The Annals of Statistics
1994, Vol. 22, No. 4, 1701–1762

MARKOV CHAINS FOR EXPLORING POSTERIOR DISTRIBUTIONS

BY LUKE TIERNEY[1]

University of Minnesota

Several Markov chain methods are available for sampling from a posterior distribution. Two important examples are the Gibbs sampler and the Metropolis algorithm. In addition, several strategies are available for constructing hybrid algorithms. This paper outlines some of the basic methods and strategies and discusses some related theoretical and practical issues. On the theoretical side, results from the theory of general state space Markov chains can be used to obtain convergence rates, laws of large numbers and central limit theorems for estimates obtained from Markov chain methods. These theoretical results can be used to guide the construction of more efficient algorithms. For the practical use of Markov chain methods, standard simulation methodology provides several variance reduction techniques and also gives guidance on the choice of sample size and allocation.

1. Introduction. Suppose we are given a posterior distribution π on a quantity θ with values in a space E. Usually E will be a subset of $\mathbb{R}^k$ and π will have a density with respect to a σ-finite measure μ,

$$\pi(dx) = \pi(x)\mu(dx).$$

For simplicity, π will be used to denote both the distribution and the density. We may be interested in computing a particular numerical characteristic of π, or more generally in developing an understanding of what information π contains about θ.

Several methods for computing characteristics of posterior distributions are now available. These include asymptotic approximations, numerical integration and sampling or Monte Carlo methods. Sampling methods for examining posterior distributions provide ways of generating samples with the property that the empirical distribution of the sample, or an appropriately weighted empirical distribution, approximates the posterior distribution. Using such samples, it is easy to estimate characteristics such as the mean or standard deviation of a function of θ. Marginal distributions can be estimated using smoothing or, in some cases, variance reduction methods. In addition, for equally weighted samples, methods for viewing point clouds, such as rotating plots and Grand Tours, can be used to examine the joint uncertainty about three or more components or features of θ.

A number of different sampling methods are available. In rare cases it is possible to sample directly from the posterior distribution and thus obtain an

Received August 1991; revised August 1992.

[1]Research supported in part by NSF Grant DMS-90-05858.

AMS 1991 *subject classifications.* 60J05, 62-04, 65C05.

Key words and phrases. Monte Carlo, Metropolis–Hastings algorithm, Gibbs sampler, variance reduction.

i.i.d. sample from π. In most problems this is not possible. Either the sample has to be dependent, or the distribution used to generate the sample has to be different from π. A method that uses independent samples from a distribution similar to π is importance sampling. The sample is then weighted to make up for the difference between π and the distribution used to generate the sample. Over the past decade, most work on sampling methods for exploring posterior distributions has centered on importance sampling [Geweke (1989), Stewart (1979), Zellner and Rossi (1984) among others]. More recently, results of Gelfand and Smith (1990) on the Gibbs sampler have rekindled interest in the use of dependent samples generated using Markov chains with equilibrium distribution π. Gelfand and Smith extend the Gibbs sampling algorithm of Geman and Geman (1984), originally developed for Bayesian image reconstruction, to continuous distributions and show how the algorithm can be used in a wide variety of problems. Other methods of generating Markov chains with a specified equilibrium distribution include the Metropolis algorithm [Metropolis, Rosenbluth, Rosenbluth, Teller and Teller (1953)] and some of its variants and extensions.

This paper outlines a number of the basic Markov chain algorithms that are available, and describes several ways in which the algorithms can be combined to form hybrid algorithms. Results from the theory of general state space Markov chains [Nummelin (1984) and Revuz (1975)] are used to derive properties of these algorithms and to suggest improvements to some algorithms. Finally, some issues that arise in the implementation of Markov chain methods are discussed.

2. Some Markov chains.

2.1. *Notation and definitions.* This subsection gives some informal definitions of concepts that are needed in the remainder of the section. More careful definitions of some of these concepts are given in Section 3.

A time-homogeneous Markov chain with invariant distribution π is a sequence of E-valued random variables $\{X_n;\ n \geq 0\}$ such that the transition kernel P defined by

$$P(X_n, A) = P\{X_{n+1} \in A \mid X_0, \ldots, X_n\}$$

satisfies

$$\pi(A) = \int \pi(dx)P(x, A)$$

for all measurable sets A. The distribution of X_0 is the initial distribution of the chain. The conditional distribution of X_n given X_0 is written as

$$P\{X_n \in A \mid X_0\} = P^n(X_0, A),$$

where P^n denotes the nth iterate of the kernel P.

The invariant distribution π is an equilibrium distribution for the chain if for π-almost all x,

$$\lim_{n \to \infty} P^n(x, A) = \pi(A)$$

for all measurable sets A. A Markov chain with invariant distribution π is irreducible if, for any initial state, it has positive probability of entering any set to which π assigns positive probability. A chain is periodic if there are portions of the state space it can only visit at certain regularly spaced times; otherwise, the chain is aperiodic. If a chain has a proper invariant distribution π and it is irreducible and aperiodic, then π is the unique invariant distribution and is also the equilibrium distribution of the chain (see Theorem 1 in Section 3).

Many approaches are available for constructing Markov chains with a specified invariant distribution. Several strategies are outlined in the following subsections.

2.2. *Conditioning and the Gibbs sampler.* One approach to constructing a Markov chain with invariant distribution π is to use conditioning. Suppose X has distribution π, f is a function defined on E and $Y = f(X)$. If

$$Q(y, A) = P\{X \in A \mid Y = y\},$$

then $P(x,A) = Q(f(x),A)$ is a transition kernel with invariant distribution π. Since $P(x, \cdot)$ puts all its mass on the set $f^{-1}(\{f(x)\})$, P will typically not be irreducible. But choosing several functions $f_1, \ldots, f_m$, constructing the corresponding conditioning kernels $P_1, \ldots, P_m$ and using these kernels in order produces another kernel $P = P_1 P_2 \cdots P_m$ with invariant distribution π that may be irreducible. This is the strategy used in the Gibbs sampler of Gelfand and Smith (1990). If E is a subset of a product space and $x \in E$ can be written as $x = (x_1, \ldots, x_m)$, then the Gibbs sampler uses the functions

$$f_i(x) = f_i(x_1, \ldots, x_m) = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_m)$$

for $i = 1, \ldots, m$.

The kernel obtained by conditioning on Y produces X_{n+1} by sampling from the conditional distribution $X \mid Y = f(X_n)$. As a result, for any y the conditional distribution $X \mid Y = y$ is an invariant distribution for this kernel. Any kernel that is invariant with respect to these conditional distributions for all y will have invariant distribution π. In particular, a kernel that applies one of the methods described below to the conditional distributions $X \mid Y = y$ produces a kernel with invariant distribution π. Again, a single kernel of this type is usually not irreducible, but several can be used in series to produce an irreducible kernel.

The parameterization used to construct a Gibbs sampler can have a significant effect on its performance. A discussion of these parameterization issues is given, for example, by Hills and Smith (1992). The performance of Gibbs samplers can also sometimes be improved by introducing auxiliary variables [Besag and Green (1993), Section 5]. A particularly successful example of this approach is the algorithm of Swendsen and Wang (1987) for sampling from the multicolor generalization of the Ising model.

2.3. *The Metropolis algorithm.* The Metropolis algorithm was originally introduced by Metropolis, Rosenbluth, Rosenbluth, Teller and Teller (1953)

for computing properties of substances composed of interacting individual molecules. This algorithm has been used extensively in statistical physics [Hammersley and Handscomb (1964), Section 9.3]. A variation on this algorithm was proposed by Barker (1965). A generalization was introduced by Hastings (1970) and studied further by Peskun (1973). Another generalization is known as the hit-and-run algorithm [Smith (1984) and Schmeiser and Chen (1991)]. The Metropolis algorithm forms the basis of the simulated annealing algorithm [Kirkpatrick, Gelatt and Vecchi (1983)]. In the discrete Bayesian image reconstruction problems considered by Geman and Geman (1984), the Gibbs sampler can be viewed as a special case of Barker's version of the Metropolis algorithm.

2.3.1. *The general algorithm.* To define Hastings' version of the algorithm, suppose that π has a density with respect to μ and let Q be a Markov transition kernel of the form

$$Q(x, dy) = q(x,y)\mu(dy).$$

To avoid some trivial special cases, let $E^+ = \{x\colon \pi(x) > 0\}$ and assume that $Q(x, E^+) = 1$ for $x \notin E^+$. Also assume that π is not concentrated on a single point. Then define

$$\alpha(x,y) = \begin{cases} \min\left\{\dfrac{\pi(y)q(y,x)}{\pi(x)q(x,y)}, 1\right\}, & \text{if } \pi(x)q(x,y) > 0, \\ 1, & \text{if } \pi(x)q(x,y) = 0. \end{cases}$$

If the chain is currently at a point $X_n = x$, then it generates a candidate value Y for the next location X_{n+1} from the distribution $Q(x, \cdot)$. With probability $\alpha(x,y)$ this candidate is accepted and the chain moves to $X_{n+1} = y$. Otherwise, the step is rejected and the chain remains at $X_{n+1} = x$.

This algorithm only depends on π through ratios of the form $\pi(y)/\pi(x)$; thus π only needs to be known up to a normalizing constant.

If $\pi(x)q(x,y) > 0$, then $\alpha(x,y) = 0$ if $\pi(y) = 0$. Thus the chain almost surely does not leave the set E^+ once it is entered. The restrictions imposed on Q ensure that E^+ will be entered after at most one step. In practice, the initial state will always be chosen to be in E^+. The formal specification of P is extended to the complement of E^+ for mathematical completeness.

If we define the off-diagonal density of a Metropolis kernel as

$$p(x,y) = \begin{cases} q(x,y)\alpha(x,y), & \text{if } x \neq y, \\ 0, & \text{if } x = y, \end{cases}$$

and set

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

then the Metropolis kernel P can be written as

$$P(x, dy) = p(x,y)\mu(dy) + r(x)\delta_x(dy),$$

where δ_x denotes point mass at x. The value $r(x)$ is the probability that the algorithm remains at x. Since p satisfies the reversibility condition

$$\pi(x)p(x,y) = \pi(y)p(y,x),$$

it follows that π is an invariant distribution for P: for any measurable set A,

$$\begin{aligned}
\int P(x,A)\pi(dx) &= \int \left[\int_A p(x,y)\mu(dy)\right]\pi(x)\mu(dx) + \int r(x)\delta_x(A)\pi(x)\mu(dx) \\
&= \int_A \left[\int p(x,y)\pi(x)\mu(dx)\right]\mu(dy) + \int_A r(x)\pi(x)\mu(dx) \\
&= \int_A \left[\int p(y,x)\pi(y)\mu(dx)\right]\mu(dy) + \int_A r(x)\pi(x)\mu(dx) \\
&= \int_A (1 - r(y))\pi(y)\mu(dy) + \int_A r(x)\pi(x)\mu(dx) \\
&= \int_A \pi(y)\mu(dy).
\end{aligned}$$

For P to be irreducible, it is necessary that Q be irreducible, but this is not sufficient: irreducibility of P depends on both Q and π. If P is irreducible and $\pi(\{x\colon r(x) > 0\}) > 0$, then, by the results of Nummelin (1984), Section 2.4, the Metropolis kernel is aperiodic.

The original Metropolis algorithm assumes $q(x,y) = q(y,x)$. In this case the acceptance probability simplifies to

$$\alpha(x,y) = \min\left\{\frac{\pi(y)}{\pi(x)}, 1\right\}.$$

Barker (1965) and Hastings (1970) consider several alternative forms of the acceptance probability, but results of Peskun (1973) show that the form given here is optimal among a fairly wide range of choices. The main reason is that this form rejects candidate steps less often than other forms. Hastings suggests monitoring rejections. Other forms of the acceptance probability may be useful in kernels used for conditioning as discussed in Section 4.3.2 below.

The Metropolis algorithm is very general, allowing a variety of useful special cases based on different choices of the transition kernel Q. The following subsections give a few examples of Metropolis kernels that are useful for examining posterior distributions.

2.3.2. *Random walk chains.* Suppose $E = \mathbb{R}^k$, μ is Lebesgue measure and f is a density on E. If Y is generated by drawing Z independently from f and setting $Y = x + Z$, then $q(x,y) = f(y-x)$. Thus the kernel Q driving the Metropolis chain is a random walk. Natural choices for the increment distribution include a uniform distribution on a disk, a normal distribution or perhaps a multivariate

t distribution. Split-t distributions [Geweke (1989)] may also be useful. The scale matrix for f can be taken as a constant c times the inverse information at the posterior mode. Good choices for the step size constant c are still an open problem, but $c = 1$ and $c = \frac{1}{2}$ seem to work reasonably well in a number of examples.

If the increment density f is symmetric about the origin, then q is symmetric and the simpler form of the acceptance probability can be used. The algorithm given by Müller (1991) corresponds to a random walk Metropolis chain with an increment density f that is symmetric about the origin. The hit-and-run algorithm [Smith (1984) and Schmeiser and Chen (1991)] with a uniform direction choice and step size chosen by the Metropolis method corresponds to a random walk chain with a spherically symmetric increment density.

If the density f is strictly positive on all of $\mathbb{R}^k$, then the random walk Metropolis kernel is irreducible and aperiodic. If f is not strictly positive everywhere, but is strictly positive on a neighborhood of the origin, then a sufficient condition for the Metropolis kernel to be irreducible and aperiodic is that E^+ be open and connected.

2.3.3. *Independence chains.* Candidate steps Y can also be chosen from a fixed density f. This option is discussed in Hastings (1970), Section 2.5. In this case, $q(x,y) = f(y)$ and the acceptance probability $\alpha(x,y)$ can be written as

$$\alpha(x,y) = \min\left\{\frac{w(y)}{w(x)}, 1\right\},$$

where $w(x) = \pi(x)/f(x)$. This function w is the importance weight function that would be used in importance sampling if observations were generated from the density f.

The independence Metropolis chain is closely related to the corresponding importance sampling process. Candidate steps with low weights are rarely accepted. On the other hand, candidates with high weights are usually accepted, and the process will usually remain at these points for several steps, thus using repetition to build up weight on these points within the sample. If some points have very high weight values, then the process may "get stuck" at these points for a long time. Thus, as for importance sampling, it is useful to choose f to produce a weight function that is bounded, and as close to constant as possible. If the weight function is constant, then the Metropolis algorithm will never reject candidate steps, and the chain produces an i.i.d. sample from π. Again, the best choice of f for an independence Metropolis kernel is still open to question, but, because of the close relation to importance sampling, it is reasonable to conjecture that multivariate t distributions with low degrees of freedom, split-t distributions and other distributions that have been found to be useful as importance sampling densities will be good choices in this context as well.

An independence Metropolis kernel is irreducible and aperiodic if and only if f is positive μ-almost everywhere on E^+.

For some problems, there may be a natural density f to use in an independence chain but it may be difficult or impossible to sample directly from f. Geyer

(1991) suggests constructing an auxiliary chain Y_n with equilibrium distribution f and switching the states of the X_n and Y_n chains with the Metropolis acceptance probability

$$\min\left\{\frac{\pi(y)f(x)}{\pi(x)f(y)}, 1\right\}.$$

The resulting bivariate chain (X_n, Y_n) has an invariant distribution with density proportional to $\pi(x)f(y)$.

2.3.4. *Rejection sampling chains.* An interesting special case of the independence Metropolis kernel occurs if sampling from f uses rejection sampling. Rejection sampling [see, e.g., Ripley (1987), Section 3.2.] is the basis for several algorithms for generating variates from standard univariate distributions. This method has also been suggested by Zeger and Karim (1991) and by Carlin and Gelfand (1991) for use with the Gibbs sampler in cases where it is not possible to sample directly from certain conditional distributions.

In attempting to use rejection sampling to sample directly from π, we use a density h and a constant c such that, hopefully, $\pi(x) \le c\, h(x)$ for all x. Pairs (Z, U) are generated by generating Z from h and U uniformly on the interval $[0, c\, h(Z)]$ until a pair satisfying $U < \pi(Z)$ is obtained. The final Z then has density

$$f(x) \propto \pi(x) \wedge c\, h(x).$$

If we do indeed have $\pi(x) \le c\, h(x)$, then f is proportional to π and we obtain an i.i.d. sample from π. But it is very difficult to ensure that c is large enough for $c\, h$ to dominate π without choosing c excessively large, leading to an inefficient algorithm with many rejections. And even then without extensive analysis of the tails of h and π we cannot be certain that $c\, h$ does dominate π.

Fortunately, using this rejection scheme to drive an independence Metropolis chain provides a simple remedy. If we define $C = \{x\colon \pi(x) \le c\, h(x)\}$, then the Metropolis acceptance probability is

$$\alpha(x,y) = \begin{cases} 1, & \text{for } x \in C, \\ \dfrac{c\, h(x)}{\pi(x)}, & \text{for } x \notin C,\ y \in C, \\ \min\left\{\dfrac{\pi(y)h(x)}{\pi(x)h(y)}, 1\right\}, & \text{for } x \notin C,\ y \notin C. \end{cases}$$

Thus the algorithm will occasionally reject candidate steps when the chain is at a point $x \notin C$. This repeats the point x within the sample path and thus compensates for the deficiency in the envelope at x. Dependence is introduced to make up for this deficiency and ensure that π is the invariant distribution of the Markov chain. If the envelope is in fact adequate, then no candidates are rejected, and the chain produces an i.i.d. sequence from π.

The visits to the set C form a renewal process that can be used for a regenerative analysis of a rejection sampling chain.

2.3.5. *Autoregressive chains.* If $E = \mathbb{R}^k$, μ is Lebesgue measure and f is a density on E with respect to μ, then an intermediate strategy between a random walk and an independence kernel is obtained by generating Z from f and defining candidate steps by $Y = a + b(x - a) + Z$. Here $a \in \mathbb{R}^k$ is a fixed vector and b is either a real constant or a fixed $k \times k$ matrix. Thus $q(x,y) = f(y - a - b(x - a))$. For $b = 1$ this reduces to the random walk kernel. For $a = 0$ and $b = 0$ it produces an independence kernel. If $0 < b < 1$, then this strategy shrinks the current state toward the point a before adding the increment Z.

In this autoregressive strategy, it is also possible to take $b < 0$. If $b = -1$, then the current state is reflected about the point a before adding the increment Z. This is used by Hastings (1970), Example 2, to induce negative correlations between successive states in the chain. This will often reduce the variance of estimates of expectations of linear functions. It can be viewed as a method of producing antithetic variates [Ripley (1987), Section 5.3].

The reflection strategy can be taken to a limit by using a degenerate increment density concentrated at the origin. This corresponds to generating a candidate step by just reflecting the current state x about the point a to produce $y = 2a - x$. This candidate is then accepted with probability

$$\alpha(x,y) = \min\left\{\frac{\pi(2a - x)}{\pi(x)}, 1\right\}.$$

Since the candidate generation density is not absolutely continuous with respect to Lebesgue measure, the dominating measure for π, this limiting kernel does not satisfy the definition of a Metropolis kernel given above. Nevertheless, it does have invariant distribution π. It is obviously periodic and not irreducible and therefore cannot be used alone. But it can be useful in mixtures or cycles with other kernels as discussed below.

Strategies that use reflection about a point a are most effective when the density π is approximately symmetric about a. Other symmetries, such as axial symmetries, can be exploited along similar lines.

2.3.6. *Grid-based chains.* Ritter and Tanner (1991) propose a method called the griddy Gibbs sampler for Gibbs sampling in problems where conditionals cannot be sampled directly. This method is also described in Tanner (1991). The idea is to evaluate the conditional density on a grid and use an approximate cumulative distribution function based on these grid values to generate variables with approximately the right conditional distribution.

Using this algorithm in its pure form may require quite a fine grid and thus a very large number of posterior density evaluations to control the error in the approximation. Fortunately, as in the case of rejection sampling, it is possible to embed this algorithm in a Metropolis chain to ensure that the equilibrium distribution is exactly π even for a coarse grid. To simplify the discussion, suppose that $E = \mathbb{R}$, μ is Lebesgue measure and f is a density on E with respect to μ. In higher-dimensional problems, the one-dimensional algorithm can be applied to each coordinate in turn, as in the Gibbs sampler, or along randomly chosen directions, as in the hit-and-run algorithm. Let $x_1, \ldots, x_m$ be a fixed set of points.

To generate a candidate step Y, first select a point x_k from $x_1, \ldots, x_m$ according to a distribution that is proportional to the density values $\pi(x_1), \ldots, \pi(x_m)$ at these points. Then generate an increment Z from f and add this increment to x_k to obtain $Y = x_k + Z$. Thus the density of the candidate generation kernel is

$$q(x,y) = \frac{\sum \pi(x_i) f(y - x_i)}{\sum \pi(x_i)}$$

and the probability of accepting a candidate is

$$\alpha(x,y) = \min\left\{ \frac{\pi(y) \sum \pi(x_i) f(x - x_i)}{\pi(x) \sum \pi(x_i) f(y - x_i)}, 1 \right\}.$$

If the tails of the density f are thick enough, then the acceptance probability will be bounded away from 0.

For equally spaced points the density $q(x,y)$ is a kernel-smoothed approximation to $\pi(x)$. If this approximation is good, then the acceptance probability will be close to 1, and the chain will produce approximately i.i.d. variables from π. If the approximation is not particularly good, the Metropolis chain will reject frequently to compensate for deficiencies in the approximation, but the equilibrium distribution will still be exactly equal to π.

Several variations on this approach are possible. The points can be equally spaced or unequally spaced, and different probabilities can be used. For example, the points might be obtained as a sample from a density that approximates π and the probabilities taken to be proportional to the corresponding importance weights. This approach could be applied directly in higher dimensions. It may also be useful in some cases to use different densities f at each point.

Another variation is to use a sliding lattice centered at the current location x of the chain. Thus the grid points would be of the form $x \pm ih$ for $0 \le i \le m$ for some fixed increment h. This changes the candidate generation density to

$$q(x,y) = \frac{\sum \pi(x \pm ih) f(y - x \mp ih)}{\sum \pi(x \pm ih)}$$

and the acceptance probability to

$$\alpha(x,y) = \min\left\{ \frac{\pi(y) \sum \pi(y \pm ih) f(x - y \mp ih) \sum \pi(x \pm ih)}{\pi(x) \sum \pi(x \pm ih) f(y - x \mp ih) \sum \pi(y \pm ih)}, 1 \right\}.$$

This approach is particularly useful if there is too much uncertainty about the range of the distribution to permit using a fixed grid.

For a sliding lattice grid it is possible to consider the limiting case where the increment distribution is point mass at the origin. In this case candidates are chosen from the lattice with probabilities proportional to $\pi(x \pm ih)$ and are accepted with probability

$$\alpha(x,y) = \min\left\{ \frac{\sum \pi(x \pm ih)}{\sum \pi(y \pm ih)}, 1 \right\}.$$

If the support of π is bounded and the lattice is wide enough to cover the support from any starting point in the support, then $\alpha(x,y) = 1$ for all x and y in the support. As in the reflection scheme discussed above, this limiting case has invariant distribution π but is not irreducible. Again it can be used in combination with other kernels. It also has the advantage that the conditional expectation $E[X_{n+1} \mid X_n]$ is very easy to compute; this is useful for variance reduction.

2.4. *Combining strategies.* The methods outlined above can be used in a pure form, or they can be combined into hybrid strategies. One way to form a hybrid strategy is to use conditioning, as described above, and then apply a more basic strategy to the conditional distributions. Two other basic forms of hybrid strategies are mixtures and cycles. Suppose $P_1, \ldots, P_m$ are Markov kernels with invariant distribution π. In a mixture, positive probabilities $\alpha_1, \ldots, \alpha_m$ are specified, and at each step one of the kernels is selected according to these probabilities. In a cycle, each kernel is used in turn, and when the last one is used the cycle is restarted.

Mixtures and cycles can be used in several ways. For example, a Gibbs sampler can be combined with occasional steps from an independence chain in a mixture or a cycle to "restart" the Gibbs sampler and thus reduce correlations while preserving the equilibrium distribution. As another example, suppose θ can be split into two components (θ_1, θ_2), and direct sampling from $\theta_1 \mid \theta_2$ is possible but direct sampling from $\theta_2 \mid \theta_1$ is not possible. Such a situation is considered by Zeger and Karim (1991). Then "Gibbs steps" for $\theta_1 \mid \theta_2$ can be combined with Metropolis steps for $\theta_2 \mid \theta_1$ in a mixture or a cycle.

If one of the kernels in a mixture is irreducible and aperiodic, then it is easy to see that the mixture kernel is irreducible and aperiodic as well. If one of the kernels is irreducible and aperiodic, then the cycle kernel is often also irreducible and aperiodic, but there are simple counterexamples that show that no general statement to this effect can be made. In general, irreducibility and aperiodicity of a cycle need to be verified for each case unless one of the kernels has some special structure (see Proposition 4).

Mixtures can also be constructed by selecting according to a fixed distribution from a family of transition functions indexed by a general parameter. Hit-and-run algorithms with a uniform direction distribution [Smith (1984) and Schmeiser and Chen (1991)] can thus also be viewed as mixtures.

3. Some theoretical results. This section presents some results from the theory of general state space Markov chains in a form that is useful for establishing properties of Markov chains constructed to have a specified invariant distribution. The development in this section is based primarily on Nummelin (1984).

To start off, we need some more notation and definitions. Let $\mathcal{E}$ be a countably generated σ-algebra on E. A (*Markov*) *transition kernel* on $(E, \mathcal{E})$ is a map $P\colon E \times \mathcal{E} \to [0,1]$ such that:

(i) for any fixed $A \in \mathcal{E}$, the function $P(\cdot, A)$ is measurable;
(ii) for any fixed $x \in E$, the function $P(x, \cdot)$ is a probability measure on $(E, \mathcal{E})$.

Probabilities for a Markov chain with transition kernel P started at x are denoted by P_x.

If ν is a probability, P is a transition function on $(E, \mathcal{E})$ and h is a real-valued, $\mathcal{E}$-measurable function, then νP, Ph and νh are defined by

$$(\nu P)(A) = \int P(x,A)\nu(dx),$$

$$(Ph)(x) = \int h(y)P(x,dy)$$

and

$$\nu h = \int h(y)\nu(dy)$$

for all $x \in E$ and $A \in \mathcal{E}$. A nonnegative real-valued function h is called *harmonic* for P if $h = Ph$.

Following Nummelin, the total variation norm of a bounded signed measure λ on $(E, \mathcal{E})$ is defined as

$$\|\lambda\| = \sup_{A \in \mathcal{E}} \lambda(A) - \inf_{A \in \mathcal{E}} \lambda(A).$$

The total variation distance between two such measures λ_1 and λ_2 is $\|\lambda_1 - \lambda_2\|$. Other definitions of the total variation distance may differ by a factor of 2.

The notation $\{A_n \text{ i.o.}\}$ means that the sequence A_n occurs infinitely often, that is, $\sum 1_{A_n} = \infty$.

For general state spaces, irreducibility is defined with respect to a σ-finite measure φ. A transition kernel P on $(E, \mathcal{E})$ is φ-irreducible if $\varphi(E) > 0$ and for each $x \in E$ and each $A \in \mathcal{E}$ with $\varphi(A) > 0$ there exists an integer $n = n(x,A) \geq 1$ such that $P^n(x,A) > 0$. For our purposes, it is natural to take $\varphi = \pi$. The usual notion of irreducibility for a discrete state space corresponds to irreducibility with respect to counting measure.

A π-irreducible transition kernel P is periodic if there exists an integer $d \geq 2$ and a sequence $\{E_0, E_1, \ldots, E_{d-1}\}$ of d nonempty disjoint sets in $\mathcal{E}$ such that, for all $i = 0, \ldots, d-1$ and all $x \in E_i$,

$$P(x, E_j) = 1 \quad \text{for } j = i + 1 \pmod{d}.$$

Otherwise, the kernel is aperiodic.

3.1. *Convergence of distributions.* A crucial concept in the theory of discrete state space Markov chains is *recurrence*. This concept is also crucial in the convergence theory of general state space chains. A general definition of recurrence is given in Nummelin (1984), Definition 3.5. A definition that is sufficient for the present context is as follows. A π-irreducible chain X_n with invariant distribution π is recurrent if, for each B with $\pi(B) > 0$,

$$P_x\{X_n \in B \text{ i.o.}\} > 0 \quad \text{for all } x,$$

$$P_x\{X_n \in B \text{ i.o.}\} = 1 \quad \text{for } \pi\text{-almost all } x.$$

The chain is *Harris recurrent* if $P_x\{X_n \in B \text{ i.o.}\} = 1$ for all x.

Suppose a chain P is π-irreducible and that π is an invariant distribution for the chain. To obtain a contradiction, suppose in addition that the chain is not recurrent. Then Theorem 3.6(i) and Proposition 3.9(iv) of Nummelin (1984) show that there exist sets B_i such that $E = \cup B_i$ and the B_i are transient, that is, $P^n(x, B_i) \to 0$ for all x. Since at least one of these sets must have positive π-probability, this leads to a contradiction. Thus if P is π-irreducible and has π as an invariant distribution, then P must be recurrent. By Corollary 5.2 of Nummelin (1984), a π-irreducible recurrent chain has an invariant measure that is unique up to a multiplicative constant. The chain is called *positive recurrent* if the total mass of this measure is finite; otherwise it is *null recurrent*. Thus if P is π-irreducible and π is an invariant probability distribution for P, then P must be positive recurrent, and π is its unique invariant distribution. If P is also aperiodic, then Theorem 3.7(i) and Proposition 6.3 of Nummelin (1984) show that the transition probabilities converge to π. This is summarized in the following theorem.

THEOREM 1. *Suppose P is π-irreducible and $\pi P = \pi$. Then P is positive recurrent and π is the unique invariant distribution of P. If P is also aperiodic, then, for π-almost all x,*

$$\|P^n(x, \cdot) - \pi\| \to 0,$$

with $\|\cdot\|$ denoting the total variation distance. If P is Harris recurrent, then the convergence occurs for all x.

Athreya, Doss and Sethuraman (1992) present a proof of this result from first principles.

While this result does not provide any information on the rate of convergence, its assumptions are quite minimal. In fact, the assumptions are essentially necessary and sufficient: if

$$\|P^n(x, \cdot) - \pi\| \to 0$$

for all x, then, by Proposition 6.3 of Nummelin (1984), the chain is π-irreducible, aperiodic, positive Harris recurrent and has invariant distribution π.

The exceptional null set for non-Harris recurrent chains is a nuisance. Fortunately, in our applications it is usually not hard to verify that a chain is Harris recurrent. The basic result, which follows from Theorem 3.6(i) and Theorem 3.8 of Nummelin (1984), is the following theorem.

THEOREM 2. *If P is recurrent, then it is Harris recurrent if and only if every bounded harmonic function is a constant.*

A condition that is satisfied by most irreducible Gibbs samplers is given by the following corollary.

COROLLARY 1. *Suppose P is π-irreducible and $\pi P = \pi$. If $P(x, \cdot)$ is absolutely continuous with respect to π for all x, then P is Harris recurrent.*

PROOF. Let h be a bounded harmonic function for P. The assumptions imply that P is recurrent; hence by Proposition 3.13 of Nummelin (1984) $h = \pi h$ π-almost everywhere. Absolute continuity then implies that $(Ph)(x) = \pi h$ for all x. Thus $h \equiv \pi h$, that is, h is a constant. □

For general Metropolis kernels as defined above, no additional conditions are needed.

COROLLARY 2. *Suppose P is a π-irreducible Metropolis kernel. Then P is Harris recurrent.*

PROOF. Let h be a bounded harmonic function for P. Since the assumptions imply that P is recurrent, $h = \pi h$ π-almost everywhere as in the preceding proof. Suppose $x \in E^+$, let $p(x,y)$ denote the off-diagonal density of the Metropolis kernel and let $r(x)$ be the probability that a chain starting at x remains at x. Now $\int_A \pi(y)\mu(dy) = 0$ implies $\int_A p(x,y)\mu(dy) = 0$ for $x \in E^+$. Therefore

$$\int p(x,y)h(y)\mu(dy) = \big(1 - r(x)\big)\pi h$$

and thus

$$\int P(x,dy)h(y) = \big(1 - r(x)\big)\pi h + r(x)h(x) = h(x).$$

This in turn implies

$$\big(1 - r(x)\big)\big(h(x) - \pi h\big) = 0$$

for every $x \in E^+$. Since π is not concentrated on a single point, π-irreducibility implies that $r(x) < 1$ for all x. Thus $h(x) = \pi h$ for all $x \in E^+$. Finally, if $x \notin E^+$, then, by the restrictions imposed on Q and the definition of $\alpha(x,y)$, we have $Q(x,E^+) = 1$ and $\alpha(x,y) = 1$, and therefore $h(x) = (Ph)(x) = \pi h$. Thus $h \equiv \pi h$, that is, h is a constant. □

A similar development to the one presented in this section is given in a paper by Chan (1993), which came to my attention while preparing the revision of this paper.

3.2. *Rates of convergence.* A Markov chain is called *ergodic* if it is positive Harris recurrent and aperiodic. Several stronger forms of ergodicity that provide information on the rate of convergence in Theorem 1 are available.

A rather weak form of ergodicity stronger than simple ergodicity is called *ergodicity of degree* 2. If S_B denotes the hitting time for the set B, then an ergodic chain with invariant distribution π is ergodic of degree 2 if

$$\int_B \pi(dx)E_x\big[S_B^2\big] < \infty$$

for all $B \in \mathcal{E}$ with $\pi(B) > 0$ [Nummelin (1984), Definition 5.5]. For such a chain

$$n\|P^n(x, \cdot) - \pi\| \to 0$$

for π-almost all x [Nummelin (1984), Corollary 6.9]. Ergodicity of degree 2 is typically very difficult to verify in practice.

Two stronger forms of ergodicity are called *geometric* and *uniform ergodicity*. An ergodic Markov chain with invariant distribution π is geometrically ergodic if there exists a nonnegative extended real-valued function M with $\pi|M| < \infty$ and a positive constant $r < 1$ such that

$$\|P^n(x, \cdot) - \pi\| \leq M(x)r^n$$

for all x. The chain is uniformly ergodic if there is a positive constant M and a positive constant $r < 1$ such that

$$\sup_{x \in E} \|P^n(x, \cdot) - \pi\| \leq Mr^n.$$

Uniform ergodicity implies geometric ergodicity, and geometric ergodicity implies ergodicity of degree 2.

To give more easily verified sufficient conditions for geometric and uniform ergodicity, we need the notions of a *minorization condition* and a *small set*. A π-irreducible kernel P satisfies a minorization condition $M(m, \beta, C, \nu)$ for an integer $m \geq 1$, a constant $\beta > 0$, a set $C \in \mathcal{E}$ and a probability measure ν on $\mathcal{E}$ if $\pi(C) > 0$ and

$$\beta\nu(\cdot) \leq P^m(x, \cdot) \quad \text{for all } x \in C.$$

A set C is a small set for P if P satisfies a minorization condition $M(m, \beta, C, \nu)$ for some m, β and ν. In the general theory, small sets play similar roles to individual states in discrete chain theory.

The following *drift condition* is sufficient to ensure geometric ergodicity and can sometimes also be used to verify uniform ergodicity.

PROPOSITION 1. *Suppose X_n is ergodic and there exist a nonnegative real-valued $\mathcal{E}$-measurable function g, a small set C, a constant $r > 1$ and an integer $m \geq 1$ such that*

$$\sup_{x \in C^c} E\left[rg(X_{n+m}) - g(X_n) \mid X_n = x\right] = \sup_{C^c}\left(rP^m g - g\right) < 0$$

and

$$\sup_{x \in C} E\left[g(X_{n+m}); X_{n+m} \in C^c \mid X_n = x\right] = \sup_{C} P^m(1_{C^c} g) < \infty.$$

Then X_n is geometrically ergodic. If g is bounded, then X_n is uniformly ergodic.

PROOF. Since $\|P^n(x, \cdot) - \pi\|$ is nonincreasing in n, an aperiodic chain X_n is geometrically ergodic if $Y_n = X_{nm}$ is geometrically ergodic for any integer $m \geq 1$.

So it is sufficient to consider the case $m = 1$. For this case, Proposition 5.21 and Theorem 6.14(iii) of Nummelin (1984) show that the hypotheses of the proposition imply geometric ergodicity. Finally, the proof of Theorem 1 in Chan (1989) [see also Theorem 3.1 and the following remarks in Chan (1993)] shows that under these hypotheses there exist constants a and b and a positive constant $\rho < 1$ such that

$$\|P^n(x, \cdot) - \pi\| \leq \big(a + bg(x)\big)\rho^n.$$

So if g is bounded, then X_n is uniformly ergodic. □

This condition can sometimes be used to verify geometric or uniform ergodicity of Gibbs samplers by taking C to be bounded with $\pi(x)$ positive on C and taking g to depend only on one of the coordinates. Chan (1993) gives corollaries to this result with sufficient conditions that may be easier to verify, and shows how to use these results to verify geometric ergodicity of Gibbs samplers in several examples.

An alternative condition for geometric ergodicity of a Gibbs kernel is given by Schervish and Carlin (1992). Remarks following the proof of their Corollary 1 show that their square integrability condition implies geometric ergodicity as it is defined here.

For uniform ergodicity, a simple necessary and sufficient condition is available [Nummelin (1984), Theorem 6.15].

PROPOSITION 2. *A transition kernel P is uniformly ergodic if and only if the state space E is small. Furthermore, if P satisfies a minorization condition $M(m, \beta, E, \nu)$, then the convergence rate r satisfies $r^m \leq (1 - \beta)$.*

PROOF. This follows from Theorem 6.15 of Nummelin (1984), together with the observation that a minorization condition $M(m, \beta, E, \nu)$ for P implies that P is ν-irreducible, that the conditions $M(k, \beta, E, \nu)$ hold for all $k \geq m$ and therefore that P is aperiodic. The rate result follows from the remarks after Theorem 6.15 of Nummelin and the representation of the invariant distribution given in Corollary 5.2 of Nummelin. □

Roberts and Polson (1990) give conditions for uniform ergodicity based on continuity and compactness. In addition, they give a version of the preceding theorem based on a minorization condition with $m = 1$.

For a Metropolis kernel, the following corollary gives a sufficient condition for uniform ergodicity that can often be ensured by truncating π to a suitable compact set.

COROLLARY 3. *A Metropolis kernel with $\mu(E^+) < \infty$ and q and π bounded and bounded away from 0 on E^+ satisfies a minorization condition $M(1, \beta, E, \nu)$ with ν proportional to the restriction of μ to E^+, and is therefore uniformly ergodic.*

Weaker conditions are possible.

For an independence Metropolis kernel, a sufficient condition for uniform ergodicity can be given in terms of the weight function.

COROLLARY 4. *An independence Metropolis kernel with density f and bounded weight function $w = \pi/f$ satisfies a minorization condition $M(1, \beta, E, \pi)$ with $\beta = (\sup w)^{-1}$, and is thus uniformly ergodic. The convergence rate r satisfies $r \leq (1 - \beta) = (1 - (\sup w)^{-1})$.*

Under certain conditions we can infer rates of convergence of mixtures or cycles of kernels from their components. For mixtures, if one kernel is uniformly ergodic then the mixture is as follows.

PROPOSITION 3. *Suppose P_1 and P_2 have invariant distribution π and P_1 is uniformly ergodic. Then for $0 < \alpha < 1$ the kernel $\alpha P_1 + (1 - \alpha)P_2$ is uniformly ergodic.*

PROOF. Since P_1 is uniformly ergodic, it satisfies a minorization condition $M(m, \beta, E, \nu)$ for some m, β and ν. Thus

$$\big(\alpha P_1 + (1 - \alpha)P_2\big)^m(x, \cdot) \geq \alpha^m P_1^m(x, \cdot) \geq \alpha^m \beta \nu(\cdot)$$

for all x. So $\alpha P_1 + (1 - \alpha)P_2$ satisfies a minorization condition $M(m, \alpha^m, \beta, E, \nu)$ and is uniformly ergodic. □

For cyclic combinations, a stronger hypothesis appears to be needed.

PROPOSITION 4. *Suppose P_1 and P_2 have invariant distribution π and assume that P_1 satisfies the minorization condition $M(1, \beta, E, \nu)$ for some β and ν. Then P_1P_2 and P_2P_1 are uniformly ergodic.*

PROOF. Since $P_1(x, \cdot) \geq \beta\nu(\cdot)$ for all x,

$$(P_1P_2)(x, \cdot) \geq \beta(\nu P_2)(\cdot).$$

So P_1P_2 satisfies the minorization condition $M(1, \beta, E, \nu P_2)$ and is therefore uniformly ergodic. Similarly,

$$(P_2P_1)(x, \cdot) \geq \beta\nu(\cdot)$$

and P_2P_1 satisfies the minorization condition $M(1, \beta, E, \nu)$. □

Since an independence kernel with bounded weight function $w = \pi/f$ satisfies a minorization condition $M(1, \beta, E, \nu,)$, any mixture or cycle containing such a kernel is uniformly ergodic. Thus any strategy can be made uniformly ergodic by inserting periodic or random "restart steps" using a restart distribution f with sufficiently thick tails.

3.3. *Limiting behavior of averages.* Suppose we use a single long run to estimate the expectation πf of a real-valued π-integrable function f by the sample average

$$\bar{f}_n = \frac{1}{n}\sum_{i=1}^{n} f(X_i).$$

The limiting behavior of this average is described by a law of large numbers and a central limit theorem.

A law of large numbers can be obtained from the ergodic theorem or the Chacon–Ornstein theorem. The following theorem is a corollary to Theorem 3.6 in Chapter 4 of Revuz (1975).

THEOREM 3. *Suppose X_n is ergodic with equilibrium distribution π and suppose f is real-valued and $\pi|f| < \infty$. Then for any initial distribution, $\bar{f}_n \to \pi f$ almost surely.*

The law of large numbers holds for any ergodic chain; it does not require any conditions on the rate of convergence to the stationary distribution. The central limit theorems that are available do require some assumptions on the rate of convergence. The following central limit theorem is given in Corollary 7.3 of Nummelin (1984).

THEOREM 4. *Suppose X_n is ergodic of degree 2 with equilibrium distribution π and suppose f is real-valued and bounded. Then there exists a real number $\sigma(f)$ such that the distribution of*

$$\sqrt{n}(\bar{f}_n - \pi f)$$

converges weakly to a normal distribution with mean 0 and variance $\sigma(f)^2$ for any initial distribution.

The boundedness assumption on f can be removed if the chain is uniformly ergodic.

THEOREM 5. *Suppose X_n is uniformly ergodic with equilibrium distribution π and suppose f is real-valued and $\pi(f^2) < \infty$. Then there exists a real number $\sigma(f)$ such that the distribution of*

$$\sqrt{n}(\bar{f}_n - \pi f)$$

converges weakly to a normal distribution with mean 0 and variance $\sigma(f)^2$ for any initial distribution.

This follows from Corollary 4.2(ii) of Cogburn (1972). The assumption that the chain is uniformly ergodic implies, in Cogburn's terminology, that the entire state space is uniform.

The conditions in these theorems are stronger than they need to be. Nummelin (1984), Theorem 7.6, gives a weaker condition, but it is stated in terms of expectations related to hitting times for sets with positive π-probability and appears rather difficult to verify directly. Kipnis and Varadhan (1986) give a central limit theorem for reversible Markov chains that only requires a finite limiting variance; Tóth (1986) extends their result to nonreversible chains, but has to add another condition that is not easy to verify or interpret.

Nummelin also gives an expression for $\sigma(f)$ in terms of potential operators for P. Kemeny and Snell (1976), Corollary 4.6.2, and Peskun (1973) give explicit expressions for $\sigma(f)$ for a finite state space. None of these expressions appear to be suitable for numerical determination of $\sigma(f)$ in the present context.

As a final note, the ergodicity assumptions in the theorems of this subsection imply that the chains are aperiodic. Aperiodicity is not necessary for these results on sample path averages to hold.

4. Implementation issues. In addition to examining the theoretical properties of a Markov chain, there are several implementation issues that need to be considered before using such a chain to examine a posterior distribution.

4.1. *Choosing a sampling plan.* There are two extreme approaches to using Markov chains to sample from a posterior distribution. At one extreme, a Markov chain can be used to generate n independent realizations from the posterior distribution by using n separate runs, each of length m, and retaining the final states from each chain. The run length m is to be chosen large enough to ensure that the chain has reached equilibrium. The other extreme is to use a single long run, or perhaps a small number of long runs. Experience and theoretical assessments in the simulation literature appear to favor the use of long runs [Bratley, Fox and Schrage (1987), Section 3.1.1; Kelton and Law (1984) and Whitt (1991)]. The major drawback of using short runs is that it is virtually impossible to tell when a run is long enough based on such runs. Even using long runs, determining how much of the initial series is affected by the starting state is very difficult, but some literature on the subject is available [Ripley (1987), Section 6.1]. A second drawback of short runs is that it makes inefficient use of the data: only n out of a total of nm data points are used. With a single run of length nm it is possible to use all the data, after possibly discarding a small initial fraction.

A complication that does arise from the dependence in using a single series is that variances of estimates are harder to obtain. Again, the simulation literature offers several alternatives, such as the use of batch means and time series analysis [Bratley, Fox and Schrage (1987), Chapter 3, and Ripley (1987), Chapter 6]. Some approaches designed specifically for Markov chains are described by Geyer (1992); other methods are discussed by Geweke (1992). Another approach currently being explored is to use regenerative simulation methods [Ripley (1987), Section 6.4] by identifying embedded renewal processes or by modifying Markov chain methods to have easily identified embedded renewal processes. An example of such a renewal process is given at the end of Section 2.3.4.

For some purposes it may nevertheless be useful to have an approximate

independent sample from the posterior. Using long runs this can be achieved by retaining every rth point of a sample path. The number r of points to skip in order to produce approximate independence can usually be chosen much smaller than the number m of steps needed to reach approximate equilibrium, since small amounts of correlation are usually much less serious than biases in estimates of means.

4.2. *Determining the run length.* Another consideration is to determine the total sample size or run length required for accurate estimates. For an i.i.d. sample of size n from a posterior distribution π, the standard deviation of the sample mean of a function $f(\theta)$ is $\sigma/\sqrt{n}$, where σ is the posterior standard deviation of $f(\theta)$. If a preliminary estimate of σ is available, perhaps from an asymptotic analysis, then this can be used to estimate the sample size that would be required in i.i.d. sampling. In dependent sampling, observations are generally positively correlated and a larger sample size will be required. If the series can be approximated by a first-order autoregressive process, then the asymptotic standard deviation of the sample mean is

$$\frac{\sigma}{\sqrt{n}}\sqrt{\frac{1+\rho}{1-\rho}},$$

where again σ is the posterior standard deviation of $f(\theta)$ and ρ is the autocorrelation of the series $f(X_n)$. A rough guess for ρ can thus be used to adjust the sample size for dependence in the series.

Instead of determining a fixed sample size in advance, it is also possible to use sequential or batch-sequential rules for determining when to stop sampling. Since prior information on the values of the posterior mean and standard deviation is often available from initial analysis, Bayesian sequential methods are a natural choice. Batching can be used to ensure that an assumption of normality for batched means is reasonable.

An intermediate option is to determine a sample size based on a pilot run. Raftery and Lewis (1992) suggest one possible approach.

4.3. *Variance reduction.* As with any simulation method, variance reduction techniques can often significantly reduce the sample sizes required for accurate estimates. Standard variance reduction methods such as importance sampling, conditioning, antithetic variates, control variates and common variates [Bratley, Fox and Schrage (1987), Chapter 2, and Ripley (1987), Chapter 5] can be used with any Markov chain method.

4.3.1. *Importance sampling.* Importance sampling can be used as a variance reduction method by constructing a Markov chain with equilibrium distribution π' instead of π and then weighting sample results with appropriate importance weights [Hastings (1970), Section 2.5]. To reduce the variance of an estimate of an expectation πf, the density $\pi'(x)$ would be chosen to be nearly proportional to $\pi(x)|f(x)|$.

Importance sampling is also useful if it is easier to construct a chain with equilibrium distribution π', or if a sample from such a chain is already available. For example, this approach can be used to estimate expectations for a nonconjugate prior distribution with a Gibbs sampler for a similar but conjugate prior distribution.

Using common sample paths with importance weights is particularly useful for comparing expectations under two similar distributions, perhaps corresponding to two different prior distributions or to the deletion of some observations. The positive correlations resulting form using the same series to estimate the two expectations will reduce the variance of the estimated difference. This can be viewed as an example of using common variates.

As with any application of importance sampling, unbounded weight functions should be used with caution.

4.3.2. *Conditioning.* Suppose an ergodic Markov chain produces pairs of the form (X_n, Y_n), the X margin of the equilibrium distribution is π and the conditional expectations $h(Y_n) = E[f(X_n) \mid Y_n]$ can be evaluated. Then the sample average of the conditional expectations $h(Y_n)$ will converge to the expectation πf. For a sequence of i.i.d. pairs the average of the conditional expectations has smaller variance than the average of the series $f(X_n)$. Whether this form of conditioning leads to a reduction in variance for a dependent series depends on the correlation structure. A sufficient condition is that the correlations in the $h(Y_n)$ series are no larger than the correlations in the $f(Y_n)$ series. Liu, Wong and Kong, (1991), Theorem 4.1, show that this condition holds for the random scan Gibbs sampler, in which the coordinate to update is selected independently according to a fixed distribution at each step. Another sufficient condition is given by Schmeiser and Chen (1991).

Conditioning is particularly useful for computing expectations or marginal densities of coordinate margins in Gibbs samplers, since the assumptions required for the Gibbs sampler imply that conditional means or densities of one parameter given the rest are usually available. In this case the variables Y_n consist of all coordinates of the Gibbs sampler sequence X_n except the one of interest. Gelfand and Smith (1990) refer to this use of conditioning as Rao-Blackwellization.

The pairs (X_n, Y_n) can be constructed to facilitate evaluating the conditional expectations $h(Y_n)$. In some cases these expectations are available in closed form, in others they can be approximated using asymptotic methods or low-dimensional numerical integration. A general method for constructing a chain (X_n, Y_n) is to start with a chain Y_n with invariant distribution π and another Markov kernel R with invariant distribution π. The X_n components are then generated from the distribution $R(Y_n, \cdot)$. The conditional expectations $h(Y_n) = (Rf)(Y_n)$ are particularly easy to calculate for the discrete limiting reflection and lattice kernels described in Sections 2.3.5 and 2.3.6.

A chain (X_n, Y_n) can also be formed by taking X_n to be a chain with invariant distribution π and taking Y_n to consist of X_{n-1} and some of the variates used to generate X_n from X_{n-1}. For example, Schmeiser and Chen (1991) propose

a conditioning scheme in which Y_n consists of X_{n-1} and the direction taken by the hit-and-run algorithm. The conditional expectation $h(Y_n)$ is computed by a one-dimensional integral. If this integral is evaluated numerically, then this may require a rather large number of posterior density evaluations. The number of evaluations can be reduced significantly by replacing the numerical integral by the conditional expectation of a finite grid chain with the conditional distribution as its invariant distribution.

The augmented chains (X_n, Y_n) used for conditioning are also Markov chains with known invariant distributions. Their properties can therefore be derived from the results of the preceding section.

4.3.3. *Antithetic and control variates.* Antithetic variation can be introduced into a Markov chain method by using a Metropolis step in which a candidate step is obtained by reflecting the current state of the chain through a point. If the posterior density is approximately symmetric about this point, then the sample will be also, and the resulting negative correlations will reduce variances of estimates of expectations of linear functions of θ. This technique can also be used to take advantage of approximate axial symmetries in a posterior distribution.

One way to introduce control variates into a Markov chain method is to use the sample path with importance weights to calculate estimates of normal approximations and to correct for the errors in these estimates. Another approach, described by Schmeiser and Chen (1991), is to calculate estimates of normal approximations using common variates, for example, by transforming the variates used to generate candidate steps in a Metropolis algorithm to appropriate normal variates.

4.4. *Other issues.* In using Markov chain methods, it is important to monitor the performance of the samplers to ensure that they are not exhibiting any unusual behavior. Gelfand and Smith (1990) propose the use of quantile plots to monitor performance. Monitoring sample paths of estimates is also useful for this purpose, as is monitoring autocorrelations of the parameters. Time series methods may also be useful for determining whether a series exhibits any unusual features.

For Metropolis chains it is also important to keep track of the number of candidates that are rejected. For an independence chain, the proportion of rejections can be related to the total variation distance between the posterior density π and the candidate generation density f.

Some consideration of numerical stability is needed in using any sampling-based method. Expressions used to evaluate log posterior densities obtained by translating mathematical formulas into a computer language are often reasonably stable near the posterior mode but may not be stable far away from the mode. This can lead to overflows or, on IEEE hardware, results that are NaN's or infinities. One way to avoid these problems is to carefully study the formulas for evaluating the log posterior density and modify them to be numerically stable even for extreme parameter values. The effort required to do

TABLE 1
Pump failures for pumps at the Farley 1 nuclear plant. Times are in thousands of hours

Pump	1	2	3	4	5	6	7	8	9	10
Failures	5	1	5	14	3	19	1	1	4	22
Times	94.320	15.720	62.880	125.760	5.240	31.440	1.048	1.048	2.096	10.480

this can be considerable. An expedient alternative that is often effective is to truncate the parameter space to a reasonable range that contains essentially all the posterior probability and for which the posterior density formula is numerically stable. This truncation also often ensures that a Markov chain used to sample from π is uniformly ergodic and thus improves the behavior of the Markov chain estimates.

A numerical issue that is unique to Markov chain methods is the possibility that rounding may introduce absorbing states. If this happens, results obtained from a Markov chain method may be meaningless. Again truncation away from areas of the state space where such rounding may occur can be helpful.

Another important consideration in selecting a Markov chain strategy is the cost of implementing and using the strategy. The costs can usually be broken down into three rough groups: the cost of coding the strategy, the cost of generating the chain and the cost of storing and processing the results. The importance of these costs varies from problem to problem, and as a result different chains may be optimal for different problems. There are also tradeoffs that may need to be considered. The Gibbs sampler is easier to code from scratch than most other methods, but code for a Gibbs sampler tends to contain fewer reusable components. Some strategies, such as grid-based chains, may require a rather large number of posterior density function evaluations but require a smaller sample size than other methods if these function values are used for variance reduction.

5. An example. One of the examples presented by Gelfand and Smith (1990) is a hierarchical Poisson model. Failures in pumps at a nuclear power plant are assumed to occur according to independent Poisson processes with each pump having its own failure rate $\lambda_1, \ldots, \lambda_{10}$. These rates are modeled as independent draws from a common distribution of rates. The pumps are observed for different periods of time. The counts and observation periods are given in Table 1.

Gaver and O'Muircheartaigh (1987) analyze these data using an empirical Bayes approach. Gelfand and Smith (1990) use the Gibbs sampler. They assume a conjugate prior structure in which the rates, conditional on a hyperparameter β, have a gamma distribution $G(\alpha, \beta)$ with density proportional to $x^{\alpha-1}e^{-\beta x}$, and β has a gamma distribution $G(\gamma, \delta)$ with $\gamma = 0.01$ and $\delta = 1$. For comparison with the results of Gaver and O'Muircheartaigh (1987), Gelfand and Smith set α to the method of moments estimator, $\alpha = 1.802$.

To allow for the possibility of outliers in the rates, Gaver and O'Muircheartaigh (1987) also consider several other distributions, including a t distribution

with 5 degrees of freedom for the logarithms of the rates. Carlin and Gelfand (1991) examine a prior of this form using a Gibbs sampler in which the conditional distributions are sampled by rejection sampling. We consider a variant of this model in which, conditional on a parameter θ, the logarithms of $(\lambda_i - \theta)/\sigma$ have a standard t distribution with 5 degrees of freedom, and θ has a normal distribution with mean μ and standard deviation τ. Following Carlin and Gelfand, we set $\mu = -1$ and $\tau = 1$. Like the parameter α in the gamma model, the parameter σ in the t model is a function of the conditional coefficient of variation of the rates. For a log normal rather than a log t model, the conditional means and coefficients of variation of the rates are equal if $\theta = \log(\alpha) - \frac{1}{2}\sigma^2 - \log(\beta)$ and $\sigma^2 = \log(1+1/\alpha)$. With these identifications, posterior expectations under the t model can be computed using the Gibbs sampler sequence under the gamma model and importance weights equal to the ratios of the t to the gamma density values.

The Gibbs sampler is particularly well suited to this problem under the conjugate gamma model. If the counts are denoted by s_i and the times by t_i, then, given β, the λ_i are independent $G(\alpha + s_i, t_i + \beta)$, and, given $\lambda_1, \ldots, \lambda_{10}$, the distribution of β is $G(\gamma + 10\alpha, \sum\lambda_i + \delta)$. Proposition 2 can be used to show that this Gibbs sampler is uniformly ergodic. Suppose the sampler is run by selecting first new values for the λ_i and then a new β. Then the distribution of the new β, $\beta_{(1)}$, depends on the initial values of the parameters only through the initial value of β, $\beta_{(0)}$. Let $f(\beta_{(1)} \mid \beta_{(0)})$ denote the conditional density of the new β given the initial one. It is sufficient to show that

$$h(\beta_{(1)}) = \inf_{\beta_{(0)}} f(\beta_{(1)} \mid \beta_{(0)})$$

is positive for all positive $\beta_{(1)}$. The probability distribution ν in the minorization condition with $m = 2$ needed to apply Proposition 2 can then be taken proportional to $\int f_{\beta \mid \lambda}(\beta \mid \lambda) f_{\lambda \mid \beta}(\lambda \mid u) h(u)\, du$. To show that h is positive, write

$$\begin{aligned}
f(\beta_{(1)} \mid \beta_{(0)}) &= \int \frac{1}{\Gamma(\gamma + 10\alpha)} \Big(\sum \lambda_i + \delta\Big)^{\gamma + 10\alpha} \beta_{(1)}^{\gamma + 10\alpha - 1} e^{-\beta_{(1)}(\Sigma\lambda_i + \delta)} \\
&\quad \times \left[\prod_{i=1}^{10} \frac{1}{\Gamma(\alpha + s_i)} \big(t_i + \beta_{(0)}\big)^{\alpha + s_i} \lambda_i^{\alpha + s_i - 1} e^{-\lambda_i(t_i + \beta_{(0)})}\right] d\lambda_1 \cdots d\lambda_{10} \\
&\geq \int \frac{1}{\Gamma(\gamma + 10\alpha)} \delta^{\gamma + 10\alpha} \beta_{(1)}^{\gamma + 10\alpha - 1} e^{-\beta_{(1)}(\Sigma\lambda_i + \delta)} \\
&\quad \times \left[\prod_{i=1}^{10} \frac{1}{\Gamma(\alpha + s_i)} \big(t_i + \beta_{(0)}\big)^{\alpha + s_i} \lambda_i^{\alpha + s_i - 1} e^{-\lambda_i(t_i + \beta_{(0)})}\right] d\lambda_1 \cdots d\lambda_{10} \\
&= \frac{1}{\Gamma(\gamma + 10\alpha)} \delta^{\gamma + 10\alpha} \beta_{(1)}^{\gamma + 10\alpha - 1} e^{-\beta_{(1)}\delta} \left[\prod_{i=1}^{10} \left(\frac{t_i + \beta_{(0)}}{t_i + \beta_{(0)} + \beta_{(1)}}\right)^{\alpha + s_i}\right] \\
&\geq \frac{1}{\Gamma(\gamma + 10\alpha)} \delta^{\gamma + 10\alpha} \beta_{(1)}^{\gamma + 10\alpha - 1} e^{-\beta_{(1)}\delta} \left[\prod_{i=1}^{10} \left(\frac{t_i}{t_i + \beta_{(1)}}\right)^{\alpha + s_i}\right].
\end{aligned}$$

The final right-hand side does not depend on $\beta_{(0)}$ and is positive, which completes the proof.

Four Markov chain methods were used to estimate failure rates for pumps 1, 5 and 10 under the t model. The first method used the Gibbs sampler for the gamma model with importance weights. The remaining three methods were Metropolis chains. All three were applied to the logarithms of the parameters standardized by the first-order approximate mean vector and covariance matrix. The first Metropolis algorithm was an independence chain with candidates generated by a multivariate t distribution with 2 degrees of freedom. The second algorithm was a random walk chain with increments generated from a normal distribution with independent components, zero means and standard deviations equal to 0.5. The final algorithm was a rejection independence chain with envelope density proportional to a mixture of a multivariate t distribution with 2 degrees of freedom and a standard normal distribution. The mixing probabilities were 0.2 for the t distribution and 0.8 for the normal distribution. The multiplier for the envelope was chosen based on a preliminary sample of 50 observations to produce a rejection probability in the candidate generation phase of approximately 0.85; thus the expected number of function evaluations needed per candidate should be approximately $1/0.15 = 6\frac{2}{3}$.

Conditioning was used to reduce variances of estimates from the resulting series. Since conditional means of one rate given all other parameters are not available under the t model, a sliding five-point lattice kernel as described at the end of Section 2.3.6 was used for conditioning. The lattice kernel was applied on the logarithm scale, and a lattice spacing of 1.5 asymptotic marginal posterior standard deviations was used.

The results for runs of length 5000 are shown in Table 2. Batch means based on batches of size 50 were used to estimate standard errors. Serial correlations in the batch means were negligible for all but the random walk samplers; for the random walk samplers the serial correlations among batch means were approximately 0.2 for all three rates. Standard errors for the random walk were adjusted for the correlations by modeling the batch means as first-order autoregressions.

For comparison, Table 2 also shows asymptotic approximations to the posterior means and standard deviations computed by applying the moment generating function method described in Tierney, Kass and Kadane (1989) on the logarithm scale.

The reweighted Gibbs sampler performs rather poorly for estimating expectations under the t model. The reason is that the t model has thicker tails than the gamma model, and the weight function is therefore unbounded and quite large in the tails. This approach might work more reasonably for a truncated t model.

The independence chain performs somewhat better than the random walk chain. This is not surprising, since the posterior distribution is not too different in shape from the candidate generation density for the independence chain and this chain usually performs better in such situations. Random walk chains tend to be less sensitive to the choice of the increment distribution and, as a

TABLE 2

Estimated posterior means for the t prior using the importance-weighted Gibbs sampler and independence, random walk and rejection Metropolis chains. For the Metropolis chains, R is the proportion of candidates rejected by the Metroplois algorithm. For the rejection chain, F is the average number of function evaluations per candidate step

		$\lambda_1 \times 100$	$\lambda_5 \times 10$	$\lambda_{10} \times 10$
Asymptotic approximations				
	Approx. post. mean	7.314	4.655	19.143
	Approx. post. SD	2.744	2.394	4.286
Importance-weighted Gibbs sampler				
Sample average	Est. post. mean.	6.921	4.927	16.800
	Est. stand. err.	0.404	0.302	0.539
Conditioning	Est. post. mean.	7.343	4.429	19.087
	Est. stand. err.	0.010	0.203	0.013
Independence ($R = 0.653$)				
Sample average	Est. post. mean.	7.459	4.778	19.160
	Estt. stand. err.	0.098	0.074	0.155
Conditioning	Est. post. mean.	7.292	4.646	19.166
	Est. stand. err.	0.019	0.025	0.012
Random walk ($R = 0.563$)				
Sample average	Est. post. mean.	7.436	4.685	19.592
	Estt. stand. err.	0.241	0.254	0.373
Conditioning	Est. post. mean.	7.317	4.569	19.137
	Est. stand. err.	0.020	0.065	0.017
Rejection ($R = 0.028$, $F = 7.81$)				
Sample average	Est. post. mean.	7.244	4.639	19.054
	Est. stand. err.	0.048	0.045	0.072
Conditioning	Est. post. mean.	7.295	4.630	19.125
	Est. stand. err.	0.008	0.011	0.009

result, work better than independence chains with a poorly chosen candidate generation density.

In this comparison of chains of length 5000 the rejection chain dominates all others. The Metropolis rejection rate of approximately 2.8% is very low, which means that the sequence produced by the rejection chain is almost an i.i.d. sequence from the posterior distribution. But the series required approximately 40,000 evaluations of the log posterior density function, compared to only 5000 for the random walk and independence chains. It is possible that better choices of the envelope density and constant could reduce the number of function evaluations required without seriously reducing the performance.

In almost all cases the use of the simple lattice conditioning kernel reduced variances by a factor of 10 or more. The five-point lattice requires a total of nine conditional density evaluations per observation. In the present example the conditional density of λ_i given all other components only depends on θ and is quite

simple. In other examples these evaluations would be more costly and would have to be compared to the cost of using a longer chain, perhaps with subsampling. For some strategies, such as grid-based Gibbs samplers, these conditional density values would be computed while generating the chain and would thus be available at no cost, provided storage is not a limiting consideration.

6. Conclusions. Markov chains can be used to explore posterior distributions in a variety of ways. Simple uses include estimating expectations under the posterior distribution and generating samples from the posterior distribution for constructing plots of one-, two- or three-dimensional margins. A more elaborate application is to use a Markov chain to control an animation in which a function of the parameter is viewed as the parameter is moved through the posterior distribution by the chain; an example of this approach is outlined in Tierney (1991).

Different Markov chains have different characteristics in different problems. Which characteristics are desirable can vary from one application to another. For computing averages it is usually desirable to reduce correlations, perhaps even making them negative, in order to reduce variances. For the animation example mentioned above, on the other hand, the strong positive serial correlations usually present in a random walk Metropolis chain are in fact an advantage. As a result, no single Markov chain method will dominate all others in all problems. It is important to be able to select or design a method with suitable characteristics from a range of methods and strategies for combining methods.

Hybrid algorithms provide a general way of adding certain characteristics, such as uniform ergodicity to speed convergence, to other algorithms. Another useful strategy is to incorporate approximate algorithms, such as rejection or grid-based algorithms, into a Metropolis algorithm to ensure that the invariant distribution is exactly equal to the posterior distribution.

More work is clearly needed to understand the effects of varying the parameters in Metropolis and hybrid algorithms and to determine good default values for these parameters. Strategies for adaptively setting these parameters would also be useful. Another important open issue for higher-dimensional problems is to determine when it is better to use algorithms that move through the parameter space in arbitrary directions, and when it is better to partition the parameter into components and use algorithms that, like the Gibbs sampler, change only one component at a time.

Acknowledgments. I would like to thank Charlie Geyer for helpful discussions and for bringing the work of K. S. Chan and Kipnis and Varadhan to my attention.

REFERENCES

ATHREYA, K. B., DOSS, H. and SETHURAMAN, J. (1992). A proof of convergence of the Markov chain simulation method. Technical Report 868, Dept. Statistics, Florida State Univ.

BARKER, A. A. (1965). Monte Carlo calculation of the radial distribution functions for a proton-electron plasma. *Austral. J. Phys.* **18** 119–133.

BESAG, J. and GREEN, P. J. (1993). Spatial statistics and Bayesian computation (with discussion). *J. Roy. Statist. Soc. Ser. B* **55** 25–37, 53–102.

BRATLEY, P., FOX, B. L. and SCHRAGE, L. E. (1987). *A Guide to Simulation*, 2nd ed. Springer, New York.

CARLIN, B. P. and GELFAND, A. E. (1991). An iterative Monte Carlo method for nonconjugate Bayesian analysis. *Statistics and Computing* **1** 119–128.

CHAN, K. S. (1989). A note on the geometric ergodicity of a Markov chain. *Adv. in Appl. Probab.* **21** 702–704.

CHAN, K. S. (1993). Asymptotic behavior of the Gibbs sampler. *J. Amer. Statist. Assoc.* **88** 320–326.

COGBURN, R. (1972). The central limit theorem for Markov processes. In *Proc. Sixth Berkeley Symp. Math. Statist. Probab.* **2** 485–512. Univ. California Press, Berkeley.

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. Amer. Statist. Assoc.* **85** 398–409.

GEMAN, S. and GEMAN, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6** 721–741.

GEWEKE, J. (1989). Bayesian inference in econometric models using Monte Carlo integration. *Econometrica* **57** 1317–1339.

GEWEKE, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments (with discussion). In *Bayesian Statistics 4* (J. O. Berger, J. M. Bernardo, A. P. Dawid and A. F. M. Smith, eds.) 169–193. Oxford Univ. Press.

GEYER, C. J. (1991). Monte Carlo maximum likelihood for dependent data. In *Computing Science and Statistics: 23rd Symposium on the Interface* (E. M. Keramidas, ed.) 156–163. Interface Foundation of North America, Fairfax Station, VA.

GEYER, C. (1992). Practical Markov chain Monte Carlo (with discussion). *Statist. Sci.* **7** 473–511.

HAMMERSLEY, J. M. and HANDSCOMB, D. C. (1964). *Monte Carlo Methods.* Chapman and Hall, London.

HASTINGS, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57** 97–109.

HILLS, S. E. and SMITH, A. F. M. (1992). Parameterization issues in Bayesian inference (with discussion). In *Bayesian Statistics 4* (J. O. Berger, J. M. Bernardo, A. P. Dawid and A. F. M. Smith, eds.) 227–246. Oxford Univ. Press.

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

KEMENY, J. G. and SNELL, J. L. (1976). *Finite Markov Chains.* Springer, New York.

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

KIRKPATRICK, S., GELATT, Jr., C. D. and VECCHI, M. P. (1983). Optimization by simulated annealing. *Science* **220** 671–680.

LIU, J., WONG, W. H. and KONG, A. (1991). Correlation structure and convergence rate of the Gibbs sampler: applications to the comparisons of estimators and augmentation schemes. Technical Report 299, Dept. Statistics, Univ. Chicago.

METROPOLIS, N., ROSENBLUTH, A. W., ROSENBLUTH, M. N., TELLER, A. H. and TELLER, E. (1953). Equations of state calculations by fast computing machines. *J. Chem. Phys.* **21** 1087–1091.

MÜLLER, P. (1991). A generic approach to posterior integration and Gibbs sampling. Technical Report 91-09, Dept. Statistics, Purdue Univ.

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

PESKUN, P. H. (1973). Optimum Monte-Carlo sampling using Markov chains. *Biometrika* **60** 607–612.

RAFTERY, A. E. and LEWIS, S. (1992). How many iterations in the Gibbs sampler? In *Bayesian Statistics 4* (J. O. Berger, J. M. Bernardo, A. P. Dawid and A. F. M. Smith, eds.) 763–773. Oxford Univ. Press.

REVUZ, D. (1975). *Markov Chains*. North-Holland, Amsterdam.

RIPLEY, B. D. (1987). *Stochastic Simulation*. Wiley, New York.

RITTER, C. and TANNER, M. A. (1991). The griddy Gibbs sampler. Technical Report, Div. Biostatistics, Univ. Rochester.

ROBERTS, G. and POLSON, N. (1990). A note on the geometric convergence of the Gibbs sampler. Unpublished manuscript.

SCHERVISH, M. J. and CARLIN, B. P. (1992). On the convergence of successive substitution sampling. *Journal of Computational and Graphical Statistics* **1** 111–127.

SCHMEISER, B. and CHEN, M. H. (1991). On random-direction Monte-Carlo sampling for evaluating integrals. Technical Report SMS 91-1, School of Industrial Engineering, Purdue Univ.

SMITH, R. L. (1984). Efficient Monte Carlo procedures for generating points uniformly distributed over bounded regions. *Oper. Res.* **32** 1296–1308.

STEWART, L. T. (1979). Multiparameter univariate Bayesian inference. *J. Amer. Statist. Assoc.* **74** 684–693.

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

TANNER, M. A. (1991). *Tools for Statistical Inference: Observed Data and Data Augmentation Methods. Lecture Notes in Statist.* **67**. Springer, New York.

TIERNEY, L. (1991). Exploring posterior distributions using Markov chains. In *Computer Science and Statistics: 23rd Symposium on the Interface* (E. M. Keramidas, ed.) 563–570.

TIERNEY, L., KASS, R. E. and KADANE, J. B. (1989). Fully exponential Laplace approximations to expectations and variances of nonpositive functions. *J. Amer. Statist. Assoc.* **84** 710–716.

TÓTH, B. (1986). Persistent random walks in random environment. *Probab. Theory Related Fields* **71** 615–625.

WHITT, W. (1991). The efficiency of one long run versus independent replications in steady-state simulation. *Management Sci.* **37** 645–666.

ZEGER, S. L. and KARIM, M. R. (1991). Generalized linear models with random effects: a Gibb's sampling approach. *J. Amer. Statist. Assoc.* **86** 79–86.

ZELLNER, A. and ROSSI, P. E. (1984). Bayesian analysis of dichotomous quantal response models. *J. Econometrics* **25** 365–393.

SCHOOL OF STATISTICS
UNIVERSITY OF MINNESOTA
270 VINCENT HALL
206 CHURCH STREET
MINNEAPOLIS, MINNESOTA 55455

DISCUSSION

HANI DOSS[1]

Ohio State University

0. Introduction. In my comments I discuss two topics, the basic convergence theorem (Theorem 1) and the importance-weighted Gibbs sampler, in particular, the question of assessing the variability of estimates formed by this method.

[1]Research supported by Air Force Office of Scientific Research Grant 94-1-0028.