

Optimal scaling of discrete approximations to Langevin diffusions

by

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

(June 1995; last revised, March 1997.)

Acknowledgements. We are grateful to Radford Neal for bringing references from the physics literature to our attention. We also thank Richard Gibbens, Michael Miller, and Richard Tweedie, for helpful discussions, and would like to thank the referees very much for their work. This work was supported in part by EPSRC of the U.K. and by NSERC of Canada.

Abstract. We consider the optimal scaling problem for proposal distributions in Hastings-Metropolis algorithms derived from Langevin diffusions. We prove an asymptotic diffusion limit theorem and show that the relative efficiency of the algorithm can be characterised by its overall acceptance rate, independently of the target distribution. The asymptotically optimal acceptance rate is 0.574. We show that as a function of dimension n , the complexity of the algorithm is $O(n^{1/3})$, which compares favourably with the $O(n)$ complexity of random-walk Metropolis algorithms. We illustrate this comparison with a number of example simulations.

Keywords. Langevin algorithm, Hastings-Metropolis, Markov chain Monte Carlo, weak convergence.

* Statistical Laboratory, University of Cambridge, Cambridge CB2 1SB, U.K. Internet: G.O.Roberts@statslab.cam.ac.uk.

** Department of Statistics, University of Toronto, Toronto, Ontario, Canada M5S 1A1. Internet: jeff@utstat.toronto.edu. Supported in part by NSERC of Canada.

1. Introduction.

This paper contains theoretical results related to the practical implementation of certain Metropolis-Hastings algorithms (Metropolis et al., 1953; Hastings, 1970; Smith and Roberts, 1993) as used to explore probability distributions, for example in Bayesian statistics. Specifically, we consider issues related to discrete approximations to Langevin diffusions, as proposed in Grenander and Miller (1994), Phillips and Smith (1994), and Roberts and Tweedie (1995).

Hastings-Metropolis algorithms are now routinely used in many statistical applications (see for example Smith and Roberts, 1993; Besag and Green, 1993). The most commonly used algorithm of this type is the random walk algorithm, which is largely appealing because it is not problem-specific and hence is easy to implement. However, as a result of this, it can frequently be slow to converge. Langevin algorithms use local problem-specific information, and are therefore often almost as easy to implement.

In recent work of Roberts, Gelman and Gilks (1994), the problem of optimal scaling of proposal variances for random-walk Metropolis algorithms was considered. It was proved that, for Gaussian proposals and certain target distributions, the asymptotic acceptance probability should be tuned to be approximately 0.234 for optimal performance of the algorithm. Furthermore, it was shown that the proposal variance should scale as n^{-1} as the dimension $n \rightarrow \infty$. The paper thus provided a useful heuristic for running Metropolis algorithms efficiently. Although the result applied only asymptotically as $n \rightarrow \infty$, numerical studies (Gelman, Roberts, and Gilks, 1994) indicated that the result was a good approximation even in low dimensions. However, this result does not apply to more general Hastings algorithms. If the proposal density makes use of the structure of the target density, intuition suggests that a higher acceptance probability is likely to be preferred.

In this paper we carry out a similar study for a class of algorithms given by discrete approximations to Langevin diffusions. A Langevin diffusion for a multivariate probability density function π (with respect to Lebesgue measure) is the unique (up to a speed factor) non-explosive diffusion which is reversible with respect to π . It makes use of the gradient of π to move more often in directions in which π is increasing. Thus, a discrete approximation to a Langevin diffusion should have an optimal acceptance probability which is larger than the 0.234 figure for random-walk proposals.

Roberts and Tweedie (1995) demonstrate that Langevin algorithms can converge at sub-geometric rates (at least for certain classes of target densities). However it is also well

known that Langevin algorithms can be significantly more efficient than their natural competitors, random walk Metropolis algorithms, particularly in high dimensional problems. In fact we shall be able to make direct comparisons of the efficiency of these algorithms asymptotically in dimension.

Our main results may be summarised as follows. For discrete approximations to Langevin diffusions for certain target distributions π , the optimal asymptotic scaling can be characterised as being that algorithm which has limiting acceptance probability in high dimensions (i.e. as $n \rightarrow \infty$) approximately 0.574. Furthermore, the proposal variance should scale as $n^{-1/3}$, and thus $O(n^{1/3})$ steps are required for the Langevin algorithm to converge. Therefore, Langevin algorithms are considerably more efficient than random-walk based Metropolis methods which require $O(n)$ steps for the same class of target densities.

To understand the above, we need to know by what criterion we are measuring optimality. Suppose X is our Markov chain, and f is some function of interest: that is, we wish to estimate $\pi(f) \equiv \mathbf{E}_\pi[f(X)]$. Assuming that a central limit theorem holds for X and f , a natural measure of efficiency is related to the variance of the ergodic estimate of $\pi(f)$:

$$e_f = \left(\lim_{t \rightarrow \infty} t \mathbf{Var} \left(\frac{\sum_{i=1}^t f(X_i)}{t} \right) \right)^{-1}.$$

Therefore e_f^{-1} is proportional to the number of iterations needed to achieve a particular accuracy for the ergodic estimate of $\pi(f)$. Unfortunately, the efficiency e_f varies with f , and furthermore, there is no clear relationship between e_f and other natural measures of efficiency (perhaps to do with convergence rates of algorithms). However, for the diffusion limit, there is only one sensible measure of its efficiency, its speed measure. All other measures of efficiency are equivalent (up to a normalization constant) including those described above.

We note also that, while 0.574 is the optimal acceptance probability, the speed of the algorithm remains relatively high for acceptance probabilities between, say, 0.4 and 0.8. Practical implications such as these are considered in Section 3, and a number of simulations are described that show the asymptotic behaviour of the algorithms. In particular, a comparison between the Langevin algorithm and the random walk Metropolis algorithm, and a comparison with the asymptotically optimal algorithm is given in Fig 3.1.

We prove our results formally only for target distributions of the form $\pi_n(\mathbf{x}) =$

$\prod_{i=1}^n f(x_i)$ corresponding to i.i.d. components. However, various generalisations are possible; see Section 4 and the similar discussion in Roberts, Gelman, and Gilks (1994, Section 3). Furthermore, such optimal-scaling results appear to be quite robust over changes in the model; see for example Gelman, Roberts, and Gilks (1994).

Similar algorithms have been studied in various contexts in the physics literature (Neal, 1993, Section 5.3; Neal, 1994). Algorithms similar to discrete Langevin diffusions were proposed by Rossky, Doll, and Friedman (1978). The idea that the proposal variance should scale as $n^{-1/3}$ is suggested in Kennedy and Pendleton (1991). Also, optimal acceptance probabilities are considered through simulations in Mountain and Thirumalai (1994).

Our formal definitions are as follows. The reversible Langevin diffusion for the n -dimensional density π_n , with variance σ^2 , is the diffusion process $\{\mathbf{\Lambda}_t\}$ which satisfies the stochastic differential equation

$$d\mathbf{\Lambda}_t = \sigma d\mathbf{B}_t + \frac{\sigma^2}{2} \nabla \log \pi_n(\mathbf{\Lambda}_t) dt,$$

where \mathbf{B}_t is standard n -dimensional Brownian motion. Thus, the natural discrete approximation can be written

$$\tilde{\mathbf{\Lambda}}_{t+1} = \tilde{\mathbf{\Lambda}}_t + \sigma_n \mathbf{Z}_{t+1} + \frac{\sigma_n^2}{2} \nabla \log \pi_n(\tilde{\mathbf{\Lambda}}_t)$$

where the random variables \mathbf{Z}_t are distributed as independent standard normal, and where σ_n^2 is the chosen step variance. However, such discrete approximations can have vastly different asymptotic behaviours from the diffusion process they attempt to approximate (Roberts and Tweedie, 1995). Specifically, these approximations can be transient, no matter how small the step variance σ_n^2 . Therefore, to construct a sound algorithm based on the Langevin diffusion, it is not sufficient to merely approximate the diffusion itself. It is necessary to introduce a Metropolis accept/reject step (Metropolis et al., 1953; Hastings, 1970) which serves to ensure that π_n is a stationary distribution for the process.

The algorithm proceeds as follows. Given \mathbf{X}_t , we choose a proposal random variable \mathbf{Y}_{t+1} by

$$\mathbf{Y}_{t+1} = \mathbf{X}_t + \sigma_n \mathbf{Z}_{t+1} + \frac{\sigma_n^2}{2} \nabla \log \pi_n(\mathbf{X}_t)$$

and then set $\mathbf{X}_{t+1} = \mathbf{Y}_{t+1}$ with probability

$$\alpha_n(\mathbf{X}_t, \mathbf{Y}_{t+1}) = \frac{\pi_n(\mathbf{Y}_{t+1})q_n(\mathbf{Y}_{t+1}, \mathbf{X}_t)}{\pi_n(\mathbf{X}_t)q_n(\mathbf{X}_t, \mathbf{Y}_{t+1})} \wedge 1$$

where

$$q_n(\mathbf{x}, \mathbf{y}) = \frac{1}{(2\pi\sigma_n^2)^{n/2}} \exp\left(\frac{-1}{2\sigma_n^2} \|\mathbf{y} - \mathbf{x} - \frac{\sigma_n^2}{2} \nabla \log \pi_n(\mathbf{x})\|_2^2\right) \equiv \prod_{i=1}^n q(x_i^n, y_i),$$

and $\|\cdot\|_2$ is the usual L^2 -norm. Otherwise, with probability $1 - \alpha_n(\mathbf{X}_t, \mathbf{Y}_{t+1})$, we set $\mathbf{X}_{t+1} = \mathbf{X}_t$.

Thus the discrete algorithm has the desired stationary distribution π_n . However, the practical problem of determining the size of σ_n^2 remains. Specifically, a larger value of σ_n^2 corresponds to a larger proposal step size. This potentially allows for faster mixing, but only if the acceptance probabilities do not become unacceptably small. Such issues are the subject of the present paper.

2. Main results.

We consider the Metropolis-adjusted discrete approximations $\{\mathbf{X}_t\}$ to the Langevin diffusion for π_n as above, with

$$\pi_n(\mathbf{x}) = \prod_{i=1}^n f(x_i) = \prod_{i=1}^n e^{g(x_i)} \quad (2.1)$$

a fixed probability distribution on \mathbf{R}^n . Throughout, we shall assume that \mathbf{X}_0 is distributed according to the stationary measure π . We further assume that g is a C^8 function with derivatives $g^{(i)}$ satisfying

$$|g(x)|, |g^{(i)}(x)| \leq M_0(x) \quad (2.2)$$

$1 \leq i \leq 8$, for some polynomial $M_0(\cdot)$, and that

$$\int_{\mathbf{R}} x^k f(x) dx < \infty, \quad k = 1, 2, 3, \dots \quad (2.3)$$

Finally, in order to apply standard SDE results, we assume that g' is a Lipschitz function.

In order to compare these discrete approximations to limiting continuous-time processes, it is convenient to define the discrete approximations as jump processes with exponential holding times. Specifically, we let $\{J_t\}$ be a Poisson process with rate $n^{1/3}$, and let $\mathbf{\Gamma}^n = \{\mathbf{\Gamma}_t^n\}_{t \geq 0}$ be the n -dimensional jump process defined by $\mathbf{\Gamma}_t^n = \mathbf{X}_{J_t}$ where we take $\sigma_n^2 = \ell^2 n^{-1/3}$ in the definitions from the previous section, with ℓ an arbitrary positive constant. We assume throughout that $\{\mathbf{X}_t\}$ is non-explosive. We let

$$a_n(\ell) = \int \int \pi_n(\mathbf{x}) q_n(\mathbf{x}, \mathbf{y}) \alpha_n(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y} = \mathbf{E} \left(\frac{\pi_n(\mathbf{Y}) q_n(\mathbf{Y}, \mathbf{X})}{\pi_n(\mathbf{X}) q_n(\mathbf{X}, \mathbf{Y})} \wedge 1 \right),$$

be the π_n -average acceptance rate of the algorithm which generates $\mathbf{\Gamma}$.

The two main results in this paper are the following.

Theorem 1. We have

$$\lim_{n \rightarrow \infty} a_n(\ell) = a(\ell),$$

where $a(\ell) = 2 \Phi(-K\ell^3/2)$, with $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-t^2/2} dt$ and

$$K = \sqrt{\mathbf{E} \left(\frac{5g'''(X)^2 - 3g''(X)^3}{48} \right)} > 0, \quad (2.4)$$

with the expectation taken over X having density $f = e^g$.

Theorem 1 gives a formula for the asymptotic acceptance probability of the algorithm. We note that, in the Gaussian case where $g(x) = -x^2/2 + C$, this theorem reduces precisely to equation (2.3) of Kennedy and Pendleton (1991).

Theorem 2. Let $\{U^n\}_{t \geq 0}$ be the process corresponding to the first component of $\mathbf{\Gamma}^n$. Then as $n \rightarrow \infty$, the process U^n converges weakly (in the Skorokhod topology) to the Langevin diffusion \mathbf{U} defined by

$$dU_t = (h(\ell))^{1/2} dB_t + \frac{1}{2} h(\ell) \frac{d}{dx} \log \pi_1(U_t) dt,$$

where $h(\ell) = 2\ell^2 \Phi(-K\ell^3/2)$ is the speed of the limiting diffusion. Furthermore, $h(\ell)$ is maximized at the unique value of ℓ for which $a(\ell) = 0.574$ (to three decimal places).

Theorem 2 may be interpreted as follows. For a given target density π_n as above, with n large, suppose a Metropolis-adjusted discrete approximation to the Langevin diffusion for π_n is run with proposal steps of variance σ_n^2 . Then setting $\ell_n = \sigma_n n^{1/6}$, the theorem says that the speed (which is proportional to the mixing rate) of the process is approximately given by $h(\ell_n)$. Furthermore, the optimal value $\hat{\ell}_n$ of ℓ_n which maximizes this speed is that for which the asymptotic acceptance probability $a_n(\hat{\ell}_n)$ is approximately 0.574. Hence σ_n^2 should be tuned to be approximately $\hat{\ell}_n^2 n^{-1/3}$, which will make the acceptance probability approximately 0.574. If it is discovered that the acceptance rate is substantially smaller or substantially larger than 0.574, then the value of σ_n^2 should be modified accordingly.

3. Practical implications.

In this section, we will illustrate our results with a collection of simulation studies. The asymptotic results of Section 2 hold approximately in rather low dimensional problems as we shall observe. We shall also see the efficiency advantage of the Langevin algorithm over the Metropolis algorithm ($O(n^{2/3})$ as discussed earlier).

Measurements of efficiency in low dimensional Markov chains are not unique (see Gelman et. al., 1994 for a discussion). Perhaps the most natural measure of efficiency would be the asymptotic variance of a relevant quantity of interest. Since we do not wish to specify such a quantity of interest, we have opted instead for a different (but asymptotically equivalent) criterion. We therefore measure the average squared jumping distance for the algorithm, which we call the *first order efficiency*.

Definition. First order efficiency of a multi-dimensional Markov chain \mathbf{X} with first component $X^{(1)}$ say.

$$\text{first order efficiency} = \mathbf{E} \left((X_{t+1}^{(1)} - X_t^{(1)})^2 \right)$$

where \mathbf{X}_t is assumed to be stationary .

We note that maximising this quantity is equivalent to minimizing the lag-1 auto-correlations. Recall also that all sensible measurements of efficiency of the algorithm are asymptotically equivalent (up to a normalisation constant) in the diffusion limit.

Figure 3.1 shows estimated first order efficiencies as a function of the acceptance rate (what we term the *efficiency curve*) for a product of standard normals. Dimensions 1, 5, 10, and 20 are given, and in each case two envelopes of curves are given: the upper envelope for each graph represents the first order efficiency for the Langevin algorithm; and the lower envelope, the random walk Metropolis algorithm. The solid curve represents the asymptotic efficiency curve normalised to have the same maximum value as the finite-dimensional Langevin efficiency. Each point on the graph represent an estimate of first order efficiency based on 2×10^5 iterations of the algorithm, and the roughness of the curves show that some error in our estimates of efficiency still exist, although the error is not large enough to obscure the general picture. [Note that we plot first order efficiency times dimension as this is the scale on which the Metropolis algorithm's efficiency is asymptotically constant in dimension.] The value of the first order efficiency of the Langevin algorithm is therefore an absolute measure of efficiency *relative to* the Metropolis dynamics.

Figure 3.1 First order efficiency of Langevin and random walk Metropolis algorithms as a function of overall acceptance rates (efficiency curves). The solid lines show the asymptotic curve for the Langevin algorithm; the upper envelope of points show the Langevin algorithm, and the lower envelope shows the Metropolis algorithm.

The efficiency gain for the Langevin method is immediately obvious, and is increasing with dimension. In one dimension, the optimal acceptance rate is in excess of 0.7, however in 5 or more dimensions the optimal acceptance rate is very close to the limiting value of 0.574. Furthermore, the whole efficiency curve is very well approximated by the asymptotic curve for dimensions greater than 5.

In the case where f is not symmetric, the asymptotic picture is not as clear in low dimensions. Figure 3.2 considers the 10-dimensional case where $g = -x^2/2 - \exp\{x^{-2}/2\}$, $x \geq 0$ and $g = -x^2/2$ for $x < 0$. In this case the efficiency curve is shifted towards lower acceptance rates, and in particular the optimal acceptance rate is less than 0.5. Note however that the asymptotically optimal acceptance rate (0.574) still gives a relative efficiency for

this particular problem in excess of 0.95. Furthermore, additional simulations done for the same density f , but in 1000 dimensions, show excellent agreement with the asymptotic efficiency curve.

Figure 3.2 The relative efficiency of a non-symmetric 10-dimensional density (points), with the solid line representing the asymptotic curve.

Finally in this section, a word of caution. The main results in this paper require very smooth regularity conditions in g . It is tempting to suggest that these are technical conditions required for the proofs, but perhaps inconsequential to the actual asymptotic efficiency curve in practical situations. Although it is highly unlikely that the C^8 condition we impose on g is *necessary* for the results in this paper, some smoothness conditions are necessary.

Consider, for instance, the following example. Suppose $f(x) = 2/3$ for $0 \leq x \leq 1$, and $f(x) = 1/3$ for $-1 \leq x < 0$, with $f(x) = 0$ for $|x| > 1$. In this case, since $g'(x) = 0$ almost everywhere, the Langevin algorithm coincides exactly with the random-walk Metropolis algorithm (from almost all starting values). Therefore, since the asymptotic relative efficiency curves for these algorithms are different, it is impossible that our results and those of Roberts et al. (1994) both apply to this example; this shows the discontinuity of f (and hence of g) is significant. Furthermore, additional simulations (not given here) involving continuous log densities g with discontinuous derivatives suggest that the asymptotics are very different, and asymptotically optimal acceptance rates are likely to be less than 0.574.

4. Extensions.

Our main results are proved for sequences of densities of the product form (2.1). It appears that the asymptotic efficiency curve (efficiency against acceptance rate) is robust to changes in correlation structure in a number of different extensions, although a number of the technical issues addressed in the proofs of Section 5 are even more delicate in such extensions. Some of these types of extensions are discussed in Roberts Gelman and Gilks (1994). Here we will briefly discuss one extension in an attempt to illuminate why the efficiency curve remains invariant, even in the presence of correlation. No attempt at formal proof will be made here.

Consider a sequence of densities $\{\pi_n\}$ with

$$\pi_n(\mathbf{x}) = f_0(\mathbf{x}_0) \prod_{i=0}^n f(\mathbf{x}_i)$$

where we assume that the $\{\mathbf{x}_i\}$ are a -vectors for some $a > 1$, and \mathbf{x}_0 is a vector (not necessarily a -dimensional).

Suppose that we use the usual scaling for the proposal variance, $\sigma_n^2 = \ell^2 n^{-1/3}$, and consider the process produced by just looking at the \mathbf{x}_0 coordinates. The formal asymptotics leading to Theorems 1 and 2 still apply here (although with more involved technicalities), so that analogously, the \mathbf{x}_0 process has a multidimensional diffusion limit with scaling:

$$h(\ell) = 2\ell^2 \Phi(-K_{\text{mult}} \ell^3 / 2) \tag{4.1}$$

where K_{mult} is a multidimensional analogue of K appearing in (2.4), which incorporates information about correlation in π_n . However, analogously to Theorem 1, the acceptance rate also satisfies

$$a(\ell) = 2\Phi(-K_{\text{mult}} \ell^3 / 2) . \tag{4.2}$$

It remains to observe that the efficiency curve from (4.1) and (4.2) only varies with K_{mult} via a multiplicative constant. Specifically, given a constant $0 < a_0 < 1$, and the scaling $\ell(a_0)$ which achieves acceptance rate a_0 ,

$$h(\ell(a_0)) = \frac{a \left(-2\Phi^{-1}(a_0/2) \right)^{2/3}}{K_{\text{mult}}^{2/3}} .$$

Thus, the relative efficiency curve remains unaltered.

We stress that the absolute efficiency of the algorithm varies considerably with many properties of the target density, though the relative efficiency as a function of acceptance rate remains unaltered. Similar arguments can be used to (at least heuristically) justify other extensions of the results of Section 2.

We have only considered the spherical proposal case here. For problems where components have very different scales, it will be sensible to allow differently scaled proposals in different dimensions. There are some interesting problems involved in choosing variance scaling componentwise, perhaps also in an attempt to allow some components to converge quicker than others. We do not pursue these problems here.

Finally, we note that extensions to weak convergence of multivariate components are also possible. For instance, with π_n of the form given in (2.1), it is possible (by very similar techniques to those used below) to show that the process described by the first $c > 1$ components of $\mathbf{\Gamma}^n$ converges weakly to an n -dimensional Markov process consisting of independent components, each of the type given by the \mathbf{U} of Theorem 2.

5. Theorem proofs.

Let us define the generators of the discrete approximation process $\mathbf{\Gamma}^n$ and of the (first-component) Langevin diffusion process with speed $h(\ell)$, viz.

$$G_n V(\mathbf{x}^n) = n^{1/3} \mathbf{E} \left((V(\mathbf{Y}) - V(\mathbf{x}^n)) \left(\frac{\pi_n(\mathbf{Y}) q_n(\mathbf{Y}, \mathbf{x}^n)}{\pi_n(\mathbf{x}^n) q_n(\mathbf{x}^n, \mathbf{Y})} \wedge 1 \right) \right),$$

where the expectation is taken over $\mathbf{Y} \sim q_n(\mathbf{x}^n, \cdot)$; and

$$GV(\mathbf{x}^n) = h(\ell) \left(\frac{1}{2} V''(x_1) + \frac{1}{2} g'(x_1) V'(x_1) \right)$$

(where $g(x_1) = \log f(x_1)$ as above).

To prove the weak convergence of the processes as in Theorem 2, it suffices (Ethier and Kurtz, 1986, Chapter 4, Corollary 8.7) to show that there exist events $F_n^* \subseteq \mathbf{R}^n$ such that for all t ,

$$\mathbf{P}(\mathbf{\Gamma}_s^n \in F_n^* \text{ for all } 0 \leq s \leq t) \rightarrow 1 \tag{5.1}$$

and

$$\lim_{n \rightarrow \infty} \sup_{\mathbf{x}^n \in F_n^*} |G_n V(\mathbf{x}) - GV(\mathbf{x})| = 0$$

for all test functions V in the domain of a “core” for the generator G , provided that this domain strongly separates points. In the present context, by the nature of G , we can

restrict to functions V which depend only on the first coordinate x_1 of \mathbf{x}^n . Furthermore, we may restrict (Ethier and Kurtz, 1986, Chapter 8, Theorem 2.1) to functions V which are in C_c^∞ , i.e. which are infinitely differentiable with compact support.

The essence of the proof will be showing the uniform convergence of G_n to G , as $n \rightarrow \infty$ (and hence $\sigma_n^2 \rightarrow 0$), as above. This will involve careful Taylor series expansions with uniform bounds on remainder terms. It will also involve a quantitative version of the Lindeberg Central Limit Theorem.

To proceed, we expand $G_n V(\mathbf{x})$ in a power series involving powers of $n^{-1/6}$.

Lemma 3. *Defining Z_i by $Y_i = x_i^n + \sigma_n Z_i + \frac{\sigma_n^2}{2} g'(x_i^n)$ (so that Z_i is distributed as standard normal), and recalling that $q(x, y) = \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp\left(-\frac{1}{2\sigma_n^2}(y - x - \frac{\sigma_n^2}{2}g'(x))^2\right)$, there exists a sequence of sets $F_n \in \mathbf{R}^n$, with $\lim_{n \rightarrow \infty} n^{1/3} \pi_n(F_n^C) = 0$, such that*

$$\log\left(\frac{f(Y_i)q(Y_i, x_i^n)}{f(x_i^n)q(x_i^n, Y_i)}\right) = C_3(x_i^n, Z_i)n^{-1/2} + C_4(x_i^n, Z_i)n^{-2/3} + C_5(x_i^n, Z_i)n^{-5/6} \\ + C_6(x_i^n, Z_i)n^{-1} + C_7(x_i^n, Z_i, \sigma_n),$$

where

$$C_3(x_i^n, Z_i) = \ell^3 \left(-\frac{1}{4} Z_i g'(x_i^n) g''(x_i^n) - \frac{1}{12} Z_i^3 g'''(x_i^n) \right);$$

and where $C_4(x_i^n, Z_i)$, $C_5(x_i^n, Z_i)$, and $C_6(x_i^n, Z_i)$ are (also) polynomials in Z_i and the derivatives of g . Furthermore, if \mathbf{E}_Z stands for expectation with $Z \sim N(0, 1)$, and \mathbf{E}_X stands for expectation with X having density $f(\cdot)$, then $\mathbf{E}_X \mathbf{E}_Z(C_3(X, Z)) = \mathbf{E}_X \mathbf{E}_Z(C_4(X, Z)) = \mathbf{E}_X \mathbf{E}_Z(C_5(X, Z)) = 0$, while $\mathbf{E}_X(\mathbf{E}_Z(C_3(X, Z)^2)) = \ell^6 K^2 = -2\mathbf{E}_X \mathbf{E}_Z(C_6(X, Z)) > 0$. In addition,

$$\lim_{n \rightarrow \infty} \sup_{\mathbf{x}^n \in F_n} \mathbf{E} \left| \sum_{i=2}^n \log\left(\frac{f(Y_i)q(Y_i, x_i^n)}{f(x_i^n)q(x_i^n, Y_i)}\right) - \left(\left(n^{-1/2} \sum_{i=2}^n C_3(x_i^n, Z_i) \right) - \ell^6 K^2 / 2 \right) \right| = 0.$$

Proof. The (Taylor-series) expansion follows from straightforward (but messy) computation, done using the Mathematica computation system (Wolfram, 1988). By inspection of the result, the coefficients are polynomials in Z_i and in g and its derivatives. The fact that $\mathbf{E}_X \mathbf{E}_Z(C_3(X, Z)) = \mathbf{E}_X \mathbf{E}_Z(C_5(X, Z)) = 0$ is then immediate because these coefficients contain only terms involving odd powers of Z . The facts that $\mathbf{E}_X \mathbf{E}_Z(C_4(X, Z)) = 0$ and that $\mathbf{E}_X(\mathbf{E}_Z(C_3(X, Z)^2) + 2\mathbf{E}_Z(C_6(X, Z))) = 0$ follow from first replacing the even powers of z by the appropriate moments of the standard normal distribution, and then finding

(again using Mathematica) explicit anti-derivatives of $e^{g(x)}C_4$ and $e^{g(x)}((C_3)^2 + 2C_6)$, respectively, which are of the form $e^{g(x)}$ times a polynomial in derivatives of $g(x)$, and thus clearly approach 0 as $x \rightarrow \pm\infty$. (Note: the *existence* of $\mathbf{E}_X(\mathbf{E}_Z(C_3(X, Z)^2))$, and hence of $\mathbf{E}_X\mathbf{E}_Z(C_6(X, Z))$, follows since all moments of π_n exist.)

We now construct the sequence of sets F_n on which we uniformly control each of the four terms in the expansion (excluding just the C_3 term). The only thing to be ensured is that $\lim_{n \rightarrow \infty} n^{1/3}\pi_n(F_n^c) = 0$. For $j = 4, 5, 6$, set $C_j(x) = \mathbf{E}_Z C_j(x, Z)$, and $V_j(x_i) = \mathbf{Var}(C_j(x_i, Z))$. Because of the polynomial restrictions on g , C_j and V_j are bounded by polynomials. Now,

$$\mathbf{E} \left(\left(\sum_{i=2}^n C_j(x_i, Z_i) - \mathbf{E}_X(C_j(X)) \right)^2 \right) = \sum_{i=2}^n V_j(x_i) + \left(\sum_{i=2}^n (C_j(x_i) - \mathbf{E}_X(C_j(X))) \right)^2$$

So setting

$$F_{n,j} = \left\{ x; \left| \sum_{i=2}^n (C_j(x_i^n) - \mathbf{E}_X(C_j(X))) \right| < n^{5/8} \right\} \cap \left\{ x; \left| \sum_{i=2}^n (V_j(x_i^n) - \mathbf{E}_X(V_j(X))) \right| < n^{6/5} \right\}$$

it is easy to show by Markov's inequality applied to moments of the constrained functions that $n^{1/3}\pi_n(F_{n,j}^c) \rightarrow 0$ as $n \rightarrow \infty$, and for $x \in F_{n,j}$, $\mathbf{E} \left(\left(\sum_{i=2}^n C_j(x_i, Z_i) - \mathbf{E}_X(C_j(X)) \right)^2 \right) \leq O(n^{5/4})$. L^1 convergence of the 4th 5th and 6th terms in the Taylor expansion is thus assured.

It remains to consider $C_7(x_i^n, Z_i, \sigma_n)$. However, by using the remainder formula of the Taylor series expansion, and again using (2.2), we derive the bound

$$|C_7(x_i^n, Z_i, \sigma_n)| \leq n^{-7/6} p(x_i^n, w)$$

for some polynomial p with either $0 \leq w_i \leq Z_i$ or $Z_i \leq w_i \leq 0$. Now we can always take such a bounding polynomial to be of the form $A(1 + x_i^N)(1 + w^N)$ for sufficiently large A and for a sufficiently large even integer N . Since we can bound this polynomial in turn by $A(1 + x_i^N)(1 + Z_i^N)$ and since all polynomial moments of Z_i exist, it follows that we can write

$$\mathbf{E}_Z |C_7(x_i^n, Z, \sigma_n)| \leq n^{-7/6} p(x_i^n)$$

for a suitable polynomial $p(\cdot)$. Then, setting $u_7 = \mathbf{E}_X(p(X))$, $v_7 = \mathbf{Var}_X(p(X))$, and

$$F_{n,7} = \left\{ \mathbf{x}^n \in \mathbf{R}^n ; \left| \frac{1}{n} \sum_{i=1}^n p(x_i^n) - u_7 \right| < 1 \right\},$$

Chebychev's inequality implies that

$$\pi_n(F_{n,7}^C) \leq v_7 n^{-1}.$$

Furthermore, for $\mathbf{x}^n \in F_{n,7}$,

$$\sum_{i=2}^n \mathbf{E}_Z |C_7(x_i^n, Z, \sigma_n)| \leq (u_7 + 1)n^{-1/6}.$$

We put $F_n = F_{n,4} \cap F_{n,5} \cap F_{n,6} \cap F_{n,7}$. On F_n , terms 4, 5, 6 and 7 converge uniformly on L^1 , and it follows that the third term must also. Thus the last statement of the lemma also holds. \blacksquare

The main point of this lemma is that, in the log-expansion of the proposal density components, the terms corresponding to $n^{-1/6}$ and $n^{-1/3}$ vanish, and the next three terms each have vanishing expectation. (This is in contrast to with the situation for random-walk Metropolis, in which no terms cancel and only the first has vanishing expectation.)

To continue, we define \tilde{G}_n by

$$\tilde{G}_n V(\mathbf{x}^n) = n^{1/3} \mathbf{E} \left((V(\mathbf{Y}) - V(\mathbf{x}^n)) \left(\prod_{i=2}^n \frac{f(y_i) \exp\left(-\frac{1}{2\sigma_n^2}(x_i^n - y_i - \frac{\sigma_n^2}{2}g'(y_i))^2\right)}{f(x_i^n) \exp\left(-\frac{1}{2\sigma_n^2}(y_i - x_i^n - \frac{\sigma_n^2}{2}g'(x_i^n))^2\right)} \wedge 1 \right) \right).$$

In fact, \tilde{G}_n is like G_n except that the product omits the factor corresponding to $i = 1$. The following theorem shows that this omission is unimportant.

Theorem 4. *There exists sets $S_n \subseteq \mathbf{R}^n$ with $n^{1/3}\pi_n(S_n^C) \rightarrow 0$ such that for any $V \in C_c^\infty$,*

$$\lim_{n \rightarrow \infty} \sup_{\mathbf{x}^n \in S_n} \left| G_n V(\mathbf{x}^n) - \tilde{G}_n V(\mathbf{x}^n) \right| = 0.$$

Moreover,

$$\lim_{n \rightarrow \infty} \sup_{\mathbf{x}^n \in S_n} \mathbf{E} \left| \left(\frac{\pi_n(\mathbf{Y})q_n(\mathbf{Y}, \mathbf{x}^n)}{\pi_n(\mathbf{x}^n)q_n(\mathbf{x}^n, \mathbf{Y})} \wedge 1 \right) - \left(\prod_{i=2}^n \frac{f(y_i) \exp\left(-\frac{1}{2\sigma_n^2}(x_i^n - y_i - \frac{\sigma_n^2}{2}g'(y_i))^2\right)}{f(x_i^n) \exp\left(-\frac{1}{2\sigma_n^2}(y_i - x_i^n - \frac{\sigma_n^2}{2}g'(x_i^n))^2\right)} \wedge 1 \right) \right| = 0$$

Proof. Since the function $x \mapsto e^x \wedge 1$ has Lipschitz constant 1, and since $Y_1 = x_1^n + \sigma_n Z + \frac{1}{2}\sigma_n^2 g'(x_1^n)$, where $Z \sim N(0, 1)$, it follows that

$$\left| G_n V(\mathbf{x}^n) - \tilde{G}_n V(\mathbf{x}^n) \right| \leq n^{1/3} \mathbf{E} (|V(\mathbf{Y}) - V(\mathbf{x}^n)| |R(x_1^n, Z, \sigma_n)|)$$

where

$$R(x, z, \sigma) = \frac{1}{2} \left[z^2 - \left(z + \frac{\sigma}{2} g'(x) + \frac{\sigma}{2} g'(x + \frac{\sigma^2}{2} g'(x) + \sigma z) \right)^2 \right] + g(x + \frac{\sigma^2}{2} g'(x) + \sigma z) - g(x).$$

By a first-order Taylor series expansion in σ , as in the argument for the proof of Lemma 3, with the integral form of the remainder, we obtain

$$|R(x, z, \sigma_n)| \leq M_1(x) M_2(z) n^{-1/3} \quad (5.2)$$

for suitable positive polynomials $M_1(\cdot)$ and $M_2(\cdot)$.

Finally, since $V \in C_c^\infty$, there exists a constant K_1 such that

$$|V(\mathbf{Y}) - V(\mathbf{x}^n)| \leq K_1 |Y_1 - x_1^n|.$$

Furthermore, since $|Y_1 - x_1^n| \leq \sigma_n |Z| + \frac{\sigma_n^2}{2} |g'(x_1^n)|$, and recalling that g' is assumed to be Lipschitz, we can write $g'(x) \leq K_2(1 + |x|)$ for a constant $K_2 \geq 1$. It follows that

$$n^{1/3} |V(\mathbf{Y}) - V(\mathbf{x}^n)| |R(x_1^n, Z, \sigma_n)| \leq K_1 K_2 (1 + |x|) M_1(x) \left(\sigma_n |Z| M_2(Z) + \frac{\sigma_n^2}{2} M_2(Z) \right). \quad (5.3)$$

We now set S_n to be the set on which $M_1(x_1^n) \leq n^{1/12}$. By (5.3), and recalling that $Z \sim N(0, 1)$ so that $M_2(Z)$ and $|Z| M_2(Z)$ are integrable, and using Markov's inequality, we see that

$$\pi_n(S_n^C) = \pi_n \left((M_1(x_1^n))^5 \geq n^{5/12} \right) \leq n^{-5/12} \mathbf{E} \left((M_1(x_1^n))^5 \right).$$

The first result follows. The proof of the second is virtually identical, except that the estimation of $V(\mathbf{Y}) - V(\mathbf{x}^n)$ is not necessary. \blacksquare

Lemma 5. *There exist sets $T_n \subseteq \mathbf{R}^n$ with $n^{1/3}\pi_n(T_n^C) \rightarrow 0$ such that for any $V \in C_c^\infty$,*

$$\lim_{n \rightarrow \infty} \sup_{\mathbf{x}^n \in T_n} \left| n^{1/3} \mathbf{E}(V(Y_1) - V(x_1^n)) - \frac{\ell^2}{2} (V''(x_1^n) + g'(x_1^n)V'(x_1^n)) \right| = 0.$$

Proof. We write $Y_1 = x_1^n + \sigma_n Z + \frac{1}{2}\sigma_n^2 g'(x_1^n)$, and $W(x, z, \sigma) = V(x + \sigma z + \frac{\sigma^2}{2}g'(x))$. A second-order Taylor series expansion with respect to σ_n then gives that

$$\mathbf{E}(V(Y_1) - V(x_1^n)) = \frac{\sigma_n^2}{2} (V''(x_1^n) + g'(x_1^n)V'(x_1^n)) + \mathbf{E} \left(\int_0^{\sigma_n} \frac{\partial^3 V}{(\partial \sigma)^3}(x_1^n, z, \epsilon) \frac{(\sigma_n - \epsilon)^2}{2} d\epsilon \right).$$

Therefore, by (2.2), and an argument for the remainder term similar to that used in the proof of Lemma 3, there exists a polynomial $M_3(\cdot)$ such that the remainder term is less than $n^{-1/2}M_3(x_1^n)$. Letting T_n be the set on which $M_3(x_1^n) \leq n^{1/12}$, the result follows by Markov's inequality as in the previous lemma. \blacksquare

To proceed, we make some further definitions. Let $a(x) = -\frac{1}{4}g'(x)g''(x)$ and $b(x) = -\frac{1}{12}g'''(x)$, so that with $C_3(x, z)$ as in Lemma 3 we have

$$C_3(x, z) = l^3(a(x)z + b(x)z^3).$$

Set $Q_n(\mathbf{x}^n; \cdot) = \mathcal{L}(n^{-1/2} \sum_{i=2}^n C_3(x_i^n, Z_i))$ and let $\phi_n(\mathbf{x}^n; t) = \int e^{itw} Q_n(dw)$ be the corresponding characteristic function. Finally, let $\phi(t) = e^{-t^2 K^2/2}$ be the characteristic function of the distribution $N(0, K^2)$, with K as in Theorem 1.

Lemma 6. *There exists a sequence of sets $H_n \subseteq \mathbf{R}^n$ such that*

(1)

$$\lim_{n \rightarrow \infty} n^{1/3}\pi_n(H_n^C) = 0;$$

(2) For all $t \in \mathbf{R}$,

$$\lim_{n \rightarrow \infty} \sup_{\mathbf{x}^n \in H_n} |\phi_n(\mathbf{x}^n; t) - \phi(t)| = 0;$$

(3) For all bounded continuous functions r ,

$$\lim_{n \rightarrow \infty} \sup_{\mathbf{x}^n \in H_n} \left| \int_{\mathbf{R}} Q_n(\mathbf{x}^n, dy) r(y) - \frac{1}{\sqrt{2\pi\ell^6 K^2}} \int_{\mathbf{R}} r(y) \exp\{-y^2/\ell^6 K^2\} dy \right| = 0$$

where K is as in Theorem 1.

(4)

$$\lim_{n \rightarrow \infty} \sup_{\mathbf{x}^n \in H_n} \left| \mathbf{E}_Z \left(1 \wedge \exp \left(n^{-1/2} \sum_{i=2}^n C_3(x_i^n, Z_i) - \ell^6 K^2/2 \right) \right) - 2\Phi(-\ell^3 K/2) \right| = 0.$$

Proof. We define H_n as a region on which certain functionals have average value close to their mean. Specifically, we let H_n be the set of $\mathbf{x}^n \in \mathbf{R}^n$ such that

$$\left| \frac{1}{n} \sum_{i=2}^n h(x_i^n) - \int h(x)f(x)dx \right| \leq n^{-1/4}$$

and

$$|h(x_i^n)| \leq n^{3/4}, \quad 1 \leq i \leq n, \quad (5.4)$$

for each of the functionals $h(x) = a(x)^2, b(x)^2, a(x)b(x), a(x)^4, b(x)^4, a(x)^3b(x), a(x)^2b(x)^2, a(x)b(x)^3$.

Statement (1) now follows from Chebychev's inequality together with (2.2) and (2.3).

Assuming (2) for the moment, statement (3) follows by the continuity theorem for characteristic functions (applied to an arbitrary sequence of $\{\mathbf{x}^n; n = 1, 2, \dots\} \in H_1 \times H_2 \times \dots$).

Statement (4) then follows since if $R \sim N(-\alpha, 2\alpha)$, then $\mathbf{E}(1 \wedge e^R) = 2\Phi(-\sqrt{\alpha/2})$ (cf. Roberts, Gelman, and Gilks, 1994, Proposition 2.5), and furthermore $w \mapsto 1 \wedge e^w$ is a bounded functional.

It remains to prove (2). Our proof is a quantitative modification of the standard proof of the Lindeberg Central Limit Theorem (cf. Durrett, 1991, pp. 98-99).

Taking $\{\mathbf{x}^n\}$ to be a fixed sequence in $H_1 \times H_2 \times \dots$, we set $W_i = C_3(x_i^n, Z_i)$, set

$$v(x_i^n) = \mathbf{Var}_Z(W_i) = \ell^6(a(x_i^n)^2 + 6a(x_i^n)b(x_i^n) + 15b(x_i^n)^2),$$

and decompose $\phi_n(\mathbf{x}^n, t) = \prod_{i=2}^n \theta_i^n(x_i^n, t)$ as a product of characteristic functions of $n^{-1/2}W_i$. Note that by (5.4), for any $t \in \mathbf{R}$, we have $\frac{t^2}{2n}v(x_i^n) \leq 1$ for sufficiently large n . Hence, using equation (3.6) on page 85 of Durrett (1991), for any $\epsilon > 0$, we have

$$\begin{aligned} \left| \theta_i^n(x_i^n, t) - \left(1 - \frac{t^2}{2n}v(x_i^n) \right) \right| &\leq \mathbf{E}_Z \left(\frac{|t|^3}{n^{3/2}} \frac{|W_i|^3}{3!} \wedge \frac{2t^2}{n} \frac{|W_i|^2}{2!} \right) \\ &\leq \mathbf{E}_Z \left(\frac{|t|^3}{n^{3/2}3!} |W_i|^3; |W_i| \leq n^{1/2}\epsilon \right) + \frac{t^2}{n} \mathbf{E}_Z \left(|W_i|^2; |W_i| > n^{1/2}\epsilon \right) \\ &\leq \frac{\epsilon |t|^3}{6n} \mathbf{E}_Z (|W_i|^2) + \frac{t^2}{\epsilon^2 n^2} \mathbf{E}_Z (|W_i|^4). \end{aligned}$$

Hence, since $\mathbf{x}^n \in H_n$, and using Lemma (4.3) on page 94 of Durrett (1991),

$$\begin{aligned} \left| \phi_n(\mathbf{x}^n; t) - \prod_{i=2}^n \left(1 - \frac{t^2}{2n} v(x_i^n) \right) \right| &\leq \sum_{i=2}^n \left(\frac{\epsilon |t|^3}{6n} \mathbf{E}_Z (|W_i|^2) + \frac{t^2}{\epsilon^2 n^2} \mathbf{E}_Z (|W_i|^4) \right) \\ &\leq \frac{K^2 + 22n^{-1/4}}{6} \ell^6 \epsilon |t|^3 + \frac{t^2}{\epsilon^2 n} \left(\xi + 14868 \ell^{12} n^{-1/4} \right). \end{aligned}$$

where $\xi = \mathbf{E}_X \mathbf{E}_Z (|W_i|^4)$.

Given $\delta > 0$, we choose ϵ small enough to make the first term less than $\delta/2$, and then choose n large enough to make the second term less than $\delta/2$, to get that

$$\left| \phi_n(\mathbf{x}^n; t) - \prod_{i=2}^n \left(1 - \frac{t^2}{2n} v(x_i^n) \right) \right| < \delta.$$

On the other hand,

$$\begin{aligned} &\left| \prod_{i=2}^n \left(1 - \frac{t^2}{2n} v(x_i^n) \right) - e^{-t^2 \ell^6 K^2 / 2} \right| \\ &\leq \left| e^{-t^2 \ell^6 K^2 / 2} - e^{-t^2 \sum_{i=2}^n v(x_i^n) / 2n} \right| + \left| \prod_{i=2}^n \left(1 - \frac{t^2}{2n} v(x_i^n) \right) - \prod_{i=2}^n e^{-t^2 v(x_i^n) / 2n} \right|. \end{aligned}$$

Now, the first term goes to 0 uniformly for $\mathbf{x}^n \in H_n$. Also, by Lemma 4.3 on page 94 of Durrett (1991), the second term is bounded above by $\sum_{i=2}^n t^4 v^2(x_i^n) / (4n^2)$, which goes to 0 uniformly for $\{\mathbf{x}^n\}$ sequences (such that $\mathbf{x}^n \in H_n$) since the individual terms of $v^2(x_i^n)$ converge uniformly to their respective limits. The result follows. \blacksquare

Proof of Theorem 1. Recalling that

$$a_n(\ell) = \mathbf{E} \left(\frac{\pi_n(\mathbf{y}) q_n(\mathbf{y}, \mathbf{x})}{\pi_n(\mathbf{x}) q_n(\mathbf{x}, \mathbf{y})} \wedge 1 \right),$$

Theorem 1 follows directly from the second statement in Theorems 3 and 4 and from part (4) of Lemma 6. \blacksquare

Proof of Theorem 2. We take $F_n^* = H_n \cap S_n \cap T_n \cap F_n$. Then $\lim_{n \rightarrow \infty} n^{1/3} \pi_n(F_n^{*C}) = 0$, and $\lim_{n \rightarrow \infty} \mathbf{P}(\Gamma_t^n \in F_n^*, 0 \leq t \leq T) = 1$ for any fixed T . Also, from Lemma 3, Theorem 4, and Lemmas 5 and 6, it follows that

$$\lim_{n \rightarrow \infty} \sup_{\mathbf{x}^n \in F_n^*} |G_n V(\mathbf{x}^n) - G V(\mathbf{x}^n)| = 0$$

for all $V \in C_c^\infty$ which depend only on the first coordinate. Therefore, as discussed at the beginning of this section, using Corollary 8.7 of Chapter 4 of Ethier and Kurtz (1986), the weak convergence in Theorem 2 follows.

Finally, to prove the statement about maximizing $h(\ell)$, we note that this problem amounts to finding the value $\hat{\ell}$ of ℓ which maximizes the function $2\ell^2\Phi(-K\ell^3/2)$, and then evaluating $\hat{a} = a(\hat{\ell}) = 2\Phi(-K\hat{\ell}^3/2)$. Making the substitution $u = K\ell^3/2$ shows this is the same as finding the value \hat{u} of u which maximizes $2^{5/3}K^{-2/3}u^{2/3}\Phi(-u)$, and then evaluating $\hat{a} = 2\Phi(-\hat{u})$. It follows that the value of \hat{u} , and hence also the value of \hat{a} , does not depend on the value of K (provided $K > 0$), so it suffices to take $K = 2$. For $K = 2$ we find (again using Mathematica) that $\hat{\ell} \doteq 0.82515$, so that $\hat{a} \doteq 0.57424$. This completes the proof of Theorem 2. ■

REFERENCES

- 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.
- R. Durrett (1991), *Probability: Theory and Examples*. Wadsworth & Brooks, Pacific Grove, California.
- S.N. Ethier and T.G. Kurtz (1986), *Markov processes, characterization and convergence*. Wiley, New York.
- A. Gelman, G.O. Roberts, and W.R. Gilks (1994), Efficient Metropolis jumping rules. Research Rep. No. 94-10, Statistical Laboratory, University of Cambridge.
- U. Grenander and M.I. Miller (1994), Representations of knowledge in complex systems (with discussion). *J. Roy. Stat. Soc. B* **56**, 549-604.
- W.K. Hastings (1970), Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97-109.
- A.D. Kennedy and B. Pendleton (1991), Acceptances and autocorrelations in hybrid Monte Carlo. *Nuclear Phys. B (Proc. Suppl.)* **20**, 118-121.
- 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.
- R.D. Mountain and D. Thirumalai (1994), Quantative measure of efficiency of Monte Carlo simulations. *Physica A* **210**, 453-460.

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.

D.B. Phillips and A.F.M. Smith (1994), Bayesian model comparison via jump diffusions. Tech. Rep. **94-20**, Imperial College London.

G.O. Roberts, A. Gelman, and W.R. Gilks (1994), Weak convergence and optimal scaling of random walk Metropolis algorithms. Research Rep. No. 94-16, Statistical Laboratory, University of Cambridge.

G.O. Roberts and R.L. Tweedie (1995), Exponential Convergence of Langevin Diffusions and their discrete approximations. Preprint.

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

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.

S. Wolfram (1988), *Mathematica: A system for doing mathematics by computer*. Addison-Wesley, New York.