Scaling Limits for the Transient Phase of Local Metropolis-Hastings Algorithms

by

Ole F. Christensen* and Gareth O. Roberts** and Jeffrey S. Rosenthal***

(January 2003; revised September 2004.)

Abstract. This paper considers high-dimensional Metropolis and Langevin algorithms in their initial transient phase. In stationarity, these algorithms are well-understood and it is now well-known how to scale their proposal distribution variances. For the random walk Metropolis algorithm, convergence during the transient phase is extremely regular - to the extent that the algorithm's sample path actually resembles a deterministic trajectory. In contrast, the Langevin algorithm with variance scaled to be optimal for stationarity, performs rather erratically. We give weak convergence results which explain both of these types of behaviour, and give practical guidance on implementation based on our theory.

1. Introduction.

Markov chain Monte Carlo (MCMC) algorithms are a very popular method for sampling from complicated probability distributions $\pi(\cdot)$ (see e.g. Gilks, Richardson and Spiegelhalter, 1996). One very common MCMC algorithm is the Metropolis-Hastings algorithm (Metropolis, Rosenbluth, Rosenbluth, Teller and Teller, 1953; Hastings, 1970). This algorithm requires that we choose a *proposal distribution*. A fundamental question is what

* BiRC - Bioinformatics Research Center, Aarhus University, Hoegh-Guldbergs Gade 10, 8000 Aarhus C, Denmark. Email: `olefc@birc.au.dk`.
** Department of Mathematics and Statistics, Fylde College, Lancaster University, Lancaster, LA1 4YF, England. Email: `g.o.robert@lancaster.ac.uk`.

*** Department of Statistics, University of Toronto, Toronto, Ontario, Canada M5S 3G3. Email: `jeff@math.toronto.edu`. Supported in part by NSERC of Canada.

scaling should be used for the proposal. This question has recently received considerable attention (Gelman, Roberts and Gilks, 1996; Roberts, Gelman and Gilks, 1997; Roberts and Rosenthal, 1998; Roberts, 1998; Breyer and Roberts, 2000; Roberts and Rosenthal, 2001; Roberts and Yuen, 2004; Bédard, 2004).

The result of Roberts *et al.* (1997) says that, for random-walk Metropolis algorithms, to achieve optimal mixing speed of the algorithm, the proposal variance should scale with dimension d like $O(d^{-1})$, and the optimal acceptance rate should be 0.234. By contrast, the result of Roberts and Rosenthal (1998) says that, for Langevin Metropolis-Hastings algorithms, the proposal variance should scale with dimension d like $O(d^{-1/3})$, and the optimal acceptance rate should be 0.574. These results provide very useful practical guidance.

However, these results do have limitations. Firstly, these results are limiting results as the dimension of the problem goes to infinity, and this restricts the class of target distribution for which formal scaling limits can be demonstrated. For instance, Roberts *et al.* (1997) and Roberts and Rosenthal (1998) only prove results rigorously when $\pi(\cdot)$ has components which are i.i.d. (at least asymptotically), though subsequent theoretical and empirical work shows that these results apply somewhat more generally (see Roberts and Rosenthal, 2001, for a survey). Secondly, and sometimes overlooked, the results assume that the chain is started in stationarity, i.e. they consider mixing properties only in the stationary phase of the chain.

As a further motivation for the study in this paper, it is empirically well-known that while MCMC sample path trajectories in stationarity typically resemble diffusion processes, in their transient phases, algorithms often look almost deterministic, in that the characteristic fluctuations of MCMC trajectories are absent, or dominated by a systematic drift term. Figure 1 is an example of this type of behaviour from a Bayesian posterior distribution for inference for partially observed diffusions (see Beskos *et. al.*, 2004, for details). In this example, parameters are being updated using simple Metropolis jumping rules. Further examples of this deterministic initial transient phase are considered later in the paper.

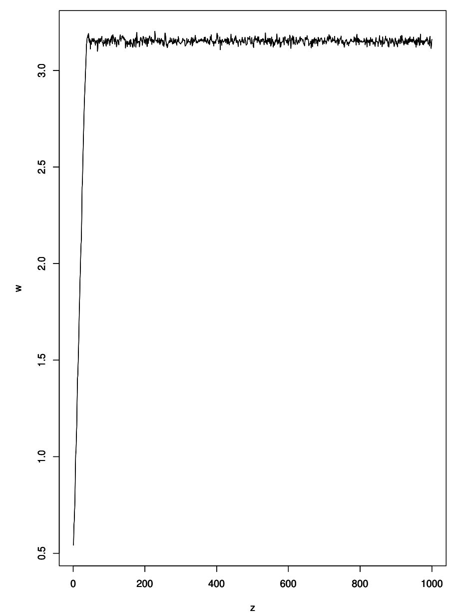

Figure 1. A sample path of a Metropolis algorithm for a Bayesian posterior distribution for partially observed diffusions, as in Beskos *et. al*, (2004).

So while the diffusion approximation is effective for studying algorithms in stationarity, it seems likely that a different theory is needed to describe the transient phase. In this paper, we consider what happens in the transient (pre-stationary) phase of the chain. Here our focus will be on Metropolis and Langevin algorithms. Our investigation will involve theory, numerical investigation, and Bayesian applications. Most of the theory just covers high-dimensional Gaussian distributions, where clear cut explicit results can be given to explain *almost deterministic transient* behaviour. However, we shall see in examples that the conclusions we draw from the theoretical study extend well beyond the Gaussian case, providing useful practical guidance for MCMC users in complex high-dimensional problems.

For the theory results, we consider chains which are started far out in the tails of $\pi(\cdot)$, and study their approach to the "center" of $\pi(\cdot)$. Asymptotically (i.e., as $d \to \infty$), this happens *deterministically* as long as the proposal variance is scaled appropriately. In particular, for Langevin algorithms, the proposal standard deviation must only scale as $O(d^{-1/2})$ or smaller. If the proposal variance recedes to zero more slowly than $O(d^{-1/2})$, then the convergence time becomes *exponentially* large as a function of dimension, exhibiting erratic behaviour and rejecting a large proportion of proposed moves in the tail region.

To illustrate that these scaling problems for the Langevin algorithm are relevant in practice, in Section 6 we will consider an example of a high-dimensional target density related to Bayesian inference for a spatial generalised linear mixed model (GLMM). Our example will involve inference for a latent log-Gaussian Cox point-process, given point process data. For this example, and as predicted by the Gaussian theory, we demonstrate that for two natural starting values (both being near the mode of the distribution) the algorithm is stuck at the starting value, whereas for a starting value chosen approximately from the stationary distribution, the algorithm mixes very well.

Metropolis methods are of course very commonly used, while Langevin methods are less well-established in the statistical literature. However, Langevin algorithms offer huge computational advantages in many problems (particularly for spatial models) and deserve to be more widely adopted. For example, for the model considered in Section 6, Metropolis methods are now established to be prohibitively slow, while Langevin methods offer a feasible and practical alternative. In fact, theory predicts (see Roberts and Rosenthal, 1998) that Langevin algorithms will almost always massively outperform Metropolis alternatives in the stationary phase, whenever they are practically feasible to run. For further examples of the use of Langevin methods for high-dimensional Bayesian problems, see Grenander and Miller (1994), Neal (1996) and Møller, Syversveen and Waagepetersen (1998).

2. Definitions.

Given a d-dimensional probability density function of interest, π (with respect to the Lebesgue measure), and a proposal kernel $Q(\mathbf{x}, \cdot)$, the Metropolis-Hastings algorithm proceeds as follows. Suppose that at the t'th iteration, the current state of the algorithm is given by $\mathbf{X}_t$. Then the algorithm proposes a new value $\mathbf{Y}_{t+1} \sim Q(\mathbf{X}_t, \cdot)$, which has proposal density with respect to the Lebesgue measure given by $q(\mathbf{X}_t, \cdot)$. The proposal $\mathbf{Y}_{t+1}$ is accepted as the new value (and we therefore set $\mathbf{X}_{t+1} = \mathbf{Y}_{t+1}$) with probability

$\alpha(\mathbf{X}_t, \mathbf{Y}_{t+1})$ where

$$\alpha(\mathbf{x}, \mathbf{y}) = \min\left\{1, \frac{\pi(\mathbf{y})q(\mathbf{y}, \mathbf{x})}{\pi(\mathbf{x})q(\mathbf{x}, \mathbf{y})}\right\} \ ,$$

and otherwise rejected, in which case we set $\mathbf{X}_{t+1} = \mathbf{X}_t$.

Throughout this paper we will consider two algorithms. The symmetric random-walk Metropolis algorithm takes $q(\mathbf{x}, \mathbf{y})$ to be a spherically symmetric function of $\|\mathbf{y} - \mathbf{x}\|$, often denoted by $q(\|\mathbf{y} - \mathbf{x}\|)$. In this case, $\alpha(\mathbf{x}, \mathbf{y})$ simplifies to

$$\alpha(\mathbf{x}, \mathbf{y}) = \min\left\{1, \frac{\pi(\mathbf{y})}{\pi(\mathbf{x})}\right\} \ .$$

Therefore left to its own devices, Q would carry out a random walk with increment density $q(\cdot)$. Most of what we describe below requires only that $q(\cdot)$ be a density of a square integrable random variable. However for simplicity we shall assume that q is multivariate Gaussian, i.e. $Q(\mathbf{x}, \cdot) \sim \mathrm{MVN}(\mathbf{x}, hI_d)$ where I_d is the d-dimensional identity matrix and h is a proposal variance parameter.

The second algorithm we consider is the Langevin algorithm (see for example Roberts and Tweedie, 1996; Roberts and Rosenthal, 1998), which is motivated by a discrete approximation to a Langevin diffusion, and takes $Q(\mathbf{x}, \cdot) \sim \mathrm{MVN}(\mathbf{x} + \nabla \log \pi(\mathbf{x})h/2, hI_d)$.

3. A Concrete Example.

To further motivate the results we shall describe later, we consider the case where $\pi(\cdot)$ is a d-dimensional standard normal distribution, so that $\pi(\mathbf{x}) \propto \exp(-\frac{1}{2}\|\mathbf{x}\|^2)$.

We consider a random-walk Metropolis algorithm with Gaussian proposal distribution where the proposal variance h is scaled to be proportional to d^{-1} in each dimension. Thus $\mathbf{Y}_{t+1} \sim \mathrm{MVN}(\mathbf{X}_t,\ (\ell^2/d)\, I_d)$. We also consider the Langevin algorithm with proposal variance h scaled to be proportional to $d^{-1/3}$. Thus in this case $\mathbf{Y}_{t+1} \sim \mathrm{MVN}(\mathbf{X}_t + \nabla \log \pi(\mathbf{X}_t)\ell^2/(2d^{1/3}),\ (\ell^2/d^{1/3})\, I_d)$. (Note that there is no relationship between ℓ as used in the Metropolis case, and ℓ as used in the Langevin case. In both cases, ℓ is just being used as a generic scale constant.) These variance scalings are optimally efficient in the stationary phase (see Roberts *et al.*, 1997, and Roberts and Rosenthal, 1998), and we also choose the constant in the proposal variance, ℓ, as the stationary optimal scaling.

Figure 2 below shows output from simulating the target density $\pi(\mathbf{x}) \propto \exp(-\|\mathbf{x}\|^2/2)$ where $\mathbf{x}$ has dimension $d = 1000$. Plots on the left correspond to two independent random walk Metropolis runs while plots on the right correspond to Langevin runs. For the Metropolis traces, we have superimposed a deterministic function which is formally defined in (3) below, and which we will discuss in more detail later.

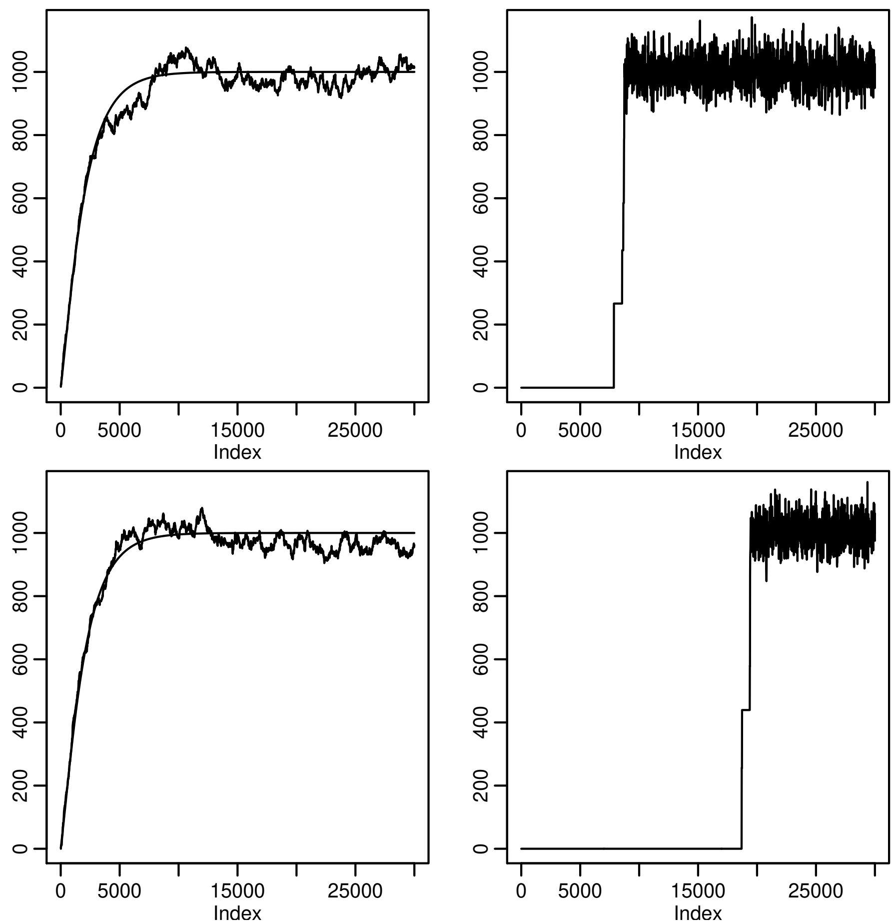

Figure 2. Trace plots of $\|\mathbf{x}\|^2$ for simulating a 1000-dimensional normal distribution, when starting at the origin. Left: Random walk Metropolis, together with solid line giving the function f defined in (3) below. Right: Langevin. Two independent simulations are made for each algorithm.

For the random walk Metropolis method, the initial convergence appears to be almost deterministic, governed by the function f. However once the algorithm reaches the stationary region, its trajectory becomes more obviously stochastic, displaying behaviour characteristic of a diffusion process. Note that convergence to the stationary region is fairly quick, but subsequent mixing is displaying high serial correlation.

In contrast, the Langevin method takes a large number of iterations to move at all, and then proceeds in a sequence of unpredictable irregular jumps to move towards the stationary region. The algorithm displays 'sticky patches' when in tail regions, where large numbers of successive iterations are rejected. However once the algorithm has succeeded in finding the stationary region, its mixing is very rapid indeed.

The theory in Roberts and Rosenthal (1998) predicts accurately the comparative performance of these two algorithms in stationarity: mixing time for the Langevin algorithm ought to be $O(d^{1/3})$ comparing favourably with $O(d)$ for the random walk Metropolis algorithm. However from this output, it is clear that the transient phase of the algorithms needs further investigation, since here, the random walk Metropolis method appears to be doing better.

4. Scaling limit for random-walk Metropolis algorithms.

We now consider the random-walk Metropolis algorithm on the above Gaussian example analytically. The Gaussian context here allows a detailed theoretical study to be performed. It appears to be very difficult to extend these theoretical arguments substantially beyond the context we shall present here, although we emphasis that in practice, the phenomenon we shall describe here is empirically observed far more generally.

We let $W_t^d = (1/d)\|\mathbf{X}_{\lfloor td \rfloor}\|^2$, so that W^d is a process which keeps track of the norm-squared of $\mathbf{X}$, shrunk by a factor d, with time speeded up by a factor d. The functional W^d is the natural functional to consider in this problem for two reasons. Firstly, W_t^d is a Markov process because of the spherical symmetry of the problem, and this allows relatively straightforward calculations to establish its limiting behaviour for d going to infinity. Secondly, since we shall be interested in particular in the starting value $\mathbf{X}_0 = \underline{0}$, the norm-squared is in some sense the only interesting functional: all angular components mix immediately in 1 iteration. Thus, the study of W^d is sufficient to characterise the convergence of the entire distribution of $\mathbf{X}$.

We have the following calculation. (The proof of this result, and most of others in this paper, have been put in the appendix, so as not to interrupt the flow of the paper.)

Lemma 1. *For $t = 0, 1, 2, \ldots$,*

$$\lim_{d\to\infty} \mathbf{E}\left[W^d_{(t+1)/d} - W^d_{t/d} \,|\, W^d_{t/d} = w\right] d \;=\; a_\ell(w)\,, \tag{1}$$

where

$$a_\ell(w) \;=\; \ell^2\Phi(N^*) + \exp((\ell^2/2)(w-1))(1-2w)\ell^2\Phi(-N^* - \ell w^{1/2})\,, \tag{2}$$

with $N^* = -\ell w^{-1/2}/2$, $\phi(s) = (2\pi)^{-1/2}\exp(-s^2/2)$, and $\Phi(s) = \int_{-\infty}^{s} \phi(u)du$. Moreover the convergence in (1) is uniform for $w \in [0, K]$ for all $K > 0$.

We shall require one more technical result which ensures that fluctuations of W^d are not too 'severe'. This will essentially allow us to show that the sample paths resemble a deterministic trajectory in an appropriate sense.

Lemma 2. *For $t = 0, 1, 2, \ldots$, for all $K \geq 0$,*

$$\limsup_{d\to\infty} \sup_{0\leq w\leq K} \mathbf{E}\left[\left(W^d_{(t+1)/d} - W^d_{t/d}\right)^2 \,|\, W^d_{t/d} = w\right] d^2 \;<\; \infty\,.$$

We are now in a position to state formally the main result of this section. From Lemmas 1 and 2, we obtain the following.

Theorem 3. *When $W^d_0 = w_0$, then as $d \to \infty$, we have $W^d \Rightarrow f$, where $\Rightarrow$ is weak convergence, and where f is a deterministic function satisfying $f(0) = w_0$ and*

$$f'(t) \;=\; a_\ell(f(t))\,, \tag{3}$$

with $a_\ell(\cdot)$ as in (2).

Theorem 3 therefore explains the deterministic sample path behaviour of the random-walk Metropolis algorithm during its transient phase, as illustrated in Figure 2.

Figure 3 shows the function $a_\ell(w)$ for different values of ℓ. Convergence of the algorithm is quicker when the modulus of a_ℓ is large. We see that there exists no ℓ which is uniformly maximising the speed of convergence.

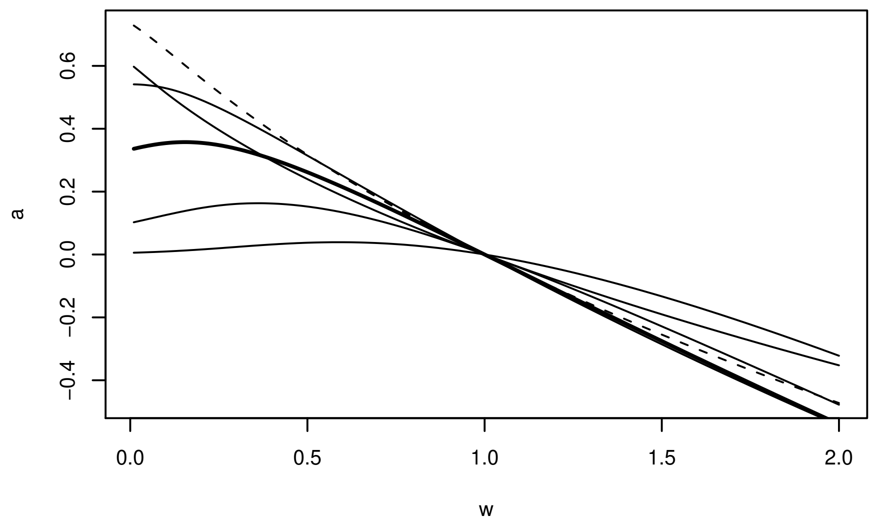

Figure 3. A collection of the $a_\ell(\cdot)$ functions defined in (2) for various different values of ℓ. Convergence is quicker when a_ℓ is as large as possible in modulus and positive for $w < 1$, negative for $w > 1$. The thick solid curve for $\ell = 2.38$, represents the scaling corresponding to optimal mixing for a chain started in stationarity. The dashed curve for $\ell = \sqrt{2}$, represents the scaling produced by optimising $a_\ell(0)$. Both thick solid and dashed curves perform close to optimally for all values of w. The four remaining curves are for $\ell = 1, \ldots, 4$.

Started in stationarity, the mixing time of optimally scaled random walk Metropolis algorithms, on reasonably behaved target distributions, is known to be $O(d)$ as $d \to \infty$. We argue informally that the convergence time of the random walk Metropolis algorithm, when started from the transient phase, is still $O(d)$. Indeed, from Theorem 3, the time taken for W to reach $1-1/d$ is $O(\log d)$. However, once W reaches $1-1/d$, then it behaves as in its stationary phase, and thus converges in distribution in $O(d)$ further iterations. Thus, W, and hence also $||\mathbf{X}||^2$, converges to stationarity in time $O(d + \log d) = O(d)$.

5. Scaling limits for Langevin algorithms.

We now move on to study in more detail the erratic behaviour of the Langevin algorithm in its transient phase, as illustrated in Figure 2. For the example in Section 3, we first provide theoretical justification for the problematic behaviour observed when the proposal variance is $h = \ell^2 d^{-1/3}$. Secondly, we motivate a different scaling limit for the algorithm.

For the Langevin algorithm with scaling $O(1/d^{1/3})$, we obtain a result analogous to, but qualitatively different from, Lemma 1.

Lemma 4. *Consider again the d-dimensional standard normal target density, and its exploration using the Langevin algorithm with scaling $h = \ell^2 d^{-1/3}$. Let $W_t^d = (1/d)\|\mathbf{X}_{\lfloor td^{1/3}\rfloor}\|^2$. Then as $d \to \infty$, for $0 \le w \le 1$, and $t = 0, 1, 2, \ldots$,*

$$\mathbf{E}\left[W^d_{(t+1)/d^{1/3}} - W^d_{t/d^{1/3}} \mid W^d_{t/d^{1/3}} = w\right] d^{1/3} \approx \ell^2 (1-w)\,\mathrm{acc}_d(w),$$

where $\mathrm{acc}_d(w) = \min\{1, \exp(-d^{1/3}\ell^4(1-w)/8)\}$ is the acceptance probability of moves from w. In particular, for $w < 1$, as $d \to \infty$, $\mathrm{acc}_d(w)$ decreases as $O(e^{-Cd^{1/3}})$ for some $C > 0$.

From Lemma 4, we see the reason for the problems reported in Section 3. The acceptance probability of moves from the origin are receding exponentially in $d^{1/3}$, leading to severe mixing problems. In particular, this explains why the Langevin algorithm remains at 0 for so long in the runs of Figure 2.

We also observe that for $w \ge 1$, the algorithm behaves well (when using the scaling $O(d^{1/3})$). Therefore, it is only starting values too close to the mode (origin) which lead to severe convergence problems.

On the other hand, we may avoid the problem of acceptance probabilities decreasing to 0, by choosing the variance scaling to be $O(d^{-1/2})$ instead of $O(d^{-1/3})$:

Lemma 5. *Consider the Langevin algorithm with scaling $h = \ell^2 d^{-1/2}$. Let $W_t^d = (1/d)\|\mathbf{X}_{\lfloor td^{1/2}\rfloor}\|^2$. Then*

$$\lim_{d\to\infty} \left(\mathbf{E}[W^d_{(t+1)/d^{1/2}} - W^d_{t/d^{1/2}} \mid W^d_{t/d^{1/2}} = w]\right) d^{1/2} = b_\ell(w),$$

where

$$b_\ell(w) = \ell^2(1-w)\min\{1, \exp(-\ell^4(1-w)/8)\}. \qquad (4)$$

Also for all $K > 0$,

$$\limsup_{d\to\infty} \sup_{0\le w\le K} \mathbf{E}\left[(W^d_{(t+1)/d^{1/2}} - W^d_{t/d^{1/2}})^2 \,|\, W^d_{t/d^{1/2}} = w\right] d \;<\; \infty\,.$$

Using this proposition, it follows that for the Langevin algorithm using a scaling $O(d^{-1/2})$, we have a result similar to Theorem 3, again providing deterministic convergence from the transient phase to the stationary phase.

Theorem 6. *When $W_0^d = w_0$, then as $d \to \infty$, we have $W^d \Rightarrow f$, where $\Rightarrow$ is weak convergence, and where f is a deterministic function satisfying $f(0) = w_0$ and*

$$f'(t) \;=\; b_\ell(f(t))\,, \qquad (5)$$

with $b_\ell(\cdot)$ as in (4).

We therefore should see deterministic sample path behaviour of the Langevin algorithm during its transient phase, when using a variance scaling of $O(d^{-1/2})$.

To use Theorem 6 for simulations, we first need to decide how to choose the constant ℓ in the proposal variance $h = \ell^2/d^{1/2}$. Figure 4 shows the function $b_\ell(w)$ for different choices of ℓ. We see that, as in the case for $a_\ell(w)$, there exists no ℓ which is uniformly maximising the speed of convergence.

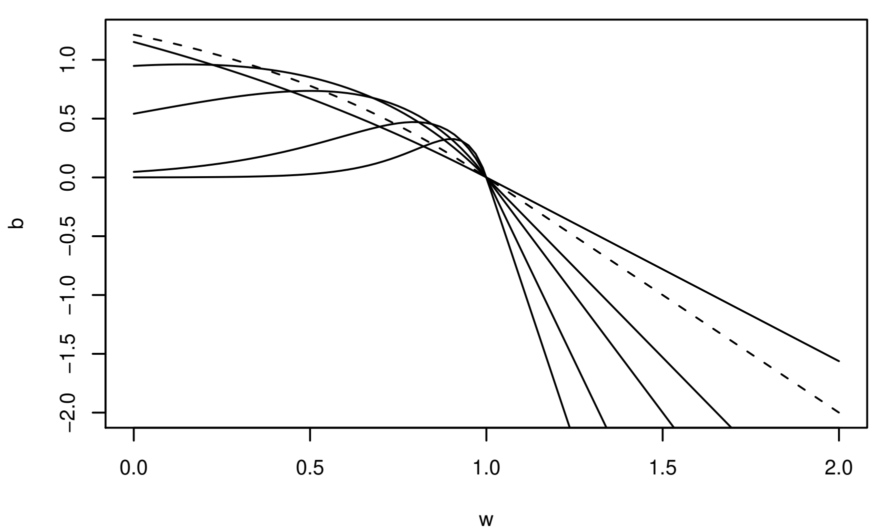

Figure 4. A collection of the $b_\ell(\cdot)$ functions defined in (4) for various different values of ℓ. Convergence is quicker when b_ℓ is as large as possible in modulus, and positive for $w < 1$, negative for $w > 1$. The dashed curve is for $\ell = \sqrt{2}$, and the five solid curves are for $\ell = 1.25, 1.75, 2, 2.5, 3$.

The value $\ell = \sqrt{2}$ maximises $b_\ell(0)$, and in Figure 5 we show a trace plot for the target density (for the example in Section 3) with this choice of ℓ.

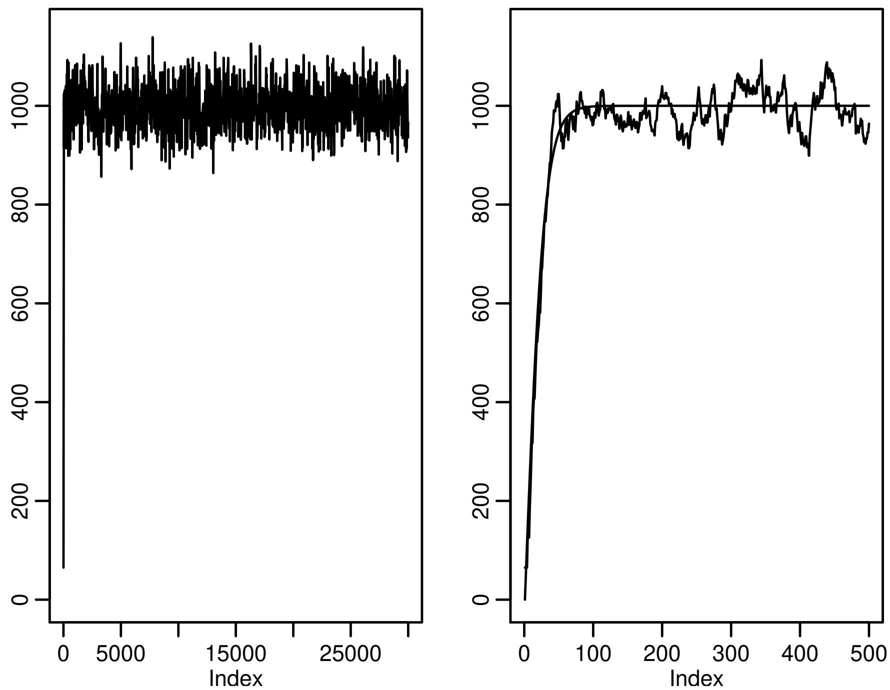

Figure 5. Trace plots of $||\mathbf{x}||^2$ for simulating a 1000 dimensional normal distribution starting at the origin using the Langevin algorithm with proposal variance $h = \ell^2/d^{1/2}$ with $\ell = \sqrt{2}$. Left: Trace plot. Right : Trace plot for the first 500 iterations, together with a solid line giving the solution to $f'(t) = b_\ell(f(t))$, where $b_\ell(\cdot)$ is as in (4).

The initial convergence from the transient phase now appears to be quick, and almost deterministic. Also, once the algorithm has reached the stationary region, its mixing is rapid. However, the mixing in the stationary region of the Langevin algorithm scaled as $O(d^{1/2})$ is still slower than that when scaled as $O(d^{1/3})$, since $O(d^{1/3})$ is the optimal choice in stationarity (Roberts and Rosenthal, 1998).

An obvious and important question is whether or not the result in Lemma 5 is specific to the normal target density. Clearly, for other target densities, the form of (4) is different, since in the derivation we have used extensively that $\frac{\partial}{\partial \mathbf{x}} \log \pi(\mathbf{x}) = -\mathbf{x}$ for the normal density. However, our conclusion about using the scaling $h = \ell^2 d^{-1/2}$ during the transient phase for the Langevin algorithm also appears to hold for other distributions. Whilst proving results as general as Theorem 6 is difficult, we can nevertheless prove the following weaker statement.

Theorem 7. *Suppose that $\pi(\mathbf{x}) = \exp(\sum_{i=1}^{d} g(x_i))$ where g is any four times differentiable function. Consider the Langevin algorithm for $\pi(\cdot)$, with variance scaling $h = \ell^2 d^{-1/2}$. Assume the algorithm's initial point $\mathbf{X}_0$ is at a mode of $\pi(\cdot)$. Let p_d be the probability that the first proposed move is accepted, i.e. that $\mathbf{X}_1 \neq \mathbf{X}_0$. Then $\lim_{d\to\infty} p_d$ exists and is strictly positive.*

Theorem 7 shows that for a general class of target distributions $\pi(\cdot)$, the Langevin algorithm with scaling $O(d^{-1/2})$, in contrast with $O(d^{-1/3})$, will not get "stuck" when starting from the mode.

6. Log-Gaussian Cox point-process example.

We now consider an example of a high-dimensional target density related to inference for a log-Gaussian Cox point-process. The example is from Møller, Syversveen and Waagepetersen (1998) and consists of locations of 126 Scots pine saplings in a natural forest in Finland. The locations are shown in the left plot in Figure 6.

The discretised version of the model used in the paper can be defined as follows. First the area of interest, $[0,1]^2$, is discretised into a 64×64 regular grid, where the random variables $\mathbf{X} = \{X_{i,j}\}$ are the number of points in grid-cells (i,j), $i,j = 1,\ldots,64$. The dimension of the problem is thus $d = 64^2 = 4096$. Note that due to the fine discretisation used, most of the grid cells contain no points, and only a few contain more than one point. Given an unobserved intensity process $\Lambda(\cdot) = \{\Lambda(i,j) \mid i,j = 1,\ldots,64\}$, the random variables $X_{i,j}$, $i,j = 1,\ldots,64$, are assumed to be conditionally independent and Poisson distributed with means $m\Lambda(i,j)$, $i,j = 1,\ldots,64$, where $m = 1/4096$ is the area of each grid-cell. The prior assumed for $\Lambda(\cdot)$ is

$$\Lambda(i,j) = \exp(Y_{i,j}),$$

where $\mathbf{Y} = (Y_{i,j}, i, j = 1, \ldots, 64)$ is multivariate Gaussian with mean $\mathrm{E}[\mathbf{Y}] = \mu\mathbf{1}$, and covariance matrix $\mathrm{Cov}(\mathbf{Y}) = \Sigma$, where

$$\Sigma_{(i,j),(i',j')} = \sigma^2 \exp(-((i-i')^2 + (j-j')^2)^{1/2}/(64\beta)).$$

In the paper they estimated parameter values $\beta = 1/33$, $\sigma^2 = 1.91$ and $\mu = \log(126) - \sigma^2/2$, which we will use here.

Møller *et al.* (1998) considered simulation of the intensity $\Lambda(\cdot)$ given the data $\mathbf{X} = \mathbf{x}$, or equivalently $\mathbf{Y}$ given $\mathbf{X} = \mathbf{x}$. Using Bayes' formula, the target density of interest is

$$f(\mathbf{y} \mid \mathbf{x}) \propto \prod_{i,j=1}^{64} \exp(x_{i,j} y_{i,j} - m \exp(y_{i,j})) \, \exp(-0.5(\mathbf{y} - \mu\mathbf{1})^T \Sigma^{-1} (\mathbf{y} - \mu\mathbf{1})) \, ,$$

where $\mathbf{y} = (y_{i,j}, i, j = 1, \ldots, 64)$. As in Papaspiliopoulos *et al.* (2003) and Christensen *et al.* (2003), we reparameterise $\mathbf{Y} = \mu\mathbf{1} + L\Gamma$, where L is obtained by Cholesky factorisation such that $\Omega = LL^T$ where $\Omega = (\Sigma^{-1} + \mathrm{diag}(\mathbf{y}))^{-1}$. The target density is now

$$f(\gamma \mid \mathbf{x}) \propto \prod_{i,j=1}^{64} \exp(x_{i,j} y_{i,j} - m \exp(y_{i,j})) \, \exp(-0.5 \gamma^T L^T \Sigma^{-1} L \gamma) \, ,$$

where the vector $\mathbf{y} = \mu\mathbf{1} + L\gamma$. The gradient of the log target density is $\nabla \log f(\gamma \mid \mathbf{x}) = -L^T \Sigma^{-1} L\gamma + L^T \{x_{i,j} - m \exp(y_{i,j})\}_{i,j}$.

As in Møller *et al.* (1998), we use a Langevin algorithm for MCMC simulation. The high dimensionality of the problem makes mixing of a Metropolis alternative prohibitively slow. (For this model, that was noticed by Christensen and Waagepetersen, 2002, and theoretical arguments explaining this can be found in Roberts and Rosenthal, 2001. We therefore omit any comparison with the random walk Metropolis algorithm for this problem.)

The right hand plot in Figure 6 shows the estimated intensity $\mathbf{E}[\Lambda(\cdot) \mid \mathbf{x}]$ based on 100000 iterations, subsampling every 10th observation for storage reasons and using the starting value II below; the plot is similar to Figure 12, upper left plot, in Møller *et al.* (1998), the only difference being that more grey-scale colours are used here.

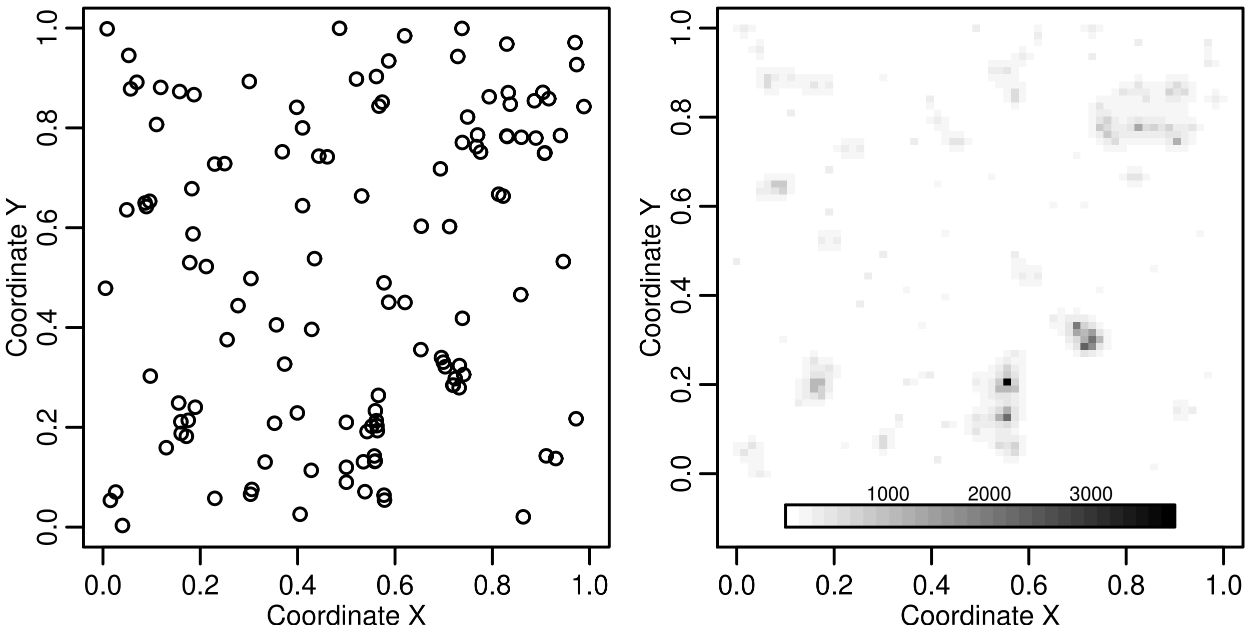

Figure 6. Scots pine saplings. Left: locations of trees. Right: the estimated intensity $\mathbf{E}[\Lambda(\cdot) \mid \mathbf{x}]$.

Note that using the Cholesky factorisation as above is computationally slow, and is only used here for simplicity of presentation. Møller *et al.* (1998) use a different reparameterisation and a circulant embedding technique, where they extend the grid to a torus and use the two-dimensional fast Fourier transform (FFT) to reduce the computational burden. We refer to their paper for further details about this.

Now we compare the performance of the Langevin algorithm above for three different starting values. The starting values expressed in terms of $\mathbf{Y}$ (which have to be transformed to starting values for Γ) are

I : $Y_{i,j} = \mu$ for $i, j = 1, \ldots, 64$.

II : a random starting value for Γ, simulated from $\Gamma \sim \mathrm{MVN}(\mathbf{0}, I_{4096})$.

III : a starting value near the posterior mode. Let $Y_{i,j}$ solve the equation $0 = x_{i,j} - \exp(Y_{i,j}) - (Y_{i,j} - \beta)/\sigma^2$.

In all three cases we use the optimal scaling for a vector of independent standard normal distributed random variates in stationarity, $\hat{\ell}^2/(4096)^{1/3} = 0.1701563$ where $\hat{\ell} = 1.65$.

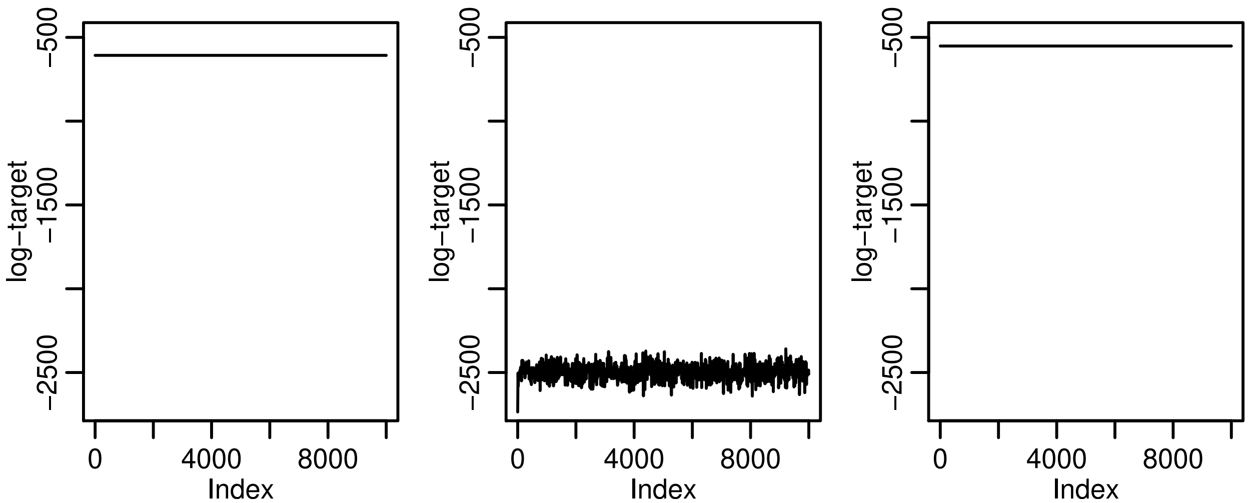

Figure 7. Scots pine saplings. Trace plots $\log f(\gamma \mid \mathbf{x})$ when using the scaling $1.65^2/(4096)^{1/3}$. Left: starting value I. Middle : starting value II. Right: starting value III.

For the starting value II, we observed that the convergence to equilibrium was fast, say less than a hundred iterations. The overall acceptance rate was approximately 54% which is close to the asymptotically optimal value of 57.4% (see Roberts and Rosenthal, 1998). For the starting values I and III all proposals were rejected. Figure 7 shows the trace plots of $\log f(\gamma \mid \mathbf{x})$ from the algorithm with these three different starting values. It can be seen that the cases where the algorithm rejects all proposed moves (I and III) correspond to starting the algorithm way out in the tail of the target distribution.

For comparison, Figure 8 shows trace plots for the starting values I, II and III where we instead used the scaling $\hat{\ell}^2/(4096)^{1/2} = 0.03125$, where $\hat{\ell} = \sqrt{2}$ was found in Section 5 to maximise the speed of convergence when starting at the mode for the standard normal distribution, i.e. maximise $b_\ell(0)$ with $b_\ell(\cdot)$ as in (4). We see that the algorithm approaches the equilibrium distribution very rapidly for all three starting values. The acceptance rate was in all three cases approximately 96%, and the scaling is therefore too small in stationarity, as expected.

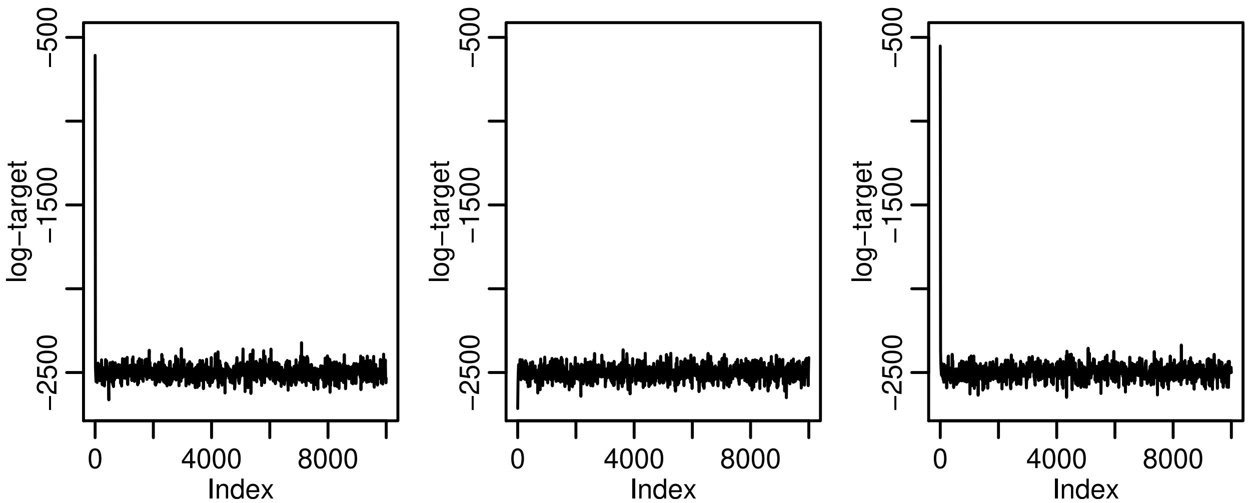

Figure 8. Scots pine saplings. Trace plots $\log f(\gamma \mid \mathbf{x})$ when using the scaling $\sqrt{2}^2/(4096)^{1/2}$. Left: starting value I. Middle: starting value II. Right: starting value III.

In practice, an analysis of this type will need to incorporate parameter uncertainty, and include MCMC update steps for the parameters as well as hidden Gaussian field updates. Since the dimensionality of the parameters is usually small (2 or 3) this can be done using fairly routine methods once careful reparameterisation has been carried out. See Christensen *et al.* (2003) for details.

7. Discussion.

The results above introduce the intriguing suggestion that proposals for Langevin algorithms should be scaled very differently in the transient and stationary phases - that is once stationarity has been reached, the proposal variance can increase from $O(d^{-1/2})$ to $O(d^{-1/3})$ (where for instance an $O(d^{-1/3})$ variance is defined as described in our example in Section 6). In practice, implementing a purely adaptive strategy which does this automatically is not a feasible solution, since it could destroy the stationarity of $\pi(\cdot)$. One simple solution to this problem is to alternate $O(d^{-1/2})$ and $O(d^{-1/3})$ moves in order to cover the possibility of being in either regime at each time-point. This strategy would resemble somewhat that of using a heavy tailed proposal distribution, a suggestion considered in Stramer and Tweedie (1999a,b) to improve the convergence of MCMC algorithms.

In contrast, optimal scalings of Metropolis algorithms have rather stable properties in the tails, and it is always optimal to set the proposal variance to be $O(d^{-1})$. However even in this case, the optimal acceptance rate of the algorithm can vary. To consider this further, we note that to maximise the speed of movement towards convergence, we need to maximise $a_\ell(w)$ as a function of ℓ (separately for each $0 \le w < 1$). The resulting optimal value, $\ell^*(w)$, can be numerically computed, but it depends explicitly on w. Fortunately, reasonable choices of ℓ, including both $\ell = \ell^*(0)$ and ℓ equal to the optimal scaling in stationarity, are not too far from being optimal for all values of $0 \le w < 1$ (see Figure 3).

These results also raise the question of how to chose default starting values for algorithms. This is a very practical issue, clearly of interest for routine use of MCMC algorithms. In particular, for the Langevin algorithm we have seen that natural candidates for default starting values are too near the mode of the distribution, leading to difficulties regarding the scaling of the algorithm.

Most of the rigorous results in this paper are proved for concrete Gaussian target distribution examples. Analogues of Theorems 3 and 6, and also Lemma 5, will no doubt hold for suitably smooth spherically symmetric densities, and Theorem 7 covers a more general class of target distributions. It is unrealistic to hope that blanket results might hold to cover all distributions obtained from complex Bayesian analyses. However, we believe that the broad conclusions of our main results hold rather generally in realistic and complex statistical problems, and we believe this is illustrated through the point-process example of Section 6.

Deterministic behaviour of sample paths might seem surprising, given the stochastic nature of the algorithms we consider. However, stochasticity is still present, it is just being masked by a drift term of relatively overwhelming magnitude. In stationarity of course this drift is not present, so that the stochastic behaviour prevails. Also, in the transient phase, it is still possible to observe stochasticity by considering different functionals. For instance in the standard Gaussian example of Section 3, observing X_1 instead of $\|\mathbf{X}\|$ will give diffusion type behaviour rather than deterministic smooth sample paths. Because our work here is based on starting distributions which are non-stationary, there is no inconsistency between this and the diffusion limit type results of, for example, Roberts and Rosenthal (1998).

As has been seen in other contexts (see for example Roberts and Rosenthal, 2001), our results indicate that simple Metropolis algorithms have rather robust properties in the following sense. They require no special adaptive scaling strategies within the transient phase, and optimal scaling for stationary algorithms turns out to be close to optimal in all regions of the state space. In contrast, the use of Langevin methods can always be quicker, and we have seen that properly scaled Langevin algorithms do hugely outperform competitor Metropolis alternatives (which is crucial in high dimensional examples such as that in Section 6), but great care is required in scaling.

Thus a broad conclusion is that Langevin algorithms are very powerful alternatives to Metropolis algorithms, but their implementation can sometimes require great care. These findings support the use of hybrid strategies which alternate between different update schemes, for instance using both Metropolis and Langevin updates, or by interspersing complicated update schemes with collections of computationally cheaper single site moves.

The focus in this paper is on global scaling of high-dimensional proposals. It should be noted however, that reparameterisation of target densities is also frequently needed for the methods we discuss to be effective. In fact, this is particularly true for Langevin methods as can be seen from Roberts and Rosenthal (2001), and our example in Section 6 uses such a reparameterisation as well.

8. Appendix: Proofs of Results.

Proof of Lemma 1. We write $\mathbf{Y}_{t+1} = \mathbf{X}_t + (\ell^2/d)^{1/2}\mathbf{Z}$, where $\mathbf{Z}$ is standard normal. Then the proposed value for $W^d_{(t+1)/d}$ (given $W^d_{t/d}$) is given by

$$\begin{aligned}\|\mathbf{X}_t + (\ell^2/d)^{1/2}\mathbf{Z}\|^2 &= \|\mathbf{X}_t\|^2 + 2(\ell^2/d)^{1/2}\mathbf{X}_t^T\mathbf{Z} + (\ell^2/d)\|\mathbf{Z}\|^2 \\ &= \|\mathbf{X}_t\|^2 + V_d + \epsilon_d,\end{aligned}$$

say, where we set $\epsilon_d = \ell^2\left(d^{-1}\|\mathbf{Z}\|^2 - 1\right)$ and $V_d = 2(\ell^2/d)^{1/2}\mathbf{X}_t^T\mathbf{Z} + \ell^2$. Thus $V_d \sim N(\ell^2, 4\ell^2\|\mathbf{X}_t\|^2/d)$, i.e. $N(\ell^2, 4\ell^2 w)$. Notice that $\epsilon_d \to 0$ in L^1 by a simple law of large numbers argument.

Now, this proposed value is then *accepted* with probability equal to the minimum of 1 and

$$\frac{\pi(\mathbf{Y}_{t+1})}{\pi(\mathbf{X}_t)} = \frac{\exp(-\|\mathbf{Y}_{t+1}\|^2/2)}{\exp(-\|\mathbf{X}_t\|^2/2)}$$
$$= \exp(-(\|\mathbf{Y}_{t+1}\|^2 - \|\mathbf{X}_t\|^2)/2)\ .$$

Therefore by the L^1 convergence of ϵ_d and the fact that the function $x \mapsto x \min\left(1, e^{-x/2}\right)$ is a contraction

$$\mathbf{E}\left[W^d_{(t+1)/d} - W^d_{t/d} \,|\, W^d_{t/d} = w\right] d = \mathbf{E}\left[(V_d + \epsilon_d) \min\left(1, \exp(-(V_d + \epsilon_d))\right]\right. \tag{6}$$
$$\to \mathbf{E}\left[(\ell^2 + 2\ell w^{1/2} U)\ \min\left(1,\ \exp(-(\ell^2 + 2\ell w^{1/2} U)/2)\right)\right]$$

with the last expectation taken with respect to $U \sim N(0,1)$. Since ϵ_d is independent of w, it is easy to see that this convergence takes place uniformly for $w \in [0, K]$ for any $K > 0$.

The remainder of the proof is concerned with computing this expectation, which is straightforward but tedious.

We first recall that $N^* = -\ell w^{-1/2}/2$, and note that the above "min" equals 1 for $U \le N^*$ only. Hence, as $d \to \infty$,

$$\mathbf{E}\left[W^d_{(t+1)/d} - W^d_{t/d} \,|\, W^d_{t/d} = w\right] d \ \approx\ \int_{N^*}^{\infty} \phi(u) \exp(-\ell^2/2 - \ell w^{1/2} u)(\ell^2 + 2\ell w^{1/2} u) du$$
$$+ \int_{-\infty}^{N^*} \phi(u)(\ell^2 + 2\ell w^{1/2} u) du \ \equiv\ I_1 + I_2\,.$$

Then

$$I_1 \ = \ \int_{N^*}^{\infty} \phi(u + \ell w^{1/2}) \exp(\ell^2 w/2 - \ell^2/2)\left(2\ell w^{1/2}(u + \ell w^{1/2}) - 2\ell^2 w + \ell^2\right) du$$
$$= \ \exp(\ell^2 w/2 - \ell^2/2)\left\{2\ell w^{1/2} \phi(N^* + \ell w^{1/2}) + (-2\ell^2 w + \ell^2)\left[1 - \Phi\left(N^* + \ell w^{1/2}\right)\right]\right\}\,,$$
$$= \ 2\ell w^{1/2}\phi(N^*) + \exp(\ell^2 w/2 - \ell^2/2)(-2\ell^2 w + \ell^2)\Phi\left(-N^* - \ell w^{1/2}\right)\,,$$

where we first used $\phi(u)\exp(-\ell w^{1/2} u) = \phi(u+\ell w^{1/2})\exp(\ell^2 w/2)$, second that $\int_{u^*}^{\infty} u\phi(u) du = \phi(u^*)$, and third that $\phi(N^* + \ell w^{1/2}) = \phi(N^*)\exp(\ell^2/2 - \ell^2 w/2)$.

Similarly,

$$I_2 \ = \ \ell^2 \Phi(N^*) - 2\ell w^{1/2} \phi(N^*)\,.$$

Combining the expressions for I_1 and I_2, the result follows. ∎

Proof of Lemma 2. Letting $U \sim N(0,1)$, then for d large

$$\mathbf{E}\left[(W^d_{(t+1)/d} - W^d_{t/d})^2 \mid W^d_{t/d} = w\right] d^2 = \mathbf{E}\left[(V_d + \epsilon_d)^2 \min\left(1, \exp(-(V_d + \epsilon_d))\right)\right]$$

$$\leq \mathbf{E}\left[(V_d + \epsilon_d)^2\right]$$

which is easily seen to be bounded as a function of w as $d \to \infty$. ∎

Proof of Theorem 3. Let G^d denote the discrete time generator of W^d and let $\mathcal{C} = \{C_c^\infty \text{ functions} : [0,\infty) \to \mathbf{R}\}$. (A function $h \in C_c^\infty$ if it is infinitely differentiable with compact support.) Then for $h \in \mathcal{C}$,

$$G^d h(w) = d\mathbf{E}\left[h(W^d_{(t+1)/d}) - h(w) | W^d_{t/d} = w\right]$$

$$= d\,\mathbf{E}\left[h'(w)(W^d_{(t+1)/d} - w) + h''(W^*)(W^d_{(t+1)/d} - w)^2/2 | W^d_{t/d} = w\right]$$

where W^* represents a value in between w and $W^d_{(t+1)/d}$. However the second term on the right hand side converges to 0 uniformly in w by Lemma 2, so that by Lemma 1

$$\lim_{d\to\infty} \sup_{0 \leq w \leq K} \left|G^d h(w) - a_\ell(w) h'(w)\right| = 0 \ . \tag{7}$$

Thus the infinitesimal behaviour of the speeded up W^d converges to that of the limiting diffusion. The remaining technicalities involve deducing that convergence of infinitesimals is sufficient to ensure weak convergence of the process. This final step follows from the following argument which is typical of that used for weak convergence results in Ethier and Kurtz (1986).

$\mathcal{C}$ is a *core* for the generator of the deterministic process described by (3) (see Section 3 of Chapter 1, and Theorem 2.1 of Chapter 8 of Ethier and Kurtz (1986)). So by Theorem 6.5 of Chapter 1 and (7), the finite dimensional distributions of W^d converge to the required limit. Moreover, from Corollary 8.7 of Chapter 4 of Ethier and Kurtz (1986), this is therefore sufficient to ensure the weak convergence statement in (3) thus proving Theorem 3. ∎

Proof of Lemma 4. We use techniques similar to the proof of Lemma 1. For the Langevin algorithm, $\mathbf{Y}_{t+1} = (1 - h/2)\mathbf{X}_t + \sqrt{h}\mathbf{Z}$ with $\mathbf{Z}$ being a standard multivariate normal random vector. From this we get

$$\|\mathbf{Y}_{t+1}\|^2 = (1-h/2)^2\|\mathbf{X}_t\|^2 + 2(1-h/2)h^{1/2}\mathbf{X}_t^T\mathbf{Z} + h\|\mathbf{Z}\|^2$$

$$\approx (1 - h + h^2/4)\|\mathbf{X}_t\|^2 + 2h^{1/2}(\|\mathbf{X}_t\|^2)^{1/2}U + hd$$

where $U \sim N(0,1)$. Using that $h \to 0$ and $hd \to \infty$ we can write

$$\|\mathbf{Y}_{t+1}\|^2 - \|\mathbf{X}_t\|^2 \approx hd(1-w)$$

The acceptance probability for $\mathbf{Y}_{t+1}$ becomes the minimum of 1 and

$$\frac{\pi(\mathbf{Y}_{t+1})}{\pi(\mathbf{X}_t)}\frac{q(\mathbf{Y}_{t+1},\mathbf{X}_t)}{q(\mathbf{X}_t,\mathbf{Y}_{t+1})} = \frac{\exp(-\|\mathbf{Y}_{t+1}\|^2/2)}{\exp(-\|\mathbf{X}_t\|^2/2)}\frac{\exp(-\|\mathbf{X}_t - (1-h/2)\mathbf{Y}_{t+1}\|^2/2h)}{\exp(-\|\mathbf{Y}_{t+1} - (1-h/2)\mathbf{X}_t\|^2/2h)}$$

$$= \exp(-h(\|\mathbf{Y}_{t+1}\|^2 - \|\mathbf{X}_t\|^2)/8) \approx \exp(-h^2 d(1-w)/8).$$

Thus for large d

$$\mathbf{E}[W^d_{(t+1)/d^{1/3}} - W^d_{t/d^{1/3}} \mid W^d_{t/d^{1/3}} = w] \approx hd(1-w)\min\{1, \exp(-h^2 d(1-w)/8)\}/d,$$

proving (by substituting $h = \ell^2 d^{-1/3}$) the result. ∎

Proof of Lemma 5. The first part of the lemma follows from the proof of Lemma 4, using $h = \ell^2/d^{1/2}$.

The second part of the lemma follows in a similar manner to the proof of Lemma 2, and the details are therefore omitted. ∎

Proof of Theorem 6. This theorem follows from Lemma 4 and Lemma 5, similar to how Theorem 3 followed from Lemmas 1 and 2. The details are therefore omitted. ∎

Proof of Theorem 7. The acceptance probability for $\mathbf{Y}_{t+1}$ equals the minimum of 1 and

$$\frac{\pi(\mathbf{Y}_{t+1})}{\pi(\mathbf{X}_t)}\frac{q(\mathbf{Y}_{t+1},\mathbf{X}_t)}{q(\mathbf{X}_t,\mathbf{Y}_{t+1})} = \frac{\exp(\sum_{i=1}^d g(Y_{t+1,i}))}{\exp(\sum_{i=1}^d g(X_{t,i}))}\frac{\exp(-\sum_{i=1}^d (X_{t,i}-Y_{t+1,i}-h/2\times g'(Y_{t+1,i}))^2/2h)}{\exp(-\sum_{i=1}^d (Y_{t+1,i}-X_{t,i}-h/2\times g'(X_{t,i}))^2/2h)}$$

$$= \exp(-(h/8)\sum_{i=1}^d (g'(Y_{t+1,i})^2 - g'(X_{t,i})^2))$$

$$\times \exp(\sum_{i=1}^d (g(Y_{t+1,i}) - g(X_{t,i}) - (g'(Y_{t+1,i}) + g'(X_{t,i}))(Y_{t+1,i}-X_{t,i})/2)).$$

This first term is similar to the term we got for the normal target density. Making a third and fourth order Taylor expansion of $g'(Y_{t+1,i})$ and $g(Y_{t+1,i})$, respectively, we get that the last term is approximately

$$\exp(-\sum_{i=1}^d (g'''(X_{t,i})(Y_{t+1,i}-X_{t,i})^3/12 + g''''(X_{t,i})(Y_{t+1,i}-X_{t,i})^4/24)).$$

Now we use the fact that $\mathbf{X}_t$ is equal to a mode of the distribution (assumed for notational simplicity to be the origin, $\mathbf{0}$).

Thus $g'(X_{t,i}) = 0$ and $X_{t,i} = 0$, for $i = 1,\ldots,d$. This gives us

$$\frac{\pi(\mathbf{Y}_{t+1})}{\pi(\mathbf{0})}\frac{q(\mathbf{Y}_{t+1},\mathbf{0})}{q(\mathbf{0},\mathbf{Y}_{t+1})} \approx \exp(-(h/8)g''(0)^2\sum_{i=1}^d Y_{t+1,i}^2)$$

$$\times \exp(-g'''(0)\sum_{i=1}^d Y_{t+1,i}^3/12 - g''''(0)\sum_{i=1}^d Y_{t+1,i}^4/24).$$

The proposal in this case is $\mathbf{Y}_{t+1} = h^{1/2}\mathbf{Z}$, i.e. $Y_{t+1,i} = h^{1/2}Z_i$, with $\mathbf{Z}$ being a standard multivariate normal random vector. Now observe that $\sum_{i=1}^d Y_{t+1,i}^3$ has mean zero, and therefore by the Central Limit Theorem is of order $d^{1/2}h^{3/2}$. Using the Law of

Large Numbers we see that the terms $\sum_{i=1}^d Y_{t+1,i}^2$ and $\sum_{i=1}^d Y_{t+1,i}^4$ are asymptotically dh and $3dh^2$, respectively. Therefore in the acceptance probability expression, the $Y_{t,i}^3$ terms are negligible as $d \to \infty$, and

$$\frac{\pi(\mathbf{Y}_{t+1})}{\pi(\mathbf{0})} \frac{q(\mathbf{Y}_{t+1}, \mathbf{0})}{q(\mathbf{0}, \mathbf{Y}_{t+1})} \approx \exp(-h^2 d(g''(0)^2 + g''''(0))/8)$$

$$= \exp(-\ell^4 (g''(0)^2 + g''''(0))/8) = c, \quad \text{say,}$$

when d is large and $h = \ell^2 d^{-1/2}$. Thus, asymptotically as $d \to \infty$, all proposed moves will be accepted with the same positive acceptance probability c. This proves Theorem 7. ∎

Acknowledgements The authors are grateful to the referees, associate editor and editor for constructive comments and suggestions. The authors also thank Alexandros Beskos for producing the plot in Figure 1.

REFERENCES

Bédard, M. (2004), On the robustness of optimal scaling for Metropolis-Hastings algorithms. Ph.D. dissertation, University of Toronto. Work in progress.

Beskos, A., Papaspiliopoulos, O., Roberts, G.O. and Fearnhead P.N. (2004). Exact likelihood-based estimation of diffusion processes. In progress.

Breyer, L. and Roberts, G.O. (2000), From Metropolis to diffusions: Gibbs states and optimal scaling. *Stoch. Proc. Appl*, **90**, 181–206.

Christensen, O.F. and Waagepetersen, R. (2002), Bayesian prediction of spatial count data using generalized linear mixed models. *Biometrics* **58**, 280–286.

Christensen, O.F., Roberts, G.O. and Sköld, M. (2003). Robust MCMC methods for spatial GLMM's. Technical Report **23**, Centre for Mathematical Sciences, Lund University.

Ethier, S.N. and Kurtz, T.G. (1986), *Markov processes: Characterization and Convergence*. Wiley, New York.

Gelman, A., Roberts, G. O., and Gilks, W.R. (1996), Efficient Metropolis jumping rules. In: *Bayesian statistics 5* (eds. J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith), Oxford University Press, 599–608.

Gilks, W.R., Richardson, S., and Spiegelhalter, D.J., eds. (1996), *Markov chain Monte Carlo in practice*. Chapman and Hall, London.

Grenander U. and Miller, M.I. (1994), Representations of knowledge in complex systems, *J. Roy. Stat. Soc.* B **56**, 3, 549–603.

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

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.

Møller, J., Syversveen A.R. and Waagepetersen, R. (1998), Log Gaussian Cox processes. *Scand. J. Statist.* **25**, 451–482.

Neal, R. (1996) Bayesian Learning for Neural Networks, Lecture Notes in Statistics No. 118, New York: Springer-Verlag.

Papaspiliopoulos, O., Roberts, G.O. and Sköld, M. (2003). Non-centered parameterisations for hierarchical models and data augmentation. In: *Bayesian statistics 7* (eds. J.M. Bernardo, S. Bayarri, J.O. Berger, A.P. Dawid, D. Heckerman, A.F.M. Smith and M. West), Oxford University Press, 307–326.

Roberts, G.O. (1998), Optimal Metropolis algorithms for product measures on the vertices of a hypercube. *Stochastics and Stochastic Reports* **62**, 275–283.

Roberts, G.O., Gelman, A. and Gilks, W.R. (1997), Weak convergence and optimal scaling of random walk Metropolis algorithms. *Ann. Applied Probability* **7**, 110–120.

Roberts, G.O. and Rosenthal, J.S. (1998), Optimal scaling of discrete approximations to Langevin diffusions. *J. Roy. Stat. Soc.* B **60**, 255–268.

Roberts, G.O. and Rosenthal, J.S. (2001), Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science* **16**, 351–367.

Roberts, G.O. and Tweedie, R.L. (1996), Exponential convergence of Langevin diffusions and their discrete approximations, *Bernoulli*, **2**, 4, 341–363.

Roberts, G.O. and Yuen, W.K. (2004), Optimal scaling of Metropolis algorithms for discontinuous densities. Work in progress.

Stramer, O. and Tweedie, R.L. (1999a), Langevin-Type Models I: Diffusions with Given Stationary Distributions, and Their Discretizations. Methodology & Computing in Applied Probability **1**, 283–306.

Stramer, O. and Tweedie, R.L. (1999b), Langevin-Type Models II: Self-Targeting Candidates for MCMC Algorithms. Methodology & Computing in Applied Probability **1**, 307–328.