Bernoulli **2**(4), 1996, 341–363

Exponential convergence of Langevin distributions and their discrete approximations

GARETH O. ROBERTS[1]* and RICHARD L. TWEEDIE[2]

[1]*Statistical Laboratory, Department of Pure Mathematics and Mathematical Statistics, 16 Mill Lane, Cambridge CB2 1SB, UK*
[2]*Department of Statistics, Colorado State University, Fort Collins CO 80523, USA*

In this paper we consider a continuous-time method of approximating a given distribution π using the Langevin diffusion $d\mathbf{L}_t = d\mathbf{W}_t + \frac{1}{2}\nabla \log \pi(\mathbf{L}_t)dt$. We find conditions under this diffusion converges exponentially quickly to π or does not: in one dimension, these are essentially that for distributions with exponential tails of the form $\pi(x) \propto \exp(-\gamma|x|^{\beta})$, $0 < \beta < \infty$, exponential convergence occurs if and only if $\beta \geq 1$.

We then consider conditions under which the discrete approximations to the diffusion converge. We first show that even when the diffusion itself converges, naive discretizations need not do so. We then consider a 'Metropolis-adjusted' version of the algorithm, and find conditions under which this also converges at an exponential rate: perhaps surprisingly, even the Metropolized version need not converge exponentially fast even if the diffusion does. We briefly discuss a truncated form of the algorithm which, in practice, should avoid the difficulties of the other forms.

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

1. The Langevin method for Markov chain Monte Carlo methods

1.1. GEOMETRIC CONVERGENCE OF MARKOV CHAIN MONTE CARLO METHODS

There has recently been a real explosion in the use of the Hastings and Metropolis algorithms, which allow simulation of a probability density $\pi(\mathbf{x})$ which is known only up to a factor: that is, when only $\pi(\mathbf{x})/\pi(\mathbf{y})$ is known. This is especially relevant when π is the posterior distribution in a Bayesian context: see Besag and Green (1993), Besag *et al.* (1995), Mengersen and Tweedie (1996), Roberts and Tweedie (1996), Smith and Roberts (1993) and Tierney (1994) for a variety of approaches and properties of such methods, and their applications in statistical modelling.

* To whom correspondence should be addressed.

1350-7265 © 1996 Chapman & Hall

The class of Hastings–Metropolis algorithms is very broad. As a consequence, the user is often faced with a choice between a 'plain vanilla' algorithm such as the random walk Metropolis algorithm, where the candidate dynamics are chosen to be those of a random walk, independently of the target distribution, and a more 'shaped' candidate distribution, designed for the particular π in the problem. The former is easy to implement, and it can be demonstrated that the random walk algorithm has rather robust theoretical properties. A more problem-specific algorithm may converge more rapidly.

We will be concerned here with one such class of algorithms. Langevin algorithms, which are derived from diffusion approximations, use information about the target density (in the form of the gradient of $\log \pi$) to construct such a problem-specific proposal distribution. We study the convergence properties of these algorithms; and as a precursor to this, we consider the convergence properties of the diffusions themselves, which have recently been suggested as a continuous-time method of approach to this simulation problem (see Grenander and Miller 1994).

In this paper, our main aim will be to study geometric convergence properties of these algorithms. For a discussion of some of the stability properties enjoyed by geometrically ergodic chains in simulation, which motivate our evaluations, we refer the reader to Roberts and Tweedie (1996). We note, for example that such chains have central limit theorems and the like available, which makes it much easier to assess the algorithms.

One particularly important consequence of our work is that, whereas the genuinely continuous-time processes often perform well on large classes of target densities, the situation is much more delicate for the approximations which would be used in practice. In particular, naive discretizations of the continuous-time models may lose not only the geometric rates of convergence but also all convergence properties, even for quite standard densities π. We indicate a truncated and 'Metropolized' form of the discretization which, in practical circumstances, will avoid such rather surprising pathologies.

We do not consider here the problem of how to choose the scaling of such discrete approximations to diffusions. This problem is considered in Roberts and Rosenthal (1995a).

In order to describe the approach, it is useful to outline the standard construction of the Hastings and Metropolis algorithms (see Metropolis *et al.* 1953; Hastings 1970). These first consider a *candidate transition kernel* with densities $q(\mathbf{x},\mathbf{y}), \mathbf{x},\mathbf{y} \in \mathsf{X}$, which generates potential transitions for a discrete-time Markov chain evolving on X. Here we will generally think of X as a subset of $\mathbb{R}^k$ equipped with the Borel σ-field $\mathscr{B}$, and both $\pi(\mathbf{y})$ and $q(\mathbf{x},\mathbf{y})$ will be densities with respect to Lebesgue measure μ^{Leb}, although more general formulations are possible.

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

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

Thus actual transitions of the Hastings chain, which we denote by Φ_n, take place according to a law P with transition probability densities

$$p(\mathbf{x},\mathbf{y}) = q(\mathbf{x},\mathbf{y})\alpha(\mathbf{x},\mathbf{y}), \qquad \mathbf{y} \neq \mathbf{x}, \tag{2}$$

and with probability of remaining at the same point given by

$$r(\mathbf{x}) = P(\mathbf{x}, \{\mathbf{x}\}) = \int q(\mathbf{x},\mathbf{y})[1 - \alpha(\mathbf{x},\mathbf{y})]\mathrm{d}\mathbf{y}. \tag{3}$$

With this choice of α we have that π is invariant: that is, satisfies $\pi(A) = \int \pi(\mathbf{x})P(\mathbf{x},A)\mathrm{d}\mathbf{x}, \mathbf{x} \in \mathsf{X}, A \in \mathscr{B}$. Provided the chain is suitably irreducible and aperiodic, it is then standard (Meyn and Tweedie 1993a, Chapter 13; Roberts and Smith 1994) that the n-step transition probabilities, defined for each $n \geq 1$ by $P^n(\mathbf{x},A) = P(\Phi_n \in A|\Phi_0 = \mathbf{x})$, $\mathbf{x} \in \mathsf{X}, A \in \mathscr{B}$, converge to π in the total variation norm: that is, for π-almost all $\mathbf{x}$

$$\|P^n(\mathbf{x},\cdot) - \pi\| := \tfrac{1}{2} \sup_{A \in \mathscr{B}} |P^n(\mathbf{x},A) - \pi(A)| \to 0. \tag{4}$$

In this paper we consider special forms of the density q based on the Langevin diffusion model below, and find conditions leading to geometric convergence in (4): in the cases where $q(\mathbf{x},\mathbf{y}) = q(|\mathbf{x} - \mathbf{y}|)$ (the Metropolis algorithm) this has been addressed in Mengersen and Tweedie (1996) and Roberts and Tweedie (1996) and our results for the Langevin-based models can be compared with those.

Other recently proposed algorithms such as hybrid-Monte Carlo algorithms (see, for example, Neal 1994), are related to the class of Langevin algorithms, although in this paper we will content ourselves with a detailed study of the simplest kind of Langevin algorithm, constructed from the natural reversible diffusion process. Methods induced by non-reversible diffusion can certainly be analysed similarly to those considered in this paper, and will suffer similar problems caused by sensitivity to the tails of the target density. However, it is worth remarking that often methods induced by non-reversible methods can be shown to converge more quickly than their reversible counterparts (see, for example, Sheu 1992).

1.2. THE LANGEVIN DIFFUSION

The form of the candidate density which we study is derived from the *Langevin diffusion*, which is itself constructed so that in continuous time it converges to π under suitable regularity conditions. In principle this should be a good choice of q, since even before being 'Metropolized' using (1), the candidate chain approximates one with stationary distribution π. The improvement in using this algorithm rather than a simple random walk candidate is also considered in Roberts and Rosenthal (1995b), from different perspectives.

We assume that π is everywhere non-zero and differentiable so that $\nabla \log \pi(\mathbf{x})$ is well defined. Then the Langevin diffusion $\mathbf{L}_t$ is defined by the k-dimensional stochastic differential equation

$$\mathrm{d}\mathbf{L}_t = \mathrm{d}\mathbf{W}_t + \tfrac{1}{2}\nabla \log \pi(\mathbf{L}_t)\mathrm{d}t,$$

where $\mathbf{W}_t$ is k-dimensional standard Brownian motion. When π is suitably smooth, it can be shown that $\mathbf{L}_t$ has π as a stationary measure, and also that

$$\|P^t_{\mathrm{L}}(\mathbf{x},\cdot) - \pi\| \to 0 \tag{5}$$

for all $\mathbf{x}$, where here $P^t_{\mathrm{L}}(\mathbf{x},A) = P(\mathbf{L}_t \in A|\mathbf{L}_0 = \mathbf{x}), t \geq 0$: we give details in Section 2.1.

We will be interested in when the convergence in (5) is exponentially fast, and also in when such exponential convergence occurs for higher moments of the process: this last can be seen as a useful byproduct of our approach.

1.3. EXPONENTIAL EXAMPLES

Our results are probably most easily demonstrated by considering the following examples, which we use repeatedly.

1.3.1. The one-dimensional class $\mathscr{E}(\beta,\gamma)$

Suppose π is one-dimensional. We will say that $\pi \in \mathscr{E}(\beta,\gamma)$ if for some x_0, and some constants $\gamma > 0$ and $0 < \beta < \infty$, π takes the form

$$\pi(x) \propto \mathrm{e}^{-\gamma|x|^\beta}, \qquad |x| \geq x_0, \tag{6}$$

so that for $|x| > x_0$

$$\nabla \log \pi(x) = -\gamma\beta \operatorname{sgn}(x)|x|^{\beta-1} \tag{7}$$

We will also assume that π is smooth enough for $|x| \leq x_0$ so that any assumptions on differentiability are satisfied, although in practice we expect that these can be avoided in specific examples.

Then we shall see in Section 2.3 that the diffusion L_t converges to π exponentially quickly when $\pi \in \mathscr{E}(\beta,\gamma)$ if and only if $\beta \geq 1$: that is, if and only if the tails of π are no heavier than exponential. This is exactly similar behaviour to the symmetric or random walk algorithm as shown in Theorem 3.5 of Mengersen and Tweedie (1996).

1.3.2. The multidimensional exponential class $\mathscr{P}_m$

For higher-dimensional models we consider the exponential family $\mathscr{P}_m$ introduced and studied in the context of the random walk Metropolis algorithm in Roberts and Tweedie (1996), and consisting of sufficiently smooth densities with the form (at least for large $|\mathbf{x}|$)

$$\pi(\mathbf{x}) \propto \mathrm{e}^{-p(\mathbf{x})}, \tag{8}$$

where p is a polynomial of degree m of the following type. Decompose p as

$$p = p_m + q_{m-1}, \tag{9}$$

where q_{m-1} is a polynomial of degree $\leq m-1$, so that p_m consists of the full-degree terms in p. Then we say that $\pi \in \mathscr{P}_m$ if $p_m(\mathbf{x}) \to \infty$ as $|\mathbf{x}| \to \infty$: this is a positive definiteness condition, and we note that this condition requires that $m \geq 2$.

We will see that there is exponential convergence of the multidimensional diffusion if $\pi \in \mathscr{P}_m$. This behaviour is identical to that exhibited by the multidimensional random walk algorithm, as shown in Roberts and Tweedie (1996).

1.4. DISCRETE APPROXIMATIONS TO $\mathbf{L}_t$

In practice, of course, one implements a discrete approximation to the diffusion $\mathbf{L}_t$, and we will also consider when such discrete approximations converge to π, and when they do so geometrically fast. Two different 'h-approximations' are defined below for fixed $h > 0$: we shall see in Section 3 that the more naive approximation (ULA) may not even converge for light-tailed π (that is, when $\beta > 2$ in $\mathscr{E}(\beta, \gamma)$), and certainly need not always converge geometrically; and the Metropolized algorithm (MALA), although it must converge, can fail to do so geometrically quickly even when the diffusion itself converges at an exponential rate.

1.4.1. The unadjusted Langevin algorithm

The unadjusted Langevin algorithm (ULA) is a discrete-time Markov chain $\mathbf{U}_n$ which is the natural discretization of the ordinary Langevin diffusion $\mathbf{L}_t$. Any naive algorithm using (11) below might be constructed in this way, as in Parisi (1981) or Grenander and Miller (1994). We shall see that the algorithm may have some undesirable convergence properties, although since its implementation may involve less computational expense than some of its more robust alternatives, it may still have practical merit.

To form this chain, given $\mathbf{U}_{n-1}$, we simply construct $\mathbf{U}_n$ according to

$$N(\mathbf{U}_{n-1} + \tfrac{1}{2}\nabla \log \pi(\mathbf{U}_{n-1}), hI_k).$$

As noted by Besag (1994), this chain only approximately maintains the invariance of π: as a graphic example, if π is itself $N(0, 1)$ on $\mathbb{R}$, then when $h = 2$, we have each $U_n \sim N(0, 2)$ so that clearly if the discretization step h is this coarse then we get immediate 'convergence', but to a quite unintended distribution.

We show below in Section 3 that the ULA chain may in fact behave quite badly: for example, it may converge but not geometrically quickly even when the original diffusion is exponentially ergodic, or quite startlingly it may actually be a transient chain even though $\mathbf{L}_t$ has a very well-behaved invariant distribution.

1.4.2. The Metropolis-adjusted Langevin algorithm

Following Besag (1994), we therefore introduce a further modification, and follow the structure in (1) and (2) to construct a Metropolis-based Langevin algorithm (MALA). This is a Hastings–Metropolis chain $\mathbf{M}_n$ which uses ULA to construct the candidate chain. Thus $\mathbf{U}_n$ given $\mathbf{M}_{n-1}$ is first taken as a variable distributed as

$$N(\mathbf{M}_{n-1} + \tfrac{1}{2}h\nabla \log \pi(\mathbf{M}_{n-1}), hI_k)$$

Call this proposal density $q(\mathbf{M}_{n-1}, \mathbf{U}_n)$. Now carry out an accept/reject step, accepting $\mathbf{U}_n$ with probability

$$\alpha(\mathbf{M}_{n-1}, \mathbf{U}_n) = 1 \wedge \frac{\pi(\mathbf{U}_n)q(\mathbf{U}_n, \mathbf{M}_{n-1})}{\pi(\mathbf{M}_{n-1})q(\mathbf{M}_{n-1}, \mathbf{U}_n)}. \tag{10}$$

If $\mathbf{U}_n$ is accepted then set $\mathbf{M}_n = \mathbf{U}_n$, otherwise, let $\mathbf{M}_n = \mathbf{M}_{n-1}$.

By the Hastings construction as in (2) and (3), the MALA chain converges to π, in the sense that

$$\|P_{\mathrm{M}}^n(\mathbf{x}, \cdot) - \pi\| \to 0 \tag{11}$$

for π-almost all $\mathbf{x}$ where we write $P_{\mathrm{M}}^n(\mathbf{x}, A) = P(\mathbf{M}_n \in A | \mathbf{M}_0 = \mathbf{x})$: this follows since the chain is clearly μ^{Leb}-irreducible and aperiodic from Roberts and Tweedie (1996). As a minor but useful by-product of our results we show that in the geometrically ergodic case the convergence also holds from all starting points.

Our interest is again in finding conditions under which convergence in (11) occurs geometrically quickly and from every starting point. We will prove that (roughly speaking) when ULA is transient MALA is not exponentially ergodic; but that it is geometrically ergodic otherwise unless the tails of the target density are heavier than exponential.

1.4.3. The Metropolis-adjusted Langevin truncated algorithm

Finally, we mention briefly a simple adjustment to the algorithm which is designed to try to capture the best properties of both the random walk Metropolis algorithm, and the 'targeted' Langevin candidate ULA. We call this MALTA (the Metropolis-adjusted Langevin truncated algorithm). This revised algorithm involves replacing the first ULA approximation by choosing the *truncated* candidate distribution

$$\mathbf{T}_n \sim N(\mathbf{M}_{n-1} + R(\mathbf{M}_{n-1}), hI_k), \tag{12}$$

where the drift term is now

$$R(\mathbf{M}_n) = \frac{D\nabla \log \pi(\mathbf{x})}{2(D \vee |\nabla \log \pi(\mathbf{x})|)} \tag{13}$$

for some constant $D > 0$. The candidate jump $\mathbf{T}_n$ is then adjusted to ensure the correct stationary distribution holds, as in (10).

With MALTA, the chain has much more robust geometric ergodicity properties. We do not pursue a detailed analysis of MALTA, merely pointing out that the methods employed in this paper and in Roberts and Tweedie (1996) are readily transportable to the analysis of this algorithm.

2. Exponential convergence of the Langevin diffusion algorithm

2.1. GENERAL CONVERGENCE RESULTS

In this section we apply convergence properties of general diffusions to the Langevin diffusion. For the concepts of μ^{Leb}-irreducibility, aperiodicity and small sets, the reader should see Meyn and Tweedie (1993b).

Theorem 2.1. *Suppose that* $\nabla \log \pi(\mathbf{x})$ *is continuously differentiable and that, for some* $N, a, b < \infty$,

$$\nabla \log \pi(\mathbf{x}) \cdot \mathbf{x} \leq a|x|^2 + b, \qquad |\mathbf{x}| > N. \tag{14}$$

Then the Langevin diffusion $\mathbf{L}_t$ *satisfies the following:*

(a) *The diffusion is non-explosive,* μ^{Leb}*-irreducible, aperiodic, strong Feller and all compact sets are small.*
(b) *The measure* π *is invariant for* $\mathbf{L}$ *and, moreover, for all* $\mathbf{x}$,

$$\|P_{\text{L}}^t(\mathbf{x}, \cdot) - \pi\| \to 0. \tag{15}$$

Proof. Under (14) a simple comparison argument, comparing $|\mathbf{L}_t|$ with an appropriate Ornstein–Uhlenbeck process, demonstrates that the radial component of $\mathbf{L}_t$ is non-explosive. It follows that $\mathbf{L}_t$ is also non-explosive. From the conditions on π and the constant diffusion coefficient we then have that the diffusion drift is locally bounded. Therefore the chain is μ^{Leb}-irreducible and strong Feller, by a straightforward extension (which is possible by non-explosivity) of Theorem 2.1 of Bhattacharya (1978), which is due to Stroock and Varadhan. The strong Feller result plus the irreducibility gives that all compact sets are small (see Tweedie 1994, Theorem 5.1). The aperiodicity is then obvious since all skeleton chains are also μ^{Leb}-irreducible.

Under these conditions it follows that π is invariant for $\mathbf{L}_t$ from Section 5.4 of Ikeda and Watanabe (1989), that is, π is invariant for P_t since by construction it is invariant for the generator of the Langevin diffusion given by

$$\mathscr{A}_{\text{L}} f(\mathbf{x}) = (\tfrac{1}{2}\nabla \log \pi(\mathbf{x}))\nabla f(\mathbf{x}) + \tfrac{1}{2}\nabla^2 f(\mathbf{x}) \tag{16}$$

for any twice continuously differentiable function f. (These functions form a distribution-determining class for a non-explosive diffusion.)

Since the process has a stationary distribution it is at least recurrent, from Tweedie (1994, Theorem 2.3), and the continuity of the sample paths ensures that this extends to Harris recurrence. The total variation norm convergence in (15) then follows from Meyn and Tweedie (1993b, Theorem 6.1) for all $\mathbf{x}$, and we have the result. □

The operator $\mathscr{A}_{\text{L}}$, over a domain that contains at least all functions satisfying (16), is easily checked to be the *extended generator* of $\mathbf{L}_t$ as described in Davis (1993). For our purposes we do not need the exact form of the domain, but merely the form (16) of $\mathscr{A}_{\text{L}}$.

Note that (14) is certainly not necessary for non-explosivity. However, it should be appropriate for virtually all commonly encountered target densities.

The limiting result (15) is basically the result that justifies the use of the Langevin diffusion model: the drift on this diffusion is specifically designed to ensure convergence to π. We now turn to new results in this area that ensure that the limit is exponentially fast.

2.2. EXPONENTIAL ERGODICITY OF $\mathbf{L}_t$

We will use the following approach for exponential ergodicity (see Meyn and Tweedie 1993b; Down *et al.* 1995). When $V \geq 1$ is a measurable function on X, we define *V-uniform ergodicity* by requiring that for all $\mathbf{x}$

$$\|P^t_{\mathrm{L}}(\mathbf{x}, \cdot) - \pi\|_V \leq V(\mathbf{x})R\rho^t, \qquad t \geq 0, \tag{17}$$

for some $R < \infty$, $\rho < 1$, where

$$\|A\|_V = \sup_{\{f;|f|\leq V\}} \int f(\mathbf{x})A(\mathrm{d}\mathbf{x})$$

for any signed measure A. As is shown in Meyn and Tweedie (1993b), this strong form of exponential ergodicity is in fact implied (for some $V \geq 1$) by the seemingly simpler requirement that

$$\|P^t_{\mathrm{L}}(\mathbf{x}, \cdot) - \pi\| \leq R_{\mathbf{x}}\rho^t, \qquad t \geq 0, \tag{18}$$

for some $R_{\mathbf{x}} < \infty, \rho < 1$ and all $\mathbf{x}$.

We will use relationships between geometric ergodicity, 'exponential recurrence' and the existence of an exponential form of 'drift function' equation involving the generator of the process. Many more of these are given in Meyn and Tweedie (1993b), Down *et al.* (1995), and the full force of (17) is described there in detail. These results, in this specific case, give us the following theorem.

Theorem 2.2. *Suppose that* $\mathbf{L}_t$ *satisfies the conditions of Theorem* 2.1.

(a) *The Langevin diffusion* $\mathbf{L}_t$ *is V-uniformly ergodic for any twice continuously differentiable* $V \geq 1$, *such that*

$$\mathscr{A}_{\mathrm{L}}V \leq -cV + b\mathbb{1}_C \tag{19}$$

for some constants $b, c > 0$, *and some compact non-empty set C, where the functional operator* $\mathscr{A}_{\mathrm{L}}$ *is defined as in* (16).

(b) *If the Langevin diffusion is exponentially ergodic in the sense that* (18) *holds, then for some compact non-empty set C there exist constants* $\kappa > 1, \delta > 0$, *such that*

$$\sup_{\mathbf{x}\in C} \mathrm{E}[\kappa^{\tau^\delta_C}] < \infty \tag{20}$$

where

$$\tau^\delta_C = \inf\{t \geq \delta : \mathbf{L}_t \in C\} \tag{21}$$

for any $\mathbf{x} \in$ X.

Proof. Recall that a set C is small if there exists a positive constant ε, an integer n_0 and a probability measure ν such that

$$P^{n_0}(\mathbf{x}, \cdot) \geq \varepsilon\nu(\cdot).$$

Since C is small, (a) follows from Theorem 5.2 of Down *et al.* (1995) or Theorem 6.1 of Meyn and Tweedie (1993b): note that since the chain is non-explosive from Theorem 2.1(a) above, we do not need V to be 'normlike' as in that theorem.

To see (b), note that if (18) holds, then any δ-skeleton is also geometrically ergodic. The hitting time on C for the δ-skeleton is at least as large as τ_C^δ and since any compact non-empty C is small for the skeleton because $\mathbf{L}_t$ is strong Feller, we have that (20) follows from Theorem 15.0.1 of Meyn and Tweedie (1993a). □

We now use these results to classify the behaviour of $\mathbf{L}_t$ in much more concrete terms.

Theorem 2.3. *Suppose there exists $S > 0$ such that $|\pi(\mathbf{x})|$ is bounded for $|\mathbf{x}| \geq S$. Then a sufficient condition for the Langevin diffusion $\mathbf{L}_t$ to be exponentially ergodic is that there exists $0 < d < 1$ such that*

$$\liminf_{|\mathbf{x}| \to \infty} (1 - d)|\nabla \log \pi(\mathbf{x})|^2 + \nabla^2 \log \pi(\mathbf{x}) > 0. \tag{22}$$

Proof. To use Theorem 2.2(a), we try the test function $V = \pi^{-d}$ for some fixed $0 < d < 1$. Then

$$2\mathscr{A}_{\mathrm{L}} V = V(|\nabla \log \pi|^2 (d^2 - d) - d\nabla^2 \log \pi). \tag{23}$$

Since V is bounded away from zero for large $|\mathbf{x}|$ by hypothesis, it is sufficient for (19) that (22) holds for the chosen d. □

In the other direction we can show the following:

Theorem 2.4. *If $|\nabla\, log\, \pi(\mathbf{x})| \to 0$, then $\mathbf{L}_t$ is not exponentially ergodic.*

Proof. Suppose that $\mathbf{L}_t$ is exponentially ergodic. Then from Theorem 2.2(b) there exists a compact set C such that (20) holds. Choose R large enough so that $|\nabla \log \pi(\mathbf{x})| \leq 2(\log \kappa)^{1/2}$ for $|\mathbf{x}| \geq R$, and fix

$$S \geq \sup_{\mathbf{x} \in C} |\mathbf{x}| \vee R.$$

Now consider $\tau_C = \tau_C^0$ for starting points $|\mathbf{y}| \geq 2S$. Also define the radial process for $\mathbf{L}_t, Z_t = |\mathbf{L}_t|$ which satisfies the stochastic differential equation

$$\mathrm{d}Z_t = \mathrm{d}W_t + a(\mathbf{L}_t)\mathrm{d}t$$

for some standard Brownian motion W, and a drift coefficient $a(\cdot)$ satisfying $a(\mathbf{L}_t > -(\log \kappa)^{1/2}$ for $|\mathbf{L}_t| > S$. So if B_t denotes $W_t - (\log \kappa)^{1/2} t$, and $\sigma(X)$ denotes the first hitting time of S by any process X, then $\sigma(Z) \geq \sigma(B)$ almost surely. Therefore,

$$\begin{aligned} P_{\mathbf{y}}[\tau_C > t] &\geq P_{\mathbf{y}}[\sigma(Z) > t] \\ &\geq P_{\mathbf{y}}[\sigma(B) > t] \\ &\geq \Phi\left(\frac{-S + t(\log \kappa)^{1/2}}{\sqrt{t}}\right) - \mathrm{e}^{2S(\log \kappa)^{1/2}} \Phi\left(\frac{-S - t(\log \kappa)^{1/2}}{\sqrt{t}}\right), \end{aligned} \tag{24}$$

where Φ denotes the standard normal distribution function, and the final inequality follows from the Bachelier–Lévy formula (see, for example, Lerche 1986). Denoting the density of $\sigma(B)$ by f, it is therefore easy to check that

$$\frac{\log f(t)}{t} \to -\frac{\log \kappa}{2}$$

which contradicts (20). Therefore the process is not exponentially ergodic. □

2.3. EXPONENTIAL EXAMPLES

For distributions π with essentially exponentially decreasing 'tails' the results above give a fairly thorough classification of the behaviour of the Langevin algorithm.

2.3.1. The one-dimensional class $\mathscr{E}(\beta, \gamma)$

Let us apply these results to the class of densities $\mathscr{E}(\beta, \gamma)$ given by (6). As in (7), we have that

$$(1-d)|\nabla \log \pi(x)|^2 + \nabla^2 \log \pi(x) = (1-d)\alpha^2\beta^2 x^{2\beta-2} - \alpha\beta(\beta-1)x^{\beta-2}.$$

Therefore:

(a) for $1 \leq \beta < \infty$, the diffusion is exponentially ergodic by Theorem 2.3;
(b) for $0 < \beta < 1$, $|\nabla \log \pi(x)| \to 0$ so that, by Theorem 2.4, the diffusion is not exponentially ergodic.

This mimics exactly the behaviour found in Section 3 of Mengersen and Tweedie (1996) for Metropolis algorithms in one dimension: the exponential convergence is governed by the tails of the target density, and occurs if and only if those tails decrease at least exponentially quickly. We remark that there are a number of ways in which the Langevin diffusion can be considered to be the weak limit of an appropriate sequence of Metropolis algorithms (see, for example, Roberts *et al.* 1994).

2.3.2. The multidimensional class $\mathscr{P}_m$

For higher-dimensional models our results are not as complete. We consider the exponential class $\mathscr{P}_m$ introduced in (8).

Now, it is easy to show that, by the positive definiteness condition,

$$\liminf_{|\mathbf{x}| \to \infty} \frac{|\nabla \log \pi(\mathbf{x})|^2}{|\nabla^2 \log \pi(\mathbf{x})|} = \infty \tag{25}$$

and

$$\liminf_{|\mathbf{x}| \to \infty} (1-d)|\nabla \log \pi(\mathbf{x})|^2 > 0 \tag{26}$$

for all $0 < d < 1$. Exponential convergence of the Langevin diffusion for all $\pi \in \mathscr{P}_m$ therefore follows from Theorem 2.3.

Continuing to follow the type of example contained in Roberts and Tweedie (1996), we examine the situation where π has the exponential form (8), but where the positive definiteness condition does not hold. Specifically, consider such a two-dimensional density where

$$p(z,y) = \alpha z^2 + z^2y^2 + y^2.$$

Note that here the dominant term is z^2y^2, which does not go to ∞ along either of the rays $(z,0)$ or $(0,y)$. Now, $|\nabla p|^2 = 4z^2(\alpha + y^2)^2 + 4y^2(1+z^2)^2$, and $\nabla^2 p = 2(\alpha + 1 + z^2 + y^2)$. Hence $|\nabla p|^2$ will dominate $\nabla^2(-p)$ except along the coordinate axes.

Setting $y = 0$, therefore, we find that if $2(1-d)\alpha^2 > 1$ for some $0 < d < 1$, then we can ensure exponential convergence by Theorem 2.3: hence $\alpha > 1/\sqrt{2}$ is sufficient for exponential ergodicity.

In contrast, other polynomials failing to satisfy the positive definiteness condition of $\mathscr{P}_m$ above will never satisfy the hypothesis of Theorem 2.3: consider, for instance,

$$p(x,y) = x^2 + x^2y^4 + y^2, \tag{27}$$

which can never achieve (22) along the ray $y = 0$. Hence, if we are to show that such models are exponentially ergodic, some other method of proof needs to be found.

3. The unadjusted Langevin algorithm

3.1. CONVERGENCE OF ULA

The naive way to implement the diffusion algorithms in practice is to use the unadjusted Langevin algorithm (ULA): that is, use a first-order Gaussian approximation to the diffusion distributions on a grid of size h and construct

$$\mathbf{U}_n|\mathbf{U}_{n-1} \sim N(\mathbf{U}_{n-1} + \tfrac{1}{2}h\nabla \log \pi(\mathbf{U}_{n-1}), hI_k) \tag{28}$$

as in Parisi (1981).

This is easily seen to be μ^{Leb}-irreducible and weak Feller, provided $\nabla \log \pi(\mathbf{x})$ is continuous, and hence, as in Chapter 6 of Meyn and Tweedie (1993a), all compact sets are small, and it suffices for geometric ergodicity from Theorem 15.0.1 of Meyn and Tweedie (1993a) to find a function $V \geq 1$ such that for some compact set C and some $\lambda < 1, b < \infty$,

$$\int P(\mathbf{x}, \mathrm{d}\mathbf{y})V(\mathbf{y}) \leq \lambda V(\mathbf{x}) + b\mathbb{1}_C(\mathbf{x}). \tag{29}$$

Note that this is the discrete version of (19).

Under some circumstances the discrete approximation is well behaved and under others it is not. We will only describe the range of behaviour when the chain is on $\mathbb{R}$ rather than in higher dimensions: since the ULA is only an approximation, and since it can be shown to converge in general to a stationary distribution which is *not* π, there is a limit to the extent of our interest in the rate of such convergence. However, it is the simple and natural discretization and thus its behaviour warrants careful analysis.

Conditions under which the chain can be evaluated are mostly easily described when, for some fixed d, the limits

$$\lim_{x\to\infty} \tfrac{1}{2}h\nabla \log \pi(x)x^{-d} = S_d^+ \tag{30}$$

and

$$\lim_{x\to-\infty} \tfrac{1}{2}h\nabla \log \pi(x)|x|^{-d} = S_d^- \tag{31}$$

exist. Clearly the results below can easily be extended to other situations where S_d^+, S_d^- are replaced by lim sup or lim inf in appropriate combinations, and where the exponent d varies for positive and negative x.

Theorem 3.1. *The ULA chain U_n or $\mathbb{R}$ is geometrically ergodic if one of the following holds:*

(a) *for some $d \in [0,1)$, both $S_d^+ < 0$ and $S_d^- > 0$ exist;*
(b) *for $d = 1$, both $S_d^+ < 0$ and $S_d^- > 0$ exist and*

$$(1 + S_d^+)(1 - S_d^-) < 1. \tag{32}$$

Proof. In the first instance for (a), consider $d = 0$. We compare the ULA model with a random walk $[0,\infty)$ for positive x: the result follows by symmetry for negative x.

We have that, if W_1 is an $N(0,h)$ variable, then, for all large enough positive $U_0 = x$,

$$U_1 = U_0 + S_x + W_1,$$

where $S_x \in (S_d^+ - \varepsilon, S_d^+ + \varepsilon)$ and ε is small enough that $0 > S_d^+ + \varepsilon$. We can then show, exactly as in the argument in Meyn and Tweedie (1993a, pp. 318–319), that for some sufficiently small s the function $V(y) = \mathrm{e}^{s|y|}$ satisfies (29), and geometric ergodicity follows.

For $d \in (0,1)$ the argument is virtually identical. The mean increment at x is now even more negative, with the new mean position being approximately $x + S_d^+ x^d$; nevertheless, since $d \in (0,1)$, for large enough x this will still be sufficiently positive that the truncation approximations needed to emulate the proof above still go through, and we omit details.

For (b), note again that, for large $U_0 > x_0 > 0$, we can write

$$U_1 = U_0[1 + S_x] + W_1,$$

while for $U_0 < -x_0 < 0$ we have

$$U_1 = U_0[1 - B_x] + W_1$$

where now $S_x \in (S_d^+ - \varepsilon, S_d^+ + \varepsilon) < 0, B_x \in (S_d^- - \varepsilon, S_d^- + \varepsilon) < 0$ and this time we choose ε such that $(1 + S_x)(1 - B_x) < 1$, which is possible from (32).

Now, rather than the random walk, the analogue we use in this case is the SETAR (self-exciting threshold autoregression) model in nonlinear time series. Following the proof of Proposition 11.4.5 of Meyn and Tweedie (1993a) (see also Meyn and Tweedie, 1993a, p. 505), let us choose $V(x) = ax, x > x_0$ and $V(x) = b|x|, x \le -x_0$. Then we have that (29) again holds because of (32), and so the chain is again geometrically ergodic. □

The interesting case in (b) is perhaps not when $S_d^+ \in (-2, 0), S_d^- \in (0, 2)$, for then both negative and positive sides of the algorithm behave like simple autoregressions; rather, it is when one of the values (say, S_d^+) is strictly less than -2, so that from the positive side one makes a long negative jump; but then the next value satisfies (32) and so compensates either by drifting back towards zero like a positive-coefficient simple autoregression or by making a jump back over the origin while being centred around a smaller absolute value than the initial point, after these two steps.

Thus we have that geometric ergodicity is possible in a range of cases of ULA.

Conversely, we can show that in selected cases the ULA chain is not geometrically ergodic, and, of rather more concern, may not even be ergodic at all. We illustrate this with two different results.

Theorem 3.2. (a) *The ULA chain on* $\mathbb{R}$ *is ergodic but not geometrically ergodic if, for some* $d \in (-1, 0)$, *both* $S_d^+ < 0$ *and* $S_d^- > 0$ *exist.*

(b) *The ULA chain on* $\mathbb{R}$ *is not ergodic, and is indeed transient, if, for some* $d > 1$, *both* $S_d^+ < 0$ *and* $S_d^- > 0$ *exist, or if, for* $d = 1$, *both* $S_d^+ < -2$ *and* $S_d^- > 2$ *exist.*

Proof. The second, transient, case (b) is rather obvious, although disturbing. When $d = 1$, from a large positive value of x the next position is approximately $(1 + S_d^+)x < -x$, and then the next oscillation is to a positive but still more extreme value, and so on; while for $d > 1$ and x sufficiently large the same pattern repeats but more strongly. Again the formal verification follows the proof of transience for the SETAR model: see Meyn and Tweedie (1993a, p. 222).

The other case (a) requires rather more subtlety.

To see the chain cannot be geometrically ergodic, we use Theorem 15.0.1 of Meyn and Tweedie (1993a). Note first that, for large x, from $U_0 = x$ the expected increment is approximately $S_d^+ x^d \to 0$, $x \to \infty$. Thus for any $\varepsilon > 0$, there is a large enough x_0 that the mean next step is to the right of $x - \varepsilon$ when we start above x_0. From $x \geq x_0$ the ULA chain is therefore always stochastically larger than a random walk with increments $N(-\varepsilon, h)$, at least until the first time to hit the set $C_0 = (-\infty, x_0)$. The time taken by the random walk to hit C_0 is also correspondingly longer than that of a Brownian motion with constant drift $-\varepsilon/h$, since the random walk can be viewed as the embedded h-skeleton of the Brownian motion. Finally, since ε is arbitrary we can use the Bachelier–Lévy formula as in (24) to show that these hitting times do not have exponential tails. So none of the chains above are geometrically ergodic.

And yet, in this case the ULA model is indeed ergodic. We take $V(x) = x^2$ in the Foster drift criterion (see, for example, Meyn and Tweedie 1993a, p. 262) to show this: following the argument in Lamperti (1963), and being careful with truncations, we see that ergodicity will follow if, writing $\mu_k(x) = \mathrm{E}[(U_{n+1} - U_n)^k | U_n = x]$, we can show

$$2x\mu_1(x) + \mu_2(x) \leq -\varepsilon, \tag{33}$$

for large enough $x > 0$ and some $\varepsilon > 0$ (with a symmetric drift for negative x). But now we have that for $d \in (-1, 0)$, $2x\mu_1(x) \approx 2S_d^+ x^{1+d}$, which is increasingly large and negative, while $\mu_2(x) \approx h + [S_d^+ x^d]^2 \to h$, for large x. Thus ergodicity is indeed maintained in this model. □

Note that if, for $d = 0$, we have $S_d^+ = 0$ or $S_d^- = 0$, then by the same proof there can be no geometric ergodicity. Depending on finer structure, one may find that the chain is ergodic from a condition such as (33); or conversely, is not ergodic if the mean drift becomes very quickly close to zero, so the chain behaves too closely like a zero-drift random walk.

Clearly other behaviour is possible for sufficiently pathologically constructed tail behaviour of π: as in the SETAR examples in Meyn and Tweedie (1993a), various combinations of null recurrence and positive recurrence may occur. We do not pursue this here: our goal was to show that the ULA model is not guaranteed to behave well, more or less independently in most cases of the choice of h.

We conclude from this analysis that the ULA model is not to be recommended without considerable care and knowledge of the behaviour of π.

3.2. ULA FOR THE ONE-DIMENSIONAL CLASS $\mathscr{E}(\beta, \gamma)$

Again for a one-dimensional distribution $\pi \in \mathscr{E}(\beta, \gamma)$, as in (6), we have a rather complete evaluation of the categories above. We find, since $\nabla \log \pi(x) = -\gamma\beta x^{\beta-1}$ for positive x:

(a) For $0 < \beta < 1$, when the tails are heavy, it follows from Theorem 3.2 that the ULA chain is ergodic but not exponentially ergodic; thus the ULA approximation mimics the behaviour of the diffusion.

(b) For $1 \leq \beta < 2$, the ULA is geometrically ergodic, again as for the diffusion; the exponential case $\beta = 1$ follows from Theorem 3.1(a) with $d = 0$, and the cases between exponential and Gaussian tails follow from Theorem 3.1(a) with $d \in (0, 1)$.

(c) For $\beta = 2$, when the tails are like those of a symmetric Gaussian, the behaviour is surprisingly mixed. From Theorem 3.1(b) we have that if $\gamma h < 2$ then the chain is exponentially ergodic; but from Theorem 3.2, if $\gamma h > 2$ then the chain is transient. Thus the choice of h is crucial here, as it is not in most models. We have not classified the chain at $\gamma h = 2$, though we would expect that this might be null recurrent.

(d) For $\beta > 2$, when the tails are light, it follows from Theorem 3.2 that (perhaps surprisingly at first sight) the ULA chain is transient: this is in contrast to the case with the random walk Metropolis algorithm in Mengersen and Tweedie (1996), where such chains are shown to be geometrically ergodic. In this situation the ULA over-corrects for the light tails by throwing the chain in increasing oscillations. The underlying diffusion, since it has continuous sample paths, cannot of course have such aberrant behaviour even though it does 'drift' very quickly from the tail regions, but is then forced to 'slow down' in the centre of the space and is exponentially ergodic, as shown in Section 2.

4. The Metropolis-adjusted Langevin algorithm

4.1. CONDITIONS FOR EXPONENTIAL CONVERGENCE OF MALA

The need for some form of correction of the simple ULA models is now quite apparent since even the finest discrete approximation of the Langevin diffusion can lead to the approximating Markov chain behaving radically differently from the diffusion process it is trying to approximate.

One way of preserving the stationarity of π in any discrete approximation is to introduce a Hastings–Metropolis accept–reject step, and, for any fixed $h > 0$, this algorithm is described in Section 1.4.

In order to develop a positive result on geometric convergence of MALA algorithms, especially in higher dimensions, we need some constraints on the way in which proposed moves are accepted. The following covers many standard examples, although in other situations variations on the approach will be needed: we do not strive for total generality here.

We write $A(\mathbf{x})$ for the acceptance region of MALA from the point $\mathbf{x}$: that is, $A(\mathbf{x})$ is the region in which proposed moves are always accepted. Thus

$$A(\mathbf{x}) = \{\mathbf{y} : \pi(\mathbf{x})q(\mathbf{x},\mathbf{y}) \leq \pi(\mathbf{y})q(\mathbf{y},\mathbf{x})\}, \tag{34}$$

where q is the ULA density for a candidate move, given by (28). We also write $R(\mathbf{x})$ for $A(\mathbf{x})^{\mathrm{c}}$, the 'potential rejection' region.

Denote by $I(\mathbf{x})$ the points interior to $\mathbf{x}$: that is,

$$I(\mathbf{x}) = \{\mathbf{y}; |\mathbf{y}| \leq |\mathbf{x}|\}.$$

We say that $A(\cdot)$ *converges inwards in* q if

$$\lim_{|\mathbf{x}|\to\infty} \int_{A(\mathbf{x})\Delta I(\mathbf{x})} q(\mathbf{x},\mathbf{y})\mathrm{d}\mathbf{y} = 0, \tag{35}$$

where we denote the symmetric difference $(A \cup B)\backslash(A \cap B)$ by $A\Delta B$. For densities π that have this property we have a simple and intuitively plausible condition that guarantees geometric ergodicity.

Theorem 4.1. *Suppose that* $c(\mathbf{x}) = \mathbf{x} + \frac{1}{2}h\nabla \log \pi(\mathbf{x})$ *is the mean 'next candidate position' and that*

$$\eta \equiv \liminf_{|\mathbf{x}|\to\infty}(|\mathbf{x}| - |c(\mathbf{x})|) > 0. \tag{36}$$

Assume $A(\cdot)$ *converges inwards in* q. *If* $V_s(\mathbf{x}) = \mathrm{e}^{s|x|}$, *then the MALA chain is* V_s*-uniformly ergodic for* $s < 2h\eta$.

Proof. We will check (29) for the function V_s. Splitting the integral over obvious regions, we find:

$$\begin{aligned} PV_s(\mathbf{x})/V_s(\mathbf{x}) &= (2\pi h)^{-k/2} \int_{A(\mathbf{x})} \exp\left\{-\frac{1}{2h}|\mathbf{y} - c(\mathbf{x})|^2 + s(|\mathbf{y}| - |\mathbf{x}|)\right\}\mathrm{d}\mathbf{y} \\ &\quad + (2\pi h)^{-k/2} \int_{R(\mathbf{x})} \exp\left\{-\frac{1}{2h}|\mathbf{y} - c(\mathbf{x})|^2 + s(|\mathbf{y}| - |\mathbf{x}|)\right\}\alpha(\mathbf{x},\mathbf{y})\mathrm{d}\mathbf{y} \end{aligned}$$

$$+ (2\pi h)^{-k/2}\int_{R(\mathbf{x})}\exp\left\{-\frac{1}{2h}|\mathbf{y}-c(\mathbf{x})|^2\right\}(1-\alpha(\mathbf{x},\mathbf{y}))\mathrm{d}\mathbf{y},$$

$$\leq (2\pi h)^{-k/2}\int_{\mathbb{R}^k}\exp\left\{-\frac{1}{2h}|\mathbf{y}-c(\mathbf{x})|^2+s(|\mathbf{y}|-|\mathbf{x}|)\right\}\mathrm{d}\mathbf{y}$$

$$+ (2\pi h)^{-k/2}\int_{R(x)\cap I(\mathbf{x})}\exp\left\{-\frac{1}{2h}|\mathbf{y}-c(\mathbf{x})|^2\right\}\mathrm{d}\mathbf{y}. \tag{37}$$

Now $\exp(s(|\mathbf{x}|-|c(\mathbf{x})|))$ times the first term on the right-hand side asymptotes to $\exp(s^2/2h)$, so that lim sup of the first term is less than 1. Moreover, the second term asymptotically converges to zero since $A(\cdot)$ converges inwards in q. Hence

$$\limsup_{|\mathbf{x}|\to\infty}\frac{PV_s(\mathbf{x})}{V_s(\mathbf{x})} < 1,$$

so that, noting that compact sets are small from Roberts and Tweedie (1996), geometric convergence in V_s-norm is guaranteed by Theorem 15.0.1 of Meyn and Tweedie (1993a). □

We note from the steps of the proof that condition (35) is far from necessary. In fact it is quite easy to relax (35) to find that the following condition is sufficient for geometric convergence in V_s-norm for sufficiently small s: there exists $\varepsilon > 0$ such that

$$I(\mathbf{x}) = \{\mathbf{y} : \alpha(\mathbf{x},\mathbf{y}) \geq \varepsilon\}$$

asymptotically has q-measure zero. We omit the details of this.

Condition (36) is implied by some obvious conditions on $\nabla\log\pi(\mathbf{x})$. For example, the following two conditions together imply (36):

$$\limsup_{|\mathbf{x}|\to\infty}\mathbf{n}_{\mathbf{x}}\cdot\nabla\log\pi(\mathbf{x}) < 0, \tag{38}$$

and

$$\lim_{|\mathbf{x}|\to\infty}\mathbf{n}_{\mathbf{x}}\cdot\nabla\log\pi(\mathbf{x}) - |\mathbf{x}| = 0, \tag{39}$$

where $\mathbf{n}_{\mathbf{x}} = \mathbf{x}/|\mathbf{x}|$ denotes the outward normal vector at $\mathbf{x}$. Other conditions are clearly available, although they are more usefully pursued when specific models are being considered.

The 'inward converging' property is also often easy to evaluate. It is possible to rewrite $A(\mathbf{x})$ as

$$A(\mathbf{x}) = \left\{\mathbf{y} : \int_{\mathbf{y}}^{\mathbf{x}}\nabla\log\pi(\mathbf{z})\mathrm{d}\mathbf{z} \leq \tfrac{1}{2}(\mathbf{x}-\mathbf{y})\cdot(\nabla\log\pi(\mathbf{x})+\nabla\log\pi(\mathbf{y})) + \frac{h}{8}(|\nabla\log\pi(\mathbf{x})|^2 - |\nabla\log\pi(\mathbf{y})|^2)\right\}. \tag{40}$$

Here the line integral can be interpreted along any curve between $\mathbf{x}$ and $\mathbf{y}$, but most conveniently along the straight line, where this expression for $A(\cdot)$ offers interpretation in terms of notions of convexity.

We will illustrate this interpretation in Section 4.3 below. Before doing so, we consider conditions under which MALA does not converge geometrically quickly.

4.2. CONDITIONS FOR NON-EXPONENTIAL CONVERGENCE OF MALA

The following theorem implies (essentially) that when ULA is transient (for example, when the tails of π are lighter than Gaussian), MALA is not exponentially ergodic.

Theorem 4.2. *If π is bounded, and*

$$\liminf_{|\mathbf{x}|\to\infty} \frac{|\nabla \log \pi(\mathbf{x})|}{|\mathbf{x}|} > \frac{4}{h} \tag{41}$$

then the MALA chain is not exponentially ergodic.

Proof. Let $r(\mathbf{x})$ to be the rejection probability from each point as in (3): we know from Roberts and Tweedie (1996) that if $\operatorname{ess\,sup} r(\mathbf{x}) = 1$ then the algorithm is not geometrically ergodic.

Assume, then, for contradiction that the algorithm is exponentially ergodic, so that by the continuity of π and q, there exists $\varepsilon > 0$ such that

$$r(\mathbf{x}) \leq 1 - \varepsilon,$$

for all $\mathbf{x} \in \mathbb{R}^k$. Now choose T such that

$$P[|N(\mathbf{0}, hI_k)| \geq T] \leq \frac{\varepsilon}{2},$$

and S large enough such that there exists $M > 4/h$ such that

$$\inf_{|\mathbf{x}|\geq S} \frac{|\nabla \log \pi(\mathbf{x})|}{|\mathbf{x}|} \geq M$$

and $(M - 4/h)S > T$. Define

$$B(\mathbf{x}) = \{\mathbf{y}; |\mathbf{y} - \mathbf{x} - \tfrac{1}{2}h\nabla \log \pi(\mathbf{x})| \leq T\} :$$

we will show that

$$\lim_{|\mathbf{x}|\to\infty} \sup_{\mathbf{y}\in B(\mathbf{x})} \frac{q(\mathbf{y}, \mathbf{x})}{q(\mathbf{x}, \mathbf{y})} = 0. \tag{42}$$

To see this, first note that the denominator in (42) is uniformly bounded away from zero. Therefore it is sufficient to consider the numerator. Now for $|\mathbf{x}| > S$,

$$\begin{aligned} q(\mathbf{y}, \mathbf{x}) &= (2\pi h)^{-k/2} \exp\left\{-\frac{1}{2h}|\mathbf{x} - \mathbf{y} - \tfrac{1}{2}h\nabla \log \pi(\mathbf{y})|^2\right\} \\ &\leq (2\pi h)^{-k/2} \exp\left\{-\frac{1}{2h}(|\tfrac{1}{2}h\nabla \log \pi(\mathbf{y})|^2 - |\mathbf{x} - \mathbf{y}|^2)\right\} \\ &\leq (2\pi h)^{-k/2} \exp\left\{-\frac{1}{2h}((\tfrac{1}{2}hM|\mathbf{y}|)^2 - 4|\mathbf{y}|^2)\right\} \end{aligned}$$

since $|\mathbf{y}| > |\mathbf{x}|$ for $\mathbf{y} \in B(\mathbf{x})$. Therefore since $[\frac{1}{2}hM]^2 - 4 > 0$,

$$\begin{aligned}\sup_{\mathbf{y}\in B(\mathbf{x})} q(\mathbf{y},\mathbf{x}) &\le \sup_{\mathbf{y}\in B(\mathbf{x})} (2\pi h)^{-k/2} \exp\left\{-\frac{1}{2h}([\tfrac{1}{2}hM]^2 - 4)|\mathbf{y}|^2\right\} \\ &\le (2\pi h)^{-k/2} \exp\left\{-\frac{1}{2h}([\tfrac{1}{2}hM]^2 - 4)|\mathbf{x}|^2\right\} \\ &\to 0,\end{aligned}$$

as $|\mathbf{x}| \to \infty$, as required.

Now define a sequence of points recursively in the following way. Let $|\mathbf{x}_0| > S$ be such that $\pi(\mathbf{x}_0) > 0$, and let

$$\mathbf{x}_n = \arg\sup\{\pi(\mathbf{y}); \mathbf{y} \in B(\mathbf{x}_{n-1})\}.$$

It is easy to check that $\{\mathbf{x}_n\} \to \infty$ so that

$$\lim_{|\mathbf{x}|\to\infty} \sup_{\mathbf{y}\in B(\mathbf{x}_n)} \frac{q(\mathbf{y},\mathbf{x}_n)}{q(\mathbf{x}_n,\mathbf{y})} = 0.$$

Now

$$1 - r(\mathbf{x}_n) \le \frac{\varepsilon}{2} + \int_{B(\mathbf{x}_n)} \left(1 \wedge \frac{q(\mathbf{y},\mathbf{x}_n)}{q(\mathbf{x}_n,\mathbf{y})} \frac{\pi(\mathbf{y})}{\pi(\mathbf{x}_n)}\right) q(\mathbf{x}_n,\mathbf{y})\mathrm{d}\mathbf{y},$$

so choose N large enough so that

$$\sup_{y\in B(\mathbf{x})} \frac{q(\mathbf{y},\mathbf{x}_n)}{q(x_n,\mathbf{y})} < \frac{\varepsilon}{4},$$

for $n \ge N$. Then we have

$$\begin{aligned}\tfrac{1}{2}\varepsilon &\le \int_{B(\mathbf{x}_n)} \left(1 \wedge \frac{\varepsilon}{4}\frac{\pi(\mathbf{y})}{\pi(\mathbf{x}_n)}\right) q(\mathbf{x}_n,\mathbf{y})\mathrm{d}\mathbf{y} \\ &\le \int_{B(\mathbf{x}_n)} \left(1 \wedge \frac{\varepsilon}{4}\frac{\pi(\mathbf{x}_{n+1})}{\pi(\mathbf{x}_n)}\right) q(\mathbf{x}_n,\mathbf{y})\mathrm{d}\mathbf{y} \\ &= 1 \wedge \frac{\varepsilon}{4}\frac{\pi(\mathbf{x}_{n+1})}{\pi(\mathbf{x}_n)} \\ &\le \frac{\varepsilon\pi(\mathbf{x}_{n+1})}{4\pi(\mathbf{x}_n)},\end{aligned}$$

so that $\pi(\mathbf{x}_{n+1}) \ge 2\pi(\mathbf{x}_n)$, demonstrating that π is unbounded: and thus we have our contradiction. □

This result covers light-tailed densities. In the other direction, if π has heavy tails, MALA looks too much like a random walk on $\mathbb{R}^k$ to be geometrically ergodic. Specifically, we will be able to use the following theorem.

Theorem 4.3. *If* $\nabla \log \pi(\mathbf{x}) \to 0$, *then MALA is not geometrically ergodic.*

Proof. This follows by a similar argument to that used in the proof of Theorem 2.4. We only sketch the ideas here. Suppose MALA is geometrically ergodic. Then, for some bounded set of positive measure under π, C say, there exists $\kappa > 1$ such that for all $\mathbf{x} \in \mathbb{R}^k$,

$$\mathrm{E}_{\mathbf{x}}[\kappa^{\tau_C}] < \infty.$$

However if $\nabla \log \pi(\mathbf{x}) \to 0$, then it is easy to check that $r(\mathbf{x}) \to 0$ as $|\mathbf{x}| \to \infty$, so that the process behaves like a random walk with normally distributed increments. More specifically, we can find N large enough such that $\mathrm{e}^{\beta|\mathbf{M}|}$ is a submartingale for $|\mathbf{M}| \geq N$, and such that β is small enough so that, when τ_N denotes $\inf\{n; |\mathbf{M_n}| \leq N\}$, the collection of random variables $\{\mathrm{e}^{\beta X_{n \wedge \tau_N}}, n \geq 1\}$ are uniformly integrable, so that optional stopping applies for a contradiction. □

4.3. EXPONENTIAL MODELS

We conclude by applying the results above to the exponential models that we have analysed for our previous algorithms.

4.3.1. The one-dimensional class $\mathscr{E}(\beta, \gamma)$

In order to use Theorem 4.1, we need to consider the orientation of $A(\mathbf{x})$, and whether it 'converges inwards': this depends on the convexity properties of $\pi(\mathbf{x})$, as shown in (40). In order to give some intuition about convergence inwards, we will indicate how this behaviour occurs in one dimension, even for those chains for which Theorem 4.1 fails.

Let $\pi \in \mathscr{E}(\beta, \gamma)$ and recall that π is bounded over compacta when assessing $A(x)$ for large $|x|$. Note also that because the ULA candidate is a normal distribution with fixed variance, to check if (35) will be satisfied we need to evaluate whether the ULA step is centred near values in $A(\mathbf{x})\Delta I(\mathbf{x})$ or not.

We then have the following description:

(a) For $0 < \beta < 1$, we have $\lim_{n\to\infty} |\nabla \log \pi(\mathbf{x})| = 0$ and we get non-geometric convergence by Theorem 4.3, as we did with ULA. In this case, using (40), we have that the set $\{|y| > |x|\}$ approximates $A(x)$ for $|x|$ large: since the candidate density $q(x, y)$ is concentrated near x for $|x|$ large, we see that the set $A(x)$ does not converge inwards in q.

(b) At $\beta = 1$, for $x > 0$, we see that $q(x, y)$ is concentrated around $x - h\gamma/2$ for x positive. Using (40), we have that, for large $|x|$, we accept jumps in $A(x) = \{y > -x\}$, so that in this situation $A(\mathbf{x})\Delta I(\mathbf{x}) = \{y > x\}$ and so (35) fails since the integral is constant and positive for all x; thus the model does not converge inward in q, and we cannot use Theorem 4.1.

However, since we accept essentially all proposals in this case, the MALA chain from positive x is in effect just a random walk with negative drift on the positive half-line and with (more than) exponentially decreasing right tails; and this can be shown to be geometrically ergodic using the argument in Meyn and Tweedie (1993a, Section 16.1.3).

(c) For $1 < \beta < 2$ and for $x > 0$, the candidate density $q(x,y)$ is concentrated near $x - [h\gamma\beta/2]x^{\beta-1}$, and this is a value between 0 and x for x large. Since $\nabla \log \pi(x)$ is convex for x positive, and concave for x negative, $A(\cdot)$ converges inwards as $|x| \to \infty$. From Theorem 4.1 we have geometric ergodicity for this case.

(d) The Gaussian case $\beta = 2$ is again a threshold case, as it was for ULA. From (40) we have that, for $|x|$ large, $A(x) = \{|y| \leq |x|\}$ so that $A(x)$ always converges inwards in q. Now if $h\gamma < 2$ Theorem 4.1 shows that we have geometric ergodicity. If $h\gamma > 2$, however, (36) is violated; but we can now use Theorem 4.2, since (41) holds in exactly this case, to see that the chain is not geometrically ergodic.

(e) Finally, in the light-tailed case, $\beta > 2$, the term involving h dominates (40) and $A(\cdot)$ converges inwards as $|x| \to \infty$. But again (36) is violated, and indeed in this case $\liminf_{n\to\infty} |\nabla \log \pi(\mathbf{x})|/|\mathbf{x}| = \infty$, so that the Markov chain is not geometrically ergodic by Theorem 4.2.

4.3.2. *The multidimensional class $\mathscr{P}_m$*

Next consider the multidimensional case with $\pi(\mathbf{x}) \in \mathscr{P}_m$, $m > 2$. Then

$$\liminf_{n\to\infty} \frac{|\nabla \log \pi(\mathbf{x})|}{|\mathbf{x}|} = \infty,$$

so that MALA is not geometrically ergodic, no matter how small h is chosen to be. Note that in contrast, the random walk Metropolis algorithm is always geometric for $\pi \in \mathscr{P}_m$ (see Roberts and Tweedie 1996).

4.3.3. *A slowly converging Bayesian algorithm*

The examples we have used so far are all simple, although they do indicate the range of good and bad behaviours we might expect. We conclude with an example of a density π that occurs naturally as a Bayesian posterior density and where the arguments above show lack of geometric convergence of the MALA chain $\mathbf{M}_n$.

Let $Y_0, Y_1, \ldots$ be conditionally i.i.d. $N(\mu, \tau^{-1})$ variables and let τ, μ have the conditional conjugate priors

$$\mu \sim N(\mu_0, \tau_0^{-1}), \qquad \tau \sim \Gamma(\alpha_0, \beta_0).$$

Then the posterior density π is given by

$$\pi(\mu,\tau) \propto \exp\left\{ -\frac{\tau}{2}\sum_{i=1}^{n}(y_i - \mu)^2 - \frac{\tau_0}{2}(\mu - \mu_0)^2 \right.$$
$$\left. + \left(\beta_0 + \frac{n}{2} - 1\right)\log\tau - \alpha_o\tau \right\}, \qquad \mu \in \mathbb{R}, \tau \in \mathbb{R}^+,$$

so that

$$\nabla \log \pi(\mu, \tau) = \left\{ -\tau \sum_{i=1}^{n} (\mu - y_i) - \tau_0(\mu - \mu_0) \right\} \mathbf{e}_\mu$$
$$+ \left\{ \left(\beta_0 + \frac{n}{2} - 1\right) / \tau - \alpha_0 - \tfrac{1}{2} \sum_{i=1}^{n} (y_i - \mu)^2 \right\} \mathbf{e}_\tau.$$

This is typical of the structure in hierarchical models.

By the same arguments used above, the MALA chain would perform badly here. This example is not quite covered explicitly above but it is easy to show that geometric convergence will not hold: to see this, merely choose τ and μ large enough that the algorithm proposes steps further into the tail, by setting τh and $h \sum (y_i - \mu)^2$ large, say. The detailed justification that we can make the rejection probabilities $r(\cdot, \cdot)$ as close to unity as we wish in this case then follows as in the proof of Theorem 4.2, and we leave the details to the reader.

5. Concluding remarks

Although using the Langevin candidate ULA seems like a good strategy for generating a ‘targeted’ initial dynamic, and although Metropolizing to get the MALA chain guarantees convergence, we have shown that in many cases this gives an algorithm which is not guaranteed to converge geometrically fast even in situations when the simpler random walk candidate is known to do so. This is not very appealing, but we do note that it is a product of bad behaviour in the far reaches of the space.

We therefore conclude by recalling the MALTA algorithm with drift defined by (12). Like MALA and the random walk Metropolis algorithm, MALTA is μ^{Leb}-irreducible and Feller, and therefore converges in total variation to the target density. But since the MALTA drift is truncated, the problems with MALA are not encountered, and MALTA enjoys more stable geometrically ergodic properties, central limit theorems and the like, just as the random walk-based algorithms do, for target densities that are not heavy-tailed.

Moreover, since the algorithm behaves like MALA except in extreme situations, it will inherit most of the other desirable rapid convergence properties of MALA, and, for example, the complexity result of Roberts and Rosenthal (1995a) for high-dimensional MALA algorithms will also hold for MALTA.

For practical purposes, this truncated algorithm thus seems the most desirable version of this class of algorithms.

Acknowledgements

Work on this paper was supported in part by NSF grants DMS-9205687 and DMS-9504561, and by Nuffield Foundation grant SCI/180/91/84/G. We are grateful to Julian

Besag for raising with us the question of convergence of the Langevin diffusion and the related chains, and to Osnat Stramer for conversations on the general ergodic behaviour of diffusions. We are grateful to the referees for useful comments.

References

Besag, J.E. (1994) Comments on ‘Representations of knowledge in complex systems’ by U. Grenander and M.I. Miller. *J. Roy. Statist. Soc. Ser. B*, **56**, 591–592.

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

Besag, J.E., Green, P.J., Higdon, D. and Mengersen, K.L. (1995) Bayesian computation and stochastic systems (with discussion). *Statist. Sci.*, **10**, 3–66.

Bhattacharya, R.N. (1978) Criteria for recurrence and existence of invariant measures for multidimensional diffusions. *Ann. Probab.*, **6**, 541–553.

Davis, M.H.A. (1993) *Markov Models and Optimizations.* London: Chapman & Hall.

Down, D., Meyn, S.P. and Tweedie, R.L. (1995) Geometric and uniform ergodicity of Markov processes. *Ann. Probab.*, **23**, 1671–1691.

Grenander, U. and Miller, M.I. (1994) Representations of knowledge in complex systems (with discussion). *J. Roy. Statist. Soc. Ser. B*, **56**, 549–603.

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

Ikeda, N. and Watanabe, S. (1989) *Stochastic Differential Equations and Diffusion Processes.* Amsterdam: Elsevier.

Lamperti, J. (1963) Criteria for stochastic processes II: Passage time moments. *J. Math. Anal. Appl.*, **7**, 127–145.

Lerche, H.R. (1986) *Boundary Crossings of Brownian Motion.* Berlin: Springer-Verlag.

Mengersen, K.L. and Tweedie, R.L. (1996) Rates of convergence of the Hastings and Metropolis algorithms. *Ann. Statist.*, **24**, 101–121.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller , A. and Teller, E. (1953) Equations of state calculations by fast computing machines. *J. Chem. Phys.*, **21**, 1087–1091.

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

Meyn, S.P. and Tweedie, R.L. (1993b) Stability of Markovian processes II: Continuous time processes and sampled chains. *Adv. Appl. Probab.*, **25**, 487–517.

Neal, R. (1994) An improved acceptance procedure for the hybrid Monte Carlo algorithm. *J. Comp. Phys.*, **11**, 194–203.

Parisi, G. (1981) Correlation functions and computer simulations. *Nuclear Phys. B*, **180**, 378–384.

Roberts, G.O. and Rosenthal, J.S. (1995a) Optimal scaling of discrete approximations to Langevin diffusions. Submitted for publication.

Roberts, G.O. and Rosenthal, J.S. (1995b) Shift-coupling and convergence rates of ergodic averages. Submitted for publication.

Roberts, G.O. and Smith, A.F.M. (1994) Simple conditions for the convergence of the Gibbs sampler and Hastings–Metropolis algorithms. *Stochastic Process. Appl.*, **49**, 207–216.

Roberts, G.O. and Tweedie, R.L. (1996) Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika*, **83**, 96–110.

Roberts, G.O., Gelman, A. and Gilks, W.R. (1994) Weak convergence and optimal scaling of Hastings–Metropolis algorithms. *Ann. Appl. Prob.* To appear.

Sheu, S.-J. (1992) Some estimates of the transition density of a non-degenerate diffusion Markov process. *Ann. Probab.*, **19**, 538–561.

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

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

Tweedie, R.L. (1994) Topological conditions enabling use of Harris methods in discrete and continuous time. *Acta Appl. Math.*, **34**, 175–188.

Received October 1995 and revised April 1996