

# Hastings-Metropolis Algorithms and Reference Measures \*

Pei-de Chen,  
Department of Statistics, Colorado State University<sup>†</sup>

February 21, 1997

## Abstract

In this note, we discuss the possible existence and effect of different reference measures on the Hastings-Metropolis (H-M) algorithm. We prove that the standard H-M algorithm as a Markov chain does not depend on the choice of the reference measure, and we obtain a sufficient and necessary condition for the existence of a reference measure.

## 1 Introduction

The Hastings and Metropolis (H-M) algorithms allow simulation of a probability distribution  $\Pi$  which is only known up to a constant (normalising) factor. This is surprisingly widely relevant, occurring especially when  $\Pi$  is a Bayesian posterior distribution, but in many other contexts also [1].

In the standard construction [3, 2] of the H-M algorithm on a space  $S$ , one first considers a *candidate transition kernel*  $Q(x, \cdot)$ ,  $x \in S$ , which generates potential transitions for a discrete time Markov chain evolving on  $S$ . Traditionally  $S$  is either a countable space or  $\mathbb{R}^k$  equipped with the Borel  $\sigma$ -field  $\mathcal{B}(S)$ , and both  $\Pi$  and  $Q(x, \cdot)$  have densities  $\pi(y)$  and  $q(x, y)$  with respect to a reference measure taken either as counting measure or as Lebesgue measure. Much more general formulations are possible (see [7, 1]) on a general space  $(S, \mathcal{B}(S))$  if there is some existing reference measure  $\mu$  available on the space.

---

\*From a dissertation submitted to the Academic Faculty of Colorado State University in partial fulfillment of the requirements for the degree of Ph. D., supervised by Prof. R. Tweedie. Work supported in part by NSF Grant DMS 9504561

<sup>†</sup>Postal Address: Department of Statistics, Colorado State University, Fort Collins CO 80523, USA; email: pchen@stat.colostate.edu

The role of the reference measure is fundamental in the definition of the H-M algorithm, which works as follows. A “candidate transition” to  $y$ , generated according to the density  $q(x, y)$ , is accepted with probability  $\alpha(x, y)$ , given by

$$\alpha(x, y) = \begin{cases} \min\{\frac{\pi(y)}{\pi(x)} \frac{q(y, x)}{q(x, y)}, 1\} & \pi(x)q(x, y) > 0 \\ 1 & \pi(x)q(x, y) = 0 \end{cases} \quad (1)$$

Thus actual transitions of the H-M chain take place according to a law  $P(x, \cdot)$  with transition densities  $p(x, y) = q(x, y)\alpha(x, y)$ ,  $y \neq x$  and with probability of remaining at the same point given by  $r(x) = P(x, \{x\}) = \int q(x, y)[1 - \alpha(x, y)]dy$ .

The crucial property of the H-M algorithm is that, with this choice of  $\alpha$ , the target  $\Pi$  is invariant for the operator  $P$ . It is then standard [4, 5] that under some weak conditions, including irreducibility and aperiodicity, the  $n$ -step transition probabilities  $P^n(x, A)$  converge to  $\Pi$  in the total variation norm.

Recently Tierney [6] extended the classical H-M acceptance probability defined by (1) to more general rules which still lead to the property that  $\Pi$  is the invariant measure for the resulting chain. His construction does not require reference measures at all.

In this paper our goal is two-fold: firstly, to see what effect a change in the choice of reference measure has on the H-M algorithm (or the more general extension by Tierney); and secondly, to find conditions under which there exists such a reference measure for the algorithm.

To achieve these goals we consider the standard H-M algorithm in the following way. We start from a target probability  $\Pi$  and a candidate probability measure  $\lambda$  on the product state space  $(S \times S, \mathcal{B}(S) \times \mathcal{B}(S))$ . We will call a measure  $\mu$  an available reference measure if there is a measurable function  $q$  such that for each measurable set  $A, B \in \mathcal{B}(S)$  we have

$$\lambda(A \times B) = \int_A \int_B \mu(dx)q(x, y)\mu(dy);$$

the candidate kernel  $Q$  is then given by  $Q(x, A) = \int_A q(x, y)\mu(dy)$  as usual. This approach allows us to keep the right to choose an available reference measure (if any) freely instead

of assuming a fixed one as a prerequisite.

**Remark** The target distribution does not effect the existence of an available reference measure. In fact, if  $\mu$  is available for  $Q$ , then given any target distribution  $\Pi$ ,  $\nu = \frac{1}{2}(\mu + \Pi)$  is available for  $Q$  and  $\Pi$ . So in the discussion of the existence of a reference measure, we may ignore the target  $\Pi$ .

An easy example shows that an available reference measure does not always exist, and in Section 3 we find a sufficient and necessary condition for the existence of an available reference measure. Clearly if available reference measures exist, they form a large class (if  $\mu$  is a member, any  $\sigma$ -finite measure  $\nu$  such that  $\mu \ll \nu$  is also a member). But fortunately, the standard H-M algorithm as a Markov chain (or its transition probability kernel) does not depend on the choice of the reference measure, and we now prove this.

## 2 The Effect of the Reference Measure

Before discussing the existence of available reference measures, we first prove that the algorithm essentially does not depend on the choice of available reference measures.

**Theorem 2.1** *Suppose that an available reference measure exists for a given target  $\Pi$  and candidate transition law  $Q$ . Then the Hastings algorithm essentially does not depend on the choice of reference measures. More exactly, for different reference measures  $\mu_1$  and  $\mu_2$ , the corresponding transition kernels  $P_1$  and  $P_2$  coincide except on a  $\Pi$ -negligible subset  $\Lambda$ :*

$$P_1(x, \cdot) = P_2(x, \cdot) \quad \text{for all } x \notin \Lambda \in \mathcal{B}(S), \quad \Pi(\Lambda) = 0.$$

**PROOF** Without loss of generality, we may assume that  $\mu_1 \ll \mu_2$ . Otherwise let  $\mu_3 = \frac{\mu_1 + \mu_2}{2}$ , then  $\mu_1 \ll \mu_3$ , and  $\mu_2 \ll \mu_3$ . If  $\mu_1$  and  $\mu_3$  provide the same chain, and so do  $\mu_2$  and  $\mu_3$ , then  $\mu_1$  and  $\mu_2$  also provide the same chain.

Let  $\pi_1 = \frac{d\Pi}{d\mu_1}$ ,  $\pi_2 = \frac{d\Pi}{d\mu_2}$ ,  $\phi = \frac{d\mu_1}{d\mu_2}$ , then  $\pi_2 = \pi_1\phi$ . Denote

$$q_1(x, \cdot) = \frac{dQ(x, \cdot \setminus \{x\})}{d\mu_1}(\cdot), \quad q_2(x, \cdot) = \frac{dQ(x, \cdot \setminus \{x\})}{d\mu_2}(\cdot).$$

We have  $q_2(x, y) = q_1(x, y)\phi(y)$ . Thus

$$\begin{aligned} \alpha_2(x, y) &= \frac{\pi_2(y)q_2(y, x)}{\pi_2(x)q_2(x, y)} \wedge 1 \\ &= \frac{\pi_1(y)\phi(y)q_1(y, x)\phi(x)}{\pi_1(x)\phi(x)q_1(x, y)\phi(y)} \wedge 1 \\ &= \frac{\pi_1(y)q_1(y, x)}{\pi_1(x)q_1(x, y)} \wedge 1 \\ &= \alpha_1(x, y), \end{aligned}$$

if  $\pi_2(x)q_2(x, y) > 0$  (which implies both  $\phi(x) > 0$  and  $\phi(y) > 0$ ).

Now suppose  $\pi_2(x)q_2(x, y) = 0$  and also both  $\phi(x) > 0$  and  $\phi(y) > 0$ . Then we must have  $\pi_1(x)q_1(x, y) = 0$ , so we have still  $\alpha_2(x, y) = 1 = \alpha_1(x, y)$ . Therefore

$$\begin{aligned} P_1(x, A) &= 1_A(x)P_1(x, \{x\}) + \int_A p_1(x, y)\mu_1(dy) \\ &= 1_A(x) \{ \int_S q_1(x, y) [1 - \alpha_1(x, y)] \mu_1(dy) + Q(x, \{x\}) \} \\ &\quad + \int_A \alpha_1(x, y)q_1(x, y)\mu_1(dy) \\ &= 1_A(x) \int_{S \cap \{\phi > 0\}} q_2(x, y) [1 - \alpha_2(x, y)] \mu_2(dy) \\ &\quad + 1_A(x)Q(x, \{x\}) + \int_{A \cap \{\phi > 0\}} \alpha_2(x, y)q_2(x, y)\mu_2(dy) \\ &= P_2(x, A), \end{aligned}$$

for all  $x \in \{\phi > 0\}$ . Here we use

$$q_1(x, y)\mu_1(dy) = q_1(x, y)\phi(y)\mu_2(dy) = q_2(x, y)\mu_2(dy),$$

and since the integrands are zero on  $\{y : \phi(y) = 0\}$ , we may restrict the integration only to  $\{\phi > 0\}$ , while the replacement of  $\alpha_1(x, y)$  by  $\alpha_2(x, y)$  is available for all  $x$  such that  $\phi(x) > 0$ . Hence

$$\Lambda = \{x : P_1(x, \cdot) \neq P_2(x, \cdot)\} \subset \{x : \phi(x) = 0\}$$

is  $\mu_1$ -negligible, and so is  $\Pi$ -negligible as required.  $\square$

### 3 The Existence of Reference Measures

In this section we seek conditions under which a reference measure may exist, enabling the formulation (1).

The following simple example shows that available reference measures may fail to exist

**Example 3.1** Let  $S = \mathbb{R} = (-\infty, \infty)$ ,  $Q(x, \{x+1\}) = 1$ , for all  $x \in \mathbb{R}$ . Then any available reference measure  $\mu$  should have  $x$  as an atom for any  $x \in \mathbb{R}$ , and such  $\mu$  cannot be  $\sigma$ -finite. So no available reference measure exists for such a candidate transition kernel.

We first consider the independent case, where normally it is assumed the transition kernel  $Q$  satisfies  $Q(x, \cdot) = Q(\cdot)$  for all  $x$ . We relax this slightly by assuming we have three factors, none of which is related to a pre-assigned reference measure:

- (i) An arbitrary target distribution  $\Pi$ ,
- (ii) A common staying probability  $\rho = Q(x, \{x\})$ ,  $0 \leq \rho \leq 1$ ,
- (iii) A common transition law  $Q^\circ$ , i.e., a measure on  $\mathcal{B}(S)$ , with total mass  $1 - \rho$ .

The true candidate transition probability  $Q$  is then given by

$$Q(x, A) = \rho 1_A(x) + Q^\circ(A).$$

If  $\rho = 1$ ,  $Q^\circ$  is a zero measure, and there is nothing interesting to be studied: any target distribution become a stationary distribution, but the chain is not irreducible, and so  $P^n(x, \cdot)$  does not converge to any stationary distribution.

However, assuming  $0 \leq \rho < 1$ , it is easily checked that  $\mu = \frac{1}{2}\Pi + \frac{1}{2(1-\rho)}Q^\circ$  is an available reference measure. Thus the independent case is trivial.

Let us next consider the symmetric candidate. Since  $\mathbb{R}^k$  is a linear space, one can consider the classical random walk based Metropolis algorithm [3], in which  $q(x, y) = q(y - x)$  with  $q$  a symmetric sub-probability density on  $\mathbb{R}^k$  with respect to a reference

measure  $\mu$  .

In the general case, the state space usually does not have a linear structure, and the random walk becomes meaningless. Now, if we want to avoid assuming the existence of a reference measure, the candidate transition kernel is not a reasonable object to replace the density, because we have no way to define the symmetry of a kernel. The natural object to replace the density is a measure on the self-product state space. So in the following discussion, we regard a probability measure on  $\mathcal{B}(S) \times \mathcal{B}(S)$  as a candidate to replace the candidate density. If an available reference measure exists, this will produce a candidate transition kernel in a natural way.

When a reference measure  $\mu$  exists, we let  $\tilde{Q}$  be the symmetric probability measure on  $\mathcal{B}(S) \times \mathcal{B}(S)$  defined by

$$\begin{aligned}\tilde{Q}(B \times A) &= \int_B Q(x, A)\mu(dx) \\ &= \int_B \int_A q(x, y)\mu(dy)\mu(dx) + \int_{A \cap B} \rho(x)\mu(dx).\end{aligned}$$

where  $0 \leq \rho(x) = 1 - Q(x, S) \leq 1$  is the extra probability of not moving. We will regard  $\tilde{Q}$  as a candidate measure of the Metropolis algorithm in the wide sense.

A probability measure  $\lambda$  defined on a self-product  $\sigma$ -field of the form  $\mathcal{B}(S) \times \mathcal{B}(S)$  is called symmetric if any product measurable set  $G \in \mathcal{B}(S) \times \mathcal{B}(S)$  has the same  $\lambda$ -value as its transpose  $G' = \{(x, y) : (y, x) \in G\}$ , i.e.,  $\lambda(G) = \lambda(G')$ . We are interested in distinguishing what kind of symmetric measure can be regarded as a candidate of a Metropolis algorithm in the wide sense above.

Any probability measure  $\lambda$  on  $\mathcal{B}(S) \times \mathcal{B}(S)$  can be decomposed into two parts, the diagonal part  $\lambda_d$  and the off-diagonal part  $\lambda_o$ , as follows:

$$\lambda_d(B \times A) = \lambda(B \times A \cap \Delta)$$

with the diagonal  $\Delta = \{(x, x) : x \in S\}$  as its support, and

$$\lambda_o(B \times A) = \lambda(B \times A \setminus \Delta)$$

with  $S \times S \setminus \Delta$  as its support.

Now put  $\mu(B) = \lambda(B \times S)$ . This defines a probability measure on  $\mathcal{B}(S)$  known as the marginal probability measure on the first factor space. If we project the product space  $S \times S$  to the first factor space,  $\lambda_d$  induces a measure on  $\mathcal{B}(S)$ , denoted by  $\hat{\lambda}_d$  which is absolutely continuous with respect to  $\mu$ ,  $\hat{\lambda}_d(B) = \lambda_d(B \times S) \leq \mu(B)$ . Denote  $\frac{d\hat{\lambda}_d}{d\mu}$  by  $\rho$ , then

$$\begin{aligned} \lambda_d(B \times A) &= \lambda_d((B \cap A) \times (B \cap A)) \\ &= \hat{\lambda}_d(B \times A) \\ &= \int_{B \cap A} \rho(x) \mu(dx). \end{aligned}$$

**Theorem 3.1** *A symmetric probability measure  $\lambda$  on  $\mathcal{B}(S) \times \mathcal{B}(S)$  can be regarded as a candidate of a Metropolis algorithm  $\tilde{Q}$ , if and only if its off-diagonal part  $\lambda_o \ll d\mu \times d\mu$ , where  $\mu$  is the marginal probability measure of  $\lambda$  (on either factor spaces due to the symmetry of  $\lambda$ ). In this case,  $\mu$  is an available reference probability measure,  $\rho = \frac{d\hat{\lambda}_d}{d\mu}$  defined as above is the staying probability, and  $q(x, y) = \frac{d}{d\mu} \left[ \frac{d\tilde{Q}_o(\cdot \times \cdot)}{d\mu} (x, \cdot) \right] (x, y)$  is the density part of the candidate.*

**PROOF** For necessity, notice that  $\tilde{Q}$  is naturally decomposed into the diagonal part  $\tilde{Q}_d$ :  $\tilde{Q}_d(B \times A) = \int_{A \cap B} \rho(x) \mu(dx)$  and the off-diagonal part  $\tilde{Q}_o$ :  $\tilde{Q}_o(B \times A) = \int_B \int_A q(x, y) \mu(dy) \mu(dx)$ , and obviously  $\tilde{Q}_o \ll d\mu \times d\mu$ . Here the reference measure of the algorithm  $\mu$  is just the marginal probability measure of  $\tilde{Q}$ .

Remember here that in the definition of H-M algorithms,  $q(x, x) = 0$  for all  $x \in S$ , so we need not remove the diagonal from the region of integration.

For sufficiency, if  $\lambda_o \ll d\mu \times d\mu$ , then  $q(x, y) = \frac{d\lambda_o}{d\mu \times d\mu}$  can be regarded as the density part of the algorithm with  $\mu$  as the reference measure, as required.  $\square$

Non-symmetry does not cause much more trouble. We have

**Theorem 3.2** *A probability measure  $\lambda$  on  $\mathcal{B}(S) \times \mathcal{B}(S)$  can be regarded as a candidate of a*

Hastings algorithm if and only if  $\lambda_o \ll \mu_1 \times \mu_1$ , where  $\mu_1$  is the marginal probability measure of  $\lambda$  on the first factor space.

PROOF Since

$$\begin{aligned}\tilde{Q}(B \times A) &= \tilde{Q}_o(B \times A) + \tilde{Q}_d(B \times A) \\ &= \int_B \int_A q(x, y) \mu(dy) \mu(dx) + \int_{A \cap B} \rho(x) \mu(dx)\end{aligned}$$

still defines a probability measure on  $\mathcal{B}(S) \times \mathcal{B}(S)$  with  $\mu$  as its marginal measure on the first factor space (without symmetry,  $\mu$  need not to be the marginal measure on the second factor space), the condition is necessary.

The same argument as in the proof of Theorem 3.2 shows the sufficiency.  $\square$

The main difference with the symmetric case is that the marginal probability measure of  $\lambda$  on the second factor space  $\mu_2$  may be different from  $\mu_1$ , since

$$\begin{aligned}\mu_2(A) &= \lambda(S \times A) \\ &= \int_A [\int_S q(x, y) \mu_1(dx) + \beta(y)] \mu_1(dy),\end{aligned}$$

but obviously,  $\mu_2 \ll \mu_1$  (on condition  $\lambda_o \ll \mu_1 \times \mu_1$ ).

However,  $\mu_1 \ll \mu_2$  need not hold. To see this consider the following example.

Let  $S = (0, 1]$ ,  $\mathcal{B}(S)$  be its Borel  $\sigma$ -field,  $\lambda = \lambda_d + \lambda_o$ , where  $\lambda_d$  is uniformly distributed on  $\Delta_1 = \{(x, x) : x \in (0, \frac{1}{2}]\}$  with total mass  $\frac{1}{2}$ , and  $\lambda_o$  is uniformly distributed on  $(0, 1] \times (0, \frac{1}{2}]$  with total mass  $\frac{1}{2}$ . Then  $\mu_1 = \frac{3}{2} 1_{(0, \frac{1}{2}]} \mu^{Leb} + \frac{1}{2} 1_{(\frac{1}{2}, 1]} \mu^{Leb}$ , while  $\mu_2 = 2 \cdot 1_{(0, \frac{1}{2}]} \mu^{Leb}$ , so  $\mu_1 \ll \mu_2$  does not hold. But by Theorem 3.2,  $\lambda$  is a candidate of a Hastings algorithm.

If we change the position of the factor spaces in the example above,  $\lambda$  cannot be regarded as a candidate. For then,  $\lambda_d$  is the same,  $\lambda_o$  is uniformly distributed on  $(0, \frac{1}{2}] \times (0, 1]$ , thus  $\mu_1 = 2 \cdot 1_{(0, \frac{1}{2}]} \mu^{Leb}$ , and  $\mu_1 \times \mu_1 = 4 \cdot 1_{(0, \frac{1}{2}] \times (0, \frac{1}{2}]} \mu^{Leb} \times \mu^{Leb}$  has support  $(0, \frac{1}{2}] \times (0, \frac{1}{2}]$ . But  $\lambda_o$  has support  $(0, \frac{1}{2}] \times (0, 1]$  which is larger than that of  $\mu_1 \times \mu_1$ , so  $\lambda_o \ll \mu_1 \times \mu_1$  does not hold.

## References

- [1] W.R. Gilks, S. Richardson, and D.J. Spiegelhalter, editors. *Markov Chain Monte Carlo in Practice*. Chapman and Hall, London, 1996.
- [2] W.K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109, 1970.
- [3] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller. Equations of state calculations by fast computing machines. *J. Chemical Physics*, 21:1087–1091, 1953.
- [4] S. P. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability*. Springer-Verlag, London, 1993.
- [5] G.O. Roberts and R.L. Tweedie. Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika*, 83, 1996.
- [6] L. Tierney. A note on Metropolis-Hastings kernels for general state spaces. (submitted for publication).
- [7] L. Tierney. Markov chains for exploring posterior distributions (with discussion). *Ann. Statist.*, 22:1701–1762, 1994.