Optimal scaling of discrete approximations to Langevin diffusions

by

Gareth O. Roberts
Statistical Laboratory
University of Cambridge

and

Jeffrey S. Rosenthal
Department of Statistics
University of Toronto

Technical Report No. 9507, June 19, (1995)

TECHNICAL REPORT SERIES

University of Toronto
Department of Statistics

Optimal scaling of discrete approximations to Langevin diffusions

by

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

(June, 1995.)

Acknowledgements. We thank Michael Miller, Radford Neal, and Richard Tweedie for helpful discussions. This work was supported in part by EPSRC of the U.K. and 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).

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 $O(n^{-1})$ as the dimension $n \to \infty$. The paper thus provided a useful heuristic for running Metropolis

* 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`.

algorithms efficiently. However, this result does not apply to more general Hastings algorithms. It is clear that if the proposal density makes use of the structure of the target density, then a higher acceptance probability is likely to be preferred.

In this paper we propose 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) 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.

Our main results may be summarized as follows. For discrete approximations to Langevin diffusions for certain target distributions π, the optimal asymptotic acceptance probability in high dimensions is approximately 0.574. Furthermore, the proposal variance should scale as $n^{-1/3}$. Therefore, Langevin algorithms are considerably more efficient than random-walk based Metropolis methods; the optimal proposal variances and acceptance probabilities are both substantially larger. However, care must be taken in the interpretation of these results, since Langevin algorithms can have unfortunate properties such as sub-geometric rate of convergence; see for example Roberts and Tweedie (1995).

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. A graph of the speed of the algorithm, as a function of the acceptance probability, is given in Figure 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 generalizations are possible; see 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). 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 though simulations in Mountain and Thirumalai (1994). (We are very grateful to Radford Neal for bringing these references to our attention.)

Our formal definitions are as follows. The reversible Langevin diffusion for the n-dimensional density π_n, with unit time scaling, 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. However, such discrete approximations can have vastly different asymptotic behaviours from the diffusion process they attempt to approximate (Roberts and Tweedie, 1995). Therefore it is necessary to introduce a Metropolis step (Metropolis et al., 1953; Hastings, 1970) which serves to preserve the stationarity of π_n. Specifically, 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

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

a fixed probability distribution on $\mathbf{R}^n$. We further assume that

$$|g(x)|, |g'(x)|, |g''(x)|, |g'''(x)| \leq M_0(x) \tag{$*$}$$

for some polynomial $M_0(\cdot)$, and that

$$\int_{\mathbf{R}} x^k f(x)\, dx \ < \ \infty, \qquad k = 1, 2, 3, \ldots . \tag{$**$}$$

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 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*

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

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

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

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

Theorem 2. *For any positive integer* c, *let* $\{\mathbf{U}^n\}_{t\geq 0} = \left(\Gamma^n_{t,1}, \ldots, \Gamma^n_{t,c}\right)$ *be the process corresponding to the first* c *components of* $\mathbf{\Gamma}^n$. *Then as* $n \to \infty$, *the process* $\mathbf{U}^n$ *converges weakly (in the Skorokhod topology) to the Langevin diffusion* $\mathbf{U}$ *defined by*

$$d\mathbf{U}_t = (h(\ell))^{1/2}\, d\mathbf{B}_t + \frac{1}{2}h(\ell)\nabla \log \pi_c(\mathbf{U}_t)\, dt\,,$$

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).*

This theorem 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 d_n. Then setting $\ell_n = d_n^{1/2}n^{1/6}$, the theorem says that the speed (and hence 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 d_n 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 d_n should be modified accordingly.

3. Theorem proofs.

We first note that the assumed structure of π_n implies that the Langevin diffusion process $\{\mathbf{U}_t\}$ will consist of asymptotically independent components. It thus suffices (cf. Roberts, Gelman, and Gilks, 1994, Proposition 2.1) to consider the case in which $c = 1$.

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,

$$\mathbf{P}(\mathbf{X}_n \in F_n^* \text{ for all } 0 \leq s \leq t) \to 1 \qquad (****)$$

and

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

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 $n \to \infty$ (and hence $\sigma_n^2 \to 0$), 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. *Setting* $Y_i = x_i^n + \sigma_n Z_i + \frac{\sigma_n^2}{2} g'(x_i^n)$ *(so that* Z_i *is distributed as standard normal), and recalling that* $q(x,y) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma_n^2}(y - x - \frac{\sigma_n^2}{2} g(x))^2\right)$, *there exists a sequence of sets* $F_n \in \mathbf{R}^n$, *with* $\lim_{n\to\infty} n^{1/3} \pi_n(F_n^C) = 0$, *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* $\mathbf{E}_X((E_Z(C_3(X,Z)))^2) = \ell^6 K^2 = -2\mathbf{E}_X\mathbf{E}_Z(C_6(X,Z)) > 0$. *In addition,*

$$\lim_{n\to\infty} n^{-2/3} \sup_{\mathbf{x}^n \in F_n} \mathbf{E}_Z\left|\sum_{i=1}^n C_4(x_i^n, Z)\right| = 0;$$

$$\lim_{n\to\infty} n^{-5/6} \sup_{\mathbf{x}^n \in F_n} \mathbf{E}_Z\left|\sum_{i=1}^n C_5(x_i^n, Z)\right| = 0;$$

$$\lim_{n\to\infty} n^{-5/6} \sup_{\mathbf{x}^n \in F_n} \mathbf{E}_Z\left|\sum_{i=1}^n C_6(x_i^n, Z) - \mathbf{E}(C_6(X,Z))\right| = 0;$$

$$\lim_{n\to\infty} \sup_{\mathbf{x}^n \in F_n} \mathbf{E}_Z\left|\sum_{i=1}^n C_7(x_i^n, Z, \sigma_n)\right| = 0.$$

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 $\mathbf{E}_X((\mathbf{E}_Z(C_3(X,Z)^2) + \mathbf{E}_Z(C_6(X,Z))/2)) = 0$ 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 $e^{g(x)}((C_3)^2 + C_6/2)$, respectively, which are of the form $e^{g(x)}$ times a polynomial in derivatives of $g(x)$, and thus clearly approach 0 as $x \to \pm\infty$. (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.)

For $j = 4,5,6$, setting $C_j(x) = \mathbf{E}_Z C_j(x,Z)$, and $v_j = \mathbf{E}_X((C_j(X) - \mathbf{E}_X(C_j(X)))^4)$, $(*)$ and $(**)$ imply that $v_j < \infty$. We set

$$F_{n,j} = \{x \in \mathbf{R}^n \, ; \, n^{1/3}\mathbf{E}_Z|\frac{1}{n}\sum_{i=1}^{n} C_j(x_i^n, Z) - \mathbf{E}_X\mathbf{E}_Z(C_j(X,Z))| < n^{-1/24}\} .$$

Markov's inequality then implies that

$$\pi_n(F_{n,j}^C) \leq 3v_j n^{-1/2} .$$

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 $(*)$, it is easy to derive the bound

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

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

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

Putting $F_n = F_{n,4} \cap F_{n,5} \cap F_{n,6} \cap F_{n,7}$, the stated conclusions follow. ■

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 to be compared 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

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

Intuitively, $\tilde{G}_n$ is like G_n but where 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 $n^{1/3}\pi_n(S_n^C) \to 0$ such that for any $V \in C_c^\infty$,*

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

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

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

where

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

By a first-order Taylor series expansion in d, with the integral form of the remainder, we obtain that

$$|R(x,z,\sigma)| \le \sigma \int_0^\sigma \left|\frac{\partial^2 R}{(\partial\sigma)^2}(x,z,\epsilon)\right| d\epsilon.$$

By condition $(*)$, bounding the derivative $\frac{\partial^2 R}{(\partial\sigma)^2}$ and factoring the exponential terms, this implies that

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

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

Since $V \in C_c^\infty$, and again using $(*)$, we can find suitable positive polynomials $M_3(\cdot)$ and $M_4(\cdot)$, such that

$$|V(\mathbf{Y}) - V(\mathbf{x}^n)| \le M_3(x_1^n)M_4(Z). \tag{2}$$

We now set S_n to be the set on which $M_1(x_1^n)M_3(x_1^n) \leq n^{1/12}$. Putting equations (1) and (2) together, recalling that $Z \sim N(0,1)$ so that $M_2(Z)M_4(Z)$ is integrable, and using Markov's inequality, we see that

$$\pi_n(S_n^C) = \pi_n\left((M_1(x_1^n)M_3(x_1^n))^5 \geq n^{5/12}\right) \leq n^{-5/12}\mathbf{E}\left((M_1(x_1^n)M_3(x_1^n))^5\right).$$

The result follows. ∎

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

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

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

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

Therefore, by $(*)$, there exists a polynomial $M_5(\cdot)$ 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. ∎

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

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

Set $Q_n(\mathbf{x}^n;\cdot) = \mathcal{L}\left(n^{-1/2}\ell^3\sum_{i=2}^n C_3(x_i^n, Z_i)\right)$ and let $\phi_n(\mathbf{x}^n;t) = \int e^{itw}Q_n(dw)$ be the corresponding characteristic function. Finally, let $\phi(t) = e^{-t^2K^2/2}$ 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)

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

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

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

(3)

$$\lim_{n\to\infty} \sup_{\mathbf{x}^n \in H_n} \|Q_n(\cdot) \;-\; N(0,\ell^6 K^2)\|_{\mathrm{var}} \;=\; 0,$$

where K is as in Theorem 1.

(4)

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

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

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

and

$$|h(x_i^n)| \;\le\; n^{3/4}, \qquad 1 \le i \le n, \tag{$***$}$$

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

(1) now follows from Chebychev's inequality together with $(*)$ and $(**)$.

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

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).

We set $W_i = \ell^3 C_3(x_i^n, Z_i)$, 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 $(***)$, 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

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

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

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

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,

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

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 $\sum_{i=2}^n \left|1 - \frac{t^2}{2n} - e^{-t^2\ell^6K^2/2n}\right|$, which is bounded above by $\sum_{i=2}^{n} \frac{\ell^{12} K^4 t^4}{4n^2}$ and hence goes to 0 uniformly for $\mathbf{x}^n \in H_n$. The result follows. ∎

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 Theorem 4 and from parts (3) and (4) of Lemma 6. ∎

Proof of Theorem 2. We take $F_n^* = H_n \cap S_n \cap T_n \cap F_n$. Then from Lemma 3, Theorem 4, and Lemmas 5 and 6, it follows that $(****)$ holds, and furthermore

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

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 $2^{5/2} K^{-2/3} u^{2/3} \Phi(-u)$, 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

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 (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, Univeristy 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.

Figure 1. Asymptotic relative efficiency in terms of acceptance rates