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

Chris Sherlock

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 distributions for which we rigorously prove it. Finally we conduct a simulation study on a target and family of noise distributions which together satisfy neither of the two key conditions of the limit result on which our work is based: the target has $d = 1$ and the noise distribution depends heavily on the position in the state-space. Our key conclusions are found to hold in this example also.

Classification: 65C05, 65C40.

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

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. The proposal density for

\[^1\text{Department of Mathematics and Statistics, Lancaster University, Lancaster LA1 4YF, UK. c.sherlock@lancaster.ac.uk}\]
the noise, \( g(w|x), w \in (-\infty, \infty) \) must possess the property that \( \int_{-\infty}^{\infty} dw \ e^w g(w|x) = c > 0 \). 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’.

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, \( \hat{\pi}_W^*(x^*) = \pi(x^*) e^{w^*} \) 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 \( 1 \wedge \frac{\hat{\pi}_W^*(x^*) q(x|x^*)}{\hat{\pi}_W(x) q(x^*|x)} \). 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

\[
\pi(x) g(w|x) e^w, \tag{1}
\]

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 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 \( 1 \wedge \frac{\hat{\pi}_W^*(x^*)}{\hat{\pi}_W(x)} \). It is one of the most popular forms of PsMMH (e.g. Golightly and Wilkinson, 2011; Knape and de Valpine, 2012; Sherlock et al., 2014a) because it does not require the computation or estimation of other properties of the posterior, such as 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. For a particular \( m^* \), a practitioner might find, using repeated runs, the optimal scaling, \( \hat{\lambda}^* \), that is 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.

We consider the form of efficiency derived in Sherlock et al. (2014b), which is valid in the limit as the dimension of the target approaches infinity. Provided that across the range of \( m \) values to be considered the noise density, \( g \), 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 \( \lambda \) and \( m \) values that approximately maximise the efficiency can therefore effectively be reduced to two one-dimensional optimisation problems.
The main theoretical result of this article, Theorem 1, is stated and proved in Section 2. Given that \( \int dw \, g(w)e^w \) 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 details a simulation study on a one-dimensional target over a family of noise generating mechanisms, where the distribution of the noise depends heavily on the position in the target and yet the insensitivity described in Theorem 1 is still found to hold.

2 Set-up and main theoretical result

In Sherlock et al. (2014b) the efficiency of the RWM is examined in terms of expected squared jumping distance (ESJD) on a sequence of targets of increasing dimension. The scaling, \( \lambda^{(d)} \), of the corresponding sequence of Gaussian proposals is assumed to be some factor \( \ell \) multiplied by a given function of dimension which depends on the sequence of targets but not on the distributions of the noise in the unbiased estimates thereof. The noise distribution is also assumed to be independent of the position in the target, \( g(w|x) = g(w) \). Subject to further technical conditions on the sequence of targets it is shown that the limiting ESJD has the form

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

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 \( \Phi \) represents the cumulative distribution function of a standard Gaussian random variable. Subject to further conditions on the sequence of targets, a scaled version of the first component is shown to converge in law to a diffusion, the speed of which is proportional to \( J_m(\ell) \).

Our main result refers to the situation when there is no noise in the estimate of \( \pi \), \( B = 0 \), when the limiting ESJD simplifies to

\[
J_\infty(\ell) = 2\ell^2 \Phi \left( -\frac{\ell}{2} \right).
\]

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 (2) is particularly tractable and Sherlock et al. (2014b) 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 (2) and (3). 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 \( \hat{\ell}_m \leq 2\sqrt{2} \).
3. For any two scalings, \( \ell_1 \) and \( \ell_2 \), 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 (1) 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

\[
h(b) := e^{b/2} \rho(b) = \int_{-\infty}^{\infty} dw \, g(w) e^{b/2+w} g(w+b) = \int_{-\infty}^{\infty} dw \, g(w+b/2) g(w-b/2) e^w
\]

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

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

Using (2) and (4), the squared jumping distance is

\[
J(\ell) = 2\ell^2 \int_{-\infty}^{\infty} db \, \rho(b) \, \Phi \left( \frac{b}{\ell} - \frac{\ell}{2} \right) = 2\ell^2 \int_{-\infty}^{\infty} db \, h(b) \, e^{-b/2} \Phi \left( \frac{b}{\ell} - \frac{\ell}{2} \right)
\]

\[
= 2 \int_{0}^{\infty} db \, h(b) \, f(b, \ell),
\]
by the symmetry of \( h \). From (3), straightforward differentiation gives:

\[
\frac{d}{d\ell}(\log J_\infty) = \frac{2}{\ell} - \frac{\phi(\ell/2)}{2\Phi(-\ell/2)}, \tag{7}
\]

\[
\frac{d^2}{d\ell^2}(\log J_\infty) = -\frac{2}{\ell^2} - \frac{\phi(\ell/2)}{4\Phi(-\ell/2)^2} \left[ \phi(\ell/2) - \frac{\ell}{2} \Phi(-\ell/2) \right] < 0 \forall \ell > 0, \tag{8}
\]

so that (for \( \ell > 0 \)) \( J_\infty \) 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 A.

**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 (7) gives \( f \frac{d \log J_\infty}{d\ell} < \partial f / \partial \ell < 2f / \ell \). Multiplying by \( h \), which is non-negative, integrating and using (6) we then obtain

\[
\frac{d}{d\ell}(\log J_\infty) < \frac{d}{d\ell}(\log J) < \frac{2}{\ell}. \tag{9}
\]

We now proceed with the proof of Theorem 1.

**Proof of Part 1 of Theorem 1:** by (8), for \( \ell < \hat{\ell}_\infty \), \( dJ_\infty / d\ell > 0 \) and so \( d \log J_\infty / d\ell > 0 \). The result then follows from (9).

**Proof of Part 2 of Theorem 1:** from the definition in (4),

\[
\frac{\partial h}{\partial b} = \frac{1}{2} \int_{-\infty}^{\infty} dw \left( g(w - b/2)g(w + b/2)e^w \left( \frac{g'(w + b/2)}{g(w + b/2)} - \frac{g'(w - b/2)}{g(w - b/2)} \right) \right) \leq 0 \text{ for } b \geq 0 \tag{10}
\]

by the log-concavity of \( g \).
Furthermore, $\exists g$ s.t. $g(w) \leq \overline{g} < \infty$ (since $g$ is a log-concave density) and hence by (4)

$$h(0) \leq \overline{g} \int_{-\infty}^{\infty} dw \ g(w)e^{w} = \overline{g},$$

(11)

$$h(b) \leq \overline{g} \int_{-\infty}^{\infty} dw \ g(w + b)e^{b/2 + w} = \overline{g} e^{-b/2} \int_{-\infty}^{\infty} dw \ g(w)e^{w} = \overline{g} e^{-b/2}.$$ (12)

By Part 2 of Lemma 1,

$$\frac{dJ}{d\ell} = 2\ell \int_{0}^{\infty} db \ h(b) \frac{\partial^2 f}{\partial b^2} + \left( \frac{4}{\ell} - \frac{\ell}{2} \right) \int_{0}^{\infty} db \ h(b) \ f(b, \ell).$$ (13)

The first term is

$$2\ell \left[ h(b) \frac{\partial f}{\partial b} \right]_{0}^{\infty} - 2\ell \int_{0}^{\infty} db \ \frac{\partial h}{\partial b} \ \frac{\partial f}{\partial b}.$$

Now $\left[ h(b) \frac{\partial f}{\partial b} \right]_{0}^{\infty} = 0$ by (11), (12) 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 (10); thus the first term in (13) cannot be positive. The second term in (13) is guaranteed to be negative provided $\ell > 2\sqrt{2}$.

**Proof of Part 3 of Theorem 1:** by (8), for $\ell \in [\hat{\ell}_{\infty}, 2\sqrt{2}]$ (and, indeed, above this), the lower bound in (9) 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
Bernoulli noise

$\epsilon$

$p^*$

0.0 0.2 0.4 0.6 0.8 1.0

0.0 0.2 0.4 0.6 0.8 1.0

Figure 1: Optimal scaling for the Bernoulli noise model as a function of the two noise parameters, $\epsilon$ and $p^*$.

$b \geq 0$. However it is clear from the proof that a weaker (yet still sufficient) condition for the upper bound is

$$\int_0^\infty db \frac{\partial h}{\partial b} \frac{\partial f}{\partial b} > 0.$$  

Clearly there is scope for $\partial h/\partial b > 0$ over some regions whilst the whole expression in (13) 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 $\epsilon \in (0,1)$:

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

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

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

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

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 $\ell > 2\sqrt{2}$ and we conjecture that $\ell \leq 2\sqrt{2}$ whatever distribution of $W$.

### 4 Simulation study

The purpose of this study is to examine the optimal scaling when the conditions in Sherlock et al. (2014b) do not hold. In particular we require a low-dimensional setting where the distribution of $W$ depends strongly on the position in the statespace. We therefore choose a one-dimensional target which we make Gaussian for simplicity and so that at least the tails of the (marginal) target are not an issue in empirical estimation of efficiency (see below). We use a Gaussian random-walk proposal and a Gaussian distribution for the additive noise in the estimate of the log-target as follows:

$$
\pi(\theta) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} \theta^2\right),
q(\theta^*|\theta) \sim N(\theta, \gamma^2 V),
W^*|\theta^* \sim N\left(-\frac{1}{2} \sigma^2, \sigma^2\right), \text{ where } \sigma^2 = \phi_1 \exp \theta + \phi_2.
$$

Following Sherlock et al. (2014b) $V$ is set to $2.56^2$. A separate MCMC run of $2 \times 10^6$ iterations was performed for every combination of $\gamma \in \mathcal{G} := \{0.1, 0.2, \ldots, 1.9, 2.0\}$ and $\phi \in \mathcal{F} := \{(0, 0), (\frac{1}{2}, 0), (1, 0), (0, 1), (\frac{1}{2}, 1), (1, 1)\}$.

The efficiency function in (2) is proportional to the limiting ESJD, which is equivalent to the speed of the limiting diffusion that was shown to exist in Sherlock et al. (2014b). We consider two empirical measures of efficiency: the effective sample size (ESS) and the mean-square jumping distance (MSJD). In the presence of a limiting diffusion the two are equivalent in high dimensions. MSJD only takes into account the lag-1 autocorrelation, however in low dimensions all of the lagged auto-correlations can be important, and the ESS accounts for these. Thus ESS is the more desirable measure of efficiency; however empirical estimation of the ESS is potentially subject to large Monte Carlo variability (e.g Sokal, 1997). Furthermore, since it is based upon the ESJD as $d \to \infty$, in lower dimensions (2) is likely to be more representative of MSJD than ESS. Any discrepancy between the two measures is therefore of interest.

For each $\phi \in \mathcal{F}$ and $\gamma \in \mathcal{G}$ we define the relative efficiency $ESS_{\gamma, \phi} := ESS_{\gamma, \phi}/\max_{\gamma \in \mathcal{G}} ESS_{\gamma, \phi}$. We define the relative MSJD, $MSJD$, equivalently. The left-hand panel of Figure 2 shows
ESS as a function of $\gamma$ for each value of $\phi$. The right-hand panel shows the analogous plot for MSJD. As expected, there is considerable Monte Carlo variability in the estimates of ESS; nonetheless, it is clear that for each of the scenarios considered the optimal choice of scaling, $\hat{\gamma}$, lies in $[0, 1.6]$. More importantly, for all choices of $\gamma$ in this range the relative efficiency for each curve remains above the dotted line, which represents 70% of the estimated maximum achievable efficiency.

Figure 2 also appears to show that as the amount of noise increases, the optimal scaling actually decreases slightly, rather than increasing slightly, as suggested by Part 1 of Theorem 1. The graph of MSJD suggests $\hat{\gamma} \in [0.8, 1.0]$, and shows an even more subtle decrease in the optimal scaling. Fortunately there is very little Monte Carlo variability at all in the estimates of MSJD and on examining the numerical values it is possible to discern that the $\gamma$ value which optimises the MSJD is 1.0 for $\phi = (0, 0)$ (no noise) but is 0.9 or 0.8 for all other scenarios, including $\phi = (0, 1)$, a noise scenario which is covered by Sherlock et al. (2014b). Hence this ‘decreasing scaling’ phenomenon must be due to the low dimension of our target. To confirm this, an identical set of experiments was performed on a standard multivariate Gaussian target with $d = 10$, $V = 2.56^2/d$ $I$ (where $I$ is the $d \times d$ identity matrix) and $\sigma^2 = \phi_2 + \frac{\phi_1}{d} \sum_{i=1}^{d} \exp \theta_i$. Graphs of the mean (over components) ESS showed a similar shape to those with $d = 1$, and less Monte Carlo variability; they were maximised at $\gamma = 0.8$ for $\phi \in \{(0, 0), (\frac{1}{2}, 0)\}$ and $\gamma = 0.9$ for the remainder of the $\phi$ values. The mean
MSJD was maximised at $\gamma = 0.9$ for $\phi = (0, 0)$, $\gamma = 1.0$ for $\phi \in \{(\frac{1}{2}, 0), (1, 0), (0, 1)\}$ and $\gamma = 1.1$ for $\phi \in \{(\frac{1}{2}, 1), (1, 1)\}$. Both measures suggest an increase in the optimal scaling as the amount of noise increases.

A Proof of Lemma 1

Proof. Differentiation from the definition of $f$ in (5) shows that

$$\frac{\partial f}{\partial \ell} = \frac{2}{\ell} f - \ell^2 \phi(\ell/2) e^{-\ell^2/(2\ell^2)}, \quad (14)$$

$$\frac{\partial f}{\partial b} = \frac{1}{2} \ell^2 \left[ e^{b/2} \Phi(-b/\ell - \ell/2) - e^{-b/2} \Phi(b/\ell - \ell/2) \right]. \quad (15)$$

We also note that

$$e^{b/2} \Phi \left( -\frac{b}{\ell} - \frac{\ell}{2} \right) = e^{b/2} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-b/\ell - \ell/2} dt \ e^{-t^2/2} = \phi \left( \frac{\ell}{2} \right) e^{-b^2/(2\ell^2)} \int_0^\infty du \ e^{-u^2/2-ut/2} \times e^{-ub/\ell}, \quad (16)$$

and similarly

$$e^{-b/2} \Phi \left( \frac{b}{\ell} - \frac{\ell}{2} \right) = \phi \left( \frac{\ell}{2} \right) e^{-b^2/(2\ell^2)} \int_0^\infty du \ e^{-u^2/2-ut/2} \times e^{ub/\ell}. \quad (17)$$

Proof of Part 1: combining (16) and (17) gives

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

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

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

The result follows on dividing through by $f$ in (14) and applying the above inequality.

Proof of Part 2: combine (5), (14) 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 (15).
Proof of Part 4: combining (16) and (17) gives

\[ e^{b/2} \Phi \left( \frac{b}{\ell} - \frac{\ell}{2} \right) - e^{-b/2} \Phi \left( - \frac{b}{\ell} - \frac{\ell}{2} \right) = \phi \left( \frac{\ell}{2} \right) e^{-b^2/(2\ell^2)} \int_{0}^{\infty} du \ e^{-u^2/2 - u\ell/2} \times \left( e^{-ub/\ell} - e^{ub/\ell} \right) < 0 \]

since the integrand is negative. The result then follows from this and (15).

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

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

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.

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.

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. (2014a). Bayesian inference for hybrid discrete-continuous stochastic-kinetic models. *Inverse Problems* To appear.

Sherlock, C., Thiery, A., Roberts, G. O. and Rosenthal, J. S. (2014b). On the efficiency of pseudo-marginal random walk Metropolis algorithms. *arXiv preprint arXiv:1309.7209v2*.

Sokal, A. (1997). Monte Carlo methods in statistical mechanics: Foundations and new algorithms. In: *Functional Integration* (eds. C. DeWitt-Morette, P. Cartier and A. Folacci), volume 361 of *NATO ASI Series*, Springer US, 131–192.