


OPTIMAL SCALING OF MCMC BEYOND METROPOLIS

SANKET AGRAWAL,* *University of Warwick*
DOOTIKA VATS ,** *Indian Institute of Technology Kanpur*
KRZYSZTOF ŁATUSZYŃSKI* AND
GARETH O. ROBERTS,* *University of Warwick*

Abstract

The problem of optimally scaling the proposal distribution in a Markov chain Monte Carlo algorithm is critical to the quality of the generated samples. Much work has gone into obtaining such results for various Metropolis–Hastings (MH) algorithms. Recently, acceptance probabilities other than MH are being employed in problems with intractable target distributions. There are few resources available on tuning the Gaussian proposal distributions for this situation. We obtain optimal scaling results for a general class of acceptance functions, which includes Barker’s and lazy MH. In particular, optimal values for Barker’s algorithm are derived and found to be significantly different from that obtained for the MH algorithm. Our theoretical conclusions are supported by numerical simulations indicating that when the optimal proposal variance is unknown, tuning to the optimal acceptance probability remains an effective strategy.

Keywords: Barker’s acceptance; weak convergence; Metropolis–Hastings; lazy MH; tuning

2020 Mathematics Subject Classification: Primary 65C05
Secondary 60F05

1. Introduction

Over the past few decades, Markov chain Monte Carlo (MCMC) methods have become an abundantly popular computational tool, enabling practitioners to conveniently sample from complicated target distributions [5, 21, 26]. This popularity can be attributed to easy-to-implement accept–reject-based MCMC algorithms for target densities available only up to a proportionality constant. Here, draws from a proposal kernel are accepted with a certain *acceptance probability*. The choice of the acceptance probability and the proposal kernel can yield varying performances of the MCMC samplers.

Unarguably, the most popular acceptance probability is Metropolis–Hastings (MH), of [14, 20], owing to its acknowledged optimality [4, 25]. Efficient implementation of the MH algorithm requires tuning within the chosen family of proposal kernels. For the MH acceptance function, various optimal scaling results have been obtained under assumptions on the proposal and the target distribution. This includes the works of [3, 24, 27, 28, 29, 34, 40, 41], among others.

Received 5 April 2021; revision received 17 June 2022.

* Postal address: Coventry CV4 7AL, U.K.

** Email address: dootika@iitk.ac.in

© The Author(s), 2022. Published by Cambridge University Press on behalf of Applied Probability Trust.

Despite the popularity of the MH acceptance function, other acceptance probabilities remain practically and theoretically relevant. Recently, Barker’s acceptance rule [2] and lazy MH [18] have found use in Bernoulli-factory-based MCMC algorithms for intractable posteriors [12, 13, 15, 37, 39]. Barker’s acceptance function has also proven to be optimal with respect to search efficiency [19], and it guarantees variance improvements for waste-recycled Monte Carlo estimators [7]. Further, a class of acceptance probabilities from [3] has been of independent theoretical interest. We also introduce a new family of *generalized Barker’s* acceptance probabilities and present a Bernoulli factory for use in problems with intractable posteriors.

To the best of our knowledge, there are no theoretical and practical guidelines concerning optimal scaling outside of MH and its variants (although see [35] for a discussion on delayed-acceptance MH and [8, 32, 36] for analyses pertaining to pseudo-marginal MCMC). We obtain optimal scaling results for a large class of acceptance functions; Barker’s, lazy MH, and MH are members of this class.

We restrict our attention to the framework of [27] with a random-walk Gaussian proposal kernel and a d -dimensional decomposable target distribution. Similar to MH, our general class of acceptance functions require the proposal variance to be scaled by $1/d$. We find that, typically, for lower acceptance functions, the optimal proposal variance is larger than the optimal proposal variance for MH, implying the need for larger jumps. For Barker’s acceptance rule, the asymptotically optimal acceptance rate (AOAR) is approximately 0.158, in comparison to the 0.234 rate for MH [27]. Similar AOARs are presented for other acceptances.

In Section 2 we describe our class of acceptance probabilities, with the main results presented in Section 3. AOARs for Barker’s and other functions are obtained in Section 3.1. In Section 4 we present numerical results in some settings that comply with our assumptions and others that do not. A trailing discussion on the scaling factor for different acceptance functions and generalizations of our results is provided in the last section. All proofs are in the appendices.

2. Class of acceptance functions

Let π be the target distribution, with corresponding Lebesgue density π and support \mathcal{X} , so that an MCMC algorithm aims to generate a π -ergodic Markov chain, $\{X_n\}$. Let Q be a Markov kernel with an associated Lebesgue density $q(x, \cdot)$ for each $x \in \mathcal{X}$. We assume throughout that q is symmetric. Furthermore, let the acceptance probability function be $\alpha(x, y) : \mathcal{X} \times \mathcal{X} \rightarrow [0, 1]$. Starting from an $X_0 \in \mathcal{X}$, at the n th step, a typical accept–reject MCMC algorithm proposes $y \sim q(X_{n-1}, \cdot)$. The proposed value is accepted with probability $\alpha(X_{n-1}, y)$ and rejected otherwise, implying that $X_n = X_{n-1}$. The acceptance function α is responsible for guaranteeing π -reversibility and thus π -invariance of the Markov chain. Let $a \wedge b$ denote $\min(a, b)$ and $s(x, y) = \pi(y)/\pi(x)$. We define \mathcal{A} , the class of acceptance functions for which our optimal scaling results will hold, as follows.

Definition 1. Each $\alpha \in \mathcal{A}$ is a map $\alpha(x, y) : \mathcal{X} \times \mathcal{X} \rightarrow [0, 1]$, and for every $\alpha \in \mathcal{A}$ there exists a *balancing* function $g_\alpha : [0, \infty) \rightarrow [0, 1]$ such that

$$\alpha(x, y) = g_\alpha(s(x, y)), \quad x, y \in \mathcal{X}, \tag{1}$$

$$g_\alpha(z) = zg_\alpha\left(\frac{1}{z}\right), \quad 0 \leq z < \infty, \tag{2}$$

$$g_\alpha(e^z), \quad z \in \mathbb{R}, \text{ is Lipschitz continuous.} \tag{3}$$

Properties (1) and (2) are standard and easy to verify, with (1) ensuring that intractable constants in π cancel away and (2) ensuring π -reversibility. Property (3) is not required for α to be a valid acceptance function; however, we need it for our optimal scaling results (to establish Lemma 4), and it holds true for all common acceptance probabilities. Moreover, each $\alpha \in \mathcal{A}$ can be identified by the corresponding g_α , and we will use α and g_α interchangeably. If g_{MH} denotes the balancing function for the MH acceptance function (α_{MH}), then

$$g_{MH}(z) = 1 \wedge z, \quad z \geq 0. \tag{4}$$

It is easy to see that $\alpha_{MH} \in \mathcal{A}$. The lazy MH (α_L) acceptance of [15, 18] also belongs to \mathcal{A} . For a fixed $\varepsilon \in [0, 1]$, it is defined using

$$g_L(z) = (1 - \varepsilon)(1 \wedge z), \quad z \geq 0. \tag{5}$$

Barker’s acceptance function is $\alpha_B(x, y) = g_B(s(x, y))$ for all $x, y \in \mathcal{X}$, where

$$g_B(z) = \frac{z}{1 + z}, \quad z \geq 0. \tag{6}$$

Then (2) follows immediately. For differentiable functions, Property (3), i.e. Lipschitz continuity of $g_\alpha(e^z)$, can be verified by bounding the first derivative. In particular, the derivative of $g_B(e^z)$, given by $e^z/(1 + e^z)^2$, is bounded by 1/4 for all $z \in \mathbb{R}$, and hence $\alpha_B \in \mathcal{A}$. From [25], it is well known that in the context of Monte Carlo variability of ergodic averages, MH is superior to Barker’s. Even so, Barker’s acceptance function has had a recent resurgence, aided by its use in Bernoulli-factory MCMC algorithms for Bayesian intractable posteriors where MH algorithms are not implementable.

We present a generalization of (6): for $r \geq 1$ define

$$g_r^R(z) = \begin{cases} \frac{z(z^r - 1)}{z^{r+1} - 1}, & z \neq 1, \\ \frac{r}{r + 1}, & z = 1. \end{cases}$$

For $r \in \mathbb{N}$, the above can be rewritten as

$$g_r^R(z) = \frac{z + \dots + z^r}{1 + z + \dots + z^r}, \quad z \geq 0, \quad r \in \mathbb{N}. \tag{7}$$

If α_r^R is the associated acceptance function, then $\alpha_r^R \in \mathcal{A}$ for all $r \geq 1$. Moreover, $g_1^R \equiv g_B$ and $g_r^R \uparrow g_{MH}$ as $r \rightarrow \infty$. For $r \in \mathbb{N}$, we present a natural Bernoulli factory in the spirit of [13] that generates events of probability α_r^R without explicitly evaluating it; see Appendix D. An alternative approach would be to follow the general sampling algorithm of [23] for rational functions.

Let $\Phi(\cdot)$ be the standard normal distribution function. For a theoretical exposition, [3] defines the following acceptance probability for some $h > 0$:

$$g_h^H(z) = \Phi\left(\frac{\log z - h/2}{\sqrt{h}}\right) + z \cdot \Phi\left(\frac{-\log z - h/2}{\sqrt{h}}\right), \quad z \geq 0. \tag{8}$$

For each $h > 0$, $\alpha_h^H \in \mathcal{A}$, and observe that as $h \rightarrow 0$, $g_h^H \rightarrow g_{MH}$, while as $h \rightarrow \infty$, $g_h^H \rightarrow 0$; i.e. the chain never moves. Similar examples can be constructed by considering other well-behaved distribution functions in place of Φ . Lastly, it is easy to see that \mathcal{A} is convex. Thus, it also includes situations when each update of the algorithm randomly chooses an acceptance probability. Moreover, as evidenced in (5), \mathcal{A} is also closed under scalar multiplication as long as the resulting function lies in $[0, 1]$.

3. Main theorem

Let f be a 1-dimensional density function and consider a sequence of target distributions $\{\pi_d\}$ such that for each d , the joint density is

$$\pi_d(\mathbf{x}^d) = \prod_{i=1}^d f(x_i^d), \quad \mathbf{x}^d = (x_1^d, \dots, x_d^d)^T \in \mathbb{R}^d.$$

Assumption 1. *The density f is positive and in C^2 —the class of all real-valued functions with continuous second-order derivatives. Furthermore, f'/f is Lipschitz, and the following moment conditions hold:*

$$\mathbb{E}_f \left[\left(\frac{f'(X)}{f(X)} \right)^8 \right] < \infty, \quad \mathbb{E}_f \left[\left(\frac{f''(X)}{f(X)} \right)^4 \right] < \infty. \tag{9}$$

Consider the sequence of Gaussian proposal kernels $\{Q_d(\mathbf{x}^d, \cdot)\}$ with associated density sequence $\{q_d\}$, so that $Q_d(\mathbf{x}^d, \cdot) = N(\mathbf{x}^d, \sigma_d^2 \mathbf{I}_d)$, where for some constant $l \in \mathbb{R}^+$,

$$\sigma_d^2 = l^2 / (d - 1).$$

The proposal Q_d is used to generate a d -dimensional Markov chain, $\mathbf{X}^d = \{\mathbf{X}_n^d, n \geq 0\}$, following the accept-reject mechanism with acceptance function α . Under these conditions and with $\alpha = \alpha_{MH}$, [27] established weak convergence to an appropriate Langevin diffusion for the sequence of 1-dimensional stochastic processes constructed from the first component of these Markov chains. Since the coordinates are independent and identically distributed (i.i.d.), this limit informs the limiting behaviour of the full Markov chain in high dimensions. In what follows, we extend the results of [27] to the class of acceptance functions \mathcal{A} as defined in Definition 1.

Let $\{\mathbf{Z}^d, d > 1\}$ be a sequence of processes constructed by speeding up the Markov chains by a factor of d as follows:

$$\mathbf{Z}_t^d = \mathbf{X}_{[dt]}^d = \left(X_{[dt],1}^d, X_{[dt],2}^d, \dots, X_{[dt],d}^d \right)^T; \quad t > 0.$$

Suppose $\{\eta_d : \mathbb{R}^d \rightarrow \mathbb{R}\}$ is a sequence of projection maps such that $\eta_d(\mathbf{x}^d) = x_1^d$. Define a new sequence of 1-dimensional processes $\{U^d, d > 1\}$ as follows:

$$U_t^d := \eta_d \circ \mathbf{Z}_t^d = X_{[dt],1}^d; \quad t > 0.$$

Under stationarity, we show that $\{U^d, d > 1\}$ weakly converges in the Skorokhod topology [10] to a Markovian limit U . We denote weak convergence of processes in the Skorokhod topology by ‘ \Rightarrow ’ and standard Brownian motion at time t by B_t . The proofs are in the appendices.

Theorem 1. *Let $\{\mathbf{X}^d, d \geq 1\}$ be the sequence of π_d -invariant Markov chains constructed using the acceptance function α and proposal Q_d such that $\mathbf{X}_0^d \sim \pi_d$. Further, suppose $\alpha \in \mathcal{A}$ and π_d satisfies Assumption 1. Then $U^d \Rightarrow U$, where U is a diffusion process that satisfies the Langevin stochastic differential equation,*

$$dU_t = (h_\alpha(t))^{1/2} dB_t + h_\alpha(t) \frac{f'(U_t)}{2f(U_t)} dt,$$

with $h_\alpha(l) = l^2 M_\alpha(l)$, where

$$M_\alpha(l) = \int_{\mathbb{R}} g_\alpha(e^b) \frac{1}{\sqrt{2\pi l^2 I}} \exp\left\{-\frac{(b + l^2 I/2)^2}{2l^2 I}\right\} db \tag{10}$$

and

$$I = \mathbb{E}_f \left[\left(\frac{f'(X)}{f(X)} \right)^2 \right].$$

Remark 1. Since $\alpha_{MH} \in \mathcal{A}$, our result aligns with [27], because

$$M_{MH}(l) = \int_{\mathbb{R}} g_{MH}(e^b) \frac{1}{\sqrt{2\pi l^2 I}} \exp\left\{-\frac{(b + l^2 I/2)^2}{2l^2 I}\right\} db = 2\Phi\left(-\frac{l\sqrt{I}}{2}\right).$$

Remark 2. For symmetric proposals, Definition 1 requires α to be a function of only the ratio of the target densities at the two contested points. Thus, the result is not applicable to acceptances in [1, 22, 39].

In Theorem 1, $h_\alpha(l)$ is the speed measure of the limiting diffusion process and so the optimal choice of l is l^* such that

$$l^* = \arg \max_l h_\alpha(l).$$

Denote the average acceptance probability by

$$\alpha_d(l) := \mathbb{E}_{\pi_d, Q_d} [\alpha(\mathbf{X}^d, \mathbf{Y}^d)] = \int \int \pi(\mathbf{x}^d) \alpha(\mathbf{x}^d, \mathbf{y}^d) q_d(\mathbf{x}^d, \mathbf{y}^d) d\mathbf{x}^d d\mathbf{y}^d,$$

and the asymptotic acceptance probability by $\alpha(l) := \lim_{d \rightarrow \infty} \alpha_d(l)$. The dependence on l is through the variance of the proposal kernel. We then have the following corollary.

Corollary 1. Under the setting of Theorem 1, we obtain $\alpha(l) = M_\alpha(l)$, and the asymptotically optimal acceptance probability is $M_\alpha(l^*)$.

Corollary 1 is of considerable practical relevance, since for different acceptance functions it yields the optimal target acceptance probability to tune to.

3.1. Optimal results for some acceptance functions

In Section 2, we discussed some important members of the class \mathcal{A} . Corollary 1 can then be used to obtain the AOAR for them by maximizing the speed measure of the limiting diffusion process. For Barker’s algorithm, from Theorem 1 and (6), the speed measure $h_B(l)$ of the corresponding limiting process is $h_B(l) = l^2 M_B(l)$, where

$$M_B(l) = \int_{\mathbb{R}} \frac{1}{1 + e^{-b}} \frac{1}{\sqrt{2\pi l^2 I}} \exp\left\{-\frac{(b + l^2 I/2)^2}{2l^2 I}\right\} db.$$

Maximizing $h_B(l)$, the optimal value l^* is approximately (see Appendix C)

$$l^* = \frac{2.46}{\sqrt{I}}.$$

By Corollary 1, using this l^* yields an asymptotic acceptance rate of approximately 0.158. Hence, when the optimal variance is not analytically tractable in high dimensions, one may

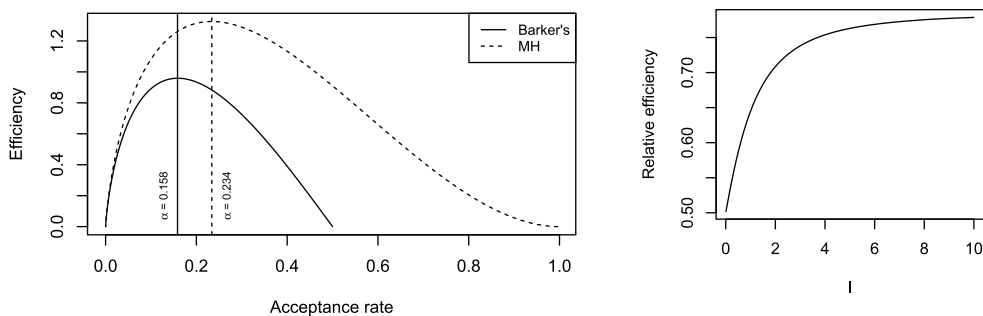


FIGURE 1. Efficiency $(h(l))$ versus acceptance rate $(\alpha(l))$ with $l = 1$ (left). Relative efficiency of Barker's versus MH $(h_B(l)/h_{MH}(l))$, plotted against l (right).

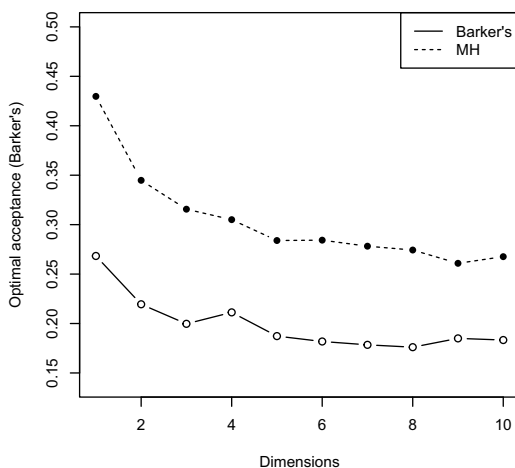


FIGURE 2. Optimal acceptance rate against number of dimensions.

consider tuning the algorithm so as to achieve an acceptance probability of approximately 0.158. Additionally, the right plot in Figure 1 verifies that the relative efficiency of Barker's versus MH, as measured by the ratio of their respective speed measures for a fixed l , remains above 0.5 [18, Theorem 4]; this relative efficiency increases as l increases. The ratio of the speed measures of Barker's versus MH at their respective optimal scaling is 0.72. This quantifies the loss in efficiency in running the best version of Barker's compared to the best version of the MH algorithm. We can also study the respective speed measures as a function of the acceptance rate; this is given in the left plot in Figure 1. We find that as the asymptotic acceptance rate increases, the speed measure for Barker's decreases more rapidly than that of MH. This suggests that there is much to gain by appropriately tuning Barker's algorithm.

For lower dimensions, the optimal acceptance rate is higher than the AOAR. Figure 2 shows optimal values for MH and Barker's algorithms on isotropic Gaussian targets in dimensions 1 to 10, the proposal kernel being the same as in the setting of Theorem 1. This plot is produced using the criterion of minimizing first-order auto-correlations in each component [11, 28, 29]. For α_{MH} and α_B , the optimal acceptance rates in one dimension are 0.43 and 0.27 respectively.

TABLE 1. Optimal proposal variance and asymptotic acceptance rates.

	α_{MH}	α_1^H	$\alpha_{1.913}^H$	α_5^H	α_{10}^R	α_5^R	α_2^R	α_B
$M_\alpha(l^*)$	0.234	0.189	0.158	0.129	0.229	0.223	0.197	0.158
$ l^* \sqrt{I} $	2.38	2.43	2.46	2.49	2.39	2.39	2.42	2.46

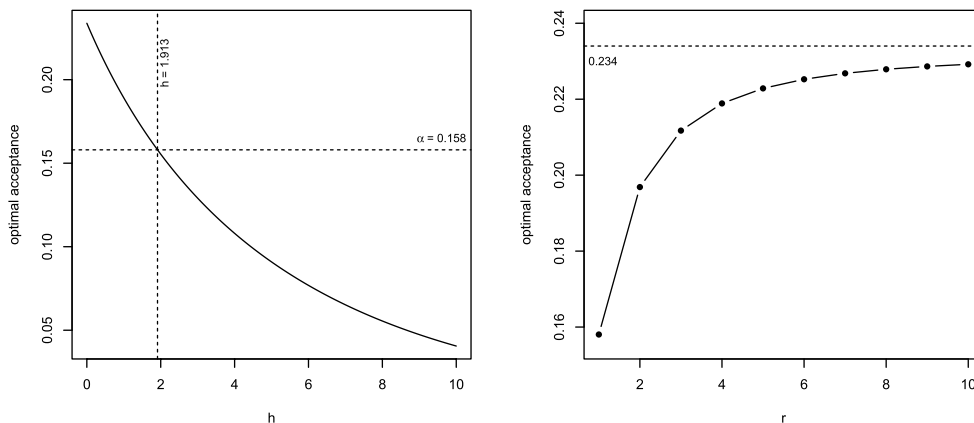


FIGURE 3. Optimal acceptance rates for α_h^H against h (left) and α_r^R against r (right).

For lazy MH with $\varepsilon \in [0, 1]$, Corollary 1 implies that the AOAR of the algorithm is $(1 - \varepsilon)0.234$ with the same optimal l^* as MH. For the acceptance functions α_h^H in (8),

$$M_h(l) = 2\Phi\left(-\frac{\sqrt{h + l^2 I}}{2}\right).$$

With $h = 0$, we obtain the result of [27] for MH. Further, the left panel of Figure 3 highlights that as $h \rightarrow 0$, the AOAR increases to 0.234 and the algorithm worsens as h increases. Moreover, for $h \approx 1.913$, the AOAR is roughly 0.158, i.e. equivalent to Barker’s acceptance function.

Lastly, the AOARs for α_r^R in (7) are available. For $r = 1, \dots, 10$, the results have been plotted in the right plot of Figure 3. As anticipated, the AOAR approaches 0.234 as r increases. Notice that α_2^R yields an AOAR of 0.197, which is a considerable increase from $\alpha_B = \alpha_1^R$. Table 1 below summarizes the results of this section. (Code for all plots and tables is available at <https://github.com/Sanket-Ag/BarkerScaling>.)

4. Numerical results

We study the estimation quality for different expectations as a function of the proposal variance (acceptance rate) for the generalized Barker acceptance function, α_r^R . We focus on $r = 1$ (Barker’s algorithm) and $r = 2$. Suppose $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is the function whose expectation

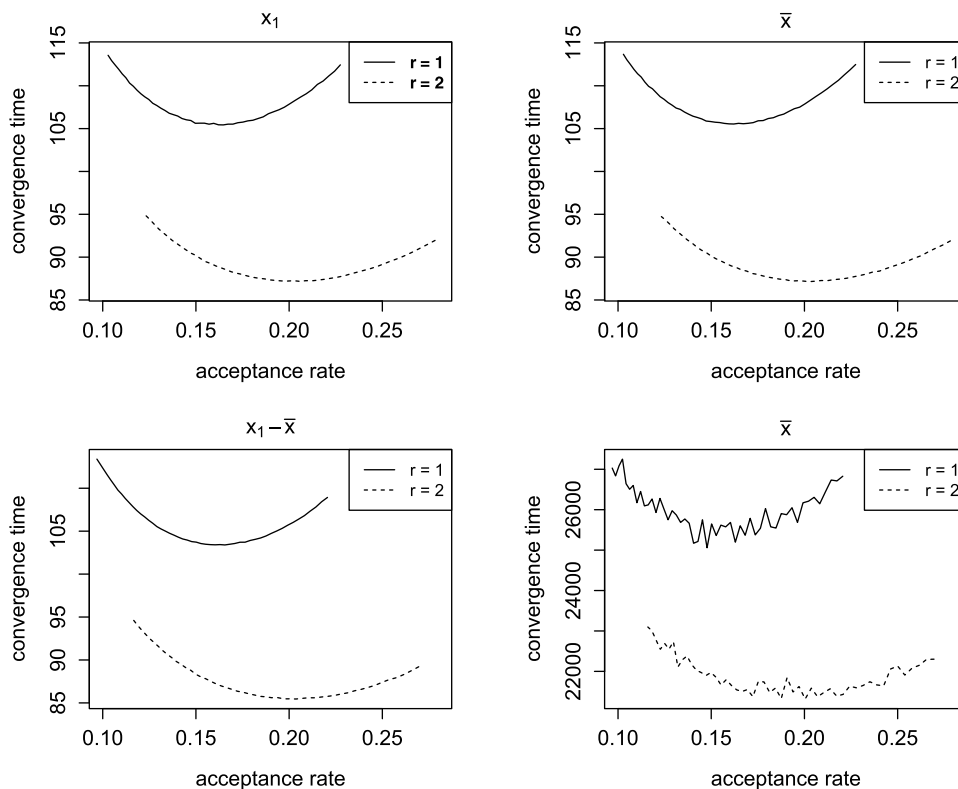


FIGURE 4. Convergence times for α_B against acceptance rate in the isotropic setting (top row) and the correlated target setting (bottom row).

with respect to π_d is of interest. Let $\{f(X_n)\}$ be the mapped process. Similarly to [29], we assess the choice of proposal variance by the convergence time:

$$\text{convergence time} := \frac{-k}{\log(\rho_k)},$$

where ρ_k is the lag- k autocorrelation in $\{f(X_n)\}$. In each of the following simulations, convergence time is estimated by averaging over 10^3 replications of Markov chains, each of length 10^6 with $k=1$. We chose a range of values of l where l is such that $\sigma_d^2 = l^2/d$ in a Gaussian proposal kernel $Q_d(x^d, \cdot) = N(x^d, \sigma_d^2 \mathbf{I}_d)$. Consider first the case of an isotropic target, $\pi_d = N_d(\mathbf{0}, \mathbf{I}_d)$ with isotropic Gaussian proposals; the conditions of Theorem 1 are satisfied. The estimated convergence time for $f(x) = x_1$ and $f(x) = \bar{x}$, where \bar{x} is the mean of all components x_1, \dots, x_d , is plotted in Figure 4 (top row). Here, $d = 50$. For both functions of interest, the optimal performance, i.e. the minimum convergence time, corresponds to an acceptance rate of approximately 0.158 for α_B and 0.197 for α_2^R ; the slight overestimation is due to the finite-dimensional setting. Next, we consider $\pi_d = N_d(\mathbf{0}, \Sigma_d)$ where Σ_d is a $d \times d$ matrix with 1 on its diagonal and all other elements are equal to some non-zero ρ . Here, the assumptions in Theorem 1 are not satisfied. For such a target and for α_{MH} , [29] showed that the rate of convergence of the algorithm is governed by the eigenvalues of Σ_d . In particular, the eigenvalues of Σ_d are $d\rho + 1 - \rho$ and $1 - \rho$, with associated eigenvectors y such that $y^T x$

yields \bar{x} and $x_i - \bar{x}$ (for $i = 1, \dots, d$), respectively. Then, it was shown that the algorithm converges quickly for functions orthogonal to \bar{x} , but much more slowly for \bar{x} . Despite the differing rates of convergence, the optimal acceptance rate, corresponding to the minimum convergence time, remains the same. We find this also to be true for α_B and α_2^R as illustrated in Figure 4 (bottom row), where we present convergence times for $x_1 - \bar{x}$ and \bar{x} . Once again, $d = 50$. The large difference between convergence times for both is quite evident from the y-axis of the two plots. The minimum again lies in a region around the asymptotic optimal. We note that because of the slow convergence rate of \bar{x} , the process demonstrates slow mixing, yielding more variable estimates of the convergence time. For both simulation settings, we see the expected improvement in the convergence time for α_2^R compared to α_B .

4.1. A Bayesian logistic regression example

We consider fitting a Bayesian logistic regression model to the famous Titanic dataset, which contains information on crew and passengers aboard the 1912 RMS Titanic. Let \mathbf{y} denote the response vector (indicating whether each person survived or not), and let \mathbf{X} denote the $n \times d$ model matrix; here $d = 10$. We assume a multivariate zero-mean Gaussian prior on $\boldsymbol{\beta}$ with covariance $100\mathbf{I}_{10}$. The resulting target density is

$$\pi(\boldsymbol{\beta} | \mathbf{y}) \propto \exp \left\{ -\frac{\boldsymbol{\beta}^T \boldsymbol{\beta}}{2} \prod_{i=1}^n \frac{\exp(-\mathbf{x}_i^T \boldsymbol{\beta})^{1-y_i}}{1 + \exp(-\mathbf{x}_i^T \boldsymbol{\beta})} \right\}.$$

For the Titanic dataset, the resulting posterior has a complicated covariance structure, with many components exhibiting an absolute mutual correlation of beyond .50. The posterior is also ill-conditioned, with the condition number of the estimated target covariance matrix being $\approx 10^5$. As seen in the bottom row of Figure 4, in such situations an isotropic proposal kernel might perform poorly for most functions. We instead consider a Gaussian proposal scheme where the proposal covariance matrix is taken to be proportional to the target covariance matrix. This is a common strategy for dealing with targets with correlated components and forms the basis for many adaptive MCMC kernels [30]. We implement Barker's algorithm to sample from the posterior. Let $\boldsymbol{\Sigma}_d$ denote the covariance matrix associated with the posterior distribution of $\boldsymbol{\beta}$; then the proposal kernel $Q_d(\mathbf{x}^d, \cdot) = N(\mathbf{x}^d, \sigma_d^2 \boldsymbol{\Sigma}_d)$. Since $\boldsymbol{\Sigma}_d$ is unavailable, we estimate it from a pilot MCMC run of size 10^7 . We then consider various values of $\sigma_d^2 = l^2/d$.

The performance of the algorithm for different functions of interest is plotted in Figure 5. Since this is a 10-dimensional problem, the optimal acceptance rate from Figure 2 is approximately 0.18. The convergence times for both, $\beta_1 - \bar{\beta}$ and $\bar{\beta}$, are similar. Furthermore, both are minimized at approximately the same acceptance rate of 0.18. It is natural here to be interested in estimating the posterior mean vector. Thus, we also study the properties of the vector $\boldsymbol{\beta}$, with efficiency measured via the multivariate effective sample size (ESS) [38]. The ESS returns the equivalent number of i.i.d. samples from π that would yield the same variability in estimating the posterior mean as the given set of MCMC samples. In Figure 5, we see that the optimal acceptance rate, corresponding to the highest ESS values, is achieved around 0.18.

5. Conclusions

We have obtained optimal scaling and acceptance rates for a large class of acceptance functions. In doing so, we have found that the scaling factor of $1/d$ for the proposal variance holds for all acceptance functions, indicating that the acceptance functions are not likely to affect

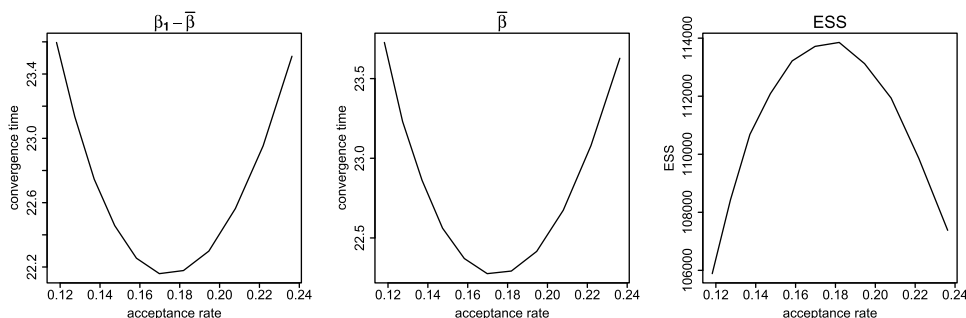


FIGURE 5. Convergence times for α_B (left and middle) and multivariate ESS for the posterior mean vector (right) against acceptance rate.

the rate of convergence, just the constants associated with that rate. Thus, practitioners need not hesitate in switching to other acceptance functions when the MH acceptance probability is not tractable, as long as Corollary 1 is used to tune their algorithm accordingly. There is also an inverse relationship between optimal variance and AOAR (see Table 1), implying that when dealing with sub-optimal acceptance functions, the algorithm seeks larger jumps. The computational cost of the Bernoulli factory that we present for α_r^R in Appendix D increases with r . Given the large jump in the optimal acceptance probability from $r = 1$ to $r = 2$, the development of more efficient Bernoulli factories is an important problem for future work. The assumption of starting from stationarity is a restrictive one. For MH with Gaussian proposals, the scaling factor of $1/d$ is still optimal when the algorithm is in the transient phase [6, 16, 17]. The optimal acceptance probability may vary depending on the starting distribution. We envision that similar results are viable for the general class of acceptance functions, and this is important future work. Our results are limited to only Gaussian proposals and trivially decomposable target densities. Other proposal distributions may make use of the gradient of the target, e.g. the Metropolis-adjusted Langevin algorithm [31] and Hamiltonian Monte Carlo [9]. In problems where α_{MH} cannot be used, the gradient of the target density is likely unavailable; thus it is reasonable to limit our attention to a Gaussian proposal. On the other hand, generalizations to other target distributions are important. For MH algorithms, [3, 34] relax the independence assumption, while [29] relax the identically distributed assumption. Additionally, [40] present a proof of weak convergence for MH for more general targets, and [33] provide optimal scaling results for general Bayesian targets using large-sample asymptotics. In these situations, extensions to other acceptance probabilities are similarly possible. Additionally, we encourage future work in optimal scaling to leverage our proof technique to demonstrate results for the wider class of acceptance probabilities.

Appendix A. Proof of Theorem 1

The proof is structurally similar to the seminal work of [27], in that we will show that the generator of the sped-up process, Z^d , converges to the generator of an appropriate Langevin diffusion. Define the discrete-time generator of Z^d as

$$G_d V(x^d) = d \cdot \mathbb{E}_{Y^d} \left[(V(Y^d) - V(x^d)) \alpha(x^d, Y^d) \right], \tag{11}$$

for all those V for which the limit exists. Since our interest is in the first component of \mathbf{Z}^d , we consider only those V which are functions of the first component only. Now, define the generator of the limiting Langevin diffusion process with speed measure $h_\alpha(l)$ as

$$GV(x) = h_\alpha(l) \left[\frac{1}{2} V''(x) + \frac{1}{2} \frac{d}{dx} (\log f)(x) V'(x) \right]. \tag{12}$$

The unique challenge in our result is identifying the speed measure $h_\alpha(l)$ for a general acceptance function $\alpha \in \mathcal{A}$. Proposition 1 is a key result that helps us obtain a form of $h_\alpha(l)$ without resorting to approximations.

To prove Theorem 1, we will show that there are events $F_d \subseteq \mathbb{R}^d$ such that for all t ,

$$\mathbb{P}[\mathbf{Z}_s^d \in F_d, 0 \leq s \leq t] \rightarrow 1 \text{ as } d \rightarrow \infty \quad \text{and}$$

$$\lim_{d \rightarrow \infty} \sup_{\mathbf{x}^d \in F_d} |G_d V(\mathbf{x}^d) - GV(x_1^d)| = 0,$$

for a suitably large class of real-valued functions V . Moreover, because of the conditions of Lipschitz continuity on f'/f , a core for the generator G has domain C_c^∞ , the class of infinitely differentiable functions with compact support [10, Theorem 2.1, Chapter 8]. Thus, we can limit our attention to only those $V \in C_c^\infty$ that are a function of the first component.

Consider now the setup of Theorem 1. Let $w = \log f$ and $\alpha \in \mathcal{A}$ with the balancing function g_α . Let w' and w'' be the first and second derivatives of w , respectively. Define the sequence of sets $\{F_d \subseteq \mathbb{R}^d, d > 1\}$ by

$$F_d = \{|R_d(x_2, \dots, x_d) - I| < d^{-1/8}\} \cap \{|S_d(x_2, \dots, x_d) - I| < d^{-1/8}\}, \quad \text{where}$$

$$R_d(x_2, \dots, x_d) = \frac{1}{d-1} \sum_{i=2}^d [\log(f(x_i))']^2 = \frac{1}{d-1} \sum_{i=2}^d [w'(x_i)]^2 \quad \text{and}$$

$$S_d(x_2, \dots, x_d) = \frac{-1}{d-1} \sum_{i=2}^d [\log(f(x_i))''] = \frac{-1}{d-1} \sum_{i=2}^d [w''(x_i)].$$

The following results from [27] will be needed.

Lemma 1. ([27].) *Let Assumption 1 hold. If $\mathbf{X}_0^d \sim \boldsymbol{\pi}_d$ for all d , then, for a fixed t , $\mathbb{P}[\mathbf{Z}_s^d \in F_d, 0 \leq s \leq t] \rightarrow 1$ as $d \rightarrow \infty$.*

Lemma 2. ([27].) *Let Assumption 1 hold. Also, let*

$$W_d(x_1, \dots, x_d) = \sum_{i=2}^d \left(\frac{1}{2} w''(x_i) (Y_i - x_i)^2 + \frac{l^2}{2(d-1)} w'(x_i)^2 \right),$$

where $Y_i \stackrel{\text{ind}}{\sim} N(x_i, \sigma_d^2), i = 2, \dots, d$. Then $\sup_{\mathbf{x}^d \in F_d} \mathbb{E}[|W_d(\mathbf{x}^d)|] \rightarrow 0$.

Lemma 3. ([27].) *For $Y \sim N(x, \sigma_d^2)$ and $V \in C_c^\infty$,*

$$\limsup_{d \rightarrow \infty} \sup_{x \in R} d |\mathbb{E}[V(Y) - V(x)]| < \infty.$$

For the following proposition, we will utilize the property (2) imposed on \mathcal{A} . This proposition is the key to obtaining our main result in such generality.

Proposition 1. Let $X \sim N(-\theta/2, \theta)$ for some $\theta > 0$. Let $\alpha \in \mathcal{A}$ with the corresponding balancing function g_α . Then $\mathbb{E}[Xg_\alpha(e^X)] = 0$.

Proof. We have

$$|\mathbb{E}[Xg_\alpha(e^X)]| \leq \mathbb{E}[|Xg_\alpha(e^X)|] \leq \mathbb{E}[|X|] < \infty;$$

the second inequality follows from the assumption that g_α lies in $[0, 1]$. Hence, the expectation exists and is equal to the integral

$$\int_{\mathbb{R}} x g_\alpha(e^x) \frac{1}{\sqrt{2\pi\theta}} \exp\left\{-\frac{(x + \theta/2)^2}{2\theta}\right\} dx =: \int_{\mathbb{R}} h(x) dx.$$

Observe that, using (2),

$$\begin{aligned} h(-x) &= -x g_\alpha(e^{-x}) \frac{1}{\sqrt{2\pi\theta}} \exp\left\{-\frac{(-x + \theta/2)^2}{2\theta}\right\} \\ &= -x g_\alpha(e^{-x}) \frac{1}{\sqrt{2\pi\theta}} \exp\left\{-\frac{1}{2\theta} \left(x^2 + \frac{\theta^2}{4} - x\theta\right)\right\} \\ &= -x e^{-x} g_\alpha(e^x) \frac{1}{\sqrt{2\pi\theta}} \exp\left\{-\frac{1}{2\theta} \left(x^2 + \frac{\theta^2}{4} - x\theta\right)\right\} \\ &= -x g_\alpha(e^x) \frac{1}{\sqrt{2\pi\theta}} \exp\left\{-\frac{1}{2\theta} \left(x^2 + \frac{\theta^2}{4} + x\theta\right)\right\} \\ &= -x g_\alpha(e^x) \frac{1}{\sqrt{2\pi\theta}} \exp\left\{-\frac{(x + \theta/2)^2}{2\theta}\right\} \\ &= -h(x). \end{aligned}$$

Hence, the result follows. □

Lemma 4. Suppose $V \in C_c^\infty$ is restricted to only the first component of \mathbf{Z}^d . Then

$$\sup_{\mathbf{x}^d \in F_d} |G_d V(\mathbf{x}^d) - G V(x_1^d)| \rightarrow 0 \quad \text{as } d \rightarrow \infty.$$

Proof. In the expression for $G_d V(\mathbf{x}^d)$ given in (11), we can decompose the proposal \mathbf{Y}^d into (Y_1^d, \mathbf{Y}^{d-}) and thus rewrite the expectation as follows:

$$G_d V(\mathbf{x}^d) = d \mathbb{E}_{Y_1^d} \left[\left(V(Y_1^d) - V(x_1^d) \right) \mathbb{E}_{\mathbf{Y}^{d-}} \left[\alpha(\mathbf{x}^d, \mathbf{Y}^d) \mid Y_1^d \right] \right]. \tag{13}$$

Let $E^{d,\alpha}$ denote the inner expectation in (13) and define $E_{lim}^{d,\alpha}$ as

$$E_{lim}^{d,\alpha} = \mathbb{E}_{\mathbf{Y}^{d-}} \left[g_\alpha \left(\exp \left\{ \log \frac{f(Y_1^d)}{f(x_1^d)} + \sum_{i=2}^d \left(w'(x_i^d)(Y_i^d - x_i^d) - \frac{l^2 w'(x_i^d)^2}{2(d-1)} \right) \right\} \right) \mid Y_1^d \right]. \tag{14}$$

Also, a Taylor series expansion of w about x_i^d for $i = 2, \dots, d$ gives

$$\begin{aligned} E^{d,\alpha} &= \mathbb{E}_{\mathbf{Y}^{d-}} \left[g_\alpha \left(\exp \left\{ \log \frac{f(Y_1^d)}{f(x_1^d)} + \sum_{i=2}^d w'(x_i^d)(Y_i^d - x_i^d) \right. \right. \right. \\ &\quad \left. \left. \left. + \frac{1}{2} w''(x_i^d)(Y_i^d - x_i^d)^2 + \frac{1}{6} w'''(Z_i)(Y_i^d - x_i^d)^3 \right\} \right) \mid Y_1^d \right] \end{aligned}$$

for Z_i lying between x_i^d and Y_i^d . Hence, the triangle inequality and Lipschitz continuity of $g(e^z)$ give, for some Lipschitz constant $K < \infty$,

$$|E^{d,\alpha} - E_{lim}^{d,\alpha}| \leq K \mathbb{E}_{Y^d} \left[\left| \sum_{i=2}^d \frac{1}{2} w''(x_i^d) (Y_i^d - x_i^d)^2 + \frac{1}{6} w'''(Z_i) (Y_i^d - x_i^d)^3 + \frac{l^2 w'(x_i^d)^2}{2(d-1)} \right| \right] \\ \leq K \mathbb{E}_{Y^d} \left[|W_d(\mathbf{x}^d)| \right] + K \sup_{z \in \mathbb{R}} |w'''(z)| \frac{l^3}{(d-1)^{1/2}}, \tag{15}$$

where $W_d(\mathbf{x}^d)$ is as defined in Lemma 2. From Lemma 2, Lemma 3, and (15),

$$\sup_{\mathbf{x}^d \in F_d} \left| G_d V(\mathbf{x}^d) - d \mathbb{E}_{Y_1^d} \left[\left(V(Y_1^d) - V(x_1^d) \right) E_{lim}^{d,\alpha} \right] \right| \rightarrow 0 \text{ as } d \rightarrow \infty. \tag{16}$$

Now let $\varepsilon(y) = \log f(y) - \log f(x_1^d)$. Also, from (14), it is clear that given $\mathbf{x}^d, E_{lim}^{d,\alpha}$ is a function of Y_1^d alone, to wit,

$$(M_{d,\alpha} \circ \varepsilon)(Y_1^d) := E_{lim}^{d,\alpha} = \mathbb{E}[g_\alpha(e^{B_d})], \tag{17}$$

where $B_d \sim N(\mu_d, \Sigma_d)$ with $\mu_d = \varepsilon(Y_1^d) - l^2 R_d / 2$ and $\Sigma_d = l^2 R_d$. Thus, by (15), it is enough to consider the asymptotic behaviour of

$$d \mathbb{E}_{Y_1^d} \left[\left(V(Y_1^d) - V(x_1^d) \right) M_{d,\alpha}(\varepsilon(Y_1^d)) \right].$$

Let $N_{d,\alpha} = M_{d,\alpha} \circ \varepsilon$ and apply Taylor series expansion on the inner term to obtain

$$\left(V(Y_1^d) - V(x_1^d) \right) M_{d,\alpha}(\varepsilon(Y_1^d)) \\ = \left(V'(x_1^d)(Y_1^d - x_1^d) + \frac{1}{2} V''(x_1^d)(Y_1^d - x_1^d)^2 + \frac{1}{6} V'''(K_d)(Y_1^d - x_1^d)^3 \right) \\ \times \left(N_{d,\alpha}(x_1^d) + N'_{d,\alpha}(x_1^d)(Y_1^d - x_1^d) + \frac{1}{2} N''_{d,\alpha}(L_d)(Y_1^d - x_1^d)^2 \right),$$

where $K_d, L_d \in [Y_1^d, x_1^d]$ or $[x_1^d, Y_1^d]$, and

$$N_{d,\alpha}(x_1^d) = M_{d,\alpha}(\varepsilon(x_1^d)) = M_{d,\alpha} \left(\log \frac{f(x_1^d)}{f(x_1^d)} \right) = M_{d,\alpha}(0), \\ N'_{d,\alpha}(x_1^d) = M'_{d,\alpha}(\varepsilon(x_1^d)) \varepsilon'(x_1^d) = M'_{d,\alpha}(0) w'(x_1^d). \tag{18}$$

Now, for all d ,

$$M_{d,\alpha}(\varepsilon) = \mathbb{E}[g_\alpha(e^{B_d})] = \int_{\mathbb{R}} g_\alpha(e^b) \frac{1}{\sqrt{2\pi l^2 R_d}} \exp \left\{ \frac{-(b - \varepsilon + l^2 R_d / 2)^2}{2l^2 R_d} \right\} db. \\ \text{So } M_{d,\alpha}(0) = \int_{\mathbb{R}} g_\alpha(e^b) \frac{1}{\sqrt{2\pi l^2 R_d}} \exp \left\{ \frac{-(b + l^2 R_d / 2)^2}{2l^2 R_d} \right\} db. \\ \text{Also, } M'_{d,\alpha}(\varepsilon) = \frac{d}{d\varepsilon} \left(\int_{\mathbb{R}} g_\alpha(e^b) \frac{1}{\sqrt{2\pi l^2 R_d}} \exp \left\{ \frac{-(b - \varepsilon + l^2 R_d / 2)^2}{2l^2 R_d} \right\} db \right).$$

The derivatives and integral here can be exchanged thanks to the dominated convergence theorem, which yields

$$\begin{aligned}
 M'_{d,\alpha}(\varepsilon) &= \int_{\mathbb{R}} g_{\alpha}(e^b) \frac{1}{\sqrt{2\pi l^2 R_d}} \left(\frac{2(b - \varepsilon + l^2 R_d/2)}{2l^2 R_d} \right) \exp \left\{ \frac{-(b - \varepsilon + l^2 R_d/2)^2}{2l^2 R_d} \right\} db. \\
 \text{So } M'_{d,\alpha}(0) &= \int_{\mathbb{R}} g_{\alpha}(e^b) \frac{1}{\sqrt{2\pi l^2 R_d}} \left(\frac{(b + l^2 R_d/2)}{l^2 R_d} \right) \exp \left\{ \frac{-(b + l^2 R_d/2)^2}{2l^2 R_d} \right\} db \\
 &= \frac{1}{l^2 R_d} \int_{\mathbb{R}} b g_{\alpha}(e^b) \frac{1}{\sqrt{2\pi l^2 R_d}} \exp \left\{ \frac{-(b + l^2 R_d/2)^2}{2l^2 R_d} \right\} db \\
 &\quad + \frac{1}{2} \int_{\mathbb{R}} g_{\alpha}(e^b) \frac{1}{\sqrt{2\pi l^2 R_d}} \exp \left\{ \frac{-(b + l^2 R_d/2)^2}{2l^2 R_d} \right\} db \\
 &= \frac{1}{2} M_{d,\alpha}(0),
 \end{aligned}$$

where the first term vanishes by Proposition 1. Hence, for all d ,

$$2M'_{d,\alpha}(0) = M_{d,\alpha}(0) = \int_{\mathbb{R}} g_{\alpha}(e^b) \frac{1}{\sqrt{2\pi l^2 R_d}} \exp \left\{ \frac{-(b + l^2 R_d/2)^2}{2l^2 R_d} \right\} db. \tag{19}$$

Now, we plug the expressions obtained above into the Taylor series expansion of $(V(Y_1^d) - V(x_1^d)) M_{d,\alpha}(\varepsilon(Y_1^d))$. The rest of the proof, with the help of Assumption 1, follows similarly as in [27, Lemma 2.6]. □

Proof of Theorem 1. From Lemma 4, we have uniform convergence of generators on the sequence of sets with limiting probability 1. Thus, by Corollary 8.7 in [10, Chapter 4], we have the required result of weak convergence (the condition that C_c^∞ separates points was verified by [27]). □

Appendix B. Proof of Corollary 1

Lemma 5. *Let $E^{d,\alpha}$ be the inner expectation in (13), and let $E_{lim}^{d,\alpha}$ be from (14). Then*

$$\mathbb{E}_{\pi_d} \left[\mathbb{E}_{Y_1} \left[E^{d,\alpha} - E_{lim}^{d,\alpha} \mid \mathbf{x}^d \right] \right] \rightarrow 0 \quad \text{as } d \rightarrow \infty.$$

Proof. Consider

$$\begin{aligned}
 \left| \mathbb{E}_{\pi_d} \left[\mathbb{E}_{Y_1^d} \left[E^{d,\alpha} - E_{lim}^{d,\alpha} \mid \mathbf{x}^d \right] \right] \right| &\leq \left| \mathbb{E}_{\pi_d} \left[\mathbb{E}_{Y_1^d} \left[E^{d,\alpha} - E_{lim}^{d,\alpha} \mid \mathbf{x}^d \in F_d \right] \right] P(\mathbf{x}^d \in F_d) \right| \\
 &\quad + \left| \mathbb{E}_{\pi_d} \left[\mathbb{E}_{Y_1^d} \left[E^{d,\alpha} - E_{lim}^{d,\alpha} \mid \mathbf{x}^d \in F_d^C \right] \right] P(\mathbf{x}^d \in F_d^C) \right|.
 \end{aligned}$$

The second term goes to 0, since the expectation is bounded and by construction $P(\mathbf{x}^d \in F_d^C) \rightarrow 0$ as $d \rightarrow \infty$. Also, following [27],

$$\sup_{\mathbf{x}^d \in F_d} |E^{d,\alpha} - E_{lim}^{d,\alpha}| \rightarrow 0 \text{ as } d \rightarrow \infty.$$

Then

$$\begin{aligned} & \left| \mathbb{E}_{\pi_d} \left[\mathbb{E}_{Y_1^d} \left[E^{d,\alpha} - E_{lim}^{d,\alpha} \mid \mathbf{x}^d \in F_d \right] \right] P(\mathbf{x}^d \in F_d) \right| \\ & \leq \mathbb{E}_{\pi_d} \left[\mathbb{E}_{Y_1^d} \left[\sup_{\mathbf{x}^d \in F_d} \left| E^{d,\alpha} - E_{lim}^{d,\alpha} \right| \mid \mathbf{x}^d \in F_d \right] \right] \rightarrow 0. \end{aligned} \quad \square$$

Proof of Corollary 1. Consider Equation (17). Using the Taylor series approximation of second order around x_1 ,

$$\mathbb{E}_{Y_1^d} [E_{lim}^{d,\alpha}] = \mathbb{E}[N_{d,\alpha}(Y_1^d)] = N_{d,\alpha}(x_1^d) + \frac{1}{2} N''_{d,\alpha}(W_{d,1}) \frac{l^2}{d-1},$$

where $W_{d,1} \in [x_1^d, Y_1^d]$ or $[Y_1^d, x_1^d]$. Since N'' is bounded [27],

$$\begin{aligned} \alpha(l) &= \lim_{d \rightarrow \infty} \mathbb{E}_{\pi_d} \left[\mathbb{E}_{Y_1^d} \left[\mathbb{E}_{Y^d} \left[\alpha(\mathbf{X}^d, \mathbf{Y}^d) \mid Y_1^d, \mathbf{x}^d \right] \mid \mathbf{x}^d \right] \right] \\ &= \lim_{d \rightarrow \infty} \mathbb{E}_{\pi_d} \left[\mathbb{E}_{Y_1^d} \left[E_{lim}^{d,\alpha} + E^{d,\alpha} - E_{lim}^{d,\alpha} \mid \mathbf{x}^d \right] \right]. \end{aligned}$$

As all expectations exist, we can split the inner expectation and use Lemma 5, so that

$$\begin{aligned} \alpha(l) &= \lim_{d \rightarrow \infty} \mathbb{E}_{\pi_d} \left[\mathbb{E}_{Y_1^d} \left[E_{lim}^{d,\alpha} \mid \mathbf{x}^d \right] \right] + \lim_{d \rightarrow \infty} \mathbb{E}_{\pi_d} \left[\mathbb{E}_{Y_1^d} \left[E^{d,\alpha} - E_{lim}^{d,\alpha} \mid \mathbf{x}^d \right] \right] \\ &= \lim_{d \rightarrow \infty} \mathbb{E}_{\pi_d} \left[M_{d,\alpha}(0) + \frac{1}{2} N''_{d,\alpha}(W_{d,1}) \frac{l^2}{d-1} \right] \\ &= \lim_{d \rightarrow \infty} \mathbb{E}_{\pi_d} \left[\int_{\mathbb{R}} g_{\alpha}(e^b) \frac{1}{\sqrt{2\pi l^2 R_d}} \exp \left\{ \frac{-(b + l^2 R_d / 2)^2}{2l^2 R_d} \right\} db \right] \\ &= \int_{\mathbb{R}} g_{\alpha}(e^b) \frac{1}{\sqrt{2\pi l^2 I}} \exp \left\{ \frac{-(b + l^2 I / 2)^2}{2l^2 I} \right\} db = M_{\alpha}(l). \end{aligned}$$

The last equality is by the law of large numbers and the continuous mapping theorem. □

Appendix C. Optimizing speed for Barker’s acceptance

We need to maximize $h_B(l) = l^2 M_B(l)$. Let I be fixed arbitrarily. Then

$$h_B(l) = \frac{1}{I} \cdot l^2 I \cdot \int_{\mathbb{R}} \frac{1}{1 + e^{-b}} \frac{1}{\sqrt{2\pi l^2 I}} \exp \left\{ \frac{-(b + l^2 I / 2)^2}{2l^2 I} \right\} db.$$

For a fixed I , we can reparametrize the function by taking $\theta = l^2 I$, and so maximizing $h_B(l)$ over positive l will be equivalent to maximizing $h_B^1(\theta)$ over positive θ , where

$$h_B^1(\theta) = \int_{\mathbb{R}} \frac{\theta}{1 + e^{-b}} \frac{1}{\sqrt{2\pi \theta}} \exp \left\{ \frac{-(b + \theta / 2)^2}{2\theta} \right\} db.$$

We make the substitution $z = (b + \theta / 2) / \sqrt{\theta}$ in the integrand to obtain

$$h_B^1(\theta) = \int_{\mathbb{R}} \frac{\theta}{1 + \exp\{-z\sqrt{\theta} + \theta/2\}} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = \mathbb{E} \left[\frac{\theta}{1 + \exp\{-Z\sqrt{\theta} + \theta/2\}} \right],$$

Algorithm 1: Die-coin algorithm for $\alpha_2^R(x, y)$

- 1: Draw $D \sim \text{Categorical} \left(\frac{c_y^2}{c_x^2 + c_x c_y + c_y^2}, \frac{c_x c_y}{c_y^2 + c_x c_y + c_x^2}, \frac{c_x^2}{c_y^2 + c_x c_y + c_x^2} \right)$
 2. **if** $D = 1$ **then**
 3. Draw $C_1 \sim \text{Bern}(p_y^2)$
 4. **if** $C_1 = 1$ **then** output 1 **else** go back to Step 1
 5. **if** $D = 2$ **then**
 6. Draw $C_1 \sim \text{Bern}(p_x p_y)$
 7. **if** $C_1 = 1$ **then** output 1 **else** go back to Step 1
 8. **if** $D = 3$ **then**
 9. Draw $C_1 \sim \text{Bern}(p_x^2)$
 10. **if** $C_1 = 1$ **then** output 0 **else** go back to Step 1
-

where the expectation is taken with respect to $Z \sim N(0, 1)$. This expectation is not available in closed form. However, standard numerical integration routines yield the optimal value of θ to be 6.028. This implies that the optimal value of l , say l^* , is approximately equal to

$$l^* \approx \frac{2.46}{\sqrt{I}} \text{ (up to 2 decimal places).}$$

Using this l^* yields an AOAR of approximately 0.158.

Appendix D. Bernoulli factory

To sample events of probability α_B , the *two-coin* algorithm, an efficient Bernoulli factory, was presented in [13]. Generalizing this to a *die-coin* algorithm, we present a Bernoulli factory for α_r^R for $r = 2$; extensions to other r can be done similarly. Let $\pi(x) = c_x p_x$ with $p_x \in [0, 1]$ and $c_x > 0$. Then

$$\alpha_2^R(x, y) = \frac{\pi(y)^2 + \pi(x)\pi(y)}{\pi(y)^2 + \pi(x)\pi(y) + \pi(x)^2} = \frac{c_y^2 p_y^2 + c_x p_x c_y p_y}{c_y^2 p_y^2 + c_x p_x c_y p_y + c_x^2 p_x^2}.$$

Acknowledgements

The authors thank the referees and the editor for comments that helped improve the work.

Funding information

D. Vats is supported by SERB grant SPG/2021/001322; K. Łatuszyński is supported by the Royal Society through the Royal Society University Research Fellowship; and G. Roberts is supported by the EPSRC grants CoSInES (EP/R034710/1) and Bayes for Health (EP/R018561/1).

Competing interests

There were no competing interests to declare which arose during the preparation or publication process of this article.

References

- [1] BANTERLE, M., GRAZIAN, C., LEE, A. AND ROBERT, C. P. (2019). Accelerating Metropolis–Hastings algorithms by delayed acceptance. *Found. Data Sci.* **1**, 103–128.
- [2] BARKER, A. A. (1965). Monte Carlo calculations of the radial distribution functions for a proton–electron plasma. *Austral. J. Phys.* **18**, 119–134.
- [3] BÉDARD, M. (2008). Optimal acceptance rates for Metropolis algorithms: moving beyond 0.234. *Stoch. Process. Appl.* **118**, 2198–2222.
- [4] BILLERA, L. J. AND DIACONIS, P. (2001). A geometric interpretation of the Metropolis–Hastings algorithm. *Statist. Sci.* 335–339.
- [5] BROOKS, S., GELMAN, A., JONES, G. AND MENG, X.-L. (2011). *Handbook of Markov Chain Monte Carlo*. CRC Press, Boca Raton.
- [6] CHRISTENSEN, O. F., ROBERTS, G. O. AND ROSENTHAL, J. S. (2005). Scaling limits for the transient phase of local Metropolis–Hastings algorithms. *J. R. Statist. Soc. B [Statist. Methodology]* **67**, 253–268.
- [7] DELMAS, J.-F. AND JOURDAIN, B. (2009). Does waste recycling really improve the multi-proposal Metropolis–Hastings algorithm? An analysis based on control variates. *J. Appl. Prob.* **46**, 938–959.
- [8] 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**, 295–313.
- [9] DUANE, S., KENNEDY, A. D., PENDLETON, B. J. AND ROWETH, D. (1987). Hybrid Monte Carlo. *Phys. Lett. B* **195**, 216–222.
- [10] ÉTHIER, S. N. AND KURTZ, T. G. (1986). *Markov Processes: Characterization and Convergence*. John Wiley, New York.
- [11] GELMAN, A., ROBERTS, G. O. AND GILKS, W. R. (1996). Efficient Metropolis jumping rules. *Bayesian Statist.* **5**, 599–608.
- [12] GONÇALVES, F. B., ŁATUSZYŃSKI, K. AND ROBERTS, G. O. (2017). Barker’s algorithm for Bayesian inference with intractable likelihoods. *Brazilian J. Prob. Statist.* **31**, 732–745.
- [13] GONÇALVES, F. B., ŁATUSZYŃSKI, K. AND ROBERTS, G. O. (2017). Exact Monte Carlo likelihood-based inference for jump-diffusion processes. Preprint. Available at <https://arxiv.org/abs/1707.00332>.
- [14] HASTINGS, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97–109.
- [15] HERBEI, R. AND BERLINER, L. M. (2014). Estimating ocean circulation: an MCMC approach with approximated likelihoods via the Bernoulli factory. *J. Amer. Statist. Assoc.* **109**, 944–954.
- [16] JOURDAIN, B., LELIÈVRE, T. AND MIAISOJEDOW, B. (2014). Optimal scaling for the transient phase of Metropolis Hastings algorithms: the longtime behavior. *Bernoulli* **20**, 1930–1978.
- [17] KUNTZ, J., OTTOBRE, M. AND STUART, A. M. (2019). Diffusion limit for the random walk Metropolis algorithm out of stationarity. *Ann. Inst. H. Poincaré Prob. Statist.* **55**, 1599–1648.
- [18] ŁATUSZYŃSKI, K. AND ROBERTS, G. O. (2013). CLTs and asymptotic variance of time-sampled Markov chains. *Methodology Comput. Appl. Prob.* **15**, 237–247.
- [19] MENEZES, A. A. AND KABAMBA, P. T. (2014). Optimal search efficiency of Barker’s algorithm with an exponential fitness function. *Optimization Lett.* **8**, 691–703.
- [20] METROPOLIS, N. *et al.* (1953). Equation of state calculations by fast computing machines. *J. Chem. Phys.* **21**, 1087–1092.
- [21] MEYN, S. P. AND TWEEDIE, R. L. (2012). *Markov Chains and Stochastic Stability*. Cambridge University Press.
- [22] MIRA, A. (2001). On Metropolis–Hastings algorithms with delayed rejection. *Metron* **59**, 231–241.
- [23] MORINA, G., ŁATUSZYŃSKI, K., NAYAR, P. AND WENDLAND, A. (2021). From the Bernoulli factory to a dice enterprise via perfect sampling of Markov chains. To appear in *Ann. Appl. Prob.*
- [24] NEAL, P. AND ROBERTS, G. O. (2006). Optimal scaling for partially updating MCMC algorithms. *Ann. Appl. Prob.* **16**, 475–515.
- [25] PESKUN, P. H. (1973). Optimum Monte-Carlo sampling using Markov chains. *Biometrika* **60**, 607–612.
- [26] ROBERT, C. AND CASELLA, G. (2013). *Monte Carlo Statistical Methods*. Springer, New York.
- [27] ROBERTS, G. O., GELMAN, A. AND GILKS, W. R. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *Ann. Appl. Prob.* **7**, 110–120.
- [28] ROBERTS, G. O. AND ROSENTHAL, J. S. (1998). Optimal scaling of discrete approximations to Langevin diffusions. *J. R. Statist. Soc. B [Statist. Methodology]* **60**, 255–268.

- [29] ROBERTS, G. O. AND ROSENTHAL, J. S. (2001). Optimal scaling for various Metropolis–Hastings algorithms. *Statist. Sci.* **16**, 351–367.
- [30] ROBERTS, G. O. AND ROSENTHAL, J. S. (2009). Examples of adaptive MCMC. *J. Comput. Graph. Statist.* **18**, 349–367.
- [31] ROBERTS, G. O. AND TWEEDIE, R. L. (1996). Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli* **2**, 341–363.
- [32] SCHMON, S. M., DELIGIANNIDIS, G., DOUCET, A. AND PITT, M. K. (2021). Large-sample asymptotics of the pseudo-marginal method. *Biometrika* **108**, 37–51.
- [33] SCHMON, S. M. AND GAGNON, P. (2022). Optimal scaling of random walk Metropolis algorithms using Bayesian large-sample asymptotics. *Statist. Comput.* **32**, 1–16.
- [34] SHERLOCK, C. AND ROBERTS, G. O. (2009). Optimal scaling of the random walk Metropolis on elliptically symmetric unimodal targets. *Bernoulli* **15**, 774–798.
- [35] SHERLOCK, C., THIERY, A. H. AND GOLIGHTLY, A. (2021). Efficiency of delayed-acceptance random walk Metropolis algorithms. *Ann. Statist.* **49**, 2972–2990.
- [36] 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**, 238–275.
- [37] SMITH, C. J. (2018). Exact Markov chain Monte Carlo with likelihood approximations for functional linear models. Doctoral Thesis, Ohio State University.
- [38] VATS, D., FLEGAL, J. M. AND JONES, G. L. (2019). Multivariate output analysis for Markov chain Monte Carlo. *Biometrika* **106**, 321–337.
- [39] VATS, D., GONÇALVES, F. B., ŁATUSZYŃSKI, K. AND ROBERTS, G. O. (2022). Efficient Bernoulli factory Markov chain Monte Carlo for intractable posteriors. *Biometrika* **109**, 369–385.
- [40] YANG, J., ROBERTS, G. O. AND ROSENTHAL, J. S. (2020). Optimal scaling of random-walk Metropolis algorithms on general target distributions. *Stoch. Process. Appl.* **130**, 6094–6132.
- [41] ZANELLA, G., BÉDARD, M. AND KENDALL, W. S. (2017). A Dirichlet form approach to MCMC optimal scaling. *Stoch. Process. Appl.* **127**, 4053–4082.