Markov chain Monte Carlo: Some practical implications of theoretical results

by

Gareth O. Roberts* and Jeffrey S. Rosenthal**

(February 1997; revised August 1997.)

(Appeared in *Canadian Journal of Statistics* **26** (1998), 5–31.)

Abstract. We review and discuss some recent progress on the theory of Markov chain Monte Carlo applications, particularly oriented to applications in statistics. We attempt to assess the relevance of this theory for practical applications.

1. Introduction.

Markov chain Monte Carlo (MCMC) algorithms – such as the Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970) and the Gibbs sampler (Geman and Geman, 1984; Gelfand and Smith, 1990) – have been an extremely popular tool in statistics (see for example the recent reviews Smith and Roberts, 1993; Tierney, 1994; Gilks, Richardson, and Spiegelhalter, 1996). In addition to the large body of applied work which uses them, there has been a substantial amount of progress on the theoretical aspects of these algorithms. To the applied user, it is often unclear what lessons (if any) can be learned from these theoretical results. This paper will attempt to bridge this gap, by describing some practical implications of various theoretical results about MCMC.

The huge complexity of these MCMC algorithms means that only partial theoretical results are feasible. Thus, in considering practical implications, it is often necessary to

* Statistical Laboratory, University of Cambridge, Cambridge CB2 1SB, U.K. Internet: G.O.Roberts@statslab.cam.ac.uk.. Supported in part by EPSRC of the U.K.

** Department of Statistics, University of Toronto, Toronto, Ontario, Canada M5S 3G3. Internet: jeff@utstat.toronto.edu. Supported in part by NSERC of Canada.

extrapolate somewhat from what can be rigorously proven. As a result, some of the suggestions made in this paper will not be strictly implied by the theory. In addition, theoretical results cannot hope to answer all questions about how to use MCMC; an applied user must also rely on intuition, experimenting, instinct, previous experience, etc.

Furthermore, we do not attempt a comprehensive review of *all* possible implications of *all* theoretical results about MCMC. Rather, we concentrate on certain specific results only. For example, we will not mention results on the development of algorithms using auxiliary variables (see for example, Besag and Green, 1993, Neal, 1993, Marinari and Parisi, 1992); on the development of adaptive methods (for example Gilks, Roberts, and George, 1994; Gilks, Roberts, and Sahu, 1996); or results on updating strategies for Gibbs sampler schemes (see Roberts and Sahu, 1997); and there are doubtless many other omissions as well.

This paper is organised as follows. The basics of MCMC algorithms are presented in Section 2. Convergence issues of various sorts are reviewed and discussed in Sections 3, 4, and 5. Optimal scaling issues are presented in Section 6, and sensitivity of MCMC to computer approximation is considered in Section 7.

2. Basic algorithms.

MCMC algorithms are required in the following context. Suppose we have a probability distribution $\pi(\cdot)$, on a state space $\mathcal{X}$ (e.g., $\mathcal{X} = \mathbf{R}^d$). The distribution $\pi(\cdot)$ often describes the posterior distribution in a Bayesian inference problem. Typically, the state space $\mathcal{X}$ is so high-dimensional, and/or $\pi(\cdot)$ is so complicated, that direct computations regarding $\pi(\cdot)$ are impossible. (Indeed, even the normalising constant for $\pi(\cdot)$ is typically unknown.)

In such situations, MCMC proceeds by constructing a Markov chain $\mathcal{X}$, with transition probabilities $P(x, \cdot)$, such that $\pi(\cdot)$ is a *stationary distribution* for this chain, i.e.,

$$\pi P(\cdot) \equiv \int \pi(dx) P(x, \cdot) = \pi(\cdot) \ .$$

One then hopes that, if this chain is simulated long enough on a computer, with resulting values $X_0, X_1, \ldots$, the distributions $\mathcal{L}(X_n)$ will eventually be approximately $\pi(\cdot)$.

It is perhaps surprising that it could ever be easier to construct and simulate such a Markov chain with stationary distribution $\pi(\cdot)$, than it is to analyse $\pi(\cdot)$ directly. However, it turns out that there are a number of quite straightforward algorithms for constructing the transition probabilities $P(x, \cdot)$ in quite general contexts.

A very useful concept in constructing such transition probabilities is *reversibility*. A Markov chain is reversible with respect to $\pi(\cdot)$ if

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

This means that, if started in stationarity, the Markov chain has the same chance of starting at x and jumping to y as starting at y and jumping to x. It follows immediately that $\pi(\cdot)$ is a stationary distribution, since $\pi P(dy) \equiv \int \pi(dx)P(x, dy) = \int \pi(dy)P(y, dx) = \pi(dy)$. Thus, the problem of satisfying an integral equation for $P(x, \cdot)$ is simplified considerably.

A further important observation is that, if $\pi(\cdot)$ is stationary for both $P_1(x, \cdot)$ and $P_2(x, \cdot)$, then it is also stationary for P_1P_2 (i.e., performing first P_1 and then P_2), for $\frac{1}{2}(P_1 + P_2)$ (i.e., performing *either* P_1 *or* P_2, with probability $\frac{1}{2}$ each), etc. In other words, it is possible to build up more complicated algorithms out of simpler ones.

Because of this observation, most algorithms used in practice are built up from the following basic "building block", the *Metropolis-Hastings algorithm*. Suppose $\pi(dx) = f(x)\mu(dx)$, where $\mu(\cdot)$ is an arbitrary reference measure (e.g. Lebesgue measure). We begin with a *proposal distribution* $Q(x, dy) = q(x, y)\mu(dy)$. The Markov chain then proceeds by at each step *proposing* a new point $y \sim Q(x, \cdot)$, and then either *accepting* the proposal and moving to it, with probability

$$\alpha(x, y) = \min(1, \ \frac{f(y)q(y, x)}{f(x)q(x, y)}) \, , \tag{1}$$

or else *rejecting* it and not moving, with probability $1 - \alpha(x, y)$. (If $f(x)q(x, y) = 0$ then we automatically set $\alpha(x, y) = 1$.) The resulting transition probability is thus

$$P(x, dy) \ = \ q(x, y)\alpha(x, y)\mu(dy) \, , \qquad y \neq x \, ,$$

with $P(x, \{x\}) = 1 - \int q(x, y)\alpha(x, y)\mu(dy)$. It is easily seen that $\alpha(x, y)$ has been defined precisely so that P is reversible with respect to π. Thus, π is a stationary distribution for this chain.

The arbitrariness of the choice of $Q(x, \cdot)$ allows us considerable freedom to design a multitude of different chains, each with stationary distribution π. Some examples include:

- The *independence sampler*: (see for example Tierney, 1994) Here $Q(x, dy) = Q(dy)$ does not depend on x.
- The *random-walk Metropolis algorithm*: (Metropolis et al., 1953) Here $Q(x, dy) = q(y - x)\mu(dy)$ depends only on the difference $y - x$.
- The *Langevin algorithm*: (see for example Rossky et al., 1978; Grenander and Miller, 1994; Neal, 1993) Here f is a C^1 function on $\mathbf{R}^d$, and $\mu(\cdot)$ is d-dimensional Lebesgue measure. The proposal is of the form

$$x + hZ + \frac{h^2}{2}\nabla \log \pi \,, \tag{2}$$

  where $h > 0$ is constant, and where $Z \sim N(0, 1)$ has a standard normal distribution. (This choice of Q is motivated by the approximating continuous-time Langevin diffusion with stationary distribution $\pi(\cdot)$.)
- The *Gibbs sampler*: (Geman and Geman, 1984, Tanner and Wong, 1987, Gelfand and Smith, 1990) Here $\mathcal{X} = \mathcal{X}_1 \times \ldots \times \mathcal{X}_d$, and $Q = Q_i$ leaves all coordinates fixed except the i^{th} one, which it proposes according to the conditional distribution $\pi(x_i \mid \{x_j\}_{j \neq i})$. This implies that $\alpha(x, y) = 1$ for all x and y, so there are no rejections. If the resulting i^{th} component Gibbs sampler is called P_i, then these components can be combined to yield the *random-scan Gibbs sampler* which is the average $P_{RS} = \frac{1}{d}(P_1 + \ldots + P_d)$, or the *deterministic-scan Gibbs sampler* which is the product $P_{DU} = P_1 \ldots P_d$.

More sophisticated algorithms, constructed by combining basic version of Metropolis-Hastings chains, have been suggested in the literature. Many involve the introduction of so-called *auxiliary variables* (see for example Duane et al., 1987; Besag and Green, 1993; Neal, 1994) which aid the mixing of the chain. A huge variety of different types of chains have been constructed in the literature. Indeed, for some of them, the implementation itself (e.g. computing the acceptance probability) is highly non-trivial.

Practical implication #1. *When studying complicated probability distributions, there are a large number of MCMC algorithms available. A variety of algorithms should be considered, to determine which one is best for the specific problem at hand.*

3. Asymptotic convergence.

It is important to note that just because $\pi(\cdot)$ is a stationary distribution for $P(x, \cdot)$, it does *not* follow that the distributions $\mathcal{L}(X_n)$ will necessarily converge to $\pi(\cdot)$, as $n \to \infty$. For example, suppose $\mathcal{X} = \mathbf{R}^2$ and the distribution $\pi(\cdot)$ satisfies that $\pi(X = Y) = 1$; then the deterministic-scan Gibbs sampler for $\pi(\cdot)$ will simply replace each coordinate by the initial value of the second coordinate, and then not move again!

Fortunately, there are some simple conditions which guarantee that $\mathcal{L}(X_n)$ *will* converge to $\pi(\cdot)$, as $n \to \infty$. Specifically, if the chain is ϕ-irreducible and aperiodic, then it follows that we will have asymptotic convergence in total variation distance from almost every starting point (cf. Tierney, 1994, p. 1758). Here "ϕ-irreducible" means that there is some non-zero measure ϕ (e.g. Lebesgue measure) such that, for every set A with $\phi(A) > 0$, there is positive probability that the chain will eventually enter the set A started from any starting value $x \in \mathcal{X}$. Also "aperiodic" means that $\mathcal{X}$ does not contain nonempty disjoint subsets $\mathcal{X}_1, \mathcal{X}_2, \ldots \mathcal{X}_j$, with $j \geq 2$, such that $P(x, \mathcal{X}_{i+1 \bmod j}) = 1$ whenever $x \in \mathcal{X}_i$. Finally, the total variation distance between two probability measures μ and ν is defined to be $\|\mu - \nu\| \equiv \sup_{A \subseteq \mathcal{X}} |\mu(A) - \nu(A)|$. In terms of these definitions, the formal statement of the convergence theorem is as follows: *Let $P(x, \cdot)$ be the transition probabilities for a Markov chain on a state space $\mathcal{X}$, with stationary distribution $\pi(\cdot)$. Suppose the chain is ϕ-irreducible and aperiodic. Then for π-almost all starting points $x \in \mathcal{X}$, we have that as $n \to \infty$,*

$$\|\mathcal{L}(X_n \,|\, X_0 = x) - \pi(\cdot)\| \to 0\,.$$

Note that the above result does allow for some exceptional starting points from which convergence will not take place. However, such exceptional points rarely arise in practice; in particular, for most ϕ-irreducible Gibbs samplers and for all ϕ-irreducible Metropolis-Hastings algorithms, the chain is *Harris recurrent* and there are no exceptional points (cf.

Tierney, 1994, Section 3.1). A detailed treatment of ϕ-irreducibility and aperiodicity may be found in Meyn and Tweedie (1993, Chapters 4 and 5).

More problem-specific criteria for ϕ-irreducibility and aperiodicity are available. For instance, if $\pi(\cdot)$ has continuous density with respect to Lebesgue measure, and if the interior of the support of $\pi(\cdot)$ is connected, then the resulting Gibbs sampler is always ϕ-irreducible where ϕ is Lebesgue measure (Roberts and Smith, 1994).

It is worth noting that hybrid chains do not necessarily inherit the irreducibility and aperiodicity properties of the constituent parts. For instance, if P_1 and P_2 are ϕ-irreducible and aperiodic and both have stationary distribution π, it does not necessarily follow that P_1P_2 is irreducible. For instance, let $\mathcal{X} = \{1, 2, 3\}$ and suppose $P_1(1, 2) = P_1(2, 3) = P_2(1, 3) = P_2(2, 1) = 1$ with $P_i(3, i) = P_i(3, 3) = 1/2$ for $i = 1, 2$. Here P_1 and P_2 are both aperiodic and irreducible, each with stationary distribution given by $\pi(1) = \pi(2) = 1/4$, $\pi(3) = 1/2$, but $P_1P_2(1, 1) = 1$ so that P_1P_2 is not irreducible.

On the other hand, random-scan hybrids of ϕ-irreducible algorithms are always ϕ-irreducible. This can be seen easily from the fact that if $P_1^n(x, A) > 0$ then the random-scan samplers always have positive probability of sampling from P_1 at each of the first n iterations, so that $[\frac{1}{2}(P_1 + P_2)]^n(x, A) > 0$ also.

In practice hybrid algorithms are often constructed from reducible component algorithms (for example the Gibbs sampler, where the component algorithms each act on just one coordinate), so that convergence properties of the hybrid algorithm have to be analysed directly.

Practical implication #2. *When considering different MCMC algorithms, it is important – and often not very difficult – to verify that the Markov chain is ϕ-irreducible and aperiodic.*

4. Geometric convergence.

Even if a chain is asymptotically convergent, questions remain about the nature and speed of this convergence. In particular, one convergence property of interest is geometric ergodicity. A chain is *geometrically ergodic* if for π-almost all $x \in \mathcal{X}$, there is $\rho < 1$ and $M(x) < \infty$, such that $\|\mathcal{L}(X_n \,|\, X_0 = x) - \pi(\cdot)\| \leq M(x)\rho^n$. (Furthermore, it may be assumed without loss of generality that ρ above is independent of x, see e.g. Nummelin and Tweedie, 1978; Roberts and Rosenthal, 1996.)

To study geometric ergodicity, we introduce the following concepts. A subset $C \subseteq \mathcal{X}$ is *small* if there exists $n_0 \in \mathbf{N}$, $\epsilon > 0$, and a probability measure $\nu(\cdot)$, such that

$$P^{n_0}(x, \cdot) \;\geq\; \epsilon\, \nu(\cdot)\,, \qquad x \in C\,. \tag{3}$$

The chain satisfies a *geometric drift condition* for the small set C, if there is a π-almost everywhere finite function $V : \mathcal{X} \to [1, \infty]$, and constants $\lambda < 1$ and $b < \infty$, such that

$$PV(x) \;\equiv\; \int V(y) P(x, dy) \;\leq\; \lambda V(x) + b\mathbf{1}_C(x)\,, \qquad x \in \mathcal{X}\,. \tag{4}$$

Then a basic result (cf. Meyn and Tweedie, 1993, Chapter 15) is that a chain is geometrically ergodic if and only if it satisfies a geometric drift condition for some small set C.

(If $\mathcal{X}$ itself is small, we say the Markov chain is *uniformly ergodic*. However, this does not often occur in statistical models with unbounded parameters.)

Often, all bounded subsets of $\mathcal{X}$ are small for P. (For example, this will be the case if some power $P^n(x, \cdot)$ has density bounded below in an ϵ-neighbourhood of x, uniformly over $x \in \mathcal{X}$, see e.g. Roberts and Tweedie, 1996a; Roberts and Rosenthal, 1997b.) In such cases, to prove geometric ergodicity, it suffices to prove that, for some function V,

$$\limsup_{|x| \to \infty} \frac{PV(x)}{V(x)} \;<\; 1\,.$$

Such ideas are used to prove geometric ergodicity for a variety of MCMC algorithms, in Chan (1993), Roberts and Tweedie (1996a, 1996b), and Roberts and Rosenthal (1997b).

For further background about drift conditions and geometric ergodicity, see Nummelin (1984) and Meyn and Tweedie (1993). For other approaches to geometric ergodicity of

MCMC algorithms under different norms, see Frieze et al. (1994), Roberts and Polson (1994), Schervish and Carlin (1992), Liu et al. (1994, 1995), Baxter and Rosenthal (1995), Polson (1996), Roberts and Rosenthal (1997a), and Holden (1996). Roberts and Rosenthal (1997a, 1997b) consider results which imply geometric ergodicity of hybrid algorithms in terms of conditions on the constituent algorithms.

Of course, geometric ergodicity is an asymptotic property and is therefore not directly connected to finite-time simulations. However, it still provides a very useful guideline in determining which algorithms are likely to perform well in practice.

In addition, geometric ergodicity implies the existence of central limit theorems for ergodic averages of functionals (Tierney, 1994; Geyer, 1992; Chan and Geyer, 1994; Roberts and Rosenthal, 1997a). While not the weakest condition to imply central limit theorems, geometric ergodicity is one of the easiest to check and leads to clean statements. For example, it follows from Roberts and Rosenthal (1997a, Theorem 4) that: *If $P(x, \cdot)$ is geometrically ergodic and reversible, and $g \in L^2(\pi)$ with $\int g(y)\pi(dy) = 0$, then there is $\sigma_g^2 < \infty$, such that*

$$\mathcal{L}\left(\frac{1}{\sqrt{n}}\sum_{j=1}^{n} g(X_j)\right) \;\Rightarrow\; N(0, \sigma_g^2)\,.$$

In fact, $\sigma_g^2 = \mathbf{Var}_\pi(g) + 2\sum_{i=1}^{\infty} \mathbf{Cov}\left(g(X_0), g(X_i)\right)$ (cf. Geyer, 1992). The finiteness of this sum is ensured by geometric ergodicity.

For non-geometrically ergodic chains, central limit theorems can easily fail to hold (see for example the results of Roberts, 1996). For a specific example, consider the independence sampler with $\mathcal{X} = \mathbf{R}^+$, $\pi(\cdot) = \mathbf{Exp}(1)$, and $Q(x, \cdot) = Q(\cdot) = \mathbf{Exp}(k)$ (i.e., with density ke^{-ky}) for some $k > 2$. It can be shown that central limit theorems for this chain do not hold. Instead, once the chain reaches a very large value, it will tend to reject subsequent proposals with high probability, and get "stuck" there.

To make these ideas more concrete, we consider the above example with two possible values for k, namely $k = 0.01$ and $k = 5$, giving rise to transition kernels P_1 and P_2 say. Clearly neither choice is particularly effective at representing π. General results about the independence sampler (see Smith and Tierney, 1996) imply that in fact P_2 is

not geometrically ergodic, while P_1 is, albeit with the somewhat slow convergence rate of $\rho = 0.99$. In fact, from Roberts and Rosenthal (1997a) a central limit theorem holds for P_1, whereas from Roberts (1996) a central limit theorem does *not* hold for P_2. So what happens in practice?

The following experiment was carried out for both P_1 and P_2 to assess the effect of running these two algorithms on their corresponding ergodic estimates. Chains of one million iterations each were simulated 55 times, for each of the two algorithms. In each case the algorithm was started at $X_0 = 1$, the mean value under π. Figure 1 gives kernel density estimates of the distribution of the ergodic average

$$10^{-6} \sum_{i=1}^{10^6} X_i$$

for both P_1 and P_2, in each case based on the 55 observed values.

Figure 1. Kernel density estimates of the distribution of the ergodic mean for P_1 (with $k = 0.01$) and P_2 (with $k = 5$), each based on 55 runs

of one million iterations each. Note that for P_1 (which is geometrically ergodic), the density is much more symmetric and is also much more concentrated around the true mean of 1.0.

It is seen that P_1 mixes somewhat slowly, but after such a long run, its ergodic estimates are approximately normally distributed about 1 with small variance. On the other hand, P_2 *appears* to converge reasonably in most of the 55 runs, though usually to values significantly below 1. Some runs however, which have managed to reach relatively high values, get "stuck" there for large numbers of iterations. The effect of this is that the ergodic estimates for P_2 have median considerably below 1 and are very heavily positively skewed.

Figure 2. Two typical simulation runs for the non-geometric chain P_2, each of length one million iterations. Note the widely different qualitative behaviour, leading to very different sample means and autocorrelation functions. This demonstrates that for non-geometric chains, different runs can have widely different characteristics, making estima-

tion hazardous.

The dangers of using P_2 are illustrated by observing traces of two selected runs and their corresponding autocorrelation plots, as shown in Figure 2. The first trace sticks for about 400 000 iterations at a high value. In no sense can the run be said to have "converged" after 10^6 iterations. However the second trace appears to have settled down sufficiently and its autocorrelation plot shows some signs of stability. (Indeed, this run would likely fool most standard convergence diagnostics into thinking convergence had occurred; see Section 5.) Its ergodic average value, on the other hand, is 0.812, considerably less than the correct value 1.

Considerations such as those occurring in this example imply the following.

Practical implication #3. *When choosing an MCMC algorithm, it is desirable if possible to find an algorithm which can be shown to be geometrically ergodic.*

5. Quantitative convergence rates.

Even geometric ergodicity of a Markov chain gives no quantitative information about how long the chain needs to be simulated until approximate stationarity is achieved. We consider those questions in this section.

In some cases it is possible to prove rigorous results about convergence times. For Markov chains on finite spaces, there has been a great deal of work in this area (see e.g. Jerrum and Sinclair, 1988; Sinclair, 1992, 1993; Frigessi et al., 1992, 1993; Frieze et al., 1994; Diaconis and Stroock, 1991; Ingrassia, 1994). Unfortunately, most statistical inference problems have uncountable parameter spaces so that these results do not directly apply (though through truncation arguments they can sometimes be used anyway to some extent, cf. Tweedie, 1996; Rosenthal, 1996). Also, the special case of the independence sampler has been solved exactly (Liu, 1996; Smith and Tierney, 1996), leading to precise information about distance to stationarity. However, this result does not generalise to other MCMC algorithms.

To consider infinite state spaces for algorithms other than the independence sampler, various authors have derived quantitative bounds on the distance to stationarity of Markov

chains after n iterations, in terms of a minorisation condition (3) and a drift condition (4). Such considerations led (Rosenthal, 1995a) to a result about asymptotic running times (with large numbers of parameters) for the Gibbs sampler for variance component models, though the result did not give clear quantitative bounds for finite numbers of parameters. This approach was extended to general models in Meyn and Tweedie (1994), where quantitative exponentially-decreasing bounds were obtained. However, to deal with near-periodicity issues it was necessary either to make very strong assumptions (e.g. strong aperiodicity), or to have the resulting bounds depend on the minorisation and drift conditions in a very complicated way, and therefore be extremely large. The method of Meyn and Tweedie was applied to Metropolis-Hastings algorithms by Mengersen and Tweedie (1996), and specialised to stochastically-ordered Markov chains by Lund et al. (1996). Related results were also developed by Baxendale (1994).

Problems of near-periodicity were circumvented in Rosenthal (1995b), by requiring that the minorisation and drift conditions satisfy $d > \frac{2b}{1-\lambda}$, where $d = \sup_{x \in C} V(x)$. This implied that in theory two chains (one started in stationarity) could be forced to couple at a finite "coupling time" whose distribution had specified tails. The coupling inequality then gave the following result. *Suppose a Markov chain satisfies the minorisation condition* (3), *and also satisfies the drift condition* (4). *Suppose further that* $C = \{x \in \mathcal{X} \, ; \, V(x) \leq d\}$ *for some* $d > \frac{2b}{1-\lambda}$. *Then for any* $0 < r < 1$, *the total variation distance to stationarity of the chain after* n *iterations is bounded above by*

$$(1-\epsilon)^{rk} + \left(\alpha^{-(1-r)}\gamma^r\right)^k \left(1 + \frac{b}{1-\lambda} + \mathbf{E}\left(V(X_0)\right)\right),$$

where

$$\alpha^{-1} \;=\; \frac{1+2b+\lambda d}{1+d} \;<\; 1\,; \quad \gamma \;=\; 1 + 2(\lambda d + b)\,.$$

This result thus gives a quantitative, exponentially-decreasing bound on the distance to stationarity for any Markov chain, provided minorisation and drift conditions can be verified (with $d > \frac{2b}{1-\lambda}$). In particular, the bounding quantities α and γ are simple functions of the constants in these two conditions. The result was applied (Rosenthal 1995b, 1996a) to some realistic, high-dimensional Gibbs samplers for certain posterior distribu-

tions, leading to useful, reasonable, quantitative bounds on the running times of these algorithms.

Concerns about near-periodic behaviour, which made the Meyn and Tweedie (1994) result much more complicated, and forced the condition $d > \frac{2b}{1-\lambda}$ in Rosenthal (1995a), were avoided altogether in Roberts and Rosenthal (1996). There, rather than considering the distributions of the individual values $\mathcal{L}(X_n)$, the ergodic averages of distributions, $\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}(X_i)$, were considered instead. This removed the restriction $d > \frac{2b}{1-\lambda}$ since it allowed for the application of *shift-coupling* rather than ordinary coupling. It gave the result that the total variation distance to stationarity of $\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}(X_i)$ was bounded above by

$$\frac{1}{n}\sum_{k=1}^{n}\left(2(1-\epsilon)^{rk} + \lambda^{(1-r)k}A^{rk}\left(\mathbf{E}\left(V(X_0)\right) + \frac{b}{1-\lambda}\right)\right),$$

where $0 < r < 1$ is arbitrary and where ϵ, λ, and b are as in (3) and (4). This result was applied (Roberts and Rosenthal, 1996) to a number of examples, giving substantially improved bounds over previous analyses.

Comparisons of these different general methods for obtaining rigorous quantitative bounds on convergence were begun in Mengersen et al. (1996). Various convergence rate theorems were directly applied to a variety of simple examples of Metropolis-Hastings algorithms. It was found there that, of the theorems considered, those of Roberts and Rosenthal (1996) proved the quickest (i.e. best) convergence times; those of Rosenthal (1995) were second quickest; those of Baxendale (1994) were third; and those of Mengersen and Tweedie (1996) were the slowest. On the other hand, for some different examples which satisfied the special condition of stochastic monotonicity, the method of Lund et al. (1996) provided substantially improved bounds.

Despite some clear successes, it must be recognised that analytic verification of minorisation and drift conditions is not feasible in most very complicated, high-dimensional problems. Moreover, even when verification is possible, the resulting computable bounds may be too large to be of practical value. Thus, it remains the case that these rigorous quantitative bounds are not available in general.

Given the difficulties in analytically verifying drift and minorisation conditions, it is

sometimes not possible to do this for complicated models of interest. In such cases, it may be possible to estimate the constants ϵ, λ, and b, in equations (3) and (4), through auxiliary simulation. This is pursued in Cowles and Rosenthal (1996), where convergence rates are estimated for several different Gibbs samplers, including for variance components models and for ordinal probit models. The method appears to hold promise for further analysis of other models.

An alternative approach, similarly bridging the gap between rigorous and non-rigorous results, is pursued in Roberts and Sahu (1996). There, rates of convergence for Gibbs samplers are approximated by rates of convergence of approximating Gaussian target densities. Since posterior distributions from most statistical models, with enough data, are approximately Gaussian, this approach does a good job of approximating convergence rates in many practical problems.

When neither rigorous nor approximation methods are available, all that remains are purely empirical convergence diagnostics. Specifically, chosen functions of a simulated Markov chain are monitored, and statistical procedures of used to assess stationarity of the monitored functions. Two of the most popular convergence diagnostics are those of Gelman and Rubin (1992) and Raftery and Lewis (1992); for reviews see Cowles and Carlin (1995) and Brooks and Roberts (1996). Unfortunately, it is well known (cf. Cowles and Carlin, 1995) that all convergence diagnostics can sometimes prematurely claim convergence, for example on the notorious "witch's hat" example (Polson, 1991; Mathews, 1993).

We also wish to draw attention to another potential problem with convergence diagnostics, which is perhaps less well known. This is the problem that, even if the Markov chain is converging perfectly, the mere act of waiting for diagnostic success may itself introduce biases in the result. To illustrate this, consider a very simplified convergence diagnostic, one which waits until two successive batch means (each of size m) are within ϵ of each other, and then outputs the resulting final batch mean. That is, we set $A_{m,i} = \frac{1}{m} \sum_{t=mi+1}^{m(i+1)} g(X_t)$, set $i^* = \inf\{i\,;\, |A_{m,i^*} - A_{m,i^*-1}| < \epsilon\}$, and consider A_{m,i^*} as an approximation of the expected value of g under $\pi(\cdot)$. This seems like a reasonable procedure, and is similar in spirit to currently used diagnostics. However, biases are introduced. To see this, suppose that the Markov chain itself is actually perfectly converging, so that $X_1, X_2, \ldots$ are in fact

i.i.d. distributed as $\pi(\cdot)$. Even still, for small ϵ, if the distribution of $A_{m,i}$ has density f, then the distribution of A_{m,i^*} will have density approximately proportional to f^2, which typically will have a different mean. For a specific example, if $\mathcal{L}(X_i) = \mathbf{Gamma}(a, b)$, then $\mathcal{L}(A_{m,i}) = \mathbf{Gamma}(ma, mb)$ and (as $\epsilon \to 0$) we have $\mathcal{L}(A_{m,i^*}) \approx \mathbf{Gamma}(2ma - 1, 2mb)$. It follows that the bias in the estimator will be $\frac{-1}{2mb}$, which could be significant if the batch size m is not sufficiently large. (Of course, because of the i.i.d. nature of this chain, if we instead use A_{m,i^*+1} then we will avoid bias. But for a more realistic non-i.i.d. chain, the bias of A_{m,i^*+1} would be comparable to that of A_{m,i^*}.)

Practical implication #4. *Quantitative rates of convergence are always an important issue. It is best to have rigorous computable bounds if possible, though this can be difficult. If not, then it may be easier to use auxiliary-simulation or approximation results. If such results are too difficult or time-consuming to obtain, then convergence diagnostics must be used. However, they should only be used with extreme caution, being careful to avoid both premature diagnoses of convergence and the introduction of bias into the result.*

6. Scaling.

One of the problems with Metropolis-Hastings algorithms is the abundance of choice available for choosing the proposal distribution $Q(x, \cdot)$. For instance even if the type of algorithm (perhaps the random walk Metropolis algorithm) has been chosen, it is necessary to scale the proposal variance to be appropriate for $\pi(\cdot)$. Such a problem is known as a *scaling* problem.

To make this question more concrete, consider the following problem. Suppose that $\pi(\cdot)$ is absolutely continuous with respect to d-dimensional Lebesgue measure, with density again denoted by π say. Suppose also that $Q(x.\cdot)$ is distributed as the d-dimensional normal distribution $N_d(x, \sigma^2 I_d)$, for some $\sigma^2 > 0$. We recall that the acceptance probabilities for this algorithm are given by (1).

For very small values of σ^2, small jumps are attempted by the algorithm, and because of the form of (1), these moves are almost always accepted. The Markov chain mixes very slowly because its increments are so small. On the other hand, if σ^2 is chosen to be very large, long distance jumps are attempted by the algorithm, most of which are rejected. The

algorithm therefore spends long periods of time in the same state, and thus the algorithm still converges slowly.

For this problem, "very large" and "very small" have to be interpreted in a way related to the particular form of π. It seems reasonable that "moderate" values of σ^2 should be preferred. However, it is difficult to see how to figure out what values are "moderate", especially if π is very complicated.

Since the efficiency of a Markov chain for estimation varies with the quantity to be estimated, to be able to make any useful statement about the Markov chain as a whole, it has been necessary to consider a suitably regular sequence of target densities on state spaces $\{\mathcal{X}_d\}$ of increasing dimension, and to consider asymptotics as the dimension $d \to \infty$. The following approach follows Gelman, Roberts, and Gilks (1996) and Roberts, Gelman, and Gilks (1997).

Set $\pi_d = \prod_{i=1}^{d} f(x_i)$ and suppose that we take σ_d^2 to be ℓ^2/d for some constant ℓ. Now consider a continuous time process $\{Z^d\}$, defined by the first component of the d-dimensional Markov chain, with time scaled by a factor of $1/d$. Now Z^d is not Markov, but as $d \to \infty$ (and under suitable regularity conditions) it converges to the limiting Langevin diffusion process:

$$dZ_t = h(\ell)^{1/2} dB_t + \frac{h(\ell) \nabla \log \pi(Z_t)}{2} dt \ ,$$

where

$$h(\ell) = \ell^2 \times 2\Phi\left(-\frac{I\ell}{2}\right),$$

with $I = E_f[((\log f(X))')^2]$ and $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-s^2/2} ds$.

Thus the optimal limiting algorithm is that which maximises $h(\ell)$ and this optimal limit is independent of which functions of π are of interest. Numerical maximisation of h is easy but leaves a solution which is a function of I. Unfortunately I is not in general available. However the optimal solution can be characterised by the solution which produces an algorithm which accepts approximately 0.234 of its proposed iterations.

Extensions of this result to larger classes of target distribution is possible; these and other practical issues are discussed in Gelman, Roberts, and Gilks (1996) and Roberts, Gelman, and Gilks (1997). The useful property of these results is that the optimal limiting

acceptance rate is not affected by the type of target density, so that its use as a guideline in practice is straightforward.

It should be stressed however that this is an asymptotic result, in two senses. First, the asymptotic acceptance rate is defined as the average acceptance rate, averaged with respect to the asymptotic (i.e., stationary) distribution $\pi(\cdot)$, which may be different from the observed rate if the chain is started in some other distribution and run for too short a time. Second, the optimality results are only proven asymptotically as the dimension $d \to \infty$. Although for some examples the convergence to the large-d limit is relatively quick as a function of dimension, in general for heavily correlated target densities, optimal scalings can be very different from 0.234. It is also worth noting that it is rarely a sensible idea to fine tune scalings too carefully, since even in the asymptotic case, reasonably efficient algorithms (relative to maximal possible efficiency) can be achieved with scalings between acceptance rates of approximately 0.15 and 0.5 (see e.g. Roberts and Rosenthal, 1995, Figure 3.1). In addition, interactive scaling of any kind is liable to alter the stationarity properties of the chain (and thus invalidate the algorithm) if continued indefinitely. Therefore any tuning of the algorithm should be carried out as a pilot sample analysis only. As a result, any efficiency gain from excessive fine-tuning is usually lost in the time spent carrying out pilot studies.

The optimal asymptotic value of 0.234 also occurs for other types of problems. For instance, if the target density is the product of *discrete* densities and a Metropolis type update is used, there is a corresponding limit theorem (though this time not to a Langevin diffusion) resulting in the same asymptotically optimal acceptance rate (see Roberts, 1997).

For more problem specific algorithms such as Langevin algorithms, similar asymptotic results are available (see Roberts and Rosenthal, 1995). In fact for the basic Langevin algorithm of equation (2), the optimal acceptance rate is approximately 0.574. Like the random-walk Metropolis case discussed above, this is proved rigorously only for i.i.d. distributions (with certain generalisations available). Again, acceptance rates near to optimal will also result in relatively good performance.

One of the most interesting aspects of these results is their implications for running-time complexity. It follows from the nature of the convergence of these algorithms to their

limiting diffusions, that the running time of random-walk Metropolis is $O(d)$, and that of the classical Langevin diffusion algorithms (as in equation (2)) is $O(d^{1/3})$, at least for sufficiently smooth densities of product form (see also Kennedy and Pendleton, 1991). Note that we do not expect this result to hold for certain multimodal sequences of densities; see Roberts, Gelman, and Gilks (1997) for a discussion of some of these possibilities.

Practical implication #5. *For certain algorithms, simple and easy-to-use rules of thumb are available as guidelines for scaling proposal distributions. These rules are supported by limiting results for high-dimensional problems. However, excessive fine tuning of proposal distributions is not necessary.*

7. Sensitivity analysis.

In practice, all algorithms are ultimately carried out by computer simulation. Therefore, the Markov chain *actually* simulated is only an approximation of the true chain. Such effects as finite precision and finite range are introduced and pose further questions about the validity of these algorithms.

Pseudo-randomness is well known to be a thorny issue requiring great care and potentially adversely affecting results (see e.g. Ripley, 1987; Hammersley and Handscomb, 1964), however to the best of our knowledge the implications of these problems in these context of Markov chain simulation are largely unexplored, and we do not pursue them here. (For some interesting and cautionary preliminary results, see Ferrenberg et al., 1992; Vattulainen et al., 1994.)

Issues related to finite-precision arithmetic (i.e., "roundoff error") were considered in Roberts, Rosenthal, and Schwartz (1995). Roundoff error was defined in terms of a function $h : \mathcal{X} \to \mathcal{X}$, with $h(x)$ "close" to x for each $x \in \mathcal{X}$. The resulting approximate Markov chain has transition kernel

$$\widehat{P}(x, A) \;=\; P(x, h^{-1}(A))\,.$$

The paper showed (Proposition 1) that for some chains – even if geometrically ergodic and strong Feller continuous – an arbitrarily small roundoff error may result in a transient chain, having no stationary distribution at all!

On the other hand, it was proven (Roberts, Rosenthal, and Schwartz, 1995, Theorems 4 and 7) that if the original chain is geometrically ergodic, with a drift function V having the property that $\log V$ is uniformly continuous on $\mathcal{X}$, then $\widehat{P}$ will automatically be geometrically ergodic, for sufficiently small $\sup |h(x) - x|$. It was further shown (Theorems 9 and 11) that the stationary distribution of $\widehat{P}$ will be close (weakly) to that of P.

The need to truncate Markov chains for their implementation on computers can also cause problems. In fact, since computer programs are likely to just crash if a value is recorded which is "out of range", it is arguable that the appropriate limit of interest for truncated MCMC is

$$\tilde{\pi} = \lim_{n \to \infty} \mathcal{L}(X_n | \tau > n) \ ,$$

where τ is the first time that the computer records an out of range value. Moreover, the correct ergodic limit for the estimation of the moment $\int_{\mathcal{X}} g(x)\pi(dx)$ is arguably the following weak limit:

$$\lim_{n \to \infty} \mathcal{L}\big(\frac{\sum_{i=1}^{n} g(X_i)}{n} \ \Big| \ \tau > n \big).$$

It is known (cf. Breyer and Roberts, 1997) that (at least for large enough range) the above limits would approximate the appropriate classical stationary limit. Such results fall into the domain of the area known as *quasi-stationarity*. Geometric ergodicity plays a role here as well, and in fact it turns out that for non-geometrically ergodic chains, such quasi-stationary limits usually do not exist.

An alternative approach to the truncation problem, without resorting to quasi-stationary methods, was considered in Tweedie (1996).

Practical implication #6. *When running MCMC algorithms, it is possible that computer limitations will adversely affect the results. However, for many chains (including those which are geometric with a log-uniformly-continuous drift function), small roundoff errors do not significantly affect the convergence properties. Truncated algorithms have unstable convergence properties for non-geometric algorithms.*

8. Conclusions.

MCMC algorithms are clearly a very exciting and widely used application of Markov chains to complicated problems of inference, estimation, and integration. While in many ways MCMC renders these problems far easier, on the other hand MCMC algorithms are themselves complicated, difficult to understand completely, and potentially problematic.

It is thus important for applied users of MCMC algorithms to understand, as far as possible, the relevant theoretical results available. Unfortunately, theoretical results at present offer only a partial understanding of these algorithms. Nevertheless, a substantial amount of available theory can be of use in guiding the practical use of MCMC.

Of perhaps greatest value are those results related to convergence issues. It is of fundamental importance to verify stationarity of the target distribution (Section 2), and asymptotic convergence of the chain (Section 3); otherwise the algorithm is simply not valid. Even if asymptotic convergence is verified, it is highly desirable to understand the qualitative (Section 4) and quantitative (Section 5) rate of this convergence; theoretical convergence-rate results (when available) are far more convincing than are traditional convergence diagnostics.

In addition to convergence issues, a number of other issues have been investigated through theoretical analysis. These include choice of optimal scaling parameters (Section 6), and sensitivity of the algorithms to certain computer limitations (Section 7).

It is by no means the case that these theoretical results will answer every question about how to implement MCMC. Every applied use of MCMC requires instinct and understanding both about the underlying model and about the Markov chain being used, and theory will never replace that. However, we do feel that theory has a lot of good advice to offer, and it would be a mistake to apply MCMC algorithms without taking this advice into account.

Acknowledgements. We would like to acknowledge the large amount of support, advice, and insight we have gained from numerous discussions and collaborations with many different researchers in the MCMC field. We are grateful to Radford Neal for his comments and corrections. We thank Nancy Reid and Rob Tibshirani for kindly inviting

us to write this paper.

REFERENCES*

P.H. Baxendale (1994), Uniform estimates for geometric ergodicity of recurrent Markov chains. Tech. Rep., Dept. of Mathematics, University of Southern California.

J.R. Baxter and J.S. Rosenthal (1995), Rates of convergence for everywhere-positive Markov chains. Stat. Prob. Lett. **22**, 333-338.

J. Besag and P.J. Green (1993), Spatial statistics and Bayesian computation (with discussions). J. Roy. Stat. Soc. Ser. B **55**, 25-37, 53-102.

L. Breyer and G.O.Roberts (1997) A quasi-ergodic theorem for evanescent processes. Preprint.

S.P. Brooks and G.O. Roberts (1996), Diagnosing Convergence of Markov Chain Monte Carlo Algorithms. Preprint.

K.S. Chan. (1993). Asymptotic behaviour of the Gibbs sampler. J. Amer. Stat. Soc., **88**, 320-326.

K.S. Chan and C.J. Geyer (1994), Discussion to Tierney (1994). Ann. Stat. **22**, 1747-1758.

M.K. Cowles and B.P. Carlin (1995), Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review. J. Amer. Stat. Assoc., to appear.

M.K. Cowles and J.S. Rosenthal (1996), A simulation approach to convergence rates for Markov chain Monte Carlo algorithms. Preprint.

P. Diaconis and D. Stroock (1991), Geometric bounds for eigenvalues of Markov chains. Ann. Appl. Prob. **1**, 36–61.

S. Duane, A.D. Kennedy, B.J. Pendleton, and D. Roweth (1987), Hybrid Monte Carlo. Phys. Lett. B **195**, 216-222.

A.M. Ferrenberg, D.P. Landau, and Y.J. Wong (1992), Monte Carlo simulations: hidden errors from "good" random number generators. Phys. Rev. Lett. **69**, 3382-3384.

* Many of the references cited in this paper but as yet unpublished appear in the MCMC preprint service, at `http://www.stats.bris.ac.uk/MCMC/`

A. Frieze, R. Kannan, and N.G. Polson (1994), Sampling from log-concave distributions. Ann. Appl. Prob. **4**, 812-837. [Correction note, p. 1255.]

A. Frigessi, C.-R. Hwang, L. Younes (1992), Optimal spectral structure of reversible stochastic matrices, Monte Carlo methods and the simulation of Markov random fields. Ann. Appl. Prob. **2**, 610-628.

A. Frigessi, C.-R. Hwang, S.J. Sheu, and P. Di Stefano (1993), Convergence rates of the Gibbs sampler, the Metropolis algorithm, and other single-site updating dynamics. J. Roy. Stat. Soc. Ser. B **55**, 205–220.

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

A. Gelman, G.O. Roberts, and W.R. Gilks (1996), Efficient Metropolis jumping rules. In Bayesian Statistics V, 599-608, Clarendon press, Oxford.

A. Gelman and D.B. Rubin (1992), Inference from iterative simulation using multiple sequences. Stat. Sci., Vol. **7**, No. **4**, 457-472.

S. Geman and D. Geman (1984), Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. IEEE Trans. on pattern analysis and machine intelligence **6**, 721-741.

C. Geyer (1992), Practical Markov chain Monte Carlo. Stat. Sci., Vol. **7**, No. **4**, 473-483.

W. R. Gilks G. O. Roberts and E. George (1994). Adaptive direction sampling, The Statistician, **43**, 179-190.

W.R. Gilks, S. Richardson, and D.J. Spiegelhalter, ed. (1996), Markov chain Monte Carlo in practice. Chapman and Hall, London.

W.R. Gilks, G.O. Roberts, and S.K. Sahu (1996), Adaptive Markov Chain Monte Carlo. Preprint.

U. Grenander and M.I. Miller (1994), Representations of knowledge in complex systems (with discussion). J. Roy. Stat. Soc. B **56**, 549-604.

J.M. Hammersley and D.C. Handscomb (1964), Monte Carlo methods. John Wiley, New York.

W.K. Hastings (1970), Monte Carlo sampling methods using Markov chains and their

applications. Biometrika **57**, 97-109.

L. Holden (1996), Geometric convergence of the Metropolis-Hastings simulation algorithm. Preprint, Norwegian Computing Center, University of Oslo.

S. Ingrassia (1994), On the rate of convergence of the Metropolis algorithm and Gibbs sampler by geometric bounds. Ann. Appl. Prob. **4**, 347–389.

M. Jerrum and A. Sinclair (1989), Approximating the permanent. SIAM J. Comput. **18**, 1149-1178.

A.D. Kennedy and B. Pendleton (1991), Acceptances and autocorrelations in hybrid Monte Carlo. Nuclear Phys. B (Proc. Suppl.) **20**, 118-121.

J.S. Liu (1996), Metropolized independent sampling with comparisons to rejection sampling and importance sampling. Stat. and Comp., to appear.

J.S. Liu, W. Wong, and A. Kong (1994), Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. Biometrika **81**, 27-40.

J.S. Liu, W. Wong, and A. Kong (1995), Covariance structure and convergence rate of the Gibbs sampler with various scans. J. Royal Stat. Sci. Ser. B **57**, 157-169.

R.B. Lund, S.P. Meyn, and R.L. Tweedie (1996), Computable exponential convergence rates for stochastically ordered Markov processes. Ann. Appl. Prob. **6**, 218-237.

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

P. Matthews (1993), A slowly mixing Markov chain with implications for Gibbs sampling. Stat. Prob. Lett. **17**, 231-236.

K.L. Mengersen, G.O. Roberts, D.J. Scott, and R.L. Tweedie (1996), Geometric convergence and Markov chain Monte Carlo. Unpublished manuscript, Queensland University of Technology.

K.L. Mengersen and R.L. Tweedie (1996), Rates of convergence of the Hastings and Metropolis algorithms. Ann. Stat. **24**, 101-121.

N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller (1953), Equations of state calculations by fast computing machines. J. Chem. Phys. **21**, 1087-1091.

S.P. Meyn and R.L. Tweedie (1993), Markov chains and stochastic stability. Springer-

Verlag, London.

S.P. Meyn and R.L. Tweedie (1994), Computable bounds for convergence rates of Markov chains. Ann. Appl. Prob. **4**, 981-1011.

R.M. Neal (1993), Probabilistic inference using Markov chain Monte Carlo methods. Tech. Rep. CRG-TR-93-1, Dept. of Computer Science, University of Toronto.

R.M. Neal (1994), An improved acceptance procedure for the hybrid Monte Carlo algorithm. J. Comp. Phys. **111**, 194-203.

E. Nummelin (1984), General irreducible Markov chains and non-negative operators. Cambridge University Press.

E. Nummelin and R.L. Tweedie (1978), Geometric ergodicity and R-positivity for general Markov chains. Ann. Prob. **6**, 404-420.

D.B. Phillips and A.F.M. Smith (1996), Bayesian model comparison via jump diffusions. In Gilks, Richardson, and Spiegelhalter (1996).

N.G. Polson (1991), Unpublished lecture.

N.G. Polson (1996), Convergence of Markov chain Monte Carlo algorithms. In Bayesian Statistics V, 599-608, Clarendon press, Oxford.

A.E. Raftery and S. Lewis (1992), How many iterations in the Gibbs sampler? In J.M. Bernado, A.F.M. Smith, A.P. Dawid, and J.O. Bergers (eds.), Proceedings of the Fourth Valencia International Meeting on Bayesian Statistics. Oxford University Press.

B.D. Ripley (1987), Stochastic simulation. Wiley, New York.

G. O Roberts (1996). A note on acceptance rate criteria for CLTs for Hastings-Metropolis algorithms. Preprint.

G. O Roberts (1997). Optimal Metropolis algorithms for product measures on the vertices of a hypercube. Preprint.

G.O. Roberts, A. Gelman, and W.R. Gilks (1997), Weak convergence and optimal scaling of random walk Metropolis algorithms. Ann. Appl. Prob., 7, 110-120.

G.O. Roberts and N.G. Polson (1994), On the geometric convergence of the Gibbs sampler. J. Royal Stat. Soc. Ser. B, 377-384.

G.O. Roberts and J.S. Rosenthal (1995), Optimal scaling of discrete approximations to Langevin diffusions. J. Roy. Stat. Soc. Ser. B, to appear.

G.O. Roberts and J.S. Rosenthal (1996), Shift-coupling and convergence rates of ergodic averages. Comm. in Stat. – Stoch. Models **13**, 1, 147–166.

G.O. Roberts and J.S. Rosenthal (1997a), Geometric ergodicity and hybrid Markov chains, *Electronic Communications in Probability* **2**, paper 2.

G.O. Roberts and J.S. Rosenthal (1997b), Two convergence properties of hybrid samplers. Preprint.

G.O. Roberts, J.S. Rosenthal, and P.O. Schwartz (1995), "Convergence properties of perturbed Markov chains". *Journal of Applied Probability*, to appear.

G.O. Roberts and S.K. Sahu (1996), Rate of Convergence of the Gibbs Sampler by Gaussian Approximation. Preprint.

G.O. Roberts and S.K. Sahu (1997) Updating schemes, Correlation Structure, Blocking and Parameterisation for the Gibbs Sampler (with S. Sahu), J. Roy. Statis. Soc., B, **59**, 291–317.

G.O. Roberts and A. F. M. Smith (1994). Simple conditions for the convergence of the Gibbs sampler and Metropolis–Hastings algorithms, Stoch. Proc. Appl., **49**, 207–216.

G.O. Roberts and R.L. Tweedie (1996), Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. Biometrika, **83**, 96-110.

G.O. Roberts and R.L. Tweedie (1996), Exponential Convergence of Langevin Diffusions and their discrete approximations, Bernoulli Vol. **2**, No. **4**.

J.S. Rosenthal (1995a), Rates of convergence for Gibbs sampler for variance components models. Ann. Stat. **23**, 740-761.

J.S. Rosenthal (1995b), Minorization conditions and convergence rates for Markov chain Monte Carlo. J. Amer. Stat. Assoc. **90**, 558-566. [Correction note, p. 1136.]

J.S. Rosenthal (1996a), Convergence of Gibbs sampler for a model related to James-Stein estimators. Stat. and Comput. **6**, 269-275.

J.S. Rosenthal (1996b), Markov chain convergence: from finite to infinite. Stoch. Proc. Appl. **62**, 55-72.

P.J. Rossky, J.D. Doll, and H.L. Friedman (1978), Brownian dynamics as smart Monte Carlo simulation. J. Chem. Phys. **69**, 4628-4633.

M.J. Schervish and B.P. Carlin (1992), On the convergence of successive substitution

sampling, J. Comp. Graph. Stat. **1**, 111–127.

A. Sinclair (1992), Improved bounds for mixing rates of Markov chains and multicommodity flow. Combinatorics, Prob., Comput. **1**, 351–370.

A. Sinclair (1993), Algorithms for random generation and counting : a Markov chain approach. Birkhauser, Boston.

A.F.M. Smith and G.O. Roberts (1993), Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods (with discussion). J. Roy. Stat. Soc. Ser. B **55**, 3-24.

R.L. Smith and L. Tierney (1996), Exact transition probabilities for the independence Metroplis sampler. Preprint, Dept. of Statistics, University of North Carolina at Chapel Hill.

M.A. Tanner and W.H. Wong (1987), The calculation of posterior distributions by data augmentation (with discussion). J. Amer. Stat. Assoc. **82**, 528-550.

L. Tierney (1994), Markov chains for exploring posterior distributions (with discussion). Ann. Stat. **22**, 1701-1762.

R.L. Tweedie (1996), Truncation approximations of invariant measures for Markov chains. Preprint, Colorado State University.

I. Vattulainen, T. Ala-Nissila, and K. Kankaala (1994), Physical tests for random numbers in simulations. Tech. Rep., Research Institute for Theoretical Physics, University of Helsinki, Finland.