Optimal scaling of the random walk Metropolis: general criteria for the 0.234 acceptance rule

Chris Sherlock

Department of Mathematics and Statistics, Lancaster University, Lancaster, United Kingdom

Introduction

Analyses of the RWM for theoretically accessible classes of high-dimensional targets has shown that in many cases the optimal scaling is achieved when the acceptance rate is ≈ 0.234*, but that there are exceptions. We present a general set of sufficient conditions which ensure that the limiting optimal acceptance rate is* 0.234*.*

The RWM algorithm creates a Markov chain with stationary distribution $\pi(\mathbf{x})$, and hence (eventually) a dependent sample with distribution $\approx \pi(\mathbf{x})$. Given the current value $\mathbf{X} \in \mathbb{R}^d$, a new value $\mathbf{X}^* = \mathbf{X} + \mathbf{Y}$ is proposed by sampling a "jump", $\mathbf{Y}$, from from a pre-specified Lebesgue density

$$q(\mathbf{y}|\mathbf{x}) = \lambda^{-d} r(\mathbf{y}/\lambda),$$

where $r(-\mathbf{y}) = r(\mathbf{y})$; the proposal is then accepted with probability $\alpha(\mathbf{x}, \mathbf{y}) = 1 \wedge (\pi(\mathbf{x}^*)/\pi(\mathbf{x}))$. If the proposed value is accepted it becomes the next current value ($\mathbf{X}' \leftarrow \mathbf{X}^*$), otherwise the current value is left unchanged ($\mathbf{X}' \leftarrow \mathbf{X}$).

Previous theoretical results

Consider exploration of targets of the form:

$$\pi_d(\mathbf{x}) = \prod_{i=1}^{d} \beta_i f_i(\beta_i x_i),$$

using a Gaussian proposal. In Roberts and Rosenthal (2001) the β_i are taken to be random, iid, and the 0.234 acceptance rate rule is shown to hold provided $\mathbb{E}\left[\beta_i^2\right] < \infty$. In Bèdard (2007) the β_i are a fixed triangular sequence, and the 0.234 acceptance rule is shown to hold provided that

$$\frac{\beta_{max}}{\sum_{i=1}^{d} \beta_i} \to 0, \quad \text{where} \quad \beta_{max} = \max_{i=1\ldots d} \beta_i. \tag{1}$$

Sherlock and Roberts (2009) considers elliptical targets $\mathbf{X}$; i.e. of the form

$$\pi_d(\mathbf{x}) := f\left(\mathbf{x}^t \mathbf{B}_d \mathbf{x}\right)$$

for a symmetric $d \times d$ matrix $\mathbf{B}_d$ with eigenvalues $\beta_1, \ldots, \beta_d$, explored using any spherically symmetric proposal $\lambda\mathbf{U}$. The 0.234 rule is shown to hold provided that there are sequences $k_x^{(d)}$ and $k_u^{(d)}$ such that

$$||\mathbf{X}|| / k_x^{(d)} \xrightarrow{p} 1 \quad \text{and} \quad ||\mathbf{U}|| / k_u^{(d)} \xrightarrow{m.s.} 1.$$

and that (1) holds. If (1) holds and $||\mathbf{U}|| / k_u^{(d)} \xrightarrow{m.s.} 1$ but $||\mathbf{X}|| / k_x^{(d)} \xrightarrow{p} R$ for some non-degenerate random variable R then the optimal acceptance rate is strictly less than 0.234.

Set-up and notation for this article

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

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

and define the following frame invariant norms of the derivatives:

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

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

Proposals $\mathbf{Y} := \lambda\mathbf{U}$ are assumed to be spherically symmetric and to satisfy $||\mathbf{U}|| / k_u^{(d)} \xrightarrow{m.s.} 1$, for some sequence $k_u^{(d)}$.

Measure of efficiency

Our efficiency criterion is the *generalised expected squared jump distance*,

$$\mathbb{E}\left[\alpha(\mathbf{X}, \mathbf{Y})\, \mathbf{Y}^t \mathbf{T} \mathbf{Y}\right].$$

where $\mathbf{T}$ is a positive definite $d \times d$ matrix and where expectation is with respect to $\pi(\mathbf{x})$ and the proposal distribution for $\mathbf{Y}$.

In order that no one component of the process dominates any of the others in its effect on the ESJD, we require that curves of constant $\mathbf{y}^t\mathbf{T}\mathbf{y}$ are not too eccentric. Specifically let τ_i $(i = 1 \ldots d)$ be the (triangular) sequence of eigenvalues associated with the (sequence of) matrices $\mathbf{T}$, and let $\tau_{max} := \max_{i=1\ldots d} \tau_i$. We require that

$$\frac{\tau_{max}}{\tilde{T}} \to 0, \quad \text{where} \quad \tilde{T} := \sum_{i=1}^{d} \tau_i.$$

We now provide conditions such that the limiting optimal acceptance rate becomes deterministic. Intuitively, this is likely to happen if the acceptance probability itself becomes, in some sense, deterministic.

Shell conditions

From position $\mathbf{X}$, split a specific proposed jump, $\mathbf{y}$, into a component, $\mathbf{y}_1$, which is parallel to $\nabla \log \pi$ and a component, $\mathbf{y}_2$, which is perpendicular to $\nabla \log \pi$. Now

$$\log[\pi(\mathbf{X} + \mathbf{y})/\pi(\mathbf{X})] = \log[\pi(\mathbf{X} + \mathbf{y}_1)/\pi(\mathbf{X})] + \log[\pi(\mathbf{X} + \mathbf{y}_1 + \mathbf{y}_2)/\pi(\mathbf{X} + \mathbf{y}_1)].$$

To first order, the first term depends on $\tilde{M}(\mathbf{x}) = ||\nabla \log \pi||$, 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 $\tilde{H}(\mathbf{X} + \mathbf{y}_1)$) and the gradient (represented by $\tilde{M}(\mathbf{X} + \mathbf{y}_1)$). If both $\tilde{M}(\mathbf{X})$ and $\tilde{H}(\mathbf{X})$ become, in some sense, deterministic, then so might the change in $\log \pi$; these requirements are embodied in the following *shell conditions*: $\exists$ sequences $\tilde{M}$ and $\tilde{H}$ such that

$$\frac{\tilde{M}(\mathbf{X})}{\tilde{M}} \xrightarrow{p} 1 \quad \text{and} \quad \frac{\tilde{H}(\mathbf{X})}{\tilde{H}} \xrightarrow{p} 1. \tag{2}$$

Relative variability conditions

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

$$\mathbf{\Delta}(\mathbf{X}, \mathbf{U}) := \mathbf{H}(\mathbf{X} + \mathbf{U}) - \mathbf{H}(\mathbf{X}),$$

and for $\mathbf{Z} \sim N(\mathbf{0}, \mathbf{I}_d)$ which is independent of $\mathbf{X}$, and any fixed $\mu > 0$ and $\delta > 0$, require that

$$\text{either} \quad P_{\mathbf{X},\mathbf{Z}}\left(\frac{1}{\tilde{H}}\left|\mathbf{Z}^t \mathbf{\Delta}\left(\mathbf{X},\ t\mu\tilde{M}/\tilde{H}\,\mathbf{Z}\right)\mathbf{Z}\right| < \delta\ \forall\ t \in [0,1]\right) \to 1, \tag{3}$$

$$\text{or} \quad P_{\mathbf{X},\mathbf{Z}}\left(\frac{\log d}{\tilde{H}}\sum_{i=1}^{d}\sum_{j=1}^{d}\left|\Delta_{ij}\left(\mathbf{X},\ t\mu\tilde{M}/\tilde{H}\,\mathbf{Z}\right)\right| < \delta\ \forall\ t \in [0,1]\right) \to 1. \tag{4}$$

Eccentricity Condition

$\tilde{H}(\mathbf{X})$ represents an "average" curvature which, intuitively, 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; in other words the scales of variability of π along each component of $\mathbf{X}$ should not be too dissimilar. The *eccentricity condition* on the target ensures that the chance of such extreme behaviour diminishes to zero.

$$\frac{\beta_{max}(\mathbf{X})}{\sum_{i=1}^{d} \beta_i(\mathbf{X})} \xrightarrow{p} 0. \tag{5}$$

Note that (5) is a generalisation of (1).

Main result

Theorem Subject to the shell conditions (2), either of the relative variability conditions (3) or (4), and the eccentricity condition (5), for fixed $\mu > 0$ set the scaling as

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

The expected acceptance rate and generalised ESJD now satisfy

$$\lim_{d\to\infty} \mathbb{E}\left[\alpha(\mathbf{X}, \mathbf{Y})\right] = 2\Phi\left(-\frac{1}{2}\mu\right), \tag{6}$$

$$\lim_{d\to\infty} \frac{\tilde{H}^2}{\tilde{M}^2\tilde{T}} \mathbb{E}\left[\mathbf{Y}^t\mathbf{T}\mathbf{Y}\ \alpha(\mathbf{X}, \mathbf{Y})\right] = 2\mu^2\Phi\left(-\frac{1}{2}\mu\right). \tag{7}$$

NB strengthening (2) and adding a regularity condition gives $\tilde{H} \sim \tilde{M}^2$.

Corollary Equation (7) is maximised at $\mu \approx 2.38$; substitution into (6) provides the limiting optimal acceptance rate of ≈ 0.234.

Example

For fixed $p > 0$, the stationary p^{th} order Markov chain

$$\pi(\mathbf{x}) := f^*(x_1, \ldots, x_p) f(x_{p+1}|x_1, \ldots, x_p) \ldots f(x_d|x_{d-p}, \ldots, x_{d-1}),$$

(with stationary distribution f^*) satisfies all of the requirements subject to certain moment conditions.

Bibliography

BÈDARD M., (2007). Weak convergence of Metropolis algorithms for non-i.i.d. target distributions. *Ann. Appl. Probab.* **17**, 1222-1244.

ROBERTS, G.O. and ROSENTHAL, J.S., (2001). Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science* **16**, 351–367.

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