Methodology and Computing in Applied Probability, 1:3, 307–328 (1999)

Langevin-Type Models II: Self-Targeting Candidates for MCMC Algorithms*

O. STRAMER stramer@stat.uiowa.edu
Department of Statistics and Actuarial Science, University of Iowa, Iowa City IA 52242, USA

R. L. TWEEDIE tweedie@biostat.umn.edu
Division of Biostatistics, University of Minnesota, Minneapolis MN 55455, USA

Received September 23, 1998; Revised March 22, 1999; Accepted April 13, 1999

Abstract. The Metropolis-Hastings algorithm for estimating a distribution π is based on choosing a candidate Markov chain and then accepting or rejecting moves of the candidate to produce a chain known to have π as the invariant measure. The traditional methods use candidates essentially unconnected to π. We show that the class of candidate distributions, developed in Part I (Stramer and Tweedie 1999), which ''self-target'' towards the high density areas of π, produce Metropolis-Hastings algorithms with convergence rates that appear to be considerably better than those known for the traditional candidate choices, such as random walk. We illustrate this behavior for examples with exponential and polynomial tails, and for a logistic regression model using a Gibbs sampling algorithm. The detailed results are given in one dimension but we indicate how they may extend successfully to higher dimensions.

Keywords: Hastings algorithms, Metropolis algorithms, Markov chain Monte Carlo, diffusions, Langevin models, discrete approximations, posterior distributions, irreducible Markov processes, geometric ergodicity, uniform ergodicity, Gibbs sampling

AMS 1991 Subject Classification: 60J05, 60J10, 60J70, 65C05, 62E17

1. Introduction

The development of the Metropolis and Hastings (M-H) algorithms represents one of the most active interactions between statistical methodology and applied probabilistic techniques. The M-H algorithms allow simulation of a probability distribution π which is only known up to a constant (normalising) factor. This is surprisingly widely relevant, occurring especially when π is a Bayesian posterior distribution, but in many other contexts also (Besag and Green (1993), Besag *et al.* (1995), Gilks *et al.* (1996)). Recent work on the probabilistic structure of M-H algorithms includes criteria for convergence and approaches to the speed of convergence of the algorithm (Mengersen and Tweedie (1996); Roberts and Tweedie (1996); Smith and Roberts (1993); Tierney (1994); Roberts and Tweedie (1996)).

In this paper we show that a new class of algorithms, based on the Langevin-type diffusions introduced in Part I (Stramer and Tweedie (1999)), converge noticeably faster in

*Work supported in part by NSF Grants DMS 9805598 and DMS 9504561

many contexts than traditional versions. We give detailed results for models in one dimension only, where the underlying diffusion theory is more complete; in Section 9 we indicate how the approach we use may extend successfully to higher dimensions.

In the standard construction (Metropolis *et al.* (1953), Hastings (1970)) of the M-H algorithm on a space X, one first considers a *candidate transition kernel* $Q(x, \cdot)$, $x \in \mathsf{X}$, which generates potential transitions for a discrete time Markov chain evolving on X. To avoid technical difficulties and assumptions, we assume here that X is $\mathbb{R} = (-\infty, \infty)$ equipped with the Borel σ-field $\mathscr{B}$, and both π and $Q(x, \cdot)$ have densities $\pi(y)$ and $q(x, y)$ with respect to Lebesgue measure μ^{Leb}: much more general formulations are possible (see Tierney (1994), Gilks *et al.* (1996)) and our methods can be adapted to them, albeit with different degrees of difficulty.

A ''candidate transition'' to y, generated according to the density *q(x,y)*, is then accepted with probability $\alpha(x,y)$, given by

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

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

$$r(x) = P(x, \{x\}) = \int q(x, y)[1 - \alpha(x, y)]dy. \tag{2}$$

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 $x \in X$, $A \in \mathscr{B}$.

As exemplified in Roberts and Tweedie (1996), the user is often faced with a choice between a traditional version of the M-H algorithm, where the candidate is, say, a random walk which moves independently of the shape of π, and a more ''targeted'' candidate distribution, designed for the particular π in the problem. The former is often easy to implement: however, a more π-specific algorithm may converge more rapidly.

The idea of shaping the candidate density based on the target was introduced as long ago as Doll *et al.* (1978) and in the probabilistic literature has been recently studied in Roberts and Tweedie (1996): they consider the candidate distribution

$$Q_L(x, \cdot) = N(x + \frac{1}{2}h\nabla \log \pi(x), h) \tag{3}$$

where $h > 0$, N is the standard normal distribution, and ∇ is the differential operator $\nabla f(x) = df/dx$. This choice is motivated by the fact that π is the stationary measure for a Langevin diffusion process, and (Doll *et al.* (1978), Roberts and Tweedie (1996), Stramer and Tweedie (1999) Q_L are the approximate h-step transition probabilities for this diffusion, based on a discretization known as the Euler scheme.

The rationale for using this candidate is therefore simple: since the Langevin diffusion already converges to π, then the chain Q_L should also converge in the ''right direction'' (if not quite to π), and should require only minimal corrections through a Metropolis step (1) in order to converge correctly (Besag (1993)).

Given this argument, one particularly surprising result in Roberts and Tweedie (1996) is therefore that, even when the Langevin diffusions converge quickly, the Metropolis-Hastings algorithms based on naive Euler discretizations may lose this geometric rate of convergence, even for quite simple densities π such as $\pi(x) \propto \exp(-|x|^{\beta})$ with $\beta > 2$, for which the classical random walk candidates perform much better (Mengersen and Tweedie (1996)).

A second problem with the algorithm based on the Euler discretization of the Langevin diffusion is that, while it may be efficient for sampling within a single mode, it may not be efficient for sampling from multimodal distributions, and over any reasonable length of time it often converges only locally to the distribution, in the vicinity of a single mode.

One way to improve the simple Langevin-Euler scheme is to use higher order schemes for the Langevin diffusion, based on stochastic Taylor expansions which reduce the error due to discretization (Kloeden and Platen (1992)). These might lead, for example, to consideration of candidate distributions using a second order scheme, such as

$$Q_{L2}(x,\,\cdot\,) = N(x + b(x) + \frac{1}{2}(b(x)b'(x) + \frac{1}{2}b''(x))h^2, h + \frac{1}{3}h^3 + \frac{1}{2}b'(x)h^2) \tag{4}$$

where $b(x) = \frac{1}{2}h\nabla \log \pi(x), h > 0$, and N is again the standard normal distribution. We compare this with other methods in Section 8. However, these more complex algorithms, just like the naive Euler scheme, may also lose the geometric rate of the underlying diffusion, and may also be inefficient for sampling multimodal distributions.

In this paper our approach is rather different. We start from the M–H algorithm, and we use candidate densities based on a wider class of diffusions, developed in Part I (Stramer and Tweedie (1999)), rather than attempting a closer approximation to the Langevin diffusion. We then prove that pathological behavior can be avoided with a better choice from this class, and indeed convergence can be substantially improved.

As in Part I we assume from here on that π is an arbitrary probability density on $\mathbb{R}$ which is positive and twice differentiable almost everywhere, and base our choice of candidate on functions b, σ which are bounded on compact subintervals of $\mathbb{R}$ and satisfy the equations

$$b(x) = \left[\frac{1}{2}\nabla \log \pi(x)\right]\sigma^2(x) + \sigma(x)\nabla\sigma(x). \tag{5}$$

Given any solutions of (5), we define the *self-targeting candidate* Q_{ST} by taking, for some $h > 0$ (which we suppress in notation at this point),

$$Q_{ST}(x,\,\cdot\,) = N(\mu_{x,h}, \sigma^2_{x,h}) \tag{6}$$

where the mean and variance of the jump from x are given by

$$\mu_{x,h} = x + \frac{b(x)}{b'(x)}(\exp(b'(x)h) - 1)$$

$$\sigma^2_{x,h} = \frac{\sigma^2(x)}{2b'(x)}[\exp(2b'(x)h) - 1], \tag{7}$$

when $x \in C$, some compact interval around zero, and more simply by

$$\mu_{x,h} = x\exp(b(x)h/x),$$

$$\sigma^2_{x,h} = \frac{x\sigma^2(x)}{2b(x)}[\exp(2b(x)h/x) - 1], \tag{8}$$

when $x \notin C$. This is the ''hybrid'' discretization introduced in Section 5 of Part I. Overall, we will show that by being somewhat more sophisticated in choosing Q, the effectiveness of the M-H procedure can be substantially enhanced.

2. The Effectiveness of Self-Targeting: An Example

To motivate what follows, we show here the effectiveness of the diffusion approach using as the target the distribution

$$\pi(x) \propto \begin{cases} (5 + (x-10)^2)^{-3} & \text{if } x \geq 0 \\ (5 + (x+10)^2)^{-3} & \text{otherwise.} \end{cases}$$

This target presents two common problems: potentially slow convergence, due to the heavy tails, and multimodality, which is likely to cause inaccurate convergence over finite runs.

We carry out the M-H algorithm with two choices of candidate. The first, denoted $B(i)$ in Figures 1 and 2, is Q_L, given by (3): this is the ''Langevin-Euler'' scheme. We know that this is not geometrically ergodic from Roberts and Tweedie (1996).

The second, denoted $B(ii)$, uses the more carefully chosen self-targeting diffusion-based candidate $\overline{Q}_{ST}$, defined in (15), with the choice of parameters given in Example 4 in Section 7: this is uniformly ergodic from Theorem 5.2.

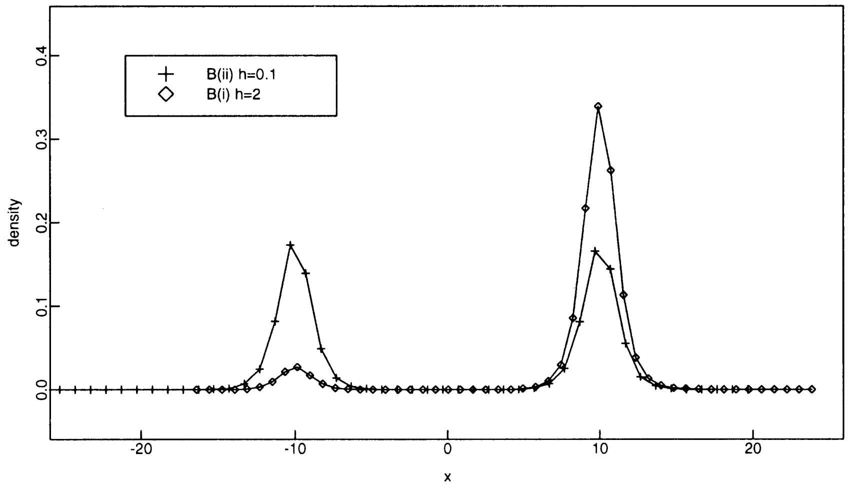

Figure 1. Comparison of density estimates for $B(i)$ with $h = 2$ and $B(ii)$ with $h = 0.1$.

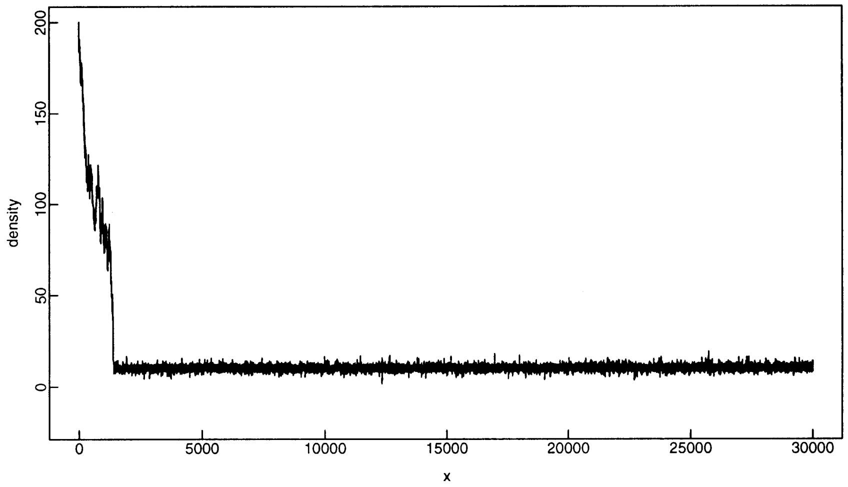

Figure 2. Trace plot of 30000 values from method $B(i)$ with $h = 2$.

We assess their behavior using a single long series: details are in Section 7. We simulate 100,000 steps with initial points $x_0 = 0$ and $x_0 = 200$. To eliminate the effect of the initial point, we have discarded 300 points.

In Figure 1 we show the density estimate when using method $B(i)$ with $h = 2$ and $x_0 = 0$ and method $B(ii)$ with $h = 0.1$ and $x_0 = 0$. These simulations show that algorithm $B(i)$ may tend to ''stick'' in the vicinity of one mode for long periods of time. Worse results were obtained for other values of the parameter h: over 100,000 iterations we found the percentage of time in each mode given by

method $B(i)$ with $h = 0.1$: 1.45% and 98.55% respectively;
method $B(i)$ with $h = 2$: 7.53% and 92.47% respectively;
method $B(i)$ with $h = 5$: 91.49% and 8.51% respectively;
method $B(i)$ with $h = 10$: 5.8% and 94.2% respectively;

In contrast, method $B(ii)$ appears not only to follow each mode well but also to estimate the relative weights of the modes well. The relative weights of the modes for method $B(ii)$ with $h = 0.1$ were 49.63% and 50.37% respectively.

The convergence rate as well as the mode-swapping behavior are shown well by trace plots. Figure 2 is a trace plot of the steps taken by algorithm $B(i)$ with $h = 2$ and a starting point $x_0 = 200$, and shows that model $B(i)$ exhibits the classic problems of burn-in and of poor mode-swapping: the lower mode is never hit in these first 30,000 points. Figure 3 is a trace plot of the steps taken by algorithm $B(ii)$ with $h = 0.1$ and a starting point $x_0 = 200$, and shows that when using this approach we need to discard only a few points to be ''close'' to one of the modes, and mode-swapping occurs every 1,000 or so points in this sample.

Thus, by using the non-Langevin diffusion choice, we obtain a scheme which describes each mode well and also estimates well the relative weights of the modes.

In Section 7, we give a number of other examples of the improvement that using self-

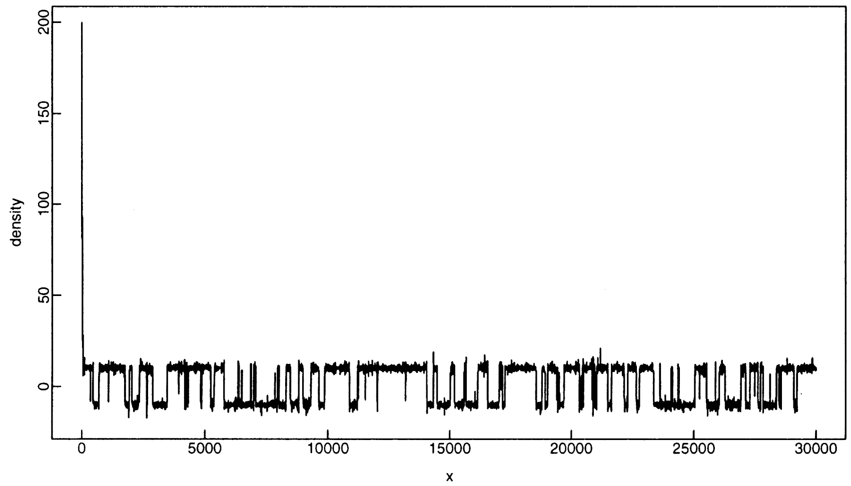

Figure 3. Trace plot of 30000 values from method $B(ii)$ with $h = 0.1$.

targeting algorithms may provide: we now develop the theoretical basis on which one might construct such algorithms.

3. Linking the Convergence Properties of Discretizations and MADA Chains

Since the candidate Q_{ST} does not actually converge to the ‘‘right’’ measure π, even when (5) holds, the need for some form of correction is apparent. Thus we introduce a Metropolis accept-reject step, as described in (1). Following Part I, we will write the chain corresponding to Q_{ST} as G_h, and we write M_h for the Metropolized version of G_h. We will call M_h the Metropolis-adjusted diffusion algorithm, or MADA chain. A key result that enables us to use results from Part I links the convergence properties of discretizations and MADA chains.

THEOREM 3.1. *Suppose $\pi(x)$ is positive and continuous, and $q(x, y)$ is positive and continuous in both variables. Let P be the transition law of the Metropolised chain formed from Q. If $\alpha(x, y)$ is such that*

$$r(x) = P(x, \{x\}) \to 0, \quad |x| \to \infty \tag{9}$$

then

(i) If Q is geometrically ergodic then P is geometrically ergodic.
(i) If Q is uniformly ergodic then P is uniformly ergodic.

Proof: Under these continuity conditions it follows from Mengersen and Tweedie, (1996) that all compact sets are small for *P*, as they also are for *Q* from [11, Chapter 6].

If Q is geometrically ergodic, we have from Theorem 15.0.1 of Meyn and Tweedie, (1993) that there exists a function $V \geq 1$, bounded on compact sets, and a $\lambda < 1, b < \infty$ such that for all sufficiently large compact sets C

$$\int Q(x, dy)V(y) \leq \lambda V(x) + b\mathbb{1}_C(x). \tag{10}$$

Choose ϵ small enough that $\lambda + \epsilon = \lambda' < 1$. If C is large enough that $r(x) \leq \epsilon, x \in C^c$ then we have immediately

$$\begin{aligned}\int P(x, dy)V(y) &\leq \int Q(x, dy)V(y) + r(x)V(x)\\ &\leq \lambda' V(x) + [b + \sup_{w \in C} V(w)]\,\mathbb{1}_C(x).\end{aligned} \tag{11}$$

Thus using the sufficiency of (11) as in Theorem 15.0.1 of Meyn and Tweedie (1993), we have that P is also geometrically ergodic.

If Q is uniformly ergodic we have similarly from Theorem 16.0.2 of Meyn and Tweedie, (1993) that there is a bounded V such that (10) holds; as in (11), this inequality is maintained for the same bounded V for P, and so from the sufficiency in Theorem 16.0.2 of Meyn and Tweedie (1993) we have that P is also uniformly ergodic. ■

The continuity conditions can certainly be weakened: all we need is that both Q and P are T-chains in the sense of Meyn and Tweedie (1993), Chapter 6. However, they hold in all of the examples below, and although some further fine-tuning of Q_{ST} may be needed to ensure (9) holds, in general the MADA chain M_h will retain the convergence properties of G_h developed in Part I.

4. Uniform Convergence of the MADA Algorithms

We now consider geometric convergence properties of the chains M_h. In particular we show that for a broad class of densities π, one can always construct some MADA chain M_h that converges uniformly fast to π as its stationary distribution.

Our first result shows that for light-tailed distributions the MADA chain converges rapidly, no matter how coarse the discretization. To state this we need the following condition from Part I:

Condition A1: π is in the class $\mathscr{E}^*$ of densities that have asymptotically exponential tails such that

$$-\frac{\log \pi(x)}{|x|^{\beta}} \to \alpha, \quad x \to \infty, \tag{12}$$

for some α, $\beta > 0$, and $\sigma^2(x)$ is chosen such that

$$\sigma^2(x)/(-\log \pi(x))^{\frac{\gamma s}{\beta}} \to a, \quad |x| \to \infty,$$

where $\gamma_s \geq 0$ and $a > 0$.

THEOREM 4.1 *Assume that Condition A1 holds with $\beta > 2$. Then the MADA chain M_h is uniformly ergodic for any h.*

Proof: We note that for large $|x|$, $\sigma^2(x) \approx \alpha_s |x|^{\gamma_s}$, where $\alpha_s > 0$ and from (5), $b(x) \approx -\alpha_b |x|^{\gamma_b} \mathrm{sgn}(x)$, where $\alpha_b > 0$ and $\gamma_b - \gamma_s + 1 = \beta$. We have from Theorem 6.2 of Part I that under these conditions, the law Q_{ST} is uniformly ergodic. We thus need only check (9).

Let $\varepsilon > 0$ be arbitrary. It is easy to verify under the assumptions of the theorem that there exist $1 < M_1 < M$ such that

(i) for all $|x| > M \mu_{x,h}$ and $\sigma_{x,h}$ are defined as in (8);
(ii) for all $|x| > M$ and $|y| < M_1$, $\alpha(x, y) \geq 1 - \epsilon$;
(iii) $P(|Z| < \frac{M_1 - 1}{\ell}) \geq 1 - \epsilon$, where $Z \sim N(0, 1)$, and $\ell = \frac{\alpha_s}{2\alpha_b}$;
(iv) for all $|x| > M$, $|\mu_{x,h}| < 1$ and $\sigma^2_{x,h} \leq \ell$.

Set $C_1 = [-M_1, M_1]$ and $C = [-M, M]$. Now from (i) we have that for all $|x| > M$,

$$\int_{C_1} q(\dot{x}, y)(1 - \alpha(x, y))dy \leq \epsilon.$$

Moreover, for all $|x| > M$, from (ii) and (iii),

$$\begin{aligned} \int_{C_1^c} q(x, y)dy &= P(|\mu_{x,h} + \sigma_{x,h} Z| > M_1) \\ &= P(Z > \frac{M_1 - \mu_{x,h}}{\sigma_{x,h}}) + P(Z < \frac{-M_1 - \mu_{x,h}}{\sigma_{x,h}}) \qquad (13) \\ &\leq P(Z > (M_1 - 1)/\ell)) + P(Z < (-M + 1)/\ell)) \\ &\leq \varepsilon. \end{aligned}$$

Thus for all $|x| > M$,

$$P(x, \{x\}) \leq \int_{C_1^c} q(x, y) + \int_{C_1} q(x, y)(1 - \alpha(x, y))dy \leq 2\varepsilon, \qquad (14)$$

and the result follows. ■

Note that as in Theorem 6.2 of Part I, we could directly show that $\mathsf{E}\,[\tau_C]$ is bounded using the argument above, where $\tau_C = \inf\{t > 0 : M_h(t) \in C\}$ is the first hitting time on C: this is an alternative criterion for uniform ergodicity.

It is clear what is happening here. For large values of $|x|$ the MADA chain M_h is close to an independent candidate chain (i.e. $q(x_1, \cdot) \approx q(x_2, \cdot)$ for large values of $|x_1|$, $|x_2|$) and we can use arguments such as those in Mengersen and Tweedie (1996) to show directly that the chain is uniformly ergodic. However, for smaller values of x the MADA chain will follow the contours of π more closely, thereby removing one of the unattractive features of the independent model.

Without light tails we have a converse result.

PROPOSITION 4.2 *Assume that Condition A1 holds with $\beta \leq 2$ and $\gamma_s + \beta - 2 > 0$. Then M_h is not geometrically ergodic.*

Proof: This follows since under these conditions it is easy though somewhat tedious to show that the rejection probability $r(x) \to 1$ as $x \to \infty$. The result thus follows from Theorem 5.1 of Roberts and Tweedie (1996). ■

Example: To illustrate that $r(x) \to 1$ as $x \to \infty$ as in Proposition 4.2, we consider $\pi(x) \propto \exp(-\sqrt{x^2+1})$ and $\sigma^2(x) = x^2 + 1$. For this example $\beta = 1$ and $\gamma_s = 2$. Then from (5), $b(x) = -0.5x\sqrt{x^2+1} + x$. We now have from (8) that

$$\mu_{x,h} = x + x(\exp((1 - 0.5\sqrt{x^2+1})h) - 1)$$

$$\sigma^2_{x,h} = \frac{x^2+1}{2(1 - 0.5\sqrt{x^2+1})}(\exp(2(1 - 0.5\sqrt{x^2+1})h) - 1)$$

When this was simulated with $G_h(0) = M_h(0) = 1000$ and $h = 0.1$, the first draw from $N(0,1)$ was a reasonable value of -0.424896. Then $G_h(1) = -13.449848$,

$$\frac{\pi(G_h(1))}{\pi(G_h(0))} = \frac{\exp(-13.48)}{\exp(-1000.00)}, \frac{q(G_h(1), G_h(0))}{q(G_h(0), G_h(1))} = \frac{\exp(-46938.60)}{\exp(-4.46)}$$

and so $\alpha(G_h(0), G_h(1)) \approx 0$. Thus we almost certainly reject the move, and $M_h(1) = 1000$ again.

5. Using a *t*-Distribution to Improve Convergence

It is perhaps surprising that the boundary case $\beta = 2$ and $\gamma_s > 0$ is not actually uniformly ergodic, since Theorem 6.2 of Part I shows that the candidate chain G_h is uniformly ergodic in this case. Similarly, although the condition $\gamma_s + \beta - 2 \geq 0$ for $\pi \in \varepsilon^*$ is sufficient for exponential ergodicity of the candidate chain G_h (see Theorem 6.1 in Stramer and Tweedie (1999)), from Theorem 4.2 the MADA chain M_h is not geometrically ergodic in this situation if $\beta \leq 2$. Thus the properties of M_h are not always those of the candidate G_h, and we need to address this by developing some variations to the discrete approximation G_h.

We do this by spreading the candidate Q to have heavier tails, using a t-distribution. Recall that $Q_{ST}(x, \cdot)$ is defined by taking

$$\frac{G_h(h) - \mu_{x,h}}{\sigma_{x,h}} \sim N(0,1).$$

Let us define the related chain $\overline{G}_h$ by setting, when $\overline{G}_h(0) = x$,

$$\frac{\overline{G}_h(h) - \mu_{x,h}}{\sigma_{x,h}}\sqrt{\frac{n}{n-2}} \sim \mathscr{T}_n \tag{15}$$

where $\mathscr{T}_n$ denotes the t distribution with $n > 2$ degrees of freedom, and call the resulting candidate $\overline{Q}_{ST}$, with density $\overline{q}(x, y)$. We assume that $n > 2$ so that the variance is finite (see Remark 5.1 in Part I).

We will show that this variation guarantees that $\overline{P}(x, \{x\}) \to 0$ as $x \to \infty$.

THEOREM 5.1 *Assume that Condition A1 holds with*

$$\gamma_s + 2\beta - 2 > 0. \tag{16}$$

Then the MADA chain $\overline{M}_h$ obtained from the candidate $\overline{Q}_{ST}$ is always geometrically ergodic and is uniformly ergodic if $\beta = 2$ and $\gamma_s > 0$.

Proof: Let $\epsilon > 0$ be arbitrary. We show that there exists $M > 0$ such that

$$\overline{P}(x, \{x\}) \le 2\epsilon, \quad |x| > M, \tag{17}$$

where $\overline{P}$ is the transition law of $\overline{G}_h$.

It is easy to check that if $\gamma_s + 2\beta - 2 > 0$ then there exist $\alpha > 0$ and $0 < a < 1$ such that

$$\overline{P}\left(\frac{-b_1(x) - \mu_{x,h}}{\sigma_{x,h}} < \mathscr{T}_n\sqrt{\frac{n-2}{n}} < \frac{b_1(x) - \mu_{x,h}}{\sigma_{x,h}}\right) \to 1$$

as $|x| \to \infty$, where $b_1(x) = |x - \alpha|x|^a \operatorname{sgn}(x)|$. Thus, there exists $M_1 > 0$ such that

$$\int_{|y| < b_1(x)} \overline{q}(x, y)dy \ge 1 - \epsilon$$

for all $|x| > M_1$. We then note that $\overline{\alpha}(x, y) \to 1$ as $x \to \infty$ for all $|y| < b_1(x)$ and hence there exists $M > M_1$ such that for all $|x| > M$ and $|y| < b_1(x)$, $1 - \overline{\alpha}(x, y) \le \epsilon$. It now follows as in (14) (with $C_1 = [-b_1(x), b_1(x)]$) that (17) holds.

The result now follows from Theorem 3.1, and the geometric ergodicity of $\overline{G}_h$ under (16) (and its uniform ergodicity when $\beta = 2$ and $\gamma_s > 0$): this follows exactly as in the proof of Theorem 6.1, Theorem 6.2 of Stramer and Tweedie (1999). ■

In this case, the t distribution has a longer tail than the normal distribution and thus is more appropriate for densities with heavier tails.

Example: We again illustrate this for $\pi(x) \propto \exp(-\sqrt{x^2+1})$ and $\sigma^2(x) = x^2 + 1$. Again we assume that $\overline{G}_h(0) = 1000$. The first draw from the rescaled $\mathscr{T}_3$ distribution in our simulation of this system was -0.148320. Then $\overline{G}_h(1) = -4.694999$,

$$\frac{\pi(G_h(1))}{\pi(G_h(0))} = \frac{\exp(-4.80)}{\exp(-1000.00)}, \frac{q(G_h(1), G_h(0))}{q(G_h(0), G_h(1))} = \frac{\exp(-28.24)}{\exp(-5.14)}$$

and so $\alpha(G_h(0), G_h(1)) = \exp(972.09)$. Thus we accept the move, and so $M_h(1) = -4.694999$.

We next show that by a further modification of Q_{ST} we can again produce a uniformly ergodic algorithm, even in this case.

An appropriate change is to truncate $\sigma_{x,h}$ for large values of $\sigma_{x,h}$. Let us define the chain $\hat{G}_h$ by setting, when $\hat{G}_h(0) = x$,

$$\frac{\hat{G}_h(h) - \mu_{x,h}}{\hat{\sigma}_{x,h}} \sqrt{\frac{n}{n-2}} \sim \mathscr{T}_n, \quad n \geq 3 \tag{18}$$

where

$$\hat{\sigma}_{x,h} = \begin{cases} \sigma_{x,h} & \text{if } \sigma_{x,h} \leq K \\ K & \text{otherwise,} \end{cases}$$

and K is some positive number. We call the resulting candidate $\hat{Q}_{ST}$.

We now have

THEOREM 5.2 *Assume that Condition A1 holds with $\gamma_s + \beta - 2 > 0$. Then the MADA chain $\hat{M}_h$ obtained from the variation $\hat{Q}_{ST}$ is uniformly ergodic.*

Proof: Using the same argument as in Theorem 6.2 of Part I, we have that the candidate chain $\hat{G}_h$ is uniformly ergodic. Using the same argument as in Theorem 5.1, $\hat{P}(x, \{x\}) \to 0$ as $x \to \infty$, and the result follows from Theorem 3.1 as before. ■

Finally in this section we comment on the results when π only has polynomial tails. Our result shows that for polynomial-tailed distributions the modified MADA chain converges rapidly, no matter how coarse the discretization. To state this formally, we first define π to be in the class $\mathscr{P}^*$ of densities with asymptotically polynomial tails if for some $\alpha > 0$, $\eta > 2$

$$\pi(x)|x|^{\eta} \to \alpha, \quad x \to \infty. \tag{19}$$

The following condition was used in Part I.

Condition A2: π is in $\mathscr{P}^*$ and $\sigma^2(x)$ is such that

$$\sigma^2(x)/\pi(x)^{\frac{-\gamma_s}{\eta}} \to a, \quad |x| \to \infty$$

where $0 \leq \gamma_s \leq \eta$, and $a > 0$.

Using a similar argument as in Theorem 6.1 of Part I, we can easily show that under Condition A2 with $\eta > 3$ and $2 \leq \gamma_s < \eta - 1$, $\overline{G}_h$ is geometrically ergodic. However, $\overline{M}_h$ based on $\overline{Q}_{ST}$ fails to satisfy the condition $\overline{P}(x, \{x\}) \to 0$ as $x \to \infty$, so we do not know if $\overline{M}_h$ retains the geometric ergodicity.

Therefore we once again truncate $\sigma^2_{x,h}(x)$ for large $\sigma_{x,h}$ and show that the resulting diffusion $\hat{G}_h$ is geometrically ergodic if $\gamma_s = 2$ and uniformly ergodic if $2 < \gamma_s < \eta$, and that its Metropolized version retains these properties.

Using Condition A2 with $0 \leq \gamma_s < \eta$, we have that for large $|x|$,

$$\sigma^2(x) \approx \alpha_s |x|^{\gamma_s}; \quad b(x) \approx -\alpha_b |x|^{\gamma_s - 1} \operatorname{sgn}(x),$$

where $\alpha_b = \frac{\alpha_s \eta}{2}(1 - \frac{\gamma_s}{\eta})$. Thus, for large $|x|$

$$\mu_{x,h} \approx x + x(\exp(-\alpha_b|x|^{\gamma_s-2}) - 1), \quad \hat{\sigma}_{x,h} = K.$$

It is easy now to check that $\hat{G}_h$ is geometrically ergodic if $\gamma_s = 2$ and uniformly ergodic if $2 < \gamma_s < \eta$. To guarantee that $\hat{P}(x, \{x\}) \to 0$ as $x \to \infty$, it is sufficient that the distribution $\mathscr{T}_n$ of

$$\frac{\hat{G}_h(h) - \mu_{x,h}}{\hat{\sigma}_{x,h}} \sqrt{\frac{n}{n-2}}$$

has a longer tail than the distribution of the target density π. Clearly, this will happen if $n - 1 < \eta$, which implies that we must have $\eta > 3$, since $\mathscr{T}_n$ is the t distribution with $n > 2$ degrees of freedom. Thus if we assume that $\eta > 3$, the result follows.

We conclude that if we choose $\sigma^2(x) \approx \alpha_s|x|^{\gamma_s}$ for large $|x|$ we always get geometric rates of convergence if $\gamma_s \geq 2$ and uniform rates of convergence if $\gamma_s > 2$, for all π in class $\mathscr{E}^*$, and for all π in the class $\mathscr{P}^*$ with $\eta > 3$ provided also $\eta > \gamma_s$. In contrast, essentially as noted also in Roberts and Tweedie (1996), the MADA chain based on the Euler approximation to the Langevin diffusion is exponentially ergodic if and only if π is in class $\mathscr{E}^*$ and $1 \leq \beta \leq 2$.

We illustrate these results in Section 7.

6. Polynomial Convergence of the Langevin Schemes

As we have shown, one can find geometrically or even uniformly converging schemes for virtually all distributions π. However, implementing these may require more computations than for the simple Langevin diffusion with the Euler discretization.

Thus we sometimes prefer to use the Euler scheme to the Langevin diffusion if its convergence properties are not too poor, especially if we can start from the ''center'' of π in some sense.

It is shown in Roberts and Tweedie (1996) that when $\gamma_s = 0$ and $b(x) \to 0$ when $|x| \to \infty$, the Euler scheme (with and without Metropolis adjustment) is not geometrically ergodic. We now will show that it still has a polynomial or better rate of convergence to π, for the f-norm

$$\|P^n(x, \cdot) - \pi\|_f := \sup_{|g| \leq f} |\mathsf{E}_x[g(X_n)] - \mathsf{E}_\pi[g(X_n)]|,$$

where f is a polynomial of appropriate order. Guided by the results in Tuominen and Tweedie (1994), Stramer and Tweedie (1999), we find conditions under which such a polynomial rate of convergence will also hold for the MADA chain.

THEOREM 6.1. *Suppose b is continuous, and that either Condition A1 holds with* $\beta < 1$ *(in which case choose K below as any positive integer) or that Condition A2 holds with* $\gamma_s < 2$ *(in which case choose* $K = \eta$*). Then*

$$r(n)\|P_L^n(x, \cdot) - \pi\|_f \to 0$$

where P_L is the transition law of the MADA chain obtained from the Langevin-Euler scheme Q_L, and for any integer k with $2 \le k \le K$

$$f(x) = |x|^{k-2} \vee 1\, x \in \mathbb{R},$$

$$r(n) = n^{K-k} \vee 1\, n \ge 1.$$

Proof: For simplicity and in order to show the main idea of the proof we will assume that condition A1 holds with $\beta = 0.5$ and $\alpha = 1$ (i.e. $\pi(x) \propto e^{-\sqrt{|x|}}$). Thus if $G_h(0) = x$, then

$$G_h(h) = x - \frac{\operatorname{sgn}(x)}{4|x|^{0.5}} h + \sqrt{h} Z$$

where Z has standard normal distribution. It is easy to check that there exists some $c > 0$ and $x_0 > 0$ such that

$$\int Q_L(x, dy)|y|^k \le |x|^k - c|x|^{k-0.5}, \quad |x| > x_0, \tag{20}$$

where Q_L is the transition probability of the Euler chain. As in (11) it is now sufficient for (20) to hold for P as well as for Q_L to show that $r(x)|x|^k \to 0$ as $|x| \to \infty$.

Set $C^x = [x - 2|x|^a, x + 2|x|^a]$ where $0 < a < 0.5$. Let $h > 0$. It is easy to check that for any $\varepsilon > 0$ there exists $M > 0$ such that

(i) if $|x| > M$ then

$$\int_{y \notin C^x} Q_L(x, dy) \le P(\sqrt{h}|Z| > 2|x|^a) < \frac{1}{|x|^N},$$

where $N > k$ is some integer;

(ii) if $|x| > M$ and $y \in C^x$ then $Q_L(y, dx)$ is the density from $N(y - \frac{1}{4\sqrt{y}}, h)$ at x, which is approximately the same as $N(y,\ h)$ at x and $Q_L(x, dy)$ is approximately the same as $N(x,\ h)$ at y. Thus for $|x| > M$ and $y \in C^x$, $Q_L(y, dx) \approx Q_L(x, dy)$. In other words,

$$\frac{Q_L(y, dx)}{Q_L(x, dy)} \ge 1 - \varepsilon.$$

Let $R^x = \{y \in C^x : \alpha(x, y) \le 1\}$. We now have that for $x > M$

$$\begin{aligned} r(x)x^k &= x^k \int (1 - \alpha(x, y)) Q_L(x, dy) \\ &\le \frac{x^k}{x^N} + x^k \int_{R^x} \left(1 - \frac{e^{-\sqrt{y}}}{e^{-\sqrt{x}}} \frac{Q_L(y, dx)}{Q_L(x, dy)}\right) Q_L(x, dy) \\ &\le \frac{x^k}{x^N} + x^k \int_{R^x} (1 - (1 - \varepsilon) e^{\frac{(x-y)}{\sqrt{x}+\sqrt{y}}}) Q_L(x, dy) \le \frac{x^k}{x^N} + x^k (1 - (1 - \varepsilon) e^{\frac{-2x^a}{\sqrt{x}}}). \end{aligned} \tag{21}$$

Thus using the same argument for $x < -M$, we have that $r(x)|x|^k \to 0$ as $|x| \to \infty$. The proof follows now from (20) by using the same argument as in the proof of Proposition 5.2 of Tuominen and Tweedie (1994). ■

7. Exponential and Polynomial Examples

We can summarize the results above for the various types of tail behavior of π and choices of γ_s as in Table 1. Note that in general we get better rates of convergence for lighter tailed π (i.e. for larger β), and for larger values of γ_s.

Example 1: The exponential class $\mathscr{E}$

We say that $\pi \in \mathscr{E}$, as introduced in Roberts and Tweedie (1996); Stramer and Tweedie (1999), if for some x_0, and some constants $\gamma > 0$ and $0 < \beta < \infty$, π takes the form $\pi(x) \propto e^{-\gamma|x|^\beta}$, $|x| \geq x_0$. We now compare the MADA chain properties in Table 1 to those based on the Euler approximations to the Langevin diffusion:

$\beta > 2$: When the tails of π are lighter than Gaussian we have uniform convergence (see Theorem 4.1) of the algorithm with candidate Q_{ST} (6), as opposed to non-geometric convergence of the algorithm with candidate Q_L (3).

$\beta = 2$: When the tails of π are Gaussian we can get uniform convergence (see Theorem 5.1) of the algorithm with candidate $\overline{Q}_{ST}$ (15) if $\gamma_s > 0$, while we can only get a geometric rate of convergence for the algorithm with candidate Q_L.

$1 \leq \beta < 2$: When the tails of π are heavier than Gaussian but at least exponential we get uniform convergence (see Theorem 5.2) of the algorithm with candidate $\hat{Q}_{ST}$ (18) if $\gamma_s + \beta - 2 > 0$, while we can only get a geometric rate of convergence for the algorithm with candidate Q_L.

$0 < \beta < 1$: When the tails of π are heavier than exponential we still have uniform convergence (see Theorem 5.2) of the algorithm with candidate $\hat{Q}_{ST}$, as opposed to non-geometric convergence of the algorithm with candidate Q_L.

Example 2: Polynomial convergence: Suppose that $\pi \sim \mathscr{T}_n \in \mathscr{P}^*$, the t distribution with $n \geq 3$ degrees of freedom: that is, $\pi(x) \propto (1 + \frac{x^2}{n})^{-\frac{n+1}{2}}$, $x \in \mathbb{R}$. Choose

$$\sigma^2(x) = (n + x^2)^{\frac{\gamma_s}{2}}.$$

Table 1. Qualitative behavior of various candidate transition laws.

π	β	γ_s	Candidate	Convergence	Reference
Light tails	$\beta > 2$	All γ_s	Q_{ST} (6)	Uniform	Theorem 4.1
Heavier tails	$\beta \leq 2$	$\gamma_s + \beta - 2 > 0$	Q_{ST} (6)	Not geometric	Proposition 4.2
Gaussian	$\beta = 2$	$\gamma_s > 0$	$\bar{Q}_{ST}$ (15)	Uniform	Theorem 5.1
All tails	$\beta > 0$	$\gamma_s + 2\beta - 2 > 0$	$\bar{Q}_{ST}$ (15)	Geometric	Theorem 5.1
All tails	$\beta > 0$	$\gamma_s + \beta - 2 > 0$	$\hat{Q}_{ST}$ (18)	Uniform	Theorem 5.2

Then $\widehat{M}_h$ is uniformly ergodic if $2 < \gamma_s < n+1$ and geometrically ergodic if $2 = \gamma_s < n+1$. In Roberts and Tweedie (1996), it is shown that the MALA (Metropolis-adjusted Langevin algorithm) chain based on Q_L is not geometrically ergodic for this class of densities.

To illustrate these rates, we now carry out these computations in a little more detail in one specific case.

Example 3: Specific polynomial convergence: Consider the model

$$\pi(x) \propto \frac{1}{(5+x^2)^3}, \quad x \in \mathbb{R}.$$

We compare three algorithms:

A(i). The Metropolis algorithm with the candidate distribution $N(x,\ 1)$ (i.e. the random walk algorithm);

A(ii). The MALA chain obtained from Q_L with $h=1$, (i.e. $\sigma^2(x) \equiv 1$ $b(x) = -3x/(5+x^2)$);

A(iii). The MADA chain $\widehat{M}_h$ with $\sigma^2(x) = (5+x^2)^{1.5}$, $b(x) = -1.5x(5+x^2)^{0.5}$ and $h = 0.1$; here $\gamma_s = 3$.

We will evaluate convergence through the behavior of the conditional variances $v(x,t) = \text{Var}(M_h(t)|M_h(0) = x)$ with $x=1$, $x=5$ and $x=9$. In Figure 4, we show $v(x,t)$ for the Metropolis algorithm using *A(i);* in this case as shown in Mengersen and Tweedie, (1996) the chain does not converge exponentially fast and indeed convergence clearly depends on the starting point.

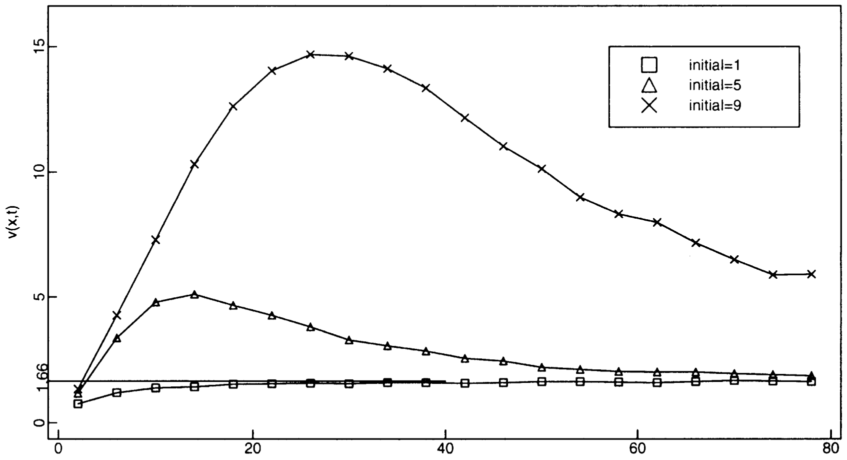

Figure 4. Conditional variance $v(x,t)$ with $x(0) = 1, x(0) = 5$ and $x(0) = 9$ for $A(i)$.

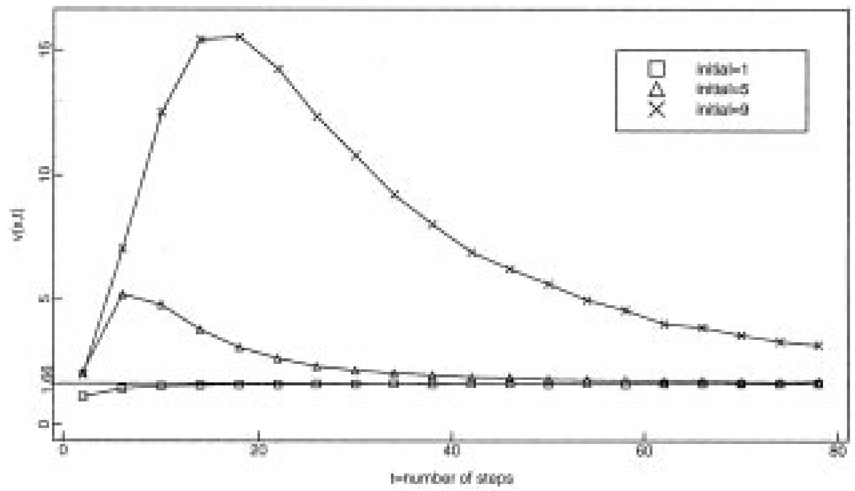

Figure 5. Conditional variance $v(x, t)$ with $x(0) = 1, x(0) = 5$ and $x(0) = 9$ for $A(ii)$.

In Figure 5, we show $v(x, t)$ for the MADA chain M_h using $A(ii)$; in this case as shown in Roberts and Tweedie (1996) the chain does not converge exponentially fast but from Theorem 6.1 it still has a third order polynomial subgeometric rate of convergence to π. Convergence is still slow and depends on the starting point but it is faster than for the Metropolis algorithm using $A(i)$.

In Figure 6, we show $v(x, t)$ for the MADA chain $\widehat{M}_h$ using $A(iii)$ which is uniformly ergodic; the conditional variance converges quite rapidly for all starting points.

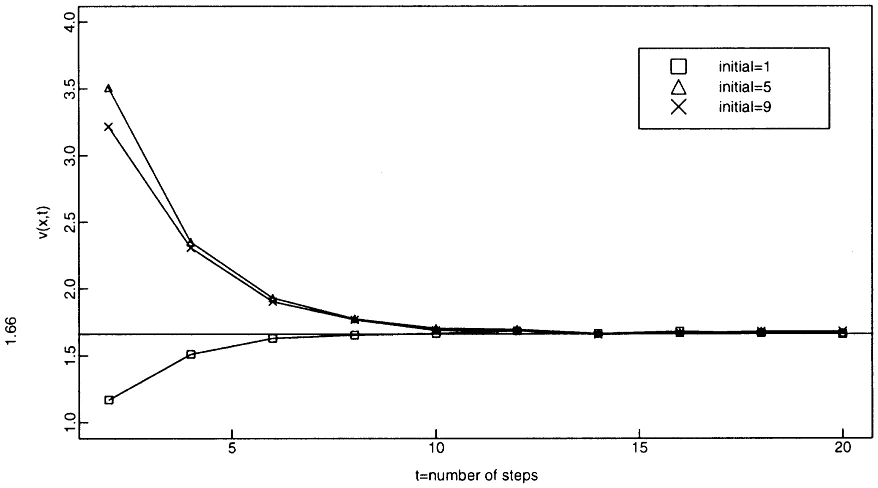

Figure 6. Conditional variance $v(x, t)$ with $x(0) = 1, x(0) = 5$ and $x(0) = 9$ for $A(iii)$.

Example 4 : Estimating Mixtures

We now describe in detail the model discussed in the introduction. We take

$$\pi(x) \propto \begin{cases} (5+(x-10)^2)^{-3} & \text{if } x \geq 0 \\ (5+(x+10)^2)^{-3} & \text{otherwise,} \end{cases}$$

and we compare two algorithms:

$B(i)$ This uses the transition law Q_L with $h = 2$ and $\sigma^2(x) \equiv 1$, and

$$b(x) = \begin{cases} \frac{-3(x-10)}{5+(x-10)^2} & \text{if } x \geq 0 \\ \frac{-3(x+10)}{5+(x+10)^2} & \text{otherwise.} \end{cases}$$

$B(ii)$ This uses the MADA chain $\overline{M}_h$ obtained from $\overline{Q}_{ST}$ with $h = 0.1$, and

$$\sigma^2(x) = \begin{cases} (5+(x-10)^2)^{3-\varepsilon} & \text{if } x \geq 0 \\ (5+(x+10)^2)^{3-\varepsilon} & \text{otherwise;} \end{cases}$$

$$b(x) = \begin{cases} -(x-10)(5+(x-10)^2)^{2-\varepsilon}\varepsilon & \text{if } x \geq 0 \\ -(x+10)(5+(x+10)^2)^{2-\varepsilon}\varepsilon & \text{otherwise;} \end{cases}$$

here $0 < \varepsilon \approx 0$, and thus $\gamma_s \approx 6$.

For method $B(ii)$ we have chosen σ^2 in such a way that (i) $\mu_{x,h}$ is small when π is not too small and is bigger when π is very small and $\sigma^2_{x,h}$ is larger as π decreases, so that each mode is well described (see Figure 1, Figure 3); and (ii) the algorithm is uniformly ergodic (see Figure 3).

In contrast, as shown in Mengersen and Tweedie (1996) when using method $B(i)$ the chain does not converge exponentially fast and indeed convergence clearly depends on the starting point (see Figure 2); and the chain also exhibits the problem of poor mode-swapping for different choices of h (see Figure 1, Figure 2).

8. A Logistic Regression Model

We next consider a Bayesian logistic regression model (see Gilks *et al.*, (1996)). We suppose that the distribution of y given the covariate z is

$$y_i \sim \text{Bernoulli } ((1+e^{-(\mu+\alpha z_i)})^{-1}, 1), \quad i = 1, \ldots, n;$$

where the parameters α, μ have the distributions

$$\alpha \sim N(0,1); \quad \mu \sim N(0,1).$$

We assume conditional independence between the $\{y_i\}$ given the model parameters and covariates, and independence between the parameters themselves. Thus the full conditional for α is

$$\pi(\alpha|\mu) \propto e^{-\frac{1}{2}\alpha^2}\Pi_{i=1}^{n}\{1+e^{-(\mu+\alpha z_i)}\}^{-y_i}\{1+e^{\mu+\alpha z_i}\}^{y_i-1},$$

and for μ is

$$\pi(\mu|\alpha) \propto e^{-\frac{1}{2}\mu^2}\Pi_{i=1}^{n}\{1+e^{-(\mu+\alpha z_i)}\}^{-y_i}\{1+e^{\mu+\alpha z_i}\}^{y_i-1},$$

which do not simplify.

We wish to simulate the two dimensional posterior distribution $\pi(\alpha,\mu)$. One way to do this is to use the Gibbs sampler, where we successively update each component with a value picked from its distribution conditional on the current value of the other component (Gilks *et al.* (1996)). We simulate 1000 observations from this logistic model with $\mu = 1.10096$ and $\alpha = 2.276053$: these were random draws of μ and α from the prior which is N(0, 1). For the sake of illustration we fix $\mu = 1.10096$ and simulate $\pi(\alpha|\mu = 1.10096)$.

We compare the four algorithms:

$C(i)$ The Metropolis (random walk) algorithm: we use the candidate distribution $N(x,(2.4)^2\ 0.016)$, which is optimal in some ways for reasons explained below;

$C(ii)$ The Langevin diffusion with the Euler scheme: we use $\sigma^2(x)\equiv 1$, $b(x) = \frac{1}{2}\frac{\partial}{\partial x}\log(\pi(x|\mu = 1.10096))$, with the MADA chain obtained from Q_L with $h = 0.01$;

$C(iii)$ The Langevin choice with the second order scheme (4): here b and σ are defined as in $C(ii)$ with the MADA chain obtained from Q_{L2} with $h = 0.01$;

C(iv) A MADA chain $\widehat{M}_h$ with $h = 0.01$: we took $\sigma^2(x) = -d\log(\pi(x|\mu = 1.10096))$ with $d = 0.005$, $b(x) = -\frac{d}{2}\frac{\partial}{\partial x}\log(\pi(\alpha|\mu = 1.10096))(\log\pi + 1)$, so that $\gamma_s = 4$.

We first assess the behavior of a single long series for case C(iv) with 50,000 steps; we start from $x_0 = 10$ and discard 300 steps to eliminate the effect of the initial point. Figure 7 shows the estimated histogram for π.

Using Figure 7 we have that $\pi(\alpha|\mu)$ is approximately normal with variance 0.016 and thus from Roberts *et al.* (1995) we have that among the class of symmetric normal candidate distributions, the most efficient candidate distribution is $N(x, 2.4^2\ 0.016)$, which we use in $C(i)$.

In Figure 8 we estimate the conditional mean $m(x,\ t)$, where t denotes the number of steps and x is the starting point. for $C(i)$, $C(ii)$, $C(iii)$ and $C(iv)$ with $x = 10$, and 10,000 replications.

Cases $C(i)$, C(ii) and $C(iii)$ are geometrically ergodic, from Theorem 5.1 for $C(ii)$ and $C(iii)$ and from Mengersen and Tweedie (1996) for case $C(i)$. It is perhaps surprising that there is almost no improvement moving from the MADA chain based on the first order (Euler) scheme for the Langevin diffusion as in $C(ii)$ to the second order scheme as in $C(iii)$; but both converge more rapidly than the traditional random walk algorithm $C(i)$, for which convergence could be even slower without the information obtained by using algorithm C(iv).

All three of these are much slower than $C(iv)$, which is uniformly ergodic and very clearly outperforms all other choices.

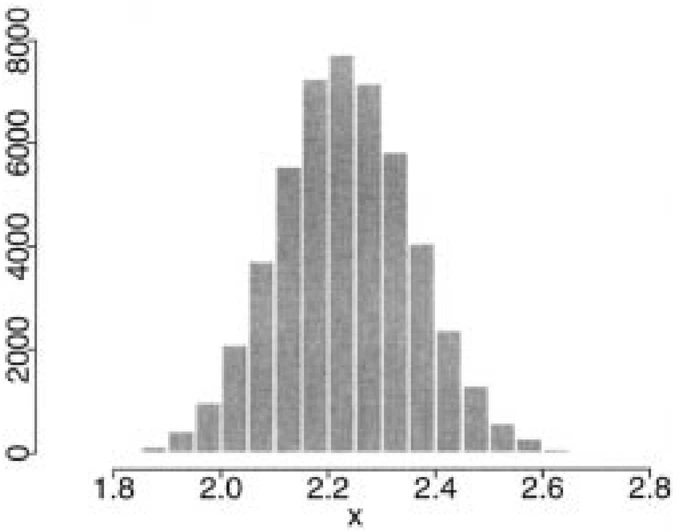

Figure 7. Histogram for *C(iv):* the sample mean is 2.232 and the sample variance is 0.016.

9. Metropolis Adjusted Multi-Dimensional Algorithms

We have considered in detail the effect of a Metropolis-Hastings adjustment to diffusion based algorithms only in one dimension. In practice, of course, the multi-dimensional case is of even greater interest. In this section we sketch the possible effects of adding an accept-reject step to discretizations of the multi-dimensional diffusions such as those described in Section 9 of Part I. These results are only sketched here, but they do indicate

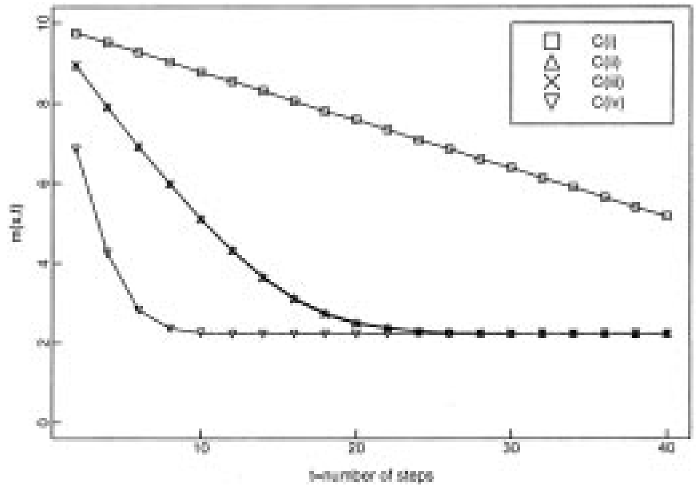

Figure 8. Conditional mean *m(x,t)* with $x(0) = 10$, for *C(i), C(ii), C(iii)* and $C(iv)$.

that the MADA approach can be expected to work well in higher dimensions under reasonable circumstances.

As for the one dimensional case, we write M_h for the Metropolised version of the discretization D_h, defined in (37) of Part I; and again call M_h the Metropolis-adjusted diffusion algorithm, or MADA chain. Our goal is to show that by choosing the candidate transition probability from a broader class of self-targeting candidates, the effectiveness of the M-H procedure can be substantially enhanced. We will illustrate this in just two examples.

Example: As in Section 9 of Part I, we consider

$$\pi(x) \propto \left(\frac{1}{2+x_1^2+x_2^2}\right)^2, x = (x_1, x_2) \in \mathbb{R}^2.$$

For this example, the Metropolis-adjusted Langevin algorithm (MALA) is not exponentially ergodic (see Theorem 4.3 of Roberts and Tweedie (1996)).

Now let us choose $a(x) = (2 + x_1^2 + x_2^2)I$, where I is the identity matrix, so that $b(x) = (-x_1, -x_2)'$ and the Jacobian $J(x)$ of $b(x)$ is

$$J(x) = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}.$$

In this case the local linearization scheme is defined by

$$D_h((k+1)h) | D_h(kh) = (x_1, x_2)' \sim N(\mu_{x,h}, \sigma^2_{x,h}),$$

where

$$\mu_{x,h} = \begin{bmatrix} x_1 \exp(-h) \\ x_2 \exp(-h) \end{bmatrix}, \quad a_{x,h} = \frac{2+x_1^2+x_2^2}{2}(\exp(-2h) - 1)I.$$

As in the one dimensional case, to obtain a geometric rate of convergence of the MADA chain we need some modification of the candidate D_h. In practice we would propose for this example to use both truncation of $a_{x,h}$ and a t distribution instead of the normal distribution above.

Example: Secondly, also as in Section 9 of Part I, we consider

$$\pi(x) \propto \exp(-x_1^4 - x_2^4 - x_1^2 x_2^2 - x_1^2 - x_2^2), \; x = (x_1, x_2) \in \mathbb{R}^2.$$

For this example, MALA is not exponentially ergodic from Theorem 4.2 of Roberts and Tweedie (1996).

In contrast, if we choose the local linearization scheme for the Langevin diffusion (Section 9 of Part I), then the MADA algorithm is exponentially ergodic. We omit the proof, but illustrate the exponentially ergodic behavior of M_h in the following two figures. Figure 9 is a trace plot of the steps taken by $M_h(kh)$ for $h = 0.1$, $k = 1, \ldots, 5000$ and a starting point $(100, -40)$. The arrows indicate the end of each step. Figure 10 is a trace plot of the steps taken by $M_h(kh)$ for $k = 301, \ldots, 5000$. It is clear from Figure 9 that the

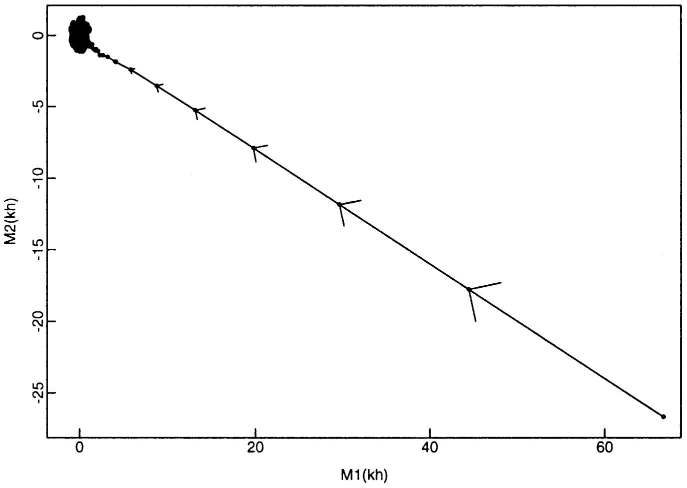

Figure 9. Trace plot of 5000 values from $M_h(kh)$ with $h = 0.1$. This illustrates the speed of convergence from a distant starting point.

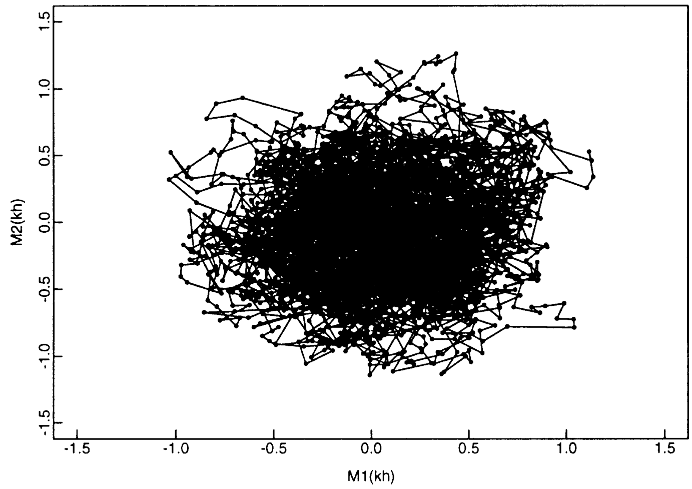

Figure 10. Trace plot of $M_h(kh)$ for $k = 301, \ldots, 5000$. This shows the shape of the target density π.

process M_h hits a neighborhood of [0,0] (the mode of π) rapidly, and then proceeds to approximate the target density π. Over a longer run of 50,000, the rate of rejection for this example is just 1243/50000, so that clearly, this scheme performs better than the MALA scheme.

Acknowledgment

We are grateful to the referees for a very careful review which improved the original presentation considerably.

References

J. E. Besag, Comments on ‘‘Representations of knowledge in complex systems,’’ by U. Grenander and M. I. Miller, *J. Roy. Statist. Soc. Ser. B* 56, 1994.

J. E. Besag and P. J. Green, ‘‘Spatial statistics and Bayesian computation (with discussion),’’ *J. Roy. Statist. Soc. Ser. B* vol. 55 pp. 25–38, 1993.

J. E. Besag, P. J. Green, D. Higdon, and K. L. Mengersen, ‘‘Bayesian computation and stochastic systems (with discussion),’’ *Statistical Science* vol. 10 pp. 3–66, 1995

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.

A. Gelman, G. O. Roberts, and W. R. Gilks, *Efficient Metropolis jumping rules*, In Bayesian statistics 5, ed. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, Oxford University Press: New York, 1995.

W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, *Markov Chain Monte Carlo in practice*, Chapman and Hall: London, 1996.

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

P. E. Kloeden and E. Platen, *Numerical solution of stochastic differential equations*, Springer-Verlag, Berlin, 1992.

K. L. Mengersen and R. L. Tweedie, ‘‘Rates of convergence of the Hastings and Metropolis algorithms,’’ *Annals of Statistics* vol. 24 pp. 101–121, 1996.

N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller, ‘‘Equations of state calculations by fast computing machines,’’ *J. Chemical Physics* vol. 21 pp. 1087–1091, 1953.

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

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

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, 1996.

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

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, 1999.

L. Tierney, ‘‘Markov chains for exploring posterior distributions (with discussion),’’ *Ann. Statist.* vol. 22 pp. 1701–1762, 1994.

P. Tuominen and R. L. Tweedie, ‘‘Subgeometric rates of convergence of *f*-ergodic Markov chains,’’ *Adv. Appl. Probab.* vol. 26 pp. 775–798, 1994.