Technical Report No. 0411, Department of Statistics, University of Toronto

Taking Bigger Metropolis Steps by Dragging Fast Variables

Radford M. Neal
Department of Statistics and Department of Computer Science
University of Toronto, Toronto, Ontario, Canada
http://www.cs.utoronto.ca/~radford/
radford@stat.utoronto.ca

26 October 2004

Abstract. I show how Markov chain sampling with the Metropolis-Hasting algorithm can be modified so as to take bigger steps when the distribution being sampled from has the characteristic that its density can be quickly recomputed for a new point if this point differs from a previous point only with respect to a subset of "fast" variables. I show empirically that when using this method, the efficiency of sampling for the remaining "slow" variables can approach what would be possible using Metropolis updates based on the marginal distribution for the slow variables.

1 Introduction

Suppose we wish to sample from a distribution $\pi(x, y) \propto \exp(-E(x, y))$, where E is a given "energy" function, by simulating a Markov chain with π as its equilibrium distribution. Let's suppose that x is a "slow" variable and y is a "fast" variable, so that once $E(x, y)$ has been computed (and intermediate quantities cached), we can compute $E(x, y')$ much faster than we can compute $E(x', y')$ for some x' for which we haven't previously calculated E.

I was led to consider this problem because it arises with Bayesian models that attempt to infer cosmological parameters from data on the cosmic microwave background radiation (Lewis and Bridle 2002), for which recomputation after changing only fast variables can be around a thousand times faster than recomputation after changing a slow variable. Similarly large differences between fast and slow variables can arise with Gaussian process classification models (Neal 1999), in which updating the latent variables is fast, while updating the parameters of the covariance matrix is slow, since the new covariance matrix must then be inverted. Computationally equivalent problems also arise in geostatistics (Diggle, Tawn, and Moyeed 1998), and for what are called "generalized linear mixed effects models". Many other statistical problems also have some variables that are faster than others, though not always by such a large factor.

Ideally, we would like to do Metropolis-Hastings updates for x (Metropolis, *et al* 1953; Hastings 1970), using some proposal distribution, $S(x^*|x)$, and accepting or rejecting x^* based on its marginal distribution, $\pi(x)$. The acceptance probability for such a proposal would be

$$a(x, x^*) \;\; = \;\; \min\left[1,\; \frac{S(x|x^*)\,\pi(x^*)}{S(x^*|x)\,\pi(x)}\right] \tag{1}$$

Suppose, however, that we can't feasibly compute the marginal distribution, $\pi(x)$, so that this approach is not possible. Instead we will have to use a Metropolis-Hastings algorithm that operates on the joint distribution for x and y. If we could sample directly from the conditional distribution for y, $\pi(y|x)$, we could generate x^* from $S(x^*|x)$ and then y^* from $\pi(y^*|x^*)$, and the resulting acceptance probability for (x^*, y^*) would be the same (due to cancellation) as that above using the marginal distribution for x. However, let's assume that sampling from $\pi(y|x)$ is also infeasible. We might hope to approximate $\pi(y|x)$ by some transition distribution $T(y^*|y;x)$ that we can sample from. To use this approximation in a Metropolis-Hasting proposal, however, we would need to be able to compute the probability of proposing y^*, which will likely be impossible if we have to resort to iterative methods (eg, Markov chain simulation) in order to obtain a good approximation.

This paper describes a way in which these problems can be bypassed when recomputing $E(x, y)$ after changing only the "fast' variable y is much quicker than recomputing $E(x, y)$ after changing x. In this method, changes to x are made in conjunction with changes to y that are found by "dragging" y with the help of intermediate transitions that involve only fast re-computations of E. In the limit as the number of such intermediate transitions increases, I show empirically (but haven't proved) that the method is equivalent to using the marginal distribution of x. Since the intermediate transitions involve only fast computations, we hope to be able to do quite a few intermediate transitions, and get close to the effect of using the marginal probabilities for x.

The method can be seen as a generalization of "tempered transitions" (Neal 1996), and could be expressed in greater generality than I have done here, where I concentrate on the context with fast and slow variables. To begin, I'll describe the method when there is only one intermediate transition, since this is easier to work with, but I expect that one would use many intermediate transitions in practice, as described later.

2 The method with one intermediate transition

If the current state is (x, y), we start by generating a proposed new value x^* according to the probabilities $S(x^*|x)$. We then define a distribution, ρ, over values for y that is intermediate between $\pi(y|x)$ and $\pi(y|x^*)$, as follows:

$$\rho(y; x, x^*) \;\; \propto \;\; \exp(-(E(x, y) + E(x^*, y))/2) \tag{2}$$

Here, the dependence of ρ on x and x^* has been made explicit, but note that this is a distribution over y only, not x and y jointly. We chose some transition probabilities, T, for updating y so as to leave ρ invariant. These probabilities must of course depend on x and x^*. We write them as $T(y'|y; x, x^*)$. We require that they satisfy detailed balance, so that for all x, x^*, y, and y',

$$\rho(y; x, x^*)\, T(y'|y; x, x^*) \;\; = \;\; \rho(y'; x, x^*)\, T(y|y'; x, x^*) \tag{3}$$

We also require that T depend symmetrically on the two x values — for all x, x^*, y and y':

$$T(y'|y; x, x^*) \;\; = \;\; T(y'|y; x^*, x) \tag{4}$$

T might, for example, be a Metropolis-Hastings update, or a series of such updates. We apply this transition once, to sample a value y^* from $T(y^*|y; x, x^*)$. We then accept (x^*, y^*) as the next state with probability $a(x, y, x^*, y^*)$, defined as follows:

$$a(x,y,x^*,y^*) \;=\; \min\left[1,\ \frac{S(x|x^*)\,\pi(x^*,y^*)\,\rho(y;x,x^*)}{S(x^*|x)\,\pi(x,y)\,\rho(y^*;x,x^*)}\right] \tag{5}$$

$$= \min\left[1,\ \frac{S(x|x^*)\ \exp(-E(x^*,y^*))\ \exp(-(E(x,y)+E(x^*,y))/2)}{S(x^*|x)\ \exp(-E(x,y))\ \exp(-(E(x,y^*)+E(x^*,y^*))/2)}\right] \tag{6}$$

$$= \min\left[1,\ \frac{S(x|x^*)}{S(x^*|x)}\ \exp\left(\frac{E(x,y)+E(x,y^*)}{2} - \frac{E(x^*,y)+E(x^*,y^*)}{2}\right)\right] \tag{7}$$

If we don't accept, the next state is the current state, (x, y).

Although this expression for $a(x, y, x^*, y^*)$ has four occurrences of $E(\cdot, \cdot)$, only two slow evaluations are needed. In fact, only one slow evaluation is needed if we assume that an evaluation was done previously for the current state, when it was proposed. Note also that we would often choose a symmetric proposal distribution for x, so that $S(x^*|x)/S(x|x^*) = 1$.

To show that this is a valid update, I will prove that it satisfies detailed balance. The probability of moving from (x, y) to a different state (x^*, y^*) when in equilibrium is

$$\pi(x,y)\,S(x^*|x)\,T(y^*|y;x,x^*)\ \min\left[1,\ \frac{S(x|x^*)\,\pi(x^*,y^*)\,\rho(y;x,x^*)}{S(x^*|x)\,\pi(x,y)\,\rho(y^*;x,x^*)}\right]$$

$$= \min\left[S(x^*|x)\,\pi(x,y)\,T(y^*|y;x,x^*),\ \frac{S(x|x^*)\,\pi(x^*,y^*)\,\rho(y;x,x^*)\,T(y^*|y;x,x^*)}{\rho(y^*;x,x^*)}\right] \tag{8}$$

$$= \min\left[S(x^*|x)\,\pi(x,y)\,T(y^*|y;x,x^*),\ S(x|x^*)\,\pi(x^*,y^*)\,T(y|y^*;x,x^*)\right] \tag{9}$$

$$= \min\left[S(x^*|x)\,\pi(x,y)\,T(y^*|y;x,x^*),\ S(x|x^*)\,\pi(x^*,y^*)\,T(y|y^*;x^*,x)\right] \tag{10}$$

Here, the detailed balance condition (3) and symmetry condition (4) have been used. Examination of the above shows that swapping (x, y) and (x^*, y^*) leaves it unchanged, showing the detailed balance holds.

I would expect this method to work better than the simple method of just proposing to change from x to x^* while keeping y unchanged. The latter method will work well only if the old y is often suitable for the new x^* — ie, if the old y is typical of $\pi(y|x^*)$. This will often be true only if the change from x to x^* is small. The new method changes y to a y^* that is drawn approximately (if T works well) from a distribution that is halfway between $\pi(y|x)$ and $\pi(y|x^*)$. Such a y^* should have a better chance of being suitable for x^*, allowing the change from x to x^* to be greater while still maintaining a good acceptance probability. If we propose an x^* that is a really big change from x, however, even a y^* that comes from a distribution halfway to $\pi(y|x^*)$ may not be good enough.

3 The method with many intermediate transitions

We can try to take bigger steps in x by "dragging" y through a series of intermediate distributions interpolating between $\pi(y|x)$ and $\pi(y|x^*)$. Given some integer $n > 1$, we define the following distributions, for $i = 0, \ldots, n$:

$$\rho_i(y; x, x^*) \quad \propto \quad \exp(-((1-i/n)E(x,y) \,+\, (i/n)E(x^*,y))) \tag{11}$$

Notice that $\rho_0(y; x, x^*) = \pi(y|x)$ and $\rho_n(y; x, x^*) = \pi(y|x^*)$. When $n = 2$, ρ_1 is the same as the ρ defined above in (2). Finally, note that $\rho_i(y; x, x^*) = \rho_{n-i}(y; x^*, x)$.

For each ρ_i, we need to choose transition probabilities, T_i, which may depend on x and x^*. We require that they satisfy detailed balance, so that for all x, x^*, y, and y',

$$\rho_i(y; x, x^*)\, T_i(y'|y; x, x^*) \quad = \quad \rho_i(y'; x, x^*)\, T_i(y|y'; x, x^*) \tag{12}$$

We also require of each opposite pair of transitions, T_i and T_{n-i}, that for all x, x^*, y and y',

$$T_i(y'|y; x, x^*) \quad = \quad T_{n-i}(y'|y; x^*, x) \tag{13}$$

These conditions will be satisfied if the T_i are standard Metropolis updates with respect to the ρ_i, with T_i and T_{n-i} using the same proposal distribution.

The update procedure using $n-1$ intermediate distributions is as follows. If the current state is x, we first propose a new x^* according to the probabilities $S(x^*|x)$. We then generate a series of values $y_1, \ldots, y_{n-1}$, with y_i being drawn according to the probabilities $T_i(y_i|y_{i-1}; x, x^*)$. Let $y^* = y_{n-1}$, and define $y_0 = y$. We accept (x^*, y^*) as the new state of the Markov chain with the following probability:

$$\begin{aligned} a(x, y, x^*, y^*, y_1, \ldots, y_{n-2}) \quad &= \quad \min\left[1,\ \frac{S(x|x^*)\,\pi(x^*, y^*)}{S(x^*|x)\,\pi(x, y)} \prod_{i=1}^{n-1} \frac{\rho_i(y_{i-1}; x, x^*)}{\rho_i(y_i; x, x^*)}\right] &(14)\\ &= \quad \min\left[1,\ \frac{S(x|x^*)}{S(x^*|x)}\, \exp\left(\frac{1}{n}\sum_{i=0}^{n-1} E(x, y_i) \,-\, \frac{1}{n}\sum_{i=0}^{n-1} E(x^*, y_i)\right)\right] &(15)\end{aligned}$$

To show that this is a valid update, I will show that the probability in equilibrium of the chain moving from (x, y) to a different state (x^*, y^*) while generating intermediate states $y_1, \ldots, y_{n-2}$ is equal to the probability of the chain moving from (x^*, y^*) to (x, y) while generating intermediate states $y_{n-2}, \ldots, y_1$. Detailed balance then follows by summing over possible sequences of intermediate states. The probability of moving from (x, y) to (x^*, y^*) via $y_1, \ldots, y_{n-2}$ can be written as

$$\begin{aligned} &\pi(x, y)\, S(x^*|x) \left[\prod_{i=1}^{n-1} T_i(y_i|y_{i-1}; x, x^*)\right] a(x, y, x^*, y^*, y_1, \ldots, y_{n-2}) \\ &\quad = \quad \min\Bigg[S(x^*|x)\,\pi(x, y) \prod_{i=1}^{n-1} T_i(y_i|y_{i-1}; x, x^*), \\ &\qquad\qquad S(x|x^*)\,\pi(x^*, y^*) \prod_{i=1}^{n-1} \frac{\rho_i(y_{i-1}; x, x^*)\, T_i(y_i|y_{i-1}; x, x^*)}{\rho_i(y_i; x, x^*)} \Bigg] &(16)\\ &\quad = \quad \min\left[S(x^*|x)\,\pi(x, y) \prod_{i=1}^{n-1} T_i(y_i|y_{i-1}; x, x^*),\ \ S(x|x^*)\,\pi(x^*, y^*) \prod_{i=1}^{n-1} T_i(y_{i-1}|y_i; x, x^*)\right] &(17)\\ &\quad = \quad \min\left[S(x^*|x)\,\pi(x, y) \prod_{i=1}^{n-1} T_i(y_i|y_{i-1}; x, x^*),\ \ S(x|x^*)\,\pi(x^*, y^*) \prod_{i=1}^{n-1} T_{n-i}(y_{i-1}|y_i; x^*, x)\right] &(18)\end{aligned}$$

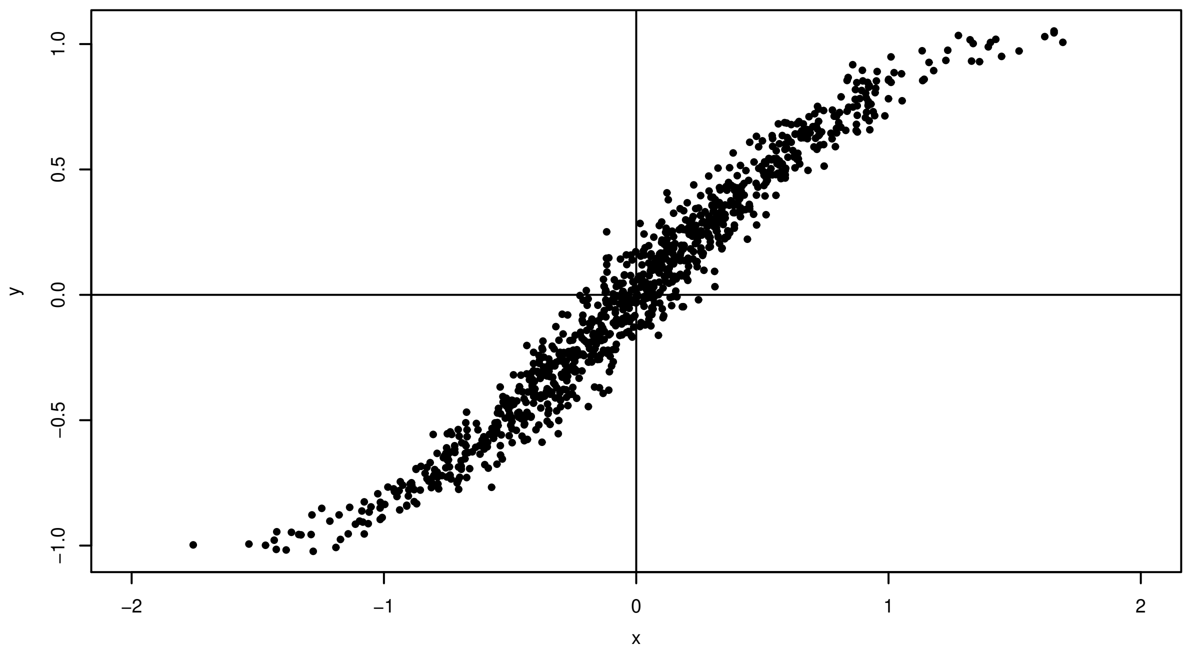

Figure 1: A sample of 1000 points from the first test distribution.

If we swap x and x^*, y and y^*, and y_i and y_{n-i-1}, reverse the order of the two products, and swap the arguments of min, we see that this expression is unchanged, showing that the reverse transition from (x^*, y^*) to (x, y) via $y_{n-2}, \ldots, y_1$ is equally likely.

4 Tests on simple distributions

I first tested the dragging method on a simple distribution in which x and y are both one-dimensional, with $\pi(x, y)$ defined by the following energy function:

$$E(x, y) \;=\; x^2 \;+\; 50\,(1 + x^2)^2\,(y - \sin(x))^2 \tag{19}$$

Examination of this shows that the conditional distribution for y given x is Gaussian with mean $\sin(x)$ and standard deviation $1/(1+x^2)$. From this, one can deduce that the marginal distribution for x can be obtained with an energy function of $x^2 + \log(1+x^2)$. For this test problem, we can therefore compare performance using dragging transitions to the "ideal" performance when doing Metropolis updates based on this marginal distribution. Figure 1 shows a sample of points obtained in this way, with y values filled in randomly from their conditional distribution given x.

For purposes of this test, we can pretend that computing $\sin(x)$ is much slower than any of the other computations involved in evaluating $E(x, y)$, or in the mechanics of performing Markov chain updates. This will make x a "slow" variable, whereas y will be a "fast" variable. We also pretend that we don't know that x and y are positively correlated. This mimics situations in which we are first exploring the distribution, or in which the relationship between x and y is non-monotonic, so that no linear transformation is helpful.

Figure 2 shows the efficiency of six sampling methods applied to this distribution, as measured by the autocorrelations for x at lags up to 30. All the methods are based on the Metropolis algorithm with Gaussian proposals centred on the current state. In all cases, the standard deviation of the Gaussian proposals was adjusted to be approximately optimal. All the methods require only one slow computation of $\sin(x)$ for each iteration (for the Marginal Metropolis method, this would be needed only when filling in y values to go with the x values).

In the Joint Metropolis method, the proposals change x and y simultaneously and independently, with the standard deviations for each being 0.5. The rejection rate for these proposals was 87%. In the Single-variable Metropolis method, two Metropolis updates are done each iteration, one for x only, the other for y only. The standard deviations for these proposals were both 0.25. The rejection rates were 59% for x and 64% for y. For the Marginal Metropolis method, where the state consists of x alone, the proposals had standard deviation of 1.0, and the rejection rate was 47%. Clearly, the Marginal Metropolis method performs much better than the other two, though in real problems it would typically be infeasible.

The remaining plots show the autocorrelations when sampling using updates that drag y while changing x, with 20, 100, and 500 intermediate distributions. For all three plots, the proposal distribution for x had standard deviation 1.0, while the proposal distributions for y during the intermediate transitions had standard deviation 0.2. The rejection rate for the "inner" updates of y was around 60% for all three runs. The rejection rates for the "outer" updates of x were 76%, 63%, and 52% for 20, 100, and 500 intermediate distributions. Both the rejection rate and the autocorrelations seem to be approaching the "ideal" values seen with the Marginal Metropolis method. Provided that recomputing $E(x, y)$ after changing y is around a thousand times faster than recomputing it after changing x, updates for x using dragging transitions will be almost as good as updates based on the marginal distribution of x.

To see how sensitive these results are to the dimensionality of the fast parameter, I did a second test introducing another fast parameter, z. The energy function used was

$$E(x, y) \;=\; x^2 \;+\; 50\,(1+x^2)^2\,(y-\sin(x))^2 \;+\; 12.5\,(z-y)^2 \tag{20}$$

This produces marginal distributions for (x, y) and for x that are the same as for the first test.

Figure 3 shows the efficiency of the six sampling methods applied to this distribution. The same proposal standard deviations were used as in the first test, except that for the Joint Metropolis updates, the standard deviations were 0.3, producing a rejection rate of 85%. The dragging transitions were done using joint Metropolis updates for y and z as the inner transitions, with proposal standard deviations of 0.2.

As can be seen, all methods perform less well with the extra variable, except for the Marginal Metropolis method, which is the same as in the first test. The dragging transitions are less affected, however. The autocorrelation time (one plus twice the sum of the autocorrelations at all lags) when using 500 intermediate distributions increased from approximately 7.4 to approximately 9.3 with the addition of z. In contrast, the autocorrelation time for the joint Metropolis updates increased from approximately 75 to approximately 205, and that for the single-variable Metropolis updates went from approximately 230 to approximately 365.

The programs (written in R) used for these tests are available from my web page.

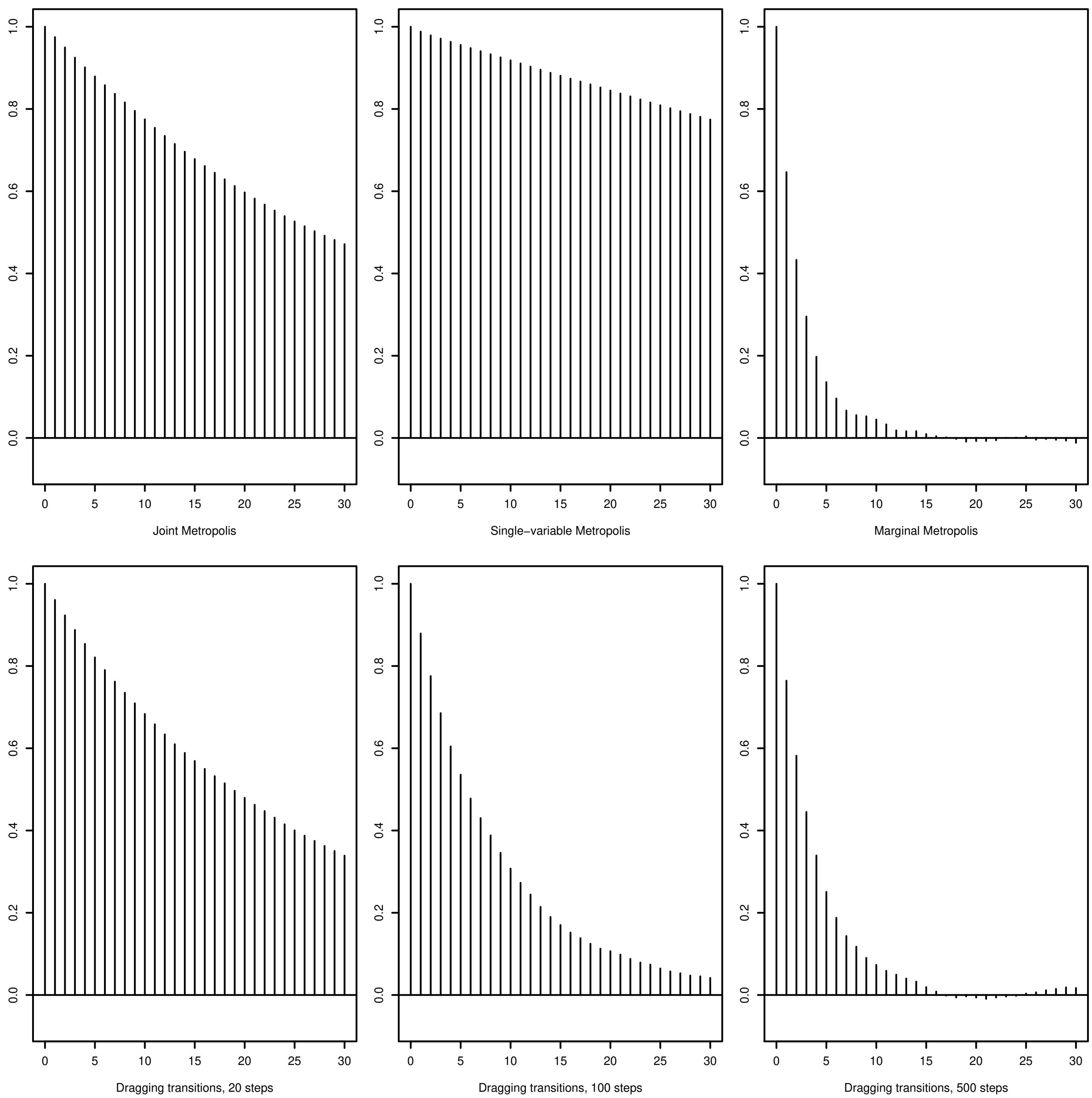

Figure 2: Estimated autocorrelations for x at lags up to 30 when sampling for the first test problem using six methods.

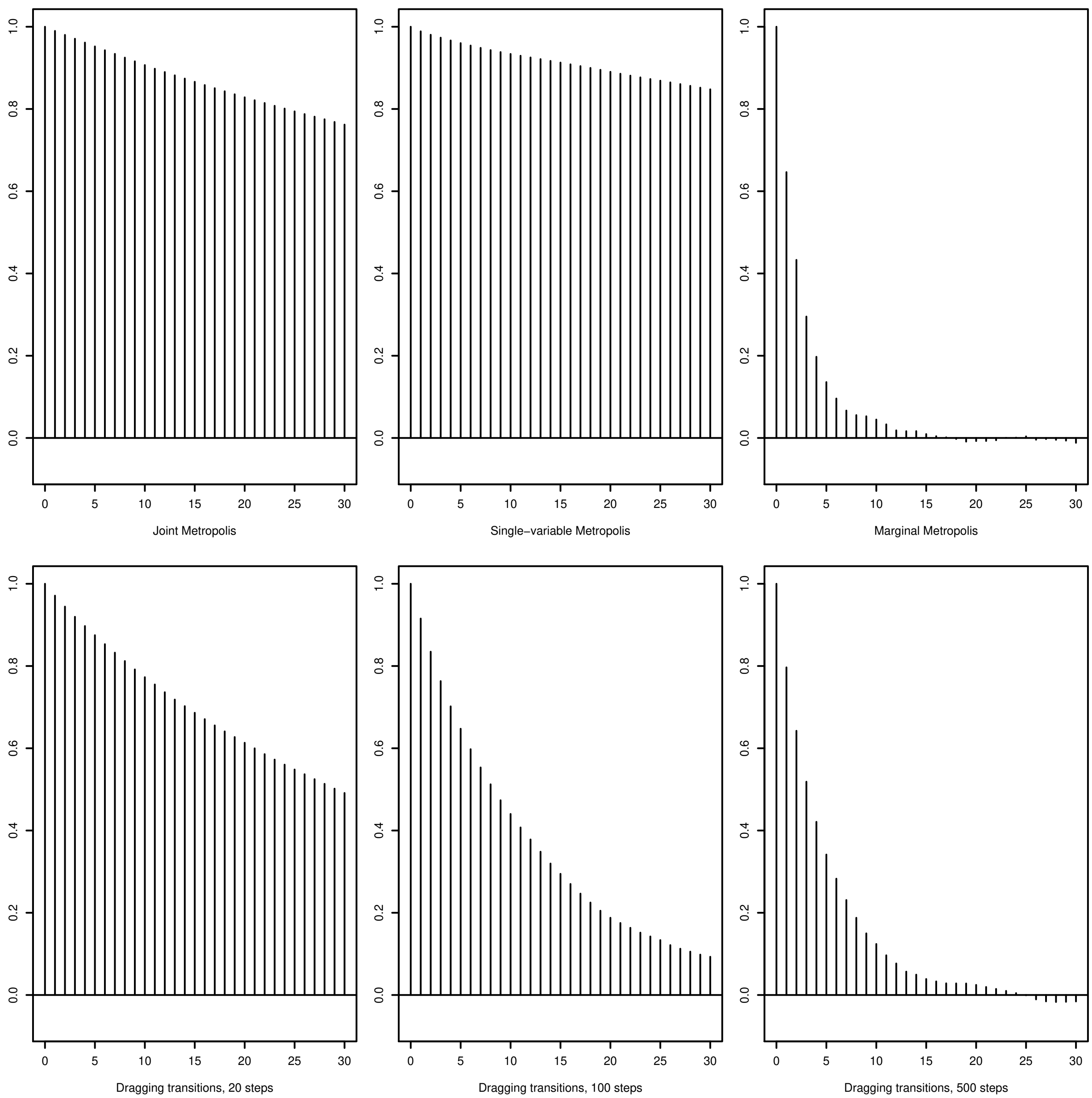

Figure 3: Estimated autocorrelations for x at lags up to 30 when sampling for the second test problem using six methods.

Acknowledgements

I thank Antony Lewis and Sarah Bridle for introducing me to the CMB application, and for comments on a draft of this paper. This research was supported by the Natural Sciences and Engineering Research Council on Canada. I hold a Canada Research Chair in Statistics and Machine Learning.

References

Diggle, P. J., Tawn, J. A., and Moyeed, R. A. (1998) "Model-based geostatistics", *Applied Statistics*, vol. 47, pp. 299-350.

Hastings, W. K. (1970) "Monte Carlo sampling methods using Markov chains and their applications", *Biometrika*, vol. 57, pp. 97-109.

Lewis, A. and Bridle, S. (2002) "Cosmological parameters from CMB and other data: a Monte-Carlo approach", `http://arxiv.org/abs/astro-ph/0205436`.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953) "Equation of state calculations by fast computing machines", *Journal of Chemical Physics*, vol. 21, pp. 1087-1092.

Neal, R. M. (1996) "Sampling from multimodal distributions using tempered transitions", *Statistics and Computing*, vol. 6, pp. 353-366.

Neal, R. M. (1999) "Regression and classification using Gaussian process priors" (with discussion), in J. M. Bernardo, *et al* (editors) *Bayesian Statistics 6*, Oxford University Press, pp. 475-501.