EFFICIENT METROPOLIS JUMPING RULES

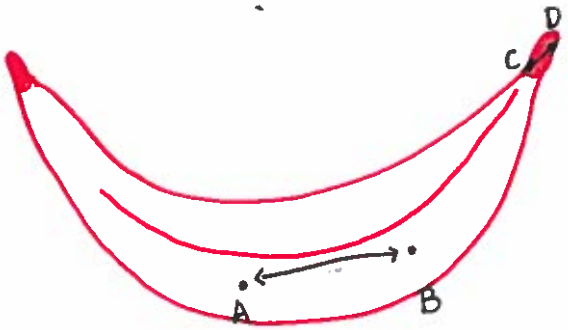
María Lomelí
Tea Talk

(based on paper:

Efficient Metropolis Jumping Rules

A. GELMAN*, G. O. ROBERTS** and W. R. GILKS***
*University of California, USA, **University of Cambridge, UK and
***Medical Research Council, UK

1 Motivation.



- "Bold" or large jump from A to B
- Small jump from C to D

`Banana shaped' density

Goal of MCMC: estimate a (typically multivariate) target distribution $\pi(\theta)$ by generating a Markov chain $\theta^{(1)}, \theta^{(2)}, \ldots$ whose stationary distribution is π.

Metropolis Algorithm Assumes a symmetric jumping density

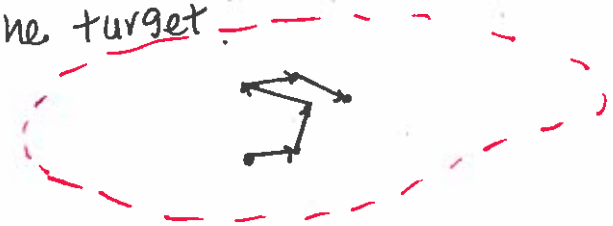
$$J(\theta, \theta') = J(\theta', \theta)$$

Accept/Reject candidate point according to:

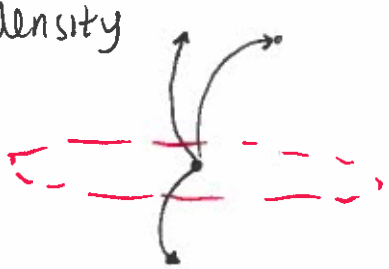
$$\alpha(\theta, \theta') = \min\left\{ \frac{\pi(\theta')}{\pi(\theta)}, 1 \right\}$$

Slow Mixing due to:

1) Jumps are so short that the simulation moves slowly through the target.



2) The jumps are nearly all into low-probability areas of the target density



It is often possible to improve the mixing by adjusting the jump distribution.

Heuristic rules such as : monitoring the distance of each jump or frequency of acceptance.

Case $J(\theta', \theta) = J(|\theta' - \theta|)$

Famous heuristic strategy :

Choose the scaling of $J(\cdot,\cdot)$ so that the average acceptance rate of the algorithm is roughly 1/4.

2. Univariate Examples.

$\theta \in \mathbb{R}$.

2 measures of efficiency :

a) $\text{Eff}_\theta = \dfrac{\chi^2}{V_{\bar\theta}}$ where χ^2 corresponds to the empirical variance of an iid sample from θ. and

$$V_{\bar\theta} = \lim_{N\to\infty} N \,\text{Var}\left(\frac{1}{N}\sum_{t=1}^{N} \theta^{(t)}\right)$$

limiting scaled variance from the Markov chain output.

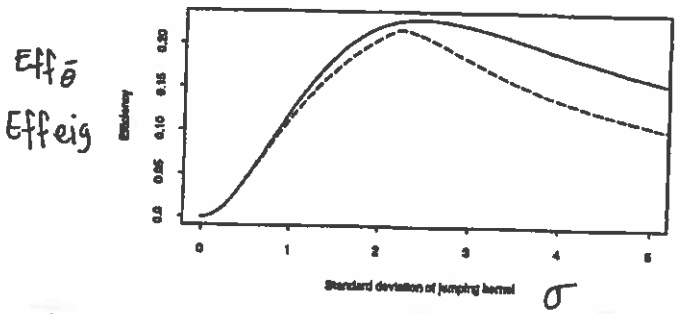
b) $\text{Eff}_{\bar\theta} = \left\{ \sum a(i) \dfrac{1+\lambda_i}{1-\lambda_i} \right\}^{-1} \geq \dfrac{1-\lambda_2}{1+\lambda_2} = \text{Eff eig}$

Efficiency based on the 2nd eigenvalue.

where $\lambda_1, \lambda_2, \ldots$ are the eigenvalues of the transition kernel

$$J(\theta, \theta')$$

One of the easiest characteristics of a Metropolis algorithm to monitor is the frequency of "acceptance" in the Metropolis step—which we label p_{jump}. It has been claimed that, for a wide variety of problems, optimal rules have acceptance probabilities near 0.5 (see, for example, Muller, 1993).



The optimal efficiency, using either measure, is just below 0.25. (The "corner" at the maximum of the eff_{eig} line occurs when the second and third largest eigenvalues are equal.)

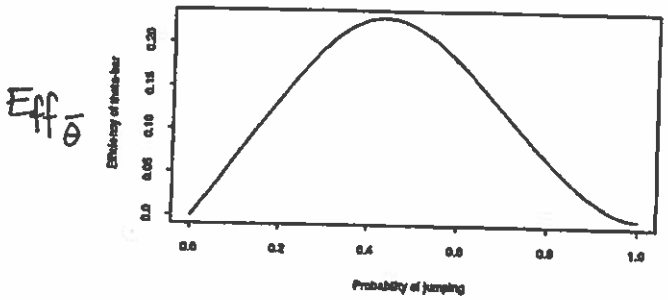


Figure 1b plots the efficiency measure $\text{eff}_{\bar{\theta}}$ as a function of acceptance rate; the leftmost point on the curve corresponds to $\sigma \to \infty$, and the rightmost point to $\sigma = 0$. At least for this example, the folklore seems correct; an acceptance rate near (but slightly below) 0.5 is optimal.

where P_{jump} : frequency of acceptance in the Metropolis step.

3. Numerical illustrations.

i) Discrete approximation of Π

Let = linspace $(-6, 6, 100)$;

$p(\cdot) = \text{normpdf}(\cdot)$;

$p(\cdot) = \dfrac{p(\cdot)}{\text{sum}(p(\cdot))}$;

ii) Transition density : $J(\cdot,\cdot) \propto e^{-\frac{1}{2}\left(\frac{\cdot - \cdot}{\sigma}\right)^2}$

$$P = \begin{matrix} -6 \\ -5 \\ \vdots \\ 0 \\ 1 \\ \vdots \\ 6 \end{matrix} \begin{bmatrix} -6 & -5 & \cdots & & 6 \\ & & & & \\ & & J(i,j)_{ij} & & \end{bmatrix}$$

eig (P); to compute Eff eig & Eff$_{\bar{\theta}}$ using an asymptotic formula

If σ is too low, the Metropolis steps are too short and move too slowly through the target distribution; if σ is too high, the algorithm almost always rejects and stays in the same place. The optimal σ is somewhere in between.

Interestingly, if one cannot be optimal, it seems better to use too high a value of σ than too low; $\sigma = 5$ is better than $\sigma = 1$.

3. Multivariate Target Distributions

π is d-dimensional (not neccesarily Normal) but factorizes:

$$\pi(\theta) = \prod_{i=1}^{d} f(\theta_i)$$

Proposal distribution $N_d\left(\theta' \mid \theta, \left(\frac{\phi^2}{d}\right) I_d\right)$

θ : current point

As $d \to \infty$, assuming that $\theta, \theta_1, \theta_2, \ldots$ are all iid $\sim f$ then Y^d

Theo. $\xrightarrow{w}$ a limiting Langevin diffusion which satisfies:

$$dY_t = \frac{f'(Y_t)\, h(\phi)}{2 f(Y_t)}\, dt + h(\phi)^{1/2}\, dB_t$$

speed of diffusion $\quad h(\phi) = 2\phi^2 \Phi\left(\frac{-\phi F^{1/2}}{2}\right)$

$$h'(\phi) = 4\phi \Phi\left(\frac{-\phi F^{1/2}}{2}\right) + 2\phi^2 \Phi'\left(\frac{-\phi F^{1/2}}{2}\right) \times \left(\frac{-F^{1/2}}{2}\right)$$

$$\Phi\left(\frac{-\phi}{2}\right) = \frac{1}{4\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{\phi}{2}\right)^2}$$

$$F = \int_{-\infty}^{\infty} \frac{(f'(x))^2}{f(x)}\, dx$$ Fisher info measure for f

($F = 1$ if f is standard Normal)

$Y_d := \theta_1^{[td]}$ a speed up continuous time version of the d-dimensional Metropolis algorithm.

i.e. continuous time process which remains constant for a time interval $1/d$ and then jumps according to the Metropolis algorithm

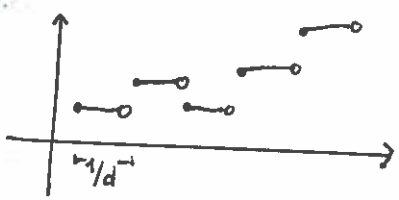


Figure 2 shows how h varies with the jumping kernel scale factor $\phi = \sigma_d\sqrt{d}$, and with the acceptance rate p_{jump}, assuming $F = 1$. Here we see clearly that efficiency is maximised by setting $\phi = 2.38$ (Figure 2a) or by setting $p_{jump} = 0.234$ (Figure 2b).

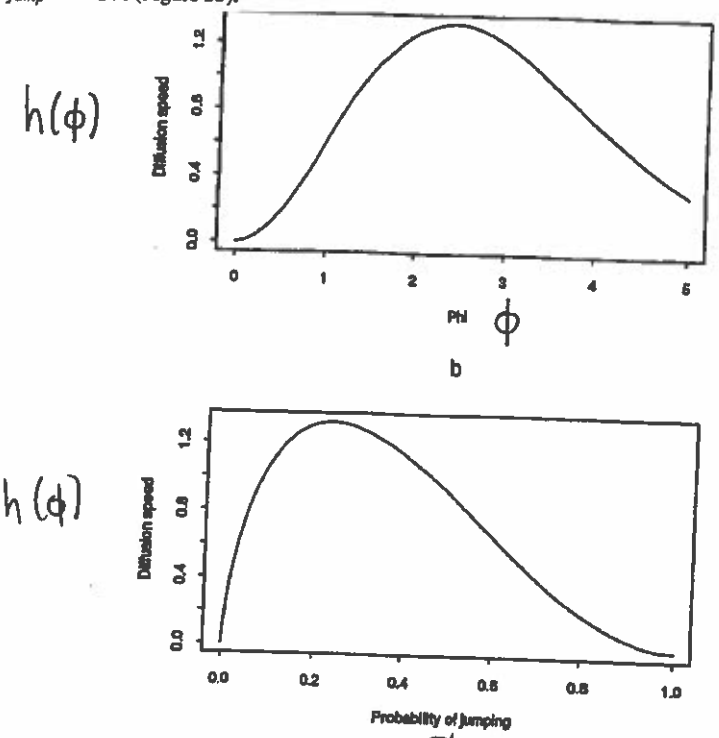


Figure 2. *Diffusion speed $h(\phi)$ related to jumping kernel scale factor ($\sigma = \phi/\sqrt{d}$) and acceptance rate p_{jump}.*

The limiting value of p_{jump} is $\dfrac{h(\phi)}{\phi^2}$

$h(\phi)$ is maximized at $\tilde{\phi} = \dfrac{2.38}{F^{1/2}}$

$$\alpha = \frac{h(\tilde{\phi})}{\tilde{\phi}^2} = 0.234$$

The optimal jumping kernel has variance-covariance matrix $\left(\dfrac{\tilde{\phi}^2}{d}\right) * Id.$

What is the relevance for finite dimensional problems?

The simulation study below demonstrates that the asymptotic optimality of accepting approximately 1/4 of proposed moves is approximately true for dimension as low as 6.

Table 1. *Optimal scale factor σ_d and optimal efficiency for normal jumping kernel and standard normal target distribution in low dimensions, compared to theoretical values based on Theorem 3.1.*

Dimension, d	Optimal σ_d	eff$_1$	p_{jump}	$2.38/\sqrt{d}$	$0.331/d$
1	2.40	0.233	0.441	2.38	0.331
2	1.70	0.136	0.352	1.68	0.166
3	1.39	0.098	0.316	1.37	0.110
4	1.25	0.076	0.279	1.19	0.083
5	1.10	0.062	0.275	1.06	0.066
6	1.00	0.053	0.266	0.97	0.055
7	0.93	0.047	0.261	0.90	0.047
8	0.87	0.041	0.255	0.84	0.041
9	0.80	0.037	0.261	0.79	0.037
10	0.74	0.034	0.267	0.75	0.033

The results show that the asymptotically optimal $\sigma_d = 2.38/\sqrt{d}$ (from Section 3.1) applies for d as low as 1, and the asymptotic acceptance rate of 0.234 and efficiency of $0.331/\sqrt{d}$ are attained approximately by $d = 6$. Thus Theorem 3.1 accurately predicts the behavior of the optimal spherically symmetric multivariate normal jumping kernel in low dimensions.

The theory and the simulation study both support the use of an over-dispersed proposal distribution, as recommended by Besag and Green (1993) for one-dimensional sampling in multivariate problems. However for higher dimensional problems, it is advisable to have proposals with smaller variances in relation to those of the target density.

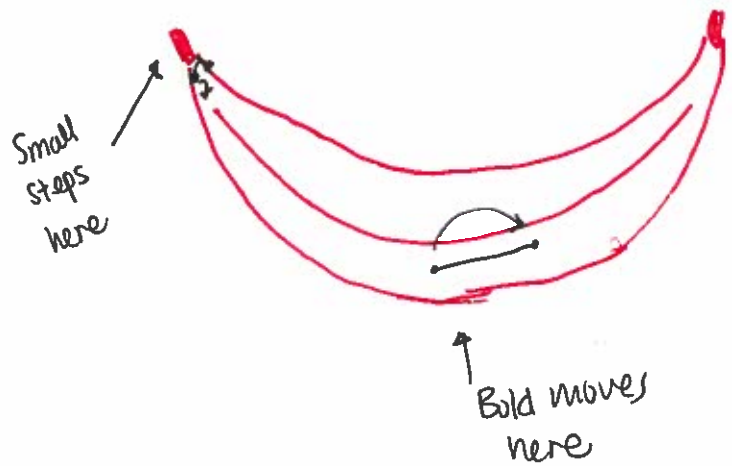
4 Practical Implications.

- Heuristics for adaptive Metropolis scheme.
- After a few iterations we could monitor the convergence of the chain and maybe try to improve efficiency using whatever info is available from the simulations been produced so far. :)

→ Adaptively altering a metropolis rule

Incidentally, the simulations produced by an adaptive "Markov chain" simulation are not, in general, themselves a Markov chain, because the transition probabilities can depend on the results of earlier iterations (see, for example, Gelfand and Sahu, 1993).

So we could use this for the Banana distribution



Family of proposals parameterized by scaling factor

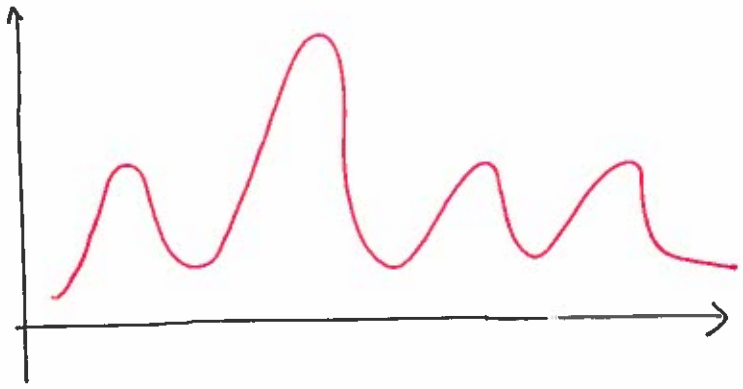
$\{\gamma : p_\gamma(x^*|y)\}$

Fine tune the Metropolis algorithm while it's running by increasing or decreasing variance

Care has to be taken when adopting this approach, since adaptation to information from previous iterations can compromise the stationarity of the target density. However, such an approach is acceptable as part of a pilot sample analysis, where adaptation stops after a fixed number of exploratory iterations.

Our computations provide some justification for such an adaptive approach. For higher dimensional jumping rules, however, a lower acceptance rate near 0.25 is preferable. Moreover, Theorem 3.1 implies that an average acceptance rate of between 0.15 and 0.4 yields at least 80% of the maximum efficiency obtainable (see Figure 2). In practice therefore, adaptation cannot be recommended when acceptance rates are within this range. Even the folklore figure of 0.5 produces reasonable results (approximately 75% of maximum possible efficiency)

Still valid in a Metropolis within Gibbs scheme.
How about a multimodal target?



Finally, we emphasize that an acceptance rate of around 0.25 does not guarantee efficiency of the algorithm. In particular, a different approach may be required to sample efficiently from highly multimodal distributions. However, when an efficient scaling does exist, it is often sufficient to only loosely tune the proposal distribution in order to obtain satisfactory results.