Adaptive Markov Chain Monte Carlo through Regeneration

Walter R. Gilks; Gareth O. Roberts; Sujit K. Sahu

Journal of the American Statistical Association, Vol. 93, No. 443 (Sep., 1998), 1045-1054.

Stable URL:
http://links.jstor.org/sici?sici=0162-1459%28199809%2993%3A443%3C1045%3AAMCMCT%3E2.0.CO%3B2-X

Journal of the American Statistical Association is currently published by American Statistical Association.

Adaptive Markov Chain Monte Carlo Through Regeneration

Walter R. GILKS, Gareth O. ROBERTS, and Sujit K. SAHU

Markov chain Monte Carlo (MCMC) is used for evaluating expectations of functions of interest under a target distribution π. This is done by calculating averages over the sample path of a Markov chain having π as its stationary distribution. For computational efficiency, the Markov chain should be rapidly mixing. This sometimes can be achieved only by careful design of the transition kernel of the chain, on the basis of a detailed preliminary exploratory analysis of π. An alternative approach might be to allow the transition kernel to adapt whenever new features of π are encountered during the MCMC run. However, if such adaptation occurs infinitely often, then the stationary distribution of the chain may be disturbed. We describe a framework, based on the concept of Markov chain regeneration, which allows adaptation to occur infinitely often but does not disturb the stationary distribution of the chain or the consistency of sample path averages.

KEY WORDS: Adaptive method; Bayesian inference; Gibbs sampling; Markov chain Monte Carlo; Metropolis–Hastings algorithm; Mixing rate; Regeneration; Splitting.

1. INTRODUCTION

Markov chain Monte Carlo (MCMC) has had a profound influence on Bayesian statistical analysis, enabling inference to be made with data and models previously considered intractable. Non-Bayesian applications of MCMC have also been developed. In all applications, the aim is to estimate the expectation $\mathbb{E}_\pi[g]$ of functions of interest $g(\mathbf{x})$ under a target distribution π. A Markov chain $\{\mathbf{X}_1, \ldots, \mathbf{X}_N\}$ with stationary distribution π is run, and $\mathbb{E}_\pi[g]$ is estimated by the delayed average,

$$\bar{g}_N = \frac{1}{N} \sum_{n=M+1}^{M+N} g(\mathbf{X}_n), \tag{1}$$

where the first M iterations (the "burn-in") are discarded. If the chain is irreducible, then $\bar{g}_N$ is a consistent estimator of $\mathbb{E}_\pi[g]$ (see, e.g., Tierney 1994).

One practical difficulty in using $\bar{g}_N$ as an estimator of $\mathbb{E}_\pi[g]$ comes from the dependence within the sequence $\mathbf{X}_n$. Approximating the series $g(\mathbf{X}_n)$ as a first order autoregressive process with autocorrelation ρ, the variance of $\bar{g}_N$ can be written as

$$\operatorname{var}(\bar{g}_N) = \frac{\sigma^2}{N} \frac{1+\rho}{1-\rho},$$

where σ^2 is the variance of $g(\mathbf{X})$ under π. Hence high positive autocorrelations will substantially reduce efficiency, and we may be forced to use a large number of iterations N to achieve adequate accuracy in $\bar{g}_N$. High autocorrelations result from a slow mixing Markov chain, that is, $\Pr[\mathbf{X}_n \in B \mid \mathbf{X}_0 \in A]$ converges slowly to $\pi(B)$.

In many applications, mixing is rapid for untuned MCMC methods such as the Gibbs sampler (Gelfand and Smith 1990, Geman and Geman 1984). This is demonstrated by the huge range of problems routinely handled by the Gibbs sampling software BUGS (Spiegelhalter, Thomas, and Best 1996). However, in some applications of Gibbs sampling, model re-parameterization may be necessary to achieve rapid mixing (see for example, Gelfand, Sahu, and Carlin 1995). Usually, analytic intractability of the target distribution prevents determination of the best parameterization in advance. Consequently it may be necessary to conduct preliminary experiments to determine an acceptable parameterization. Similarly, in the more general Metropolis–Hastings (M–H) framework for MCMC (see Sec. 3), preliminary experiments are often required to determine efficient proposal distributions.

To avoid preliminary exploratory work, it is tempting to use all or part of the history of the Markov chain to dynamically construct improved parameterizations or proposal distributions. For example, if the chain wanders into the vicinity of a mode of π that it has not previously encountered, then an additional proposal distribution might be constructed to generate candidate points near this mode. However, allowing such adaptation to take place infinitely often will in general disturb the stationary distribution of the chain and the consistency of sample path averages (1). The problem is that the process is no longer Markov, because $\mathbb{P}[\mathbf{X}_n \mid \mathbf{X}_0, \mathbf{X}_1, \ldots, \mathbf{X}_{n-1}] \neq \mathbb{P}[\mathbf{X}_n \mid \mathbf{X}_{n-1}]$, so the consistency of (1) is no longer assured. Gelfand and Sahu (1994) provided an example where infinite adaptation disturbs the ergodicity of the chain, despite each participating transition kernel having the same stationary distribution. There are several remedies to this problem. The most obvious would be to stop adapting after a prechosen iteration or after a fixed number of adaptations, and commence the burn-in after the last adaptation (Gelfand and Sahu 1994). However, such strategies cannot guarantee that all the important features of π will be discovered during the adaptation phase. We call this the pilot adaptation scheme (PAS).

Gilks and Roberts (1996) and Gilks, Roberts, and George (1994) proposed methods of *adaptive direction sampling*

Walter R. Gilks is Senior Scientist, Medical Research Council, Biostatistics Unit, Cambridge, CB2 2SR, U.K. (E-mail: wally.gilks@mrc-bsu.cam.ac.uk). Gareth O. Roberts is Lecturer, Statistical Laboratory, University of Cambridge, Cambridge, CB2 1SB, U.K. (E-mail: g.o.roberts@statslab.cam.ac.uk). Sujit K. Sahu is Lecturer, School of Mathematics, University of Wales, Cardiff, CF2 4YH, U.K. (E-mail: smasks@d168.math.cf.ac.uk).

Journal of the American Statistical Association
September 1998, Vol. 93, No. 443, Theory and Methods

(ADS) and *adaptive Metropolis sampling* (AMS), which aim to gather information about π as the chain proceeds. The adaptation involves a replicated state space, but in high-dimensional problems, the replication may be prohibitively large.

Here we propose a new strategy for adaptation, based on the concept of Markov chain *regeneration.* For a discrete state-space Markov chain, regeneration times are the iterations at which the chain revisits a nominated state. For continuous state-space Markov chains, a technique for regeneration due to Nummelin (1984) can be used. This technique has been applied to MCMC samplers by Mykland, Tierney, and Yu (1995) and Robert (1995); see also Section 2.1. Our adaptive strategy is that at each regeneration time the transition kernel of the chain is modified, based on the history of the chain. The modified transition kernel is constructed such that π is retained as its stationary distribution. We show that such adaptation can continue indefinitely without affecting the consistency of (1). The method allows an increasing amount of information from the chain's history to be used in constructing proposal distributions, unlike PAS, ADS, and AMS. The extra computational burden required for adaptation will in general be small.

Figure 1 illustrates the process of adaptation through regeneration. At each regeneration time, shown in the figure by vertical dotted lines, the variance σ_i^2 of the proposal distribution is adapted to improve mixing, based on the chain's output to that time. The mixing rate of the chain clearly increases with each adaptation. This example is described in more detail in Section 4.1.1.

Thus adaptation can occur at each regeneration time and can be based on the entire history of the chain up to that time. This gives immense freedom in modifying the transition mechanism for the chain. However, the practicality of our methodology is currently limited by the available technology for delivering regeneration times: as we show, this is often difficult in high-dimensional problems. Nevertheless, we regard our framework for adaptation as an important theoretical advance, motivating further research into techniques for regenerative simulation.

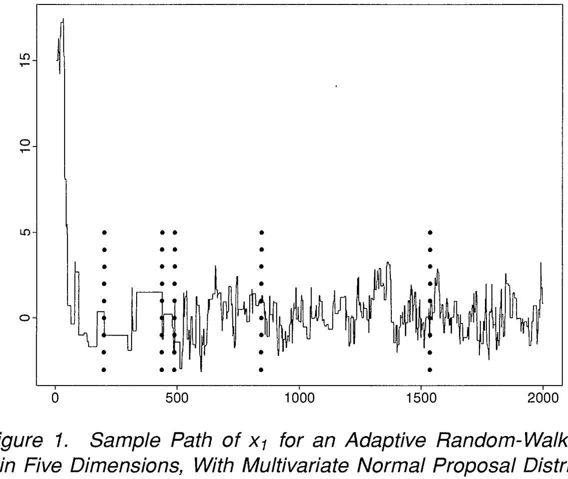

Figure 1. Sample Path of x_1 for an Adaptive Random-Walk Sampler in Five Dimensions, With Multivariate Normal Proposal Distribution $N_5(\mathbf{x}, \sigma_i^2\mathbf{I}_5)$ and Stationary Distribution $N_5(\mathbf{0}, \mathbf{I}_5)$. Regeneration times are indicated by vertical dotted lines.

The rest of the article is organized as follows. In Section 2 we introduce the concept of regeneration and adaptation at regeneration, and provide theoretical support. In Section 3 we review the splitting techniques required for adaptation. We present four illustrations of adaptive MCMC in Section 4, and provide some of the proofs from Sections 2 and 3 in the Appendix.

2. REGENERATION: A FRAMEWORK FOR ADAPTATION

2.1 Regeneration and Splitting

Let $\{\mathbf{X}_n\colon n = 0, 1, \ldots\}$ be an irreducible Markov chain on a state space $(E, \mathcal{E})$ with transition kernel $P = P(\mathbf{x}, d\mathbf{y})$ and invariant distribution π. Suppose that we can find a set $A \in \mathcal{E}$ with $\pi(A) > 0$ such that $\mathbf{X}_{n+1}, \mathbf{X}_{n+2}, \ldots$ is conditionally independent of $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ given $\mathbf{X}_n \in A$. Then A is called a *proper atom* for the Markov chain, and whenever the chain enters A, the chain is said to *regenerate.* Regeneration times divide the chain into sections, called *tours*, and the sample paths of tours are independent. Such independence can be exploited for both theoretical and practical purposes, as we demonstrate. For a discrete state-space Markov chain, any individual state can be chosen to represent A. However, in more general state spaces, proper atoms will not usually exist. Nevertheless, regenerations might still be defined using a technique due to Nummelin (1984), which we now describe.

Suppose that it is possible to find a function $s^*(\mathbf{x}) \le 1$ and a probability measure $\nu^*(d\mathbf{y})$ such that $\pi(s^*) = \int s^*(\mathbf{x})\pi(d\mathbf{x}) > 0$ and

$$P(\mathbf{x}, A) \ge s^*(\mathbf{x})\nu^*(A) \qquad (2)$$

for all $\mathbf{x} \in E$ and all $A \in \mathcal{E}$. Then we can write

$$P(\mathbf{x}, d\mathbf{y}) = s^*(\mathbf{x})\nu^*(d\mathbf{y}) + (1 - s^*(\mathbf{x}))\Lambda(\mathbf{x}, d\mathbf{y}),$$

where

$$\Lambda(\mathbf{x}, d\mathbf{y}) = \begin{cases} \frac{P(\mathbf{x},d\mathbf{y}) - s^*(\mathbf{x})\nu^*(d\mathbf{y})}{1 - s^*(\mathbf{x})} & \text{if } s^*(\mathbf{x}) < 1 \\ \nu^*(d\mathbf{y}) & \text{if } s^*(\mathbf{x}) = 1. \end{cases} \qquad (3)$$

The pair (s^*, ν^*) is called an *atom* for the transition kernel P.

We can now construct a Markov chain on an augmented state space, as follows. Suppose that the chain is currently at $\mathbf{X}_n$. First, generate a Bernoulli variable S_{n+1} with success probability $s^*(\mathbf{X}_n)$. If $S_{n+1} = 1$, then generate $\mathbf{X}_{n+1}$ according to ν^*; otherwise, generate $\mathbf{X}_{n+1}$ from Λ. Then $(\mathbf{X}_n, S_n)$ forms a Markov chain called the *split* chain, where the marginal sequence $\{\mathbf{X}_n\}$ is a Markov chain with transition kernel P and stationary distribution π.

The essential point in the foregoing construction is that when $S_{n+1} = 1$, the transition mechanism ν^* is independent of the current state $\mathbf{X}_n$. Consequently, the augmented set $E \times \{1\}$ is a proper atom for the Markov chain $(\mathbf{X}_n, S_n)$, and the times at which $S_n = 1$ are regeneration times for the split chain.

2.2 Retrospective Regeneration

The foregoing scheme for identifying regeneration times is prospective; the regeneration indicator S_{n+1} is sampled before sampling $\mathbf{X}_{n+1}$. In practice, however, it is often more convenient and computationally efficient to determine regeneration times retrospectively; that is, to sample S_{n+1} after sampling $\mathbf{X}_{n+1}$. In particular, this method does not require the density ν^* to be normalized. The retrospective method of regeneration is used in all of the examples in Section 4.

We define a nonnormalized atom to be a pair (s, ν) such that for all sets A,

$$P(\mathbf{x}, A) \geq s(\mathbf{x})\nu(A) , \tag{4}$$

where $\int_E \nu(d\mathbf{y})$ need not be unity. Nonnormalized atoms are of course equivalent to normalized atoms by the scaling $\nu^*(d\mathbf{y}) = \nu(d\mathbf{y})/\int_E \nu(d\mathbf{z})$ and $s^*(\mathbf{x}) = \mathbf{s}(\mathbf{x}) \int_\mathbf{E} \nu(\mathbf{dz})$. Assuming that the chain is at stationarity, the regeneration probability is

$$r(s, \nu) = \mathbb{E}_\pi [s(\mathbf{X})] \int_E \nu(d\mathbf{y}). \tag{5}$$

In practice we try to choose s, ν to maximize r, subject to (4).

Nonnormalized atoms can be used to construct a split chain as follows. Suppose that the chain is currently at $(\mathbf{X}_n, S_n)$. First, generate $\mathbf{X}_{n+1}$ according to $P(\mathbf{X}_n, .)$; then sample S_{n+1} from a Bernoulli distribution with retrospective success probability

$$r^A(\mathbf{X}_n, \mathbf{X}_{n+1}) = \frac{s(\mathbf{X}_n)\nu(d\mathbf{X}_{n+1})}{P(\mathbf{X}_n, d\mathbf{X}_{n+1})}. \tag{6}$$

This construction is probabilistically equivalent to that given in Section 2.1. Details for splitting canonical MCMC samplers are given in Section 3.

2.3 Adaptation Through Regeneration

In general, we will not know a good MCMC sampler design at the outset. We propose to modify the sampler on the basis of the past sample path of the chain, as the simulation proceeds. Such an adaptive process can be dangerous; it is no longer Markov, so we lose the support of Markov chain theory, which guarantees ergodicity. Even if the adaptive process exhibits some kind of stationary behavior, its stationary measure may not be the target distribution π. However, by modifying the sampler only at regeneration times, convergence of (1) to $\mathbb{E}_\pi[g]$ is preserved. Informally, because regeneration tours are independent in the unadaptive case, information from one tour can be used freely to construct dynamics for subsequent tours.

Our adaptive process is as follows. Let $T_1, T_2, \dots$ denote the regeneration times of the adaptive chain.

Iteration 1. Let P_1 denote the initial irreducible transition kernel P_1 having stationary distribution π and a normalized atom (s_1^*, ν_1^*). The first iteration is a regeneration; set $S_1 = 1$ and $T_1 = 1$ and sample $\mathbf{X}_1$ from ν_1^*.

Iteration n + 1. Suppose that after n iterations, the chain is at $\mathbf{X}_n$. Suppose that i regenerations have occurred and the current transition kernel is P_i, with a normalized atom (s_i^*, ν_i^*). First, generate a Bernoulli variable S_{n+1} with success probability $s_i^*(\mathbf{X}_n)$.

- If $S_{n+1} = 1$, a regeneration has occurred; set $T_{i+1} = n$. We may now update the Markov chain dynamics to produce a modified irreducible transition kernel P_{i+1} with stationary distribution π and a normalized atom (s_{i+1}^*, ν_{i+1}^*). We may determine $P_{i+1}, s_{i+1}^*, \nu_{i+1}^*$ in almost any way we feel appropriate, using the past sample path $\{\mathbf{X}_1, \dots, \mathbf{X}_n\}$. Complete the iteration by sampling $\mathbf{X}_{n+1}$ from ν_{i+1}^*.
- If $S_{n+1} = 0$, proceed as in the nonadaptive split chain, sampling $\mathbf{X}_{n+1}$ from the probability measure proportional to

$$P_i(\mathbf{X}_n, \cdot) - s_i^*(\mathbf{X}_n)\nu_i^*(\cdot),$$

as in (3). An equivalent construction utilizing nonnormalized atoms is as follows. Suppose that the chain is currently at $\mathbf{X}_n$, with current transition kernel P_i and a nonnormalized atom (s_i, ν_i). First sample $\mathbf{y}$ from $P_i(\mathbf{X}_n, \cdot)$, then sample S_{n+1} from a Bernoulli distribution with retrospective success probability

$$r_i^A(\mathbf{X}_n, \mathbf{y}) = \frac{s_i(\mathbf{X}_n)\nu_i(d\mathbf{y})}{P_i(\mathbf{X}_n, d\mathbf{y})}. \tag{7}$$

- If $S_{n+1} = 1$, a regeneration has occurred; set $T_{i+1} = n$. Discard $\mathbf{y}$, determine new dynamics $P_{i+1}, s_{i+1}, \nu_{i+1}$ using the past sample path, as above, and sample $\mathbf{X}_{n+1}$ from the probability measure proportional to ν_{i+1}.
- If $S_{n+1} = 0$, set $\mathbf{X}_{n+1} = \mathbf{y}$.

Note that both of the foregoing methods involve sampling from the measure ν^*. This is avoidable in the nonadaptive methods for regenerative simulation described by Mykland et al. (1995) and Robert (1995).

We now provide some theoretical support for our method.

2.4 Theoretical Results

For each $i > 1$, (P_i, s_i, ν_i) depends on the history of the process $\{\mathbf{X}_1, \dots, \mathbf{X}_{T_i - 1}\}$ before the ith regeneration. Let $\mathcal{F}_{i-1}$ denote the σ algebra generated by $\{\mathbf{X}_1, \dots, \mathbf{X}_{T_i-1}\}$. Let $N_i = T_{i+1} - T_i$ denote the length of the ith tour.

Suppose that we wish to estimate $\mathbb{E}_\pi[g] = \int g(\mathbf{x})\pi(d\mathbf{x})$ for some function of interest g. Let $G_i = \sum_{n=T_i}^{T_{i+1}-1} g(\mathbf{X}_n)$ and let $Z_i = G_i - \mathbb{E}_\pi[g]N_i$. From renewal theory, we have, $\mathbb{E}[Z_i \mid \mathcal{F}_{i-1}] = 0$. Therefore, a natural estimator of $\mathbb{E}_\pi[g]$ is

$$R_n = \frac{\sum_{i=1}^n G_i}{\sum_{i=1}^n N_i}. \tag{8}$$

We show that R_n is mean squared error (MSE) consistent for $\mathbb{E}_\pi[g]$. In general, the purpose of adapting (P_i, s_i, ν_i) to $\mathcal{F}_{i-1}$ will be to reduce tour lengths and improve the mixing of P_i. However, we do not assume that the adaptation will achieve these aims. We only assume that by controlling the variance of Z_i, the adaptation will not make matters arbitrarily worse (see condition a of Theorem 2). Robert (1995) gave some results on consistency and also a central limit theorem for the nonadaptive samplers.

Throughout, we denote weak convergence by $\Rightarrow$ and convergence in probability by $\overset{P}{\rightarrow}$.

Theorem 1. Assume that there exists a constant $b_1 < \infty$ such that $\mathbb{E}[Z_i^2] < b_1^2$, for all i. Then R_n is MSE consistent for $\mathbb{E}_\pi[g]$.

Proof. See the Appendix.

To establish a central limit theorem for R_n, we make assumptions concerning the limiting behavior of the adaptation.

Theorem 2. Assume the following:

a. There exists a constant $b_2 < \infty$ such that $\mathbb{E}\left[Z_i^{2+\varepsilon} \mid \mathcal{F}_{i-1}\right] < b_2^{2+\varepsilon}$, for some $\varepsilon > 0$ and for all i.

b. $1/n \sum_{i=1}^n \mathbb{E}\left[Z_i^2 \mid \mathcal{F}_{i-1}\right] \overset{P}{\rightarrow} V^2$, where V is an $\mathcal{F}_\infty$ random variable with $\Pr[V > 0] = 1$.

c. $1/n \sum_{i=1}^n N_i \overset{P}{\rightarrow} C$, where C is an $\mathcal{F}_\infty$ random variable.

Then the following holds:

$$\frac{\sqrt{n}C}{V}\left(R_n - \mathbb{E}_\pi[g]\right) \Rightarrow \mathrm{N}(0, 1). \tag{9}$$

Proof. See the Appendix.

It is important that we allow V and C to be random in the foregoing. Effectively this allows the adaptation to achieve better results on some occasions that on others, so that perhaps after a good start, the algorithm would settle down to a particularly efficient limiting algorithm. Theorem 2 does not depend on achieving a good start, and allows the ultimate efficiency of the algorithm to depend on early regeneration tours.

As an example, consider a Gibbs sampler with varying parameterization, adapting to reduce correlations between components but in a way that the level of adaptation diminishes as the simulation proceeds, so that a limiting but random parameterization exists. The asymptotic efficiency of the algorithm thus is random, but Theorem 2 still holds. Of course, Theorem 2 can still apply to adaptation that continues indefinitely.

We now show that

$$\widehat{\mathrm{MSE}}(R_n) \equiv \frac{\sum Z_i^2}{\left(\sum N_i\right)^2}$$

is a consistent estimator of the conditional asymptotic variance, $V^2/(nC^2)$.

Theorem 3. Assume that there exists a constant $b_3 < \infty$ such that $\mathbb{E}\left[Z_i^4 \mid \mathcal{F}_{i-1}\right] < b_3^4$, for all i. Further, with assumptions b and c of Theorem 2, as $n \rightarrow \infty$, the following holds:

$$n\,\widehat{\mathrm{MSE}}(R_n) \overset{P}{\rightarrow} \frac{V^2}{C^2}.$$

Proof. See the Appendix.

The regularity conditions in Theorems 1–3 will be hard to check in most applications, and we have not attempted to do this in the examples that follow. However, loosely interpreted, these conditions require only that successive adaptations do not make the mixing of the sampler arbitrarily worse. The aim of the adaptation is of course to improve mixing, so provided that this is done sensibly, the regularity conditions generally will be satisfied. It is important, however, to guard against overzealous adaptation. For example, if a short run of the sampler suggests a single mode in π, and if all proposal distributions are then focused on this mode, then other major modes may go undetected.

3. SPLITTING MCMC SAMPLERS

The techniques of Nummelin (1984) can be used to split the MCMC samplers (Mykland et al. 1995). This requires calculating the transition kernel of the underlying Markov chain. The Gibbs sampler transition kernel is analytically intractable except for conjugate problems. However, clever updating schemes for the Gibbs sampler depending on the particular application can be used so that the splitting mechanism does not require the calculation of the full transition kernel. (See Mykland et al. 1995 for some illustrations.) In the rest of this section we give details for the M-H algorithm.

The most general form of the M-H algorithm is as follows. Let Q be a Markov transition kernel. We assume that $Q(\mathbf{x}, d\mathbf{y})$ has a density $q(\mathbf{x}, \mathbf{y})$ with respect to a measure $\mu(d\mathbf{y})$. Suppose that the chain is currently at a point $\mathbf{X}_n = \mathbf{x}$. A candidate point $\mathbf{y}$ is sampled from the distribution $Q(\mathbf{x}, \cdot)$, which is accepted with probability

$$\alpha(\mathbf{x}, \mathbf{y}) = \min\left\{\frac{\pi(\mathbf{y})q(\mathbf{y}, \mathbf{x})}{\pi(\mathbf{x})q(\mathbf{x}, \mathbf{y})}, 1\right\}; \tag{10}$$

that is, we set $\mathbf{X}_{n+1} = \mathbf{y}$. Otherwise, $\mathbf{y}$ is discarded and we set $\mathbf{X}_{n+1} = \mathbf{x}$.

Different choices of the proposal kernel $Q(\mathbf{x}, d\mathbf{y})$ give different versions of the algorithm. The choice $Q(\mathbf{x}, d\mathbf{y}) = f(\mathbf{y})\mu(d\mathbf{y})$ defines the *independence sampler.* The Metropolis, Rosenbluth, Rosenbluth, Teller, and Teller (1953) version of the algorithm is given by choosing q to be symmetric; that is, $q(\mathbf{x}, \mathbf{y}) = q(\mathbf{y}, \mathbf{x})$. A special case of this is the *random-walk* Metropolis algorithm, for which $q(\mathbf{x}, \mathbf{y}) = q(|\mathbf{x} - \mathbf{y}|)$. We now present atoms for the independence and random-walk samplers, as proposed by Mykland et al. (1995). For simplicity of exposition, we present only the simplest schemes; Mykland et al. (1995) provided a more general treatment.

3.1 The Independence Sampler

To be effective, the independence proposal distribution, f, should be similar to π but with heavier tails; a poor choice for f can produce a nongeometrically ergodic chain (Roberts and Tweedie 1996). In practice, a suitable f is not usually known. This suggests using the adaptive strategy outlined in Section 2.3 to adapt on f.

The independence sampler is the easiest MCMC sampler to split. For any $c > 0$, set

$$s(\mathbf{x}) = \min\left\{\frac{c}{w(\mathbf{x})}, 1\right\} \tag{11a}$$

and

$$\nu(d\mathbf{y}) = f(\mathbf{y}) \min\left\{\frac{w(\mathbf{y})}{c}, 1\right\} \mu(d\mathbf{y}) \tag{11b}$$

in (4), where $w(\mathbf{x}) = \pi(\mathbf{x})/f(\mathbf{x})$. Mykland et al. suggested setting c to a central value of $w(\mathbf{x})$. However, it may be difficult to determine such a value for c in advance. Our adaptive framework can be used to adapt splitting parameters such as c at each regeneration time, using past values of $w(\mathbf{x})$, as described.

We can implement the adaptative strategy for the independence sampler with the following pseudocode, which describes the calculations for iteration $n+1$, assuming Lebesgue measure μ. This algorithm uses the retrospective method of regeneration, as described in Section 2.2. Note that the denominator of the retrospective regeneration probability $r^A(\mathbf{x}, \mathbf{y})$ for the independence sampler is $f(\mathbf{y})\min\{[\pi(\mathbf{y})]/[\pi(\mathbf{x})], 1\}\, d\mathbf{y}$. Substituting this, together with equation (10), into equation (6) gives the formulas for r^A used in the algorithm. We begin the iteration with $\mathbf{X}_n = \mathbf{x}$.

```
Sample $y \sim f(y)$    // generates a candidate point
// Perform Metropolis-Hastings acceptance test:
Sample $U_1 \sim U(0,1)$    // generates a Uniform r.v.
If $U_1 < \min\left(1, \frac{w(y)}{w(x)}\right)$ {    // (Provisionally) accept candidate, $y$:
  // The retrospective regeneration probability (7) reduces to:
  If $w(\mathbf{x}) > c$ and $w(\mathbf{y}) > c$, set $r^A = \max\{c/w(\mathbf{x}), c/w(\mathbf{y})\}$
  Else if $w(\mathbf{x}) < c$ and $w(\mathbf{y}) < c$, set $r^A = \max\{w(\mathbf{x})/c, w(\mathbf{y})/c\}$
  Else set $r^A = 1$
  // Perform conditional regeneration test:
  Sample $U_2 \sim U(0,1)$
  If $U_2 < r^A$ {    // Regeneration has occurred. Insert code here
    // to adapt on $f(\cdot)$, eg heavier tails. Also adapt on $c$:
    Set $c = w(\tilde{x})/2$    // $\tilde{x}$ is the mode of $\pi(x)$ discovered so far
    // Discard $y$ and resample from new $\nu$ using rejection sampling:
    Repeat {  Sample $y \sim f(y)$;    Sample $U_3 \sim U(0,1)$;
    } Until $U_3 < \min\{w(y)/c, 1\}$
  }
  Set $\mathbf{X}_{n+1} = y$    // finally accepts current candidate $y$
} Else {    // Reject candidate $y$:

  Set $\mathbf{X}_{n+1} = \mathbf{X}_n$
}
```

The foregoing provides a good scheme for identifying regeneration times. However, independence samplers are not usually used on their own, because of their potential for nongeometric ergodicity; see Section 3.3.

3.2 The Random-Walk Sampler

To split the random-walk sampler, we first need to find an atom (s_q, ν_q) for the transition kernel Q. One approach is to choose a *distinguished* point $\tilde{\mathbf{x}} \in E$ and a set $D \in \mathcal{E}$, and define

$$s_q(\mathbf{x}) = \inf\left\{\frac{q(\mathbf{x}, \mathbf{y})}{q(\tilde{\mathbf{x}}, \mathbf{y})} : \mathbf{y} \in D\right\} \tag{12a}$$

and

$$\nu_q(d\mathbf{y}) = q(\tilde{\mathbf{x}}, \mathbf{y}) 1(\mathbf{y} \in D) \mu(d\mathbf{y}). \tag{12b}$$

Then the pair

$$s(\mathbf{x}) = s_q(\mathbf{x}) \min\left\{\frac{\pi(\tilde{\mathbf{x}})}{\pi(\mathbf{x})}, 1\right\}$$

and

$$\nu(d\mathbf{y}) = \nu_q(d\mathbf{y}) \min\left\{\frac{\pi(\mathbf{y})}{\pi(\tilde{\mathbf{x}})}, 1\right\} \tag{13}$$

provides a splitting for the Markov chain. In many applications Q is Gaussian, for which calculation of (s_q, ν_q) is easy. The choices of the implicit parameter $\tilde{\mathbf{x}}$ and the set D need experimentation. We can set $\tilde{\mathbf{x}}$ at the mode of the distribution discovered so far and update it at regeneration points. Similarly, we can experiment with the set D and update it adaptively. However, the practicality of this splitting is rather limited, as we illustrate with the following result.

Theorem 4. Let $\pi(\mathbf{x}) = \mathrm{N}_m(\mathbf{0}, \mathbf{I}_m)$, where $\mathbf{x}$ is $m \times 1$, N_m denotes a multivariate normal distribution, and I_m is the identity matrix of order m. Let $\tilde{\mathbf{x}} = \mathbf{0}$ and $D = \{\mathbf{y}: \mathbf{y}'\mathbf{y} \leq d\}$, where $d > 0$ is a scalar. Then, for the Metropolis algorithm with proposal distribution $q(\mathbf{x}, \mathbf{y}) = \mathrm{N}_m(\mathbf{x}, k\mathbf{I}_m)$ and splitting as defined by (13),

$$\max_{d>0} r(s, \nu) \to 0 \text{ exponentially as } m \to \infty,$$

where k is set at its asymptotic optimal value $2.38^2/m$ (Gelman et al. 1996) and $r(s, \nu)$ is the regeneration rate as given in (5).

Proof. See the Appendix.

Hence, although the method will work for low-dimensional problems (see Sec. 4), for high-dimensional problems this method will fail to identify regeneration times. This problem is common in regenerative simulation by splitting. Ideally, we would like to perform splitting based on the n-step transition kernel of the Markov chain, where we allow n to vary (and probably increase) with dimension. Unfortunately, it is generally analytically intractable to perform n-step splitting for $n > 1$ (although see Cowles and Rosenthal 1996 for a numerical alternative).

3.3 Hybrid Samplers

Several proposal distributions can be used in a hybrid sampler. At each iteration of the Markov chain, one of these proposal distributions can be chosen according to some random or systematic scheme (Tierney 1994). If regenerations are easily achieved for the independence sampler but difficult to achieve for other proposal distributions, then we can adapt any or all of the proposal distributions whenever a regeneration is obtained on an independence-sampler step.

4. EXAMPLES

4.1 Random-Walk Metropolis Algorithms

The regenerative framework for adaptation provides a method for tuning many designs of the MCMC sampler; for example, updating schemes, blocking, and parameterization on-line. In this section we concentrate on adapting the proposal distributions of M-H algorithms. The optimal proposal distribution for any given target distribution is generally unknown, and ad-hoc tuning methods (based mostly on PAS) are often performed. The examples here demonstrate how such ad hoc methods can be replaced by more attractive on-line adaptation.

For the random-walk algorithm, often the proposal density is of the form $q(\mathbf{x},\mathbf{y}) = \mathrm{N}_m(\mathbf{x},\sigma^2\mathbf{I}_m)$, where σ is a constant. Clearly, some values of σ will give rise to better mixing than others. Values of σ that are too small will result in most candidates $\mathbf{y}$ being accepted, but the steps $|\mathbf{X}_{n+1}-\mathbf{X}_n|$ will be small. Values of σ which are too large will result in large proposed moves $|\mathbf{y}-\mathbf{X}_n|$, most of which will be rejected.

Gelman, Roberts, and Gilks (1996) considered the problem of choosing σ when the target distribution π has the exchangeable form $\prod_{i=1}^m \pi(x_i)$, where x_i denotes the ith element of $\mathbf{x}$. They show that for large m, the variance of $\bar{g}_N$ in (1) is minimized by choosing σ such that 23.4% of candidates are accepted overall. In general, there is no theoretical optimal value of this acceptance rate. However, theoretical results and empirical evidence suggest that for most cases overall, 15%–50% of the proposed moves should be accepted for optimal performance, (see, e.g., Roberts 1996). Besag, Green, Higdon, and Mengersen (1995, sec. 2.3.3) also discussed such matters.

Thus the optimal scaling σ might be determined empirically, through monitoring candidate acceptance rates in an adaptive MCMC run. If the empirical acceptance rate during the ith tour is A_i, then the scaling σ_{i+1} for tour $i+1$ could be set as

$$\log \sigma_{i+1} = \log \sigma_i + (\mathrm{logit}(A_i) - \mathrm{logit}(a))/m, \qquad (14)$$

where a is the target acceptance rate. Thus if $A_i > a$, then the scaling σ will be increased, which will reduce the acceptance rate during tour $i+1$. Similarly, if $A_i < a$, then the acceptance rate will be increased. Under (14), σ_i will not converge to the optimum, but the theory of Section 2.4 does not require this. Instead, σ_i will tend to oscillate around the optimal scaling. Convergence of σ_i to its optimum could be easily achieved by replacing m in (14) by mi^β for some small $\beta > 0$. Other updating equations can also be considered. However, (14) works well in the examples here.

4.1.1 Example 1. We consider a five-dimensional standard normal target distribution for the Metropolis algorithm. The optimal acceptance rate is .275, with proposal densities of the form $q(\mathbf{x},\mathbf{y}) = \mathrm{N}_5(\mathbf{x},\sigma^2\mathbf{I}_5)$ with the the optimal value of σ as 1.10 (Gelman et al. 1996).

To implement adaptation, we use the regenerative scheme described in Section 3.2, with $\tilde{\mathbf{x}} = \mathbf{0}$ and $D = \{\mathbf{x}\colon \mathbf{x}'\mathbf{x} < d\}$, setting $d = 16$. Note that in theory, any positive value of d will work. It is tuned to produce an acceptable number of regenerations and can be updated during the adaptation phase. However, in this example we keep it fixed all of the time. We set the initial scaling at $\sigma_1 = 10$, and at each regeneration we updated σ_i according to (14), with $a = .275$.

The adaptive chain began with a very low acceptance rate. The first adaptation resulted in a much smaller value of σ, which gave a better but still too low acceptance rate of about .12. By the fifth adaptation, acceptance rates were fairly close to optimal. This demonstrates that even with a very inefficient starting proposal variance, we can adapt to the optimal value quite easily, although of course a more judicious choice of σ_1 would have obviated the need for adaptation in this case. Figure 1 shows the sample path from the adaptive chain and clearly shows the acceleration in mixing with the first few adaptations.

4.1.2 Example 2. We consider a dataset given by Bates and Watts (1988, p. 307), modeled by Newton and Raftery (1994) as follows. Response y_i is modeled as

$$y_i = \beta_1 + \frac{\beta_2}{1+\exp\{-\beta_4(x_i-\beta_3)\}} + \varepsilon_i, \qquad i = 1,2,\ldots,n,$$

where ε_i is assumed to be normally distributed with mean 0 and variance σ^2. The prior for σ^2 is taken as $\pi(\sigma^2) \propto \sigma^{-2}$ with a design invariant prior $\pi(\beta) \propto |\mathbf{V}^T\mathbf{V}|^{1/2}$ for β, where $\mathbf{V}$ is an $n \times 4$ matrix with elements $[\partial E(y_i|\beta)]/\partial\beta_j, i = 1,\ldots,n; j = 1,\ldots,4$.

For this model we first integrate out σ^2 analytically. Hence the target posterior distribution is four dimensional. The full conditional distributions are not easy to sample from, so we use the random-walk M-H algorithm. We take the maximum likelihood estimate (MLE) to be the starting point as well as the distinguished point $\tilde{\beta}$ for splitting the sampler. We take $d = .95$, again noting that this can be adapted.

The proposal dispersion matrix is chosen to be .1 times a diagonal matrix having the variances of the MLE's along its diagonal. The acceptance rate of the nonadaptive sampler is about 54%. This acceptance rate is much higher than the "optimal" rate of 27.9% proposed by Gelman et al. (1996) for a simpler situation, so we use the adaptive sampler to tune the scaling. At regeneration, we update each variance of the proposal distribution using (14). The adaptive sampler quickly adapts the proposal scalings to have acceptance rate near 28%, and ergodic averages behave substantially better. Figure 2 plots the kernel density estimate of the rate parameter β_4.

4.2 Independence Samplers

4.2.1 Example 3. Here we give an illustration of how the proposal distribution for the independence sampler can be adapted. Our target distribution is a mixture of three bivariate normal distributions:

$$.34 \times \mathrm{N}_2\{\mathbf{0}, \mathbf{I}_2\} + .33 \times \mathrm{N}_2\left\{-\begin{pmatrix} 3 \\ 3 \end{pmatrix}, \begin{pmatrix} 1 & .9 \\ .9 & 1 \end{pmatrix}\right\}$$

$$+ .33 \times \mathrm{N}_2\left\{\begin{pmatrix} 2 \\ 2 \end{pmatrix}, \begin{pmatrix} 1 & -.9 \\ -.9 & 1 \end{pmatrix}\right\}.$$

The contours of this distribution are plotted in Figure 3. We consider independence samplers with a normal proposal distribution with mean $\mathbf{0}$ and dispersion $2 \times \mathbf{I}_2$. (This is justified on the grounds that preliminary investigation revealed only one mode, and we take an overdispersed proposal density around that mode.) We first consider a PAS with one adaptation. Based on the first 5,000 iterations, we calculate the mean and dispersion of the target distribution and use those as the parameters of the proposal distribution.

To implement the fully adaptive sampler, we follow the details outlined in Section 3.1. Starting with the same initial proposal distribution, we adapt the mean and dispersion at regeneration points. The independence chain produced many regenerations, and we decided to adapt the parameters if the number of iterations from the last adaptation exceeded a threshold value of 100 (primarily to avoid the computations needed for possibly unnecessary adaptation).

The autocorrelation plots of the two schemes are given in Figure 4. Clearly, the fully adaptive scheme provides faster mixing than the PAS. We have also experimented with other combinations of starting parameters and number of iterations and burn-ins. In all cases the adaptive sampler provided a better and faster reconstruction of the target distribution than the PAS. More ingenious adaptation schemes might also be tried here for better results. However, we have not pursued those, as the simple adaptation scheme does a substantially better job than the PAS.

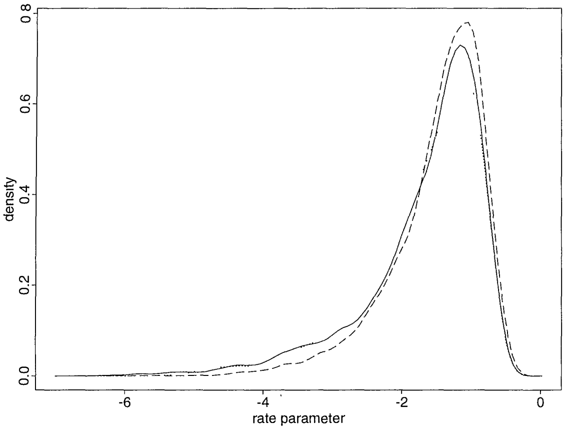

Figure 2. Estimates of the Marginal Posterior Density for β_4. ——, Represents the adaptive sampler; · · ·, the nonadaptive sampler; – – –, the weighted likelihood bootstrap method of Raftery and Newton (1994).

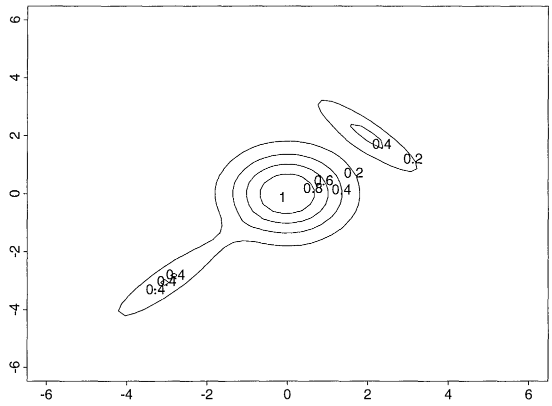

Figure 3. Contours of the Target Distribution of Section 4.2.1.

4.2.2 Example 4. We consider a real data example on age and length measurements on $n = 27$ dugongs (sea cows). Carlin and Gelfand (1991) provided a Bayesian analysis of the dataset originally given by Ratkowsky (1983). The length y_i given the age x_i for the ith individual is assumed to follow the following nonlinear growth curve model:

$$y_i \sim \mathrm{N}(\alpha - \beta\gamma^{x_i}, \sigma^2),$$

where $\alpha, \beta > 1$, $0 < \gamma < 1$, and $i = 1, 2, \ldots, n$.

Following the implementation of this problem in the BUGS software (Speigelhalter et al. 1996), we assume that $\tau = \sigma^{-2}$ follows a gamma prior with density proportional to $\tau^{a-1}e^{-a\tau}$ with $a = 10^{-3}$. Flat priors are assumed for the remaining parameters. We integrate out τ analytically and work with the resulting marginal density of α, β, and γ as follows:

$$\pi(\alpha, \beta, \gamma | y_1, y_2, \ldots y_n)$$
$$\propto \left\{2a + \sum_{i=1}^{n} (y_i - \alpha + \beta\gamma^{x_i})^2\right\}^{-a-n/2}$$

We are interested in the marginal density of γ. The marginal density can be found exactly by integrating out α, β, and τ (in that order) from the full joint posterior density. We want to compare the performance of the MCMC methods by reconstructing the marginal density using the output of the MCMC samplers.

We attempt an independence sampler with a normal proposal distribution with its mean at the MLE, $\tilde{\boldsymbol{\beta}}$, say of α, β, and γ and a covariance matrix with all off-diagonal entries 0 and diagonal entries equal to the variance estimates of the MLE. For the adaptive scheme, we first take c to be $w(\tilde{\boldsymbol{\beta}})/2$ in the splitting construction presented in Section 3.1. At regeneration points, we calculate the mean and covariance matrix of the values sampled so far and use a normal distribution with these updated parameters as the next proposal distribution for the independence sampler. We also update the parameter c by replacing it with $w(\boldsymbol{\beta})/2$, evaluated at the largest value of $\pi(\boldsymbol{\beta})$ discovered so far.

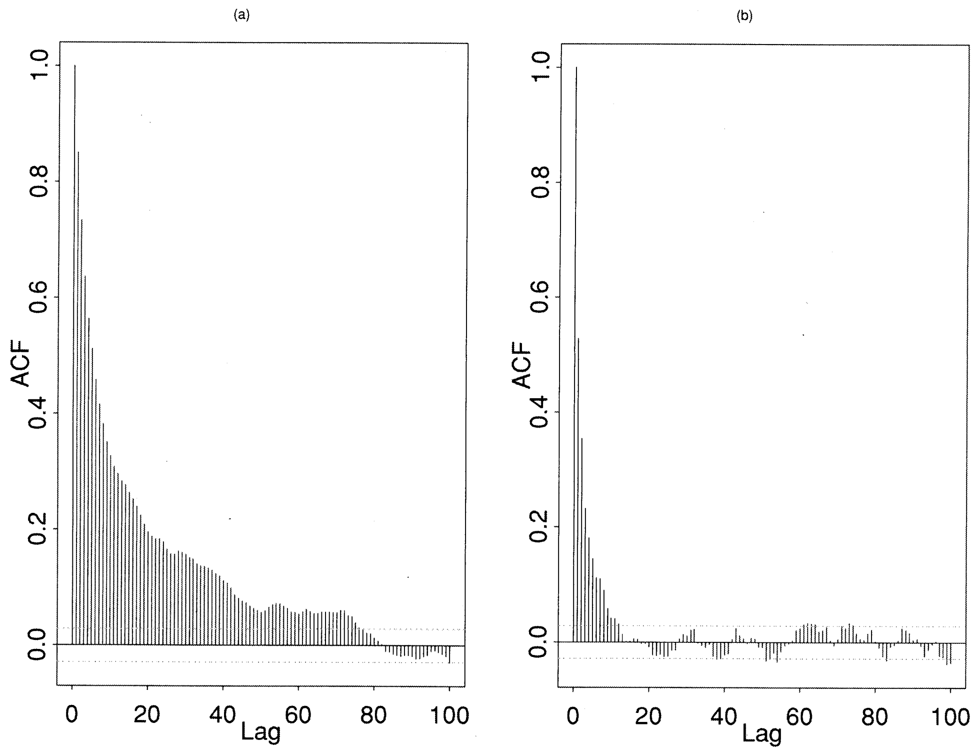

Figure 4. AutoCorrelation Plots. (a) PAS scheme; (b) fully adaptive scheme.

To compare the adaptive schemes to the untuned methods, we also consider a random-walk sampler for this problem with the PAS. The starting proposal distribution is normal with the same covariance matrix as earlier. We update this covariance matrix only once after a burn-in period of 5,000 iterations. We calculate an unbiased estimate of the target covariance matrix based on the 5,000 sampled values, then replace the starting proposal variance matrix by the foregoing estimate.

Figure 5 plots the marginal density of γ from these two schemes. The dotted curve is the result from 10,000 iterates of the adaptive process after a burn-in period of 5,000 iterations. The dashed curve is based on the same number of iterations from the aforementioned PAS; the solid line is the actual density. All densities are scaled to have maximum 1, and the two kernel density estimates are based on the same value of the smoothing parameter. Figure 5 shows that the adaptive scheme is better than the nonadaptive scheme in approximating the true density. However, it may not be a huge improvement over the nonadaptive scheme. This is because this is a relatively simple problem for the MCMC methods, and the BUGS software also produces similar estimates. However, the point of this example is that even for simpler problems, the adaptive schemes that we propose can improve the MCMC samplers.

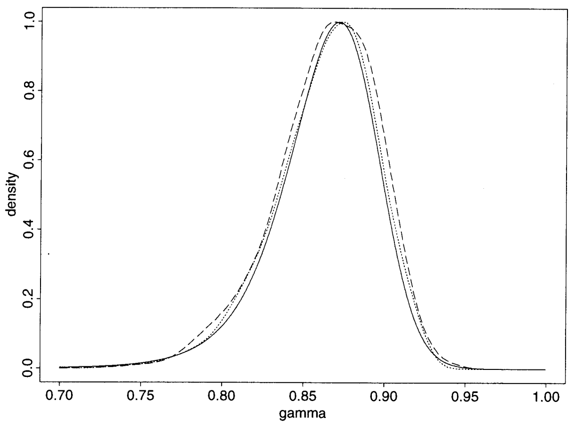

Figure 5. Marginal Posterior Density for γ. ——, Exact density; · · ·, kernel density estimate from the output of the adaptive sampler; – – –, kernel density estimate from the output of the PAS.

5. DISCUSSION

We have provided a theoretical framework for adaptation in MCMC samplers based on regenerative simulation. When a regeneration is obtained, proposal distributions can be adapted on the basis of the output from the sampler obtained so far. How these opportunities for adaptation are exploited depends on the application. For example, the scaling of proposal distributions might be adjusted, or if multiple modes in π are a problem, additional mode-hopping proposal distributions might be introduced. The methodology is very general and works well in low-dimensional problems, as we have illustrated.

In high dimensions, the practicality of our methodology is limited by the practicality of regenerative simulation itself; it is generally difficult to obtain frequent regenerations in high dimensions. In particular, we have shown for a simple but high-dimensional π that the splitting for the random-walk sampler is inadequate. Thus we anticipate that this splitting will be unlikely to work in more complex applications. This result does not apply to the independence sampler; indeed, with $f = \pi$, we can regenerate at every iteration. However, regeneration probabilities are bounded above by acceptance probabilities (10), and unless f is carefully tailored to π, acceptance probabilities may be quite small in high dimensions. Nevertheless, with the available regenerative technology, splitting the independence sampler may be the most promising option, especially because regenerations so obtained provide the opportunity to adapt any or all of the proposal distributions in a hybrid sampler, as discussed in Section 3.3. Perhaps splitting schemes other than those described in Section 3, or indeed regenerative schemes other than Nummelin's splitting strategy, described in Section 2.1, might provide more opportunities for adaptation. Further work in this area is needed.

APPENDIX: PROOFS

Proof of Theorem 1

First, note that $\mathbb{E}[Z_i Z_j] = 0$ for $i \neq j$, because $\mathbb{E}[Z_i|\mathcal{F}_{i-1}] = 0$. Then we see that,

$$\begin{aligned}\mathrm{MSE}(R_n) &= \mathbb{E}[(R_n - \mathbb{E}_\pi[g])^2] = \mathbb{E}\left[\left(\frac{\sum Z_i}{\sum N_i}\right)^2\right]\\ &\leq \frac{1}{n^2}\mathbb{E}\left[\left(\sum Z_i\right)^2\right]\\ &= \frac{1}{n^2}\sum \mathbb{E}\left[Z_i^2\right] < b_1^2/n \to 0 \quad \text{as} \quad n \to \infty.\end{aligned}$$

To establish a limiting distribution for R_n, we use the following standard result, which we state without proof.

Lemma A.1. Let X_n, X, Y_n be random variables such that $X_n \Rightarrow X$, and $Y_n \xrightarrow{P} y$ for some nonzero constant y. Then,

$$\frac{X_n}{Y_n} \Rightarrow \frac{X}{y}.$$

Proof of Central Limit Theorem 2

We first prove a uniform integrability condition. Letting $\mathrm{I}_{[\,]}$ denote the indicator function,

$$\begin{aligned}&\mathbb{E}\left[Z_i^2\,\mathrm{I}_{[|Z_i|>\varepsilon\sqrt{n}]} \mid \mathcal{F}_{i-1}\right]\\ &\leq \left\{\mathbb{E}\left[Z_i^{2+\varepsilon} \mid \mathcal{F}_{i-1}\right]\right\}^{[2/(2+\varepsilon)]}\\ &\quad \times \left\{\Pr\left[|Z_i| > \varepsilon\sqrt{n} \mid \mathcal{F}_{i-1}\right]\right\}^{[\varepsilon/(2+\varepsilon)]}\\ &\quad < b_2^2\left\{\frac{1}{\varepsilon^2 n}\,\mathbb{E}\left[Z_i^2 \mid \mathcal{F}_{i-1}\right]\right\}^{[\varepsilon/(2+\varepsilon)]}\\ &\quad < b_2^{2+[2\varepsilon/(2+\varepsilon)]}\left(\varepsilon^2 n\right)^{-[\varepsilon/(2+\varepsilon)]},\end{aligned}$$

by the Hölder, Markov and Jensen inequalities. Therefore, unconditionally we have

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[Z_i^2\mathrm{I}_{[|Z_i|>\varepsilon\sqrt{n}]}\right] < b_2^{2+\varepsilon}\left(\varepsilon^2 n\right)^{-[\varepsilon/(2+\varepsilon)]} \to 0 \qquad \text{as} \quad n \to \infty.$$

With this uniform integrability result and assumption b, a martingale central limit theorem (essentially taken from thm. 1.6 of Basawa and Prakasa Rao, 1980, p.387) gives

$$\frac{1}{\mathbf{V}\sqrt{n}}\sum Z_i \;\Rightarrow\; \mathrm{N}(0,1).$$

Therefore, from Lemma A.1 and assumption c of Theorem 2, we have

$$\frac{C\sqrt{n}}{\mathbf{V}}\left(R_n - \mathbb{E}_\pi[g]\right) = \frac{\frac{1}{V\sqrt{n}}\sum Z_i}{\frac{1}{Cn}\sum N_i} \Rightarrow \mathrm{N}(0,1).$$

Proof of Theorem 3

Let $Y_i = Z_i^2 - \mathbb{E}\left[Z_i^2 \mid \mathcal{F}_{i-1}\right]$. Then we have $\mathbb{E}[Y_i] = 0$, the $\{Y_i\}$ are uncorrelated, and, from the assumption on b_3, $\mathrm{var}[Y_i] < b_3^4$, using Jensen's inequality. So by the weak law (e.g., Durrett 1991, p. 29), we can prove $n^{-1}\sum Y_i \Rightarrow 0$. With assumption b of Theorem 2, this gives $n^{-1}\sum Z_i^2 \xrightarrow{P} \mathbf{V}^2$, by Slutsky's theorem. So by Lemma A.1, we have

$$n\widehat{\mathrm{MSE}}(R_n) = \frac{\frac{1}{n}\sum Z_i^2}{\left(\frac{1}{n}\sum N_i\right)^2} \xrightarrow{P} \frac{\mathbf{V}^2}{C^2}.$$

Before proving Theorem 4, we note the following general results.

Lemma A.2. It is easy to use Lagrange optimization to verify

$$\inf_{\mathbf{y}\in D}\mathbf{y}'\mathbf{\Gamma}^{-1}\mathbf{x} = -\sqrt{d}\sqrt{\mathbf{x}'\mathbf{\Gamma}^{-2}\mathbf{x}}.$$

Suppose that $q(\mathbf{x},\mathbf{y})$ is a normal density with mean $\mathbf{x}$ and dispersion matrix $\mathbf{\Gamma}$; that is,

$$q(\mathbf{x},\mathbf{y}) = |\mathbf{\Gamma}|^{-1/2}(2\pi)^{-m/2} \times \exp\left\{-\frac{1}{2}(\mathbf{y}-\mathbf{x})'\mathbf{\Gamma}^{-1}(\mathbf{y}-\mathbf{x})\right\}. \quad \text{(A.1)}$$

We can split the foregoing normal transition kernel (A.1) using (13) and Lemma 2. Let the distinguished point $\tilde{\mathbf{x}} = \mathbf{0}$ and the set D be $\{\mathbf{y}\colon\ \mathbf{y}'\mathbf{y} \leq d\}$, where $d > 0$ is a scalar.

Lemma A.3. The pair (s_q, ν_q), where

$$s_q(\mathbf{x}) = \exp\left\{-\frac{1}{2}\mathbf{x}'\mathbf{\Gamma}^{-1}\mathbf{x} - \sqrt{d}\sqrt{\mathbf{x}'\mathbf{\Gamma}^{-2}\mathbf{x}}\right\}$$

and

$$\nu_q(\mathbf{y}) = q(\mathbf{0},\mathbf{y})\,1(\mathbf{y} \in D),$$

provides a splitting for the transition density $q(\mathbf{x},\mathbf{y})$ as given in (A.1).

Proof of Theorem 4

It is easy to see that $r(s,\nu) \leq K_1 \times K_2$ where $K_1 .= \mathbb{E}_\pi\left[s_q(\mathbf{X})\right]$ and $K_2 = \int \nu_q(d\mathbf{y})$. Choosing $\mathbf{\Gamma} = k\mathbf{I}$ in Lemma A.3, we have

$$\begin{aligned}K_1 &= (2\pi)^{-m/2}\int \exp\left\{-\frac{1}{2k}\,\mathbf{x}'\mathbf{x} - \frac{\sqrt{d}}{k}\sqrt{\mathbf{x}'\mathbf{x}} - \frac{1}{2}\,\mathbf{x}'\mathbf{x}\right\}\,d\mathbf{x}\\ &= (2\pi)^{-m/2}\left(\frac{k}{k+1}\right)^{m/2}\\ &\quad \times \int \exp\left\{-\frac{1}{2}\,\mathbf{z}'\mathbf{z} - \frac{\sqrt{d}}{k}\left(\frac{k}{k+1}\right)^{1/2}\sqrt{\mathbf{z}'\mathbf{z}}\right\}\,d\mathbf{z}\\ &= \left(\frac{k}{k+1}\right)^{m/2}\mathbb{E}\left[\exp\left\{-\sqrt{\frac{d}{k(k+1)}}\,\sqrt{\chi_m^2}\right\}\right]\end{aligned}$$

and

$$K_2 = \int \nu_q(d\mathbf{y}) = \mathbb{E}\left[\chi_m^2 \leq \frac{d}{k}\right],$$

where $\chi_m^2 \sim \chi^2$ distribution with m df. Now we have

$$\begin{aligned} K_2 &= 2^{-m/2}\{\mathbf{\Gamma}(m/2)\}^{-1} \int_0^{d/k} \exp\left(-\frac{u}{2}\right) u^{m/2-1}\, du \\ &\leq 2^{-m/2}\{\mathbf{\Gamma}(m/2)\}^{-1} \int_0^{d/k} \exp\left(-\frac{u}{2}\right) \left(\frac{d}{k}\right)^{m/2-1} du \\ &= 2^{-m/2+1}\{\mathbf{\Gamma}(m/2)\}^{-1} \left(\frac{d}{k}\right)^{m/2-1} \left(1 - \exp\left(-\frac{d}{2k}\right)\right) \\ &\leq 2^{-m/2+1}\{\mathbf{\Gamma}(m/2)\}^{-1} \left(\frac{d}{k}\right)^{m/2-1}. \end{aligned}$$

The foregoing expressions for K_1 and K_2 yield

$$\begin{aligned} r(s,\nu) \leq{}& \frac{k(k+1)^{-m/2}}{2^{m-1}\{\Gamma(m/2)\}^2} \\ &\times \int_0^\infty u^{m/2-1} \exp\left\{-\frac{u}{2} - \sqrt{\frac{d}{k(k+1)}}\sqrt{u}\right\} d^{m/2-1}\, du. \end{aligned}$$

The maximum of the foregoing integrand with respect to d will be achieved at $ud = k(k+1)(m-2)^2$. Therefore, we see that

$$\begin{aligned} r(s,\nu) \leq{}& k^{m/2}(k+1)^{-1} \exp\{2-m\} \\ &\times (m-2)^{m-2} 2^{2-m}\{\mathbf{\Gamma}(m/2)\}^{-2}. \quad \text{(A.2)} \end{aligned}$$

Using Stirling's approximation for gamma function, we can approximate the upper bound in (A.2) by

$$r_m = k^{m/2}(k+1)^{-1}\pi^{-1}(m-2)^{-1}.$$

Setting $k = 2.38^2/m$, it is easy to see that the theorem follows.

[Received March 1996. Revised March 1998.]

REFERENCES

Basawa, I. V., and Prakasa Rao, B. L. S. (1980), *Statistical Inference for Stochastic Processes*, London: Academic Press.

Bates, D. M., and Watts, D. G. (1988), *Nonlinear Regression Analysis and Its Applications*, New York: Wiley.

Besag, J., Green, E., Higdon, D., and Mengersen, K. (1995), "Bayesian Computation and Stochastic Systems" (with discussion), *Statistical Science*, 10, 3–66.

Carlin, B. P. and Gelfand, A. E. (1991), "An Iterative Monte Carlo Method for Nonconjugate Bayesian Analysis," *Statistics and Computing*, 1, 119–128.

Cowles, M. K., and Rosenthal, J. S. (1996), "A Simulation Approach to Convergence Rates for Markov Chain Monte Carlo," technical report, University of Toronto, Dept. of Statistics.

Durrett, R. (1991), *Probability: Theory and Examples*, Belmont, CA: Wadsworth.

Gelfand, A. E., and Sahu, S. K. (1994), "On Markov Chain Monte Carlo Acceleration," *J. Comp. Graph. Statist.* , 3, 261–276.

Gelfand, A. E., Sahu, S. K., and Carlin, B. P. (1995), "Efficient Parameterizations for Normal Linear Mixed Models," *Biometrika* 82, 479–488:

Gelfand, A. E., and Smith, A. F. M. (1990), "Sampling-Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association*, 85, 398–409.

Gelman, A., Roberts, G. O., and Gilks, W. R. (1996), "Efficient Metropolis Jumping Rules," in *Bayesian Statistics 5*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, Oxford, U.K.: Oxford University Press, pp. 599–608.

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.

Gilks, W. R., and Roberts, G. O. (1996), "Strategies for Improving MCMC," in *Markov Chain Monte Carlo in Practice*, eds. W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, London: Chapman and Hall, pp. 89–114.

——— (1996), "Strategies for Improving MCMC," in *MCMC in Practice*, eds. W. R. Gilks, D. J. Spiegelhalter, and S. Richardson, London: Chapman and Hall, pp. 89–114.

Gilks, W. R., Roberts, G. O., and George, E. I. (1994), "Adaptive Direction Sampling," *The Statistician*, 43, 179–189.

Hastings, W. K. (1970), "Monte Carlo Sampling Methods Using Markov Chains and Their Applications," *Biometrika*, 57, 97–109.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953), "Equations of State Calculations by Fast Computing Machine," *The Journal of Chemical Physics*, 21, 1087–1091.

Mykland, P., Tierney, L., and Yu, B. (1995), "Regeneration in Markov Chain Samplers," *Journal of the American Statistical Association*, 90, 233–241.

Newton, M. A., and Raftery, A. E. (1994), "Approximate Bayesian Inference With the Weighted Likelihood Bootstrap" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 56, 3–48.

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

Ratkowsky, D. (1983), *Nonlinear Regression Modelling*, New York: Marcel Dekker.

Robert, C. P. (1995), "Convergence Control Methods for Markov Chain Monte Carlo Algorithms," *Statistical Science*, 10, 231–253.

Roberts, G. O. (1996), "Markov Chain Concepts Related to Sampling Algorithms," in *MCMC in Practice* eds. W. R. Gilks, D. J. Spiegelhalter, and S. Richardson, London: Chapman and Hall, pp. 45–57.

Roberts, G. O., and Tweedie, R. L. (1996), "Geometric Convergence and Central Limit Theorems for Multidimensional Hastings–Metropolis Algorithms," *Biometrika*, 83, 1, 1996, 96–110.

Spiegelhalter, D. J., Thomas, A., and Best, N. G. (1996), "Computation on Bayesian Graphical Models," in *Bayesian Statistics 5*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, Oxford U.K.: Oxford University Press, pp. 407–426.

Tierney, L. (1994), "Markov Chains for Exploring Posterior Distributions" (with discussion), *The Annals of Statistics*, 22, 1701–1762.