Optimal scaling for the pseudo-marginal random walk Metropolis: insensitivity to the noise generating mechanism.

Chris Sherlock [1]

Abstract

We examine the optimal scaling and the efficiency of the pseudo-marginal random walk Metropolis algorithm using a recently-derived result on the limiting efficiency as the dimension, $d \to \infty$. We prove that the optimal scaling for a given target varies by less than 20% across a wide range of distributions for the noise in the estimate of the target, and that any scaling that is within 20% of the optimal one will be at least 70% efficient. We demonstrate that this phenomenon occurs even outside the range of noise distributions for which we rigorously prove it. We then conduct a simulation study on an example with $d = 10$ where importance sampling is used to estimate the target density; we also examine results available from an existing simulation study with $d = 5$ and where a particle filter was used. Our key conclusions are found to hold in these examples also.

Classification: 65C05, 65C40.

Keywords: Pseudo marginal Markov chain Monte Carlo, random walk Metropolis, optimal scaling, Particle MCMC, robustness.

[2]

[1]Department of Mathematics and Statistics, Lancaster University, Lancaster LA1 4YF, UK. c.sherlock@lancaster.ac.uk

[2]The final publication is available at Springer via http://dx.doi.org/10.1007/s11009-015-9471-6

1 Introduction

The pseudo-marginal Metropolis-Hastings algorithm (PsMMH) (Beaumont, 2003; Andrieu and Roberts, 2009) supposes that it is impossible or infeasible to evaluate a target density, $\pi(x),\ x \in \mathcal{X} \subseteq \mathbb{R}^d$, but that an estimator $\hat{\pi}_W(x) = \pi(x)e^W$ can be constructed.

A Markov chain is created from an initial value $x^{(0)}$ and a noisy estimate of the target $\hat{\pi}_{w^{(0)}}(x^{(0)})$ as follows. At iteration i, given the current value x and $\hat{\pi}_w(x)$, a new value x^* is proposed from some density $q(x^*|x)$. An estimate,

$$\hat{\pi}_{w^*}(x^*) = \pi(x^*)e^{w^*} \tag{1}$$

is then constructed by, effectively, sampling from $g(w^*|x^*)$. The proposed value, x^*, and the estimate, $\hat{\pi}_{w^*}(x^*)$, are then accepted with probability $1 \wedge [\hat{\pi}_{w^*}(x^*)q(x|x^*)] \,/\, [\hat{\pi}_w(x)q(x^*|x)]$. The proposal density for the noise, $g(w|x),\ w \in (-\infty, \infty)$ must possess the property that $\int_{-\infty}^{\infty} \mathrm{d}w\ e^w g(w|x) = c > 0$. Provided that $c > 0$ its exact value is irrelevant in all that follows and so without loss of generality we take $c = 1$ and refer to $\hat{\pi}_{W^*}(x^*)$ as 'the unbiased estimator of the target'. Both w and w^* are unknown since $\pi(x)$ and $\pi(x^*)$ are unknown; nevertheless, the above algorithm can be viewed as constructing a Markov chain $\{(X_k, W_k)\}_{k\geq 0}$. The stationary density of this Markov chain is

$$\pi(x)g(w|x)e^w, \tag{2}$$

which admits $\pi(x)$ as a marginal. Samples from the Markov chain may therefore be used to approximately compute expectations with respect to $\pi(x)$. The additive noises in the log-target at the current and proposed values, respectively W and W^*, are henceforth simply referred to as *additive noises*.

The pseudo-marginal random walk Metropolis (PsMRWM) is a special case of the PsMMH with $q(x^*|x) = q(x^* - x) = q(x - x^*)$, so that the acceptance probability simplifies to $1 \wedge \frac{\hat{\pi}_{w^*}(x^*)}{\hat{\pi}_w(x)}$. One common practice is to set

$$X^*|x \sim N(x, \lambda^2 \hat{V}), \tag{3}$$

for a scaling parameter, λ, and where $\hat{V}$ is an estimate of the posterior variance, obtained from an initial run of the algorithm. The PsMRWM is one of the most popular forms of PsMMH (e.g. Golightly and Wilkinson, 2011; Knape and de Valpine, 2012; Sherlock *et al.*, 2014) because it does not require the computation or estimation of other properties of the target, such as local gradients.

Often the method of producing an unbiased estimator of the target has a tuning parameter, m, such as the number of particles in a particle filter (Andrieu *et al.*, 2010) or the number of Monte Carlo samples when importance sampling. As part of a general tuning strategy for optimising the efficiency of the algorithm, for a particular m^*, a practitioner might find, using repeated runs, the optimal scaling, $\hat{\lambda}^*$: the scaling which maximises the efficiency of the algorithm. They would then wish to know whether or not $\hat{\lambda}^*$ might be a sensible value to use for other choices of m, or whether 'retuning' would be necessary.

Sherlock *et al.* (2015) derive an expression, which is valid in the limit as the dimension of the target approaches infinity (see Section 2.1), for the efficiency of a pseudo-marginal RWM algorithm as a function of the scaling and the form of the additive noise: the limiting expected squared jumping distance (ESJD). Sherlock *et al.* (2015) then examine two particular forms for the distribution of the additive noise in the estimate of the logarithm of the target, Gaussian and Laplace, and find that the theoretical optimal scaling is insensitive to the variance of the noise and even to which of the two distributions is used.

We consider the form of efficiency derived in Sherlock *et al.* (2015). Provided that across the range of m values to be considered the density of the additive noise, g, is always log-concave, our theoretical result implies that $\hat{\lambda}^*$ will be within 20% of the optimal scaling for any other choice of m. Furthermore, for any given m, the efficiency at $\hat{\lambda}^*$ will be at least 70% of the maximum achievable efficiency. The two-dimensional optimisation problem of choosing λ and m values that approximately maximise the efficiency can therefore effectively be reduced to two one-dimensional optimisation problems.

This introduction concludes with Section 1.1 which discusses a property that is, in a sense, the converse to the insensitivity of optimal scaling to the noise distribution: the insensitivity of the optimal variance of the noise distribution to the proposal kernel $q(x^*|x)$. The main theoretical result of this article, Theorem 1, is stated and proved in Section 2. Given that $\int \mathrm{d}w \; g(w)e^w$ is finite, g cannot, at least in terms of its tail behaviour, be 'too far' from log-concave. In Section 3 we demonstrate empirically that the statement in Theorem 1 that relies on the log-concavity appears to hold more generally. The efficiency measure upon which Theorem 1 is based relies on several assumptions; in particular it is a limit result for high-dimensional targets and it relies on the noise in the proposal and the proposed position in the target being independent. Section 4 examines two simulation studies for the insensitivity properties predicted by Theorem 1. Firstly, the simulation study of Sherlock *et al.* (2015), where the estimate of the target was obtained from a particle filter, then a new simulation study where the estimate of the target is obtained by importance sampling; both studies support the heuristics of Theorem 1. The article concludes with a discussion.

1.1 Sensitivity of noise choice to the form of MH proposal

The key contribution of this article is on the robustness of the choice of scaling for a PsM-RWM algorithm, or equivalently of the optimal choice from a particular restricted class of Metropolis-Hastings (MH) kernels, to the form of the noise distribution. However, it is natural to ask whether or not a converse property might hold: an insensitivity of the optimal choice from a restricted class of noise distributions to the choice of Metropolis Hastings (MH) kernel, q. The example function $f(x, y) = -(x-y)^2 - 9 \times (x-1)^2$ demonstrates that neither insensitivity need imply the other, yet both are of interest.

Research into the choice of noise distribution was initiated in Pitt *et al.* (2012) using particle filters to generate the noisy approximations to the likelihood. Several assumptions were made on the distribution of the noise in the logarithm of a new estimate of the target, $g(w^*|x^*)$.

1. The Markov chain on (X, W) is stationary.
2. The distribution is independent of position: $g(w^*|x^*) = g(w^*)$.
3. $g(w^*) = N(w^*; -\sigma^2/2, \sigma^2)$, for some variance, σ^2.
4. σ^2 is inversely proportional to the computational cost of the algorithm.

Assumption 2 was made for tractability and an heuristic argument was provided as to why it should hold in the large data limit provided the posterior for all parameters becomes tight. It has also been found to hold approximately in simuation studies on real statistical examples (Sherlock *et al.*, 2015; Doucet *et al.*, 2015). Assumptions 3 and 4 were found to hold empirically and have since been justified formally for a particle filter with a large number of particles and a large number of observations and where the computational cost is proportional to the number of particles (Bérard *et al.*, 2014). Considering a perfect independence sampler and minimising the variance of an arbitrary functional of the Markov chain in this case leads to an optimal variance of 0.85. Simulations showed this to be the case, but also that the optimal variance for a tuned RWM (in two dimensions) was only slightly larger, suggesting an insensitivity of the optimal variance to the choice of algorithm. Following this line of research, Doucet *et al.* (2015) considered an upper bound on the mixing efficiency of *any* pseudo-marginal MH algorithm. Combined with the same noise assumptions as in Pitt *et al.* (2012) this leads to a bound on the overall efficiency of the algorithm. The bound on overall efficiency is a function of the noise variance and it was shown that the variance at which it is optimised lies between 0.85 and 2.82, with the exact value depending on the efficiency of the idealised marginal algorithm. Since this applies to any algorithm it therefore applies to the PsMRWM across any range of scalings and, subject to the assumptions on the noise, implies a degree of insensitivity of the optimal variance to the choice of scaling.

Andrieu and Vihola (2014) studied the relative mixing efficiencies of any pair of pseudo-marginal algorithms which have the same MH proposal kernel $q(x^*|x)$ but different noise generating mechanisms $g_1(w|x)$ and $g_2(w|x)$. It was shown that provided the expectation of any convex function $\phi(W)$ under g_2 exceeds that under g_1 then the Markov chain that uses g_1 will mix more efficiently. Bornn *et al.* (2014) leveraged this result for a particular special case of pseudo-marginal MCMC where importance sampling is used to estimate the (smoothed) posterior: approximate Bayesian computation with a positive MH kernel and a uniform error distribution. It was shown that in this scenario the optimal number of importance samples was either 1 or 2 *whatever* the variance of the noise. This result complements investigations into importance sampling for the pseudo-marginal RWM under a gamma noise distribution as $d \to \infty$ in an early version of Sherlock *et al.* (2015) (`http://arxiv.org/abs/1309.7209v1`) where it is shown that a single importance sample is always best.

The results described in the first and second paragraphs do not contradict each other since the former rely, in particular, on Assumptions 3 and 4, a set up which is justified for the particle filter as already discussed. When importance sampling is used instead of a particle filter then, using the Central Limit Theorem and the delta method it can be shown (e.g. Pitt *et al.*, 2012, Lemma 2) that as the number of samples $m \to \infty$, the distribution of the estimate of the log-posterior will be approximately Gaussian, $N(-\sigma^2/2, \sigma^2)$, with $\sigma^2 \propto 1/m$ as, apparently, required. However, straightforward examination of the error terms shows that the delta method requires $\sigma^2 << 1$, so that the assumptions on the form and cost are not appropriate for the optimal variances of $\mathcal{O}(1)$ that are suggested in Pitt *et al.* (2012), Doucet *et al.* (2015) and Sherlock *et al.* (2015).

This converse insensitivity, or its lack, is investigated further in our simulation studies and is discussed further in Section 5.

2 Set-up and main theoretical result

2.1 The efficiency function

Sherlock *et al.* (2015) consider a sequence of targets $\left\{\pi^{(d)}(x^{(d)})\right\}_{d=1}^{\infty}$. In each dimension, d, an unbiased estimator is available, exactly as described in and around Equation (1). It is assumed that there exists a constant, $s^{(d)}$ such that

$$\lim_{d\to\infty}\frac{1}{s^{(d)}}\left|\left|\nabla\log\pi^{(d)}(X^{(d)})\right|\right|^2=1\quad\text{and}\quad\lim_{d\to\infty}\frac{1}{s^{(d)}}\nabla^2\log\pi^{(d)}(X^{(d)})=-1,$$

where $X^{(d)}\sim\pi^{(d)}$, and a regularity condition on the target allows same $s^{(d)}$ to be used in both expressions. The constant $s^{(d)}$ is a measure of the roughness of $\log\pi^{(d)}$; for example, if

$$\pi^{(d)}(x^{(d)})=\exp\left(\sum_{i=1}^{d}f\left(x_i^{(d)}\right)\right)\tag{4}$$

then $s^{(d)}=-d/\mathbb{E}\left[f''(X)\right]$, where X has density $\exp(f(x))$. The scaling for the RWM algorithm in dimension d is then set to

$$\lambda^{(d)}=\ell/\sqrt{s^{(d)}},\tag{5}$$

for some fixed ℓ, and the proposal is $X^{*(d)}=x^{(d)}+\lambda^{(d)}Z^{(d)},\ Z^{(d)}\sim N(0,I)$. Assumptions 1 and 2 from Section 1.1 are also made.

Perhaps the most natural measure of efficiency of an MCMC algorithm is the effective sample size (ESS, e.g. Carlin and Louis, 2009, Ch.3) of each component; the number of independent samples that would lead to the same variance in the estimator of the posterior mean of the component as that arising from the correlated sample of points obtained from the MCMC algorithm. Even this measure, however, has its drawbacks, since it is not invariant to the functional of the target that is being considered. Sherlock *et al.* (2015) examine the efficiency of the RWM in terms of expected squared jumping distance (ESJD) on the sequence of targets. Subject to further technical conditions on the sequence it is shown that the limiting ESJD has the form

$$J_m(\ell)=2\ell^2\mathbb{E}\left[\Phi\left(\frac{B}{\ell}-\frac{\ell}{2}\right)\right].\tag{6}$$

Here $B:=W^*-W$ is the difference in the additive noise in the estimate of $\log\pi$ at the proposed value and at the current value, and Φ denotes the cumulative distribution function of a standard Gaussian random variable. Maximising ESJD is equivalent to minimising the lag-1 autocorrelation of the chain. The following result (proved in Appendix A) extends results on the positivity of Metropolis-Hastings algorithms in Lemma 3.1 of Baxendale (2005) and Proposition 3 of Doucet *et al.* (2015) to the pseudo-marginal RWM. It shows that for jump proposal distributions such as the Gaussian or Student-t all of the eigenvalues of the algorithm are non-negative; hence, minimising the lag-1 autocorrelation is a sensible goal.

Proposition 1. *If the proposal in a Metropolis-Hastings algorithm satisfies*

$$q(x^*|x) = \int r(x,z)r(x^*,z)\,dz, \tag{7}$$

then the corresponding pseudo-marginal Metropolis-Hastings algorithm is positive.

Further justification for the use of ESJD as a measure of efficiency is provided in Sherlock *et al.* (2015) where it is shown that for the product target in (4), and subject to further technical conditions, as $d \to \infty$ a scaled version of the first component of each element in the sequence of Markov chains converges to a diffusion, the speed of which is proportional to $J_m(\ell)$. When a limiting diffusion exists, then in that limit $J_m(\ell)$ is also proportional to the ESS and is invariant (up to a multiplicative constant) to any transformation, hence $J_m(\ell)$ is unambiguously the right measure of efficiency.

2.2 Insensitivity

Our main result refers to the situation when there is no noise in the estimate of π, $B = 0$, when the limiting ESJD simplifies to

$$J_\infty(\ell) = 2\ell^2\Phi\left(-\frac{\ell}{2}\right). \tag{8}$$

In this case, as noted in Roberts *et al.* (1997), the optimal scaling is $\hat{\ell}_\infty \approx 2.38$.

When the additive noise in the log-target is Gaussian then (6) is particularly tractable and Sherlock *et al.* (2015) suggest through a plot and an asymptotic argument that $\hat{\ell}$ is between $\hat{\ell}_\infty$ and $2\sqrt{2}$, where the exact value depends on the variance of the Gaussian distribution. We show this rigorously, and for a more general form of noise distribution. We also provide bounds on the potential loss of efficiency suffered by choosing a different scaling between $\hat{\ell}_\infty$ and $2\sqrt{2}$.

Theorem 1. *Let $\hat{\ell}_m$ and $\hat{\ell}_\infty \approx 2.38$ be the values which optimise the efficiency functions $J_m(\ell)$ and $J_\infty(\ell)$ that are defined in (6) and (8). Let $g(w^*)$ be the density of W^*, the noise* *in the log-target at a proposed new target value, and assume that W^* is independent of that target value. Then*

1. $\hat{\ell}_m \geq \hat{\ell}_\infty$.
2. *If $g(w^*)$ is log-concave then* $\hat{l}_m \leq 2\sqrt{2}$.
3. *For any two scalings, ℓ_1 and ℓ_2, both in* $[\hat{\ell}_\infty, 2\sqrt{2}]$, $J_m(\ell_1)/J_m(\ell_2) > 0.70$.

Proof of Theorem 1

For simplicity of notation we suppress the subscript m throughout this proof. From (2) and the independence of W^* from X^*, the density of the noise in the log-target at the current value, W, is $e^w g(w)$. Let B have density $\rho(b)$ and note that

$$h(b) := e^{b/2}\rho(b) = \int_{-\infty}^{\infty} \mathrm{d}w\ g(w)e^{b/2+w}g(w+b) = \int_{-\infty}^{\infty} \mathrm{d}w\ g(w+b/2)g(w-b/2)e^w \tag{9}$$

is a symmetric function, $h(b) = h(-b)$. Define

$$f(b,\ell) := \ell^2\left[e^{-b/2}\Phi\left(\frac{b}{\ell}-\frac{\ell}{2}\right) + e^{b/2}\Phi\left(-\frac{b}{\ell}-\frac{\ell}{2}\right)\right]. \tag{10}$$

Using (6) and (9), the squared jumping distance is

$$\begin{aligned} J(\ell) &= 2\ell^2\int_{-\infty}^{\infty} \mathrm{d}b\ \rho(b)\ \Phi\left(\frac{b}{\ell}-\frac{\ell}{2}\right) = 2\ell^2\int_{-\infty}^{\infty}\mathrm{d}b\ h(b)\ e^{-b/2}\Phi\left(\frac{b}{\ell}-\frac{\ell}{2}\right) \\ &= 2\int_0^{\infty}\mathrm{d}b\ h(b)\ f(b,\ell), \end{aligned} \tag{11}$$

by the symmetry of h. From (8), straightforward differentiation gives:

$$\frac{d}{d\ell}(\log J_\infty) = \frac{2}{\ell} - \frac{\phi(\ell/2)}{2\Phi(-\ell/2)}, \tag{12}$$

$$\frac{d^2}{d\ell^2}(\log J_\infty) = -\frac{2}{\ell^2} - \frac{\phi(\ell/2)}{4\Phi(-\ell/2)^2}\left[\phi(\ell/2) - \frac{\ell}{2}\Phi(-\ell/2)\right] < 0\ \forall\ \ell > 0, \tag{13}$$

so that (for $\ell > 0$) J_∞ has a single stationary point (at $\hat{\ell}_\infty$), which is a maximum.

Lemma 1 provides key properties of f. Its proof is non-trivial but uninteresting and so is deferred to Appendix B.

Lemma 1. *For all $b \geq 0$, the following hold.*

1. $$\frac{2}{\ell} - \frac{\phi(\ell/2)}{2\Phi(-\ell/2)} < \frac{1}{f}\frac{\partial f}{\partial \ell} < \frac{2}{\ell}.$$

2. $$\frac{\partial f}{\partial \ell} = \ell \frac{\partial^2 f}{\partial b^2} + \left(\frac{2}{\ell} - \frac{\ell}{4}\right) f.$$

3. *$\partial f/\partial b \to 0$ as $b \to 0$ and as $b \to \infty$, whatever the value of $\ell > 0$.*

4. *For all $\ell > 0$, $\partial f/\partial b \leq 0$.*

Combining Part 1 of Lemma 1 with (12) gives $f \; d\log J_\infty/d\ell < \partial f/\partial \ell < 2f/\ell$. Multiplying by h, which is non-negative, integrating and using (11) we then obtain

$$\frac{d}{d\ell}(\log J_\infty) < \frac{d}{d\ell}(\log J) < \frac{2}{l}. \tag{14}$$

We now proceed with the proof of Theorem 1.

Proof of Part 1 of Theorem 1: by (13), for $\ell < \hat{\ell}_\infty$, $dJ_\infty/d\ell > 0$ and so $d\log J_\infty/d\ell > 0$. The result then follows from (14).

Proof of Part 2 of Theorem 1: from the definition in (9),

$$\frac{\partial h}{\partial b} = \frac{1}{2}\int_{-\infty}^{\infty} \mathrm{d}w \; g(w-b/2)g(w+b/2)e^w \left(\frac{g'(w+b/2)}{g(w+b/2)} - \frac{g'(w-b/2)}{g(w-b/2)}\right) \leq 0 \text{ for } b \geq 0 \tag{15}$$

by the log-concavity of g.

Furthermore, $\exists \; \overline{g}$ *s.t.* $g(w) \leq \overline{g} < \infty$ (since g is a log-concave density) and hence by (9)

$$h(0) \;\leq\; \overline{g}\int_{-\infty}^{\infty} \mathrm{d}w \; g(w)e^w = \overline{g}, \text{ and} \tag{16}$$

$$h(b) \;\leq\; \overline{g}\int_{-\infty}^{\infty} \mathrm{d}w \; g(w+b)e^{b/2+w} = \overline{g}e^{-b/2}\int_{-\infty}^{\infty} \mathrm{d}w \; g(w)e^w = \overline{g}e^{-b/2}. \tag{17}$$

By Part 2 of Lemma 1,

$$\frac{\mathrm{d}J}{\mathrm{d}\ell} = 2\ell\int_0^\infty \mathrm{d}b \; h(b)\frac{\partial^2 f}{\partial b^2} + \left(\frac{4}{\ell} - \frac{\ell}{2}\right)\int_0^\infty \mathrm{d}b \; h(b) \; f(b,\ell). \tag{18}$$

The first term is

$$2\ell\left[h(b)\frac{\partial f}{\partial b}\right]_0^\infty - 2\ell\int_0^\infty \mathrm{d}b\;\frac{\partial h}{\partial b}\frac{\partial f}{\partial b}.$$

Now $\left[h(b)\frac{\partial f}{\partial b}\right]_0^\infty = 0$ by (16), (17) and Part 3 of Lemma 1. Also $\partial f/\partial b \leq 0$ by Part 4 of Lemma 1, and $\partial h/\partial b \leq 0$ by (15); thus the first term in (18) cannot be positive. The second term in (18) is guaranteed to be negative provided $\ell > 2\sqrt{2}$.

Proof of Part 3 of Theorem 1: by (13), for $\ell \in [\hat{\ell}_\infty, 2\sqrt{2}]$ (and, indeed, above this), the lower bound in (14) is always negative; also the upper bound is always positive. Supposing, without loss of generality, that $\ell_2 > \ell_1$, we therefore have

$$[\log J_\infty]_{\hat{\ell}_\infty}^{2\sqrt{2}} \leq [\log J_\infty]_{\ell_1}^{\ell_2} < [\log J]_{\ell_1}^{\ell_2} < [2\log \ell]_{\ell_1}^{\ell_2} \leq [2\log \ell]_{\hat{\ell}_\infty}^{2\sqrt{2}}.$$

Evaluating the outer-most terms and exponentiating gives (to 3dp)

$$0.949\;J(\ell_1) < J(\ell_2) < 1.411\;J(\ell_1).$$

3 The log-concavity condition

The lower bound for $\hat{\ell}$ in Theorem 1 holds for all noise distributions whereas the upper bound has only been shown to hold when W^* has a log-concave density. This condition is weaker than might be thought, holding, for example, when the unbiased multiplicative noise, e^{W^*}, has a (left-truncated) t distribution or a Gamma distribution, even if the Gamma shape parameter is less than unity. Nonetheless it is natural to ask whether or not the upper bound holds more generally. The key consequence of the log-concavity of g_* is that $\partial h/\partial b \leq 0$ for $b \geq 0$. However it is clear from the proof that a weaker (yet still sufficient) condition for the upper bound is

$$\int_0^\infty \mathrm{d}b\;\frac{\partial h}{\partial b}\frac{\partial f}{\partial b} > 0.$$

Clearly there is scope for $\partial h/\partial b > 0$ over some regions whilst the whole expression in (18) remains negative, so log-concavity is certainly not a necessary condition.

We investigate the following set of discrete noise distributions, indexed by $p \in (0,1)$ and $\epsilon \in (0,1)$:

$$e^{W^*} = \begin{cases} \epsilon & w.p.\;\; p^* \\ a & w.p.\;\; 1-p^*. \end{cases},$$

where $a = (1 - p^*\epsilon)/(1 - p^*)$. In this case

$$J_{\epsilon,p^*} = 2\ell^2 \left[p^*(1-p)\Phi(-k/\ell - \ell/2) + (p^*p + (1-p^*)(1-p))\Phi(-\ell/2) + (1-p^*)p\Phi(k/\ell - \ell/2) \right],$$

where $k = \log a - \log \epsilon$, and $p = p^*\epsilon$.

The top-left panel of Figure 1 shows the optimal scaling as a function of the two noise parameters $\hat{\ell}(\epsilon, p^*)$ and demonstrates that for this set of noise distributions $2.38 < \hat{\ell} < 2.64$. Indeed, we have not been able to find a model for W^* where $\hat{\ell} > 2\sqrt{2}$ and we conjecture that $\hat{\ell} \leq 2\sqrt{2}$ whatever the distribution of W.

4 Simulation study

The efficiency measure upon which Theorem 1 is based relies on several assumptions; in particular it is a limit result for high-dimensional targets and it relies on the noise in the proposal and the proposed position in the target being independent. Furthermore, as discussed in Section 2.1, in low dimensions our theoretical measure of mixing efficiency, ESJD, and a more practically relevant measure, ESS, are no longer necessarily equivalent and it is possible that insensitivity when optimising ESJD may not translate to insensitivity when optimising ESS. To test the applicability of Theorem 1 in practice we now examine two real examples where the parameter space has a dimension of 5 and 10 respectively. We first describe the evidence of insensitivity arising from the simulation study in Sherlock *et al.* (2015), which used a particle filter, before describing a new simulation study that uses importance sampling.

Sherlock *et al.* (2015) examined the five-dimensional target distribution that arises from a continuous-time Markov jump process (the Lotka-Volterra predator-prey model), noisy observations of which are available at a set of 50 time points. A pilot run provided an estimate of the posterior variance matrix, $\hat{V}$, for the five parameters, and the jump proposal was as in (3).

Since $\ell \propto \lambda$, with the constant of proportionality unknown for any real target, to test Parts 1 and 2 of Theorem 1 we must consider the ratio of upper and lower end points and compare against $2\sqrt{2}/\hat{\ell}_\infty \approx 1.19$. There is considerable Monte Carlo variability in the efficiencies displayed in Figure 6 in Sherlock *et al.* (2015); nonetheless, over the large range of m values considered, the largest optimal scaling was no more than twice the smallest optimal scaling.

It is also clear from the same figure that over the range of optimal scalings, for each m the efficiency over this range is at least 70% of the maximum. Finally, the insensitivity result of Doucet *et al.* (2015) is also supported as the optimal variance (estimated at the posterior mean for x) ranges between 0.97 and 2.15.

The simulation study of Sherlock *et al.* (2015) had $d = 5$ and an additive noise distribution that was close to Gaussian and with a variance that was inversely proportional to the computational cost. The theory in Sherlock *et al.* (2015) is strictly valid in the limit as $d \to \infty$, yet even with this low dimension there is evidence that the optimal scaling was relatively insensitive to the choice of m. The range of variation was not as narrow as predicted by Theorem 1, although some of the excess could have been due to Monte Carlo error.

We wish to investigate the applicability of Theorem 1 further. We therefore conduct a simulation study based on a real statistical model but using importance sampling rather than a particle filter so that the additive noise is not expected to be Gaussian nor, indeed, is its variance expected to be inversely proportional to the computational cost.

4.1 Logistic regression using a latent Gaussian process

Filippone and Girolami (2014) use pseudo-marginal Metropolis-Hastings to obtain the posterior distribution of the parameters of a latent Gaussian process (GP) where the observed response is Bernoulli with a success probability determined from the GP via the probit link function. Giorgi *et al.* (2015) use Monte Carlo maximum likelihood to estimate the parameters of a generalised linear geostatistical model for binomial data where the success probability depends on a latent GP and on fixed effects via the logistic link function. In both of the above articles the likelihood for a particular set of parameter values is estimated using importance sampling with the proposal based upon the Laplace approximation or the Expectation Propagation algorithm (Filippone and Girolami, 2014) or a variation on the Laplace approximation (Giorgi *et al.*, 2015). Our statistical model is motivated by these two applications.

Let z_i $(i = 1, \ldots, l)$ be a set of points in $\mathbb{R}^a$ with components z_{ik}, $(k = 1, \ldots a)$ and let Z be the $l \times a$ matrix with ith row z_i'. We use the logistic link function and denote the overall mean on the logit scale by $\mu \in \mathbb{R}$ and covariate effects by $\beta \in \mathbb{R}^a$. The variance of the GP is $\tau^2 \in \mathbb{R}^+$ and the range parameters (one for each dimension of the process) are $\phi \in (\mathbb{R}^+)^a$, so that the correlation between the values of the GP at the l points is the $l \times l$ matrix R with elements

$$R_{ij} = \exp\left(-\sqrt{\sum_{k=1}^{a}\left(\frac{z_{ik}-z_{jk}}{\phi_k}\right)^2}\right).$$

We consider the following statistical model:

$$\begin{aligned} S|\phi,\tau^2 &\sim N_l(0,\tau^2 R) \\ p_i &= \frac{\exp(s_i+\mu+z_i^t\beta)}{1+\exp(s_i+\mu+z_i^t\beta)} \\ Y_i|s_i,\mu,\beta &\sim \mathrm{Bin}(n,p_i). \end{aligned}$$

Since all of the importance sampling algorithms in (Filippone and Girolami, 2014) and (Giorgi *et al.*, 2015) require an iterative scheme to obtain the proposal distribution, we opt instead for a simpler approach based on ideas for Poisson data in Haran and Tierney (2012) and Lampaki (2015). We first transform the data as follows:

$$y_i^+ = \begin{cases} 1/2 & \text{if } y_i = 0, \\ n-1/2 & \text{if } y_i = n, \\ y_i & \text{otherwise} \end{cases}, \quad y_i^* = \mathrm{logit}\left(\frac{y_i^+}{n}\right).$$

Using the delta method, the expectation and variance of Y_i^* given the GP are respectively

$$\mathbb{E}\left[Y_i^*|s_i\right] \approx s_i+\mu+z_i'\beta \quad \text{and} \quad \mathrm{Var}\left[Y_i^*|s_i\right] \approx \frac{1}{np_i(1-p_i)}.$$

For tractability we approximate p_i in the variance term using the observed data: $p_i \approx y_i^+/n$. This leads to a Gaussian approximation of

$$Y^*|s \sim N_l\left(s+\mu\mathbf{1}+Z\beta, D\right),$$

where D is a diagonal matrix with $1/D_{ii} = y_i^+(1-y_i^+/n)$, and $\mathbf{1}$ is an l-vector of ones. Combining this with the Gaussian prior for S leads to a Gaussian approximation for $S|y^*$ with mean μ_c and variance Σ_c, obtained via standard formulae. The proposal distribution for our importance sampler is a Student-t distribution with $\nu = 20$ degrees of freedom and density

$$q(s|y) \propto \left(1+\frac{1}{\nu}(s-\mu_c)'\Sigma_c^{-1}(s-\mu_c)\right)^{-\frac{\nu+\ell}{2}}.$$

We consider $a = 4$ so that $d = 10$, and apply the following map:

$$(\mu,\beta_1,\ldots,\beta_4,\log(\tau^2),\log(\phi_1),\ldots,\log(\phi_4)) \leftrightarrow x.$$

We place 81 points, z_i, uniformly on a hypergrid with oppposite corners at $(-\frac{1}{2},-\frac{1}{2},-\frac{1}{2},-\frac{1}{2})$ and $(\frac{1}{2},\frac{1}{2},\frac{1}{2},\frac{1}{2})$. A data set was simulated using $n = 10$ and $x = (\frac{1}{2},-1,0,0,1,0,0,0,0,0)$.

For the analysis we assume *a priori* $X \sim N_{10}(0, I)$; this prior is tight enough to prohibit difficult tail behaviour (the investigation of which is not the point of this simulation study), yet relaxed enough that the main influence is due to the likelihood (the mean diagonal term of the posterior variance matrix was 0.35, and none of the terms was larger than 0.5).

Define the sets of possible scalings, Λ, and number of importance samples, $\mathcal{M}$ as

$$\Lambda := \{0.2, 0.4, 0.6, 0.7, 0.8, 1.0, 1.2, 1.4, 1.6\}, \text{ and } \mathcal{M} := \{10, 20, 40, 100, 200, 400, 1000\}.$$

The posterior variance matrix, $\hat{V}$, was estimated from a trial run and for each combination of $\lambda \in \Lambda$ and $m \in \mathcal{M}$, a pseudo marginal RWM was run using the proposal in (3). At least 2×10^5 iterations were used, with the number increasing as m decreased so as to ensure that the effective sample size of any component was always greater than 1000.

For each $m \in \mathcal{M}$ and $\lambda \in \Lambda$ define the relative efficiencies over λ and over m, respectively as

$$\text{ESS}^*_{m,\lambda} := \frac{\text{ESS}_{m,\lambda}/T_{m,\lambda}}{\max_{\lambda\in\Lambda}(\text{ESS}_{m,\lambda}/T_{m,\lambda})} \quad \text{and} \quad \text{ESS}^{**}_{m,\lambda} := \frac{\text{ESS}_{m,\lambda}/T_{m,\lambda}}{\max_{m\in\mathcal{M}}(\text{ESS}_{m,\lambda}/T_{m,\lambda})},$$

where $\text{ESS}_{m,\lambda}$ is the minimum effective sample size over the $d = 10$ components of x, and $T_{m,\lambda}$ is the CPU time for the run.

The top-right panel in Figure 1 shows, for each $m \in \mathcal{M}$, a plot of $\text{ESS}^*_{m,\lambda}$ against λ. For each m, the optimal scaling always lies in the narrow range between 0.6 and 0.8. Furthermore, the efficiency is always at least 70% of the optimal obtainable efficiency over a much wider range than this, approximately between 0.4 and 1.0. This provides evidence that the insensitivity and robustness predicted by Theorem 1 can continue hold for moderate dimensions and when the target is not estimated using a particle filter.

The bottom left panel in Figure 1 shows kernel density plots of the estimated log-posterior at the posterior mean for x, when $m = 40$ and $m = 200$, two values that bound the range of sensible values for m for this problem (see the discussion of the bottom-right panel, below). Unlike the discrepancy from a Gaussian distribution that was found in the particle filter example in Sherlock *et al.* (2015) (and indeed in the particle filter example in Doucet *et al.* (2015) with $m = 4$) it is the right tail that is too heavy and the left tail that is too light (skewness=0.22 and 0.62 respectively), and this persists across the range of useful m values. To gauge the variability of the variance and skewness across the posterior for one of the most efficient m values, 1000 independent samples of x from the posterior were obtained by thinning a run of 10^6 iterations, which had a minimum ESS of 12 634, by a factor of 1000. For each x value, the log-target was estimated a thousand times using $m = 100$,

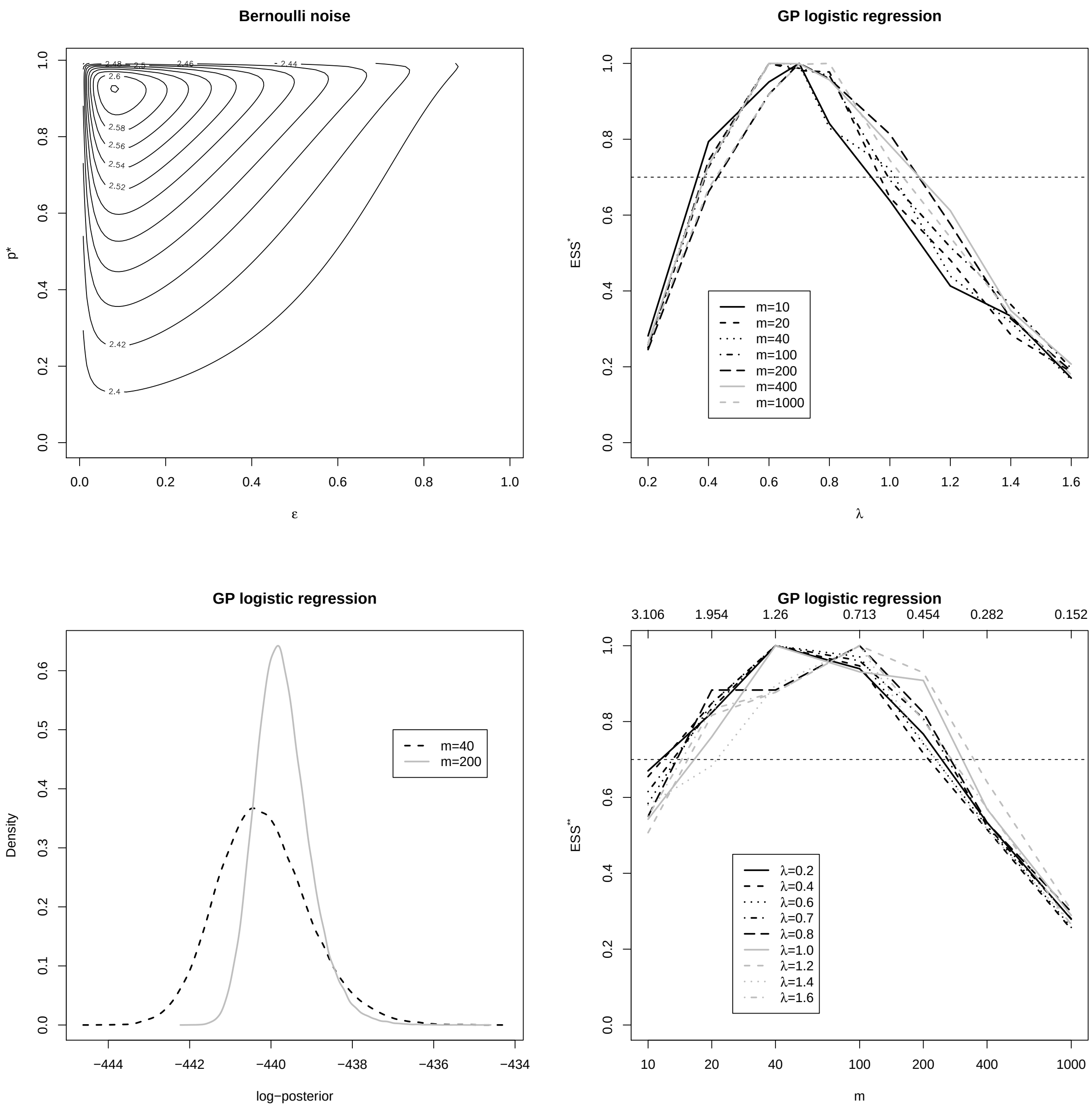

Figure 1: Top-left panel: optimal scaling for the Bernoulli noise model as a function of the two noise parameters, ϵ and p^*. Remaining panels: results from the Gaussian Process regression; top-right: ESS^* against scaling for each $m \in \mathcal{M}$; bottom-left: kernel density estimate of the distribution of the noise in the log-posterior (at the posterior mean) for $m = 40$ and $m = 200$ using 10^5 samples; bottom-right: ESS^{**} against m (bottom axis) and variance of the additive noise (top axis) for each value of $\lambda \in \Lambda$.

and the variance and skewness were noted. The $(0.025, 0.5, 0.975)$ quantiles for the variance and skewness were, respectively, $(0.54, 0.74, 0.96)$ and $(0.32, 0.52, 0.75)$, showing a moderate amount of variability over the main posterior mass.

The bottom right panel in Figure 1 shows, for each $\lambda \in \Lambda$, a plot of $\text{ESS}^{**}_{m,\lambda}$ against m. For each scaling, the optimal value of m lies between 40 and 100, corresponding to variances of approximately 1.26 or 0.71 respectively. Interestingly, also, the efficiency is around 70% or higher for all m between 20 and 200. This provides evidence that the insensitivity predicted in Doucet *et al.* (2015) can continue to hold even when the target is moderately skewed and, as is clear from the parallel scales for m and the variance of the additive noise, σ^2, that the variance is *not* inversely proportional to m (indeed a log-log plot and a simple linear regression show that, approximately, $\sigma^2 \propto m^{-0.65}$).

5 Discussion

The thrust of this article is that the optimal scaling of a pseudo-marginal RWM algorithm is insensitive to the noise distribution, and hence, when the noise is generated by an importance sampler or a particle filter, it is insensitive to the number of samples or particles, m. Moreover, for a particular m, the loss in efficiency over the range of optimal scalings, compared with the optimal efficiency for that m is small.

Theorem 1 is limited to the pseudo-marginal RWM and is strictly only proved in the limiting regime of Sherlock *et al.* (2015) which specifies, in particular, that the distribution of the additive noise in the proposal should be independent of the proposed position. However Theorem 1 requires only the mild log-concavity assumptions on the form of the noise distribution. There is an implicit assumption that, for any fixed noise generating mechanism (e.g. choice of m), the computational cost of the algorithm does not depend on the scaling. This is certainly true for the example considered in Section 4.1 (and similarly in Filippone and Girolami (2014)) and for many other examples such as inference for partially observed stochastic differential equations using a particle filter (e.g. Golightly and Wilkinson, 2011); however it is unlikely to hold in other scenarios such as inference for a Markov jump process (e.g. Golightly and Wilkinson, 2011; Sherlock *et al.*, 2015), where doubling all of the rate parameters effectively doubles the CPU time required for simulations. Even in this scenario, however, the dependence on scaling of the total CPU time for a run will be small provided the scaling is much smaller than the width of the main posterior mass, as happens in moderate to high dimensions. This is because the average CPU time per iteration is an average of the costs over a smoothed version of π:

$$\int \pi(x)q(x^*|x;\lambda)c(x^*)\mathrm{d}x\mathrm{d}x^*,$$

where $c(x^*)$ is the computational cost of estimating the target at x^*.

A simulation study in the literature (Sherlock *et al.*, 2015) with $d = 5$ and where the likelihood was estimated using a particle filter showed the optimal scaling to exhibit an insensitivity to the number of particles similar to, though weaker than, that predicted. For each m value, the CPU time varied with λ by less than 1% from its mean value and no trend was evident (personal communication), suggesting that the mechanism discussed above played no role in the larger-than-expected variability; we conjecture that Monte Carlo variability is at least partly responsible. A new simulation study in this article chose $d = 10$ and used importance sampling to estimate the likelihood; here both the variance and skewness of the distribution of the additive noise were shown to vary by a factor of approximately 2 over the main posterior mass, yet the insensitivity of the optimal scaling to the number of importance samples was striking.

In Section 1.1 we discussed a converse result to Theorem 1, that the optimal choice of the variance of the estimate in the log-posterior is insensitive to the MCMC algorithm and hence, for an RWM algorithm, to the choice of scaling. We noted that current theory and simulation results suggest that this holds when the noise is generated using a particle filter, so that the noise in the log-posterior is approximately Gaussian with a variance that is inversely proportional to the computational cost, but that it might not hold when importance sampling is used and where the computational cost is proportional to the number of samples. In the latter setting $m = 1$ or $m = 2$ might be optimal whatever the variance. We examined the simulation study in Sherlock *et al.* (2015), which used a particle filter, and we found evidence of this insensitivity. Curiously, in our new simulation study, which uses importance sampling and where the additive noise and the relationship between variance and computational cost do not satisfy these assumptions, the insensitivity of the optimal choice of m to λ still appears to hold. With the particular example used, there is a considerable start-up cost at each iteration of the MCMC algorithm; this is the same whatever the value of m and a comparison of timings show that it is approximately equal to the cost of subsequently generating approximately 200 samples. Hence in our example it *is* worthwhile increasing the number of importance samples beyond 1 or 2.

As pointed out by a reviewer, as the dimension of the target increases, typically, many more observations are required to maintain a certain 'tightness' in the posterior. With a relatively broad posterior and a relatively large range of possible values for x, although it need not be the case, there is far greater potential for substantial changes in the distribution of the noise in the estimate of the log-posterior as the Markov chain moves around the posterior, violating Assumption 2 of Section 1.1. One might, therefore, wish to imagine hypothetical sets of observations increasing suitably quickly in size as our series of hypothetical targets increases in dimension. In practice there is typically one data set, one associated model on which inference is to be performed, and hence once particular dimension for the parameter space. As with all limit results, when using our result in practice one hopes that the model is sufficiently close to one that is sufficiently far along our hypothetical limiting sequence for the limit to be approximately applicable. In the real examples we have studied this appears to be the case.

Acknowledgements I am grateful for the comments of two anonymous referees which undoubtedly improved the simulation study and the overall clarity of this article and increased its breadth to include detailed discussion of the converse insensitivity to that in the title.

A Proof of Proposition 1

Following (2), define the extended target as $\tilde{\pi}(x,w) := \pi(x)g(w|x)e^w$ and let

$$c := \int \mathrm{d}x\mathrm{d}w\ \tilde{\pi}(x,w)[1-\overline{\alpha}(x,w)]f(x,w)^2 \geq 0,$$

where $\overline{\alpha}(x,w)$ is the average acceptance probability from (x,w).

As in Baxendale (2005), note that for $a \geq 0$ and $b \geq 0$, $a \wedge b = \int_0^\infty \mathrm{d}t\ \mathbb{I}_{[0,a]}(t)\mathbb{I}_{[0,b]}(t)$. Denoting the pseudo-marginal MH kernel by $P(x,w;x^*,w^*)$, for any $f \in L^2(\tilde{\pi})$ we have

$$\begin{aligned}&\int \mathrm{d}x\mathrm{d}w\mathrm{d}x^*\mathrm{d}w^*\ \tilde{\pi}(x,w)P(x,w;x^*,w^*)f(x,w)f(x^*,w^*)\\&\qquad = c + \int \mathrm{d}x\mathrm{d}w\mathrm{d}x^*\mathrm{d}w^*\ g(w|x)q(x^*|x)g(w^*|x^*)\left[e^w\pi(x)\wedge e^{w^*}\pi(x^*)\right]f(x,w)f(x^*,w^*)\\&\qquad = c + \int \mathrm{d}z \int_0^\infty \mathrm{d}t\ b(t,z)^2 \geq 0,\end{aligned}$$

where

$$b(t,z) \ := \ \int \mathrm{d}x\mathrm{d}w\ g(w|x)f(x,w)r(z,x)\mathbb{I}_{[0,v\pi(x)]}(t).$$

B Proof of Lemma 1

Proof. Differentiation from the definition of f in (10) shows that

$$\frac{\partial f}{\partial \ell} = \frac{2}{\ell} f - \ell^2 \phi(\ell/2) e^{-b^2/(2\ell^2)}, \quad (19)$$

$$\frac{\partial f}{\partial b} = \frac{1}{2}\ell^2 \left[e^{b/2}\Phi(-b/\ell - \ell/2) - e^{-b/2}\Phi(b/\ell - \ell/2) \right]. \quad (20)$$

We also note that

$$\begin{aligned} e^{b/2}\Phi\left(-\frac{b}{\ell} - \frac{\ell}{2}\right) &= e^{b/2}\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{-b/\ell-\ell/2} \mathrm{d}t\ e^{-t^2/2} \\ &= \phi\left(\frac{\ell}{2}\right) e^{-b^2/(2\ell^2)} \int_0^\infty \mathrm{d}u\ e^{-u^2/2-u\ell/2} \times e^{-ub/\ell}, \end{aligned} \quad (21)$$

and similarly

$$e^{-b/2}\Phi\left(\frac{b}{\ell} - \frac{\ell}{2}\right) = \phi\left(\frac{\ell}{2}\right) e^{-b^2/(2\ell^2)} \int_0^\infty \mathrm{d}u\ e^{-u^2/2-u\ell/2} \times e^{ub/\ell}. \quad (22)$$

Proof of Part 1: combining (21) and (22) gives

$$f(b,\ell) = 2\ell^2\phi\left(\frac{\ell}{2}\right) e^{-b^2/(2\ell^2)} \int_0^\infty \mathrm{d}u\ e^{-u^2/2-u\ell/2} \times \cosh\left(ub/\ell\right).$$

Thus, $f(b,\ell) = 2\ell^2\phi\left(\frac{\ell}{2}\right) e^{-b^2/(2\ell^2)} \times I(b,\ell)$, where

$$I(b,\ell) \geq \int_0^\infty \mathrm{d}u\ e^{-u^2/2-u\ell/2} = \frac{\Phi(-\ell/2)}{\phi(\ell/2)}.$$

The result follows on dividing through by f in (19) and applying the above inequality.

Proof of Part 2: combine (10), (19) and the fact that

$$\frac{\partial^2 f}{\partial b^2} = \frac{1}{4} f - \ell\phi(\ell/2) e^{-b^2/(2\ell^2)}.$$

Proof of Part 3: this follows directly from (20).

Proof of Part 4: combining (21) and (22) gives

$$e^{b/2}\Phi\left(\frac{b}{\ell} - \frac{\ell}{2}\right) - e^{-b/2}\Phi\left(-\frac{b}{\ell} - \frac{\ell}{2}\right) = \phi\left(\frac{\ell}{2}\right) e^{-b^2/(2\ell^2)} \int_0^\infty \mathrm{d}u\, e^{-u^2/2-u\ell/2} \times \left(e^{-ub/\ell} - e^{ub/\ell}\right) < 0$$

since the integrand is negative. The result then follows from this and (20).

□

References

Andrieu, C. and Roberts, G. O. (2009). The pseudo-marginal approach for efficient Monte Carlo computations. *Ann. Statist.* **37**(2), 697–725.

Andrieu, C. and Vihola, M. (2014). Establishing some order amongst exact approximations of MCMCs. *ArXiv e-prints* .

Andrieu, C., Doucet, A. and Holenstein, R. (2010). Particle Markov chain Monte Carlo methods. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **72**(3), 269–342.

Baxendale, P. H. (2005). Renewal theory and computable convergence rates for geometrically ergodic Markov chains. *Ann. Appl. Probab.* **15**(1B), 700–738.

Beaumont, M. A. (2003). Estimation of population growth or decline in genetically monitored populations. *Genetics* **164**, 1139–1160.

Bérard, J., Del Moral, P. and Doucet, A. (2014). A lognormal central limit theorem for particle approximations of normalizing constants. *Electron. J. Probab.* **19**, no. 94, 28.

Bornn, L., Pillai, N., Smith, A. and Woodard, D. (2014). The Use of a Single Pseudo-Sample in Approximate Bayesian Computation. *ArXiv e-prints* .

Carlin, B. P. and Louis, T. A. (2009). *Bayesian methods for data analysis*. Texts in Statistical Science Series, CRC Press, Boca Raton, FL, 3rd edition.

Doucet, A., Pitt, M. K., Deligiannidis, G. and Kohn, R. (2015). Efficient implementation of Markov chain Monte Carlo when using an unbiased likelihood estimator. *Biometrika* **102**(2), 295–313.

Filippone, M. and Girolami, M. (2014). Pseudo-marginal Bayesian inference for Gaussian processes. *IEEE Tran. Pattern Anal. Mach. Intell.* **36**(11), 2214–2226.

Giorgi, E., Sesay, S. S. S. and Terlouw, D. J. D. P. J. (2015). Combining data from multiple spatially referenced prevalence surveys using generalized linear geostatistical models. *Statistics in Society* **178**(2), 445–464.

Golightly, A. and Wilkinson, D. J. (2011). Bayesian parameter inference for stochastic biochemical network models using particle Markov chain Monte Carlo. *Interface Focus* **1**(6), 807–820.

Haran, M. and Tierney, L. (2012). On automating Markov chain Monte Carlo for a class of spatial models. *arXiv preprint arXiv:1205.0499* , 1–26.

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

Lampaki, I. (2015). Markov chain Monte Carlo methodology for inference on generalised linear spatial models PhD Thesis; in preparation.

Pitt, M. K., dos Santos Silva, R., Giordani, P. and Kohn, R. (2012). On some properties of Markov chain Monte Carlo simulation methods based on the particle filter. *Journal of Econometrics* **171**(2), 134 – 151.

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

Sherlock, C., Golightly, A. and Gillespie, C. (2014). Bayesian inference for hybrid discrete-continuous stochastic-kinetic models. *Inverse Problems* **30**(11), 1–22.

Sherlock, C., Thiery, A. H., Roberts, G. O. and Rosenthal, J. S. (2015). On the efficiency of pseudo-marginal random walk Metropolis algorithms. *Ann. Statist.* **43**(1), 238–275.