OPTIMAL SCALING OF THE RANDOM WALK METROPOLIS: GENERAL CRITERIA FOR THE 0.234 ACCEPTANCE RULE

CHRIS SHERLOCK,* *Lancaster University*

Abstract

Scaling of proposals for Metropolis algorithms is an important practical problem in Markov chain Monte Carlo implementation. Analyses of the random walk Metropolis for high-dimensional targets with specific functional forms have shown that in many cases the optimal scaling is achieved when the acceptance rate is approximately 0.234, but that there are exceptions. We present a general set of sufficient conditions which are invariant to orthonormal transformation of the coordinate axes and which ensure that the limiting optimal acceptance rate is 0.234. The criteria are shown to hold for the joint distribution of successive elements of a stationary pth-order multivariate Markov process.

Keywords: Random walk Metropolis; optimal scaling; optimal acceptance rate

2010 Mathematics Subject Classification: Primary 60F17
Secondary 65U05

1. Introduction

The Metropolis–Hastings updating scheme provides a very general class of algorithms for obtaining an approximate dependent sample from a target distribution, $\pi(\cdot)$, by constructing a Markov chain with π as its limiting distribution. Given the current value X, a new value X^* is proposed from a prespecified Lebesgue density $q(x^* \mid x)$ and is then accepted with probability $\alpha(x, x^*) = 1 \wedge (\pi(x^*)q(x \mid x^*))/(\pi(x)q(x^* \mid x))$. If the proposed value is accepted, it becomes the next current value ($X' \leftarrow X^*$); otherwise, the current value is left unchanged ($X' \leftarrow X$).

The d-dimensional *random walk Metropolis* (RWM) [8] has

$$q(x^* \mid x) = \lambda^{-d} r\left(\frac{x^* - x}{\lambda}\right) = \lambda^{-d} r\left(\frac{y^*}{\lambda}\right),$$

where $y^* := x^* - x$ is the proposed jump, and $r(y) = r(-y)$ for all y. In this case the acceptance probability simplifies to $\alpha(x, x^*) = 1 \wedge \pi(x^*)/\pi(x)$. Consider the behaviour of the RWM as a function of the scale of proposed jumps, λ, and some measure of the scale of variability of the target distribution, s. If $\lambda \ll s$ then, although proposed jumps are often accepted, the chain moves slowly and exploration of the target distribution is relatively inefficient. If $\lambda \gg s$ then many proposed jumps are not accepted, the chain rarely moves, and exploration is again inefficient. This suggests that, given a particular target and form for the jump proposal distribution, there may exist a finite scale parameter for the proposal with which the algorithm will explore the target as efficiently as is possible.

Previous theoretical work on this 'optimal scaling' problem has involved specific simple forms for the target distribution (see Section 1.1); more complex examples have been examined via simulation studies (see, e.g. [9]). For many of the targets considered, it has been found that, in the limit as $d \to \infty$, the optimal scaling is achieved when the acceptance probability is $\alpha_{\text{opt}} \approx 0.234$; although, for some specific classes of target, it has been shown that the limiting optimal acceptance rate is different to 0.234.

In this article we present general criteria which ensure that in the limit the optimal scaling will be achieved when an algorithm is tuned so that the acceptance rate is 0.234; the form of this scaling is also supplied. Checking these criteria for a given target is relatively simple, compared to proving the 0.234 rule from scratch. Their use is demonstrated for a multivariate pth-order Markov process with finite p.

1.1. Existing results for optimal scaling of the RWM

Much of the existing literature on the optimal scaling problem examines a sequence of stationary Metropolis algorithms on targets with increasing dimension, d. This program was first followed in [12] for product targets, $\pi(\boldsymbol{x}) = \prod_{i=1}^{d} f(x_i)$, explored by Gaussian jump proposals, $\boldsymbol{Y}^{(d)} \sim N(\boldsymbol{0}, \lambda_d^2 \boldsymbol{I}_d)$, where $\boldsymbol{I}_d$ denotes the d-dimensional identity matrix. Subject to certain moment conditions, as $d \to \infty$, a time-scaled, continuous-time limit, U_t, of the first component of the target satisfies a Langevin stochastic differential equation with stationary density f:

$$\mathrm{d}U_t = \tfrac{1}{2}\sigma^2 (\log f(U_t))' \mathrm{d}t + \sigma \, \mathrm{d}W_t. \tag{1}$$

Here $\mathrm{d}W_t$ represents a Brownian motion increment. The speed of this diffusion, σ^2, depends on the sequence of scale parameters, λ_d. Maximising σ leads to the following optimal value of the scale parameter:

$$\hat{\lambda}_d := \frac{\mu_{\text{opt}}}{(\tilde{M}^{(d)})^{1/2}}, \quad \text{where} \quad \tilde{M}^{(d)} := d\mathbb{E}[(\log f(X))'^2] \text{ and } \mu_{\text{opt}} \approx 2.38.$$

This result generalises to targets with components which are independent and which are identically distributed up to an arbitrary set of scalings; the 0.234 rule continues to hold provided that no one scale parameter is (see [2]), or is likely to be (see [11]), much smaller than any of the others. It is shown in [3] that if this condition fails then the optimal acceptance rate need not be 0.234, and indeed that the path of the limiting process need not even be continuous. Further specific classes of targets are considered in [5], [9], and [10], and in these cases the 0.234 rule is also shown to hold.

In [15] unimodal elliptically symmetric targets are explored using any proposal distribution, but using the expected squared jumping distance (ESJD) as a measure of efficiency. For consecutive draws from the chain, X and X', ESJD is defined to be $\mathbb{E}[\|X' - X\|^2]$, where expectation is with respect to the joint distribution of X and X' in a stationary chain (here and throughout this article, $\| \cdot \|$ refers to the Euclidean norm). The ESJD, which is also used as the efficiency criterion in [1] and [4], is proportional to the speed of any limiting diffusion (e.g. consider $\mathbb{E}[\mathrm{d}U_t^2]$ in (1)). Furthermore, since, for any stationary chain, $\mathbb{E}[\|X' - X\|^2] = 2 \sum_{i=1}^{n} \text{var}[X_i](1 - \text{cor}[X_i, X_i'])$, maximising the ESJD is equivalent to minimising a weighted sum of the lag-1 autocorrelations for the different components. The 0.234 rule is again shown to hold provided that the elliptical target is not too eccentric (defined according to a criterion which is almost identical to that in [2]). A further property is also required of the target: that

there exists a sequence of positive values $\{k_x^{(d)}\}$ such that

$$\frac{\|\boldsymbol{X}^{(d)}\|}{k_x^{(d)}} \overset{\mathbb{P}}{\to} 1. \tag{2}$$

Here the subscript x simply denotes that the sequence of positive values corresponds to the target sequence $\{\boldsymbol{X}^{(d)}\}$. In high dimensions (2) effectively confines the target to a hyperspherical shell, the width of which is much smaller than the radius. A similar condition (with a different sequence of positive values) is required on the proposal, although with mean-square convergence rather than convergence in probability. The key idea is that if both the size of a proposal and the curvature of the target 'shell' become deterministic, then so too will certain aspects of the behaviour of the RWM algorithm. Sherlock and Roberts [15] also considered targets for which the convergence in (2) is to some fixed and nondegenerate radial distribution. In this case it was shown that the limiting optimal acceptance rate is strictly less than 0.234.

It is clear from the above review that the 0.234 rule is common but not ubiquitous. In this article we use a generalisation of the ESJD (see Section 2.3) as the efficiency criterion and draw on the 'shell' intuition of [15] to provide general criteria such that the 0.234 rule is satisfied on exploration with any spherically symmetric proposal.

1.2. Outline of this article

The next section is devoted to a complete introduction to the notation which will be used in this article. All of the theoretical results are contained in Section 3 with all proofs deferred to Appendix A. In Section 4 we apply the criteria to a class of pth-order Markov processes (with p finite), of which the target with independent and identically distributed (i.i.d.) components considered in [12] is a special case; the results are discussed and placed in context in Section 5. Throughout this article, proposed jumps are assumed to arise from a spherically symmetric distribution which satisfies a shell condition.

2. Notation

2.1. Sequences of random variables

In this article $\{\boldsymbol{X}^{(d)}\}$ denotes the sequence of d-dimensional target distributions, and $\{\boldsymbol{U}^{(d)}\}$ the sequence of d-dimensional spherically symmetric proposals with unit scale parameter. The sequence of scale parameters will be denoted by $\{\lambda^{(d)}\}$, and the sequence of scaled proposals is therefore $\{\boldsymbol{Y}^{(d)}\} := \{\lambda^{(d)}\boldsymbol{U}^{(d)}\}$. Convergence in expectation is denoted by '$\overset{\mathbb{E}}{\to}$', and convergence in mean square by '$\overset{\text{m.s.}}{\to}$'. The distribution function of a standard Gaussian will be denoted by $\Phi(\cdot)$.

2.2. Derivatives of the target

For a given posterior $\pi(\boldsymbol{x})$, denote the first two derivatives of the log posterior as

$$M_i(\boldsymbol{x}) := \left.\frac{\partial \log \pi}{\partial x_i}\right|_{\boldsymbol{x}} \quad \text{and} \quad H_{ij}(\boldsymbol{x}) := -\left.\frac{\partial^2 \log \pi}{\partial x_i \partial x_j}\right|_{\boldsymbol{x}},$$

and the matrix of perturbations of the negative Hessian as

$$\boldsymbol{\Delta}(\boldsymbol{x}, \boldsymbol{u}) := \boldsymbol{H}(\boldsymbol{x} + \boldsymbol{u}) - \boldsymbol{H}(\boldsymbol{x}).$$

We require the following coordinate-free norms of the derivatives:

$$\tilde{M}(\boldsymbol{x}) := \|\nabla \log \pi\| = \|\boldsymbol{M}(\boldsymbol{x})\|, \qquad \tilde{H}(\boldsymbol{x}) := -\nabla^2 \log \pi = \text{trace}(\boldsymbol{H}(\boldsymbol{x})).$$

The eigenvalues of $\boldsymbol{H}(\boldsymbol{x})$ will be denoted by $\beta_1(\boldsymbol{x}), \ldots, \beta_d(\boldsymbol{x})$, and the maximum modulus as $\beta_{\max}(\boldsymbol{x}) := \max_{i=1,\ldots,d} |\beta_i(\boldsymbol{x})|$; note that $\sum_{i=1}^d \beta_i(\boldsymbol{x}) = \tilde{H}(\boldsymbol{x})$.

The superscript '(d)' should be applied to all of the above quantities since each can be applied to every random variable in the sequence $\{\boldsymbol{X}^{(d)}\}$; however, to simplify the notation, the superscript will be omitted from equations whenever its use is clear from the context.

Remark 1. The current position of the chain, $\boldsymbol{x}$, is a vector in the particular Cartesian coordinate system in which the problem is defined. However, the position and the movement of the chain is absolute in space and does not depend on the particular choice of coordinate system through which it is viewed; neither does the acceptance rate. Provided there has been no deformation of the coordinate grid, properties of angle, distance, and, thus, our criterion of optimality, the ESJD, will also be unaffected. Sufficient conditions for the 0.234 rule should therefore be invariant to orthonormal transformations of the coordinate system. This is not the case in, for example, [2], [11], or [12], where the form of the target is *defined* with respect to a particular Cartesian coordinate system. All of the conditions in this article are invariant to orthonormal transformations of the coordinate axes.

2.3. Acceptance rate and expected square jump distance

We denote the acceptance probability for a proposed jump of size $\boldsymbol{y}$ from position $\boldsymbol{x}$ by $\alpha(\boldsymbol{x}, \boldsymbol{y})$. We also denote a generalised potential square jump distance arising from proposal $\boldsymbol{y}$ by $S(\boldsymbol{y}) := \boldsymbol{y}^\top \boldsymbol{T} \boldsymbol{y}$, where $\boldsymbol{T}$ is some $d \times d$ positive definite and symmetric matrix, which we refer to as the jump distance matrix. The trace of this matrix will be denoted by $\tilde{T}$. The expected generalised squared jump distance then arises naturally as $\mathbb{E}[S(\boldsymbol{Y})\alpha(\boldsymbol{X}, \boldsymbol{Y})]$, where expectation is with respect to $\pi(\boldsymbol{x})q(\boldsymbol{y})$. We are interested in the limit as $d \to \infty$ of the acceptance rate and of some sensible rescaling $(c^{(d)})$ of the expected generalised squared jump distance

$$\lim_{d \to \infty} \mathbb{E}[\alpha^{(d)}(\boldsymbol{X}^{(d)}, \boldsymbol{Y}^{(d)})]$$

and

$$\lim_{d \to \infty} c^{(d)} \mathbb{E}[S^{(d)}(\boldsymbol{Y}^{(d)})\alpha^{(d)}(\boldsymbol{X}^{(d)}, \boldsymbol{Y}^{(d)})].$$

The generalisation from the Euclidean distance $\boldsymbol{y}^\top \boldsymbol{y}$ to $\boldsymbol{y}^\top \boldsymbol{T} \boldsymbol{y}$ effectively allows the speed of each component of the chain to be down-weighted by a measure of the scale of variability of the target along that component, so that the contribution to the total is proportional to the *relative* speed with which that component is explored.

3. Theoretical results

Two main theorems are presented in this section: the first provides sufficient conditions such that the limiting optimal acceptance rate is 0.234; the second simplifies the limiting formula for the ESJD subject to a slightly stronger set of conditions. Proofs of the results appear in Appendix A.

Throughout this article, it is assumed that the Markov chain is ergodic and that it has reached equilibrium. We consider additional criteria which ensure that the limiting expected acceptance rate of an optimally scaled stationary chain is 0.234.

Our approach uses a second-order Taylor expansion of the log-likelihood, and for this to be valid we assume, without further reference, that $\tilde{H}^{(d)}(\boldsymbol{X}^{(d)})$ is nonzero $\boldsymbol{X}^{(d)}$-almost everywhere, and that $H_{ij}^{(d)}(\boldsymbol{x}^{(d)})$ is a continuous function of $\boldsymbol{x}^{(d)}$ (see, e.g. [6]).

3.1. Fundamental results

From position X, split a specific proposed jump, y, into a component, y_1, which is parallel to $\nabla \log \pi$ and a component, y_2, which is perpendicular to $\nabla \log \pi$. Now

$$\log\left[\frac{\pi(X+y)}{\pi(X)}\right] = \log\left[\frac{\pi(X+y_1)}{\pi(X)}\right] + \log\left[\frac{\pi(X+y_1+y_2)}{\pi(X+y_1)}\right].$$

To first order, the first term depends on $\tilde{M}(x) = \|\nabla \log \pi\|$, whereas the second depends on 'how many contours' a tangential move is likely to cross, which in turn depends on both the curvature (represented by $\tilde{H}(X+y_1)$) and the gradient (represented by $\tilde{M}(X+y_1)$). If both $\tilde{M}(X)$ and $\tilde{H}(X)$ become, in some sense, deterministic, then so might the change in $\log \pi$; these requirements are embodied in the *weak shell conditions* on the target.

Use of the curvature and gradient at the current position to model movement to a new position is unlikely to be valid if these quantities change significantly on the scale of a proposed jump (e.g. if $\tilde{H}(x)$ and $\tilde{H}(x+y)$ are very different). The requirement that the quantities at x be representative of values over the likely jump region is embodied in the *relative variability condition* (5) below.

The term $\tilde{H}(X)$ represents an 'average' curvature. Intuitively, this should be applicable provided there is no particular direction where the effect on the target of a unit move in that direction is much larger than the effect of movement in any other direction. Thus, the scales of variability of π along each component of X should not be too dissimilar. The *first eccentricity condition* on the target ensures that the chance of such extreme behaviour diminishes to 0.

Condition 1. (Weak shell conditions for the target.) *We require that there are sequences $\{\tilde{M}^{(d)}\}$ and $\{\tilde{H}^{(d)}\}$ such that*

$$\frac{\tilde{M}^{(d)}(X^{(d)})}{\tilde{M}^{(d)}} \xrightarrow{\mathbb{P}} 1, \tag{3}$$

$$\frac{\tilde{H}^{(d)}(X^{(d)})}{\tilde{H}^{(d)}} \xrightarrow{\mathbb{P}} 1. \tag{4}$$

Condition 2. (Relative variability condition on H.) *Let $Z^{(d)} \sim N(0, I_d)$ be independent of all targets, X. For any fixed $\mu > 0$ and $\delta > 0$, we require that*

$$\mathbb{P}_{X,Z}\left(\frac{1}{\tilde{H}}\left|Z^\top \Delta\left(X, t\mu\left(\frac{\tilde{M}}{\tilde{H}}\right)Z\right)Z\right| < \delta \text{ for all } t \in [0, 1]\right) \to 1. \tag{5}$$

Condition 3. (First target eccentricity condition.) *We require that*

$$\frac{\beta_{\max}(X^{(d)})}{\sum_{i=1}^d \beta_i(X^{(d)})} \xrightarrow{\mathbb{P}} 0. \tag{6}$$

For a deterministic change in $\log \pi$, we also require that the distance of any tangential jump become deterministic, and that the distribution of the component of the proposal parallel to $\nabla \log \pi$ approach some known distribution function. Both of these requirements are met if a *shell condition for the proposal* is imposed. We also require an *eccentricity condition on the jump distance matrix* which ensures that there is no one component whose contribution to the ESJD outweighs that of all of the others. The ESJD is an expectation of a quadratic form on the proposal, and, hence, the shell condition on the proposal must be slightly stronger than for the target.

Condition 4. (Shell condition for the proposal.) *We assume that there exists a sequence $\{k_u^{(d)}\}$ such that the sequence of spherically symmetric proposals satisfies the following shell condition:*

$$\frac{\|\boldsymbol{U}^{(d)}\|}{k_u^{(d)}} \overset{\text{m.s.}}{\to} 1. \tag{7}$$

Condition 5. (Eccentricty of the jump distance matrix.) *Let $\{\tau_i(d), i = 1, \dots, d\}$ be the triangular sequence of eigenvalues associated with the sequence of matrices $\boldsymbol{T}^{(d)}$, and let $\tau_{\max}(d) := \max_{i=1,\dots,d} \tau_i(d)$. We require that*

$$\frac{\tau_{\max}(d)}{\sum_{i=1}^d \tau_i(d)} \to 0. \tag{8}$$

We may now state our first result.

Theorem 1. *Let the target satisfy the weak shell conditions (3) and (4), the relative variability condition (5), and the eccentricity condition (6). Let the distance matrix satisfy the eccentricity condition (8), and let the proposal satisfy the strong shell condition (7). For fixed μ, set*

$$\lambda^{(d)} = \mu \frac{d^{1/2} \tilde{M}^{(d)}}{k_u^{(d)} \tilde{H}^{(d)}}.$$

The expected acceptance rate and generalised ESJD now satisfy

$$\lim_{d\to\infty} \mathbb{E}[\alpha(\boldsymbol{X}^{(d)}, \boldsymbol{Y}^{(d)})] = 2\Phi\left(-\tfrac{1}{2}\mu\right), \tag{9}$$

$$\lim_{d\to\infty} \frac{(\tilde{H}^{(d)})^2}{(\tilde{M}^{(d)})^2 \tilde{T}^{(d)}} \mathbb{E}[S^{(d)}(\boldsymbol{Y}^{(d)})\alpha(\boldsymbol{X}^{(d)}, \boldsymbol{Y}^{(d)})] = 2\mu^2\Phi\left(-\frac{1}{2}\mu\right). \tag{10}$$

Equation (9) is identical to the acceptance rate equation in [12] and many of the other references in Section 1.1, as (up to a multiplicative constant) is the efficiency equation (10). Optimising (10) for μ and substituting into (9) therefore leads to the same limiting optimal scaling and acceptance rate.

Corollary 1. *Subject to the conditions in Theorem 1, the limiting optimal acceptance rate for the RWM algorithm is approximately 0.234, which is achieved when $\mu = \mu_{\text{opt}} \approx 2.38$.*

3.2. Stronger assumptions

Strengthening the two target shell conditions, and adding a regularity condition allows us to relate $\tilde{M}$ and $\tilde{H}$.

Condition 6. (Strong shell conditions for the target.) *We require that*

$$\frac{\tilde{M}^{(d)}(\boldsymbol{X}^{(d)})}{\tilde{M}^{(d)}} \overset{\text{m.s.}}{\to} 1, \tag{11}$$

$$\frac{\tilde{H}^{(d)}(\boldsymbol{X}^{(d)})}{\tilde{H}^{(d)}} \overset{\mathbb{E}}{\to} 1. \tag{12}$$

Condition 7. (Regularity condition on the target.) *For each i, with all the components of $\boldsymbol{x}$ fixed except the ith, we require that, as $|x_i| \to \infty$,*

$$\frac{\partial \pi}{\partial x_i} \to 0. \tag{13}$$

This condition implies that $\mathbb{E}[(\partial \log \pi / \partial x_i)^2] = -\mathbb{E}[\partial^2 \log \pi / \partial x_i^2]$, provided that either side exists and is finite. Trivially, we then have the following result.

Proposition 1. *Subject to (11), (12), and (13), $(\tilde{M}^{(d)})^2 / \tilde{H}^{(d)} \to 1$.*

Given (11), and the Cauchy–Schwarz inequality if $i \neq j$, the left-hand expectation must be finite for all i, j, and, hence (given (13)), so is the right-hand expectation. Thus, we may define $\boldsymbol{H} := \mathbb{E}[\boldsymbol{H}(\boldsymbol{X})]$, with all of the components finite. Given (11) and (12), we also choose

$$\tilde{M} := \mathbb{E}_X[\tilde{M}(\boldsymbol{X})] \quad \text{and} \quad \tilde{H} := \text{trace}(\boldsymbol{H}) = \mathbb{E}_X[\tilde{H}(\boldsymbol{X})].$$

Finally, let $\beta_1, \ldots, \beta_d$ be the eigenvalues of $\boldsymbol{H}$, with $\beta_{\max} := \max_{i=1,\ldots,d} \beta_i$; clearly, $\sum_{i=1}^d \beta_i = \tilde{H}$. Given that the Hessian of the log target now has a finite first moment, we might consider using the expected Hessian matrix $\boldsymbol{H} := \mathbb{E}[\boldsymbol{H}(\boldsymbol{X})]$ as our jump distance matrix $\boldsymbol{T}$. The eccentricity condition on the Hessian (6) ensures that the probability that $\boldsymbol{H}(\boldsymbol{X})$ is too eccentric dwindles to 0 as $d \to \infty$; in addition, we now forbid the modulus of any one eigenvalue to exceed some fixed multiple of the modulus of their total.

Condition 8. (Second target eccentricity condition.) *We assume that there exists $k > 0$ such that, for each d,*

$$\mathbb{P}\left(\left| \frac{\beta_{\max}(\boldsymbol{X}^{(d)})}{\sum_{i=1}^d \beta_i(\boldsymbol{X}^{(d)})} \right| \leq k \right) = 1. \tag{14}$$

This condition is satisfied, for instance, if all of the eigenvalues are nonnegative $\boldsymbol{X}^{(d)}$-almost everywhere. It is not satisfied on well-behaved targets with local minima; however, $\boldsymbol{H}$ would not be representative of any global structure on such targets and so would not make a sensible jump distance matrix in any case.

We may now obtain a neater form for the limiting optimal scaling and ESJD.

Theorem 2. (a) *Let the target satisfy the strong shell conditions (11) and (12), the relative variability condition (5), the first eccentricity condition (6), and the regularity condition (13). Let the distance matrix $\boldsymbol{T}^{(d)}$ satisfy the distance matrix eccentricity condition (8), and let the proposal satisfy the shell condition (7).*

For fixed $\mu > 0$, set

$$\lambda^{(d)} = \mu \frac{d^{1/2}}{k_u^{(d)}} \frac{1}{(\tilde{H}^{(d)})^{1/2}}.$$

Subject to the above conditions, the expected acceptance rate satisfies (9), and the generalised ESJD satisfies

$$\lim_{d \to \infty} \frac{\tilde{H}^{(d)}}{\tilde{T}^{(d)}} \mathbb{E}[S^{(d)}(\boldsymbol{Y}^{(d)}) \alpha(\boldsymbol{X}^{(d)}, \boldsymbol{Y}^{(d)})] = 2\mu^2 \Phi\left(-\frac{1}{2}\mu \right). \tag{15}$$

(b) *If, in addition, the target satisfies the second eccentricity condition (14) then the distance matrix can be chosen to be $\boldsymbol{H}^{(d)}$. In this case the expected acceptance rate continues to satisfy (9), and the generalised ESJD satisfies (15) with $\tilde{T}^{(d)} = \tilde{H}^{(d)}$.*

Remark 2. In some sense $\boldsymbol{H}^{(d)}$ therefore provides a natural scale for the expected square jump distance. Indeed, for an elliptical target, choosing $\boldsymbol{T} = \boldsymbol{H}$ is equivalent to using the expected squared Mahlanobis distance of a jump as the measure of efficiency.

4. Example

We now consider an example class of target, the stationary pth-order univariate Markov process, and show that the strong shell conditions (11) and (12), the eccentricity condition (6), and the relative variability condition (5) all hold. With an assumption of the regularity condition (13), the target then satisfies all of the conditions required for part (a) of Theorem 2.

The univariate Markov process is a special case of the Gibbs distribution considered in [5]; however, we require much weaker conditions than those given in [5]; in particular we do not need a 'strongly mixing' condition, nor do we need to specify an absence of phase transitions. We also allow the more general form of proposal discussed in Section 3, and then generalise the whole result to a multivariate pth-order Markov process.

Consider a pth-order univariate Markov process with conditional density $f(x_{p+1} \mid x_1, \ldots, x_p)$, and denote the stationary distribution of any p adjacent components as $f^*(x_1, \ldots, x_p)$. We will consider the d-dimensional target

$$\pi(\boldsymbol{x}) = f^*(x_1, \ldots, x_p) f(x_{p+1} \mid x_1, \ldots, x_p) \cdots f(x_d \mid x_{d-p}, \ldots, x_{d-1}).$$

Define $\boldsymbol{x}_m := (x_m, \ldots, x_{m+p})$ and $g(\boldsymbol{x}_m) := \log f(x_{m+p} \mid x_m, \ldots, x_{m+p-1})$, and, hence,

$$g_k(\boldsymbol{x}_1) := \frac{\partial g}{\partial x_k} \quad \text{and} \quad g_{km}(\boldsymbol{x}_1) := -\frac{\partial^2 g}{\partial x_k \partial x_m}, \qquad 1 \le k, m \le p+1.$$

The following definitions will be helpful:

$$E_{km} := \mathbb{E}[g_k(\boldsymbol{X}_1) g_m(\boldsymbol{X}_{1+|m-k|})], \qquad \tilde{E}_{km} := \mathbb{E}[g_{km}(\boldsymbol{X}_1)],$$

$$E_1^* := \sum_{k=0}^{p} \sum_{m=0}^{p} E_{k+1,m+1}, \quad \text{and} \quad E_2^* := \sum_{k=0}^{p} \tilde{E}_{k+1,k+1}.$$

Here the expectations are taken over the stationary distribution of the process. We require the *moment conditions*

$$\mathbb{E}[|g_k(\boldsymbol{X})|^2] < \infty \quad \text{and} \quad \mathbb{E}[|g_{km}(\boldsymbol{X})|^2] < \infty.$$

We also require a *Lipschitz condition* on each of the second derivatives: $|g_{km}(\boldsymbol{x}_1) - g_{km}(\boldsymbol{x}_1')| \le b\|\boldsymbol{x}_1 - \boldsymbol{x}_1'\|$, $1 \le k, m \le p+1$. We make similar moment and Lipschitz assumptions on the derivatives of $\log f^*$. Since f^* depends only on the first p components, it does not contribute to the limits which we consider, and is ignored henceforth. Now

$$M_i(\boldsymbol{x}) = \frac{\partial \log \pi}{\partial x_i} = \sum_{k=0}^{(p \wedge (i-1))} g_{k+1}(\boldsymbol{x}_{i-k})$$

$$\text{and} \quad H_{ij}(\boldsymbol{x}) = \frac{\partial^2 \log \pi}{\partial x_i \partial x_j} = \sum_{k=0}^{(p-|i-j|) \wedge ((i \wedge j)-1)} g_{k+1}(\boldsymbol{x}_{(i \wedge j)-k}).$$

However, slight differences in form when $i \wedge (d-i) \le p$ or $j \wedge (d-j) \le p$ play no part in the limit results, and for clarity of exposition are henceforth ignored. With this notational simplification,

$$\tilde{M}(\boldsymbol{X})^2 = \sum_{i=1}^{d} \sum_{k=0}^{p} \sum_{m=0}^{p} g_{k+1}(\boldsymbol{X}_{i-k}) g_{m+1}(\boldsymbol{X}_{i-m}) \quad \text{and} \quad \tilde{H}(\boldsymbol{X}) = \sum_{i=1}^{d} \sum_{k=0}^{p} g_{k+1,k+1}(\boldsymbol{X}_{i-k}).$$

The finite order, p, of the Markov chain leads to two important properties.

1. For $|i - j| > p$, $H_{ij} = 0$ and, hence, $\Delta_{ij} = 0$. Both $\boldsymbol{H}$ and $\boldsymbol{\Delta}$ therefore have at most $(2p + 1)d$ nonzero terms rather than d^2.

2. For $|i - j| \le p$, H_{ij} depends only on the components of $\boldsymbol{x}$ between $(i \vee j) - p$ and $(i \wedge j) + p$. Thus, the Lipschitz condition on the second derivatives implies that

$$\left| \Delta_{ij}\left(\boldsymbol{X}, t\mu\left(\frac{\tilde{M}}{\tilde{H}}\right)\boldsymbol{Z} \right) \right| \le bt\mu \frac{\tilde{M}}{\tilde{H}} \|\boldsymbol{Z}_i\|,$$

where $\boldsymbol{Z}_i = (Z_{i-p}, \ldots, Z_{i+p})$. This estimate is conservative when $p \ne 0$.

Shell conditions. By the assumption of stationarity, the distribution of $\boldsymbol{X}_i$ is independent of i, and so the moment assumptions together with Scheffe's lemma give

$$\frac{1}{d^{1/2}} \tilde{M}(\boldsymbol{X}) \overset{\text{m.s.}}{\to} E_1^{*1/2} \quad \text{and} \quad \frac{1}{d} \tilde{H}(\boldsymbol{X}) \overset{\mathbb{E}}{\to} E_2^*.$$

Hence, the strong shell conditions hold with $\tilde{M}^{(d)} := d^{1/2} E_1^{*1/2}$ and $\tilde{H}^{(d)} := d E_2^*$.

Eccentricity condition. From property 1, define

$$Q_i(\boldsymbol{X}) := \sum_{j=1}^d |H_{ij}(\boldsymbol{X})| = \sum_{j=i-p}^{i+p} \left| \sum_{k=0}^{p-|i-j|} g_{k+1, k+1+|i-j|}(\boldsymbol{x}_{(i \wedge j)-k}) \right|,$$

and let $Q^* := \mathbb{E}[Q_i(\boldsymbol{X})]$. Note that $Q^{**} := \mathbb{E}[Q_i(\boldsymbol{X})^2] < \infty$ by the moment assumptions above, since p is fixed and finite. By Chebyshev's inequality, $\mathbb{P}(|Q_i - Q^*| > d^\alpha) < Q^{**} d^{-2\alpha}$, and, hence,

$$\mathbb{P}\left(\max_{i=1,\ldots,d-p} |Q_i - Q^*| > d^\alpha \right) < \frac{Q^{**}}{d^{2\alpha-1}}.$$

Thus,

$$\mathbb{P}\left(\max_{i=1,\ldots,d} \sum_{j=1}^d |H_{ij}(\boldsymbol{X})| > Q^* + d^\alpha \right) < \frac{Q^{**}}{d^{2\alpha-1}}.$$

The largest magnitude of any eigenvalue of $\boldsymbol{H}(\boldsymbol{X})$ cannot be greater than

$$\max_{i=1,\ldots,d} \sum_{j=1}^d |H_{ij}(\boldsymbol{X})|.$$

Also, $\mathbb{P}(2\tilde{H}(\boldsymbol{X}) > dE_2^*) \to 1$, from the second shell condition, (4). Choose any $\alpha \in (0.5, 1)$ to complete the proof, since this implies that

$$\mathbb{P}\left(\frac{\beta_{\max}(\boldsymbol{X})}{\tilde{H}(\boldsymbol{X})} \le \frac{Q^* + d^\alpha}{dE_2^*/2} \right) \to 1.$$

Relative variability condition. Define

$$T := \sum_{i=1}^d \sum_{j=i-p}^{i+p} |Z_i||Z_j| \|\boldsymbol{Z}_i\|.$$

The general inequality $(\sum_{j=i-p}^{i+p} |Z_j|)^2 \leq (2p+1) \sum_{j=i-p}^{i+p} |Z_j|^2$ implies that

$$T \leq (2p+1)^{1/2} \sum_{i=1}^{d} |Z_i| \sum_{j=i-p}^{i+p} |Z_j|^2 = (2p+1)^{1/2} \sum_{j=-p}^{p} \sum_{i=1}^{d} |Z_i||Z_{i+j}|^2.$$

For fixed j,

$$\frac{1}{d} \sum_{i=1}^{d} |Z_i||Z_{i+j}|^2 \xrightarrow{\mathbb{P}} \begin{cases} \mathbb{E}[|Z_1|^3], & i = j, \\ \mathbb{E}[|Z_1|Z_2^2], & i \neq j, \end{cases}$$

and so $T/d^\alpha \xrightarrow{\mathbb{P}} 0$ for any $\alpha > 1$. Thus, using properties 1 and 2, and since $t \in [0, 1]$,

$$\frac{1}{\tilde{H}} \left| \mathbf{Z}^\top \mathbf{\Delta}\left(X, t\mu\left(\frac{\tilde{M}}{\tilde{H}}\right)\mathbf{Z}\right)\mathbf{Z} \right| \leq b\mu\left(\frac{\tilde{M}}{\tilde{H}^2}\right)T = b\mu \frac{E_1^{*1/2}}{E_2^{*2}} \frac{1}{d^{3/2}} T \xrightarrow{\mathbb{P}} 0,$$

and the relative variability condition (5) is satisfied.

4.1. Extension

Consider a pth-order *multivariate* Markov chain where now the ith element is $(x_i^{(1)}, \ldots, x^{(r^*)})$ for some fixed $r^* \in \mathbb{Z}^+$, and the vector $\mathbf{x}_i$ now has $(p+1)r^*$ elements. The derivatives are now defined as

$$g_k^{(r)}(\mathbf{x}_1) := \frac{\partial g}{\partial x_k^{(r)}} \quad \text{and} \quad g_{km}^{(rs)}(\mathbf{x}_1) := -\frac{\partial^2 g}{\partial x_k^{(r)} \partial x_m^{(s)}}, \qquad 1 \leq k, m \leq p+1; \ 1 \leq r, s \leq r^*.$$

For each combination of r, s, k, and m, each derivative is subject to an analogous moment condition to that required for the univariate Markov chain for each combination of k and m, and the second derivatives are required to satisfy an analogous *Lipschitz* condition. The former ensures that the strong shell conditions are satisfied and the latter contributes towards the satisfaction of the relative variability condition. The important aspects of properties 1 and 2 do not change. For example, the row of the Hessian that corresponds to a given scalar component of the multivariate element of the Markov chain is now nonzero for at most $(2p+1)r^*$ (rather than $2p+1$) other scalar elements; this number is still fixed and finite. Thus, the proof that the stationary multivariate pth-order Markov chain also satisfies conditions (5), (6), (11), and (12) involves no new theoretical considerations, and merely an increased notational burden; it is therefore omitted.

5. Discussion

We have presented criteria on a sequence of targets and spherically symmetric proposals for random walk Metropolis algorithms which ensure that the limiting optimal acceptance rate is 0.234. Unlike previous conditions (except those in [15]), these criteria are invariant to orthonormal transformations of the coordinate system.

As $d \to \infty$, certain scalar summaries of the gradient and curvature of the log-target should become, in some sense, deterministic; furthermore, the curvature in any one direction should not be much more extreme than that in any of the other directions; finally, a scalar summary of curvature should show little relative variability over the distance of a typical proposed jump. The proposal should also be confined (in probability) to a shell which becomes infinitesimally thin

relative to its radius, and the quadratic form used to define distance (for efficiency calculations) should not be too biased in favour of any one direction.

Strengthening of the 'shell' conditions, and adding a regularity condition allow the expected negative Hessian to be used as in the quadratic form which defines the jump distance; this was shown to be a 'natural' scale on which to measure efficiency.

The above conditions generalise the 'shell' intuition provided in [15] for spherically symmetric targets, and clarify traits which are sufficient to guarantee deterministic behaviour in terms of scaling and acceptance rate.

The criteria are relatively easy to check for a given class of targets, when compared to proving from scratch that the deterministic behaviour holds. This is demonstrated for a finite-order multivariate Markov process, of which the i.i.d. target of [12] is a special case.

Appendix A. Proof of Theorem 1

A.1. Some preliminary results

The following limit results are proved in [14] for a sequence of d-dimensional spherically symmetric random variables $\{U^{(d)}\}$, subject to (7) or to the weaker condition

$$\frac{\|U^{(d)}\|}{k_u^{(d)}} \overset{\mathbb{P}}{\to} 1. \tag{16}$$

Lemma 1. *Subject to (16), for any unit vector $e^{(d)}$, which may be random, but is independent of $U^{(d)}$,*

$$\frac{d^{1/2}}{k_u^{(d)}} U^{(d)} \cdot e^{(d)} \overset{\mathrm{D}}{\to} N(0, 1).$$

Lemma 2. *Subject to (16), if the target eccentricity condition (6) holds then*

$$\frac{d}{(k_u^{(d)})^2} \frac{U^{(d)\top} H(X^{(d)}) U^{(d)}}{\tilde{H}(X^{(d)})} \overset{\mathbb{P}}{\to} 1.$$

Lemma 3. *Provided the strong shell condition on the proposal (7), and the eccentricity condition on the distance matrix (8) hold then*

$$\frac{\|U^{(d)}\|}{k_u^{(d)}} \overset{\mathrm{m.s.}}{\to} 1 \quad \Longleftrightarrow \quad \frac{d^{1/2}(U^{(d)\top} T^{(d)} U^{(d)})^{1/2}}{k_u^{(d)}(\tilde{T}^{(d)})^{1/2}} \overset{\mathrm{m.s.}}{\to} 1.$$

The following results are given in [13].

Lemma 4. *Let $\{U_d\}$ and $\{W_d\}$ be sequences of random variables such that*

$$U_d \overset{\mathrm{m.s.}}{\to} 1 \quad and \quad W_d \overset{\mathbb{P}}{\to} 1,$$

and let $\{G_d(\cdot)\}$ be a sequence of functions with $0 \leq G_d(w) \leq 1$. Then

$$\mathbb{E}[G_d(W_d)] \to c \quad \Longrightarrow \quad \mathbb{E}[U_d^2 G_d(W_d)] \to c.$$

The acceptance region of a Metropolis–Hastings algorithm is defined in [15] as $R_a(x) := \{y \in \mathbb{R}^d : \alpha(x, y) = 1\}$. The reversibilty of the Metropolis–Hastings algorithm is then shown to imply the following.

Lemma 5. *For a Metropolis–Hastings Markov chain which has proposal $q(\boldsymbol{y} \mid \boldsymbol{x})$, and which has reached equilibrium,*

$$\mathbb{E}[\alpha(\boldsymbol{X}, \boldsymbol{Y})] = 2 \int d\boldsymbol{x} \int_{\boldsymbol{y} \in R_a(\boldsymbol{x})} d\boldsymbol{y}\, \pi(\boldsymbol{x}) q(\boldsymbol{y} \mid \boldsymbol{x}). \tag{17}$$

Also, for any symmetric function $S(\boldsymbol{y})$,

$$\mathbb{E}[\alpha(\boldsymbol{X}, \boldsymbol{Y}) S(\boldsymbol{Y})] = 2 \int d\boldsymbol{x} \int_{\boldsymbol{y} \in R_a(\boldsymbol{x})} d\boldsymbol{y}\, S(\boldsymbol{y}) \pi(\boldsymbol{x}) q(\boldsymbol{y} \mid \boldsymbol{x}).$$

Expectations are with respect to $\pi(\boldsymbol{x}) q(\boldsymbol{y} \mid \boldsymbol{x})$.

Define k_u, $\tilde{M}$, and $\tilde{H}$ as in Section 3.1, and, hence,

$$R(\boldsymbol{X}, \boldsymbol{U}) := \frac{d}{k_u^2} \frac{\tilde{M}}{\tilde{M}(\boldsymbol{X})} \frac{1}{\tilde{H}} \int_0^1 dt\, (1 - t) \boldsymbol{U}^\top \boldsymbol{\Delta}\left(\boldsymbol{X}, \frac{d^{1/2}\tilde{M}}{k_u \tilde{H}} \boldsymbol{U}\right) \boldsymbol{U}. \tag{18}$$

Here $R(\boldsymbol{X}, \boldsymbol{U})$ is a multiple of the remainder term from a Taylor expansion of the log-posterior and, in the proof of Theorem 1, must be shown to vanish in probability. We now show that the relative variability condition (5) ensures this.

Lemma 6. *If the weak shell condition for the target (3) and the relative variability condition (5) are satisfied, and if the proposal satisfies (16), then $R^{(d)}(\boldsymbol{X}^{(d)}, \boldsymbol{U}^{(d)}) \xrightarrow{\mathbb{P}} 0$.*

Proof. Any spherically symmetric random variable, $\boldsymbol{U}$, can be decomposed into a uniform angular component and a radial distribution. We may therefore create an invertible map from any $\boldsymbol{U}$ with a continuous radial distribution function to a standard Gaussian of the same dimension, $\boldsymbol{Z}$, for example by

$$\|\boldsymbol{Z}^{(d)}\| = F_{\|\boldsymbol{Z}\|}^{-1}(F_{\|\boldsymbol{U}\|}(\|\boldsymbol{U}\|)),$$

where $F_{\|\boldsymbol{U}\|}(\cdot)$ and $F_{\|\boldsymbol{Z}\|}(\cdot)$ are the distribution functions of $\|\boldsymbol{U}\|$ and $\|\boldsymbol{Z}\|$, respectively. We then fix

$$\boldsymbol{Z} = \frac{\|\boldsymbol{Z}\|}{\|\boldsymbol{U}\|} \boldsymbol{U}. \tag{19}$$

In requiring that the remainder term (18) converge in probability to 0, the fraction $\tilde{M}/\tilde{M}(\boldsymbol{X})$ may be ignored provided (3) holds. Furthermore, by (19),

$$\frac{d}{(k_u^{(d)})^2} U_i U_j = \frac{\|\boldsymbol{U}\|^2}{k_u^{(d)}} \frac{d}{\|\boldsymbol{Z}\|^2} Z_i Z_j.$$

Thus, if the proposal shell condition (7) holds, and since $\|\boldsymbol{Z}\|/d^{1/2} \xrightarrow{\mathbb{P}} 1$, the requirement $R^{(d)}(\boldsymbol{X}^{(d)}, \boldsymbol{U}^{(d)}) \xrightarrow{\mathbb{P}} 0$ is equivalent to

$$R^{*(d)} := \frac{1}{\tilde{H}} \int_0^1 dt\, (1 - t) \boldsymbol{Z}^\top \boldsymbol{\Delta}\left(\boldsymbol{X}, t\mu\left(\frac{\tilde{M}}{\tilde{H}}\right)\boldsymbol{Z}\right) \boldsymbol{Z} \xrightarrow{\mathbb{P}} 0,$$

for which (5) is sufficient.

A.2. Proof of Theorem 1

It will be helpful to define

$$V^{(d)}(\mathbf{X}^{(d)}, \mathbf{U}^{(d)}) := \frac{d}{(k_u^{(d)})^2} \frac{\tilde{M}^{(d)}}{\tilde{M}^{(d)}(\mathbf{X}^{(d)})} \frac{1}{\tilde{H}^{(d)}} \mathbf{U}^{(d)\top} \mathbf{H}^{(d)}(\mathbf{X}^{(d)}) \mathbf{U}^{(d)}.$$

Combining (3), (4), (6), and (7), and applying Lemma 2, we find that $V^{(d)} \xrightarrow{\mathbb{P}} 1$. Now, from (17),

$$\mathbb{E}[\alpha] = 2\mathbb{E}_X[\mathbb{P}(Y \in R_a(X))],$$

and the acceptance region $R_a(\mathbf{x})$ has $\log \pi(\mathbf{x} + \mathbf{y}) - \log \pi(\mathbf{x}) > 0$. Taylor expansion with the integral form for the remainder (see, for example, [6, p. 86]) gives

$$\log \pi(\mathbf{x} + \mathbf{y}) - \log \pi(\mathbf{x}) = \mathbf{y} \cdot \mathbf{M}(\mathbf{x}) - \int_0^1 dt(1-t)\mathbf{y}^\top \mathbf{H}(\mathbf{x} + t\mathbf{y})\mathbf{y}$$

$$= \mathbf{y} \cdot \mathbf{M}(\mathbf{x}) - \frac{1}{2}\mathbf{y}^\top \mathbf{H}(\mathbf{x})\mathbf{y} - \int_0^1 dt(1-t)\mathbf{y}^\top \mathbf{\Delta}(\mathbf{x}, t\mathbf{y})\mathbf{y},$$

$$\frac{\log \pi(\mathbf{x} + \mathbf{y}) - \log \pi(\mathbf{x})}{\lambda \tilde{M}} = \mathbf{u} \cdot \hat{\mathbf{M}}(\mathbf{x}) - \frac{\lambda}{2\tilde{M}(\mathbf{x})}\mathbf{u}^\top \mathbf{H}(\mathbf{x})\mathbf{u}$$

$$- \frac{\lambda}{\tilde{M}(\mathbf{x})} \int_0^1 dt(1-t)\mathbf{u}^\top \mathbf{\Delta}(\mathbf{x}, t\lambda \mathbf{u},)\mathbf{u}$$

$$= \frac{k_u}{d^{1/2}} \left(\frac{d^{1/2}}{k_u} \mathbf{u} \cdot \hat{\mathbf{M}}(\mathbf{x}) - \frac{1}{2}\mu V(\mathbf{x}, \mathbf{u}) - R(\mathbf{x}, \mathbf{u}) \right),$$

where $\hat{\mathbf{M}} = \mathbf{M}/\tilde{M}$. So

$$\mathbb{P}(Y \in R_a(\mathbf{x})) = \mathbb{P}\left(\frac{d^{1/2}}{k_u} \mathbf{U} \cdot \hat{\mathbf{M}}(\mathbf{x}) > \frac{1}{2}\mu V(\mathbf{x}, \mathbf{U}) + R(\mathbf{x}, \mathbf{U}) \right).$$

Let $G^{(d)}(\cdot)$ be the distribution function of $(d^{1/2}/k_u^{(d)})\mathbf{U}^{(d)} \cdot \hat{\mathbf{M}}^{(d)}(\mathbf{X}^{(d)})$. Now Lemma 1 gives

$$\frac{d^{1/2}}{k_u^{(d)}} \mathbf{U}^{(d)} \cdot \hat{\mathbf{M}}^{(d)}(\mathbf{X}^{(d)}) \xrightarrow{\mathrm{D}} N(0, 1),$$

and so $G^{(d)}(\cdot) \to \Phi(\cdot)$. Since the $G^{(d)}$ are monotonic with identical finite upper and lower bounds, this convergence is uniform in the argument, and so, for a given $\varepsilon > 0$, we may choose d_1 such that, for all $d > d_1$ and any $t \in \mathbb{R}$, $|G_d(t) - \Phi(t)| < \varepsilon$.

For all bounded continuous functions $g(\cdot)$,

$$X_n \xrightarrow{\mathrm{D}} X \quad \Longleftrightarrow \quad \mathbb{E}[g(X_n)] \to \mathbb{E}[g(X)]$$

(see, e.g. [7, p. 316]). Hence, since $V^{(d)}(\mathbf{X}^{(d)}, \mathbf{U}^{(d)}) \xrightarrow{\mathbb{P}} 1$, and using Lemma 6, there must be a d_2 such that, for all $d > d_2$,

$$\mathbb{E}_{\mathbf{X}^{(d)}, \mathbf{U}^{(d)}} \left[\left| \Phi\left(-\tfrac{1}{2}\mu V^{(d)} - R^{(d)}\right) - \Phi\left(-\tfrac{1}{2}\mu\right) \right| \right] < \varepsilon.$$

For $d > \max(d_1, d_2)$, consider

$$t_d := \mathbb{E}_{\mathbf{X}^{(d)}\mathbf{Y}^{(d)}}[\alpha(\mathbf{X}^{(d)}\mathbf{Y}^{(d)})] - 2\Phi\left(-\tfrac{1}{2}\mu\right).$$

Application of the triangle inequality gives

$$
\begin{aligned}
|t_d| &= 2\big|\mathbb{E}_{\boldsymbol{X}^{(d)}, \boldsymbol{U}^{(d)}}\big[G_d\big(-\tfrac{1}{2}\mu V^{(d)}(\boldsymbol{X}^{(d)}, \boldsymbol{U}^{(d)}) - R^{(d)}(\boldsymbol{X}^{(d)}, \boldsymbol{U}^{(d)})\big) - \Phi\big(-\tfrac{1}{2}\mu\big)\big]\big| \\
&\leq 2\mathbb{E}_{\boldsymbol{X}^{(d)}, \boldsymbol{U}^{(d)}}\big[\big|G_d\big(-\tfrac{1}{2}\mu V^{(d)} - R^{(d)}\big) - \Phi\big(-\tfrac{1}{2}\mu V^{(d)} - R^{(d)}\big)\big|\big] \\
&\quad + 2\mathbb{E}_{\boldsymbol{X}^{(d)}, \boldsymbol{U}^{(d)}}\big[\big|\Phi\big(-\tfrac{1}{2}\mu V^{(d)} - R^{(d)}\big) - \Phi\big(-\tfrac{1}{2}\mu\big)\big|\big] \\
&< 2\varepsilon,
\end{aligned}
$$

proving (9). Now, by assumptions (7) and Lemma 3,

$$
\frac{1}{\mu} \frac{\tilde{H}^{(d)}}{\tilde{M}^{(d)}(\tilde{T}^{(d)})^{1/2}} S^{1/2}(\boldsymbol{Y}^{(d)}) = \frac{d^{1/2} S^{1/2}(\boldsymbol{U}^{(d)})}{k_u^{(d)}(\tilde{T}^{(d)})^{1/2}} \overset{\text{m.s.}}{\to} 1.
$$

Therefore, by (9) and Lemma 4,

$$
\frac{1}{\mu^2 \alpha_*} \frac{(\tilde{H}^{(d)})^2}{(\tilde{M}^{(d)})^2 \tilde{T}^{(d)}} \mathbb{E}_{\boldsymbol{X}^{(d)} \boldsymbol{Z}^{(d)}}\left[S(\boldsymbol{Z}^{(d)}) G_d\left(-\frac{1}{2}\mu V^{(d)}(\boldsymbol{X}^{(d)}, \boldsymbol{Z}^{(d)}) - R^{(d)}(\boldsymbol{X}^{(d)}, \boldsymbol{Z}^{(d)})\right)\right] \to 1,
$$

where $\alpha_* := \lim_{d\to\infty} \mathbb{E}[\alpha(\boldsymbol{X}^{(d)}, \boldsymbol{Y}^{(d)})]$. Equation (10) then follows directly.

A.3. Proof of Theorem 2

We require two simple preparatory results.

Proposition 2. *Consider two sequences of random variables A_d and B_d with $0 < A_d < 1$ almost everywhere. If*

$$
A_d \overset{\mathbb{P}}{\to} 0 \quad \text{and} \quad B_d \overset{\mathbb{E}}{\to} 1
$$

then $\mathbb{E}[A_d B_d] \to 0$.

Proof. By Holder's inequality,

$$
|\mathbb{E}[A_d(B_d - 1)]| \leq \operatorname{ess\,sup} |A_d| \mathbb{E}[|B_d - 1|] \leq \mathbb{E}[|B_d - 1|] \to 0.
$$

But, $|\mathbb{E}[A_d(B_d - 1)]| = |\mathbb{E}[A_d B_d] - \mathbb{E}[A_d]|$, and $\mathbb{E}[A_d] \to 0$.

Lemma 7. *If the first and second target eccentricity conditions (6) and (14), and the strong shell condition (12) hold, then*

$$
\frac{\beta_{\max}}{\sum_{i=1}^d \beta_i} \to 0.
$$

Proof. Let $\boldsymbol{e}_{\max}$ be the eigenvector associated with $\beta_{\max}$. Then

$$
\beta_{\max} = |\beta_{\max} \boldsymbol{e}_{\max}| = |\boldsymbol{H} \boldsymbol{e}_{\max}| = |\mathbb{E}[\boldsymbol{H}(\boldsymbol{X}) \boldsymbol{e}_{\max}]| \leq \mathbb{E}[|\beta_{\max}(\boldsymbol{X}) \boldsymbol{e}_{\max}|] = \mathbb{E}[\beta_{\max}(\boldsymbol{X})].
$$

Therefore,

$$
\frac{\beta_{\max}}{\mathbb{E}[\tilde{H}(\boldsymbol{X})]} \leq \frac{\mathbb{E}[\beta_{\max}(\boldsymbol{X})]}{\mathbb{E}[\tilde{H}(\boldsymbol{X})]} = \mathbb{E}\left[\frac{\beta_{\max}(\boldsymbol{X})}{\tilde{H}(\boldsymbol{X})} \frac{\tilde{H}(\boldsymbol{X})}{\tilde{H}}\right] \frac{\tilde{H}}{\mathbb{E}[\tilde{H}(\boldsymbol{X})]}.
$$

But, (12), (14), and Proposition 2 imply that

$$
\mathbb{E}\left[\frac{\beta_{\max}(\boldsymbol{X})}{\tilde{H}(\boldsymbol{X})} \frac{\tilde{H}(\boldsymbol{X})}{\tilde{H}}\right] \to 0.
$$

The result then follows by (12).

Lemma 7 shows that the eigenvalues of $\boldsymbol{H}^{(d)}$ satisfy the eccentricity condition on the distance matrix (8), and Proposition 1 relates the resulting $\tilde{T}^{(d)}$ (which is $\tilde{H}^{(d)}$) to $\tilde{M}^{(d)}$. The result then follows from Theorem 1.

Acknowledgements

References

[1] ATCHADÉ, Y. F., ROBERTS, G. O. AND ROSENTHAL, J. S. (2011). Towards optimal scaling of Metropolis-coupled Markov chain Monte Carlo. *Statist. Comput.* **21,** 555–568.

[2] BÉDARD, M. (2007). Weak convergence of Metropolis algorithms for non-i.i.d. target distributions. *Ann. Appl. Prob.* **17,** 1222–1244.

[3] BÉDARD, M. (2008). Optimal acceptance rates for Metropolis algorithms: moving beyond 0.234. *Stoch. Process. Appl.* **118,** 2198–2222.

[4] BESKOS, A., ROBERTS, G. AND STUART, A. (2009). Optimal scalings for local Metropolis–Hastings chains on nonproduct targets in high dimensions. *Ann. Appl. Prob.* **19,** 863–898.

[5] BREYER, L. A. AND ROBERTS, G. O. (2000). From Metropolis to diffusions: Gibbs states and optimal scaling. *Stoch. Process. Appl.* **90,** 181–206.

[6] FOLLAND, G. B. (2002). *Advanced Calculus*. Prentice-Hall, Upper Saddle River, NJ.

[7] GRIMMETT, G. R. AND STIRZAKER, D. R. (2001). *Probability and Random Processes*, 3rd edn. Oxford University Press, New York.

[8] METROPOLIS, N. *et al.* (1953). Equations of state calculations by fast computing machines. *J. Chem. Phys.* **21,** 1087–1091.

[9] NEAL, P. AND ROBERTS, G. (2008). Optimal scaling for random walk Metropolis on spherically constrained target densities. *Methodology Comput. Appl. Prob.* **10,** 277–297.

[10] ROBERTS, G. O. (1998). Optimal Metropolis algorithms for product measures on the vertices of a hypercube. *Stoch. Stoch.* **62,** 275–283.

[11] ROBERTS, G. O. AND ROSENTHAL, J. S. (2001). Optimal scaling for various Metropolis–Hastings algorithms. *Statist. Sci.* **16,** 351–367.

[12] 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.

[13] SHERLOCK, C. (2006). Methodology for inference on the Markov modulated Poisson process and theory for optimal scaling of the random walk Metropolis. Doctoral Thesis, Lancaster University. Available at http://eprints.lancs.ac.uk/850/.

[14] SHERLOCK, C. AND ELTON, D. (2012). A subclass of spherical and elliptical distributions with Gaussian-like limit properties. *J. Prob. Statist.* **2012,** 17pp.

[15] SHERLOCK, C. AND ROBERTS, G. (2009). Optimal scaling of the random walk Metropolis on elliptically symmetric unimodal targets. *Bernoulli* **15,** 774–798.