Submitted to the Annals of Statistics

ON THE EFFICIENCY OF PSEUDO-MARGINAL RANDOM WALK METROPOLIS ALGORITHMS

By Chris Sherlock, Alexandre H. Thiery, Gareth O. Roberts, and Jeffrey S. Rosenthal

Lancaster University, National University of Singapore, University of Warwick and University of Toronto

E-mail: c.sherlock@lancaster.ac.uk; a.h.thiery@nus.edu.sg

We examine the behaviour of the pseudo-marginal random walk Metropolis algorithm, where evaluations of the target density for the accept/reject probability are estimated rather than computed precisely. Under relatively general conditions on the target distribution, we obtain limiting formulae for the acceptance rate and for the expected squared jump distance, as the dimension of the target approaches infinity, under the assumption that the noise in the estimate of the log-target is additive and is independent of the position. For targets with independent and identically distributed components, we also obtain a limiting diffusion for the first component.

We then consider the overall efficiency of the algorithm, in terms of both speed of mixing and computational time. Assuming the additive noise is Gaussian and is inversely proportional to the number of unbiased estimates that are used, we prove that the algorithm is optimally efficient when the variance of the noise is approximately 3.283 and the acceptance rate is approximately 7.001%. We also find that the optimal scaling is insensitive to the noise and that the optimal variance of the noise is insensitive to the scaling. The theory is illustrated with a simulation study using the particle marginal random walk Metropolis.

1. Introduction. Markov chain Monte Carlo (MCMC) algorithms have proved particularly successful in statistics for investigating posterior distributions in Bayesian analysis of complex models (see e.g. [SR93, Tie94, BGJM11]). Almost all MCMC methods are based on the Metropolis-Hastings (MH) algorithm which owes much of its success to its tremendous flexibility. However, in order to use the classical MH algorithm, it must be possible to evaluate the target density up to a fixed constant of proportionality. While this is often possible, it is increasingly common for exact pointwise likelihood evaluation to be prohibitively expensive, perhaps due to the sheer size of the data set being analysed. In these situations, classical MH is rendered inapplicable.

The *pseudo-marginal Metropolis-Hastings algorithm* (PsMMH) [Bea03, AR09] provides a general recipe for circumventing the need for target density evaluation. Instead it is required only to be able to unbiasedly *estimate* this density. The target densities in the numerator and denominator of the MH accept/reject ratio are then replaced by their unbiased estimates. Remarkably, this yields an algorithm which still has the target as its invariant distribution. One possible choice of algorithm, the *pseudo-marginal random walk Metropolis* (PsMRWM), is popular in practice (e.g.

MSC 2010 subject classifications: Primary 65C05,65C40,60F05

Keywords and phrases: Markov chain Monte Carlo, MCMC, Pseudo-marginal random walk Metropolis, Optimal scaling, Diffusion limit, Particle methods

[GW11, KdV12]) because it requires no further information about the target, such as the local gradient or Hessian, which are generally more computationally expensive to approximate than the target itself [PDS11].

Broadly speaking, the mixing rate of any PsMMH algorithm decreases as the dispersion in the estimation of the target density increases [AR09]. In particular, if the target density happens to be substantially over-estimated then the chain will be overly reluctant to move from that state leading to a long run of successive rejections (a *sticky patch*). Now, in PsMMH algorithms, the target estimate is usually computed using an average of some number, m, of approximations (see Sections 1.1 and 3). This leads to a trade off, with increasing m leading to better mixing of the chain, but also to larger computational expense. We shall consider the problem of optimising m.

It is well known (e.g. [RR01, SFR10]) that the efficiency of the random-walk Metropolis (RWM) algorithm varies enormously with the scale of the proposed jumps. Small proposed jumps lead to high acceptance rates but little movement across the state space, whereas large proposed jumps lead to low acceptance rates and again to inefficient exploration of the state space. The problem of choosing the optimal scale of the RWM proposal has been tackled for various shapes of target (e.g. [RGG97, RR01, Béd07, BR08, BRS09, SR09, She13, BR00]) and has led to the following rule of thumb: choose the scale so that the acceptance rate is approximately 0.234. Although nearly all of the theoretical results are based upon limiting arguments in high dimension, the rule of thumb appears to be applicable even in relatively low dimensions (e.g. [SFR10]).

This article focusses on the efficiency of the PsMRWM as the dimension of the target density diverges to infinity. For relatively general forms of the target distribution, under the assumption of additive independent noise in the log-target, we obtain (Theorem 1) expressions for the limiting expected squared jump distance (ESJD) and asymptotic acceptance rate. ESJD is now well-established as a pragmatic and useful measure of mixing for MCMC algorithms in many contexts (see e.g. [PG10]), and is particularly relevant when diffusion limits can be established (see for example the discussion in [RR13]). We then prove a diffusion limit for a rescaling of the first component, in the case of a target with independent and identically distributed components (Theorem 2); the efficiency of the algorithm is then given by the speed of this limiting diffusion, which is equivalent to the limiting ESJD. We examine the relationship between efficiency, scaling, and the distributional form of the noise, and consider the *joint* optimisation of the efficiency of the PsMRWM algorithm (taking computational time into account) with respect to m, and the RWM scale parameter. Exact analytical results are obtained (Corollary 1) under an assumption of Gaussian noise in the estimate of the log-target, with a variance that is inversely proportional to m. In that case, we prove that the optimal noise variance is 3.283, and the corresponding optimal asymptotic acceptance rate is 7.001%, thus extending the previous 23.4% result of [RGG97]. Finally, we illustrate the use of these theoretical results in a simulation study (Section 4).

1.1. *The PsMRWM.* Consider a state space $\mathcal{X} \subseteq \mathbb{R}^d$ and let $\pi(\cdot)$ be a distribution on $\mathcal{X}$, whose density (with respect to Lebesgue measure) will be referred to as $\pi(\mathbf{x})$. The MH updating scheme provides a very general class of algorithms for obtaining an approximate dependent sample from a target distribution, $\pi(\cdot)$, by constructing a Markov chain with $\pi(\cdot)$ as its limiting distribution. Given the current value $\mathbf{x}$, a new value $\mathbf{x}^*$ is proposed from a pre-specified Lebesgue density $q(\mathbf{x}, \mathbf{x}^*)$ and is then accepted with probability $\alpha(\mathbf{x}, \mathbf{x}^*) = 1 \wedge [\pi(\mathbf{x}^*)\, q(\mathbf{x}^*, \mathbf{x})]/[\pi(\mathbf{x})\, q(\mathbf{x}, \mathbf{x}^*)]$. If the proposed value is accepted then it becomes the next current value, otherwise the current value is left unchanged.

The PsMMH algorithm [AR09] presumes the computational infeasibility of evaluating $\pi(\mathbf{x})$ and uses an approximation $\hat{\pi}_{\mathbf{v}}(\mathbf{x})$ that depends on some auxiliary variable, $\mathbf{v}$. The auxiliary variable is sampled from some distribution $q_{\text{aux}}(\mathbf{v}|\mathbf{x})$, and the approximation $\hat{\pi}_{\mathbf{v}}(\mathbf{x})$ is assumed to satisfy that $\mathbb{E}_{q_{aux}}[\hat{\pi}_{\mathbf{V}}(\mathbf{x})] = c\pi(\mathbf{x})$, for some constant $c > 0$. The value of the constant is irrelevant to all that

follows, and so, without loss of generality, we assume that $c = 1$. We also assume that $\hat{\pi}_{\mathbf{v}} > 0$.

The PsMMH algorithm creates a Markov chain; with a stationary density (since $c = 1$) of

$$\tilde{\pi}(\mathbf{x}, \mathbf{v}) = q_{\text{aux}}(\mathbf{x}, \mathbf{v})\hat{\pi}_{\mathbf{v}}(\mathbf{x}), \tag{1.1}$$

which has $\pi(\mathbf{x})$ as its $\mathbf{x}$ marginal. When a new value, $\mathbf{X}^*$, is proposed via the MH algorithm, a new auxiliary variable, $\mathbf{V}^*$, is proposed from the density $q_{\text{aux}}(\mathbf{x}^*, \mathbf{v}^*)$. The pair $(\mathbf{x}^*, \mathbf{v}^*)$ are then jointly accepted or rejected. The acceptance probability for this MH algorithm on $(\mathbf{x}, \mathbf{v})$ is

$$1 \wedge \frac{\hat{\pi}_{\mathbf{v}^*}(\mathbf{x}^*)\ q\left(\mathbf{x}^*, \mathbf{x}\right)}{\hat{\pi}_{\mathbf{v}}(\mathbf{x})\ q\left(\mathbf{x}, \mathbf{x}^*\right)}.$$

We are thus able to substitute the estimated density for the true density, and still obtain the desired stationary distribution for $\mathbf{x}$. Note that for symmetric proposals, this simplifies to $1 \wedge \left[\hat{\pi}_{\mathbf{v}^*}(\mathbf{x}^*)/\hat{\pi}_{\mathbf{v}}(\mathbf{x})\right]$.

Different strategies exist for producing unbiased estimators, for instance using importance sampling or latent variable representations, as in [FPR08], or using particle filters ([GSS93], [DM04]) as in [ADH10]. We shall illustrate our theory in the context of Bayesian analysis of a partially observed Markov jump process.

1.2. *Previous related literature.* [PdSSGK12] and [DPDK14] both examine the efficiency of pseudo-marginal algorithms using bounds on the integrated autocorrelation time (I_{ACT}) and under the assumptions that the chain is stationary and the distribution of the additive noise in the log-target is independent of $\mathbf{x}$ (our Assumptions 1). Under the further assumption that this additive noise is Gaussian and the computing time inversely proportional to its variance (our Assumptions 4), both articles then seek information on the optimal variance of this additive noise. [PdSSGK12] consider the (unrealistic) case where the Metropolis-Hastings algorithm is an independence sampler which proposes from the desired target distribution for x, and obtain an optimal variance of 0.92^2. [DPDK14] consider a general Metropolis-Hastings algorithm and define a parallel hypothetical kernel Q^* with the same proposal mechanism as the original kernel, Q, but where the acceptance rate separates into the product of that of the idealised marginal algorithm (if the true target were known) and that of an independence sampler which proposes from the assumed distribution for the noise. This kernel can never be more efficient than the true kernel. Upper and lower bounds are obtained for the I_{ACT} for Q^* in terms of the of I_{ACT} of the exact chain and the I_{ACT} and a particular lag-1 autocorrelation of the independence sampler on the noise. These bounds are examined under the assumption that the additive noise is Gaussian and the optimal variance for the noise is estimated to lie between 0.92^2 and 1.68^2.

Other theoretical properties of pseudo-marginal algorithms are considered in [AV14], which gives qualitative (geometric and polynomial ergodicity) results for the method and some results concerning the loss in efficiency caused by having to estimate the target density.

1.3. *Notation.* In this paper, we follow the standard convention whereby capital letters denote random variables, and lower case letters denote their actual values. Bold characters are used to denote vectors or matrices.

2. Studying the Pseudo Marginal Random Walk Metropolis in high dimensions.

2.1. *Proposal distribution.* We focus on the case where the proposal, $\mathbf{x}^*$, for an update to $\mathbf{x}$ is assumed to arise from a random walk Metropolis algorithm with an isotropic Gaussian proposal

$$\mathbf{X}^* = \mathbf{x} + \lambda\mathbf{Z}, \qquad \text{where} \qquad \mathbf{Z} \overset{\mathcal{D}}{\sim} \mathbf{N}(\mathbf{0}, \mathbf{I}), \tag{2.1}$$

and $\mathbf{I}$ is the $d \times d$ identity matrix, and $\lambda > 0$ is the scaling parameter for the proposal. The results presented in this article extend easily to a more general correlation matrix by simply considering the linear co-ordinate transformation which maps this correlation matrix to the identity matrix and examining the target in this transformed space. In proving the limiting results we consider a sequence of d-dimensional target probabilities $\pi^{(d)}$. In dimension d the proposal is $\mathbf{X}^{(d)*} \overset{\mathcal{D}}{\sim} \mathbf{N}(\mathbf{x}^{(d)}, \lambda^{(d)2}\mathbf{I}^{(d)})$.

2.2. *Noise in the estimate of the log-target.* We will work throughout with the log-density of the target, and it will be convenient to consider the difference between the estimated log-target ($\log \hat{\pi}_V(\mathbf{x})$) and the true log-target ($\log \pi(\mathbf{x})$) at both the proposed values $(\mathbf{x}^*, V^*)$ and the current values $(\mathbf{x}, V)$, as well as the difference between these two differences:

$$\begin{cases} W & := \log \hat{\pi}_V(\mathbf{x}) - \log \pi(\mathbf{x}), \\ W^* & := \log \hat{\pi}_{V^*}(\mathbf{x}^*) - \log \pi(\mathbf{x}^*), \\ B & := W^* - W. \end{cases} \tag{2.2}$$

Throughout this article we assume the following.

ASSUMPTIONS 1. *The Markov chain $(\mathbf{X}, W) = \{(\mathbf{X}_k, W_k)\}_{k \geq 0}$ is stationary, and the distribution of the additive noise in the estimated log-target at the proposal, W^*, is independent of the proposal itself, $\mathbf{X}^*$.*

REMARK 1. *It is unrealistic to believe that the second part of Assumptions 1 should hold in practice. Pragmatically, this assumption is necessary in order to make progress with the theory presented herein; however, in our simulation study in Section 4 we provide evidence that, in the scenarios considered, the variation in the noise distribution is relatively small.*

Note that the noise term within the Markov chain, W, does not have the same distribution as the noise in the proposal, W^*, since, for example, moves away from positive values of W will be more likely to be rejected than moves away from negative values of W. In the notation of Section 1.1, since W^* is a function of $\mathbf{V}$, $q_{\text{aux}}(\mathbf{x}^*, \mathbf{v})$ now gives rise to $g^*(w^*)$, the density of the noise in the estimate of the log-target, which is independent of $\mathbf{x}^*$. Integrating (1.1) gives the joint stationary density of the Markov chain $(\mathbf{X}, W)$ as

$$g^*(w)e^w \pi(\mathbf{x}). \tag{2.3}$$

This is Lemma 1 of [PdSSGK12]. Under Assumptions 1, W and $\mathbf{X}$ are therefore independent, and the stationary density of W is $g^*(w)e^w$.

2.3. *High dimensional target distribution.* We describe in this section conditions on the sequence of target densities $\pi^{(d)}$ that ensure that the quantity $\log\left[\pi^{(d)}(\mathbf{X}^*)/\pi^{(d)}(\mathbf{X})\right]$ behaves asymptotically as a Gaussian distribution under an appropriate choice of jump scaling $\lambda^{(d)}$. The main assumption is that there exist sequences of scalings $s_g^{(d)} > 0$ and $s_L^{(d)} > 0$ for the gradient and the Laplacian of the log-likelihood $\log \pi^{(d)}$ such that the following two limits hold in probability,

$$\lim_{d\to\infty} \frac{\left\|\nabla \log \pi^{(d)}(\mathbf{X}^{(d)})\right\|}{s_g^{(d)}} = 1 \qquad \text{and} \qquad \lim_{d\to\infty} \frac{\Delta \log \pi^{(d)}(\mathbf{X}^{(d)})}{s_L^{(d)}} = -1, \tag{2.4}$$

for $\mathbf{X}^{(d)} \overset{\mathcal{D}}{\sim} \pi^{(d)}$. In the rest of this article we assume that the sequence of densities $\pi^{(d)}$ is such that for each index $i \geq 1$, with all components of $\mathbf{x}$ fixed except the i^{th}, the i^{th} component satisfies

$$\frac{\partial \pi^{(d)}}{\partial x_i} \to 0 \qquad \text{as} \qquad |x_i| \to \infty. \tag{2.5}$$

Under this regularity condition, an integration by parts shows that $\mathbb{E}\big[\|\nabla \log \pi^{(d)}(\mathbf{X}^{(d)})\|^2\big] = -\mathbb{E}\big[\Delta \log \pi^{(d)}(\mathbf{X}^{(d)})\big]$. Equation (2.4) thus yields $\lim_{d\to\infty}(s_g^{(d)})^2/s_L^{(d)} = 1$. We will suppose from now on, without loss of generality, that $s_g^{(d)} = \sqrt{s_L^{(d)}} =: s^{(d)}$. We also require that no single component of the local Hessian $H^{(d)}(\mathbf{x}) := \big[\partial^2_{ij} \log \pi^{(d)}(\mathbf{x})\big]_{0\le i,j\le d}$ dominate the others in the sense that the limit

$$\lim_{d\to\infty} \frac{\mathrm{Trace}\big[\big(H^{(d)}\big)^2(\mathbf{X}^{(d)})\big]}{\big(s^{(d)}\big)^4} = 0 \tag{2.6}$$

holds in probability. We also assume that the Hessian matrix is sufficiently regular so that for any $\sigma^2, \varepsilon > 0$ and $\mathbf{Z}^{(d)} \overset{\mathcal{D}}{\sim} \mathbf{N}(0, \mathbf{I}^{(d)})$

$$\lim_{d\to\infty} \mathbb{P}\Big(\sup_{t\in(0,1)} \Big| \frac{\langle \mathbf{Z}^{(d)}, [H^{(d)}(\mathbf{X}^{(d)} + t\,\sigma\,\mathbf{Z}^{(d)}/s^{(d)}) - H^{(d)}(\mathbf{X}^{(d)})]\,\mathbf{Z}^{(d)}\rangle}{(s^{(d)})^2} \Big| > \varepsilon \Big) = 0. \tag{2.7}$$

These conditions are discussed in detail in [She13] where they are shown to hold, for example, when the target is the joint distribution of successive elements of a class of finite-order order multivariate Markov processes. The targets considered in [RGG97], [RR01] and Section 2.5 all satisfy the conditions with $s^{(d)} \propto d^{1/2}$. We record the conditions formally as:

ASSUMPTIONS 2. *The sequence of densities $\pi^{(d)}$ satisfies Equations (2.4),(2.6),(2.7) and the regularity condition (2.5).*

We shall show in next section that under these assumptions the choice of jump size

$$\lambda^{(d)} := \frac{\ell}{s^{(d)}} \tag{2.8}$$

for a parameter $\ell > 0$ leads to a Gaussian asymptotic behaviour for $\log[\pi^{(d)}(\mathbf{X}^*)/\pi^{(d)}(\mathbf{X})]$. This ensures that for high dimensions, the mean acceptance probability $\alpha^{(d)}(\ell)$ of the MCMC algorithm,

$$\alpha^{(d)}(\ell) := \mathbb{E}\Big[1 \wedge \frac{\pi^{(d)}\left(\mathbf{X}^{(d)} + \lambda^{(d)}\mathbf{Z}^{(d)}\right) e^{W^*}}{\pi^{(d)}\left(\mathbf{X}^{(d)}\right) e^{W}}\Big],$$

stays bounded away from zero and one.

2.4. *Expected squared jump distance.* A standard measure of efficiency for local algorithms is the Euclidian Expected Squared Jumping Distance (e.g. [SR09], [BRS09], [She13]) usually defined as $\mathbb{E}\|\mathbf{X}_{k+1} - \mathbf{X}_k\|^2$. Consider, for example, a target with elliptical contours, or one which has components which are independent and identically distributed up to a scale parameter. In such situations the Euclidean ESJD is dominated by those components with a larger scale. We would prefer an efficiency criterion which weights components at least approximately equally, so that moves along each component are considered relative to the scale of variability of that component. A squared Mahalanobis distance is the natural extension of Euclidean ESJD, and in the case of the two example targets mentioned above, it is exactly the correct generalisation of Euclidean ESJD. We therefore define a generalised potential squared jump distance for a single iteration with respect to some $d \times d$ positive definite symmetric matrix $\mathbf{T}^{(d)}$, $\mathbb{E}\Big[\,\|\mathbf{X}^{(d)}_{k+1} - \mathbf{X}^{(d)}_k\|^2_{\mathbf{T}^{(d)}}\,\Big]$, where the Markov

chain $\{\mathbf{X}_k^{(d)}\}_{k\geq 0}$ is assumed to evolve at stationarity and $\|z\|^2_{\mathbf{T}^{(d)}} := \langle z, \mathbf{T}^{(d)} z\rangle$. We will require that, in the limit as $d \to \infty$, no one principal component of $\mathbf{T}^{(d)}$ dominates the others in the sense that

$$\text{Trace}\left[\left(\mathbf{T}^{(d)}\right)^2\right] / \text{Trace}\left[\mathbf{T}^{(d)}\right]^2 \to 0. \tag{2.9}$$

Clearly, (2.9) is satisfied when $\mathbf{T}^{(d)} = I_d$ (*i.e.* Euclidian ESJD).

THEOREM 1. *Consider a PsMRWM algorithm. Assume that the additive noise satisfies Assumptions 1, the sequence of densities $\pi^{(d)}$ satisfy Assumptions 2, and the sequence of jump distance matrices $\mathbf{T}^{(d)}$ satisfy (2.9). Assume further that the jump size $\lambda^{(d)}$ is given by (2.8) for some fixed $\ell > 0$.*

1. **Acceptance probability.**
   The mean acceptance probabilities $\alpha^{(d)}(\ell)$ converge as $d \to \infty$ to a non-trivial value $\alpha(\ell)$,

$$\lim_{d\to\infty} \alpha^{(d)}(\ell) = 2 \times \mathbb{E}\left[\Phi\left(\frac{B}{\ell} - \frac{\ell}{2}\right)\right] =: \alpha(\ell), \tag{2.10}$$

   with B as in (2.2), where Φ is the cumulative distribution of a standard Gaussian distribution.
2. **Expected Square Jump Distance.**
   A rescaled expected squared jump distance converges as $d \to \infty$ to a related limit,

$$\lim_{d\to\infty} \frac{\left(s^{(d)}\right)^2}{\mathit{Trace}\left[\mathbf{T}^{(d)}\right]} \times \mathbb{E}\left\|\mathbf{X}_{k+1}^{(d)} - \mathbf{X}_k^{(d)}\right\|^2_{\mathbf{T}^{(d)}} = \ell^2 \times \alpha(\ell) =: J(\ell). \tag{2.11}$$

Theorem 1 is proved in Section 5.1. It establishes limiting values for the acceptance probability and expected squared jump distance, and more importantly for the relationship between them, which is crucial to establishing optimality results as we shall see. Further, (2.11) shows that, as is common in scaling problems for MCMC algorithms (for example in [RGG97, RR98]), the ESJD decomposes into the product of the acceptance probability $\alpha(\ell)$ and the expected squared *proposed* jumping distance ℓ^2, implying an asymptotic independence between the size of the proposed move and the acceptance event. As in the RWM case, we wish to be able to consider $J(\ell)$ to be a function of the asymptotic acceptance rate $\alpha(\ell)$. Our next result, which is proved in Section 5.2, shows that this is indeed possible.

PROPOSITION 1. *For a PsMRWM algorithm with noise difference B as in (2.2), with jump size determined by $\ell > 0$ as in (2.8), and with limiting asymptotic acceptance rate $\alpha(\ell)$ as in (2.10), the mapping $\ell \mapsto \alpha(\ell)$ is a continuous decreasing bijection from $(0, +\infty)$ to $(0, \alpha_{max}]$, where*

$$\alpha_{max} := \lim_{\ell\to 0} \alpha(\ell) = 2 \times \mathbb{P}[B > 0].$$

Proposition 1 yields that $\alpha_{\max} = \sup_{\ell>0} \alpha(\ell)$. When there is no noise in the estimate of the target, as already proved in [RGG97], the acceptance rate simplifies to $\alpha_0(\ell) := 2\,\Phi(-\ell/2)$ and the associated expected squared jump distance reads $J_0(\ell) = \ell^2\,\alpha_0(\ell)$. Thus we may also consider the asymptotic efficiency of a pseudo-marginal algorithm relative to the idealised algorithm if the target were known precisely by defining $J_{rel}(\ell) = J(\ell)/J_0(\ell)$, which also reads

$$J_{\text{rel}}(\ell) = \frac{1}{\Phi\left(-\ell/2\right)} \mathbb{E}\left[\Phi\left(\frac{B}{\ell} - \frac{\ell}{2}\right)\right]. \tag{2.12}$$

The following proposition, which is proved in Section 5.3, shows that the relative efficiency can never exceed unity and that it is bounded below by the acceptance rate in the limit as $\ell \to 0$.

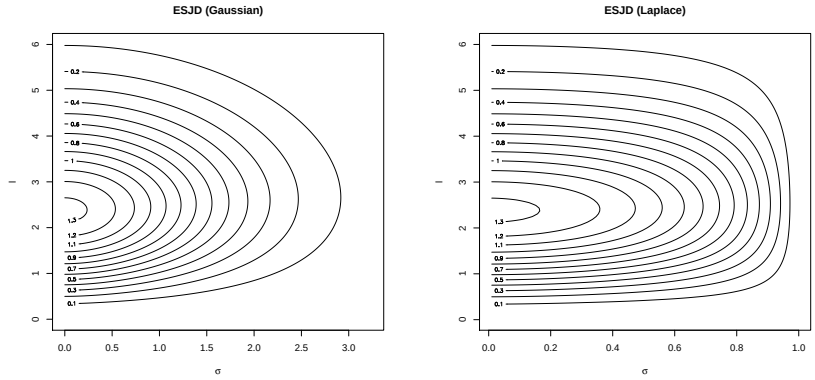

FIG 1. *Contour plots of the asymptotic expected squared jump distance $J(\ell)$ from (2.11) plotted as a function of the scaling parameter ℓ and of the standard deviation, σ, of the additive noise. In the left-hand panel the additive noise in the log-target is assumed to be Gaussian, and in right-hand panel it is assumed to have a Laplace distribution.*

PROPOSITION 2. *With $\alpha(\ell)$ and $J_{rel}(\ell)$ as defined in (2.10) and (2.12) respectively,*

$$\alpha_{max} \leq J_{rel}(\ell) \leq 1.$$

The quantities $\alpha(\ell)$, $J(\ell)$ and $J_{\text{rel}}(\ell)$ depend upon the distribution of B, and hence on the distribution of the additive noise W from (2.2). Figure 1 considers two particular cases: where the distribution of the additive noise is Gaussian, i.e. $W^* \sim \mathbf{N}(-\sigma^2/2, \sigma^2)$ (which we shall consider further in Section 3), and where the distribution of the additive noise is Laplace (i.e., double-exponential), with mean $\log(1-\sigma^2/2)$ and scale parameter $\sigma/\sqrt{2}$. For each of these two cases, it shows a *contour plot* of $J(\ell)$ as a function of the proposal scaling parameter ℓ and of the standard deviation of the additive noise, σ. Figure 2 shows the equivalent plots for $J_{\text{rel}}(\ell)$.

Our ultimate goal is often to choose ℓ to *maximise* $J(\ell)$, and thus obtain an *optimal* limiting diffusion (and hence an approximately optimal algorithm for finite d too). We shall use Theorem 1 to establish an optimal acceptance rate in a particular limiting regime, in Section 3.2 below.

Figure 2 illustrates that, except for small values of the scaling, the relative efficiency for a given noise distribution is relatively insensitive to the scaling. Related to this, from Figure 1 it appears that the optimal scaling (i.e., the value ℓ which maximises $J(\ell)$) is relatively insensitive to the variance of the additive noise. When there is no noise, the optimum is $\hat{\ell}_0 \approx 2.38$ as first noted in [RGG97], however the optimum remains close to 2.5 across a range of variances for both choices of noise distribution.

For these two examples, as might be expected, for any given scaling of the random walk proposal, the efficiency relative to the idealised algorithm decreases as the standard deviation of the noise increases, a phenomenon that is investigated more generally in [AV14]. Thus there is an implicit *cost* of having to estimate the target density. As a result of this, we should not expect the optimal acceptance probability for RWM of 0.234 to hold here.

2.5. *Diffusion limit.* We next prove that PsMRWM in high dimensions can be well-approximated by an appropriate diffusion limit (obtained as $d \to \infty$). This provides further justification for measuring efficiency by the ESJD, as discussed in detail in [RR13]. Briefly, the limiting ESJD (suitably scaled) is equal to the square of the limiting process's diffusion coefficient, h say. By a simple time

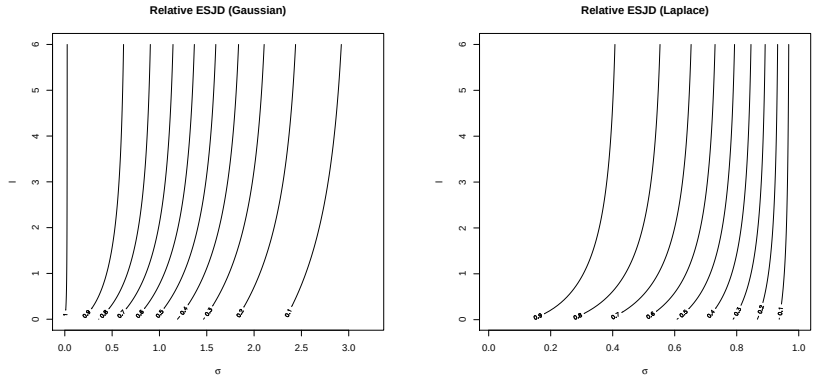

FIG 2. *Contour plots of $J_{rel}(\ell)$ from (2.12), the asymptotic expected squared jump distance relative to the idealised algorithm, plotted as a function of the scaling parameter ℓ and of the standard deviation, σ, of the additive noise. In the left-hand panel the additive noise in the log-target is assumed to be Gaussian, and in the right-hand panel it is assumed to have a Laplace distribution.*

change argument, the asymptotic variance of *any* Monte Carlo estimate of interest is inversely proportional to h. Minimising variance is thus equivalent to maximising h, i.e. h becomes (at least in the limit) unambiguously the right quantity to optimise. By constrast, MCMC algorithms which have non-diffusion limits can behave in very different ways, and ESJD may not be an appropriate way to compare algorithms in such cases.

We shall consider in this section the PsMRWM algorithm applied to a sequence of simple i.i.d target densities

$$\pi^{(d)}(x_1, \ldots, x_d) = \prod_{i=1}^{d} f(x_i)$$

where f is a one dimensional probability density. We assume throughout this section that the following regularity assumptions hold.

ASSUMPTIONS 3. *The first four moments of the distribution with density f are finite. The log-likelihood mapping $x \mapsto \log f(x)$ is smooth with second, third and fourth derivatives globally bounded.*

One can verify that under Assumptions 3, the target $\pi^{(d)}$ satisfies Assumptions 2. It is important to stress that the ESJD analysis of Section 2.4 only relies on the weaker Assumptions 2 and, as discussed at the end of the previous section, is valid for much more general target distributions than the ones with i.i.d. coordinates considered in this section. The stronger Assumptions 3 are standard in the diffusion-limit literature and are, perhaps, the simplest from which a diffusion limit is expected to result [RGG97]. However, these i.i.d. assumptions have been relaxed in various directions [BPS+04, Béd07, BR08, BRS09, PST12] and we believe that our diffusion limit Theorem 2 could also be extended to similar settings at the cost of considerably less transparent proofs.

In the remainder of this article we consider the sequences of scaling functions $\sqrt{s_L^{(d)}} = s_g^{(d)} := \sqrt{I \times d}$, with

$$I := \mathbb{E}\big[\big\{\big(\log f(X)\big)'\big\}^2\big] = -\mathbb{E}\big[\big(\log f(X)\big)''\big] \tag{2.13}$$

and $X \overset{\mathcal{D}}{\sim} f(x)\,dx$. Indeed, Equation (2.4) is satisfied; consequently, for a tuning parameter $\ell > 0$, we consider d-dimensional RWM proposals with scaling

$$\lambda^{(d)} := \ell\, I^{-1/2}\, \delta^{1/2} \qquad \text{with} \qquad \delta = 1/d \tag{2.14}$$

as in (2.8). The quantity I, which quantifies the roughness and the scale of the marginal density $f(x)\,dx$, has been introduced in the definition of the RWM jump-size (2.14) so that all our limiting results on the *optimal* choice of parameter ℓ are independent of $f(x)\,dx$. The main result of this section is a diffusion limit for a rescaled version $V^{(d)}$ of the first coordinate process. For time $t \geq 0$ we define the piecewise-constant continuous-time process

$$V^{(d)}(t) := X^{(d)}_{\lfloor dt \rfloor, 1}$$

with the notation $\mathbf{X}^{(d)}_k = (X^{(d)}_{k,1}, \ldots, X^{(d)}_{k,d}) \in \mathbb{R}^d$ so that $V^{(d)}(t)$ is the first coordinate of $\mathbf{X}^{(d)}_{\lfloor dt \rfloor}$. Note that in general the process $V^{(d)}$ is not Markovian. The next theorem shows that nevertheless, in the limit $d \to \infty$, the process $V^{(d)}$ converges weakly to an explicit Langevin diffusion. This result thus generalises the original RWM diffusion limit proved in [RGG97].

THEOREM 2. *Let $T > 0$ be a finite time horizon. For all $d \geq 1$ let each Markov chain and the additive noise satisfy Assumptions 1, let the sequence of product form densities $\pi^{(d)}$ satisfy the regularity Assumptions 3 and set the scale of the jump proposals as in Equation (2.14). Then, as $d \to \infty$,*

$$V^{(d)} \Rightarrow V$$

in the Skorokhod topology on $D([0,T])$, where V satisfies the Langevin SDE

$$dV_t = h^{1/2}(\ell)\, dB_t + \frac{1}{2}\, h(\ell)\, \nabla \log f(V_t)\, dt \tag{2.15}$$

with initial distribution $V_0 \overset{\mathcal{D}}{\sim} f$ and B_t a standard Brownian motion. The speed function h is proportional to the asymptotic rescaled ESJD function J,

$$h(\ell) \;=\; J(\ell)/I,$$

with the constant of proportionality I defined by Equation (2.13).

The time change argument discussed before Theorem 2 shows that the quantity J_{rel} exactly measure the loss of mixing efficiency (computational time not taken into consideration) when exact evaluations of the target density are replaced by unbiased estimates; as already mentioned, the pseudo-marginal algorithm always has worse mixing properties than the idealised algorithm.

3. Optimising the PsMRWM. We next consider the question of optimising the PsMRWM. Now, when examining the efficiency of a standard RWM, the expected computation (CPU) time is usually not taken into account since it is implicitly assumed to be independent of the choice of tuning parameter(s). This may indeed be approximately true for the RWM. However, for the PsMRWM the expected CPU time for a single iteration of the algorithm is usually approximately inversely proportional to the variance of the estimator $\hat{\pi}(x)$. For this reason, we measure the efficiency of the PsMRWM through a rescaled version of the ESJD:

$$\text{(Efficiency)} := \frac{\text{(Expected Square Jump Distance)}}{\text{(Expected one-step computing time)}} \tag{3.1}$$

Of course, for any increasing function F, the quantity F(ESJD)/(Expected one-step computing time) is a possible measure of efficiency. However, the discussion at the start of Section 2.5 indicates that (3.1) is the appropriate measure of efficiency in the high-dimensional asymptotic regime considered in this article.

In the remainder of this section, we implicitly assume that the target distributions satisfy Assumptions 2.

3.1. *Standard (Gaussian) regime.* We shall restrict attention to the case in which the additive noise follows a Gaussian distribution. More precisely, we shall assume the following, which we shall refer to for brevity as 'the standard asymptotic regime' (SAR):

ASSUMPTIONS 4. *For each $x \in \mathcal{X}$ and $\sigma^2 > 0$, we have an unbiased estimator $\hat{\pi}(x)$ of $\pi(x)$, such that $\log \hat{\pi}(x)$ follows a Gaussian distribution with variance σ^2. Furthermore, the expected one-step computing time is inversely proportional to σ^2.*

Intuitively, Assumptions 4 are designed to model the situation where $\pi(x)$ is estimated as a product of n averages of m i.i.d. samples in the limit as $n \to \infty$ and with $m \propto n$. For a fixed large n, approximate normality follows from the central limit theorem; moreover $\sigma^2 \approx c/m$ for some $c > 0$, and the computational time is proportional to m and hence to $1/\sigma^2$. Assumptions 4 have recently been shown to hold more generally, in the context of particle filtering for a hidden Markov model (see [BDMD13]). There are other natural situations where multiplicative forms for the importance sampling estimator of the likelihood might make the estimator well-approximated as a log-Gaussian, for example in correcting for a PAC likelihood approximation (see [LM03]).

Under the SAR of Assumptions 4, we will prove an optimality result in Section 3.2 which specifies a particular optimal variance for the estimate of the log-target.

3.2. *Optimisation under the standard asymptotic regime.* In this section we consider a sequence $\pi^{(d)}$ of target distributions satisfying Assumptions 2 and assume that each unbiased estimator satisfies the independence in Assumptions 1. Under these assumptions, the rescaled ESJD of the PsMRWM algorithm with jump size (2.8) is described by Theorem 1. Under the SAR, i.e. Assumptions 4, and with $\mathrm{Var}\left[\log \hat{\pi}(x)\right] = \sigma^2$, the noise difference is $B \overset{\mathcal{D}}{\sim} \mathbf{N}(-\sigma^2, 2\sigma^2)$. Since the mean one step computing time is assumed to be inversely proportional to the variance, σ^2, the asymptotic efficiency, as $d \to \infty$, is proportional to

$$\sigma^2 \times J_{\sigma^2}(\ell) \;=:\; \mathbf{Eff}_{\sigma^2}(\ell) \tag{3.2}$$

where $J_{\sigma^2}(\ell)$ stands for the asymptotic rescaled ESJD identified in Theorem 1, i.e. $J(\ell)$, in the special case where $B \overset{\mathcal{D}}{\sim} \mathbf{N}(-\sigma^2, 2\sigma^2)$.

Figure 3 provides a contour plot of this efficiency $\mathbf{Eff}_{\sigma^2}(\ell)$, relative to the highest achievable efficiency, and of the logarithm of the asymptotic acceptance rate $\alpha(\ell)$, both as functions of the scaling parameter ℓ and of the standard deviation, σ. It also provides a plot of the profile $\mathbf{Eff}_{\sigma^2_{opt}(\ell)}(\ell)$ as a function of ℓ, again relative to the highest achievable value.

As previously suggested by Figure 1, we see that the conditional optimal value of ℓ is relatively insensitive to the value of σ.

The point at which the maximal efficiency is achieved is detailed precisely in Corollary 1 below.

COROLLARY 1. *The efficiency $\boldsymbol{Eff}_{\sigma^2}(\ell)$ is maximised (to three decimal places) when the variance σ^2 of the log-noise is*

$$\sigma^2_{\mathrm{opt}} \;=\; 3.283$$

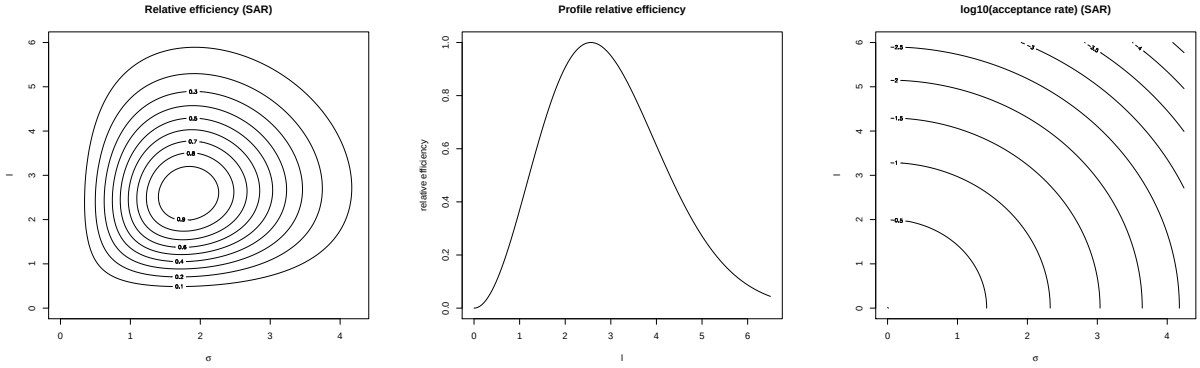

FIG 3. *Contour plots of the theoretical relative efficiency* $\boldsymbol{Eff}_{\sigma^2}(\ell)/\boldsymbol{Eff}_{\sigma^2_{opt}}(l_{opt})$*, and of the base-10 logarithm of the asymptotic acceptance probability* $\alpha(\ell)$*, and a plot of the profile relative efficiency* $\boldsymbol{Eff}_{\sigma^2_{opt}(\ell)}(\ell)/\boldsymbol{Eff}_{\sigma^2_{opt}}(l_{opt})$*, all for the scenario where the additive noise arises from the SAR.*

and the scaling parameter ℓ *is*

$$\ell_{\text{opt}} \;=\; 2.562\,,$$

at which point the corresponding asymptotic acceptance rate is

$$\alpha_{\text{opt}} = 7.001\,\%.$$

As $\sigma^2 \to \infty$ *the optimal scaling satisfies* $\ell_{opt}(\sigma) \to 2\sqrt{2}$ *and as* $\ell \to \infty$ *the optimal variance satisfies* $\sigma^2_{opt}(\ell) \to 4$*.*

PROOF. For convenience, write $\tau^2 := 2\sigma^2$ and introduce three independent standard Gaussian random variables $U, V, Z \overset{\mathcal{D}}{\sim} \mathbf{N}(0,1)$. Notice that $B \overset{\mathcal{D}}{\sim} -\tau^2/2 + \tau\, U$ and

$$\begin{aligned}\mathbf{Eff}_{\sigma^2}(\ell) &= \tau^2\,\ell^2\,\mathbb{E}\big[\Phi(B/\ell - \ell/2)\big] = \tau^2\,\ell^2\,\mathbb{P}\big[V < \big(-\tau^2/2 + \tau\, U\big)/\ell - \ell/2\big] \\ &= \tau^2\,\ell^2\,\mathbb{P}(\ell\, V - \tau\, U < -\big(\tau^2 + \ell^2\big)/2\big] = \tau^2\,\ell^2\,\mathbb{P}\big[\sqrt{\ell^2+\tau^2}\, Z < -\big(\tau^2+\ell^2\big)/2\big] \\ &= \tau^2\,\ell^2\,\Phi\left(-\frac{1}{2}\sqrt{\tau^2+\ell^2}\right).\end{aligned} \tag{3.3}$$

For fixed $\tau^2 + \ell^2$, the quantity $\tau^2\ell^2$ is maximised when $\tau^2 = \ell^2$, at which point the efficiency is $\tau^4\Phi(-\tau/\sqrt{2}) \propto \sigma^4\Phi(-\sigma)$. This is maximised numerically when $\sigma^2 = \sigma^2_{\text{opt}} = 3.283$ (to three decimal places), and at this point $\ell_{\text{opt}} = \sigma_{\text{opt}}\sqrt{2}$ and $\alpha_{\text{opt}} = 2\Phi(-\sigma_{\text{opt}})$ with the corresponding numerical values as stated.

Differentiating (3.3) with respect to ℓ we find that the optimal scaling satisfies

$$\Phi\left(-\frac{1}{2}\sqrt{\ell^2+\tau^2}\right) = \frac{1}{4}\ell^2\varphi\left(-\frac{1}{2}\sqrt{\ell^2+\tau^2}\right)\Big/\sqrt{\ell^2+\tau^2}$$

The result for large τ^2 follows from the relationship $\Phi(-x) \sim \varphi(x)/x$ as $x \to \infty$. The symmetry of the function $(\ell^2, \tau^2) \mapsto \mathbf{Eff}_{\sigma^2}(\ell)$ in τ and ℓ then provides the result for large ℓ. □

Remarks

1. This leads to a new optimal scaling for standard Gaussian targets of $\lambda \approx \ell_{\text{opt}}/\sqrt{d}$ with $\ell_{\text{opt}} \approx 2.562$, and contrasts with the corresponding formula $\hat{\ell}_0/\sqrt{d}$, with $\hat{\ell}_0 \approx 2.38$, for the usual random walk Metropolis algorithm [RGG97]; recall that $\hat{\ell}_0$ satisfies $\hat{\ell}_0 = \text{argmin}_{\ell>0}\, \ell^2\, \Phi(-\ell/2)$.
2. In the discussion of Figure 1 it was noted that for a Gaussian or Laplace noise regime the optimal scaling at a particular noise variance, σ^2, is insensitive to the value of σ^2. From Figure 3 and from the symmetry of expression (3.3), the optimal variance at a particular scaling ℓ is also insensitive to the value of ℓ. Moreover as $\ell \to 0$ the optimal variance is $\hat{\ell}_0^2/2 \approx 2.83$, which corresponds (at least to 2dp) with the value obtained in [DPDK14].
3. In practice, σ^2 might be a function of a *discrete* number m of samples or particles and, hence, only take a discrete set of values. In particular, if the variance in the noise using $m = 1$ is already lower than 3.283, then there can be little gain in increasing m.
4. In many problems the computational cost of obtaining an unbiased estimate of the target is much larger than the cost of the remainder of the algorithm, but this is not always the case. Consider therefore the more general problem where the cost of obtaining a single unbiased estimate is t_{rat} times the cost of the remainder of the algorithm. In this case the efficiency functional should be expressed as (Efficiency) $= J_{\sigma^2}(\ell)/(1 + t_{rat}\sigma^{-2})$ and the optimal acceptance rate is a function of t_{rat} which varies between between 7.0% (as $t_{rat} \to \infty$) and 23.4% (as $t_{rat} \to 0$).

Figure 3 shows that in contrast to the insensitivity of the optimal scaling to the variance of the noise, the acceptance rate at this optimum could potentially vary by a factor of 3 or more. Thus if a particular scaling of the jump proposals maximises $J(\ell)$ for some particular noise distribution and variance, then that scaling should be close to optimal across a wide range of noise distributions and variances. However, tuning to a particular acceptance rate, whilst more straightforward in practice, could lead to a sub-optimal scaling if the noise distributions encountered in the tuning runs are not entirely representative of the distributions that will be encountered during the main run.

Our theory applies in the limit when the dimension d of the (marginal) target $\mathbf{X}$ goes to infinity. However, using a similar argument to that in [SR09], when $\mathbf{X} \sim \mathbf{N}(0, \mathbf{I}_d)$ it can be shown that under the SAR with the proposal as in (2.1) the ESJD and acceptance rate are:

$$\text{ESJD}(\lambda, d) = 2\lambda^2\, \mathbb{E}\left[\|\mathbf{Z}\|^2 \Phi\left(-\frac{\lambda}{2}\|\mathbf{Z}\| + \frac{B}{\lambda\|\mathbf{Z}\|}\right)\right] \quad \text{and} \quad \alpha(\lambda, d) = \mathbb{E}\left[\Phi\left(-\frac{\lambda}{2}\|\mathbf{Z}\| + \frac{B}{\lambda\|\mathbf{Z}\|}\right)\right],$$

where $\mathbf{Z} \overset{\mathcal{D}}{\sim} \mathbf{N}(0, \mathbf{I}_d)$ and $B \overset{\mathcal{D}}{\sim} \mathbf{N}(-\sigma^2, 2\sigma^2)$. Numerical optimisation of the efficiency function, $\sigma^2 \times \text{ESJD}(\lambda, d)$ for $d = 1, 2, 3, 5$ and 10 produces a steady decrease in $\hat{\ell} = \hat{\lambda}\sqrt{d}$ from 2.59 to 2.57 and in $\hat{\alpha}$ from 11.5% to 7.7%, and a similarly steady increase in $\hat{\sigma}^2$ from 3.23 to 3.27. Thus, at least for Gaussian targets and with efficiency measured by ESJD, the asymptotic results for the optimal scaling and optimal variance are applicable in any dimension but there may be a small increase in the optimal acceptance rate, as is found for the non-pseudo-marginal RWM (e.g. [RR01, SR09]).

In the simulation study of Section 4, below, we find that Corollary 1 and its associated formulae provide a good description of the optimal settings for a particle filter with $T = 50$ and $d = 5$.

4. Simulation Study. In this section we restrict attention to the SAR of Section 3.1. Corollary 1 suggests that the optimal efficiency should be obtained by choosing the number of unbiased estimates, m, such that the variance in the log-target is approximately 3.3. The scale parameter, λ, should be set so that the acceptance rate is approximately 7%. Since the constant of proportionality relating λ and ℓ is unknown in practice, we cannot simply set $\ell \approx 2.56$.

In practice the assumptions underlying this result may not hold: the dimension of the parameter space is finite, the distribution of the noise, W^*, may not be Gaussian, and it is likely to also vary

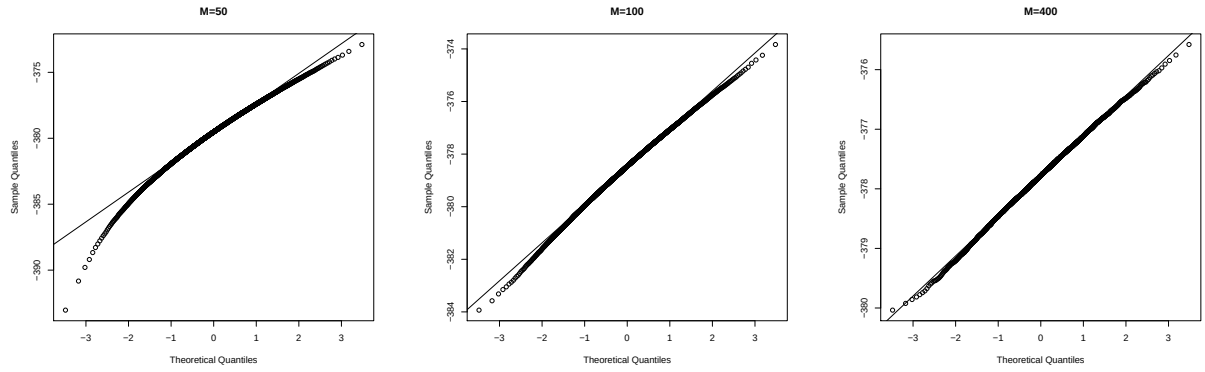

FIG 4. *Normal QQplots of the noise in the estimate of the log-target at the a proposed value of the posterior median,* $\hat{\mathbf{x}}$*, when* $m = 50$ *(left panel),* $m = 100$ *(centre) and* $m = 400$ *(right).*

with position, $\mathbf{x}^*$. We conduct a simulation study to provide an indication of both the extent of and the effect of such deviations.

We use the Particle Marginal RWM algorithm (PMRWM) of [ADH10] to perform exact inference for the Lotka-Volterra predator-prey model; see [GW11] for a more detailed description of the PMRWM which focusses on this particular class of applications. Starting from an initial value, which is, for simplicity, assumed known, the two-dimensional latent variable $\mathbf{U}$ evolves according to a Markov Jump Process (MJP). Each component is observed at regular intervals with Gaussian error of an unknown variance. Appendix B provides details of the observation regime and of the transitions of the MJP and their associated rates. It also provides the parameter values, the priors and the lengths of the MCMC runs.

An initial run provided an estimate of a central value, $\hat{\mathbf{x}}$ (the vector of posterior medians), and the posterior variance matrix, $\widehat{\mathrm{Var}}(\mathbf{X})$. Since the shape of the target distribution, and hence the optimal shape of the proposal, is unknown we follow the frequently used strategy for the RWM (e.g. [SFR10]) of setting the proposal covariance matrix to be proportional to $\widehat{\mathrm{Var}}(\mathbf{X})$. ¿From Remark 1 following Corollary 1, we set $\mathbf{V}_{\mathrm{prop}} = \gamma^2 \times (2.56^2/d) \times \widehat{\mathrm{Var}}(\mathbf{X})$ with $\gamma = 1$ corresponding to an optimal tuning for a Gaussian target.

Let $\mathbb{M} := \{50, 80, 100, 150, 200, 300, 400\}$ define the set of choices for the number of particles, m, and let $\mathbb{G} := \{0.4, 0.6, 0.8, 1.0, 1.2, 1.4, 1.6\}$ define the set of choices for the relative scaling, γ. For each (m, γ) in $\mathbb{M} \times \mathbb{G}$ an MCMC run of at least 2.5×10^5 iterations was performed starting from $\hat{\mathbf{x}}$. For diagnostic purposes runs of at least 10^4 iterations were performed with $m \in \mathbb{M}$ and $\gamma = 0$ (so $\mathbf{x} = \hat{\mathbf{x}}$ throughout).

We perform three checks on our assumptions. The diagnostic runs provide a sample from the distribution of W^*, the estimate of the log-target at a proposed value; this allows us to investigate the second part of Assumptions 1 and both parts of Assumptions 4. We first examine the SAR Assumptions 4. Figure 4 shows QQplots for $m = 50$, $m = 100$ and $m = 400$ against a Gaussian distribution; it is clear that at $m = 50$ the right-hand tail is slightly too light and the left-hand tail is much heavier than that of a Gaussian. Similar but much smaller discrepancies are present at $m = 100$, whilst at $m = 400$ the noise distribution is almost indistinguishable from that of a Gaussian. The left hand panel in Figure 5 plots $\log \mathrm{Var}[W^*]$ against $\log m$ and includes a line with the theoretical slope of -1 and passing through an additional point at $m = 1600$. The heavy left-hand tail at $m = 50$ leads to a considerably higher variance than that which would arise under

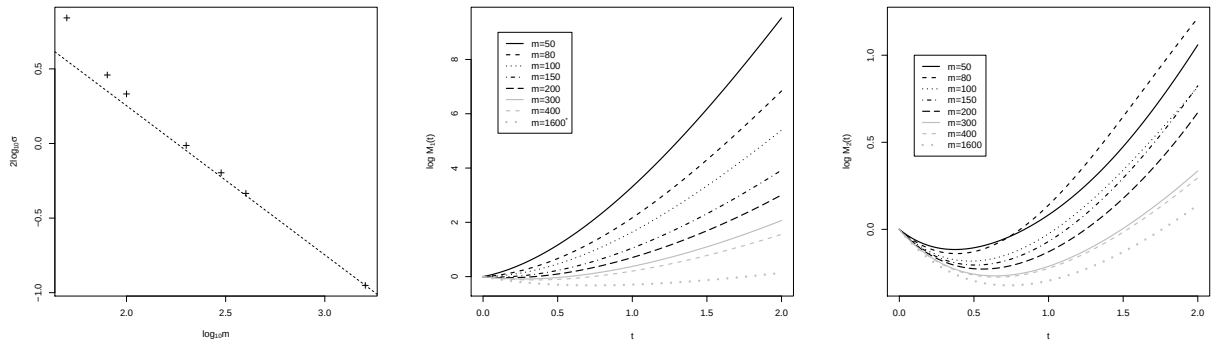

FIG 5. *In the left panel the logarithm of the empirical variance of the noise in the estimate of the log-proposal sampled at* $\mathbf{x} = \hat{\mathbf{x}}$ *is plotted against the logarithm of the number of particles used; the centre and right panels are plots of the logarithms of the empirical estimates of the moment generating functions of* $\hat{L}$ *and* L *(*$M_1(t)$ *and* $M_2(t)$*, respectively) against* t*. The additional lowest curve in the centre panel* * *and in the right-hand panel is the logarithm of* $M_2(t)$ *with* $m = 1600$*, and constitutes our best estimate of 'truth'.*

the SAR, however even by $m = 80$ the fit is reasonably close.

We assess the degree of dependence of the distribution of W^* on the position $\mathbf{x}$ by considering the joint distribution of W^* and $L := (\log \pi)(\mathbf{X})$, the true log-target evaluated at $\mathbf{X}$, where $\mathbf{X}$ is distributed according to the target. For a particular m, all of the runs with $\gamma > 0$ provide a combined sample of size n_1 from the distribution of the estimate of the log-target at the current value, $\hat{L} = L + W$, whereas (after scaling so that $\frac{1}{n_2}\sum_{i=1}^{n_2} \exp w^{*(i)} = 1$) each run with $\gamma = 0$ provides a sample of size n_2 from the distribution of W^* at $\mathbf{x} = \hat{\mathbf{x}}$. Equation (2.3) shows that subject to Assumptions 1, W and L are independent and that the density of W is an exponentially tilted version of the density of W^*. These two properties lead directly to the following.

PROPOSITION 3. *If Assumptions 1 hold, the identity*

$$\mathbb{E}\big[\exp(t\hat{L})\big]/\mathbb{E}\big[\exp\big\{(t+1)W^*\big\}\big] \;=\; \mathbb{E}\big[\exp(tL)\big] \tag{4.1}$$

holds for any $t \in \mathbb{R}$ *such that all the above three expectations are well defined.*

The right hand side of (4.1) is independent of the noise distribution, or equivalently of the number of particles, m. Moreover, if the noise is small enough then the ratio on the left hand side should provide a good estimator of the true moment generating function (MGF) of L even if there is dependence (since the impact of any dependence will be small).

In our scenario, realisations of L are typically between -385 and -375 with a mode at approximately -379, so the MGFs of L and $\hat{L}$ are dominated by the term e^{-379t}, whatever the noise distribution. To be able to discern any differences we therefore consider for each value of m, shifted estimators of the MGFs of $\hat{L}$ and of L.

$$M_1(t) := \frac{1}{n_1}\sum_{i=1}^{n_1} \exp\Big[t(\hat{L}^{(i)} + 379)\Big] \quad \text{and} \quad M_2(t) := M_1(t)\left(\frac{1}{n_2}\sum_{i=1}^{n} \exp\Big[(t+1)W^{*(i)}\Big]\right)^{-1}.$$

The central panel of Figure 5 shows $M_1(t)$ with a separate curve for each value of m; the lowest curve is our best estimate of the true MGF of L ($M_2(t)$ from $m = 1600$). The right hand panel

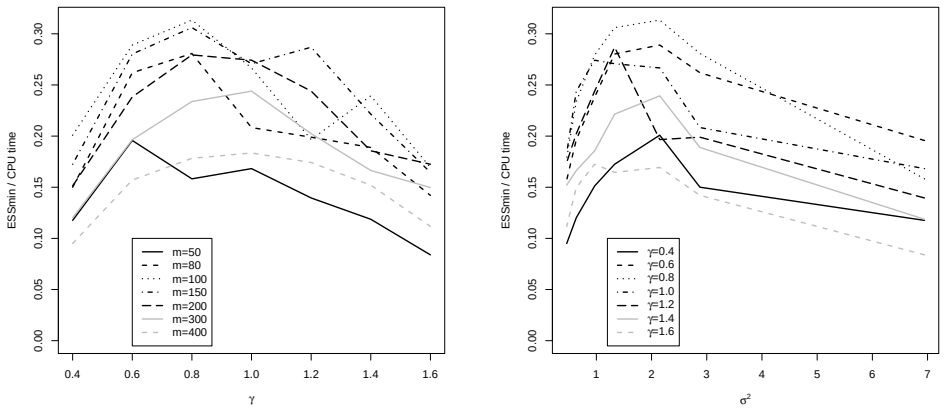

FIG 6. *Empirical efficiency, $\widehat{\text{eff}}$, measured in terms of minimum effective sample size per CPU second, plotted against (left panel) γ for different values of m and (right panel) σ^2 (estimated from the sample of W^* at the posterior median, $\hat{\mathbf{x}}$) for different values of γ.*

shows $M_2(t)$ for each value of m. Clearly the curves in the right hand panel do not coincide and so the assumption of independence does not hold precisely. However it is clear from the very different vertical scales of the two figures that *most* of the difference between the distribution of $\hat{L}$ for any given m and the distribution of L *can* be explained by Assumptions 1.

We now consider an empirical measure of efficiency $\widehat{\text{eff}}$, the quotient of the minimum (over the parameters) effective sample size and the CPU time. The left-hand panel of Figure 6 shows $\widehat{\text{eff}}$ plotted against γ for different values of m, whilst the right-hand panel shows $\widehat{\text{eff}}$ plotted against m for different values of γ. The optimal (over $\mathbb{G}$) value for γ is either 0.8 or 1.0 whatever the value of m, which is consistent with the expected insensitivity of the optimal scaling and suggests that the target is at least approximately Gaussian. The optimal (over $\mathbb{M}$) value for m is either $m = 200$, $m = 150$ or $m = 100$, corresponding to an optimal σ^2 (estimated from the sample for W^*) of either 1.0, 1.3 or 2.1, again (as far as can be discerned) showing no strong sensitivity to γ. Finally the overall optimum occurs at $\sigma^2 = 2.1$ and $\gamma = 0.8$ with an acceptance rate of 15.39%. The optimal σ^2 is slightly lower than the theoretically optimal value of 3.3. Further theoretical investigations (using numerical integration) for a true 5-dimensional Gaussian target corrupted by noise subject to the SAR show that ESJD per second is still optimised at $\sigma^2 \approx 3.3$; however empirical investigations show that the ESS/sec for this target is optimised at a value of $\sigma^2 \approx 2$. The discrepancy between the theory and our simulation study is therefore likely to be attributable to this discrepancy between ESS and ESJD in low dimensional settings. The relatively high acceptance rate is a consequence of this lower variance and fits with our theory since from (3.3) the acceptance rate should be $2\,\Phi(-\frac{1}{2}\sqrt{2\sigma^2 + \gamma^2 \times 2.56^2}) = 14.7\%$.

5. Proofs of results. Equation (2.3) yields that $B = W^* - W$ has density ρ satisfying

$$\rho(b) := \int_{w \in \mathbf{R}} g^*(w) g^*(w+b) e^w \, dw = \int_{w^* \in \mathbf{R}} g^*(w^* - b) g^*(w^*) e^{w^* - b} \, dw^* = e^{-b} \rho(-b).$$

Thus

$$\rho(b) = e^{-b/2} \, h(b), \text{where } h \text{ is a symmetric function, } h(b) = h(-b). \tag{5.1}$$

This fact will be used in the proofs of Theorem 1 and Proposition 1.

5.1. *Proof of Theorem 1.* For notational convenience, we drop the index $[\cdot]^{(d)}$ when the context is clear. As in Section 2.3, the Hessian matrix of the log-likelihood $L(\mathbf{x}) := \log \pi^{(d)}(\mathbf{x})$ at $\mathbf{x} \in \mathbf{R}^d$ is denoted by $H(\mathbf{x}) = \left[\partial^2_{ij} L(\mathbf{x})\right]_{1 \le i,j \le d}$.

- **Proof of Equation (2.10).**
  The mean acceptance probability equals

$$\alpha^{(d)}(\ell) := \mathbb{E}\Big[1 \wedge \exp\Big(L(\mathbf{X} + \lambda^{(d)}\mathbf{Z}) - L(\mathbf{X}) + B\Big)\Big] = \mathbb{E}\Big[F\Big(L(\mathbf{X} + \lambda^{(d)}\mathbf{Z}) - L(\mathbf{X}) + B\Big)\Big]$$

  with $\mathbf{X} \overset{\mathcal{D}}{\sim} \pi^{(d)}$, jump scale $\lambda^{(d)} := \ell / s^{(d)}$, random variable $\mathbf{Z} \overset{\mathcal{D}}{\sim} \mathbf{N}(0, \mathbf{I}_d)$ independent from $\mathbf{X}$ and accept-reject function $F(u) := 1 \wedge \exp(u)$. Algebra shows that for any $b \in \mathbf{R}$ and $V \overset{\mathcal{D}}{\sim} \mathbf{N}(-\ell^2/2, \ell^2)$ we have $\mathbb{E}[1 \wedge \exp(V + b)] = \Phi(-\ell/2 + b/\ell) + e^b \Phi(-\ell/2 - b/\ell)$. By (5.1)

$$\begin{aligned} \mathbb{E}\left[1 \wedge \exp(V + B)\right] &= \int_{-\infty}^{\infty} h(b) \left(e^{-b/2}\Phi(-\ell/2 + b/\ell) + e^{b/2}\Phi(-\ell/2 - b/\ell)\right) db \\ &= 2\int_{-\infty}^{\infty} h(b)\; e^{-b/2}\Phi\left(-\ell/2 + b/\ell\right)\; db = 2\,\mathbb{E}\left[\Phi\left(-\ell/2 + B/\ell\right)\right]. \end{aligned}$$

  Since F is continuous and bounded, in order to prove Equation (2.10) it, therefore, suffices to show that $L(\mathbf{X} + \lambda^{(d)}\mathbf{Z}) - L(\mathbf{X})$ converges in law to a Gaussian distribution with mean $-\ell^2/2$ and variance ℓ^2. A second order expansion yields

$$L(\mathbf{X} + \lambda^{(d)}\mathbf{Z}) - L(\mathbf{X}) = \lambda^{(d)}\langle \nabla L(\mathbf{X}), \mathbf{Z}\rangle + \frac{1}{2}\big(\lambda^{(d)}\big)^2 \langle \mathbf{Z}, H(\mathbf{X})\,\mathbf{Z}\rangle + R(\mathbf{X}, \mathbf{Z}, \lambda^{(d)})$$

  with remainder $R(\mathbf{X}, \mathbf{Z}, \lambda^{(d)}) := \big(\lambda^{(d)}\big)^2 \int_0^1 (1-t)\,\langle \mathbf{Z},\, [H(\mathbf{X} + t\lambda^{(d)}\mathbf{Z}) - H(\mathbf{X})]\,\mathbf{Z}\rangle\, dt$. Slutsky's Lemma shows that to finish the proof of (2.10) it suffices to verify that $\lambda^{(d)}\langle \nabla L(\mathbf{X}), \mathbf{Z}\rangle$ converges in law to a centred Gaussian distribution with variance ℓ^2 and that

$$\lim_{d\to\infty} \frac{1}{2}\big(\lambda^{(d)}\big)^2 \langle \mathbf{Z}, H(\mathbf{X})\,\mathbf{Z}\rangle = -\ell^2/2 \qquad \text{and} \qquad \lim_{d\to\infty} R(\mathbf{X}, \mathbf{Z}, \lambda^{(d)}) = 0$$

  in probability.

  - Note that conditionally upon $\mathbf{X} = \mathbf{x} \in \mathbf{R}^d$ the quantity $\lambda^{(d)}\langle \nabla L(\mathbf{X}), \mathbf{Z}\rangle$ has a centred Gaussian distribution with variance $\ell^2 \left\|\nabla L(\mathbf{x})\right\|^2 / \big(s_G^{(d)}\big)^2$. Equation (2.4) shows that $\lambda^{(d)}\langle \nabla L(\mathbf{X}), Z\rangle$ converges in law to a Gaussian distribution with variance ℓ^2.
  - Conditionally upon $\mathbf{X} = \mathbf{x}$ the quantity $\big(\lambda^{(d)}\big)^2 \langle \mathbf{Z}, H(\mathbf{X})\,\mathbf{Z}\rangle$ has the same distribution as $\ell^2 \left(\sum_{i=1}^d \beta_i(\mathbf{x})\, Z_i^2\right) / s_L^{(d)}$ where $\big(\beta_1(\mathbf{x}), \ldots, \beta_d(\mathbf{x})\big)$ is the spectrum of the Hessian matrix $H(\mathbf{x})$. The conditional mean thus equals the re-scaled Laplacian $\ell^2\, \Delta L(\mathbf{x}) / s_L^{(d)}$ and the conditional variance is

$$2\,\ell^4 \sum_{i=1}^d \beta_i(\mathbf{x})^2 / \big(s_L^{(d)}\big)^2 = 2\,\ell^4\, \mathrm{Trace}\big[H^2(\mathbf{x})\big] / \big(s_L^{(d)}\big)^2.$$

    Markov inequality, Equations (2.4) and (2.6), and the hypothesis $s_L^{(d)} = \big(s_g^{(d)}\big)^2$ yield that $\frac{1}{2}\left(\lambda^{(d)}\right)^2 \langle \mathbf{Z}, H(\mathbf{X}), \mathbf{Z}\rangle$ converges in probability to $-\ell^2/2$.

– Equation (2.7) shows that the remainder $R(\mathbf{X}, \mathbf{Z}, \lambda^{(d)})$ converges to zero in probability.

- **Proof of Equation** (2.11).
The proof of Equation (2.11) follows from Equation (2.10). Note that we have

$$\frac{\big(s^{(d)}\big)^2}{\mathrm{Trace}\left[\mathbf{T}^{(d)}\right]} \times \mathbb{E}\big\|\mathbf{X}_{k+1}^{(d)} - \mathbf{X}_k^{(d)}\big\|_{\mathbf{T}^{(d)}}^2 := \ell^2\,\mathbb{E}\Big[\frac{\|\mathbf{Z}\|_{\mathbf{T}^{(d)}}^2}{\mathrm{Trace}\left[\mathbf{T}^{(d)}\right]} \times F\Big(L(\mathbf{X}+\lambda^{(d)}\mathbf{Z}) - L(\mathbf{X}) + B\Big)\Big].$$

Since $\lim_{d\to\infty}\mathbb{E}\Big[F\Big(L(\mathbf{X}+\lambda^{(d)}\mathbf{Z}) - L(\mathbf{X}) + B\Big)\Big] = \alpha(\ell)$, to prove Equation (2.11) it suffices to verify that

$$\mathbb{E}\left[\Big\{\frac{\|\mathbf{Z}\|_{\mathbf{T}^{(d)}}^2}{\mathrm{Trace}\left[\mathbf{T}^{(d)}\right]} - 1\Big\} \times F\Big(L(\mathbf{X}+\lambda^{(d)}\mathbf{Z}) - L(\mathbf{X}) + B\Big)\right]$$

converges to zero as $d \to \infty$. Since the function F is bounded, the conclusion follows once we have proved that $\mathbb{E}\left[\Big(\frac{\|\mathbf{Z}\|_{\mathbf{T}^{(d)}}^2}{\mathrm{Trace}\left[\mathbf{T}^{(d)}\right]} - 1\Big)^2\right]$ converges to zero. Diagonalisation of the symmetric matrix $\mathbf{T}^{(d)}$ in an orthonormal basis shows that this last quantity equals $2 \times \mathrm{Trace}\left[\big(\mathbf{T}^{(d)}\big)^2\right] / \mathrm{Trace}\left[\mathbf{T}^{(d)}\right]^2$ so that the conclusion directly follows from Equation (2.9).

5.2. *Proof of Proposition 1.* The dominated convergence theorem shows that $\ell \mapsto \alpha(\ell) = 2 \times \mathbb{E}\big[\Phi(B/\ell - \ell/2)\big]$ is continuous and converges to zero as ℓ tends to infinity. Since the limiting acceptance probability can also be expressed as $\alpha(\ell) = 2\,\mathbb{P}(\ell\,\xi + \ell^2/2 < B)$ for $\xi \overset{\mathcal{D}}{\sim} \mathbf{N}(0,1)$ independent from all other sources of randomness, it also follows that the limiting acceptance probability $\alpha(\ell)$ converges to $2\,\mathbb{P}(B > 0)$ as ℓ converges to zero. To finish the proof of Proposition 1, it remains to verify that the function $\ell \to \alpha(\ell)$ is strictly decreasing. To this end, we will establish that the derivative $\frac{d}{d\ell}\alpha(\ell)$ is strictly negative. Applying (5.1), the derivative of $\ell \mapsto \alpha(\ell)$ is

$$\frac{d\alpha}{d\ell}(\ell) = \frac{d}{d\ell}\int_{b\in\mathbf{R}} 2\,\Phi\Big[-\ell/2 + b/\ell\Big]\,e^{-b/2}\,h(b)\,db = -\int_{b\in\mathbf{R}} \varphi\Big[-\ell/2 + b/\ell\Big]\Big\{1 + \frac{2b}{\ell^2}\Big\}\,e^{-b/2}\,h(b)\,db$$

with $\varphi(x) = \Phi'(x) = e^{-x^2/2}/\sqrt{2\pi}$ the density of a standard Gaussian distribution. Algebra shows that the function $b \mapsto b\,e^{-b/2}\,\varphi[-\ell/2 + b/\ell]$ is odd so that the derivative simplifies,

$$\frac{d\alpha}{d\ell}(\ell) = -\int_{b\in\mathbf{R}} \varphi\Big[-\ell/2 + b/\ell\Big]\,e^{-b/2}\,h(b)\,db.$$

This quantity is clearly strictly negative, concluding the proof of Proposition 1.

5.3. *Proof of Proposition 2.* The upper bound follows from a similar argument to that in [AV14]. Let $\widetilde{W}$ be an independent copy of W^* and let $V \overset{\mathcal{D}}{\sim} \mathbf{N}(-\ell^2/2, \ell^2)$ be independent from any other source of randomness. Relating $\widetilde{W}$ to W through (2.3) yields

$$\mathbb{E}\left[1 \wedge \exp(V+B)\right] = \mathbb{E}\left[\exp(\widetilde{W}) \wedge \exp(V)\exp(W^*)\right] \le \mathbb{E}\left[1 \wedge \exp(V)\right] = 2 \times \Phi(-\ell/2);$$

we have applied Jensen's inequality twice to the function $(x, y) \mapsto x \wedge \exp(V)y$ which is concave in both x and y. Since $J(\ell) = \mathbb{E}\left[1 \wedge \exp(V+B)\right]$, the upper bound follows.

The lower bound follows from a similar argument to that used in [DPDK14]. We note that $(1 \wedge e^V)(1 \wedge e^B) \le 1 \wedge e^{V+B}$. V and B are independent by assumption; as $\alpha_{\max} = \mathbb{E}[1 \wedge e^B]$, the result follows on taking expectations with respect to both of these variables.

5.4. *Proof of Theorem 2.* In this section we use the following notations. We write $u_n \lesssim v_n$ when the absolute value of the quotient u_n/v_n is bounded above by a constant which is independent of the index n; we write $u_n \asymp v_n$ if $u \lesssim v_n$ and $v_n \lesssim u_n$. For $(\mathbf{x}, w) \in \mathbf{R}^d \times \mathbf{R}$ we write $\mathbb{E}_{\mathbf{x},w}[\,\cdot\,]$ instead of $\mathbb{E}[\,\cdot\,|(\mathbf{X}_0^{(d)}, W_0^{(d)}) = (\mathbf{x}, w)]$. The Metropolis-Hastings accept-reject function is the globally Lipschitz function $F(u) = 1 \wedge e^u$. The log-likelihood function is denoted by $A := \log f$ in this section. We drop the index $(\cdot)^{(d)}$ when the context is clear.

The proof follows ideas from [Béd07], which itself is an adaptation of the original paper [RGG97]. It is based on ([EK86], Theorem 8.2, Chapter 4) which gives conditions under which the finite dimensional distributions of a sequence of processes converge weakly to those of some Markov process. ([EK86], Corollary 8.6, Chapter 8) provides further conditions for this sequence of processes to be relatively compact in the appropriate topology and thus establish weak convergence of the stochastic processes themselves.

The situation is slightly more involved than the one presented in [RGG97, Béd07]; the proof needs an homogenisation argument since the processes $\mathbf{X}^{(d)}$ and $W^{(d)}$ evolve on two different time scales. Indeed, it will become apparent from the proof that the process $\mathbf{X}^{(d)}$ takes $\mathcal{O}(d)$ steps to mix while the process $W^{(d)}$ takes $\mathcal{O}(1)$ steps to mix. In order to exploit this time-scales separation, we introduce an intermediary time scale $T_d = \lfloor d^\gamma \rfloor$ where $0 < \gamma < 1/4$ is an exponent whose exact value is not important to the proof. The intuition is that after $\mathcal{O}(T_d)$ steps the process $W^{(d)}$ has mixed while each coordinate of $\mathbf{X}^{(d)}$ has only moved by an infinitesimal quantity. We introduce the subsampled processes $\widetilde{\mathbf{X}}^{(d)}$ and $\widetilde{W}^{(d)}$ defined by

$$\widetilde{\mathbf{X}}_k^{(d)} = \mathbf{X}_{kT_d}^{(d)} \qquad \text{and} \qquad \widetilde{W}_k^{(d)} = W_{kT_d}^{(d)}.$$

One step of the process $\widetilde{\mathbf{X}}^{(d)}$ (resp. $\widetilde{W}^{(d)}$) corresponds to T_d steps of the process $\mathbf{X}^{(d)}$ (resp. $W^{(d)}$). We then define an accelerated version $\widetilde{V}^{(d)}$ of the subsampled first coordinate process $k \mapsto \widetilde{X}_{k,1}^{(d)}$. In order to prove a diffusion limit for the first coordinate of the process $\mathbf{X}^{(d)}$, one needs to accelerate time by a factor of d; consequently, in order to prove a diffusion limit for the process $\widetilde{\mathbf{X}}^{(d)}$, one needs to accelerate time by a factor d/T_d and thus define $\widetilde{V}^{(d)}$ by

$$\widetilde{V}^{(d)}(t) := \widetilde{X}_{\lfloor td/T_d \rfloor,1}^{(d)}.$$

The proof then consists of showing that the sequence $\widetilde{V}^{(d)}$ converges weakly in the Skorohod topology towards the limiting diffusion (2.15) and verifying that $\|\widetilde{V}^{(d)} - V^{(d)}\|_{\infty,[0,T]}$ converges to zero in probability; this is enough to prove that the sequence $V^{(d)}$ converges weakly in the Skorohod topology towards the limiting diffusion (2.15). The proof is divided into three main steps. First, we show that the finite dimensional marginals of the process $\widetilde{V}^{(d)}$ converge to those of the limiting diffusion (2.15). Second, we establish that the sequence $\widetilde{V}^{(d)}$ is weakly relatively compact. These two steps prove that the sequence $\widetilde{V}^{(d)}$ converges weakly in the Skorohod topology towards the diffusion (2.15). As a final step, we prove that the quantity $\|\widetilde{V}^{(d)} - V^{(d)}\|_{\infty,[0,T]}$ converges to zero in probability, establishing the weak convergence of the sequence $V^{(d)}$ towards the diffusion (2.15). Before embarking on the proof we define several quantities that will be needed in the sequel. We denote by $\mathcal{L}$ the generator of the limiting diffusion (2.15). Similarly, we define $\mathcal{L}^{(d)}$ and $\widetilde{\mathcal{L}}^{(d)}$ the approximate generators of the first coordinate process $\{X_{k,1}^{(d)}\}_{k\geq 0}$ and its accelerated version $\{\widetilde{X}_{k,1}^{(d)}\}_{k\geq 0}$; for any smooth and compactly supported test function $\varphi : \mathbf{R} \to \mathbf{R}$, vector

$\mathbf{x} = (x_1, \ldots, x_d) \in \mathbf{R}^d$ and scalar $w \in \mathbf{R}$ we have

$$\begin{cases} \mathcal{L}\varphi(x_1) & = \frac{1}{2} h(\ell) \big[\varphi''(x_1) + A(x_1)\, \varphi'(x_1) \big], \\ \mathcal{L}^{(d)}\varphi(\mathbf{x}, w) & = \mathbb{E}_{\mathbf{X}^{(d)}, W} \big[\varphi(X_{1,1}^{(d)}) - \varphi(x_1) \big] / \delta, \\ \widetilde{\mathcal{L}}^{(d)}\varphi(\mathbf{x}, w) & = \mathbb{E}_{\mathbf{X}^{(d)}, W} \big[\varphi(\widetilde{X}_{1,1}^{(d)}) - \varphi(x_1) \big] / \left(T_d \times \delta \right) \end{cases}$$

with $\delta = 1/d$. Note that although φ is a scalar function, the functions $\mathcal{L}^{(d)}\varphi$ and $\widetilde{\mathcal{L}}^{(d)}\varphi$ are defined on $\mathbf{R}^d \times \mathbf{R}$. In the sequel we sometimes write $\widetilde{\mathcal{L}}^{(d)}\varphi(x_1, \ldots, x_d, w)$ instead of $\widetilde{\mathcal{L}}^{(d)}\varphi(\mathbf{x}, w)$.

5.4.1. *Convergence of the finite dimensional distributions of $\widetilde{V}^{(d)}$.* In this section we prove that the finite dimensional distributions of the sequence of processes $\widetilde{V}^{(d)}$ converge weakly to those of the diffusion (2.15). Since the limiting process is a scalar diffusion, the set of smooth and compactly supported functions is a core for the generator of the limiting diffusion ([EK86],Theorem 2.1, Chapter 8); in the sequel, one can thus work with test functions belonging to this core only. To prove the convergence of the finite dimensional marginals, one can apply ([EK86],Chapter 4, Theorem 8.2, Corollary 8.4) to the pair $(\xi^{(d)}, \varphi^{(d)})$ defined by

$$\xi^{(d)}(t) = \frac{1}{\delta\, T_d} \int_t^{t + \delta\, T_d} \varphi[\widetilde{V}^{(d)}(s)]\, ds \quad \text{and} \quad \varphi^{(d)}(t) = \widetilde{\mathcal{L}}^{(d)}\varphi\Big(\widetilde{\mathbf{X}}^{(d)}_{\lfloor td/T_d \rfloor}, \widetilde{W}^{(d)}_{\lfloor td/T_d \rfloor} \Big) \tag{5.2}$$

To establish that this result applies, we will concentrate on proving that for any smooth and compactly supported function $\varphi : \mathbf{R} \to \mathbf{R}$ the following limit holds,

$$\lim_{d \to \infty} \mathbb{E} \big| \widetilde{\mathcal{L}}^{(d)}\varphi(X_1, \ldots, X_d, W) - \mathcal{L}\varphi(X_1) \big| = 0, \tag{5.3}$$

for $\{X_k\}_{k \geq 1}$ an i.i.d sequence of random variables distributed according to $f(x)\, dx$ and $W \overset{\mathcal{D}}{\sim} e^w g^*(w)\, dw$. Equation (5.3) implies Equation (8.11) of ([EK86],Chapter 4) and the stationarity Assumption implies Equations (8.8) and (8.9) of ([EK86],Chapter 4). To verifify that Equation (8.10) of ([EK86],Chapter 4) holds, one can notice that for any index $k \geq 1$ we have $\mathbb{E}\big[\big| \varphi(X_{k,1}^{(d)}) - \varphi(X_{0,1}^{(d)}) \big| \big] \lesssim k\, \delta^{1/2}$, which is a direct consequence of the triangle inequality and the fact that φ is a Lipschitz function. The proof of (5.3) is based on an averaging argument that exploits the following relationship between the generators $\mathcal{L}^{(d)}$ and $\widetilde{\mathcal{L}}^{(d)}$,

$$\widetilde{\mathcal{L}}^{(d)}\varphi(\mathbf{x}, w) = \mathbb{E}_{\mathbf{x}, w} \Big[\frac{1}{T_d} \sum_{k=0}^{T_d - 1} \mathcal{L}^{(d)}\varphi(\mathbf{X}_k^{(d)}, W_k^{(d)}) \Big]. \tag{5.4}$$

Equation (5.4) follows from the telescoping expansion $\varphi(\mathbf{X}_{T_d}^{(d)}) - \varphi(\mathbf{X}_0^{(d)}) = \sum_{k=0}^{T_d - 1} \varphi(\mathbf{X}_{k+1}^{(d)}) - \varphi(\mathbf{X}_k^{(d)})$ and the law of iterated conditional expectations. The following lemma is crucial

LEMMA 1. **(Asymptotic expansion of $\mathcal{L}^{(d)}\varphi$)**
Let Assumptions 1 and 3 be satisfied. There exist two bounded and continuous functions $a, b : \mathbf{R} \to \mathbf{R}$ satisfying the following properties.

1. *Let W be a random variable distributed as the stationary distribution of the log-noise, $W \overset{\mathcal{D}}{\sim} e^w\, g^*(w)\, dw$, and $\alpha(\ell)$ be the asymptotic mean acceptance probability identified in Theorem 1. The following identity holds,*

$$\mathbb{E}[a(W)] = \mathbb{E}[b(W)] = \frac{1}{2}\, \alpha(\ell). \tag{5.5}$$

2. *For any smooth and compactly supported function* $\varphi : \mathbf{R} \to \mathbf{R}$ *the averaged generator* $\mathcal{G}\varphi$ *defined for any* $(x_1, w) \in \mathbf{R}^2$ *by*

$$\mathcal{G}\varphi(x_1, w) := \frac{\ell^2}{I}\Big[\, a(w)\, A'(x_1)\, \varphi'(x_1) + \, b(w)\, \varphi''(x_1)\Big]$$

satisfies

$$\lim_{d\to\infty} \mathbb{E}\Big|\mathcal{L}^{(d)}\varphi(X_1, \ldots, X_d, W) - \mathcal{G}\varphi(X_1, W)\Big|^2 = 0$$

for an i.i.d sequence $\{X_k\}_{k\geq 1}$ *marginally distributed as* $f(x)\,dx$ *and constant* I *defined by* (2.13).

The above Lemma thus shows that the approximate generator $\mathbb{E}_{\mathbf{X}^{(d)},W}\big[\varphi(X_{1,1}^{(d)}) - \varphi(x_1)\big]/\delta$ asymptotically only depends on the first coordinate $x_1 \in \mathbf{R}$ and the log-noise $w \in \mathbf{R}$. The proof is an averaging argument for the $(d-1)$ coordinates $(x_2, \ldots, x_d)$; this is mainly technical and details can be found in Appendix A.1. The next step consists in exploiting the separation of scales between the processes $\{\mathbf{X}_k^{(d)}\}_{k\geq 0}$ and $\{W_k^{(d)}\}_{k\geq 0}$.

LEMMA 2. *Let* $h : \mathbf{R} \to \mathbf{R}$ *be a bounded measurable function. Suppose that for any* $d \geq 1$ *the Markov chain* $\{(\mathbf{X}_k^{(d)}, W_k^{(d)})\}_{k\geq 0}$ *is started at stationarity. The following limit holds,*

$$\lim_{d\to\infty}\ \mathbb{E}\Big|\frac{1}{T_d}\sum_{k=0}^{T_d-1} h(W_k^{(d)}) - \mathbb{E}[h(W)]\ \Big|\ =\ 0,$$

with W *distributed according to the stationary distribution* $W \overset{\mathcal{D}}{\sim} e^w g^*(w)\,dw$.

The above lemma thus shows that $T_d = \lfloor d^\gamma \rfloor$ steps, with $0 < \gamma < 1/4$, are enough for the process $W^{(d)}$ to mix. The proof relies on a coupling argument and the ergodic theorem for Markov chains. Details can be found Appendix A.2. We now have all the tools in hands to prove Equation (5.3). First, with the notation $\mathbf{X}^{(d)} = (X_1, \ldots, X_d)$, the telescoping expansion (5.4) and conditional Jensen's inequality yields

$$\begin{aligned}\mathbb{E}\big|\widetilde{\mathcal{L}}^{(d)}\varphi(\mathbf{X}^{(d)}, W) - \mathcal{L}\varphi(X_1)\big| &= \mathbb{E}\Big|\mathbb{E}_{\mathbf{X}^{(d)},W}\Big[\frac{1}{T_d}\sum_{k=0}^{T_d-1}\mathcal{L}^{(d)}\varphi(\mathbf{X}_k^{(d)}, W_k^{(d)}) - \mathcal{L}\varphi(X_{0,1}^{(d)})\Big]\Big| \\ &\leq \mathbb{E}\Big|\frac{1}{T_d}\sum_{k=0}^{T_d-1}\mathcal{L}^{(d)}\varphi(\mathbf{X}_k^{(d)}, W_k^{(d)}) - \mathcal{L}\varphi(X_{0,1}^{(d)})\Big|.\end{aligned}$$

One can then use the triangle inequality to obtain the bound

$$\begin{aligned}\mathbb{E}\big|\widetilde{\mathcal{L}}^{(d)}\varphi(\mathbf{X}^{(d)},W) \quad - \quad \mathcal{L}\varphi(X_1)\big| &\leq \mathbb{E}\Big|\frac{1}{T_d}\sum_{k=0}^{T_d-1}\mathcal{L}^{(d)}\varphi(\mathbf{X}_k^{(d)},W_k^{(d)}) - \mathcal{L}\varphi(X_{0,1}^{(d)})\Big| \\ &\leq \frac{1}{T_d}\sum_{k=0}^{T_d-1}\mathbb{E}\Big|\mathcal{L}^{(d)}\varphi(\mathbf{X}_k^{(d)},W_k^{(d)}) - \mathcal{G}\varphi(X_{k,1}^{(d)},W_k^{(d)})\Big| \\ &\quad +\mathbb{E}\Big|\frac{1}{T_d}\sum_{k=0}^{T_d-1}\mathcal{G}\varphi(X_{k,1}^{(d)},W_k^{(d)}) - \mathcal{G}\varphi(X_{0,1}^{(d)},W_k^{(d)})\Big| \\ &\quad +\mathbb{E}\Big|\frac{1}{T_d}\sum_{k=0}^{T_d-1}\mathcal{G}\varphi(X_{0,1}^{(d)},W_k^{(d)}) - \mathcal{L}\varphi(X_{0,1}^{(d)})\Big| \\ &=: E_1(d)+E_2(d)+E_3(d).\end{aligned}$$

To complete the proof of the convergence of the finite dimensional distributions of $\widetilde{V}^{(d)}$ towards those of the limiting diffusion (2.15), it remains to prove that $E_i(d)\to 0$ as $d\to\infty$ for $i=1,2,3$.

- Since the Markov chain $\{(\mathbf{X}_k^{(d)},W^{k,})\}_{k\geq 0}$ is assumed to be stationary the quantity $E_1(d)$ also equals $\mathbb{E}\big|\,\mathcal{L}^{(d)}\varphi(X_1,\ldots,X_d,W)-\mathcal{G}\varphi(X_1,W)\,\big|$. Lemma 1 shows that $E_1(d)\to 0$ as $d\to\infty$.
- The formula for the quantity $\mathcal{G}\varphi(x,w)$ shows that the expectation $E_2(d)$ also reads

$$\begin{aligned}\frac{\ell^2}{I}\times\mathbb{E}\Big|\frac{1}{T_d}\sum_{k=0}^{T_d-1}a(W_k^{(d)})\Big\{A'(X_{k,1}^{(d)})\,\varphi'(X_{k,1}^{(d)}) - A'(X_{0,1}^{(d)})\,\varphi'(X_{0,1}^{(d)})\Big\} \\ +\frac{1}{T_d}\sum_{k=0}^{T_d-1}b(W_k^{(d)})\Big\{\varphi^{''}(X_{k,1}^{(d)})-\varphi^{''}(X_{0,1}^{(d)})\Big\}\Big|.\end{aligned} \tag{5.6}$$

  Under Assumptions 3 the function A' is globally Lipschitz; since φ is smooth with compact support the functions $x\mapsto A'(x)\varphi'(x)$ and $x\mapsto\varphi^{''}$ are both globally Lipschitz. Using the boundedness of the functions a and b, this yields that the quantity in Equation (5.6) is bounded by a constant multiple of $\frac{1}{T_d}\sum_{k=0}^{T_d-1}\mathbb{E}\big|X_{k,1}^{(d)}-X_{0,1}^{(d)}\big|$. For any index $k\geq 0$ we have $\mathbb{E}\big|X_{k+1,1}^{(d)}-X_{k,1}^{(d)}\big|\lesssim\delta^{1/2}$ so that $\mathbb{E}\big|X_{k,1}^{(d)}-X_{0,1}^{(d)}\big|\lesssim k\,\delta^{1/2}$. Since $T_d/d^{1/2}\to 0$, the conclusion follows.
- Lemma 1 shows that one can express the generator of the limiting diffusion (2.15) as $\mathcal{L}\varphi(x)=\frac{\ell^2}{I}\mathbb{E}[a(W)]\,A'(x)\,\varphi'(x)+\frac{\ell^2}{I}\mathbb{E}[b(W)]\,\varphi^{''}(x)$. The expectation $E_3(d)$ thus also reads

$$\begin{aligned}\frac{\ell^2}{I}\times\mathbb{E}\Big|\Big\{\frac{1}{T_d}\sum_{k=0}^{T_d-1}a(W_k^{(d)})-\mathbb{E}[a(W)]\Big\}A'(X_{0,1}^{(d)})\,\varphi'(X_{0,1}^{(d)}) \\ +\Big\{\frac{1}{T_d}\sum_{k=0}^{T_d-1}b(W_k^{(d)})-\mathbb{E}[b(W)]\Big\}\varphi^{''}(X_{0,1}^{(d)})\Big|.\end{aligned}$$

  Because the function φ is smooth with compact support, it follows (Cauchy-Schwarz) that this quantity is less than a constant multiple of

$$\mathbb{E}\Big[\Big\{\frac{1}{T_d}\sum_{k=0}^{T_d-1}a(W_k^{(d)})-\mathbb{E}[a(W)]\Big\}^2\Big]^{1/2}\times\mathbb{E}\big[A'(X)^2\big]^{1/2}+\mathbb{E}\Big|\frac{1}{T_d}\sum_{k=0}^{T_d-1}b(W_k^{(d)})-\mathbb{E}[b(W)]\Big|.$$

Lemma 2 shows that $\mathbb{E}\big|\frac{1}{T_d}\sum_{k=0}^{T_d-1} b(W_k^{(d)}) - \mathbb{E}[b(W)]\big| \to 0$ and under Assumptions 3 the expectation $\mathbb{E}\big[A'(X)^2\big]$ is finite. Therefore, to finish the proof of the limit $E_3(d) \to 0$, one needs to verify that $\mathbb{E}\big[\big\{\frac{1}{T_d}\sum_{k=0}^{T_d-1} a(W_k^{(d)}) - \mathbb{E}[a(W)]\big\}^2\big] \to 0$. According to Lemma 2, the sequence $\Big(\frac{1}{T_d}\sum_{k=0}^{T_d-1} a(W_k^{(d)}) - \mathbb{E}[a(W)]\Big)$ converges in L^1 to zero. The sequence is also bounded in L^∞ since the function a is bounded. A sequence bounded in L^∞ that converges to zero in L^1 also converges to zero in any L^p for $1 \le p < \infty$. The conclusion follows.

5.4.2. *Relative weak compactness of the sequence* $\widetilde{V}^{(d)}$. The process $\widetilde{V}^{(d)}$ is started at stationarity and the space of smooth functions with compact support is an algebra that strongly separates points. ([EK86]Chapter 4,Corollary 8.6) shows that in order to prove that the sequence $\widetilde{V}^{(d)}$ is relatively weakly compact in the Skorohod topology it suffices to verify that Equation (8.33) and (8.34) of ([EK86],Chapter 4) hold.

- To prove (8.34) it suffices to show that for any smooth and compactly supported test function φ the sequence $d \mapsto \mathbb{E}\big|\widetilde{\mathcal{L}}^{(d)}\varphi(X_1,\dots,X_d,W)\big|^2$ is bounded. One can use the telescoping expansion (5.4), Lemma 1 and the stationarity of the Markov chain $\{(\mathbf{X}_k^{(d)}, W_k^{(d)})\}_{k\ge 0}$ and obtain that

$$\begin{aligned}
\mathbb{E}\big|\widetilde{\mathcal{L}}^{(d)}\varphi(\mathbf{X}^{(d)},W)\big|^2 &\lesssim \mathbb{E}\big|\frac{1}{T_d}\sum_{k=0}^{T_d-1}\mathcal{L}^{(d)}\varphi(\mathbf{X}_k^{(d)},W_k^{(d)}) - \mathcal{G}\varphi(X_{k,1}^{(d)},W_k^{(d)})\big|^2 \\
&\quad + \mathbb{E}\big|\frac{1}{T_d}\sum_{k=0}^{T_d-1}\mathcal{G}\varphi(X_{k,1}^{(d)},W_k^{(d)})\big|^2 \\
&\le \frac{1}{T_d}\sum_{k=0}^{T_d-1}\mathbb{E}\big|\mathcal{L}^{(d)}\varphi(\mathbf{X}^{(d)},W_k^{(d)}) - \mathcal{G}\varphi(X_{k,1}^{(d)},W_k^{(d)})\big|^2 + \frac{1}{T_d}\sum_{k=0}^{T_d-1}\mathbb{E}\big|\mathcal{G}\varphi(X_{k,1}^{(d)},W_k^{(d)})\big|^2 \\
&= \mathbb{E}\big|\mathcal{L}^{(d)}\varphi(\mathbf{X}^{(d)},W) - \mathcal{G}\varphi(X_1,W)\big|^2 + \mathbb{E}\big|\mathcal{G}\varphi(X_1,W)\big|^2 = o(1)+\mathcal{O}(1).
\end{aligned}$$

  This proves Equation (8.34).
- To prove (8.33) one needs to show that the expectation of $\sup\big\{\,\big|\xi_d(t) - \widetilde{V}^{(d)}(t)\big| \;:\; t\in[0,T]\big\}$ converges to zero as $d\to\infty$, where the process ξ_d is defined in Equation (5.2). Note that the supremum is less than

$$\|\varphi\|_{\mathrm{Lip}} \times \sup\Big\{\delta \times \sum_{k=i}^{j}\big|X_{j,1}^{(d)} - X_{i,1}^{(d)}\big| \;:\; 0\le i<j\le d\times T \quad\text{and}\quad |i-j|\le T_d\Big\} \tag{5.7}$$

  where $\|\varphi\|_{\mathrm{Lip}}$ is the Lipschitz constant of φ. Therefore, since $\big|X_{j,1}^{(d)} - X_{i,1}^{(d)}\big| \lesssim \delta\sum_{k=i}^{j-1}|Z_k|$ where $\{Z_k\}_{k\ge 0}$ are i.i.d standard Gaussian random variables such that $X_{i,1}^{(d),*} = X_{i,1}^{(d)} + \ell\, I^{-1/2}\,\delta\, Z_k$, the following Lemma gives the conclusion.

LEMMA 3. *Let $\{\xi_k\}_{k\ge 1}$ an i.i.d sequence of standard Gaussian random variables $\boldsymbol{N}(0,1)$. We have*

$$\lim_{d\to\infty}\mathbb{E}\Big[\;\sup\Big\{\delta\times\sum_{k=i}^{j}|\xi_k| \;:\; 0\le i<j\le d\times T \quad \textit{and} \quad |i-j|\le T_d\Big\}\;\Big] \;=\; 0.$$

PROOF. Indeed, it suffices to prove that δ times the expectation of the supremum $\sup\{S(i,d) : i \leq d/T_d\}$, with $S(i,d) = \sum_{k=iT_d}^{(i+1)T_d} |\xi_k|$, converges to zero; this follows from Markov's inequality and standard Gaussian computations. □

This concludes the proof of the relative weak compactness in the Skorohod topology. The sequence of processes $\widetilde{V}^{(d)}$ is weakly compact in the Skorohod topology and the finite dimensional distributions of $\widetilde{V}^{(d)}$ converge to the finite dimensional distribution of the diffusion (2.15). Consequently, the sequence of processes $\widetilde{V}^{(d)}$ converges weakly in the Skorohod space $D([0,T])$ to the diffusion (2.15). The next section shows that the discrepancy between $V^{(d)}$ and $\widetilde{V}^{(d)}$ is small and thus proves that the sequence of processes $V^{(d)}$ also converges to the diffusion (2.15).

5.4.3. *Discrepancy between $V^{(d)}$ and $\widetilde{V}^{(d)}$.* Since $\sup_{t\leq T} \left|V_t^{(d)} - \widetilde{V}_t^{(d)}\right|$ is less than the supremum of Equation (5.7), Lemma (3) yields that $\|\widetilde{V} - V^{(d)}\|_{\infty,[0,T]}$ converges to zero in probability. This ends the proof of Theorem 2.

6. Discussion. We have examined the behaviour of the pseudo-marginal random walk Metropolis algorithm in the limit as the dimension of the target approaches infinity, under the assumption that the noise in the estimate of the log-target at a proposed new value, $\mathbf{x}$, is additive and independent of $\mathbf{x}$.

Subject to relatively general conditions on the target, limiting forms for the acceptance rate and for the efficiency, in terms of expected squared jump distance (ESJD), have been obtained. We examined two different noise distributions (Gaussian and Laplace), and found that the optimal scaling of the proposal is insensitive to the variance of the noise and to whether the noise has a Gaussian or a Laplace distribution.

We then examined the behaviour of the Markov chain on the target, $\mathbf{x}$, and the noise, obtaining a limiting diffusion for the first component of a target with independent and identically distributed components. The efficiency function in this case is proportional to the speed of the diffusion, thus further justifying the use of ESJD in this context.

We identified a 'standard asymptotic regime' under which the additive noise is Gaussian with variance inversely proportional to the number of unbiased estimates that are used. In this regime the efficiency function is especially tractable, and we showed that it is maximised when the acceptance rate is approximately 7.0% and the variance of the Gaussian noise is approximately 3.3. We noted that in this regime the optimal noise variance is also insensitive to the choice of scaling.

A detailed simulation study on a Lotka Volterra Markov jump process using a particle filter suggested that in the scenario considered the assumptions of the standard asymptotic regime are reasonable provided the number of particles is not too low. Furthermore, whilst the assumption that the distribution of the noise does not depend on the current position is not true, variations in the distribution have a small effect on the distribution of the estimates of the log-target compared with the effect of the noise itself. The optimal scaling was found to be insensitive to the noise variance (or equivalently the number of particles) and the optimal noise variance was relatively insensitive to the choice of scaling. The overall optimal scaling was consistent with the theoretical value obtained; however the optimal variance was a little lower than the theoretically optimal value. Investigations showed that this discrepancy can be explained by the differences between our theoretical measure of efficiency (ESJD) and empirical measures used in the simulation study (ESS).

The results from the simulation study suggest that in low dimension a safer option than tuning to a particular variance and acceptance rate might be to take advantage of the insensitivity of the optimal scaling to the variance and vice versa and optimise scaling and variance independently.

The diffusion limit provides strong support for the optimisation strategies suggested by the ESJD criterion. However, in an ideal world it would be good to show that the sequence of algorithms which achieves the minimal optimal integrated autocorrelation time for a given functional might converge to the optimal diffusion. This is a generic question which is relevant to all diffusion limits for MCMC algorithms, and there are still important open questions regarding the relationships between ESJD, diffusion limits and limiting optimal integrated autocorrelation. In this direction, a recent paper [**?**] has shown that diffusion limits can be translated into *complexity* results, thus demonstrating that at least the order of magnitude of the number of iterations to "converge" can be read off from the diffusion limit.

The optimal variance of 3.28 under the standard asymptotic regime is similar to the value of 2.83 obtained in [DPDK14] under the same noise assumptions and for a scenario where the component of the Markov chain on $\mathcal{X}$ mixes infinitely more slowly than the noise component. Indeed, as noted in a remark following Corollary 1, 2.83 is (to two decimal places) the optimal variance that we obtain when $\ell = 0$. There are many differences between the approaches in [DPDK14] and this article. For example, we optimise a limiting efficiency for the random walk Metropolis with respect to both the scaling and the variance whereas [DPDK14] considers the univariate optimisation of a bound on the efficiency of Metropolis-Hastings kernels which satisfy a positivity condition. That a similar conclusion may be drawn from two very different approaches is encouraging.

APPENDIX A: PROOF OF TECHNICAL LEMMAS

Let $\{X_j\}_{j\geq 1}$ be an i.i.d sequence of random variables distributed as $f(x)\,dx$, $W \overset{\mathcal{D}}{\sim} e^w g^*(w)\,dw$, $\{Z_{k,j}\}_{k\geq 0, j\geq 1}$ an i.i.d sequence of $\mathbf{N}(0,1)$ random variables, $\{U_k\}_{k\geq 0}$ an i.i.d sequence of random variables uniformly distributed on $(0,1)$ and $\{W_k^*\}_{k\geq 0}$ an i.i.d sequence distributed as $g^*(w)\,dw$. All these random variables are assumed to be independent from one another. For all integers $1 \leq j \leq d$ we set $X_{0,j}^{(d)} = X_j$ and $W_0^{(d)} = W$. We introduce the proposals $X_{k,j}^{(d),*} = X_{k,j}^{(d)} + \ell\, I^{-1/2}\, d^{-\frac{1}{2}}\, Z_{k,j}$ and define $\big(X_{k+1}^{(d)}, W_{k+1}^{(d)}\big) = \big(X_k^{(d),*}, W_k^*\big)$ if

$$U_k < F\Big(W_k^* - W_k^{(d)} + \sum_{j=1}^{d} A(X_{k,j}^{(d),*}) - A(X_{k,j}^{(d)})\Big)$$

and $\big(X_{k+1}^{(d)}, W_{k+1}^{(d)}\big) = \big(X_k^{(d)}, W_k^{(d)}\big)$ otherwise. We define $\mathbf{X}^{(d)} = (X_{k,1}^{(d)}, \ldots, X_{k,d}^{(d)})$. For any dimension $d \geq 1$ the process $\{\mathbf{X}_k^{(d)}, W_k^{(d)})\}_{k\geq 1}$ is a Metropolis-Hastings Markov chain started at stationarity, i.e. $(\mathbf{X}_0^{(d)}, W_0^{(d)}) = (X_1, \ldots, X_d, W) \overset{\mathcal{D}}{\sim} \pi^{(d)}$, targeting the distribution $\pi^{(d)}$.

A.1. Proof of Lemma 1. In this section, for notational convenience, we write Z_j instead of $Z_{0,j}$ and W^* instead of W_0^*. We set

$$a(w) := \mathbb{E}\big[F'(\Omega + W^* - w)\big] \quad \text{and} \quad b(w) := \frac{1}{2}\,\mathbb{E}\big[F(\Omega + W^* - w)\big] \tag{A.1}$$

with $F'(u) = e^u\, \mathrm{I}_{\{u<0\}}$ and $\Omega \overset{\mathcal{D}}{\sim} \mathbf{N}(-\ell^2/2, \ell^2)$ independent from all other sources of randomness. To prove Lemma 1, it suffices to show that the function a and b are continuous, bounded, satisfy identity (5.5) and that the following two limits hold,

$$\begin{cases} \lim_{d\to\infty} & \mathbb{E}\big|\mathbb{E}_d\big[(X_1^{1,d} - X_1)/\delta\big] - \ell^2\, I^{-1}\, a(W) A'(X_1)\big|^2 = 0 \\ \lim_{d\to\infty} & \mathbb{E}\big|\frac{1}{2}\,\mathbb{E}_d\big[(X_1^{1,d} - X_1)^2/\delta\big] - \ell^2\, I^{-1}\, b(W)\big|^2 = 0. \end{cases} \tag{A.2}$$

We have used the notation $\mathbb{E}_d[\cdots]$ for $\mathbb{E}[\cdots | X_1, \ldots, X_d, W]$. The fact that the functions a and b are bounded and continuous follows from the dominated convergence theorem.

- *Proof of Equation* (5.5). Note that $\mathbb{E}[b(W)] = \frac{1}{2}\mathbb{E}[1 \wedge \exp(\Omega + B)]$ with $B := W^* - W$. A standard computation show that for any $\beta \in \mathbf{R}$ we have $\mathbb{E}[1 \wedge \exp(\Omega + \beta)] = 2\Phi(-\ell/2 + \beta/\ell)$, so that the identity $\mathbb{E}[b(W)] = \frac{1}{2}\alpha(\ell)$ directly follows from the definition of α in Theorem 1. For proving the identity $\mathbb{E}[a(W)] = \frac{1}{2}\alpha(\ell)$, note that the expectation $\mathbb{E}[a(W)]$ equals

$$\begin{aligned}
&\iiint_{(z,w,w^*)\in\mathbf{R}^3} e^{-\ell^2/2+\ell z+w^*-w}\,\mathrm{I}_{\{-\ell^2/2+\ell z+w^*-w<0\}}\,e^w g^*(w)g^*(w^*)\frac{e^{-z^2/2}}{\sqrt{2\pi}}\,dw\,dw^*\,dz\\
&\quad= \iiint_{(z,w,w^*)\in\mathbf{R}^3} \mathrm{I}_{\{-\ell^2/2+\ell(-z+\ell)+w-w^*>0\}}\,e^{w^*} g^*(w)g^*(w^*)\frac{e^{-(-z+\ell)^2/2}}{\sqrt{2\pi}}\,dw\,dw^*\,dz\\
&\quad= \iiint_{(z,w,w^*)\in\mathbf{R}^3} \mathrm{I}_{\{-\ell^2/2+\ell z+w^*-w>0\}}\,e^{w^*} g^*(w)g^*(w^*)\frac{e^{-z^2/2}}{\sqrt{2\pi}}\,dw\,dw^*\,dz\\
&\quad= \mathbb{E}[\mathrm{I}_{\{\Omega+W^*-W>0\}}].
\end{aligned}$$

We have used the change of variable $(z, w^*, w) \to (-z + \ell, w, w^*)$ to go from the second line to the third. This computation shows that $\mathbb{E}[a(W)] := \mathbb{E}[e^{\Omega+W^*-W}\,\mathrm{I}_{\{\Omega+W^*-W<0\}}] = \mathbb{E}[\mathrm{I}_{\{\Omega+W^*-W>0\}}]$. Since $F(u) = 1 \wedge e^u = e^u\,\mathrm{I}_{\{u<0\}} + \mathrm{I}_{u\geq 0}$ it follows that

$$\alpha(\ell) = \mathbb{E}[F(\Omega + W^* - W)] = \mathbb{E}[e^{\Omega+W^*-W}\,\mathrm{I}_{\{\Omega+W^*-W<0\}}] + \mathbb{E}[\mathrm{I}_{\{\Omega+W^*-W>0\}}],$$

and therefore $\mathbb{E}[a(W)] = \alpha(\ell)/2$.

- *Proof of Equation* (A.2) We will only verify that the first limit in Equation (A.2) holds. The proof of the second limit is similar but easier. In other words, we will focus on proving that the sequence $\mathbb{E}_d\big[(X_1^{1,d} - X_1)/\delta\big]$ converges in L^2 to $\ell^2\, I^{-1}\, a(W)A'(X_1)$. An integration by parts shows that for any continuous function $g : \mathbf{R} \to \mathbf{R}$ such that g' has a finite number of discontinuities, if $g(Z)$ and $g'(Z)$ have a finite first moment for $Z \overset{\mathcal{D}}{\sim} \mathbf{N}(0, 1)$, the identity $\mathbb{E}[Z \times g(Z)] = \mathbb{E}[g'(Z)]$ holds. It follows that

$$\begin{aligned}
\mathbb{E}_d\big[(X_1^{1,d} - X_1)/\delta\big] &= \ell\, I^{-1/2}\,\delta^{1/2}\,\mathbb{E}_d[Z_1 \times F(\Omega^{(d)} + W^* - W)\,]\\
&= \ell^2\, I^{-1}\,\mathbb{E}_d\big[\,F'(\Omega^{(d)} + W^* - W)\;\times\; A'(x_1 + \ell\, I^{-1/2}\,\delta^{1/2}\,Z_1)\big]
\end{aligned}$$

with $\Omega^{(d)} = \sum_{i=1}^d A(X_i + \ell\, I^{-1/2}\,\delta^{1/2}\,Z_i) - A(X_i)$. Under Assumption 3 the function $A' = (\log f)'$ is globally Lipschitz so that, since the function F' is bounded, one can focus on proving that

$$\mathbb{E}_d\big[\,F'(\Omega^{(d)} + W^* - W)\big]\;\times\; A'(X_1)$$

converges in L^2 to $a(W)A'(X_1)$. By the Cauchy-Schwarz inequality, this reduces to proving that

$$\lim_{d\to\infty}\ \mathbb{E}\Big[\Big|\mathbb{E}_d[\,F'(\Omega^{(d)} + W^* - W)\,] - \mathbb{E}_d[\,F'(\Omega + W^* - W)\,]\Big|^4\Big]\ =\ 0.$$

By the Portmanteau's theorem and the dominated convergence theorem and the definition of $\Omega^{(d)}$, this reduces to proving that for almost every realisation $\{x_i\}_{i\geq 1}$ of the i.i.d sequence $\{X_i\}_{i\geq 1}$ the following limit holds in distribution,

$$\lim_{d\to\infty}\ \sum_{i=1}^{d} A(x_i + \ell\, I^{-1/2}\,\delta^{1/2}\,Z_i) - A(x_i) = \Omega.$$

Under Assumption 3 the third derivative of A is bounded so that a second order Taylor expansion yields that the difference $A(x_i+\ell\, I^{-1/2}\,\delta^{1/2}\,Z_i)-A(x_i)$ equals $A'(x_i)\,\ell\, I^{-1/2}\,\delta^{1/2}\,Z_i+(1/2)\,A''(x_i)\,\ell^2\,I^{-1}\,\delta\,Z_i^2+\mathcal{O}(d^{-3/2})$; consequently,

$$\sum_{i=1}^{d} A(x_i+\ell\, I^{-1/2}\,\delta^{1/2}\,Z_i)-A(x_i) \stackrel{\text{law}}{=} \frac{\ell^2}{2\,I}\Big\{\frac{\sum_{i=1}^d A''(x_i)}{d}\Big\}+\ell\, I^{-1/2}\,\Big\{\frac{\sum_{i=1}^d A'(x_i)^2}{d}\Big\}^{1/2}\xi + \frac{\ell^2}{2\,I}\Big\{\frac{\sum_{i=1}^d A''(x_i)\,(Z_i^2-1)}{d}\Big\}+\mathcal{O}(d^{-1/2})$$

for $\xi \stackrel{\mathcal{D}}{\sim} \mathbf{N}(0,1)$ independent from all other sources of randomness. The law of large numbers shows that for almost every realisation $\{x_i\}_{i\geq 1}$ the right-hand-side of the above equation converges in distribution towards $\Omega \stackrel{\mathcal{D}}{\sim} \mathbf{N}(-\ell^2/2,\ell^2)$.

A.2. Proof of Lemma 2. For convenience, we first give a high level description of the reasoning. We construct processes $\{\widehat{W}_k^{(d)}\}_{k\geq 0}$, $\{\widehat{Y}_k^{(d)}\}_{k\geq 0}$ and $\{Y_k\}_{k\geq}$ satisfying the following.

- With high probability $\widehat{W}_k^{(d)}=W_k^{(d)}$ for all $k\leq T_d$.
- The process $\{\widehat{Y}_k^{(d)}\}_{k\geq 0}$ has the same law as the process $\{\widehat{W}_k^{(d)}\}_{k\geq 0}$.
- With high probability $\widehat{Y}_k^{(d)}=Y_k$ for all $k\leq T_d$.
- The process $\{Y_k\}_{k\geq 0}$ is a Markov chain that is ergodic with invariant distribution $e^w\,g^*(w)\,dw$.

One can thus use an approximation of the type

$$\mathbb{E}\Big|\frac{1}{T_d}\sum_{k=0}^{T_d-1} h(W_k^{(d)})-\mathbb{E}[h(W)]\Big| \;\approx\; \mathbb{E}\Big|\frac{1}{T_d}\sum_{k=0}^{T_d-1} h(Y_k)-\mathbb{E}[h(W)]\Big|$$

and the usual ergodic theorem gives the conclusion. We use at several places the following elementary lemma.

LEMMA 4. *Let $T_d=\lfloor d^\gamma\rfloor$ with $0<\gamma<\frac{1}{4}$. Let $\{P_k^{(d)}\}_{k,d\geq 0}$ and $\{Q_k^{(d)}\}_{k,d\geq 0}$ be two arrays of $(0,1)$-valued random variables. Let $\{U_k\}_{k\geq 0}$ be a sequence of random variables uniformly distributed on the interval $(0,1)$. We suppose that for all dimension $d\geq 1$ the random variable U_k is independent from $\{P_j^{(d)}\}_{j=0}^{k-1}$ and $\{Q^{(d)}j\}_{j=0}^{k-1}$. Consider the event*

$$E_k^{(d)}:=\Big\{\omega:\quad \mathrm{I}_{\{U_j<P_j^{(d)}\}}=\mathrm{I}_{\{U_j<Q_j^{(d)}\}}\quad \textit{for all}\quad 0\leq j\leq k\Big\}.$$

Under the assumption that $\mathbb{E}\Big[|P_k^{(d)}-Q_k^{(d)}|\,\big|E_{k-1}^{(d)}\Big]\lesssim k/\sqrt{d}$ we have $\lim_{d\to\infty}\mathbb{P}\big(E_{T_d}^{(d)}\big)=1$.

PROOF. Note that $\mathbb{P}(E_k^{(d)})=\mathbb{P}(E_0^{(d)})\prod_{j=1}^k\mathbb{P}\big[\mathrm{I}_{\{U_j<P_j^{(d)}\}}=\mathrm{I}_{\{U_j<Q_j^{(d)}\}}\,\big|E_{j-1}^{(d)}\big]$. Since U_j is supposed to be independent from the event $E_{j-1}^{(d)}$ it follows that $\mathbb{P}\big[\mathrm{I}_{\{U_j<P_j^{(d)}\}}=\mathrm{I}_{\{U_j<Q_j^{(d)}\}}\,\big|E_{j-1}^{(d)}\big]=1-\mathbb{E}\big[|P_j^{(d)}-Q_j^{(d)}|\,\big|E_{j-1}^{(d)}\big]$. The conclusion then directly follows from the bound $\mathbb{E}\Big[|P_k^{(d)}-Q_k^{(d)}|\,\big|E_{k-1}^{(d)}\Big]\lesssim k/\sqrt{d}$ and $\gamma<1/4$. □

We now describe the construction of the processes $\{\widehat{W}_k^{(d)}\}_{k\geq 0}$, $\{\widehat{Y}_k^{(d)}\}_{k\geq 0}$ and $\{Y_k\}_{k\geq 0}$. To this end, we need an i.i.d sequence $\{\xi_k\}_{k\geq 0}$ of standard $\mathbf{N}(0,1)$ Gaussian random variables independent

from all other sources of randomness. All the processes start at the same position $W_0^{(d)} = \widehat{W}_0^{(d)} = \widehat{Y}_0^{(d)} = Y_0 = W$. We define $\widehat{W}_{k+1}^{(d)} = W_k^*$ if

$$U_k < F\Big[\frac{\ell}{\sqrt{d\,I}}\sum_{j=1}^{d} A'(X_j)Z_{k,j} - \ell^2/2 + W_k^* - \widehat{W}_k^{(d)}\Big]$$

and $\widehat{W}_{k+1}^{(d)} = \widehat{W}_k^{(d)}$ otherwise. We define $\widehat{Y}_{k+1}^{(d)} = W_k^*$ if

$$U_k < F\Big[\ell\, I^{-1/2}\Big\{d^{-1}\sum_{j=1}^{d} A'(X_j)^2\Big\}^{1/2}\xi_k - \ell^2/2 + W_k^* - \widehat{Y}_k^{(d)}\Big]$$

and $\widehat{Y}_{k+1}^{(d)} = \widehat{Y}_k^{(d)}$ otherwise. We define $Y_{k+1} = W_k^*$ if

$$U_k < F\Big[\ell\,\xi_k - \ell^2/2 + W_k^* - Y_k\Big]$$

and $Y_{k+1} = Y_k$ otherwise.

- $W_k^{(d)} = \widehat{W}_k^{(d)}$ *with high probability.*
  We prove that $\lim_{d\to\infty}\mathbb{P}\big[W_k^{(d)} = \widehat{W}_k^{(d)} \,:\, k = 1,\dots,T_d\big] = 1$. Because the Metropolis-Hastings function F is globally Lipschitz, Lemma 4 shows that it suffices to verify that

$$\mathbb{E}\Big|\sum_{j=1}^{d} A(X_{k,j}^{(d),*}) - A(X_{k,j}^{(d)}) - A'(X_j)\ell\, I^{-1/2}\, Z_{k,j}/\sqrt{d} + \frac{\ell^2}{2}\Big| \lesssim k/\sqrt{d}. \tag{A.3}$$

  Under Assumptions 3 the second and third derivatives of A are bounded so that the bound (A.3) follows from a second order Taylor expansion,

$$\begin{aligned}
&\mathbb{E}\Big|\sum_{j=1}^{d} A(X_{k,j}^{(d),*}) - A(X_{k,j}^{(d)}) - A'(X_j)\ell\, I^{-1/2}\, Z_{k,j}/\sqrt{d} + \ell^2/2\Big| \\
&\quad\lesssim \mathbb{E}\Big|\sum_{j=1}^{d} A(X_{k,j}^{(d),*}) - A(X_{k,j}^{(d)}) - \frac{\ell}{\sqrt{d\,I}}A'(X_{k,j}^{(d)})Z_{k,j} - \frac{\ell^2}{2\,I\,d}A''(X_{k,j}^{(d)})Z_{k,j}^2\Big| \\
&\qquad + \frac{\ell}{\sqrt{d\,I}}\mathbb{E}\Big|\sum_{j=1}^{d}\big(A'(X_{k,j}^{(d)}) - A'(X_j)\big)Z_{k,j}\Big| + \frac{\ell^2}{2\,I\,d}\mathbb{E}\Big|\sum_{j=1}^{d}\big(A''(X_{k,j}^{(d)}) - A''(X_j)\big)Z_{k,j}^2\Big| \\
&\qquad + \frac{\ell^2}{2\,I}\mathbb{E}\Big|\frac{1}{d}\sum_{j=1}^{d} A''(X_j) + I\Big| \\
&\quad\lesssim \frac{1}{\sqrt{d}} + \frac{1}{\sqrt{d}}\Big\{\sum_{j=1}^{d}\mathbb{E}\big|A'(X_{k,j}^{(d)}) - A'(X_j)\big|^2\Big\}^{1/2} + \frac{1}{2d}\sum_{j=1}^{d}\mathbb{E}\big|A''(X_{k,j}^{(d)}) - A''(X_j)\big| \\
&\qquad + \mathbb{E}\Big|d^{-1}\sum_{j=1}^{d} A''(X_j) + I\Big| \lesssim \frac{1}{\sqrt{d}} + \frac{k}{\sqrt{d}} + \frac{k}{\sqrt{d}} + \frac{1}{\sqrt{d}}.
\end{aligned}$$

  We have used the bound $\mathbb{E}|X_{k,j}^{(d)} - X_j|^2 \lesssim \frac{k^2}{d}$.

- $\widehat{W}^{(d)}$ *and* $\widehat{Y}^{(d)}$ *have same law.*
  It is straightforward to verify that the processes $\{\widehat{W}_k^{(d)}\}_{k\geq 0}$ and $\{\widehat{Y}_k^{(d)}\}_{k\geq 0}$ have the same law.
- $\widehat{Y}_k^{(d)} = Y_k$ *with high probability.*
  We prove that $\lim_{d\to\infty} \mathbb{P}\big[\widehat{Y}_k^{(d)} = Y_k \ : \ k = 1, \dots, T_d\big] = 1$. Lemma 4 shows that this follows from the elementary bound $\mathbb{E}\Big|\Big\{d^{-1}\sum_{j=1}^d A'(X_j)^2\Big\}^{1/2} - I^{1/2}\Big| \lesssim 1/\sqrt{d}$.

We now show that the Markov chain $\{Y_k\}_{k\geq 0}$ is a Markov chain that is reversible with respect to the distribution $e^w g^*(w)\, dw$,

$$e^x g^*(x) g^*(y)\, \mathbb{E}\big[\mathbb{E}[F(\Omega + y - x)]\big] = e^y g^*(y)\, g^*(x)\, \mathbb{E}\big[\mathbb{E}[F(\Omega + x - y)]\big]$$

for all $x, y \in \mathbf{R}^2$. This boils down to verifying that the function $(x, y) \mapsto e^x\, \mathbb{E}[F(\Omega + y - x)]$ is symmetric; Proposition 2.4 of [RGG97] shows that this quantity can be expressed as

$$e^x\, \Phi\Big(\frac{-\frac{1}{2}\ell^2 + y - x}{\ell}\Big) + e^y\, \Phi\Big(\frac{-\frac{1}{2}\ell^2 + x - y}{\ell}\Big),$$

which is indeed symmetric. Note that this Markov chain corresponds to the *penalty method* of [CD99]; see also [NFW12] for a discussion of this algorithm. The ergodic theorem for Markov chains applies; for any bounded and measurable function $h : \mathbf{R} \to \mathbf{R}$ we have

$$\lim_{N\to\infty} \mathbb{E}\Big|\frac{1}{N}\sum_{k=0}^{N-1} h(Y_k) - \mathbb{E}[h(W)]\Big| = 0.$$

One can thus use the triangle inequality several times to see that for any bounded and measurable function $h : \mathbf{R} \to \mathbf{R}$ we have

$$\begin{aligned}
\mathbb{E}\Big|\frac{1}{T_d}\sum_{k=0}^{T_d-1} h(W_k^{(d)}) - \mathbb{E}[h(W)]\Big| &\leq \frac{1}{T_d}\sum_{k=0}^{T_d-1} \mathbb{E}\Big|h(W_k^{(d)}) - h(\widehat{W}_k^{(d)})\Big| + \mathbb{E}\Big|\frac{1}{T_d}\sum_{k=0}^{T_d-1} h(\widehat{W}_k^{(d)}) - \mathbb{E}[h(W)]\Big| \\
&\lesssim (1 - \mathbb{P}[W_k^{(d)} = \widehat{W}_k^{(d)} \ : \ k = 1, \dots, T_d]) + \mathbb{E}\Big|\frac{1}{T_d}\sum_{k=0}^{T_d-1} h(\widehat{Y}_k^{(d)}) - \mathbb{E}[h(W)]\Big| \\
&\lesssim o(1) + \frac{1}{T_d}\sum_{k=0}^{T_d-1} \mathbb{E}\Big|h(\widehat{Y}_k^{(d)}) - h(Y_k)\Big| + \mathbb{E}\Big|\frac{1}{T_d}\sum_{k=0}^{T_d-1} h(Y_k) - \mathbb{E}[h(W)]\Big| \\
&= o(1) + o(1) + o(1),
\end{aligned}$$

which finishes the proof of Lemma 2.

APPENDIX B: DETAILS OF THE LOTKA VOLTERRA MODEL

In this Appendix, we present details of the Lotka-Volterra model used in the simulation study of Section 4. The Lotka-Volterra model is a continuous-time Markov chain on $\mathbb{N}_0^2$. The transitions and associated rates for this model are:

$$(u_1, u_2) \xrightarrow{x_1 u_1 u_2} (u_1 + 1, u_2 - 1), \quad (u_1, u_2) \xrightarrow{x_2 u_1} (u_1 - 1, u_2) \ \text{ and } \ (u_1, u_2) \xrightarrow{x_3 u_2} (u_1, u_2 + 1);$$

the rate for any other transition is zero. Observations of the Markov chain, when they occur, are subject to Gaussian error:

$$\mathbf{Y}(t) \sim \mathbf{N}\left(\begin{bmatrix} u_1(t) \\ u_2(t) \end{bmatrix}, \begin{bmatrix} x_4 & 0 \\ 0 & x_5 \end{bmatrix}\right).$$

Using $\mathbf{x} = (0.006, 0.6, 0.3, 25, 49)$, a realisation of the stochastic process was simulated from initial value $\mathbf{u}(0) = (70, 70)$ for $T = 50$ time units. The state, perturbed with Gaussian noise, $\mathbf{y}(t)$, was recorded at $t = 1, 2, \ldots, T$. For inference, $X_1, \ldots, X_5$ were assumed to be independent, *a priori* with $\log X_i \sim \text{Unif}\,[-8, 8]$, $(i = 1, \ldots, 5)$.

The initial value for each chain was a vector of estimates of the posterior median for each parameter, obtained from the initial run; hence no "burn-in" was required. Each algorithm was run for 2.5×10^5 iterations, except with $m = 50$ and $m = 80$, where 10^6 iterations were used. Output was thinned by a factor of 10 for storage.

ACKNOWLEDGEMENTS

We are grateful to the associate editor and three referees for their comments, which helped improve both the presentation and the content of this article.

REFERENCES

[ADH10] Christophe Andrieu, Arnaud Doucet, and Roman Holenstein. Particle Markov chain Monte Carlo methods. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 72(3):269–342, 2010.

[AR09] Christophe Andrieu and Gareth O. Roberts. The pseudo-marginal approach for efficient Monte Carlo computations. *Ann. Statist.*, 37(2):697–725, 2009.

[AV14] Christophe Andrieu and Matti Vihola. Convergence properties of pseudo marginal Markov chain Monte Carlo algorithms. *arXiv preprint arXiv:1210.1484*, 2014.

[BDMD13] Jean Bérard, Pierre Del-Moral, and Arnaud Doucet. A lognormal central limit theorem for particle approximations of normalizing constants. *arXiv preprint arXiv:1307.0181*, 2013.

[Bea03] Mark A. Beaumont. Estimation of population growth or decline in genetically monitored populations. *Genetics*, 164:1139–1160, 2003.

[Béd07] Mylène Bédard. Weak convergence of Metropolis algorithms for non-i.i.d. target distributions. *Ann. Appl. Probab.*, 17(4):1222–1244, 2007.

[BGJM11] Steve Brooks, Andrew Gelman, Galin L. Jones, and Xiao-Li Meng, editors. *Handbook of Markov chain Monte Carlo*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. CRC Press, Boca Raton, FL, 2011.

[BPS+04] Laird Arnault Breyer, Mauro Piccioni, Sergio Scarlatti, et al. Optimal scaling of mala for nonlinear regression. *The Annals of Applied Probability*, 14(3):1479–1505, 2004.

[BR00] Laird Breyer and Gareth O. Roberts. From Metropolis to diffusions: Gibbs states and optimal scaling. *Stoch.Proc. Appl.*, 90:181–206, 2000.

[BR08] Mylène Bédard and Jeffrey S. Rosenthal. Optimal scaling of Metropolis algorithms: heading toward general target distributions. *Canad. J. Stat.*, 36:483–503, 2008.

[BRS09] Alexandros Beskos, Gareth O. Roberts, and Andrew Stuart. Optimal scalings for local Metropolis-Hastings chains on nonproduct targets in high dimensions. *Ann. Appl. Probab.*, 19(3):863–898, 2009.

[CD99] DM Ceperley and M Dewing. The penalty method for random walks with uncertain energies. *The Journal of chemical physics*, 110:9812, 1999.

[DM04] Pierre Del Moral. *Feynman-Kac formulae*. Probability and its Applications (New York). Springer-Verlag, New York, 2004. Genealogical and interacting particle systems with applications.

[DPDK14] Arnaud Doucet, Michael Pitt, George Deligiannidis, and Robert Kohn. Efficient implementation of Markov chain Monte Carlo when using an unbiased likelihood estimator. *arXiv preprint arXiv:1210.1871v3*, 2014.

[EK86] S.N. Ethier and T.G. Kurtz. *Markov processes: Characterization and convergence*, volume 6. Wiley New York, 1986.

[FPR08] Paul Fearnhead, Omiros Papaspiliopoulos, and Gareth O. Roberts. Particle filters for partially observed diffusions. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 70(4):755–777, 2008.

[GSS93] N. J. Gordon, D. J. Salmond, and A. F. M. Smith. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *Radar and Signal Processing, IEE Proceedings F*, 140(2):107–113, April 1993.

[GW11] Andrew Golightly and Darren J. Wilkinson. Bayesian parameter inference for stochastic biochemical network models using particle Markov chain Monte Carlo. *Interface Focus*, 1(6):807–820, 2011.

[KdV12] J. Knape and P. de Valpine. Fitting complex population models by combining particle filters with Markov chain Monte Carlo. *Ecology*, 93(2):256–263, 2012.

[LM03] N Li and Stephens M. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, pages 2213–2233, 2003.

[NFW12] Geoff K Nicholls, Colin Fox, and Alexis Muir Watt. Coupled mcmc with a randomized acceptance probability. *arXiv preprint arXiv:1205.6857*, 2012.

[PDS11] George Poyiadjis, Arnaud Doucet, and Sumeetpal S. Singh. Particle approximations of the score and observed information matrix in state space models with application to parameter estimation. *Biometrika*, 98(1):65–80, 2011.

[PdSSGK12] Michael K. Pitt, Ralph dos Santos Silva, Paolo Giordani, and Robert Kohn. On some properties of Markov chain Monte Carlo simulation methods based on the particle filter. *Journal of Econometrics*, 171(2):134 – 151, 2012.

[PG10] Cristian Pasarica and Andrew Gelman. Adaptively scaling the Metropolis algorithm using expected squared jumped distance. *Statistica Sinica*, 20:343–364, 2010.

[PST12] Natesh S. Pillai, Andrew M. Stuart, and Alexandre H. Thiéy. Optimal scaling and diffusion limits for the langevin algorithm in high dimensions. *Annals of Applied Probability*, 22(6):2320–2356, 2012.

[RGG97] Gareth O. Roberts, Andrew Gelman, and Walter R. Gilks. Weak convergence and optimal scaling of random walk Metropolis algorithms. *The Annals of Applied Probability*, 7:110–120, 1997.

[RR98] Gareth O. Roberts and Jeffrey S. Rosenthal. Optimal scaling of discrete approximations to Langevin diffusions. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 60(1):255–268, 1998.

[RR01] Gareth O Roberts and Jeffrey S Rosenthal. Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science*, 16:351–367, 2001.

[RR13] Gareth O. Roberts and Jeffrey S. Rosenthal. Minimising MCMC variance via diffusion limits, with an application to simulated tempering. *Annals of Applied Probability*, 2013. To appear.

[SFR10] Chris Sherlock, Paul Fearnhead, and Gareth O. Roberts. The random walk Metropolis: linking theory and practice through a case study. *Statist. Sci.*, 25(2):172–190, 2010.

[She13] Chris Sherlock. Optimal scaling of the random walk Metropolis: general criteria for the 0.234 acceptance rule. *J. App. Prob.*, 50(1):1–15, 2013.

[SR93] Adrian F.M. Smith and Gareth O. Roberts. Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods (with discussion). *J. Royal Statistical Society, series B*, pages 4–23, 1993.

[SR09] Chris Sherlock and Gareth O. Roberts. Optimal scaling of the random walk Metropolis on elliptically symmetric unimodal targets. *Bernoulli*, 15(3):774–798, 2009.

[Tie94] Luke Tierney. Markov chains for exploring posterior distributions. *Ann. Statist.*, 22(4):1701–1762, 1994. With discussion and a rejoinder by the author.

Department of Mathematics and Statistics, Lancaster University, LA1 4YF, UK
E-mail: c.sherlock@lancaster.ac.uk

National University of Singapore, Singapore,
E-mail: a.h.thiery@nus.edu.sg

Department of Statistics, University of Warwick, UK, and University of Toronto, Canada