Statistical Science
2001, Vol. 16, No. 4, 335–339

A Geometric Interpretation of the Metropolis–Hastings Algorithm

Louis J. Billera and Persi Diaconis

Abstract. The Metropolis–Hastings algorithm transforms a given stochastic matrix into a reversible stochastic matrix with a prescribed stationary distribution. We show that this transformation gives the minimum distance solution in an L^1 metric.

1. INTRODUCTION

Let $\mathbf{X}$ be a finite set and suppose that $\pi(x) > 0$, $\sum \pi(x) = 1$ is a probability distribution on $\mathbf{X}$ from which we wish to draw samples. The Metropolis (or Metropolis–Hastings) algorithm is a widely used procedure for drawing approximate samples from π, which works by finding a Markov chain on $\mathbf{X}$ with π as stationary distribution and using the well-known fact that after running the chain for a long time it is approximately distributed as π.

An important feature of the Metropolis–Hastings algorithm is that it can be applied when π is known only through ratios $\pi(x)/\pi(y)$. This allows the algorithm to be used without knowing the normalizing constant: for example, if π is specified as $\pi(x) \propto e^{-\beta H(x)}$, which is in practice impossible to sum if $\mathbf{X}$ is very large as happens in many problems in statistical physics, or $\pi(\theta) \propto P(\theta)L(x, \theta)$ as happens in Bayesian posterior distributions when P is a prior and L is a likelihood.

The algorithm is widely used for simulations in physics, chemistry, biology and statistics. It appears as the first entry of a recent list of great algorithms of 20th-century scientific computing [4]. Yet for many people (including the present authors) the Metropolis–Hastings algorithm seems like a magic trick. It is hard to see where it comes from or why it works. In this paper we give a development which explains at least some of the properties of the algorithm and we also show how there is a natural class of related algorithms, amongst which it is optimal.

Louis J. Billera is Professor of Mathematics and Operations Research, Department of Mathematics, 501 Malott Hall, Cornell University, Ithaca, New York 14853-4201. Persi Diaconis is Mary V. Sunseri Professor of Statistics and Professor of Mathematics, Stanford University, Stanford, California 94305.

The mechanics of the algorithm are simple to explain. The Metropolis–Hastings algorithm takes a base chain given by an absolutely arbitrary stochastic matrix $K(x, y)$ and transforms this into a stochastic matrix $M(x, y)$ that is reversible with respect to π, that is, one such that

$$\pi(x)M(x, y) = \pi(y)M(y, x). \tag{1.1}$$

This implies that π is a stationary distribution for M, and so we have found the required Markov chain.

The transformation is easy to implement. If

$$R(x, y) = \frac{\pi(y)K(y, x)}{\pi(x)K(x, y)}, \tag{1.2}$$

then

$$M(x,y)=\begin{cases} K(x,x)\min(1,R(x,y)), & \text{if } x\neq y, \\ K(x,y)+\sum_z K(x,z) & \\ \quad\times(1-\min(1,R(x,z))) & \text{else.}\end{cases} \tag{1.3}$$

The transformation (1.3) has a simple probabilistic interpretation: from x, choose y with probability $K(x, y)$. With probability $\min(1, R(x, y))$ accept this choice and move to y. Failing to accept, stay at x. It is easy to check directly that $M(x, y)$ satisfies the reversibility condition (1.1). If the base chain K is irreducible, then M is irreducible.

The Metropolis algorithm was introduced in 1953 by Metropolis, Rosenbluth, Rosenbluth, Teller and Teller (1953). In the original version K was taken as symmetric, so $K(x, y) = K(y, x)$ cancels out of (1.2), which makes the rejection step easier to understand: the chain moves to higher π-density regions automatically but only with appropriate probabilities to lower regions. The extension to more general nonsymmetric chains

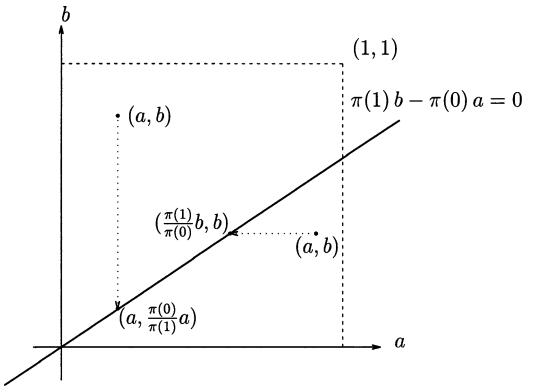

FIG. 1. *Metropolis map in the* 2×2 *case.*

appears in Hastings (1970). Textbook descriptions are in Hammersley and Handscomb (1964), Fishman (1996) and Liu (2001). Peskun (2001) shows that the Metropolis–Hastings algorithm leads to estimates with smallest asymptotic variance in a class of variants. Some approaches to convergence rates are summarized in Diaconis and Saloff-Coste (1998). Further recent references are Mengersen and Tweedie (1996) and Robert et al. (1997).

In this paper we show that the Metropolis–Hastings algorithm can be characterized geometrically as a projection in a particular L^1 norm onto $R(\pi)$, the set of π-reversible Markov chains as in (1.1). Because projections minimize distances, this shows that M is in this sense the closest reversible kernel to the original kernel, and hence it is a natural way in which to construct a good chain with π as its limiting distribution.

As an example, consider the 2×2 case. Then a stochastic matrix

$$\begin{pmatrix} 1-a & a \\ b & 1-b \end{pmatrix}, \qquad 0 \le a, b \le 1,$$

may be represented as a point (a, b) in the unit square. The space $R(\pi)$ of π-reversible matrices is given by the intersection of a line $\pi(1)b - \pi(0)a = 0$ with the unit square. If $\pi(1) > \pi(0)$ the picture is illustrated in Figure 1. The Metropolis–Hastings algorithm moves a point (a, b) to the line by projecting vertically for the top region and horizontally for the bottom region, as in the diagram.

In the general case, the space of stochastic matrices is partitioned into chambers (i.e., connected components of the complement of a finite union of hyperplanes in Euclidean space) and the Metropolis–Hastings algorithm is seen as piecewise linear over this partition, that is, a linear projection on each piece. Details are given in Section 2, which also illustrates the difficulty of replacing L^1 projections by L^2 projections. Section 3 gives a motivated development of the recipe (1.3). Section 4 gives an extension of our main result to general spaces.

2. THE METROPOLIS MAP

Let $\mathbf{X}$ be a finite set. Let $\mathcal{S}(\mathbf{X})$ be the set of stochastic matrices indexed by $\mathbf{X}$. Thus $K \in \mathcal{S}(\mathbf{X})$ has $K(x, y) \ge 0$ and $\sum_y K(x, y) = 1$ for each $x \in \mathbf{X}$. Note that $\mathcal{S}(\mathbf{X})$ is a convex set of dimension $|\mathbf{X}|(|\mathbf{X}| - \mathbf{1})$. Fixing a stationary distribution $\pi(x) > 0$, $\sum_x \pi(x) = 1$, let $R(\pi)$ be the subset of $\mathcal{S}(\mathbf{X})$ consisting of π-reversible Markov chains as in (1.1). Note that $R(\pi)$ is a convex set of dimension $|\mathbf{X}|(|\mathbf{X}|-\mathbf{1})/\mathbf{2}$. The Metropolis–Hastings algorithm gives a map M from $\mathcal{S}(\mathbf{X})$ onto $R(\pi)$:

$$M(K)(x, y) = \min\left(K(x, y), \frac{\pi(y)}{\pi(x)} K(y, x)\right) \tag{2.1}$$

for $x \ne y$. Observe that, for $x \ne y$, M is coordinatewise decreasing. In particular, $K(x, y) = 0$ implies $M(K)(x, y) = 0$. Thus M (weakly) increases the set of off-diagonal zeros.

Introduce a metric on $\mathcal{S}(\mathbf{X})$ via

$$d(K, K') = \sum_x \sum_{y \ne x} \pi(x) |K(x, y) - K'(x, y)|. \tag{2.2}$$

This is a metric on $\mathcal{S}(\mathbf{X})$: if $d(K, K') = 0$, then $K(x, y) = K'(x, y)$ for $x \ne y$ and by row stochasticity $K(x, x) = K'(x, x)$.

The main result of this paper follows.

THEOREM 1. *The Metropolis map* M *of* (2.1) *minimizes the distance* (2.2) *from* K *to* $R(\pi)$. *In fact,* $M(K)$ *is the unique closest element of* $R(\pi)$ *that is coordinatewise smaller than* K *on its off-diagonal entries.*

REMARK 2.1. In $\mathbb{R}^{\mathbf{X} \times \mathbf{X}}$, consider the hyperplanes

$$H_{xy} = \left\{K : \pi(x)K(x, y) = \pi(y)K(y, x)\right\}$$

and their corresponding half spaces

$$H_{xy}^- = \left\{K : \pi(x)K(x, y) < \pi(y)K(y, x)\right\},$$

$$H_{xy}^+ = \left\{K : \pi(x)K(x, y) > \pi(y)K(y, x)\right\}.$$

These divide the matrix space into chambers. Furthermore, $(\bigcap_{x \ne y} H_{xy}) \bigcap \mathcal{S}(\mathbf{X}) = \mathbb{R}(\pi)$. From (2.1) the map M is a diagonal linear map on each chamber of this hyperplane arrangement.

REMARK 2.2. A related natural metric is $d'(K, K') = \sum_{x,y} \pi(x)|K(x, y) - K'(x, y)|$. This is the total variation distance between the rows of K and K' weighted by the stationary distribution. The following example shows that the Metropolis chain does

not minimize d'. The example has three states and $\pi(x) = \frac{1}{3}$ for all x. With

$$K = \begin{pmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ \frac{3}{4} & \frac{1}{4} & 0 \\ \frac{1}{8} & 0 & \frac{7}{8} \end{pmatrix},$$

$$M(K) = \begin{pmatrix} \frac{5}{8} & \frac{1}{4} & \frac{1}{8} \\ \frac{1}{4} & \frac{3}{4} & 0 \\ \frac{1}{8} & 0 & \frac{7}{8} \end{pmatrix},$$

$$N = \begin{pmatrix} \frac{1}{8} & \frac{3}{4} & \frac{1}{8} \\ \frac{3}{4} & \frac{1}{4} & 0 \\ \frac{1}{8} & 0 & \frac{7}{8} \end{pmatrix},$$

we have $d'(K, M(K)) = \frac{5}{12}$ and $d'(K, N) = \frac{1}{3}$. Note that, for the metric d of (2.2), $d(K, M(K)) = d(K, N) = \frac{5}{24}$. This does not violate the claimed uniqueness in Theorem 1; the matrix N is not coordinatewise smaller than K in the (1, 2) entry. The following proof shows that $M(K)$ minimizes d' among elements of $R(\pi)$ coordinatewise less than K on off-diagonal entries.

REMARK 2.3. It is natural to seek an L^2 minimizer. A glance at Figure 1 shows the problem. Near the point $(1, 1)$ the unconstrained L^2 minimizer falls outside the box; it increases a and causes the diagonal entry in its row to become negative. Thus, to implement an L^2 minimizing projection, one would have to take into account *all* the entries in a row in order to limit the amount an entry might be increased. This would require a complete knowledge of all previous moves in the algorithm.

REMARK 2.4. The distance between $K \in \mathscr{S}(\mathbf{X})$ and $R(\pi)$ has a simple interpretation. Let $\overline{\pi K}(x, y) = \pi(x)K(x, y)$, $\overline{\pi K}(x, y) = \pi(y)K(y, x)$. Both are probabilities on $\mathbf{X}\times\mathbf{X}$ and K is π-reversible if and only if $\pi K = \overline{\pi K}$. With this notation, an easy computation shows

$$d(K, R(\pi)) = \|\pi K - \overline{\pi K}\|_{TV}.$$

PROOF OF THEOREM 1. We first argue that $d(K, R(\pi)) = d(K, M(K))$ by showing that $d(K, N) \geq d(K, M(K))$ for every N in $R(\pi)$. For this, note that

$$d(K, N) \geq \sum_{x,\, y:\, K \in H_{xy}^+} \Big(\pi(x)\,|K(x, y) - N(x, y)| + \pi(y)\,|K(y, x) - N(y, x)| \Big).$$

Since N is in $R(\pi)$, if $N(x, y) = K(x, y) + \varepsilon_{xy}$, $N(y, x) = \frac{\pi(x)}{\pi(y)}(K(x, y) + \varepsilon_{xy})$. Thus

$$\begin{aligned} d(K, N) &\geq \sum_{x,\, y:\, K \in H_{xy}^+} \pi(x)\,|\varepsilon_{xy}| + \pi(y) \left| K(y, x) - \frac{\pi(x)}{\pi(y)}(K(x, y) + \varepsilon_{xy}) \right| \\ &= \sum_{x,\, y:\, K \in H_{xy}^+} \pi(x)\,|\varepsilon_{xy}| + \big|(\pi(x)K(x, y) - \pi(y)K(y, x)) - \pi(x)\varepsilon_{xy}\big|. \end{aligned}$$

Using $|a - b| \geq |a| - |b|$ gives

$$\begin{aligned} d(K, N) &\geq \sum_{x,\, y:\, K \in H_{xy}^+} \big|\pi(x)K(x, y) - \pi(y)K(y, x)\big| \\ &= d(K, M(K)). \end{aligned}$$

When N is coordinatewise less than or equal to K on the off-diagonal entries, we have all $\varepsilon_{xy} \leq 0$. If any one is negative, we can conclude $d(K, N) > d(K, M(K))$, proving uniqueness. □

3. DEVELOPING THE ALGORITHM

The following development makes the Metropolis–Hastings algorithm seem rather natural and, in a sense, gives all possible variations.

Let $K(x, y)$ be a Markov chain on a finite state space $\mathbf{X}$. The goal is to change K to a chain with stationary distribution $\pi(x)$. The change must occur as follows: from x, choose y from $K(x, y)$ and decide to accept x or stay at y; this last choice may be stochastic with acceptance probability $F(x, y)$, $0 \leq F(x, y) \leq 1$. This gives a new chain with transition probabilities

$$K(x, y)F(x, y), \qquad x \neq y. \tag{3.1}$$

The diagonal entries are changed so that each row sums to 1.

The easiest way to get π-stationary is to insist on π-reversibility:

$$\pi(x)K(x, y)F(x, y) = \pi(y)K(y, x)F(y, x).$$

Set $R(x, y) = \pi(y)K(y, x)/\pi(x)K(x, y)$. The reversibility constraint becomes

$$\begin{aligned} F(x, y) &= R(x, y)F(y, x), \\ &0 \leq F(x, y) \leq 1. \end{aligned} \tag{3.2}$$

Since $F(y, x) = (1/R(x, y))F(x, y) \leq 1$ we must also have $F(x, y) \leq R(x, y)$ and so

$$0 \leq F(x, y) \leq \min(1, R(x, y)). \tag{3.3}$$

The point now is that $F(x, y)$ may be chosen *arbitrarily* to satisfy (3.3) for some fixed orientation of pairs and then $F(y, x)$ is forced by (3.2). This gives all modifications. We summarize.

PROPOSITION 3.1. *Let $K(x, y)$ be a Markov chain on a finite state space $\mathbf{X}$. For each pair $\{x, y\}$ choose an ordering (x, y). Choose*

$$0 \le F(x, y) \le \min\bigl(1, R(x, y)\bigr),$$

$$R(x, y) = \frac{\pi(y)K(y, x)}{\pi(x)K(x, y)},$$

and set $F(y, x) = (1/R(x, y))F(x, y)$. Then the Markov chain

$$M(x, y) = \begin{cases} K(x, y)F(x, y), & x \ne y, \\ K(x, x) + \sum_{y \ne x} K(x, y)\bigl(1 - F(x, y)\bigr) \end{cases}$$

satisfies $\pi(x)M(x, y) = \pi(y)M(y, x)$ for all x, y. Furthermore, this construction gives all possible functions F such that (3.1) *is π-reversible.*

REMARK 3.1. The classic Metropolis choice is $F(x, y) = \min(1, R(x, y))$. This maximizes the chance of moving from x. Suppose M and M' are π-reversible Markov chains on $\mathbf{X}$ with eigenvalues $1 = \lambda_0 \ge \lambda_1 \ge \cdots \ge \lambda_{|X|-1}$ and $1 = \lambda'_0 \ge \lambda'_1 \ge \cdots \ge \lambda'_{|X|-1}$ and also $M'(x, y) \le M(x, y)$ for all $x \ne y$. Then the minimax characterization of eigenvalues shows $\lambda_i \le \lambda'_i$ for all i. Thus, among the reversible chains of form (3.1), the Metropolis algorithm has the largest spectral gap $1 - \lambda_1$. Moreover, Peskun's theorem shows this gives the chain of form (3.1) with the minimum asymptotic variance for additive functionals (the usual form of statistical estimators as sums of a function evaluated along the chain). For contrast, chemists sometimes use "Barker dynamics," which amounts to the choice

$$F(x, y) = \frac{\pi(y)K(y, x)}{\pi(x)K(x, y) + \pi(y)K(y, x)} = \frac{R}{1 + R}.$$

Hastings (1970) introduced the equivalent form

$$F(x, y) = \frac{S(x, y)}{(1 + 1/R)}$$

for $S(x, y)$ a symmetric function satisfying

$$S(x, y) \le \min\left(1 + R(x, y), 1 + \frac{1}{R(x, y)}\right).$$

REMARK 3.2. It is natural to consider possible choices of F which are only functions of R : $F(x, y) = g(R(x, y))$. Then g must satisfy

$$g(x) = xg\left(\frac{1}{x}\right) \quad \text{for } x \in [0, \infty).$$

Such g may be chosen arbitrarily on $[0, 1]$ subject to

$$0 \le g(x) \le x, \qquad 0 \le x \le 1.$$

Then the functional equation specifies g on $[1, \infty)$. This gives Hastings' class of algorithms. The original Metropolis algorithm corresponds to $g(x) = x, 0 \le x \le 1$.

REMARK 3.3. Of course, the ergodicity of $M(x, y)$ must be checked; $K(x, x) \equiv 1$ for all x is a π-reversible chain!

REMARK 3.4. There are still mysterious features about the Metropolis–Hastings algorithm; when applied to natural generating sets of finite reflection groups, changing the stationary distribution proportional to length, the Metropolis–Hastings algorithm deforms the multiplication in the group algebra to multiplication in the Hecke algebra. See Diaconis and Hanlon (1992) and Diaconis and Ram (2000) for this.

4. MORE GENERAL SPACES

It is possible to extend Theorem 1 to general spaces. Let $\mathbf{X}$ be a complete separable metric space. Let μ be a σ-finite measure on $\mathbf{X}$ and $\pi(x)$ a strictly positive probability density with respect to μ. Let $\mathscr{S}$ be the set of Markov kernels $K(x, dy)$ on $\mathbf{X} \times \mathbf{X}$ having measurable densities $k(x, y)$ with respect to μ, with a possible atom at x allowed. The Metropolis–Hastings algorithm maps $K(x, dy)$ to $M(x, dy)$ with

$$M(x, dy) = k(x, y)\min(1, R(x, y))\mu(dy) + a(x)\delta_x(dy),$$

where

$$R(x, y) = \frac{\pi(y)k(y, x)}{\pi(x)k(x, y)}$$

and

$$a(x) = M\bigl(x, \{x\}\bigr) + \int \bigl(1 - R(x, y)\bigr)_+ k(x, y)\mu(dy).$$

This maps K to the π-reversible Markov kernels on $\mathbf{X}$. Define a metric on $\mathscr{S}$ by

$$d(K, K') = \int_{\mathbf{X}\times\mathbf{X}-\Delta} \pi(x)\bigl|k(x, y) - k'(x, y)\bigr| \times \mu(dx)\mu(dy),$$

with Δ the diagonal in $\mathbf{X} \times \mathbf{X}$. Then Theorem 1 holds as stated for this situation.

ACKNOWLEDGMENTS

We thank K. P. Choi, Thomas Yan, Richard Tweedie and a helpful referee for a careful reading and thoughtful comments on a previous version of this paper. The authors were supported in part by NSF Grants DMS-98-00910 and DMS-95-04379.

REFERENCES

DIACONIS, P. and HANLON, P. (1992). Eigenanalysis for some examples of the Metropolis algorithm. *Contemp. Math.* **138** 99–117.

DIACONIS, P. and RAM, A. (2000). Analysis of systematic scan Metropolis algorithms using Iwahori–Hecke algebra techniques. *Michigan Math. J.* **48** 157–190.

DIACONIS, P. and SALOFF-COSTE, L. (1998). What do we know about the Metropolis algorithm? *J. Comput. System. Sci.* **57** 20–36.

DONGARRA, J. and SULLIVAN, F., eds. (2000). The top 10 algorithms. *Comput. Sci. Engrg*. **2**.

FISHMAN, G. (1996). *Monte Carlo, Concepts, Algorithms and Applications*. Springer, New York.

HAMMERSLEY, J. and HANDSCOMB, D. (1964). *Monte Carlo Methods*. Chapman and Hall, New York.

HASTINGS, W. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57** 97–109.

LIU, J. (2001). *Monte Carlo Techniques in Scientific Computing*. Springer, New York.

MENGERSEN, K. and TWEEDIE, R. (1996). Rates of convergence of the Hastings and Metropolis algorithms. *Ann. Statist.* **24** 101–121.

METROPOLIS, N., ROSENBLUTH, A., ROSENBLUTH, M., TELLER, A. and TELLER, E. (1953). Equations of state calculations by fast computing machines. *J. Chem. Phys*. **21** 1087–1091.

PESKUN, P. (1973). Optimal Monte Carlo sampling using Markov chains. *Biometrika* **60** 607–612.

ROBERTS, G., GELMAN, A. and GILKS, W. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *Ann. Appl. Probab.* **7** 110–120.