Methodology and Computing in Applied Probability, 4, 337–357, 2002

Langevin Diffusions and Metropolis-Hastings Algorithms

G. O. ROBERTS
Department of Mathematics and Statistics, Lancaster University, Lancaster, LA1 4YF, England

O. STRAMER
Department of Statistics and Actuarial Science, University of Iowa, Iowa City IA 52242, USA

Received April 23, 2002; Accepted December 11, 2002

Abstract. We consider a class of Langevin diffusions with state-dependent volatility. The volatility of the diffusion is chosen so as to make the stationary distribution of the diffusion with respect to its natural clock, a heated version of the stationary density of interest. The motivation behind this construction is the desire to construct uniformly ergodic diffusions with required stationary densities. Discrete time algorithms constructed by Hastings accept reject mechanisms are constructed from discretisations of the algorithms, and the properties of these algorithms are investigated.

Keywords: MCMC, Langevin diffusions and algorithms

AMS 2000 Subject Classification: 65C40

1. Introduction

Recent interest in Langevin diffusions and their discretely simulated counterparts has been generated largely by their use as Markov chain Monte Carlo (MCMC) techniques (see, for example, Doll *et al*., 1978; Roberts and Tweedie, 1996; Stramer and Tweedie, 1999a, 1999b). Since the main motivation for this work is in MCMC, interest focuses largely on the stability of the stationary distribution of the processes concerned and robustness properties (such as geometric ergodicity) of their convergence properties to stationarity.

This paper will investigate the theoretical properties of three types of stochastic processes related to Langevin diffusions. Firstly we shall consider the properties of the diffusions themselves. As noted first in Roberts and Tweedie (1996) and later in Stramer and Tweedie (1999), properties of discrete approximations to these continuous time diffusions can be radically different from those of the original diffusion. The properties of these discretizations are investigated here.

One way of ensuring that at least the stationary distribution is stable to discretization is to introduce a Metropolis-Hastings accept/reject step. We also investigate the properties of these algorithms.

In Section 2, we discuss the convergence behavior of the general diffusions, and in Section 3, the more specific tempered diffusions are analyzed. Section 4 analyzes the discretized diffusions obtained without using accept reject mechanisms, and in Section 5, we introduce the corresponding Metropolis-Hastings algorithms. Section 6 considers the behavior of the various algorithms introduced in a multimodal context, and in Section 7 a Bayesian analysis of a multinomial logit model is carried out using the methods we introduce. Throughout, we give simulations to illustrate aspects of our methods' performance.

2. Diffusions: General Results

2.1. Definitions

We assume that π is a continuous density function on $\mathbb{R}^n$, which we know only up to a constant of proportionality. More precisely, we shall assume either that we know a function $\pi_u(x) = k\pi(x)$ for some unknown constant $k > 0$, or that we have available $\nabla \log \pi = \nabla \log \pi_u$, where ∇ is the usual differential operator $(\nabla f(x))_i = df/dx_i$. We also assume that π has locally uniformly Holder continuous[1] with second partial derivatives. We consider a broad class of diffusions on $\mathbb{R}^n$ which have a given stationary distribution with density π. Such a diffusion is defined as a solution to the stochastic differential equation

$$dX(t) = b(X(t))dt + \sigma(X(t))dB(t), \qquad X(0) = x \in \mathbb{R}^n, \tag{1}$$

where B is an n-dimensional Brownian motion, $a(x) = \sigma(x)\sigma'(x)$ is an $n \times n$ symmetric positive definite matrix with entries $\partial^2 a_{ij}/\partial x_k \partial x_l$ which are locally uniformly Holder continuous on $\mathbb{R}^n$, and

$$b_i(x) = \frac{1}{2}\sum_{j=1}^{n} a_{ij}(x)\partial \log \frac{\pi(x)}{\partial x_j} + \delta^{1/2}(x)\sum_{j=1}^{n}\frac{\partial}{\partial x_j}\left(a_{ij}(x)\delta^{-1/2}(x)\right),$$

where $\delta(x) = \det a(x)$. It is well known that X has π as its unique invariant measure (see Kent, 1978) so long as X is non-explosive. Criteria for non-explosion of diffusions are given in Stroock and Varadhan (1979).

We call X satisfying (1) a L diffusion (Langevin diffusion) for π, with scaling σ. The LC diffusion (Langevin diffusion with constant variance coefficient) takes σ to be cI, where $c > 0$ is a constant and I is the $n \times n$ identity matrix.

Another important special case of L diffusions for π, is obtained by choosing the diffusion matrix $a(x) = \sigma(x)\sigma'(x)$ as $a(x) = \pi_u^{-2d}(x)I$, where $0 \le d \le 1/2$. From (1),

$$b(x) = \frac{1-2d}{2}a(x)\nabla \log \pi_u(x). \tag{2}$$

We call these processes LT diffusions (Langevin tempered algorithms). In Section 3 we shall motivate this special case and the reason for calling these diffusions ''tempered''.

2.2. General Convergence Results

As in Meyn and Tweedie (1993) and Down *et al.* (1995), we formally define V-uniform ergodicity, when $V \ge 1$ is a measurable function on $\mathbb{R}^n$, by requiring that for all $x \in \mathbb{R}^n$

$$\|P_X^t(x,\cdot) - \pi\|_V \le V(x)R\rho^t, \qquad t \ge 0, \tag{3}$$

for some $R < \infty, \rho < 1$, where $P_X^t(x,A) = P(X_t \in A | X_0 = x), t \ge 0$. We call $X(t)$ exponentially ergodic if it is V-uniformly ergodic for some such V.

We now give sufficient conditions for the diffusion X_t defined as in (1) to be V-exponentially ergodic.

THEOREM 2.1 *Let $X(t)$ be defined as a solution to (1). If there exists $S > 0$ such that $|\pi(x)|$ is bounded for $|x| \ge S$, then $X(t)$ is V-uniformly ergodic for a $V \ge 1$ that is twice continuously differentiable if*

$$\mathscr{L}_V \le -cV + b\mathbb{1}_C, \tag{4}$$

for some constants $b, c > 0$, and some compact non-empty set C, where

$$\mathscr{L}_V(x) := \sum b_i(x)\frac{\partial V(x)}{\partial x_i} + \frac{1}{2}\sum_{i,j} a_{i,j}(x)\frac{\partial^2 V(x)}{\partial x_i \partial x_j}, \tag{5}$$

is the mean velocity of $V(X(t))$ at $X(t) = x$.

Proof: The proof follows directly from Meyn and Tweedie (1993) and using a similar argument to the proof of Theorem 2.1 in Roberts and Tweedie (1996). ■

3. Langevin Tempered (LT) Diffusions

3.1. Motivation

We now motivate the choice of $a(x) = \pi_u^{-2d}(x)I$ for some $0 \le d \le 1/2$ as a diffusion matrix. The diffusion X can be thought of as a time change of a tempered diffusion Z, defined by

$$dZ_t = \frac{1-2d}{2}\nabla \log \pi_u(Z_t)dt + dB_t, \tag{6}$$

which is the simple Langevin diffusion for the tempered (heated) density $\pi_d(x) \propto \pi^{1-2d}(x)$. It can be shown (see Rogers and Williams, 1987, p. 175) that $X_t \equiv Z_{\tau(t)}$ where $\tau(t) = \{s > 0 : \varphi_s > t\}$ and $\varphi_s = \int_0^s \pi_u^d(Z_s)ds$.

Thus X is the diffusion process satisfying the SDE

$$dX_t = \frac{1-2d}{2}\pi_u^{-2d}(X_t)\nabla \log \pi_u(X_t) + \pi_u^{-d}(X_t)dB_t. \tag{7}$$

Continuing the analogy with annealing or tempering, $(1-2d)^{-1}$ plays the role of a temperature, since Z from (6) has stationary distribution $\pi(z)^{(1-2d)}$. Heated Markov chains typically have better convergence properties than the ordinary unheated chains, for instance reducing the worst affects of multimodality. Thus X compensates for the higher temperature, and therefore the correspondingly disproportionately large time spent in areas of low density, by speeding up in these areas (see Gelatt *et al.*, 1983, Marinari and Parisi, 1992, and Neal, 1996).

Certain special cases are worth noting. The case $d = 0$ is, of course, the simple Langevin diffusion. The other extreme is the case $d = 1/2$. This is the infinite temperature case, so that Z is just Brownian motion, with no invariant probability measure. X satisfies

$$dX_t = c\pi^{-1/2}(X_t)dB_t, \tag{8}$$

for some $c > 0$, so is actually a local martingale.

For the one-dimensional case, we can justify our choice of $a(x)$ in a different way. Suppose we are interested in constructing a family of diffusions which are all non-explosive and uniformly ergodic, at least for as large a class of target densities as possible. It turns out to be more straightforward to do this on bounded domains since in this case it is natural to take $a(x) \equiv 1$. Therefore consider the following class of transformations which map $(-\infty, \infty)$ to a bounded interval. Let

$$g_d(x) = \int_{-\infty}^{x} \pi^d(z)dz, \tag{9}$$

and denote the inverse of g_d by h_d. Our aim will be to construct a uniformly ergodic diffusion with invariant measure, $\pi(g_d(\cdot))$ before transforming back to the infinite domain via the function h_d. Let $d > 0$ so that for a large family of target densities, $g_d(\infty) < \infty$ at least for a suitable collection of possible values for d. In the sequel we make this assumption.

If U is any random variable with density π then $V = g_d(U)$ has density

$$\tilde{\pi}(y) = \pi^{1-2d}(h_d(y)), \qquad y \in (g_d(-\infty), g_d(\infty)), \tag{10}$$

and in particular if Y is a diffusion with stationary distribution $\tilde{\pi}$, it follows that $h_d(Y)$ is a diffusion with stationary distribution π.

Consider the simple Langevin diffusion for $\tilde{\pi}$, which satisfies the following SDE (at least on $g_d(-\infty) < y < g_d(\infty)$):

$$dY_t = b(Y_t)dt + dB_t, \tag{11}$$

where

$$b(y) = \frac{1}{2}\frac{d(\log \tilde{\pi}(y))}{dy} = \frac{(1-2d)}{2}\nabla \log \pi(h_d(y)).$$

Now, letting $X_t = h_d(Y_t)$, a simple application of Ito's formula gives that $\{X_t\}$ is the LT diffusion defined as in (1) with $a(x) = \pi^{-2d}(x)$, and $b(x)$ is defined as in (2).

For concreteness, we shall make the following assumption for the one-dimensional case which is not strictly necessary in general, but which simplifies the exposition. When this assumption is weakened, correspondingly weaker versions of the results that follow are available but are not pursued here.

$$\int_{\mathbb{R}} \pi_u^s(x)dx < \infty \quad \text{for} \quad s > 0. \tag{12}$$

We now explain why we have only allowed d to lie in the interval $[0, 1/2]$.

LEMMA 3.1 *Let $X(t)$ be a one-dimensional diffusion satisfying (1) with $b(\cdot)$ as in (2). Assuming (12), $X(t)$ is non-explosive if and only if* $0 \leq d \leq 1/2$.

Proof: The question of non-explosivity is explored using the time-changed diffusion with unit volatility (a diffusion viewed through its natural clock). Assume also that $\tau(t) = \{u > 0 : \varphi_u > t\}$, where $\varphi_u = \int_0^u \pi_u^{-2d}(s)ds$. Then, $Z_t \equiv X_{\tau(t)}$ satisfies the SDE

$$dZ_t = \left(\frac{1-2d}{2}\nabla \log \pi_u(Z_t)\right)dt + dB_s, \tag{13}$$

which is non-explosive if and only if

$$\int_y^{\infty} \pi_u^{2d-1}(x)dx = \infty \quad \text{and} \quad \int_{-\infty}^{y} \pi_u^{2d-1}(x)dx = \infty, \tag{14}$$

for some (and hence for all) y (see, for example, Karatzas and Shreve, 1991 or Rogers and Williams, 1987). Hence, by (12), (14) holds if $d \leq 1/2$ and does not hold if $d > 1/2$. ■

This result is hardly surprising on reference to the sign of the drift in (2), since for $d > 1/2$, the process actually drifts away from the modes of the distribution.

LEMMA 3.2 *Let X be defined as in Lemma 3.1 and let $Y = g_d(X)$, where g_d is defined as in (9). Then, under the assumption that (12) holds, X and Y are uniformly ergodic if* $0 < d \leq 1/2$.

Proof: For a one-dimensional diffusion on a finite domain, Y, uniform ergodicity corresponds to showing that 0 and $g_d(\infty)$ are entrance boundaries. However, this is clear by scale and speed arguments for one-dimensional diffusions, see for example Sections 5.46 and 5.47 of Rogers and Williams (1987). ■

3.2. Exponential Rates of Convergence for LT Diffusions

We now give sufficient conditions for LT diffusions to be V-exponentially ergodic.

THEOREM 3.3 *Let $X(t)$ be a LT diffusion for a given unnormalized density π_u.*

A. *If there exists $S > 0$ such that $|\pi_u(x)|$ is bounded for $|x| \geq S$ and $0 < r < 1-2d$ such that*

$$\liminf_{|x|\to\infty} \pi_u^{-2d}(x)[((1-2d)-r)|\nabla \log \pi_u(x)|^2 + \nabla^2 \log \pi_u(x)] > 0, \tag{15}$$

then the process is exponentially ergodic with $V = \pi^{-r}$. (For f: $\mathbb{R}^n \to \mathbb{R}, \nabla^2 f = \sum_{i=1}^n \partial^2/\partial x_i^2 f$.)

B. *If there exists a positive definite matrix B such that for some $D > 0$ and $R > 0$*

$$\pi^{-2d}(x)\left[2\left(Bx, \frac{1}{2}\nabla \log \pi(x)(1-2d)\right) + tr(B)\right] \leq -D(Bx, x) \text{ for } \|x\| > R, \tag{16}$$

then the process is exponentially ergodic with $V = (Bx, x) + 1$.

Proof: If we choose the test function $V = \pi^{-r}, 0 < r < 1 - 2d$, then from the definition of $\mathscr{L}_V(x)$, (5) we have that

$$2\mathscr{L}_V(x) \propto \pi_u^{-2d}(x)V(x)\Big[(r^2 - r(1-2d))|\nabla \log \pi_u(x)|^2 - r\nabla^2 \log \pi_u(x)\Big].$$

(4) follows now directly from (15) so that by Theorem 2.1 the diffusion is exponentially ergodic, thus proving A.

For B, if we choose the test function $V(x) = (Bx, x) + 1$, then (4) follows now directly from (16).

Statement B now follows directly from Theorem 2.1. ∎

EXAMPLE 1 **The Multidimensional Exponential Class** P_m: We consider the exponential family P_m introduced and studied in Roberts and Tweedie (1996a, 1996b), and consisting of sufficiently smooth densities with the form (at least for large $|x|$)

$$\pi(x) \propto \mathrm{e}^{-p(x)}, \tag{17}$$

where p is a polynomial of degree m of the following type. Decompose p as $p = p_m + q_{m-1}$ where q_{m-1} is a polynomial of degree $\leq m-1$, and p_m consists of only the full degree terms in p. Then we say that $\pi \in P_m$ if $p(x) \to \infty$ as $|x| \to \infty$: this is a positive definiteness condition, and we note that this condition requires that $m \geq 2$, and that m is even.

We now show that if $X(t)$ is the LT diffusion with $0 < d < 1/2$ for $\pi \in P_m$, then $X(t)$ is exponentially ergodic. As noted in Roberts and Tweedie (1996b), by the positive definiteness condition

$$\liminf_{|x| \to \infty} \frac{|\nabla \log \pi(x)|^2}{|\nabla^2 \log \pi(x)|} = \infty$$

and

$$\liminf_{|x| \to \infty} ((1-2d) - r)|\nabla \log \pi(x)|^2 > 0,$$

for all $0 < d < 1/2, 0 < r < 1 - 2d$. We also note that $\lim_{|x| \to \infty} \pi_u^{-2d}(x) = \infty$. Thus condition A of Theorem 3.3 holds and $X(t)$ is exponentially ergodic.

EXAMPLE 2 **Multivariate *t* distribution:** Suppose that $\pi \sim t_\nu(\mu, \Sigma)$, the multivariate t distribution with $\nu > 2$ degrees of freedom, location $\mu = (\mu_1, \ldots, \mu_n)$ and symmetric positive definite $n \times n$ scale matrix Σ that is,

$$\pi(x) \propto (\nu + (x-\mu)^T \Sigma^{-1}(x-\mu))^{-(\nu+n)/2}, \qquad x \in \mathbb{R}^n. \tag{18}$$

From Theorem 2.4 of Roberts and Tweedie (1996a) we have that the LC diffusion is not exponentially ergodic since $|\nabla \log(x)| \to 0$ when $|x| \to \infty$.

For LT diffusions with $a(x) = \nu + (x-\mu)^T \Sigma^{-1}(x-\mu) \propto \pi^{-2d}$, and $b(x) = -(\nu+n)/2(1 - 2/\nu + n)\Sigma^{-1}(x-\mu)$ we can easily show that the LT diffusion is exponentially ergodic when $\nu + n > 2$, (16) holds with $B = I$. Thus, with $d = 1/\nu + n > 0$, we obtain exponential convergence to the density π.

4. Discretizations

In practice, in simulating the diffusion sample path we cannot follow the dynamic defined by Equation (1) exactly, but must instead discretize Equation (1). Our interest in this section is to consider the effects of this discretization on the ergodicity properties of the resulting discrete Markov chain.

4.1. Euler Discretization

The natural discretization of a L diffusion for π, with scaling σ is the Euler approximation $\{E_n\}$, defined as follows:

$$E_{n+1} = E_n + b(E_n)h + \sigma(E_n)h^{1/2}Z_{n+1}, \tag{19}$$

where $h > 0$ is a suitably small constant, and $\{Z_i, i \in \mathbb{Z}_+\}$ are independent $N(\mathbf{0}, I)$ random variables. We call $\{E_n\}$ satisfying (19) a LE discretization for a L diffusion. The LEC (LET) discretization is the Euler discretization for the LC (LT) diffusion.

4.2. Ozaki Discretization

Stramer and Tweedie (1996a) suggests the use of discretization schemes as proposed by Ozaki and Shoji (Ozaki, 1992 and Shoji and Ozaki, 1998). For the drift term, the Ozaki approximation represents a higher order approximation than the Euler scheme.

The Ozaki algorithm described in Shoji and Ozaki (1998) represents a linear approximation of the diffusion drift b, together with a constant approximation of the volatility σ over each small time interval $kh \leq t < (k+1)h, k = 0, 1, \ldots$. A Taylor expansion over the time interval $(kh, (k+1)h)$ is used to obtain that $b(X(t)) \approx b(X(kh)) + J(X(kh))(X(t) - X(kh))$ where $J(x) = \partial(b_1, \ldots, b_n)/\partial(x_1, \ldots, x_n)$ is the Jacobian of $b(x)$. It is assumed that $J(\cdot)$ is not zero, and that it is continuous through the remainder of the paper.

Thus for $t \in (kh, (k+1)h)$ with h small,

$$b(X(t)) \approx J(X(kh))X_t + c(X(kh)); \quad \sigma(X(t)) \approx \sigma(X(kh)), \tag{20}$$

where $c(X(kh)) = b(X(kh)) - J(X(kh))X(kh)$. Let $\{O_t\}$ be a solution to the linear stochastic differential equation,

$$d(O(t)) = (J(O(kh))O(t) + c(O(kh)))dt + \sigma(O(kh))dW(t) \quad kh \leq t < (k+1)h. \tag{21}$$

This can be solved explicitly, leading to a time-homogeneous diffusion approximation in continuous time, or a Markov chain if we consider the process $\{O_n\}_{n=1}^{\infty}$ which is defined as $O_n = O(nh)$. It is easy to check that the transition distribution $Q_h(x, \cdot)$ of O_{n+1} given $O_n = x, x \in \mathbb{R}^n$, is normal with mean $\mu_{x,h}$ and covariance matrix $a_{x,h}$ defined as follows:

$$\mu_{x,h} = x + J^{-1}(x)[\exp(J(x)h) - I]b(x),$$
$$a_{x,h} = \int_0^h \exp\{J(x)(u)\}a(x)\exp\{J'(x)(u)\}du. \tag{22}$$

It is shown in Shoji and Ozaki (1998) that if $J(x)$ has no pair of reverse-sign eigenvalues, (i.e., if λ is an eigenvalue of $J(x)$, then $-\lambda$ is not an eigenvalue of $J(x)$) then $a_{x,h}$ is the unique solution to the linear matrix equation

$$J(x)a_{x,h} + a_{x,h}J'(x) = \exp\{J(x)h\}a(x)\exp\{J'(x)h\} - a(x), \tag{23}$$

which simplifies to

$$a_{x,h} = \frac{1}{2}a(x)J^{-1}(x)[\exp(2J(x)h) - I], \tag{24}$$

under the condition that

$$(J(x)a_{x,h})' = a_{x,h}J(x). \tag{25}$$

We note that (25) is satisfied under our assumption that $a(x)$ is symmetric. Thus, for LC and LT diffusions we can use the explicit solution (24) for $a_{x,h}$ rather than the inexplicit solution (23) for $a_{x,h}$.

We call LO, LOC, and LOT discretizations the Ozaki discretizations for L, LC, and LT diffusions respectively.

4.3. Geometric Convergence of Discretizations

We now consider convergence properties of LO discretizations. We use the usual inner product notation $(x, y) = \sum_{i=1}^{n} x_i y_i$. Our standard methodology will be to use drift function techniques as described in Meyn and Tweedie (1993).

THEOREM 4.1 *Let $\{O_t\}$ be the Ozaki discretization as in (21). Assume that*

1. *The eigenvalues of $J(x) + J'(x)$ are all less than or equal to $-\lambda < 0$ for $|x| > M, M > 0$.*
2. *$|c(x)| = |b(x) - J(x)x|$ is bounded.*
3. *$tr(\sigma(x)\sigma'(x))$ is bounded.*

Then the LO discretization is geometrically ergodic with $V(x) = |x|^2 + 1$, for all $h > 0$.

Proof: It is easy to check that $\{O_n\}$ is μ^{Leb}-irreducible and from Proposition 6.1.2 of Meyn and Tweedie (1993) it is weak Feller. Hence, from Theorem 15.0.1 of Meyn and Tweedie (1993) it suffices for geometric ergodicity to find a test function $V \geq 1$ such that for some compact set C and some $\alpha < 1, b < \infty$

$$\int Q_h(x, dy)V(y) \leq \alpha V(x) + b\,\mathbb{1}_C(x). \tag{26}$$

Although the transition distribution of $\{O_k, k = 0, 1, 2, \ldots\} = \{O(kh), k = 0, 1, 2, \ldots\}$ is explicitly available from (22), it turns out to be more convenient to work directly with (21) in trying to derive a statement such as (26).

Now we shall write $\mathscr{L}_{V,x}(y)$ for the continuous time generator of the linear SDE given by (21) on the time interval $kh < t < (k+1)h$, conditional on $O_k = x$, applied to the function V at the point y. We shall use the drift function $V(y) = |y|^2 + 1$. From (21) it is easy to check that

$$\mathscr{L}_{V,x}(y) = (2y, J(x)y + c(x)) + tr(\sigma(x)\sigma'(x)).$$

Next note that for all $y \in \mathbb{R}^n$,

$$\begin{aligned}(2y, J(x)y + c(x)) = y'(J(x)|J'(x))y + 2y'c(x) &\leq -2\lambda|y|^2 + 2y'c(x)\\ &\leq -2\lambda|y|^2 + d|y|,\end{aligned}$$

for some constant $d > 0$. Thus

$$\mathscr{L}_{V,x}(y) \leq -\lambda_0\|y\|^2 + K, \tag{27}$$

for some $K > 0$, and $\lambda_0 > 0$. Moreover, this statement can be made uniformly in x for x outside a suitably large compact set. By the continuity of J, we can assume in fact that (27) holds uniformly for all x and y (with a possibly inflated value for K).

Standard martingale arguments now apply to (27) (as in Meyn and Tweedie, 1993) to give a statement of the form (26) with $\alpha = \mathrm{e}^{-h\lambda_0}$ and $b = K/\lambda_0$. ■

The regularity conditions needed for the application of Theorem 4.1 are very strong, essentially restricting the class of possible target distributions to have approximately Gaussian tails. The current proof does not readily generalize beyond this case, but it is clear that the result ought to hold in far greater generality.

In the next example we show that geometric ergodicity for the LE discretizations depends more on h than for the LO discretizations.

EXAMPLE 3 **Gaussian tails:** We consider a special case of the exponential family P_m introduced in Example 1. We assume that π has Gaussian tails and compare the LEC discretization with the LOC discretization. From Theorem 4.1, LOC is exponentially ergodic for all $h > 0$ while LEC is not always exponentially ergodic as is illustrated by the following simple example. Let π be the density of bivariate normal distribution, defined as,

$$\pi(x) \propto \exp\left(-\frac{x^T\Sigma^{-1}x}{2}\right), \qquad x \in \mathbb{R}^2, \tag{28}$$

where

$$\Sigma = \begin{bmatrix} 0.001 & 0 \\ 0 & 9 \end{bmatrix}.$$

Then,

$$E_{n+1}|E_n = x \sim N\left(x - \frac{\Sigma^{-1}xh}{2}, hI\right), \tag{29}$$

where $\{E_n\}$ is the LEC discretization and

$$O_{n+1}|O_n = x \sim N(\mu_{x,h}, a_{x,h}) \tag{30}$$

where

$$\mu_{x,h} = x + \Sigma\left(\exp\left(-\frac{\Sigma^{-1}h}{2}\right) - I\right)\Sigma^{-1}x, \quad a_{x,h} = -\Sigma\big(\exp(-\Sigma^{-1}h) - I\big)$$

and $\{O_n\}$ is the LOC discretization.

From Theorem 3.1 (b) in Roberts and Tweedie (1996a) we have that the LEC discretization is transient when $h \geq 0.002$.

In contrast, from Theorem 4.1 the LOC discretization is geometric ergodic for all $h > 0$. In addition, the variance of the step size for the two components of the algorithm is different. If we choose h to be big enough, then as desired, the variance is ''small'' for the first component of the LOC discretization and bigger for the second component.

Note that the problems that Euler schemes encounter in sampling from target densities with very heterogenous scales can be examined theoretically, see Roberts and Rosenthal (2001).

5. Metropolis-Hastings Algorithms: Definitions and Results

In practice, the behavior of the discrete approximations to (1) may be very different from that of the diffusion, (see, for example, Roberts and Tweedie, 1996b for the Euler approximations). Thus we use the discrete approximation as a candidate Markov chain for the Metropolis-Hastings algorithm, to compensate for the discretization, and ensure that π retains its status as the correct stationary distribution.

We denote the transition kernel of the discrete approximation to the diffusion by $Q(x, \cdot), x \in \mathbb{R}^n$. A ''candidate transition'' to y, generated according to the density $q(x, y)$, is then accepted with probability $\alpha(x, y)$, given by

$$\alpha(x, y) = \begin{cases} \min\left\{\frac{\pi_u(y)}{\pi_u(x)}\frac{q(y,x)}{q(x,y)}, 1\right\} & \pi_u(x)q(x, y) > 0 \\ 1 & \pi_u(x)q(x, y) = 0. \end{cases} \tag{31}$$

Thus actual transitions of the M-H chain take place according to a law $P(x, \cdot)$ with transition densities $p(x, y) = q(x, y)\alpha(x, y), y \neq x$ and with probability of remaining at the same point given by

$$r(x) = P(x, \{x\}) = \int q(x, y)[1 - \alpha(x, y)]dy. \tag{32}$$

The crucial property of the M-H algorithm is that, with this choice of α, the target π is invariant for the operator P: that is, $\pi(A) = \int \pi(x)P(x, A)dx$ for all $x \in \mathsf{X}, A \in \mathscr{B}$.

We will call the Metropolized version of a LEC discretization HLEC (named MALA in

Roberts and Tweedie, 1996a) and the Metropolized version of a LO discretization HLO (named MADA in Stramer and Tweedie, 1996b).

Two key results that link the convergence properties of discretizations and Metropolized discretizations chains are brought together in the following result, see, Stramer and Tweedie (1996b) and Roberts and Tweedie (1996b).

THEOREM 5.1

a. *Suppose $\pi(x)$ is positive and continuous, and the transition density $q(x, y)$ is positive and continuous in both variables. Let P be the transition law of the Metropolized chain formed from Q. If $\alpha(x, y)$ is such that*

$$r(x) = P(x, \{x\}) \to 0, \qquad |x| \to \infty \tag{33}$$

then if Q is geometrically ergodic then P is geometrically ergodic.

b. *Suppose that ess sup $r(x) = 1$ (where the essential supremum is taken with respect to π), then the algorithm is not geometrically ergodic.*

EXAMPLE 3 **Gaussian tails: continuation** We again assume that π has Gaussian tails described by (28) and compared the HLEC algorithm with the HLOC algorithm. From Theorem 4.1 and Theorem 5.1 (a), HLOC is geometrically ergodic for all $h > 0$ and from Theorem 3.1 (b) in Roberts and Tweedie (1996a) and Theorem 5.1 (a), HLEC is geometrically ergodic for all $h < 0.0002$.

We assumed that π is defined as in (28). Figure 1(a) gives trace plots for the first (left) and second (right) components of the HLEC algorithm with $h = 0.0019$ and a starting point $(100, 100)'$. It shows that convergence rate is very slow since h is ''too small'' for the second component. Figure 1(b) give trace plots for the HLEC algorithm with $h = 0.005$ and a starting point $(0, 0)'$. It shows that convergence rate is slow since $h = 0.005$ is still not ''big enough'' for the second component. Figure 1(c) give trace plots for the HLEC algorithm with $h = 0.01$ and a starting point $(0, 0)'$. It shows poor convergence since $h = 0.01$ is now ''too big'' for the first component and hence rejection rate is big.

Figure 1(d) gives trace plots for the HLOC algorithm with $h = 10$ and a starting point $(100, 100)$. It shows that, in this case, when using the HLOC algorithm with $h = 10$, the sampler appears to settle down rapidly to approximate stationarity.

EXAMPLE 2 **Multivariate *t* distribution: continuation** Consider the multivariate t distribution (22) with $\nu = 5, n = 3, \mu = (0, 0, 10)'$ and $\Sigma = I$. We now consider the following two algorithms:

1. The HLEC algorithm with $a(x) \equiv I$ and $b(x) = -\nu + n/2(1 - 2/\nu + n)\Sigma^{-1}(x - \mu)$.
2. The HLOT algorithm with $a(x) = [5 + (x - \mu)^T \Sigma^{-1}(x - \mu)]I$ and $b(x) = -5 + 2/2(1 - 2/5 + 2)\Sigma^{-1}(x - \mu)$.

Five thousand steps were simulated of the third component of both algorithms with initial point $(-10, 20, -30)$. Trace plots and auto-correlation plots of HLEC algorithms and of HLOT algorithms with $h = 0.1, 0.5, 2.0$ are in Figure 2. It is clear that the performance of the HLEC algorithm depends more on h and the starting point of the algorithm than does the HLOT algorithm.

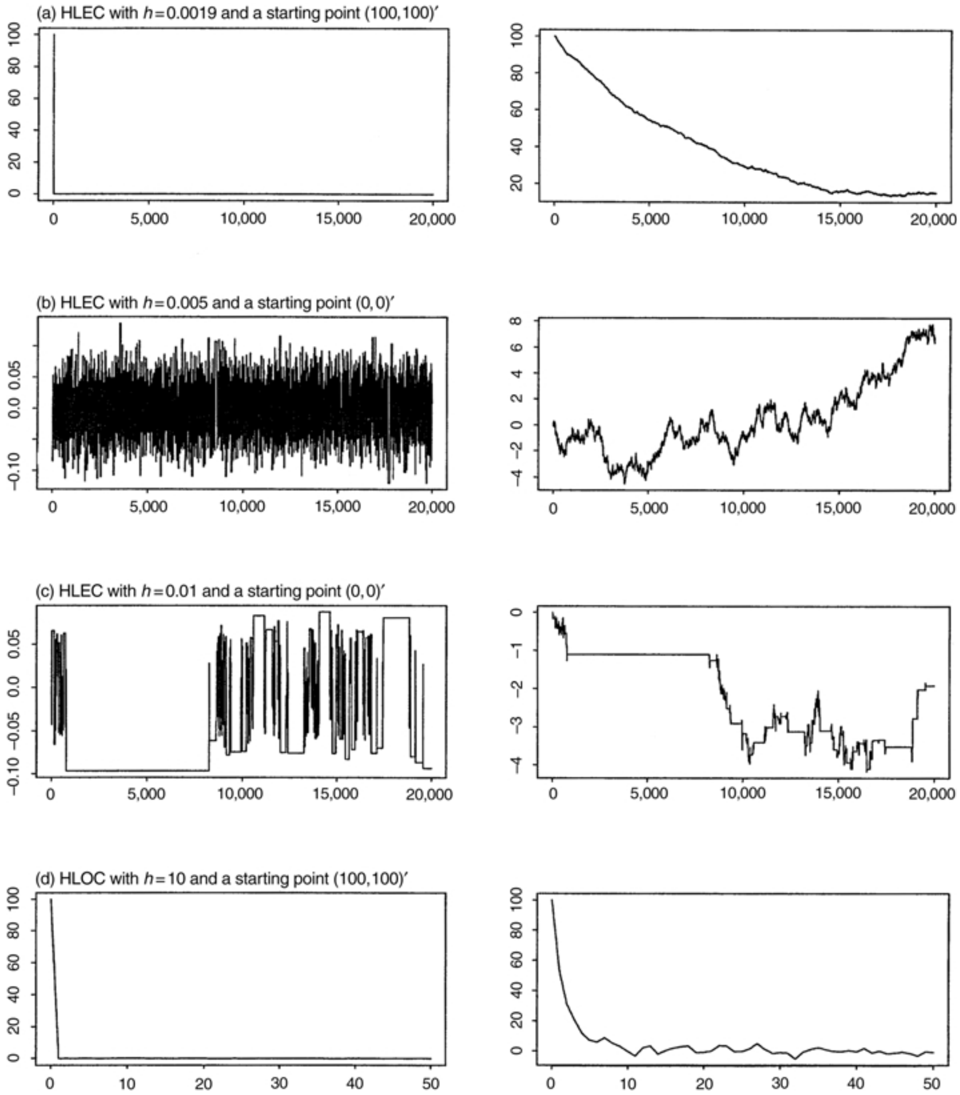

Figure 1. Bivariate normal density: Example 3 (continuation). Trace plots of the steps taken by the HLEC algorithms and the HLOC algorithm. Left: trace plots of the first component of the algorithm. Right: trace plots of the second component of the algorithm.

The computing time taken to run 5,000 iterations of the chains was 5 seconds for the HLEC algorithm and 8 seconds for the HLOT algorithm, using the language Ox developed by Doornik (1996) on a 400 MHz 64-bit RISC processor, HP workstation. For the HLOT algorithm we need to compute the exponent of a matrix. For more economical ways of computing a matrix exponential see Moler and Van Loan (1978) and Leonard (1996).

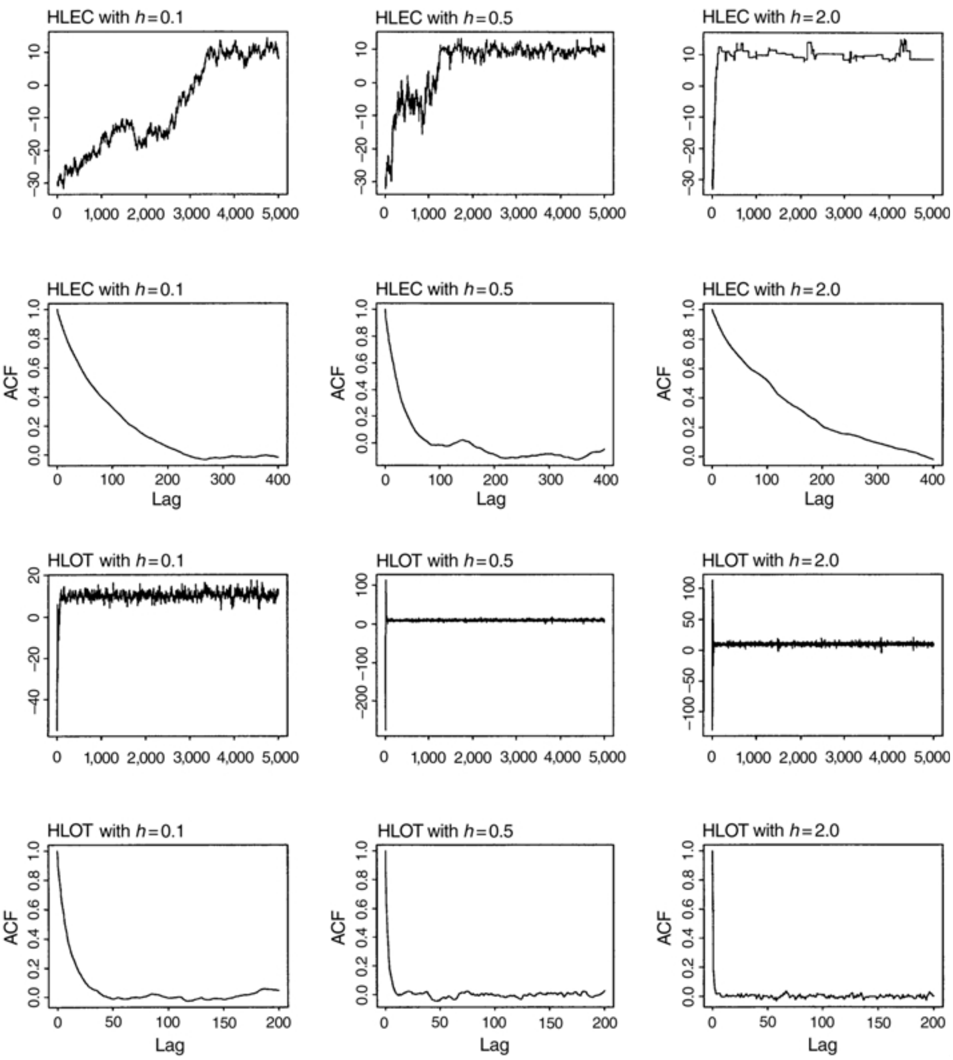

Figure 2. Trace plots and auto-correlation plots of the HLEC algorithm and the HLOT algorithm for the multidimensional t distribution. Example 2: continuation.

The HLEC appears to be unreliable even in simple examples like these. In Example 1, the algorithm is unable to cope with heterogeneous scales in different components, while in Example 2, it appears less robust to heavy tailed target distributions than its Ozaki competitor HLOT.

REMARK 5.2 In the one-dimensional case (see Stramer and Tweedie, 1999b), the use of heavy tailed proposal distributions for algorithms has distinct theoretical advantages (see Stramer and Tweedie, 1999b, Jarner and Roberts, website). It is reasonable to expect similar advantages in the multivariate setting. Here we might use multivariate t distribution with mean $\mu_{x,h}$ and variance $a_{x,h}$ defined in (22) as a candidate for the Metropolis-Hastings algorithm. The t distribution, having heavier tails than the normal distribution, is more appropriate as a proposal distribution for target densities with heavier tails. However, very little is known rigorously about the behavior of these algorithms apart from the one-dimensional case (which is investigated in Stramer and Tweedie, 1999b). Low dimensional examples of the use of heavy tailed proposals are given in Section 6.

6. Sampling from Multimodal Distributions

In practice the use of HLEC algorithms and HLOC algorithms to sample from multimodal distributions may cause problems since both algorithms will often pull the Markov chain to the closest mode. Thus, there is a need for other algorithms for estimating multi-modal distributions.

We consider the use of HLET algorithms, with $d = 1/2$; the drift of the LT diffusion is zero and the diffusion matrix is $\pi_u^{-1}(x)I$. Thus the algorithm performs like a random walk with a heterogenous increment distribution with covariance matrix $\pi_u^{-1}(x)hI$. The proposal variances are therefore ''small'' when the chain is ''close'' to one of the modes and larger when the chain is further away from the modes. However, while these algorithms increase the mobility of the chain, from Theorem 5.1 (b) they are not geometrically ergodic (at least on unbounded domains).

To improve convergence, we use hybrid MCMC strategies which alternate HLOC and HLET steps with $d = 1/2$.

Related methods of sampling from multimodal distributions include ''simulated tempering'' (see Marinari and Parisi, 1992) and the ''tempered transition'' methods (see Neal, 1996).

EXAMPLE 4 We studied the performance of diffusion algorithms on the following mixture of bivariate normal distributions,

$$\pi(x) \propto \exp[-(x-\mu_1/2)'\Sigma^{-1}(x-\mu_1)] + \exp[-(x-\mu_2/2)'\Sigma^{-1}(x-\mu_2)], \quad x \in \mathbb{R}^2$$

where $\Sigma = I, \mu_1 = (6,-5)'$ and $\mu_2 = (-2,3)'$.

To assess the behavior of these algorithms, we carried out the H-M algorithm with five choices of candidate.

1. The HLEC algorithm with $h = 6,\ 7,\ 8,\ 9$ and starting point $(0,0)$ (first row in Figure 3).

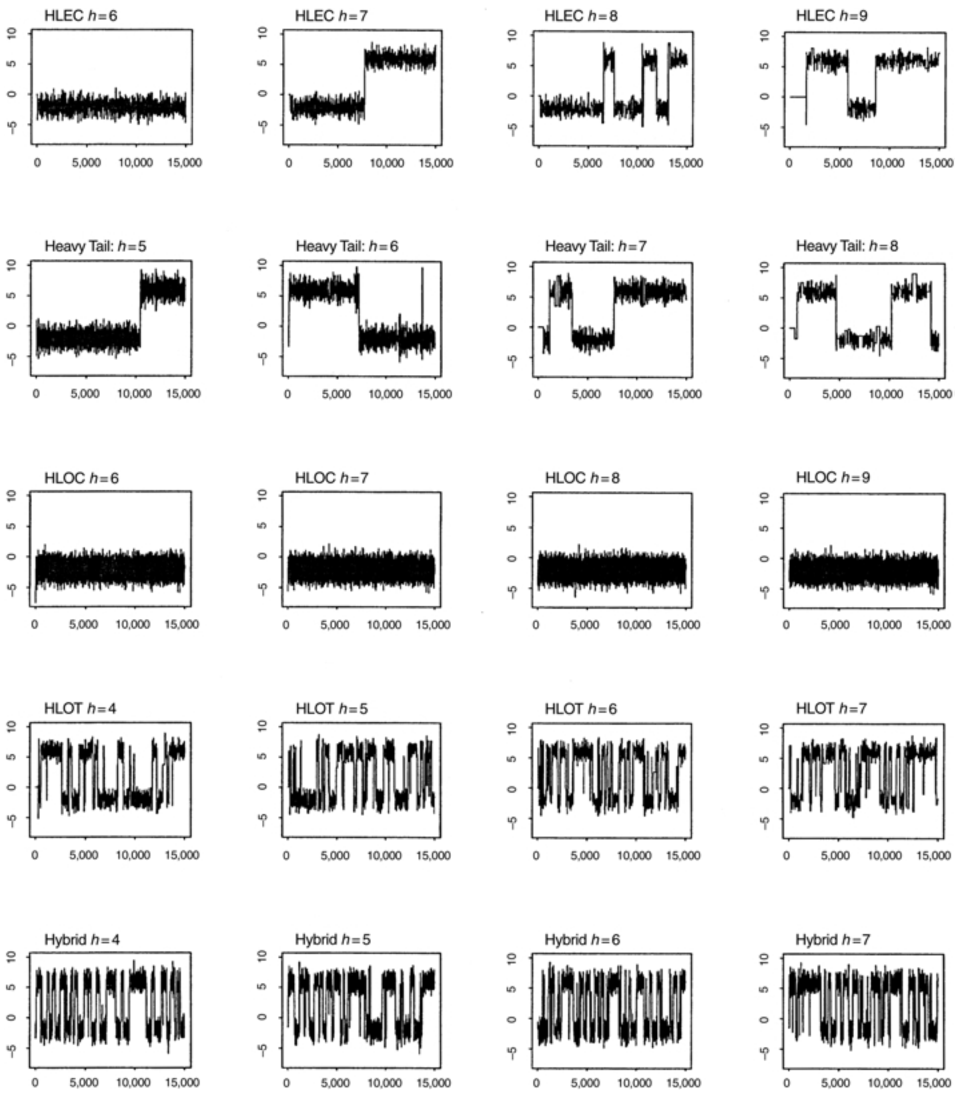

Figure 3. Langevin algorithms for the mixture of bivariate normal distribution. Trace plots of 15,000 steps of the first component of the HLEC algorithm with $h = 6, 7, 8, 9$ (first row), HLEC algorithm with t-distribution with $h = 5, 6, 7, 8$ (second row), HLOC algorithm with $h = 6, 7, 8, 9$ (third row), HLOT algorithm with $h = 4, 5, 6, 7$ (fourth row), and hybrid algorithm (fifth row).

2. The HLEC algorithm with multivariate $t_5(0, 3/5I)$-distribution, $h = 5, 6, 7, 8$ and starting point $(0, 0)$ (second row in Figure 3).
3. The HLOC algorithm with $h = 6, 7, 8, 9$ and starting point $(-100, -100)$ (third row in Figure 3).

4. The HLET algorithm with $d = 1/2, h = 4, 5, 6, 7$ and starting point $(0, 0)$ (fourth row in Figure 3).
5. The hybrid algorithm defined as follows: with probability 0.1 use the HLOC with $h = 7$ and with probability 0.9 use the HLOT algorithm with $d = 1/2$ and $h = 4, 5, 6, 7$ (fifth row in Figure 3).

The HLOC algorithm found the neighborhood of one mode started from $(-100, 100)$ very rapidly. However, it became stuck in the vicinity of one mode for a long period of time. The same results were obtained for other values of the parameter h. The HLEC algorithm with a starting point $(0, 0)$ and ''large'' h performed better than the HLOC algorithm though it still tended to ''stick'' in the vicinity of one mode for long periods of time. Worse results were obtained for other values of the parameter h. In contrast, the HLOT algorithm with $d = 1/2$ and the hybrid algorithm appeared to find both modes quite easily. The hybrid algorithm performed marginally better than the HLOT algorithm switching slightly more often between modes.

7. Multinomial Logit Model

We now illustrate our results for a Bayesian analysis of a multinomial logit model. We assume that a set C which includes all potential choices for some population can be defined and that all choices are distinct. Let N denote the sample size and A the number of choices and define, for $i = 1, \ldots, A$ and $n = 1, \ldots, N$.

$$Y_{in} = \begin{cases} 1 & \text{if observation } n \text{ choose alternative } i, \\ 0 & \text{otherwise.} \end{cases}$$

We assume that

$$(Y_{1n}, \ldots, Y_{An}) \sim \text{multinomial } ((p_n(1), \ldots, p_n(A)), 1),$$

where for a specific individual the ratio of the choice probabilities of any two alternatives is given by

$$\frac{p_n(i)}{p_n(j)} = \exp(b'(x_{in} - x_{jn})),$$

and $x_{in} = (x_{in}^1, \ldots, x_{in}^m)$ is a set of covariates associated with each alternative i and observation n. This implies that the likelihood function for a general multinomial choice model is

$$L(b|y) = \prod_{n=1}^{N} \prod_{i=1}^{A} P_n(i)^{y_{in}},$$

where

$$P_n(i) = P(Y_n(i) = 1) = \frac{\exp(b'x_{in})}{\sum_{j=1}^{A} \exp(b'x_{jn})}.$$

We assume that the prior for b is non-informative. Thus the posterior distribution $P(b|y) \propto L(b|y)$. We suggest the use of the HLOC chain obtained from the Ozaki approximation. It is easy to check that

$$\frac{\partial \log(P(b|y))}{\partial b} = \sum_{n=1}^{N} \sum_{i=1}^{A} y_{in} \left(x_{in} - \sum_{i=1}^{A} P_n(j) x_{in} \right),$$

and the Jacobian matrix $J(b)$ is

$$J(b) = -\sum_{n=1}^{N} \sum_{i=1}^{A} P_n(i) \left[x_{in} - \sum_{j=1}^{A} x_{jn} P_n(j) \right]' \frac{\left[x_{in} - \sum_{j=1}^{A} x_{jn} P_n(j) \right]}{2}.$$

We note that under the assumption that the $(NA) \times m$ matrices whose rows are $x_{in} - \sum_{j=1}^{A} x_{jn} P_n(j)$ for $i = 1, \ldots, A$ and $n = 1, \ldots, N$ is of rank m, J is the negative of a weighted moment matrix of the independent variables and hence is negative definite. Thus, from Theorem 4.1 and 5.1 the HLOC algorithm is geometrically ergodic.

We simulated 1,000 observations from multinomial logit models as follows. Let $b = [0.00, -0.95, 3.28, -5.62, 1.28, 2.68, 2.19, 1.73, 0.12, 2.03]^T$, where b is a random draw from $MVN(\mu, \Sigma)$, with $\mu = [0, 0, 0, 0, 0, 0, 0, 0, 0, 0]^T$, and $\Sigma = 9I$. We also assumed that $A = 6$, and generated $(y_{n1}, \ldots, y_{n6}), n = 1, \ldots, 1{,}000$ from multinomial $((P_n(1), \ldots, P_n(6)), 1)$.

To draw values from the posterior distribution of b, we used the HLOC algorithm and the HLEC algorithm with a starting point (20, 20, 20, 20, 20, 20, 20, 20, 20, 20). For the HLEC algorithm, we tuned h to be 0.01 to obtain acceptance rate close to 0.57 (see Roberts and Rosenthal, 1998, for the optimal scaling of HLEC algorithms). For the HLOC algorithm we chose h to be 0.1 and the acceptance rate was around 0.86. Figure 4 is a trace plot of the steps taken by the algorithms HLEC(i) and HLOC(i) for $i = 1, \ldots, 10$ at times $kh, k = 0, \ldots, 1{,}550$ for each i, where i denotes the i-th component of the chain. Figure 5 is a trace plot of the steps taken by the algorithms HLEC(i) and HLOC(i) for $i = 1, \ldots, 10$ at times $kh, k = 1{,}050, \ldots, 1{,}550$ for each i, where i denotes the i-th component of the chain.

The computing time taken to run 1,550 iterations of the chains was 4.4 hours for the HLEC algorithm and 5.3 hours for the HLOT algorithm, using Matlab on a 400 MHz 64-bit RISC processor, HP workstation.

HLOC performs better than HLEC here, but the advantage it gives is not large in this example.

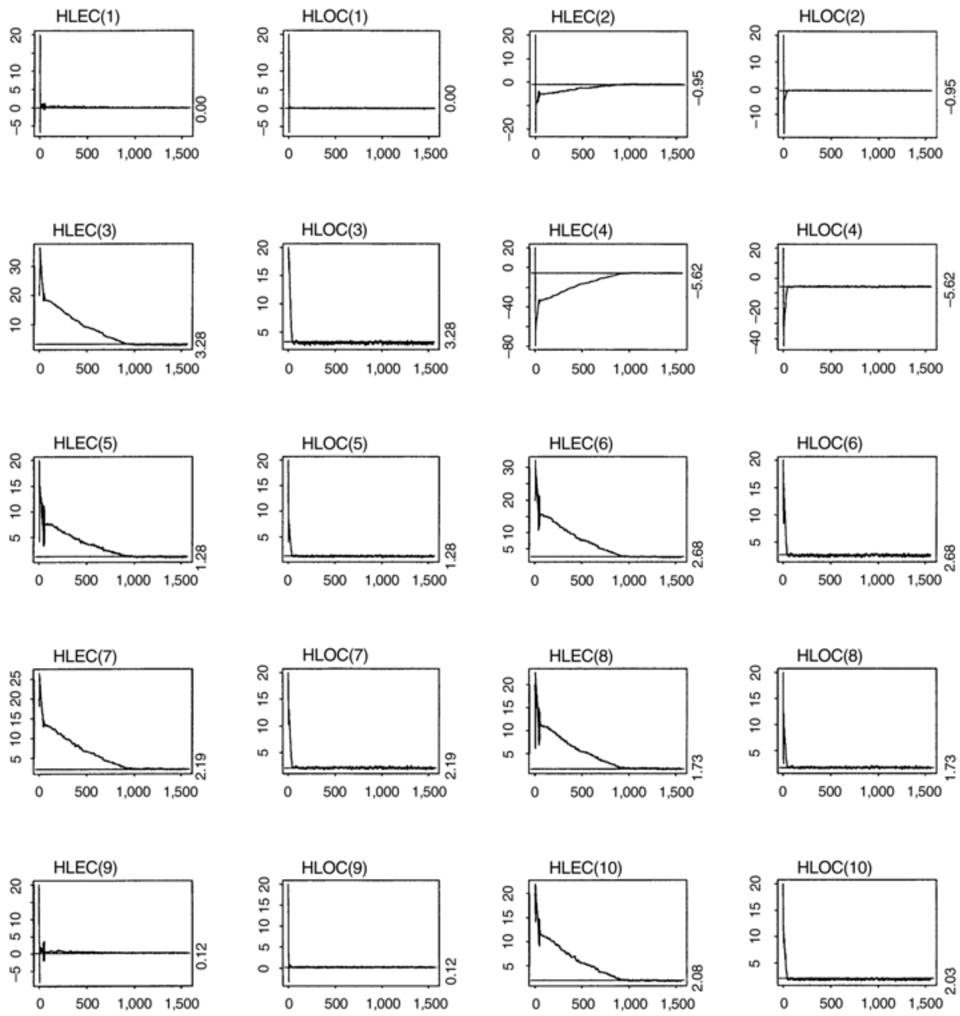

Figure 4. Trace plots of the HLEC algorithm and the HLOC algorithm. Bayesian analysis of the Multinomial logit model. The bracketed number refers to the parameter being plotted, so that for instance HLOC(i) gives a trace plot for the parameter $b(i)$ under the experiment using HLOC.

8. Conclusions

We have considered Langevin diffusions and their associated discretizations for a given target density π. Interest has largely been focused on the ergodicity and stability properties of the various processes due to the motivation from MCMC.

The Ozaki discretization provides a more stable alternative to the Euler method. This is shown in the context of a highly non-homogeneous target distribution (where it is able to adapt to different scales for different components) and also for MCMC algorithms as we

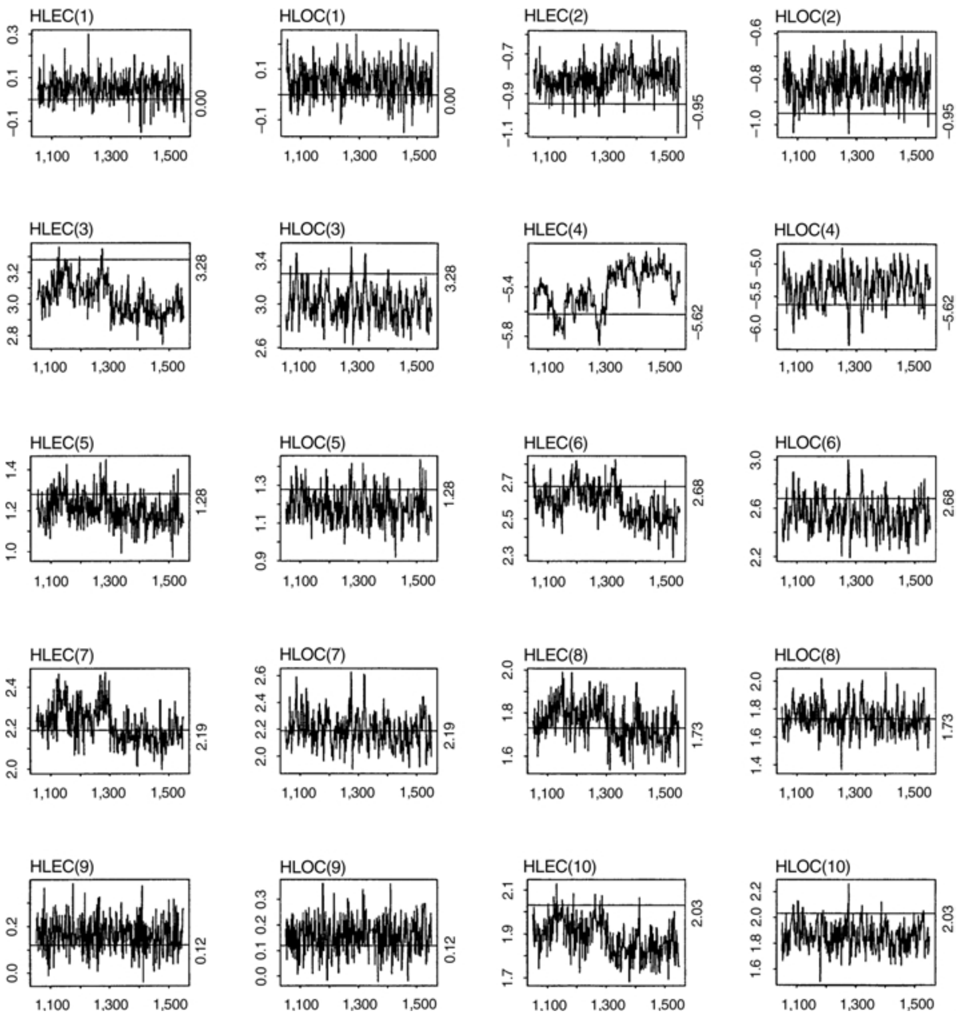

Figure 5. Trace plots of the HLEC algorithm and the HLOC algorithm for the last 500 steps. Bayesian analysis of the Multinomial logit model.

show in our Logit example. The problem with the Ozaki method is its computational cost, since the matrix exponential its requires can often be prohibitively expensive to calculate. Therefore, Langevin methods based on the simpler Euler scheme still have value.

We also found that HLEC does not perform well in multi-modal contexts. While the use of heavy tailed target densities can help somewhat in these situations, we also offer an alternative approach, the tempered Langevin diffusion and its associated Metropolis-Hastings algorithm. Though introduced using a theoretical construction, its potential appeal lies in the exploration of multi-modal target densities and our experimentation of the tempered algorithm in Section 4 on simple examples gives extremely promising results.

Acknowledgment

Work supported in part by NSF Grants DMS 9805598.

Note

1. A function $h : \mathbb{R}^n \to \mathbb{R}$ is uniformly Holder continuous if there exist positive constants a and b such that $|h(y) - h(x)| \leq a|x - y|^b$.

References

J. D. Doll, P. J. Rossky, and H. L. Friedman, ''Brownian dynamics as smart Monte Carlo simulation,'' *Journal of Chemical Physics* vol. 69 pp. 4628–4633, 1978.

J. A. Doornik, *Ox: Object Oriented Matrix Programming, 1.10*, Chapman & Hall: London, 1996.

D. Down, S. P. Meyn, and R. L. Tweedie, ''Exponential and uniform ergodicity of Markov processes,'' *Ann. Probab.* vol. 23 pp. 1671–1691, 1995.

H. Ganidis, B. Roynette, and F. Simonot, ''Convergence rate of some semi-groups to their invariant probability,'' *Stochastic Processes and their Applications* vol. 68 pp. 65–82, 1999.

C. D. Gelatt, S. Kirkpatrick, and M. P. Vecchi, ''Optimization by simulated annealing,'' *Science* vol. 220 pp. 671–680, 1983.

W. K. Hastings, ''Monte Carlo sampling methods using Markov chains and their applications,'' *Biometrika* vol. 57 pp. 97–109, 1970.

S. F. Jarner and G. O. Roberts, Convergence of heavy tailed MCMC algorithms. available at http://www.statslab.cam.ac.uk/ ~ mcmc/

I. Karatzas and S. E. Shreve, *Brownian Motion and Stochastic Calculus*, Springer-Verlag: New York, 1991.

J. Kent, ''Time-reversible diffusions,'' *Adv. Appl. Probab.* vol. 10 pp. 819–835, 1978.

I. E. Leonard, ''The matrix exponential,'' *SIAM Review* vol. 38 pp. 507–512, 1996.

E. Marinari and G. Parisi, ''Simulated tempering: A new Monte Carlo methods,'' *Europhysics Letters* vol. 19 pp. 451–458, 1992.

S. P. Meyn and R. L. Tweedie, *Markov Chains and Stochastic Stability*, Springer-Verlag: London, 1993.

S. P. Meyn and R. L. Tweedie, ''Stability of Markovian processes III: Foster-Lyapunov criteria for continuous time processes,'' *Adv. Appl. Probab.* vol. 25 pp. 518–548, 1993.

C. Moler and C. Van Loan, ''Nineteen dubious ways to compute the exponential of a matrix,'' *SIAM Review* vol. 20 pp. 801–836, 1978.

R. Neal, ''Sampling from multimodal distributions using tempered transitions,'' *Statistics and Computing* vol. 6 pp. 353–366, 1996.

T. Ozaki, ''A bridge between nonlinear time series models and nonlinear stochastic dynamical systems: A local linearization approach,'' *Statistica Sinica* vol. 2 pp. 113–135, 1992.

G. O. Roberts and J. S. Rosenthal, ''Optimal scaling of discrete approximations to Langevin diffusions,'' *Journal of the Royal Statistical Society, Series B* vol. 60 pp. 255–268, 1998.

G. O. Roberts and J. S. Rosenthal, ''Optimal scaling for various Metropolis-Hastings algorithms,'' *Statistical Science*, 2001.

G. O. Roberts and R. L. Tweedie, ''Exponential convergence of Langevin diffusions and their discrete approximations,'' *Bernouilli* vol. 2 pp. 341–364, 1996a.

G. O. Roberts and R. L. Tweedie, ''Geometric convergence and central limit theorems for multi-dimensional Hastings and Metropolis algorithms,'' *Biometrika* vol. 83 pp. 95–110, 1996b.

L. C. G. Rogers and D. Williams, *Diffusions, Markov Processes and Martingales*, John Wiley: New York, 1987.

I. Shoji and T. Ozaki, ''A statistical method of estimation and simulation for systems of stochastic differential equations,'' *Biometrika* vol. 85 pp. 240–243, 1998.

O. Stramer and R. L. Tweedie, ''Langevin-type models I: Diffusions with given stationary distributions, and their discretizations,'' *Methodology and Computing in Applied Probability* vol. 1 pp. 283–306, 1999a.

O. Stramer and R. L. Tweedie, ''Langevin-type models II: Self-targeting candidates for mcmc algorithms,'' *Methodology and Computing in Applied Probability* vol. 1 pp. 307–328, 1999b.

D. W. Stroock and S. R. S. Varadhan, *Multidimensional Diffusion Processes*, Springer-Verlag: Berlin, 1979.