Biometrika (2022), **00**, 0, *pp.* 1–17 https://doi.org/10.1093/biomet/asac056

Optimal design of the Barker proposal and other locally balanced Metropolis–Hastings algorithms

BY JURE VOGRINC

Department of Statistics, University of Warwick, Coventry CV4 7AL, U.K.
jure.vogrinc@warwick.ac.uk

SAMUEL LIVINGSTONE

Department of Statistical Science, University College London,
Gower Street, London WC1E 6BT, U.K.
samuel.livingstone@ucl.ac.uk

AND GIACOMO ZANELLA

Department of Decision Sciences, Bocconi University, via Roentgen 1, Milan 20136, Italy
giacomo.zanella@unibocconi.it

SUMMARY

We study the class of first-order locally balanced Metropolis–Hastings algorithms introduced in Livingstone & Zanella (2022). To choose a specific algorithm within the class, the user must select a balancing function $g : \mathbb{R}_+ \to \mathbb{R}_+$ satisfying $g(t) = tg(1/t)$ and a noise distribution for the proposal increment. Popular choices within the class are the Metropolis-adjusted Langevin algorithm and the recently introduced Barker proposal. We first establish a general limiting optimal acceptance rate of 57% and scaling of $n^{-1/3}$, as the dimension n tends to infinity among all members of the class under mild smoothness assumptions on g and when the target distribution for the algorithm is of product form. In particular, we obtain an explicit expression for the asymptotic efficiency of an arbitrary algorithm in the class, as measured by expected squared jumping distance. We then consider how to optimize this expression under various constraints. We derive an optimal choice of noise distribution for the Barker proposal, an optimal choice of balancing function under a Gaussian noise distribution, and an optimal choice of first-order locally balanced algorithm among the entire class, which turns out to depend on the specific target distribution. Numerical simulations confirm our theoretical findings, and in particular, show that a bimodal choice of noise distribution in the Barker proposal gives rise to a practical algorithm that is consistently more efficient than the original Gaussian version.

Some key words: Barker proposal; Locally balanced algorithm; Markov chain Monte Carlo; Metropolis–Hastings algorithm; Optimal scaling.

1. INTRODUCTION

Markov chain Monte Carlo algorithms are the workhorse of many contemporary statistical analyses and constitute an essential part of the modern data science toolkit. Despite many advances, however, reliable inference using Markov chain Monte Carlo can still be a cumbersome task. It is common for practitioners to dedicate much effort to making careful algorithm design choices, and adjusting algorithmic tuning parameters to ensure that performance is adequate for a given problem. Failure to do so can be catastrophic; examples for which a well-designed algorithm performs adequately, but a less carefully chosen alternative does not, are ubiquitous; see, e.g., Sherlock et al. (2010).

Suitable guidelines on the intelligent design and implementation of Markov chain Monte Carlo methods are therefore important. They are not always easy to formulate, however, as the best choice of method can depend on the user and the problem at hand. In some contexts, a simpler algorithm with less need for adjustment, and for which potential problems are easy to diagnose, may be preferable. In other contexts, one may be comfortable with more complex methods which can perform adequately on a larger class of problems if enough fine tuning is done.

For Metropolis–Hastings algorithms, perhaps the most celebrated guidelines concern the choice of the optimal acceptance rate (Roberts & Rosenthal, 2001). Rigorous theoretical justifications for certain values tend to be restricted to the case in which the dimension tends to infinity, and the distribution from which samples are desired has a particular structure, such as a product form, but empirically the same values are known to be appropriate in many other settings (Roberts & Rosenthal, 2001). The apparent lack of dependence of these optimal choices on the target distribution allows particularly simple recommendations to be offered to the user of a given algorithm. This approach was introduced in the work of Roberts et al. (1997), which provided the celebrated 0.234 optimal acceptance rate for the random walk Metropolis algorithm. Analogous rates have since been extracted for several other algorithms, notable examples include the Metropolis-adjusted Langevin algorithm, with optimal rate 0.574 (Roberts & Rosenthal, 1998), and Hamiltonian Monte Carlo, with optimal rate 0.651 (Beskos et al., 2013).

Livingstone & Zanella (2022) introduced a general class of gradient-based algorithms, termed first-order locally balanced Metropolis–Hastings algorithms, of which the Metropolis-adjusted Langevin algorithm (Roberts & Tweedie, 1996) is a special case. Constructing a member of the class requires a Markov kernel, which can be thought of as the initial noise distribution for the transition, together with a balancing function, which must satisfy certain properties described in § 2. Livingstone & Zanella (2022) considered different choices from within the class, and in particular, constructed a method called the Barker proposal. Empirical results in the paper show that, despite being remarkably simple to implement, the Barker algorithm enables reliable sampling in complex scenarios where other gradient-based methods may fail. The authors also established sufficient conditions for geometric ergodicity and presented some preliminary results on scaling with dimension, suggesting that relaxation times are $O(n^{1/3})$, where n is the dimension of the state. More discussion and a pedagogical derivation of the Barker algorithm is provided in Hird et al. (2022).

Several unexplored questions remain regarding locally balanced Metropolis–Hastings algorithms. The initial noise distribution in the Barker algorithm is simply chosen to be Gaussian in Livingstone & Zanella (2022), but no justification besides convenience is given for this choice. It could be that a different choice leads to a more effective algorithm. Similarly, guidelines on the optimal acceptance rate for the Barker algorithm have not been established. More generally, little discussion exists on other first-order locally balanced Metropolis–Hastings methods. It is natural to wonder whether all members of the class will exhibit $O(n^{1/3})$ relaxation times, if the Metropolis–adjusted Langevin algorithm is the most efficient choice when optimally tuned and, indeed, whether such a direct quantitative comparison of methods is possible in general. These questions are of both theoretical and practical interest, as they have direct implications for the optimal design of algorithms.

In this paper we make several contributions. First, we present general results on the optimal choice of acceptance rate, and scaling with dimension of any algorithm within the class of first-order locally balanced Markov processes under mild regularity conditions on the balancing function and a product-form assumption on the target distribution. In particular, in § 3 we show that the 57% guideline acceptance rate for the Metropolis-adjusted Langevin algorithm also holds for the Barker proposal and several other methods, as does the $O(n^{1/3})$ scaling with dimension as measured by expected squared jump distance. Despite having the same optimal acceptance rate and scaling with dimensionality, however, all such schemes have different asymptotic efficiencies, which we characterize explicitly, enabling principled and generic optimization of the algorithmic design.

Our theoretical results build on the recently introduced optimal scaling framework of Zanella et al. (2017) and Vogrinc & Kendall (2021). One powerful aspect of our approach is the ability it affords us to analyse fairly generic schemes without requiring overly case-specific calculations, while still obtaining explicit expressions for the asymptotic performances of algorithms that can directly be compared with each other. This allows characterization of the quantitative interplay between fine-scale properties of the target and proposal distributions in the resulting asymptotic efficiency, thus enabling precise methodological guidance.

2. Locally balanced Markov processes

2.1. *General framework*

Consider a Markov transition kernel Q defined on a Borel space $(\mathbb{X}, \mathcal{F})$. We restrict attention to $\mathbb{X} \subset \mathbb{R}^n$ for some finite n. We say that Q satisfies the detailed balance equations with respect to a probability measure π if

$$\int f(x)h(y)\pi(dx)Q(y,dx) = \int f(x)h(y)\pi(dy)Q(y,dx) \tag{1}$$

for any $f, h \in L^2(\pi)$. When Q does not satisfy (1), a new kernel can be constructed using the concept of a balancing function. Let $g : [0,\infty) \to [0,\infty)$ be such that $g(0) = 0$ and

$$g(t) = tg(1/t) \tag{2}$$

for $t > 0$, and recall that by Tierney (1998, Proposition 1) there exists a symmetric set $\mathcal{R} \times \mathcal{R} \in \mathbb{X} \times \mathbb{X}$ such that the Radon–Nikodym derivative

$$t(x,y) = \frac{\pi(dy)Q(y,dx)}{\pi(dx)Q(x,dy)} \tag{3}$$

is well-defined and such that $0 < t(x,y) < \infty$ if $x, y \in \mathcal{R}$ and $t(x,y) = 0$ otherwise. Then the kernel

$$\tilde{\mathcal{P}}(x,dy) = g\left\{\frac{\pi(dy)Q(y,dx)}{\pi(dx)Q(x,dy)}\right\} Q(x,dy) \tag{4}$$

satisfies (1). However, the kernel $\tilde{\mathcal{P}}$ is not necessarily Markov. One way of enforcing that (4) integrates to 1 is to restrict attention to $g \leqslant 1$, ensuring that $\tilde{\mathcal{P}}(x,\mathbb{X}) \leqslant 1$, and then combine this with $r(x,dy) = \{1 - \tilde{\mathcal{P}}(x,\mathbb{X})\}\delta_x(dy)$, where $\delta_x(A) = 1$ if $x \in A$ and 0 otherwise. The resulting kernel $\tilde{\mathcal{P}}(x,dy) + r(x,dy)$ is of Metropolis–Hastings form (e.g., Tierney, 1998).

An alternative strategy introduced by Power & Goldman (2019), Zanella (2020) and Livingstone & Zanella (2022) is to instead allow any g for which $\mathcal{Z}(x) = \tilde{\mathcal{P}}(x, \mathbb{X})$ is finite, and then set

$$\mathcal{P}(x, dy) = \frac{\tilde{\mathcal{P}}(x, dy)}{\mathcal{Z}(x)}.$$

The kernel $\mathcal{P}$ does not satisfy (1) in general; in fact, $\mathcal{P}$ is invariant with respect to the measure $\mathcal{Z}(x)\pi(dx)$. A π-invariant Markov jump process can be constructed, however, by introducing a holding time $\mathcal{Z}(x)$ at each state x and then choosing the next state according to $\mathcal{P}$. This construction is called a locally balanced Markov process; see Power & Goldman (2019) and Hird et al. (2022) for more details.

2.2. *First-order locally balanced processes*

The function $\mathcal{Z}(x)$ will not be tractable in general, so further work is needed to design a sampling algorithm based on a locally balanced Markov process. One approach is to restrict attention to symmetric Q and to π that is absolutely continuous with respect to Lebesgue measure on $\mathbb{R}^n$ with differentiable Lebesgue density $\pi(x)$. In this case, (3) reduces to $\pi(y)/\pi(x)$. From this point several natural first-order approximations of this ratio can be taken to construct a new, more tractable kernel. It is argued in Hird et al. (2022) and Livingstone & Zanella (2022) that a good choice is the componentwise approximation found by letting $Q(x, dy) = \prod_i \sigma^{-1}\mu\{(dy_i - x_i)/\sigma\}$, where μ is a centred and symmetric distribution on $\mathbb{R}$ and $\sigma > 0$, and defining

$$\tilde{P}(x, dy) = \prod_{i=1}^{n} g\big[\exp\{(y_i - x_i)\partial_i \log \pi(x)\}\big]\mu\left(\frac{dy_i - x_i}{\sigma}\right), \tag{5}$$

where $\partial_i = \partial/\partial x_i$ and $\int_A \mu\{(dy_i - x_i)/\sigma\} = \mu\{(A - x_i)/\sigma\}$ with $(A - x_i)/\sigma = \{z \in \mathbb{R} : x_i + \sigma z \in A\}$ for any event A, and its Markovian counterpart

$$P(x, dy) = \frac{\tilde{P}(x, dy)}{Z(x)}, \tag{6}$$

where $Z(x) = \tilde{P}(x, \mathbb{X})$. With this approximation, certain choices of g and each Q_i lead to familiar forms of P. Choosing $g(t) = \sqrt{t}$ and μ to be standard Gaussian, for example, leads to the unadjusted Langevin algorithm (Roberts & Tweedie, 1996). The class of kernels obtained by (6) is much broader, however, and is currently relatively unexplored.

2.3. *The choice of balancing function*

Livingstone & Zanella (2022) suggest the choice of balancing function $g(t) = 2t/(1 + t)$, as popularized by Barker (1965) in the context of the Metropolis–Hastings algorithm. With this choice, $Z(x) = 1$ and a sample from P can be easily drawn in the following manner. First sample $z_i \sim \mu$ for each i; then set $\beta_{x,i} = \partial_i \log \pi(x)$ and flip the sign of each z_i with probability $F(\beta_{x,i} z_i)$, where $F(x) = \exp(x)/\{1 + \exp(x)\}$; finally, add this to the current coordinate x_i. See Algorithm 1 for details. To construct a π-invariant Markov chain, a Metropolis–Hastings correction is then applied to this Barker proposal.

Algorithm 1. Simulating from the Barker proposal.

```
Require: current point $x \in \mathbb{X}$
 For $i = 1$ to $n$
     Draw $z_i \sim \mu$, and set $\beta_{x,i} \leftarrow \partial_i \log \pi(x)$
     Set $y_i \leftarrow x_i + z_i$ with probability $F(\beta_{x,i} z_i)$, and $y_i \leftarrow x_i - z_i$ otherwise
 Output $y = (y_1, \ldots, y_d)$
```

It is natural to wonder how many choices of g can be made. Two other simple possibilities are $\min(1, t)$ and $\max(1, t)$, the latter of which was studied in Choi (2020). The following results show that, in fact, the family of balancing functions is infinitely large.

Proposition 1. *Let $\mathcal{H} = \{h : \mathbb{R} \to [0, \infty) : h(x) = h(-x)\}$ be the space of positive even functions. Then for every $h \in \mathcal{H}$, $g_h(t) = t^{1/2} h(\log t)$ is a balancing function. Conversely, for every balancing function g, the function $h_g(x) = \exp(-x/2) g\{\exp(x)\}$ is contained in $\mathcal{H}$.*

Proposition 1 provides an explicit parameterization of g_h in terms of a specific $h \in \mathcal{H}$. The function $t^{1/2}$ can also be replaced with any other specific balancing function to give a different bijection. The goal, of course, is to find choices of g for which tractable sampling algorithms can be designed. In § 4 we design new balancing functions of this nature for specific objectives.

3. A general result on the optimal acceptance rate and scaling with dimension

3.1. *The asymptotic acceptance rate for locally balanced proposals*

Let $\pi : \mathbb{R} \to [0, \infty)$ be a probability density on $\mathbb{R}$, and for any fixed $\sigma > 0$ let $Q_\sigma : \mathbb{R} \times \mathbb{R} \to [0, 1]$ be a Markov kernel. We introduce the product measure $\pi_n(dx) = \prod_{i=1}^n \pi(x_i)\, dx_i$ on $\mathbb{R}^n$ and the product kernel $\mathcal{Q}_n(x, dy) = \prod_{i=1}^n Q_{\sigma_n}(x_i, dy_i)$, where $(\sigma_n)_{n \in \mathbb{N}}$ is a sequence of positive real numbers. On the set $\mathcal{R} \times \mathcal{R}$ as in § 2.1, let

$$\rho_n(X_{n,i}, Y_{n,i}) = \log\left\{\frac{\pi(y) Q_{\sigma_n}(y, dx)}{\pi(x) Q_{\sigma_n}(x, dy)}\right\},$$

where $X_n = (X_{n,1}, \ldots, X_{n,n}) \sim \pi_n$ and $Y_n \sim \mathcal{Q}_n(X_n, \cdot)$. The acceptance rate in a Metropolis–Hastings algorithm targeting π_n with proposal Q_n is $\alpha_n(X, Y) = 1 \wedge \exp\{\sum_{i=1}^n \rho_n(X_{n,i}, Y_{n,i})\}$. We will show that a central limit theorem holds for the ρ_n associated with first-order locally balanced Metropolis–Hastings under Assumption 1 below, and then consider optimal acceptance rates and dimension dependence in terms of the expected squared jump distance in each coordinate. Let $b(x) = \log[g\{\exp(x)\}]$, and without loss of generality set $g(1) = 1$.

Assumption 1. There exist constants $H \in (0, 1)$, $\gamma > 0$, $\beta \geqslant 0$ and $\epsilon > 0$ such that the following hold:

(i) $\pi_n(dx) = \prod_{i=1}^n \exp\{\phi(x_i)\}\, dx_i$ for some $\phi \in \mathcal{C}^{3+H}(\mathbb{R})$, and for $f = \phi''', \phi''\phi', \phi'^3$ or $\phi''|\phi'|^{1+\beta}$ the integrability condition $\int_{\mathbb{R}} f(x)^{2+\epsilon}\{1 + |\phi'(x)|^\beta\}\pi(x)\, dx < \infty$ and the mixed growth Hölder condition

$$|f(x + \delta) - f(x)| \leqslant K(x) \max(|\delta|^H, |\delta|^\gamma)$$

are satisfied, where the function K is such that $\int_{\mathbb{R}} K(x)^2\{1 + |\phi'(x)|^\beta\}\pi(x)\, dx < \infty$;

(ii) $g : \mathbb{R} \to \mathbb{R}$ is defined as in (2) with $g(1) = 1$, and $b(x) = \log[g\{\exp(x)\}]$ satisfies $b \in \mathcal{C}^3(\mathbb{R})$ with b', b'' and b''' all bounded above;

(iii) P is constructed as in (5) and (6) with g as in (ii) above and μ a centred and symmetric distribution on $\mathbb{R}$ satisfying $\int z^2 \mu(dz) = 1$ and $\int |z|^\xi \mu(dz) < \infty$ for $\xi = \max(6+3\epsilon, 2+2H, 2+2\gamma)$, and for all $a \in \mathbb{R}$ and some positive $C_\mu > 0$ one has

$$\int_{\mathbb{R}} \exp\{b(az)\}|z|^\xi \mu(dz) \leqslant C_\mu(1+|a|^\beta) \int_{\mathbb{R}} \exp\{b(az)\}\mu(dz) < \infty .$$

Part (i) of the assumption refers to the target distribution, part (ii) to the balancing function and part (iii) to the interplay between them. Part (i) is straightforwardly satisfied for many statistical models of interest, such as likelihoods from exponential families and suitably smooth priors. Part (iii) is satisfied by all the cases explicitly studied here, such as $g(t) = \surd t$ and $g(t) = 2t/(1+t)$. Part (iii) highlights the need to control the growth of g and b using the tails of μ. If g is bounded, as in the Barker case, then any μ with a moment generating function is sufficient for Assumption 1 to be satisfied, and for many targets actually only polynomial moments are required. When $g(t) = \surd t$, which is not bounded above, stronger conditions on the tails of μ are needed, such as Gaussian tails.

Part (i) of Assumption 1 is explicitly weaker than the typical smoothness assumptions made in the optimal scaling literature for product-form targets (e.g., Roberts & Rosenthal, 1998). A form of part (i) as well as the conditions $\int z^6 \mu(dz) < \infty$ and $g \in \mathcal{C}^3$ are crucial to the analysis. Part (iii) imposes uniform control with respect to x of measures $\exp[b\{\sigma_n z\phi'(x)\}]\{Z_{\sigma_n}(x)\}^{-1}\mu(z)\,dz$ in terms of only the measure $\mu(z)\,dz$. This is required so that the normalizing constants Z_{σ_n} and their second derivatives are well-defined. It may be possible to significantly relax part (ii) or (iii), especially in specific settings, at the expense of strengthening other conditions. The following proposition identifies some simple cases in which part (iii) is satisfied.

PROPOSITION 2. *Assumption* 1(iii) *is satisfied in the following cases:*

(i) *μ has a density with compact support, for any g;*

(ii) *g is bounded and nondecreasing, and $\int |z|^\xi \mu(dz) < \infty$ for ξ as in Assumption* 1*;*

(iii) *g satisfies Assumption* 1(ii) *and there exist $\tilde{C}_\mu, \tilde{\beta} > 0$ such that for all $a \in \mathbb{R}$,*

$$\int_{\mathbb{R}} \exp(az)|z|^\xi \mu(dz) \leqslant \tilde{C}_\mu(1+|a|^{\tilde{\beta}}) \int_{\mathbb{R}} \exp(az)\mu(dz) < \infty;$$

(iv) *g satisfies Assumption* 1(ii)*, μ has a density $\mu \in \mathcal{C}^1(\mathbb{R})$ such that $\lim_{z\to\pm\infty} \exp(az)\mu(z) = 0$ for any $a \in \mathbb{R}$, and there exist constants $p > 1$ and $A, B > 0$ for which*

$$|z|^p \mu(z) \leqslant A\mu(z) - Bz\mu'(z).$$

In specific examples we typically verify condition (i), (ii) or (iv) of Proposition 2. For instance, choices of the form $\mu(dz) \propto \exp(-|z|^p)\,dz$ for $p \geqslant 1$ satisfy (iv). A statement analogous to (ii), but for the function b is not valid. Even if g is bounded, b is only bounded away from infinity above, not below. In fact, since $b(x) = x + b(-x)$ holds, b can never be bounded. These conditions are required to analyse Taylor series remainder terms for the normalizing constant. It is apparent from Proposition 2 that fewer conditions on μ need to be assumed for the Barker proposal, in which g is bounded, than for the Langevin choice $g(t) = \surd t$.

THEOREM 1. *Under Assumption* 1, $\lim_{n\to\infty} \sigma_n^{-6} E(\rho_n^2) = \theta^2$ *for some* $\theta \in [0,\infty)$. *In addition, if* $\theta > 0$ *and* σ_n *is chosen such that* $\lim_{n\to\infty} n^{1/6}\sigma_n = \ell$, *then*

$$\sum_{i=1}^{n} \rho_n(X_{n,i}, Y_{n,i}) \Rightarrow N\left(-\frac{1}{2}\ell^6\theta^2,\ \ell^6\theta^2\right). \tag{7}$$

Writing $\mathfrak{g} = g''(1)$, $\mu_4 = \int_{\mathbb{R}} z^4\mu(dz)$, $\mu_6 = \int_{\mathbb{R}} z^6\mu(dz)$, $A_\phi = E_\pi\{(\phi''')^2\}$, $B_\phi = E_\pi\{(\phi'\phi'')^2\}$ *and* $C_\phi = E_\pi(\phi'\phi''\phi''')$, *the constant* θ^2 *takes the form*

$$\begin{aligned}\theta^2 = \mu_6 &\left\{\frac{1}{144}A_\phi + \left(\frac{1}{4}+\mathfrak{g}\right)^2 B_\phi - \frac{1}{6}\left(\frac{1}{4}+\mathfrak{g}\right) C_\phi\right\} \\ &+ \mu_4\left\{\frac{1}{6}\left(\frac{1}{2}+\mathfrak{g}\right) C_\phi - 2\left(\frac{1}{4}+\mathfrak{g}\right)\left(\frac{1}{2}+\mathfrak{g}\right) B_\phi\right\} + \left(\frac{1}{2}+\mathfrak{g}\right)^2 B_\phi.\end{aligned} \tag{8}$$

The specific choice of the scaling parameter $\sigma_n \propto n^{-1/6}$ in Theorem 1 is the only rate leading to a nontrivial distributional limit for $\sum_{i=1}^{n} \rho_n(X_{n,i}, Y_{n,i})$, despite the fact that $\lim_{n\to\infty} \sigma_n^{-6} E(\rho_n^2) = \theta^2$ holds for any decay rate. The expression for θ^2 depends on both the balancing function g and the distribution μ. In § 4 we consider optimal ways of choosing g and μ for certain purposes. We discuss some example choices below.

Example 1. In the Langevin case, $g(t) = \surd t$ and μ is standard Gaussian, so that $g''(1) = -1/4$, $\mu_4 = 3$ and $\mu_6 = 15$. Then

$$\theta^2 = \frac{5}{48}A_\phi + \frac{1}{8}C_\phi + \frac{1}{16}B_\phi,$$

which, if $\lim_{x\to\pm\infty} \exp\{\phi(x)\}\phi'(x)\phi''(x)^2 = 0$, can also be written using integration by parts as

$$\theta^2 = \frac{5}{48}E\{(\phi''')^2\} - \frac{1}{16}E\{(\phi'')^3\},$$

a formula that appears in Roberts & Rosenthal (1998).

Example 2. For the Barker proposal, $g(t) = 2t/(1+t)$ and μ can be any centred and symmetric distribution such that $\int z^6\mu(dz) < \infty$. With these choices, $g''(1) = -1/2$ and

$$\theta^2 = \frac{\mu_6}{144}(A_\phi + 6C_\phi + 9B_\phi). \tag{9}$$

An important consequence of Theorem 1, and in particular of (7), is a simple expression for the asymptotic acceptance rate for a first-order locally balanced Metropolis–Hastings algorithm; see, for example, Proposition 2.4 in Roberts et al. (1997).

COROLLARY 1. *Setting* $\alpha_n(X,Y) = 1 \wedge \exp\{\sum_{i=1}^{n} \rho_n(X_i, Y_i)\}$, *under the conditions of Theorem* 1,

$$\lim_{n\to\infty} E(\alpha_n) = 2\Phi(-\ell^3\theta/2),$$

where Φ *is the standard normal cumulative distribution function.*

3.2. *Optimal acceptance rates*

Given the simplified limiting expression for α_n in Corollary 1, we can examine optimal choices of the constant ℓ for a fixed θ, leading to an optimal acceptance rate. We consider optimizing the expected squared jump distance here, which is well studied and has a strong justification motivated by diffusion limits in various settings (Roberts & Rosenthal, 2001).

Using the same notation as above, denote by $(\mathcal{E}_n^{g,\mu})_{n\in\mathbb{N}}$ the sequence of expected squared jump distances for the first or any other coordinate, defined as

$$\mathcal{E}_n^{g,\mu} = E\big\{(Y_{n,1} - X_{n,1})^2\alpha(X_n, Y_n)\big\},$$

where $X_n \sim \pi_n$ and Y_n is generated from X_n using a first-order locally balanced proposal, as defined in (6), with distribution μ, balancing function g and variance parameter σ_n. Then we have the following result.

THEOREM 2. *Suppose that Assumption 1 and the conditions of Theorem 1 are satisfied for ϕ, μ, g and $\theta > 0$. Let $(\sigma_n)_{n\in\mathbb{N}}$ be a positive sequence with $\lim_{n\to\infty}\sigma_n = 0$. If either $\lim_{n\to\infty} n^{1/6}\sigma_n = 0$ or $\lim_{n\to\infty} n^{1/6}\sigma_n = \infty$, then as $n \to \infty$,*

$$n^{1/3}\mathcal{E}_n^{g,\mu} \to 0.$$

If $\lim_{n\to\infty} n^{1/6}\sigma_n = \ell$ for some $\ell \in (0,\infty)$, then as $n \to \infty$,

$$n^{1/3}\mathcal{E}_n^{g,\mu} \to h(\ell) = 2\ell^2\Phi(-\ell^3\theta/2),$$

where Φ is the standard normal cumulative distribution function on $\mathbb{R}$. Furthermore, there exists a unique optimal ℓ^*, which depends on g and μ, that maximizes $h(\ell)$, for which $2\Phi\{-(\ell^*)^3\theta/2\} \approx 0.574$. The corresponding optimal asymptotic efficiency satisfies

$$h(\ell^*) = C_h\theta^{-2/3},$$

where $C_h \approx 0.652$.

Theorem 2 shows that any first-order locally balanced Metropolis–Hastings algorithm will have the same asymptotic optimal acceptance rate of 0.57, and that algorithmic efficiency as measured by expected squared jump distance will scale as $O(n^{-1/3})$ for $n \to \infty$. This includes both the Barker and the Langevin proposals as well as many other possibilities. Theorem 2 also suggests a route to both comparison and optimal design of first-order locally balanced Metropolis–Hastings algorithms, in the former case by comparing θ^2 for different choices of μ and g, and in the latter by choosing μ and g so that θ^2 in Theorem 1 is minimized. According to the same theorem, under Assumption 1 the constant θ^2 will depend on ϕ through A_ϕ, B_ϕ and C_ϕ, on μ only through μ_4 and μ_6, and on g only through $\mathfrak{g} = g''(1)$. We explore optimal design under different constraints in the next section.

In the Langevin proposal case, the constant $h(\ell)$ was shown to correspond to the speed measure of an overdamped Langevin diffusion limit in Roberts & Rosenthal (1998). This additionally relates the Markov chain trajectories to the path of a diffusion process, unlike with the expected squared jump distance, which only optimizes one-step decorrelation of the coordinate functions. We conjecture that the same is true for locally balanced proposals in general, but proving a diffusion limit result explicitly would require additional technical assumptions and is beyond the scope of this paper.

Example 3. Consider the Gaussian target case, where $\phi(x) = -x^2/2$. Then $\phi'(x) = -x$, $\phi''(x) = -1$ and $\phi'''(x) = 0$, meaning that $A_\phi = C_\phi = 0$ and $B_\phi = E(x^2) = 1$. For Langevin proposals with $g(t) = \surd t$ and μ taken as Gaussian, the constant θ^2 in (8) becomes $\theta_{\rm L}^2 = 1/16$, whereas for the Barker choice $g(t) = 2t/(1+t)$ and the same μ we have $\theta_{\rm B}^2 = \mu_6/16$. The ratio of asymptotic expected squared jump distances is therefore $(\theta_{\rm B}/\theta_{\rm L})^{2/3} = \mu_6^{1/3}$. Here $\mu_6 = 15$, meaning that Langevin proposals are asymptotically $15^{1/3} \approx 2.47$ times more efficient than Barker proposals with Gaussian noise when optimally tuned. This is consistent with the experiments in Livingstone & Zanella (2022, § 5.2).

Example 4. Consider hyperbolic targets of the form $\phi(x) = (\delta^2 + x^2)^{1/2}$, with $\delta^2 = 0.1$ as in Livingstone & Zanella (2022). Then $A_\phi \approx 12.99$, $B_\phi \approx 0.22$ and $C_\phi \approx 1.68$. The same calculations as above imply that Langevin proposals are 1.18 times more efficient than Barker proposals with Gaussian noise when optimally tuned, which is also consistent with Livingstone & Zanella (2022, § 5.2).

4. Optimal choices among the class of locally balanced algorithms

4.1. *Optimal choice of noise in the Barker algorithm*

In this setting we fix $g(t) = 2t/(1+t)$ and minimize θ^2 with respect to μ, for a given but arbitrary choice of ϕ. In this case θ^2 is given by (9), and the only influence of μ comes from the sixth moment μ_6. The asymptotic expected squared jump distance can therefore be straightforwardly maximized by minimizing the sixth moment of μ subject to the constraint that $\mu_2 = 1$. By Jensen's inequality we have $\mu_6 \geqslant \mu_2^3 = 1$, and in fact the lower bound is uniquely attained by choosing μ to be a Rademacher distribution, such that if $W \sim \mu$ then $W = 1$ with probability 1/2 and $W = -1$ otherwise. We state this result formally as follows.

Proposition 3. *If $g(t) = 2t/(1+t)$, then θ^2 is minimized when $W \sim \mu$ is chosen to take values $+1$ and -1 each with probability* 1/2.

We can compare the relative efficiency of the Barker proposal with Rademacher versus Gaussian noise by using (9) in a similar manner to Examples 3 and 4. Doing this shows that for any ϕ the Rademacher version will be $\mu_6^{1/3} \approx 2.47$ times more efficient than the Gaussian version. It is particularly convenient that the optimal choice of μ does not depend in any way on ϕ and so generic methodological guidance can be provided for the algorithm. Comparison with the Langevin proposal is instead target dependent, as exemplified below.

Example 5. When $\phi(x) = -x^2/2$, as in Example 3, the Barker proposal with Rademacher noise will be exactly as efficient as the Langevin proposal. When $\phi(x) = (\delta^2 + x^2)^{1/2}$ with $\delta^2 = 0.1$, as in Example 4, the Rademacher proposal will be 2.08 times more efficient than the Langevin proposal.

We compare these theoretical results with empirical performances in § 5. The Rademacher version of the Barker proposal is clearly not practical, given that the resulting algorithm will not in general produce a π-irreducible Markov chain. This is an important limitation of using the expected squared jump distance as an efficiency criterion, which must be controlled for to ensure that sensible recommendations are given to the user. A pragmatic approach to the issue is to choose a distribution μ that is clearly π-irreducible, in such a way that will be visible on the time scales of a typical computer simulation, but which is similar in spirit to the Rademacher choice. One example is an evenly weighted mixture of two normal distributions centred at $\pm(1-\sigma^2)^{1/2}$, each with variance $\sigma^2 < 1$, but appreciably larger than zero so that irreducibility is no longer in question. The resulting approach, termed bimodal Barker, will satisfy $\mu_6 = 1 + 12\sigma^2 + 18\sigma^4 - 16\sigma^6$ and be $15^{1/3}\mu_6^{-1/3}$ times more efficient than the version with Gaussian noise. For small but nonnegligible σ, this is close to optimal while also being practical. For instance, for the choice $\sigma^2 = 0.1^2$, as used in the simulations below, bimodal Barker is approximately 2.37 times more efficient than the Gaussian version, compared to the optimum value of 2.47.

The result on Rademacher optimality may seem surprising at first given the lack of π-irreducibility. Similar results have, however, appeared previously; for example, it is known that the optimum expected squared jump distance for the random walk Metropolis algorithm, when the proposal distribution is spherically symmetric and the target is Gaussian is found by choosing the distribution to be uniform on a hypersphere of fixed radius from the current point (Neal & Roberts, 2011). Given the product form of π considered in this work, the Rademacher structure is therefore natural. For the random walk Metropolis algorithm, however, the benefits of choosing such an optimized proposal distribution vanish as the dimension increases (Neal & Roberts, 2011; Yang & Rodríguez, 2013), whereas in the case of the Barker and other locally balanced proposals they do not.

4.2. *Optimizing over the choice of balancing function for a fixed noise distribution*

In this subsection we turn our attention to the optimal choice of g for a fixed choice of μ. The expression (8) in this case becomes a simple quadratic in $\mathfrak{g}$, which can be straightforwardly solved to find an optimum choice for a given ϕ, as in (10) below.

PROPOSITION 4. *Given ϕ and a fixed noise distribution μ with finite fourth and sixth moments $\mu_4 < \mu_6 < \infty$, the optimum choice of $\mathfrak{g}$ is*

$$\mathfrak{g}^* = \frac{\mu_6(C_\phi - 3B_\phi) + \mu_4(9B_\phi - C_\phi) - 6B_\phi}{12B_\phi(\mu_6 - 2\mu_4 + 1)}. \tag{10}$$

Any family of balancing functions for which $\mathfrak{g} = g''(1)$ can be modified to take a desired value, could therefore in principle be used to create an optimized algorithm for a particular μ and ϕ. Consider the family

$$g_\gamma(t) = \frac{1}{2}\left(t^{1/2+\gamma} + t^{1/2-\gamma}\right), \tag{11}$$

indexed by $\gamma \geqslant 0$, where for $\gamma = 0$ we recover the Langevin case $g(t) = \sqrt{t}$. Any choice within the family is a balancing function and is such that $g_\gamma(1) = 1$ and $\mathfrak{g} = g_\gamma''(1) = \gamma^2 - 1/4$. For a given ϕ, the choice of γ can therefore be adjusted to achieve the optimum asymptotic efficiency provided that $\mathfrak{g}^*$ in (10) is larger than $-1/4$.

Given the results of the previous section, it would seem natural to set μ to be a Rademacher distribution; however, in this case it turns out that all choices of g give equivalent algorithms. This follows straightforwardly from the fact that (2) implies $g(t)/\{g(t)+g(t^{-1})\} = 1/(1+t^{-1})$, which is independent of g. In fact, Proposition 4 does not apply to the Rademacher case since $\mu_4 = \mu_6$. Another natural option is to fix μ to be standard Gaussian. In this case (10) implies that the maximum efficiency is found by choosing $\mathfrak{g} = C_\phi/(10B_\phi) - 1/5$. This scheme can be implemented using the family in (11), and sampling from the resulting first-order locally balanced proposal is straightforward as it consists of a mixture of two Gaussians; see the Supplementary

Material for details. We do not implement this scheme in the simulations, however, in favour of the more efficient alternatives discussed in the next subsection.

4.3. *Optimizing over the choice of both the noise distribution and the balancing function*

In this subsection we consider optimizing over both g and μ jointly. The following proposition identifies the best possibly achievable asymptotic efficiency with first-order locally balanced proposals for a given target.

PROPOSITION 5. *A nonnegative lower bound for θ^2 that is independent of both μ and g is*

$$\theta^2 \geqslant \frac{1}{144}\left(A_\phi - \frac{C_\phi^2}{B_\phi}\right). \tag{12}$$

Furthermore, θ^2 can be made arbitrarily close to the lower bound by choosing $\mu_4 > 1$ sufficiently close to 1*, setting $\mu_6 = \mu_4^2$ and taking*

$$\mathfrak{g} = \frac{\mu_4(C_\phi - 3B_\phi) + 6B_\phi}{12B_\phi(\mu_4 - 1)}. \tag{13}$$

Proof. Given $A_\phi > 0, B_\phi > 0$ and $C_\phi \in \mathbb{R}$, we must solve the constrained quadratic optimization problem of minimizing θ^2 subject to $1 \leqslant \mu_4 \leqslant \surd\mu_6$. The constraints on μ_4 and μ_6 are necessary because $1 = \mu_2 \leqslant \surd\mu_4$ by Jensen's inequality and $\mu_4 \leqslant (\mu_2\mu_6)^{1/2} = \surd\mu_6$ by Cauchy's inequality. Moreover, the Hamburger moment problem tells us that these constraints are sufficient: if they are satisifed, then there exists a symmetric proposal distribution on $\mathbb{R}$ that satisfies them.

Defining the new variables $m_1 = (\mathfrak{g} + 1/4)\phi'\phi'' - \phi'''/12$ and $m_2 = -(\mathfrak{g} + 1/2)\phi'\phi''$, we can rewrite θ^2 as

$$\theta^2 = (\mu_6 - \mu_4^2)E(m_1^2) + E\big\{(\mu_4 m_1 + m_2)^2\big\} \geqslant E\big\{(\mu_4 m_1 + m_2)^2\big\}, \tag{14}$$

where the inequality follows from $\surd\mu_6 \geqslant \mu_4$. Expressing this lower bound in terms of A_ϕ, B_ϕ and C_ϕ gives

$$\theta^2 \geqslant B_\phi\left\{\left(\mu_4\mathfrak{g} - \mathfrak{g} + \frac{\mu_4}{4} - \frac{1}{2}\right) - \frac{\mu_4}{12}\frac{C_\phi}{B_\phi}\right\}^2 + \frac{\mu_4^2}{144}\left(A_\phi - \frac{C_\phi^2}{B_\phi}\right),$$

which itself can be bounded from below, giving

$$\theta^2 \geqslant \frac{\mu_4^2}{144}\left(A_\phi - \frac{C_\phi^2}{B_\phi}\right) \geqslant \frac{1}{144}\left(A_\phi - \frac{C_\phi^2}{B_\phi}\right).$$

We have used three inequalities. The first, in (14), is realized if and only if $\mu_6 = \mu_4^2$; the second simply bounds a square from below by zero and is realized if and only if $\mathfrak{g}$ is defined as in (13), which requires $\mu_4 > 1$; the third relies on $\mu_4 \geqslant 1$ and is realized if and only if $\mu_4 = 1$. The last two equalities cannot be realized simultaneously. The final lower bound is always nonnegative because $B_\phi A_\phi \geqslant C_\phi^2$ by Cauchy's inequality. □

Let $\nu(a)$ for $a > 1$ denote a discrete symmetric distribution taking three possible values, $-\sqrt{a}$, 0 and $\sqrt{a}$, such that the probability of a nonzero value is $1/a$. This is the unique symmetric distribution μ with moments satisfying $\mu_2 = 1$, $\mu_4 = a$ and $\mu_6 = a^2$. Indeed, for such $W \sim \mu$ this is implied by the identity $E\{W^2(W^2-a)^2\} = \mu_6 - 2a\mu_4 + a^2 = 0$. Letting $\mathfrak{g}$ be defined by (13), choosing $\mu = \nu(\mu_4)$ and taking μ_4 arbitrarily close to 1 results in θ^2 becoming arbitrarily close to the lower bound (12). This three-point proposal results in an algorithm that achieves close-to-optimal asymptotic expected squared jump distance among the class of first-order locally balanced samplers provided that $\mathfrak{g}$ is chosen according to (13).

Remark 1. The choice of μ indicated above is in fact optimal for any fixed choice of g, but the amount of mass given to the point 0 will vary, depending on g. In the Barker case, for example, this point achieves no mass, resulting in the Rademacher choice for μ.

It is natural to consider taking the limit $\mu_4 \to 1$ and expect optimality to be reached there. When the dimension n is fixed and finite, however, this results in a Rademacher proposal, which is suboptimal. This can be seen by noting that the lower bound (12) is always smaller than $B_\phi/16 + A_\phi/144 + C_\phi/24$, the value attained by the Rademacher proposal, because

$$\frac{B_\phi}{16} + \frac{A_\phi}{144} + \frac{C_\phi}{24} = \frac{A_\phi}{144} + \left(\frac{B_\phi^{1/2}}{4} + \frac{C_\phi}{12B_\phi^{1/2}}\right)^2 - \frac{C_\phi^2}{144B_\phi} \geqslant \frac{1}{144}\left(A_\phi - \frac{C_\phi^2}{B_\phi}\right).$$

Inspecting the proof of Theorem 1 shows that $\mathfrak{g}$ must be increased sufficiently slowly as a function of n to control the remainder terms, in order for the asymptotic expression for θ^2 to be a valid representation of the expected squared jump distance. In other words, as $\mu_4 \to 1$, it takes increasingly large n for the asymptotic regime to be representative of the finite-n setting. For a finite n, it is therefore necessary to choose $\mu_4 > 1$. We explore this phenomenon further in the Supplementary Material. In all simulations below, we set $\mu_4 = 2$ unless stated otherwise.

A surprising consequence of these findings is that the three-point proposal with some mass at zero outperforms a Rademacher choice that is optimum for the Barker proposal when the freedom to choose $\mathfrak{g}$ is given. In terms of sampling, this suggests that efficiency gains can be made by allowing some components of the state to remain unchanged at each iteration of the algorithm with a probability that depends on the size of the gradient in that direction. The same family of balancing functions introduced in (11) can again be used to create this optimum sampler.

A particular case of interest is the Gaussian setting of $\phi(x) = -x^2/2$, where $\phi'''(0)$ and hence A_ϕ and C_ϕ are equal to 0. This means that by choosing any $\mu_4 > 1$ and $\mathfrak{g}$ according to (13) we can achieve zero asymptotic θ^2. The result is a super-efficient sampler whose efficiency will effectively decay at a slower rate than $n^{-1/3}$. We illustrate this surprising finding numerically in § 5, but we also stress that this property holds only when $\phi(x) = -x^2/2$, to the best of our knowledge.

5. Simulation study

5.1. *Efficiency with respect to dimension on product targets*

We examine the expected squared jump distance of the first component of two different product-form target distributions as a function of dimension. This setting is directly covered by the theoretical results of § 3 and § 4. The two target distributions considered are the multi-dimensional standard Gaussian distribution and the hyperbolic distribution of Example 4. In each case we compare the random walk Metropolis algorithm, the Metropolis-adjusted Langevin algorithm,

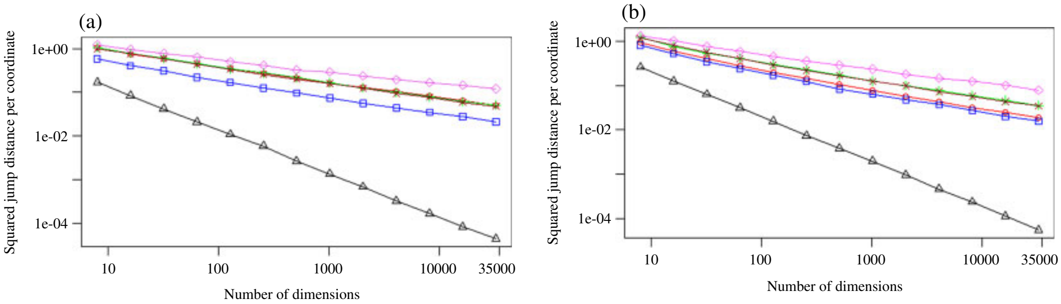

Fig. 1. Expected squared jump distance plotted against dimensionality for: (a) a Gaussian product target; and (b) a hyperbolic product target. The algorithms under comparison are the Metropolis-adjusted Langevin algorithm (circles), the random walk Metropolis algorithm (triangles), the Barker algorithm with Gaussian noise (squares), the Barker algorithm with Rademacher noise (+ symbols), the Barker algorithm with bimodal noise (× symbols) and the three-point proposal (diamonds).

the Barker proposal with Gaussian noise, Rademacher noise and bimodal noise as described in § 4.1, and the algorithm with the optimal choice over both balancing function and noise distribution as described in § 4.3, which will hereafter be referred to as the three-point proposal.

The results for the Gaussian target distribution are shown in Fig. 1(a). It is clear from the plots that, among the Barker algorithms, the Rademacher and bimodal versions are comparable and perform similarly to the Metropolis-adjusted Langevin algorithm, whereas the Barker algorithm with Gaussian noise has an expected squared jump distance that is lower by a factor of 2–2.5, in accordance with the theoretical value of 2.47. The three-point proposal performs best and appears to exhibit a slightly slower than $n^{-1/3}$ decay in expected squared jumping distance when the dimension in large. This is because in the special Gaussian case, the target θ^2 from (8) equals zero when the choices described in § 4.3 are made.

Results for the hyperbolic target are shown in Fig. 1(b). The main difference from the Gaussian example is that, now the Barker algorithms with Rademacher and bimodal noise both outperform the Langevin algorithm, as predicted by the theory in § 4.1. The three-point proposal is still the best-performing algorithm.

5.2. *Poisson random effects model*

For a realistic example in which the target distribution is not of product form, we compare algorithms on the Poisson random effects model described in Livingstone & Zanella (2022, § 6.3). We compare the Barker algorithm with bimodal noise, the Barker algorithm with Gaussian noise, the Langevin algorithm and the random walk Metropolis algorithm. The main purpose of this example is to assess whether or not the above theoretical guidelines for the noise distribution in the Barker algorithm lead to good choices, even when the target distribution does not have independent and identically distributed components.

The target distribution under consideration is a 51-dimensional posterior distribution, $p(\mu, \eta_1, \ldots, \eta_{50} \mid y)$, arising from a Poisson random effects model defined hierarchically by $\mu \sim N(0, 10^2)$, $\eta_i \mid \mu \sim N(\mu, \sigma_\eta^2)$ and $y_{ij} \mid \eta_i \sim \text{Po}\{\exp(\eta_i)\}$, independently for $i = 1, \ldots, 50$ and $j = 1, \ldots, 5$. In our experiment we generate the observed data $y = (y_{ij})_{ij}$ from the model likelihood, i.e., sampling $y_{ij} \sim \text{Po}\{\exp(\eta_i^*)\}$ independently where $\eta_1^*, \ldots, \eta_I^*$ are themselves generated independently from a $N(\mu^*, \sigma_\eta^2)$ distribution with $\mu^* = 5$. Here σ_η is a fixed value and two scenarios are considered: in the first we set $\sigma_\eta = 1$, and in the second we set $\sigma_\eta = 3$. Effectively, σ_η is a parameter that governs the heterogeneity across groups $i = 1, \ldots, 50$ in the hierarchy.

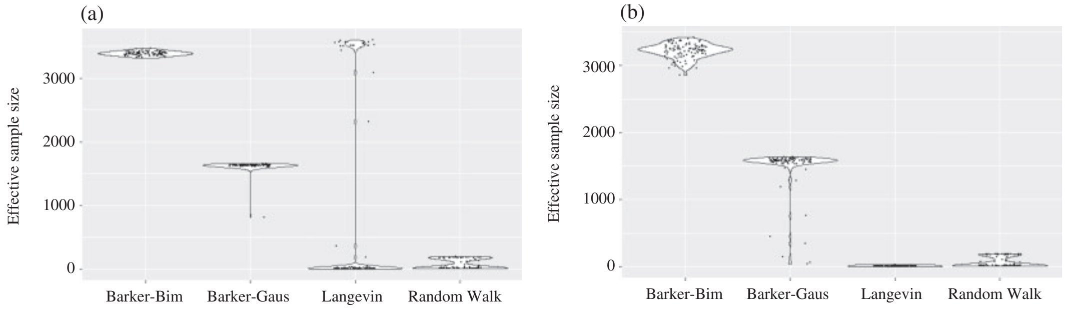

Fig. 2. Violin plots of median effective sample sizes across parameters for 100 independent repetitions of each algorithm: (a) low heterogeneity across coordinates; (b) high heterogeneity across coordinates. The algorithms under comparison are the Barker algorithm with bimodal noise (Barker-Bim), the Barker algorithm with Gaussian noise (Barker-Gaus), the Metropolis-adjusted Langevin algorithm (Langevin) and the random walk Metropolis algorithm (Random Walk).

Thus, larger values of σ_η lead to a target distribution with greater heterogeneity of scales across coordinates, which makes the adaptation and sampling process more challenging.

In each case, algorithmic tuning parameters consisting of a diagonal pre-conditioning matrix and a global scale are learned using Algorithm 4 of Andrieu & Thoms (2008), in the same manner as described in Livingstone & Zanella (2022, § 6.3). We measure efficiency in terms of the effective sample size for a given number of iterations, since all algorithms under comparison, apart from the random walk Metropolis algorithm, have a roughly equivalent cost per iteration, which is dominated by gradient computations. Figure 2 plots the median effective sample sizes across parameters for 100 independent runs of 5×10^4 iterations of each algorithm. All algorithms were randomly initialized by sampling parameter values from their prior distributions.

Both versions of the Barker algorithm appear to be more robust to different hyperparameter values than the Langevin algorithm, which in the first scenario performs well sometimes, but poorly at other times, and always performs poorly in the second scenario. This is because the Langevin algorithm is very sensitive to tuning parameter selection, and the adaptive procedure fails to converge on sensible values for these across the time scales of the simulation. The random walk Metropolis algorithm also performs poorly, which is largely explained by the dimension of the problem. The Barker algorithm with bimodal noise is approximately two times as efficient in terms of effective sample size as the version with Gaussian noise in this setting. More precisely, the median improvement in estimated effective sample size is 2.08 in the first scenario, with 10th and 90th quantiles across the 100 repetitions being 2.05 and 2.11, respectively, and 2.04 in the second scenario, with 10th and 90th quantiles 1.98 and 2.14, respectively. Similar values were obtained when looking at minimum rather than median effective sample sizes across parameters. These values suggest that the asymptotic theory developed in this paper, which quantifies bimodal Barker as being 2.37 times more efficient than Gaussian Barker, is highly predictive of behaviours observed in practice also for moderate dimensionality, and for targets that have neither independent nor identically distributed coordinates. More generally, in all our simulations, we consistently observed an improvement in efficiency when going from Gaussian to bimodal Barker with factors typically between 2 and 2.5.

5.3. *A correlated example*

Unlike the random walk or Langevin algorithms, the Barker and three-point schemes rely on a choice of coordinate system. This raises the question of how much their performance depends on

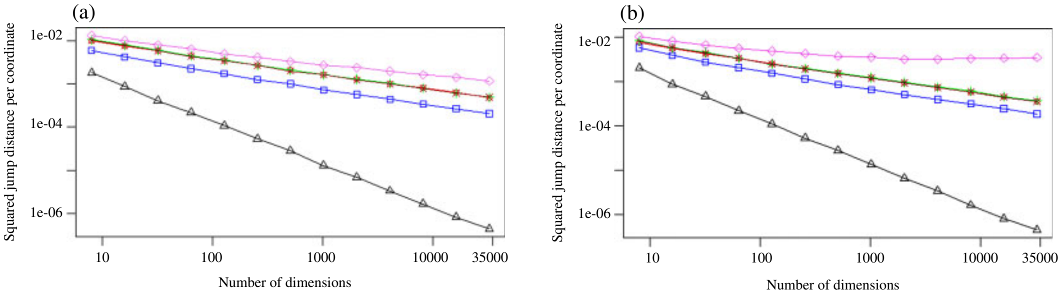

Fig. 3. Expected squared jump distance plotted against dimensionality for correlated Gaussian targets: (a) $\Sigma_{ii} = 1$ and $\Sigma_{ij} = 0.99$ for $i \neq j$; (b) $\Sigma_{ij} = 0.99^{|i-j|}$. The algorithms under comparison are the Metropolis-adjusted Langevin algorithm (circles), the random walk Metropolis algorithm (triangles), the Barker algorithm with Gaussian noise (squares), the Barker algorithm with Rademacher noise (+ symbols), the Barker algorithm with bimodal noise (× symbols) and the three-point proposal (diamonds).

the specific choice of coordinate system, and in particular, whether the $O(n^{1/3})$ scaling behaviour proved here is sensitive to the theoretical assumption that the target factorizes across the same coordinate axes as the proposal. Here we explore these issues numerically, performing high-dimensional scaling experiments similar to those in § 5.1, but for non-product-form targets with significant correlation. In particular, we consider Gaussian distributions with nondiagonal covariance matrix Σ chosen in two ways: in the first case we set $\Sigma_{ii} = 1$ for $i = 1, \ldots, n$ and $\Sigma_{ij} = \rho$ for $i \neq j$, and in the second case we take $\Sigma_{ij} = \rho^{|i-j|}$. In both cases we set $\rho = 0.99$ for a drastic departure from the independence case. As in § 5.1, we compute the expected squared jump distance per coordinate. For all algorithms under consideration we use isotropic proposals, meaning we do not use pre-conditioning to avoid aligning proposal and target axes, and we choose a step size that is numerically optimized to maximize performance as measured by expected squared jump distance. The results are reported in Fig. 3. As expected, all schemes perform worse than in the product case, as indicated by the different scales on the y-axes in Figs. 1 and 3, but the relative comparison between different schemes remains nearly unchanged and fully coherent with the theoretical predictions from § 3 and § 4. In particular, the Langevin, Barker bimodal and Barker Rademacher schemes perform nearly equivalently, while the Barker algorithm with Gaussian noise performs around 2–2.5 times worse. Overall, the experiment suggests that the relative performances of the random walk, Langevin and Barker algorithms are not particularly sensitive to correlation and to the specific choice of coordinate system.

The three-point proposal also performs well in these correlated examples and actually performs surprisingly well when $\Sigma_{ij} = \rho^{|i-j|}$. Gaining better understanding of such unexpected behaviour will be the subject of future research. However, the three-point proposal implicitly uses knowledge about the target distribution when choosing the optimal values of the tuning parameters $\mathfrak{g}$ and μ_4, and thus it has been given a somewhat unfair and potentially unrealistic advantage over the other schemes considered here. In particular, in this example $\mathfrak{g}$ was chosen according to the optimal value in (13) with $B_\phi = 1$ and $C_\phi = 0$ as given by product-form Gaussian targets.

6. Discussion

The main results of this paper rely on a product-form structure of π, and that the corresponding optimal choice of locally balanced algorithm also has a product form. We have shown in § 5 that this choice is still effective when the target distribution is no longer of product form, and therefore we recommend use of the bimodal Barker algorithm in practice. It is surprising that using a nonlocal noise distribution of this kind results in such a pronounced and consistent improvement in efficiency across multiple examples. We believe that this represents a good case study of theoretical analysis motivating new practical methodology that would otherwise be hard to devise. The product-form assumption is, as mentioned above, unrealistic in practice. It has been relaxed in various ways in the literature; see Sherlock (2013, § 1.1) for a review. Perhaps the most relevant question in the present setting is whether a different optimal locally balanced algorithm can be derived under different assumptions on the target distribution. We look forward to exploring this question in future work.

The detailed quantitative analysis and comparison of algorithms within the locally balanced class in the high-dimensional limit is made possible by the mathematical framework developed in Vogrinc & Kendall (2021, § 3). This framework identifies and uses only essential Taylor series expansions related to the limiting Kullback–Leibler divergence between a locally balanced proposal and its time reversal. Using this, we establish optimal scaling for a broad class of algorithms, including the Barker and Langevin algorithms, with a single unified proof, using significantly weaker assumptions on the smoothness and tails of the target distribution than those in Roberts & Rosenthal (1998). Our results are at present restricted to limiting expected squared jump distances, rather than diffusion limits as in Roberts et al. (1997) or Roberts & Rosenthal (1998). For the former, results can be obtained by clever, but elementary manipulations of expectations, while the latter require the study of convergence of generators or Dirichlet forms. It is therefore challenging to obtain diffusion limits under equally weak assumptions on the smoothness of the target. Establishing the latter in some sense justifies the use of the former as an efficiency metric, a point which is discussed in Roberts & Rosenthal (2001, § 2.2), since in this setting all efficiency measures are essentially equivalent. In addition, it implies pathwise convergence to a stochastic differential equation. Aside from this point, however, in practice establishing a diffusion limit would not bring much additional methodological insight. In the particular case of first-order locally balanced algorithms, we believe that the Markov chains have diffusion limits, but there are technical barriers to proving this that we are presently attempting to overcome.

One intriguing finding of this work concerns the suboptimality of the Langevin choice $g(t) = \sqrt{t}$ with noise μ chosen to be Gaussian. This is by far the most historically popular choice within the first-order locally balanced class of algorithms. The results in this paper show that, according to asymptotic efficiency as measured by expected squared jump distance, this combination of μ and g is not optimal, and in addition the optimum choice of μ when $g(t) = \sqrt{t}$ is not Gaussian, while the optimum choice of g when using Gaussian μ is not $\sqrt{t}$.

Acknowledgement

Vogrinc and Livingstone were supported by the U.K. Engineering and Physical Sciences Research Council.

Supplementary material

The Supplementary Material includes proofs of the theoretical results as well as further simulations related to the three-point proposal of § 4.3, illustrating a finite-dimensional example for which it is optimal to choose $\mu_4 > 1$.

References

Andrieu, C. & Thoms, J. (2008). A tutorial on adaptive MCMC. *Statist. Comp.* **18**, 343–73.

Barker, A. A. (1965). Monte Carlo calculations of the radial distribution functions for a proton-electron plasma. *Aust. J. Phys.* **18**, 119–34.

Beskos, A., Pillai, N., Roberts, G., Sanz-Serna, J.-M. & Stuart, A. (2013). Optimal tuning of the hybrid Monte Carlo algorithm. *Bernoulli* **19**, 1501–34.

Choi, M. C. H. (2020). Metropolis–Hastings reversiblizations of non-reversible Markov chains. *Stoch. Proces. Appl.* **130**, 1041–73.

Hird, M., Livingstone, S. & Zanella, G. (2022). A fresh take on 'Barker dynamics' for MCMC. In *International Conference on Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing (MCQMC 2020)*. Cham, Switzerland: Springer, pp. 169–84.

Livingstone, S. & Zanella, G. (2022). The Barker proposal: Combining robustness and efficiency in gradient based MCMC. *J. R. Statist. Soc.* B **84**, 496–523.

Neal, P. & Roberts, G. (2011). Optimal scaling of random walk Metropolis algorithms with non-Gaussian proposals. *Methodol. Comp. Appl. Prob.* **13**, 583–601.

Power, S. & Goldman, J. V. (2019). Accelerated sampling on discrete spaces with non-reversible Markov processes. *arXiv:* 1912.04681v2.

Roberts, G. O., Gelman, A. & Gilks, W. R. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *Ann. Appl. Prob.* **7**, 110–20.

Roberts, G. O. & Rosenthal, J. S. (1998). Optimal scaling of discrete approximations to Langevin diffusions. *J. R. Statist. Soc.* B **60**, 255–68.

Roberts, G. O. & Rosenthal, J. S. (2001). Optimal scaling for various Metropolis-Hastings algorithms. *Statist. Sci.* **16**, 351–67.

Roberts, G. O. & Tweedie, R. L. (1996). Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli* **2**, 341–63.

Sherlock, C. (2013). Optimal scaling of the random walk Metropolis: General criteria for the 0.234 acceptance rule. *J. Appl. Prob.* **50**, 1–15.

Sherlock, C., Fearnhead, P. & Roberts, G. O. (2010). The random walk Metropolis: Linking theory and practice through a case study. *Statist. Sci.* **25**, 172–90.

Tierney, L. (1998). A note on Metropolis-Hastings kernels for general state spaces. *Ann. Appl. Prob.* **8**, 1–9.

Vogrinc, J. & Kendall, W. S. (2021). Counterexamples for optimal scaling of Metropolis–Hastings chains with rough target densities. *Ann. Appl. Prob.* **31**, 972–1019.

Yang, Z. & Rodríguez, C. E. (2013). Searching for efficient Markov chain Monte Carlo proposal kernels. *Proc. Nat. Acad. Sci.* **110**, 19307–12.

Zanella, G. (2020). Informed proposals for local MCMC in discrete spaces. *J. Am. Statist. Assoc.* **115**, 852–65.

Zanella, G., Bédard, M. & Kendall, W. S. (2017). A Dirichlet form approach to MCMC optimal scaling. *Stoch. Proces. Appl.* **127**, 4053–82.

[*Received on* 5 *January* 2021. *Editorial decision on* 23 *September* 2022]