generalization, called there a *normalized family*. Let $\{ h_\theta : \theta \in \Theta \}$ be any family of unnormalized densities with respect to some measure λ on the sample space, that is for each θ, we have $h_\theta(x) \ge 0$ for all x and $\int h_\theta(x)\lambda(dx)$ is finite. Define the normalizing constant by (6.2) and the normalized densities by (6.1). These families are a natural generalization for Monte Carlo likelihood because any distribution specified by an unnormalized h_θ can be simulated by MCMC. The log likelihood of the family is

$$l(\theta) = \log \frac{f_\theta(x)}{f_\psi(x)} = \log \frac{h_\theta(x)}{h_\psi(x)} - \log \frac{c(\theta)}{c(\psi)} \tag{6.3}$$

As always we are free to add any term to the log likelihood that does not depend on θ. Here we subtract $\log f_\psi(x)$ where ψ is any point in the parameter space.

As it stands (6.2) cannot be evaluated by Monte Carlo, because it is not an integral with respect to a probability measure. Thus we rewrite it as an integral with respect to P_ψ, the probability distribution with density f_ψ.

$$\begin{aligned} c(\theta) &= \int h_\theta(x)\lambda(dx) \\ &= c(\psi) \int \frac{h_\theta(x)}{h_\psi(x)} f_\psi(x)\lambda(dx) \\ &= E_\psi \frac{h_\theta(X)}{h_\psi(X)} \end{aligned} \tag{6.4}$$

This formula is only correct if P_ψ dominates P_θ, that is, $h_\psi(x) = 0$ implies $h_\theta(x) = 0$ for all θ. Then both integrals in (6.4) can be taken over the set $\{ x : h_\psi(x) \ne 0 \}$ to avoid division by zero.

Combining (6.3) and (6.4) we get

$$l(\theta) = \log \frac{h_\theta(x)}{h_\psi(x)} - \log E_\psi \frac{h_\theta(X)}{h_\psi(X)} \tag{6.5}$$

Having written the intractable integral as an expectation, we can calculate it by Monte Carlo

$$\begin{aligned} l_n(\theta) &= \log \frac{h_\theta(x)}{h_\psi(x)} - \log \mathbb{E}_{n,\psi} \frac{h_\theta(X)}{h_\psi(X)} \\ &= \log \frac{h_\theta(x)}{h_\psi(x)} - \log \left(\frac{1}{n} \sum_{i=1}^{n} \frac{h_\theta(X_i)}{h_\psi(X_i)} \right) \end{aligned} \tag{6.6}$$

where X_1, X_2, ... are a Markov chain with stationary distribution P_ψ, and $\mathbb{E}_{n,\psi}$ denotes expection with respect to the empirical distribution of the sample, the probability measure that puts mass $1/n$ at each X_i.

If the chain is Harris recurrent, then for each θ in the parameter space, $l_n(\theta) \xrightarrow{\text{a. s.}} l(\theta)$, but the null set of sample paths for which convergence fails may depend on θ. Since a countable union of null sets is a null set, we can say that for any countable subset Θ_c of the parameter space, that for almost all sample paths

$$l_n(\theta) \to l(\theta), \qquad \text{for all } \theta \in \Theta_c.$$

That is, we have simultaneous pointwise convergence of l_n to l on Θ_c. More than that we cannot get without some continuity assumptions about the unnormalized densities. Geyer (1994) proves the following.

Theorem 12 *For a normalized family of densities $\{ h_\theta : \theta \in \Theta \}$, if the parameter set Θ is a separable metric space, the evaluation maps $\theta \mapsto h_\theta(x)$ are*

(a) *lower semicontinuous at each θ except for x in a P_ψ nullset that may depend on θ and*

(b) *upper semicontinuous for the observed x,*

and if the Markov chain sampler is Harris recurrent, then the Monte Carlo log likelihood (6.6) hypoconverges to the exact log likelihood (6.5) with probability one. Also (6.5) is upper semicontinuous and the normalizing constant (6.2) is lower semicontinuous.

The theorem as stated in Geyer (1994) contains a superfluous condition that the maps $\theta \mapsto h_\theta(x)$ be upper semicontinuous at almost all x rather than just at the observed x. This was assumed only to prove measurability of certain suprema, but this measurability is automatic when the parameter space is a Borel subset of a complete seperable metric space.

The main conclusion of the theorem is that l_n hypoconverges to l with probability one. This is a form of convergence of functions that is weaker than uniform convergence and is exactly the right form for optimization. Geyer (1994) gives several equivalent definitions, one of which is that l_n hypoconverges to l if both of the following conditions hold

(i) For every $\theta \in \Theta$ and every sequence $\theta_n \to \theta$

$$\limsup_{n \to \infty} l_n(\theta_n) \le l(\theta).$$

(ii) For every $\theta \in \Theta$ there is some sequence $\theta_n \to \theta$ such that

$$\liminf_{n\to\infty} l_n(\theta_n) \geq l(\theta).$$

More can be found in Attouch (1984). The two conditions are very different. Condition (i) is one-sided continuous convergence. It implies locally uniform convergence: for any $r > l(\theta)$, there is a neighborhood W of θ such that

$$l_n(\varphi) < r, \qquad \text{whenever } \varphi \in W.$$

Condition (ii) is weaker than pointwise convergence. It does not even imply $l_n(\theta) \to l(\theta)$.

Although

Appendix A

Computer Code

Code for examples used in the book is found in the directory

`~charlie/Isles/Stat8931.S95/Text`

(This is a local directory at the University of Minnesota School of Statistics. The code is not currently available on the web.)

The README file in this directory gives a brief explanation of how to run the code if one has `Splus`. This probably doesn't work after all this time because of changes to `Splus`.

Currently available code is described below.

Directory	*File*	*Sampler*
`VarComp`	`gibbs.f`	Gibbs sampler for variance components model from Gelfand and Smith (1990).
`VarComp`	`block.f`	Block Gibbs sampler for the same variance components model.
`VarComp`	`jump.f`	Sampler for model selection comparing the same variance components model to the model with all group means the same. Uses Gibbs within models and Metropolis-Hastings-Green to jump between models.
`Strauss`	`fix.c`	Metropolis sampler for the Strauss process with number of points fixed following Geyer and Møller (1994). Uses reversible scan described in Section 2.2.6.
`Strauss`	`fixgib.c`	Gibbs sampler for the Strauss process with number of points fixed, using rejection from uniform distribution to generate from full conditionals, following Ripley (1979).
`Strauss`	`var.c`	Metropolis-Green sampler for the Strauss with random number of points following Geyer and Møller (1994).
`Strauss`	`regen.c`	Metropolis sampler for the Strauss process with number of points fixed (like `fix.c` but without the reversible scan) using Metropolis-rejected restarts from the binomial process and splitting following Mykland, Tierney and Yu (to appear).
`Normal`	`norm.c`	Gibbs sampler for multivariate normal.
`Ising`	`ising.c`	Gibbs sampler for the Ising model.
	`st.c`	Swendsen-Wang sampler for the Ising model with periodic boundary conditions.

Bibliography

Arcones, M. A. and Yu, B. (1994). Central limit theorems for empirical and U-processes of stationary mixing sequences. *J. Theor. Probab.* **7** 47–71.

Athreya K. B. and Ney, P. (1978). A new approach to the theory of recurrent Markov chains. *Trans. Am. Math. Soc.* **245** 493–501.

Attouch, H. (1984). *Variational Convergence of Functions and Operators.* Boston: Pitman.

Bartlett, M. S. (1946). On the theoretical specification of sampling properties of autocorrelated time series. *J. Roy. Statist. Soc. Suppl.* **8** 27–41.

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., Green, P., Higdon D., and Mengersen, K. (1995). Bayesian computation and stochastic systems (with discussion). *Statistical Science* **10** 3–41.

Besag, J. and Clifford, P. (1989). Generalized Monte Carlo significance tests. *Biometrika* **76** 633–642.

Besag, J. and Green, P. (1993). Spatial statistics and Bayesian computation (with discussion). *J. Roy. Statist. Soc. Ser. B* **55** 25–37,90–95.

Bradley, R. C. (1986). Basic properties of strong mixing conditions. In *Dependence in Probability and Statistics: A Survey of Recent Results (Oberwolfach, 1985),* E. Eberlein and M. S. Taqqu (eds.) 165–192. Boston: Birkhäuser.

Breiman, L. (1968). *Probability.* Reading, Mass.: Addison-Wesley.

Chan, K.-S. (1993). On the central limit theorem for an ergodic Markov chain. *Stochastic Process. Appl.* **47** 113–117.

Chan, K. S. and Geyer, C. J. (1994). Discussion of Tierney (1994). *Ann. Statist.*, **22** 1747–1758.

Chung, K. L. (1967). *Markov Chains with Stationary Transition Probabilities*, 2nd ed. Berlin: Springer-Verlag.

Clifford, P. (1993). Discussion on the meeting on the Gibbs sampler and other Markov chain Monte Carlo methods. *J. Roy. Statist. Soc. Ser. B* **55** 53–54.

Devroye, L. (1986). *Non-Uniform Random Variate Generation*. New York: Springer-Verlag.

Doukhan, P., Massart, P. and Rio E. (1994). The functional central limit theorem for strongly mixing processes. *Ann. Inst. H. Poincaré Probab. Statist.* **30** 63–82.

Fishman, G. S. (1991). Choosing warm-up interval and sample size when generating Monte Carlo data from a Markov chain. Technical Report UNC/OR/TR 91-11, Department of Operations Research, University of North Carolina.

Gaver, D. P. and O'Muircheartaigh, I. G. (1987). Robust empirical Bayes analysis of event rates. *Technometrics* **29** 1–15.

Gelfand, A. E. and Smith A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *J. Am. Statist. Assoc.* **85** 398–409.

Gelman, A., and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science* **7** 457–511.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, 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. (1994). On the convergence of Monte Carlo maximum likelihood calculations. *J. Roy. Statist. Soc. Ser. B* **56** 261–274.

Geyer, C. J. and Møller, J. (1994). Simulation and likelihood inference for spatial point processes. *Scand. J. Statist.* **21** 359–373.

Geyer, C. J. and Thompson, E. A. (1992). Constrained Monte Carlo maximum likelihood for dependent data (with discussion). *J. Roy. Statist. Soc. Ser. B* **54** 657–699.

Geyer, C. J. and Thompson E. A. (1995). Annealing Markov chain Monte Carlo with applications to ancestral inference. *J. Am. Statist. Assoc.* **90** 909–920.

Gilks, W. R., Clayton, D. G., Spiegelhalter, D. J., Best, N. G., McNeil, A. J., Sharples, L. D., and Kirby, A. J. (1993). Modelling complexity: Applications of Gibbs sampling in medicine (with discussion). *J. Roy. Statist. Soc. Ser. B* **55** 39–52,95–97.

Gordin, M. I. and Lifšic, B. A. (1978). The 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: Limit Theorems and Applications.* 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 Linnik, Yu. V. (1971). *Independent and stationary sequences of random variables*. Groningen: Wolters-Noordhoff (English translation).

Jain, N. and Jamison, B. (1967). Contributions to Doeblin's theory of Markov processes. *Z. Wahrscheinlichkeitstheorie und Verw. Geb.* **8** 19–40.

Kelly, F. P. and Ripley, B. D. (1976). A note on Strauss's model for clustering. *Biometrika* **63** 357–360.

Kelton, D. W. and Law, A. M. (1984). An analytical evaluation of alternative strategies in steady-state simulation *Oper. Res.* **32** 169–184.

Kindermann, R. and Snell, J. L. (1980). *Markov Random Fields and Their Applications.* Providence, R. I.: American Mathematical Society.

Kipnis, C. and Varadhan, S. R. S. (1986). Central limit theorem for additive functionals of reversible Markov processes and applications to simple exclusions. *Commun. Math. Phys.* **104** 1–19.

Knuth, Donald E. (1973). *The Art of Computer Programming*, 2nd. ed., Vol. 2, *Semi-Numerical Algorithms*. Reading, Mass.: Addison-Wesley.

Liu, J., Wong, W. H., and Kong, A. (1995). Correlation structure and convergence rate of the Gibbs sampler with various scans. *J. Roy. Statist. Soc. Ser. B* **57** 157–169.

Marinari, E., and Parisi G. (1992). Simulated tempering: A new Monte Carlo scheme. *Europhys. Lett.* **19** 451–458.

Meketon, M. S. and Schmeiser B. W. (1984). Overlapping batch means: something for nothing? In S. Sheppard, U. Pooch, and D. Pegden (eds.) *Proceedings of the 1984 Winter Simulation Conference* 227–230.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *J. Chem. Phys.* **21** 1087–1092.

Meyn, S. P. and Tweedie, R. L. (1993). *Markov Chains and Stochastic Stability*. London: Springer-Verlag.

Nummelin, E. (1987). A splitting technique for Harris recurrent markov chains. *Z. Wahrscheinlichkeitstheorie und Verw. Geb.* **43** 309–318.

Nummelin, E. (1984). *General Irreducible Markov Chains and Non-Negative Operators*. Cambridge University Press.

Ogata, Y. and Tanemura, M. (1981). Estimation of interaction potentials of spatial point patterns through the maximum likelihood procedure. *Ann. Inst. Statist. Math.* **33**, Part B, 315–338.

Ogata, Y. and Tanemura, M. (1984). Likelihood analysis of spatial point patterns. *J. Roy. Statist. Soc. Ser. B* **46** 496–518.

Ogata, Y. and Tanemura, M. (1989). Likelihood estimation of soft-core interaction potentials for Gibbsian point patterns. *Ann. Inst. Statist. Math.* **41** 583–600.

Pedrosa, A. C. and Schmeiser 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* 481–488.

Penttinen, A. (1984). *Modelling Interaction in Spatial Point Patterns: Parameter Estimation by the Maximum Likelihood Method*. Number 7 in Jyväskylä Studies in Computer Science, Economics, and Statistics. University of Jyväskylä.

Potts, R. B. (1952). Some generalized order-disorder transformations. *Proc. Camb. Phil. Soc.* **48** 106–109.

Priestley, M. B. (1981). *Spectral Analysis and Time Series*. London: Academic Press.

Raymond, E. S., compiler (1993). *The New Hacker's Dictionary,* 2nd ed. Cambridge, Mass.: MIT Press. Also available in 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: John Wiley.

Roberts, G. O. and Tweedie, R. L. (1996). Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika* **83** 95–110.

Robertson, T., Wright, F. T. and Dykstra, R. L. (1988). *Order Restricted Statistical Inference*. New York: John Wiley.

Rosenthal, J. (1995). Minorization conditions and convergence rates for Markov chain Monte Carlo. *J. Am. Statist. Assoc.* **90** 558–566. Correction: **90** 1136.

Schervish, M. J. and Carlin, B. P. (1992). On the convergence rate of successive substitution sampling. *J. Comp. Graphical Statist.* **1** 111–127.

Schmeiser, B. (1982). Batch size effects in the analysis of simulation output. *Oper. Res.* **30** 556–568.

Sheehan, N. and Thomas, A. (1993). On the irreducibility of a Markov chain defined on a space of genotype configurations by a sampling scheme. *Biometrics* **49** 163–175.

Smith, A. F. M. and Roberts, G. O. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods (with discussion). *J. Roy. Statist. Soc. Ser. B* **55** 3–23, 88–90.

Strauss, D. J. (1975). A model for clustering. *Biometrika* **62** 467–75.

Swendsen, R. H. and Wang, J. S. (1987). Nonuniversal critical dynamics in Monte Carlo simulations. *Phys. Rev. Lett.* **58** 86–88.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *Ann. Statist.* **22** 1701–1762.

Torrie, G. M., and Valleau, J. P. (1977), Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *J. Computational Phys.* **23** 187–199.

Wang, J. S., and Swendsen, R. H. (1990). Cluster Monte Carlo algorithms. *Physica A* **167** 565–579.