The Metropolis-Hastings-Green Algorithm

Charles J. Geyer

December 12, 2003

1 Introduction

This note describes the simplified version of the Metropolis-Hastings-Green algorithm (Metropolis, et al., 1953; Hastings 1970; Green, 1995) that is most widely understood.

1.1 Dimension Changing

The Metropolis-Hastings-Green algorithm (as opposed to just Metropolis-Hastings with no Green) is useful for simulating probability distributions that are a mixture of distributions having supports of different dimension. An early example (predating Green's general formulation) was an MCMC algorithm for simulating spatial point processes (Geyer and Møller, 1994). More widely used examples are Bayesian change point models and Bayesian model selection (Green, 1995).

Abstractly, we are interested in a Markov chain having a state space that is a union of Euclidean spaces of different dimension. Let $\mathcal{X}$ denote the state space of the Markov chain. This is assumed to be a disjoint union

$$\mathcal{X} = \bigcup_{i \in I} \mathcal{X}_i \tag{1}$$

where each $\mathcal{X}_i$ is an open subset of $\mathbb{R}^d$ for some d. Let λ_i denote Lebesgue measure on $\mathcal{X}_i$ and let λ be the sum of the λ_i defined by

$$\lambda(B) = \sum_{i \in I} \lambda_i(B \cap \mathcal{X}_i).$$

We wish to simulate a Markov chain having stationary distribution with unnormalized probability density h with respect to λ, where h being an "unnormalized density" means

$$h(x) \ge 0, \qquad x \in \mathcal{X}$$

and

$$\int h(x) \lambda(dx) < \infty.$$

Note that the support of h, that is,

$$\{ x \in \mathcal{X} : h(x) > 0 \}$$

need not be all of $\mathcal{X}$. If we don't want the support of h to be all of $\mathcal{X}$ we just define $h(x) = 0$ for x not in the desired support.

1.2 Elementary Updates, First Version

In this section we describe an elementary update that moves the Markov chain one step. Iterating the update over and over makes a Markov chain, although more commonly the elementary update is combined in various ways with other elementary updates. The actual Markov chain of interest iterates the combined update.

The elementary update requires the following items.

- [**Augmentation**] Each subspace $\mathcal{X}_i$ is "augmented" by a subspace $\mathcal{Y}_i$, also an open subset of $\mathbb{R}^d$ for some d, such that $\mathcal{Z}_i = \mathcal{X}_i \times \mathcal{Y}_i$ has the same dimension for all i. Write
$$\mathcal{Z} = \bigcup_{i \in I} \mathcal{Z}_i$$
(the "augmented" state space).
- [**Proposal**] For each $i \in I$ and each $x \in \mathcal{X}$ there is a (proper) probability density $q(x, \cdot)$ on $\mathcal{Y}_i$ such that
  - random variates having density $q(x, \cdot)$ can be simulated for each $x \in \mathcal{X}_i$,
  - and $q(x, y)$ can be evaluated for each $(x, y) \in \mathcal{Z}$.
- [**Transformation**] There is a function $g : \mathcal{Z} \to \mathcal{Z}$ such that
  - g is its own inverse,
  - and the jacobian of g is everywhere nonsingular.

Here is the description of the update, supposing the current state is denoted by x.

1. (The "proposal"). Generate a random variate y having density $q(x, \cdot)$, and define $(x^*, y^*) = g(x, y)$.
2. (The "Green ratio"). Define
$$r(x, y) = \frac{h(x^*) q(x^*, y^*)}{h(x) q(x, y)} \cdot |\nabla g(x, y)| \tag{2}$$

3. ("Metropolis rejection") "accept" the proposal, that is, the updated state is x^*, with probability

$$a(x, y) = \min\bigl(1, r(x, y)\bigr) \tag{3}$$

and "reject" the proposal, that is, the updated state remains as it was, at x, with probability $1 - a(x, y)$.

Let us say a state x is *feasible* if $h(x) > 0$. Note that if the current state x is feasible,

- then with probability one $q(x, y) > 0$ and the denominator in (2) is nonzero, and
- if the proposal is accepted, then with probability one we have $h(x^*) > 0$ so the updated state is also feasible.

Thus if started in a feasible state, the Metropolis-Hastings-Green algorithm forever remains in the set of feasible states (with probability one). If started in a non-feasible state (2) is undefined because of division by zero and the algorithm is meaningless. Note that *proposal* of non-feasible states is allowed. They are accepted with probability zero, but that creates no problem as long as some proposal in some iteration of the algorithm is eventually accepted. What is not allowed is a *starting position* that is not feasible.

1.3 State-Dependent Mixing

Another innovation of Green's paper is what I call "state-dependent mixing". In practice one uses many elementary updates like those described in the preceding section as well as other Gibbs and Metropolis-Hastings updates.

Let the elementary updates under consideration be described by kernels P_j, $j \in J$.

The fundamental property of any MCMC update P_j is that it preserve the desired stationary distribution π, meaning $\pi P_j = \pi$ or, in words, π is invariant for P_j.

If we make a random choice of which update to do, chosing P_j with probability c_j, which does not depend on the state x, then trivially if each P_j preserves π, the kernel for the combined update, which is

$$P = \sum_{j \in J} c_j P_j$$

trivially also preserves π because

$$\pi P = \sum_{j \in J} c_j \pi P_j = \sum_{j \in J} c_j \pi = \pi$$

But this proof does not work if the probabilities are allowed to depend on the state x! We need another idea. Green's idea is the following. Define

$$K_j(x, A) = c_j(x) P_j(x, A)$$

now K_j is not a Markov transition kernel, but it is still a general kernel. It makes no sense to talk about a non-Markov kernel "preserving a distribution". But it does make sense to talk about reversibility with respect to a distribution for an arbitrary kernel. We say that a kernel K is reversible with respect to a positive finite measure η (a possibly unnormalized probability distribution) if the value of

$$\iint f_1(x) f_2(x^*) \eta(dx) K(x, dx^*)$$

is unchanged by interchanging f_1 and f_2. Now it is obvious from linearity of integration that if each K_j is reversible with respect to η, then so is $K = \sum_{j \in J} K_j$.

Thus we use *reversibility with respect to* η rather than *preserving* η and we do get what we want: if K is Markov and reversible with respect to η, then K preserves η.

Actually it is enough that K be sub-Markov, meaning

$$K(x, \mathcal{X}) \le 1, \qquad x \in \mathcal{X}$$

Because then it is easily verified that

$$P(x, A) = I(x, A)\big[1 - K(x, \mathcal{X})\big] + K(x, A) \tag{4}$$

is Markov and reversible with respect to η, where I is the identity kernel defined by

$$I(x, A) = \begin{cases} 1, & x \in A \\ 0, & x \notin A \end{cases}$$

What this sub-Markov property allows is the following algorithm (as usual, x is the current state).

- With probability $c_j(x)$ update generate the next value of the state from the distribution $P_j(x, \cdot\,)$.
- With probability $1 - \sum_j c_j(x)$ do nothing (the next value of the state is the same as the current value x).

If the $c_j(x)$ sum to one for all x, the second option is never applicable. If they don't then that is what gives rise to the first term on the right hand side of (4).

When state-dependent mixing is used, the Green ratio (2) must be modified! It is replaced by

$$r_j(x, y) = \frac{c_j(x^*) h(x^*) q_j(x^*, y^*)}{c_j(x) h(x) q_j(x, y)} \cdot |\nabla g_j(x, y)| \tag{5}$$

where we have put a subscript j on everything that depends on the update. Other than the insertion of $c_j(x)$ and $c_j(x^*)$ in the Green ratio, everything remains the same.

1.4 A Comment about Augmentation

This version of the Metropolis-Hastings-Green algorithm is remarkable in its use of "throw away" augmentation. It does not just augment the state of the Markov chain, as many people had done before Green.

Each update has its own augmentation. Say the j-th update moves between $\mathcal{X}_i$ and $\mathcal{X}_{i'}$ (what was j in the "Elementary Updates" is now i'). Then the augmentations are $\mathcal{Y}_{ij}$ and $\mathcal{Y}_{i'j}$ such that $\mathcal{X}_i \times \mathcal{Y}_{ij}$ and $\mathcal{X}_{i'} \times \mathcal{Y}_{i'j}$ have the same dimension. For $x \in \mathcal{X}_k$, $k \in \{i, i'\}$ the proposal is in $\mathcal{Y}_{kj}$ and has density $q_j(x, \cdot)$, and so forth.

It is entirely possible that different updates have the same $\mathcal{X}_i$ and $\mathcal{X}_{i'}$ but different $\mathcal{Y}_{ij}$ or $\mathcal{Y}_{i'j}$. So what is "augmented" is not the state of the Markov chain, but the imaginary state of an imaginary Markov chain that only involves the j-th elementary update. Different imaginary Markov chains (different updates) may have different augmentation.

2 Theory

The proof that the algorithm described in the preceding section preserves the distribution with unnormalized density h has two parts. First we show that if each K_i is reversible with respect to η and $K = \sum_{j \in J} K_j$ is sub-Markov, then (4) is Markov and preserves η. Second, we show that when the Green ratio is defined by (5), the K_j are reversible with respect to the measure η having unnormalized density h.

2.1 Part I

First if K is sub-Markov, then so is each K_j because the K_j are nonnegative so $K_j \le K$. As we remarked above, if each K_j is reversible with respect to η, then so is K, just by linearity of integration. So the only thing that needs to be shown in part one of the proof is that if K is sub-Markov and reversible with respect to η, then P given by (4) is Markov and reversible with respect to η and that this implies that P preserves η.

Let f_1 and f_2 be bounded measurable functions. Then

$$\begin{aligned}
\iint f_1(x) f_2(x^*) \eta(dx) P(x, dx^*) &= \iint f_1(x) f_2(x^*) \eta(dx) I(x, dx^*) \bigl[1 - K(x, \mathcal{X})\bigr] \\
&\quad + \iint f_1(x) f_2(x^*) \eta(dx) K(x, dx^*) \\
&= \int f_1(x) f_2(x) \bigl[1 - K(x, \mathcal{X})\bigr] \eta(dx) \\
&\quad + \iint f_1(x) f_2(x^*) \eta(dx) K(x, dx^*)
\end{aligned}$$

The first term in the last expression is trivially unchanged by interchanging f_1 and f_2 (because multiplication is commutative) and the second term is un-

changed by interchanging f_1 and f_2 because K is reversible with respect to η. Hence P is reversible with respect to η.

Also

$$\begin{aligned} P(x, \mathcal{X}) &= I(x, \mathcal{X})\big[1 - K(x, \mathcal{X})\big] + K(x, \mathcal{X}) \\ &= 1 - K(x, \mathcal{X}) + K(x, \mathcal{X}) \\ &= 1 \end{aligned}$$

so P is Markov.

The proof that P preserves η is then trivial. Reversibility with respect to η is

$$\iint f_1(x) f_2(x^*) \eta(dx) P(x, dx^*) = \iint f_2(x) f_1(x^*) \eta(dx) P(x, dx^*) \tag{6}$$

Take $f_1 = I_A$ and $f_2 = 1$. Then (6) becomes

$$\iint I_A(x) \eta(dx) P(x, dx^*) = \iint I_A(x^*) \eta(dx) P(x, dx^*)$$

or

$$\int_A \eta(dx) P(x, \mathcal{X}) = \int \eta(dx) P(x, A)$$

Because P is Markov $P(x, \mathcal{X}) = 1$ so the left hand side is $\int_A \eta(dx) = \eta(A)$. Hence

$$\eta(A) = \int \eta(dx) P(x, A), \qquad \text{for all } A$$

which is $\eta = \eta P$ written out in full.

2.2 Part II

Now it remains to be shown that each K_j is reversible with respect to the distribution having unnormalized density h with respect to λ. We can drop the subscript j because the argument involves only one update at a time.

We start by proving something rather different. Define $z = (x, y)$ and $w(z) = h(x)q(x, y)$. Then we can also consider this update as updating z and being reversible with respect to the distribution with unnormalized density w with respect to $\nu = \lambda \times \mu$, where μ is the sum of the μ_i.

We can rewrite (5) and (3) as

$$r(z) = \frac{c(z^*) w(z^*)}{c(z) w(z)} \cdot |\nabla g(z)| \tag{7}$$

and

$$a(z) = \min\big(1, r(z)\big) \tag{8}$$

We are to show that the kernel

$$K(z, A) = c(z) P(z, A) \tag{9}$$

where

$$P(z, A) = I(z, A)\big[1 - a(z)\big] + I(z, g(A))a(z) \tag{10}$$

is reversible with respect to the desired stationary distribution. This happens if the value of

$$\iint f_1(z) f_2(z^*) w(z) \nu(dz) K(z, dz^*) \tag{11}$$

is unchanged by interchanging f_1 and f_2, where f_1 and f_2 are any bounded measurable functions.

Now (11) is equal to

$$\int f_1(z) f_2(z) w(z)\big[1 - a(z)\big] c(z) w(z) \nu(dz) + \int f_1(z) f_2\big(g(z)\big) c(z) w(z) a(z) \nu(dz)$$

and the value of the first term is obviously unchanged by interchanging f_1 and f_2 (because multiplication is commutative). Thus we only need to check the second term, that is we need to show that the value of

$$\int f_1(z) f_2\big(g(z)\big) c(z) w(z) a(z) \nu(dz) \tag{12}$$

is unchanged by interchanging f_1 and f_2.

Now

$$\begin{aligned}
&\int f_2(z) f_1\big(g(z)\big) c(z) w(z) a(z) \nu(dz) \\
&\qquad = \int f_2\big(g(z^*)\big) f_1(z^*) c\big(g(z^*)\big) w\big(g(z^*)\big) a\big(g(z^*)\big) |\nabla g(z^*)| \nu(dz^*) \\
&\qquad\qquad = \int f_2\big(g(z)\big) f_1(z) c\big(g(z)\big) w\big(g(z)\big) a\big(g(z)\big) |\nabla g(z)| \nu(dz)
\end{aligned}$$

the first equality being the change of variable theorem for integration and the second equality being simply that the notation used for a dummy variable of integration (z^* or z) doesn't matter.

Comparing the last form with (12) we see they have the same value if

$$c(z) w(z) a(z) = c\big(g(z)\big) w\big(g(z)\big) a\big(g(z)\big) |\nabla g(z)| \tag{13}$$

almost everywhere with respect to ν (so that is all that remains to be shown).

It simplifies things if we use the notation $z^* = g(z)$ so (13) becomes

$$c(z) w(z) a(z) = c(z^*) w(z^*) a(z^*) |\nabla g(z)| \tag{14}$$

Now note that from the inverse mapping theorem and the fact that g is its own inverse we have

$$|\nabla g(z^*)| \cdot |\nabla g(z)| = 1$$

from which it follows that

$$r(z^*) = \frac{1}{r(z)} \tag{15}$$

at points where neither $r(z)$ or $r(z^*)$ is zero.

With this in hand, we prove (14) by looking at several cases.

Case I. $c(z)w(z) = 0$ from which we conclude $a(z^*) = 0$. This makes both sides of (14) zero. So (14) checks in this case.

Case II. $c(z^*)w(z^*) = 0$ from which we conclude $a(z) = 0$. Again this makes both sides of (14) zero. So (14) checks again.

Case III. $c(z)w(z) > 0$ and $c(z^*)w(z^*) > 0$ and $r(z) > 1$ from which we conclude $r(z^*) < 1$. Hence $a(z) = 1$ and $a(z^*) = r(z^*)$. So the left hand side of (14) is $c(z)w(z)$ and the right hand side is

$$c(z^*)w(z^*)r(z^*)|\nabla g(z)| = c(z^*)w(z^*)\frac{c(z)w(z)}{c(z^*)w(z^*)}|\nabla g(z^*)| \cdot |\nabla g(z)| = c(z)w(z)$$

so (14) checks in this case.

Case IV. $c(z)w(z) > 0$ and $c(z^*)w(z^*) > 0$ and $r(z) \leq 1$ from which we conclude $r(z^*) \geq 1$. Hence $a(z) = r(z)$ and $a(z^*) = 1$. So the right hand side of (14) is $c(z^*)w(z^*)|\nabla g(z)|$ and the right hand side is

$$c(z)w(z)r(z) = c(z)w(z)\frac{c(z^*)w(z^*)}{c(z)w(z)}|\nabla g(z)| = c(z^*)w(z^*)|\nabla g(z)|$$

and (14) checks in this case too.

The proof is finshed, but we add a comment that cases I and II are a bit tricky because in case I, for example, $r(z)$ and hence $a(z)$ is undefined, but it does not matter since $a(z)$ is multiplied by zero. Strictly speaking, perhaps we should define $a(z)$ to be something (say 1) in this case, so the multiplication is well defined, but it is clear that the details don't matter. Any choice, so long as some definite choice is made, will do.

Actually one might also wonder about another issue. We have shown that the update is reversible with respect to the stationary distribution with unnormalized density w, but that wasn't the original problem. What about x's alone not (x, y) pairs? Trivial. Just consider functions f_1 and f_2 that are functions of x only.

Thus the proof also shows that the update considered as a random move from x to either x or x^* depending on Metropolis rejection is reversible with respect to the distribution having unnormalized density h with respect to λ.

References

Geyer, C. J. and Møller J. (1994). Simulation and likelihood inference for spatial point processes. *Scandanavian Journal of Statistics*, 21, 359–373.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82, 711–732.

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

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*, 21, 1087–1092.