The Metropolitan-Hastings Algorithm and Extensions

S. Sawyer — Washington University — Vs. July 23, 2010

Table of Contents

1. Introduction. Suppose that we want to calculate the average value of a function $F(x)$ with respect to a probability density $\pi(x)$ on a set X, specifically

$$E(F) = \int_X F(x)\pi(x)dx \tag{1.1.1}$$

We assume here that X is endowed with an obvious, natural measure "dx". If dx is counting measure on X (that is, each point in X has mass one), then $E(F)$ is a sum instead of an integral.

We assume that we do not know the density $\pi(x)$ exactly, but that we can calculate $\pi(x)$ within the normalizing constant. That is,

$$\pi(x) = Cg(x) \tag{1.1.2}$$

where $g(x)$ is known or easy to compute but C is unknown. An important case is that of a Bayesian posterior distribution in statistics. In this case, $g(x)$ is a prior distribution times a likelihood function. In many cases, $g(x)$ is easy to write down but the normalizing constant

$$C = 1 \Big/ \int_X g(x)\,dx \tag{1.1.3}$$

is too complex to compute easily. (See Section 5 below.)

The "Metropolis" in the Metropolis-Hastings algorithm is the first author of a paper in the Journal of Chemical Physics (Metropolis *et al.* 1953; see bibliography). In Metropolis' paper, $g(x)$ is a partition function from statistical physics.

2. The Metropolis-Hastings algorithm.

2.1. The basic idea. The key idea in the paper Metropolis (1953) is first to start with a Markov chain X_n on the state space X with a fairly arbitrary Markov transition density $q(x, y)$. By definition, $q(x, y)$ is a Markov transition density if $q(x, y) \geq 0$ and $\int_{y\in X} q(x, y)dy = 1$, and X_n satisfies

$$\Pr(X_{n+1} \in y + dy \mid X_n = x) \;=\; q(x, y)dy \tag{2.1.1}$$

(We consider the case of a more general Markov transition function $q(x, A)$ below.) The second step is to modify the Markov chain X_n to define a new Markov chain X_n^* with a second transition density $p(x, y)$ that has $\pi(x)$ as a stationary probability measure. That is, such that

$$\int_X \pi(x)p(x, y)\,dx \;=\; \pi(y) \tag{2.1.2}$$

for all y. Then, by the Birkhoff ergodic theorem (see Appendix A)

$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} F(X_k^*) \quad \text{converges a.s.} \tag{2.1.3}$$

for any measurable function $F(x)$ on X with $\int_X |F(x)|\,\pi(x)dx < \infty$. The Markov chain X_n^* is called *ergodic* if the limit in (2.1.3) is constant whenever $\int_X |F(x)|\,\pi(x)dx < \infty$. In that case, one can show from (2.1.2) that

$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} F(X_k^*) \;=\; \int_X F(x)\pi(x)dx \quad \text{a.s.} \tag{2.1.4}$$

We can then, in principle, use the left-hand side of (2.1.4) for sufficiently large n to approximate the integral in (2.1.4). Sufficient conditions that X_n^* be ergodic are that X_n^* be irreducible and positive recurrent on X. The relations (2.1.2)–(2.1.4) form the basis of what are now known as Markov chain Monte Carlo (MCMC) methods.

We next show how to modify X_n with transition density $q(x, y)$ to form a Markov chain X_n^* with transtion density $p(x, y)$ that has $\pi(x)$ as a stationary density. The first step is to define an "acceptance function" $a(x, y)$, which by definition is an arbitrary function on $X \times X$ such that

$$0 \le a(x, y) \le 1 \tag{2.1.5}$$

We now define X_n^* as follows. Given $X_n^* = x$, we use the transition density $q(x, y)$ to "propose" a new state y. (That is, we choose $y \in X$ with the probability distribution (2.1.1).) With probability $a(x, y)$, the proposed state is accepted and $X_{n+1}^* = y$. Otherwise, the proposal is rejected and $X_{n+1}^* = X_n^* = x$. (That is, the process stays at the same position for one time step.)

Given $X_{n+1}^* = z$ (where $z = y$ if the proposal is accepted and $z = x$ if the proposal is rejected), the process is repeated to generate X_{n+2}^*, and so forth. This defines a new Markov chain X_n^* that is the same as X_n except for the introduction of "wait states" with $X_{n+1}^* = X_n^* = x$, which occur with probability $a(x, y)$ when $X_n^* = x$. By construction, the process X_n^* has "transition density"

$$p(x, y) \;=\; a(x, y)q(x, y) \;+\; A(x)\delta_x(dy) \tag{2.1.6}$$

where

$$A(x) \;=\; 1 \;-\; \int_{z\in X} a(x, z)q(x, z)dz$$

Here $A(x)$ is the probability that $z \in X$ is proposed but rejected, so that $X^*_{n+1} = X^*_n = x$. The expression $\delta_x(dy)$ in (2.1.6) represents a *Dirac measure* that puts mass one at the point $y = x$ and is otherwise zero. Strictly speaking, $p(x, y)$ in (2.1.6) does not represent a density with respect to dx if dx is a continuous measure on X and $A(x) > 0$. We give a better description of the algorithm below.

The final step is to find an acceptance function $a(x, y)$ so that (2.1.2) holds; that is, so that $\pi(x)$ is a stationary measure for X^*_n. Hastings (1970) proved (see below) that the transition density $p(x, y)$ in (2.1.6) has $\pi(x)$ as a stationary density if the acceptance function $a(x, y)$ satisfies

$$a(x,y) = \min\left\{1,\ \frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}\right\} = \min\left\{1,\ \frac{q(y,x)/\pi(x)}{q(x,y)/\pi(y)}\right\} \tag{2.1.7}$$

$$= \min\left\{1,\ \frac{g(y)q(y,x)}{g(x)q(x,y)}\right\} = \min\left\{1,\ \frac{q(y,x)/g(x)}{q(x,y)/g(y)}\right\}$$

where $\pi(x) = Cg(x)$ as in (1.1.2). In particular, we don't need to know the normalizing constant C in (2.1.7).

The original Metropolis (1953) algorithm assumed $q(x, y) = q(y, x)$, for which the acceptance function has the simpler form

$$a(x,y) = \min\left\{1,\ \frac{\pi(y)}{\pi(x)}\right\} = \min\left\{1,\ \frac{g(y)}{g(x)}\right\}$$

If $q(x, y) = q(y, x)$, the algorithm accepts the proposed new value y with probability one if $g(y) \geq g(x)$ and accepts it with probability $g(y)/g(x)$ if $g(y) < g(x)$. If the proposal density $q(x, y)$ is not symmetric (that is, if $q(x, y) \neq q(y, x)$), then the Metropolis-Hastings acceptance function $a(x, y)$ requires the correction factor $q(y, x)/q(x, y)$ in (2.1.7).

2.2. A question of notation. In Chapter 1 and Section 2.1, the implicit measure "dx" could be the usual volume measure (Lebesgue measure) if $X \subseteq R^d$, a counting measure that assigns measure one to each point if X is discrete, or surface measure in a subspace of R^d. In some descriptions of the Metropolis-Hastings algorithm, the precise meaning of the implicit measure dx is understood and can vary from paragraph to paragraph, and even from term to term in the same equation. This lack of a good notation can make proofs difficult to follow even when the proofs are basically correct.

For this reason, we will adopt a standard notation for measures from probability theory and write expressions like those in Section 1 in the form

$$E(F) = \int_X F(x)\pi(dx) \tag{2.2.1}$$

where $\pi(dx)$ is a probability measure on X and

$$\pi(dx) = Cg(dx) \quad \text{where} \quad C = 1 \Big/ \int_X g(dx) \tag{2.2.2}$$

where $g(dx)$ is a nonnegative measure on X with $\int_X g(dx) < \infty$. By definition, $\pi(dx)$ is a probability measure of X if it is a nonnegative measure with $\pi(X) = \int_X \pi(dx) = 1$. An expression $q(x, dy)$ is a *Markov transition function on* X if

(i) For each $x \in X$, $q(x, dy)$ (as a function of y) is a probability measure on X,

(ii) For each measurable set $A \subseteq X$, $c_A(x) = \int_A q(x, dy)$ is a measurable function of x.

It follows from (i,ii) that $\int_X h(y)q(x, dy)$ is a measurable function on X whenever $h(x)$ is a bounded measurable function on X.

Let X_n be a Markov chain X with

$$\Pr(X_{n+1} \in dy \mid X_n = x) \;=\; q(x, dy)$$

and let $a(x, y)$ be an acceptance function as in (2.1.5). Then the modified Markov chain X_n^* described in Section 2.1 has transition function

$$p(x, dy) \;=\; a(x, y)q(x, dy) + A(x)\delta_x(dy) \tag{2.2.3}$$

where

$$A(x) \;=\; 1 \;-\; \int_{z \in X} a(x, z)q(x, dz)$$

This now makes better sense since we do not have to pretend that $\delta_x(dy)$ is a function on X. Note that $p(x, dy)$ in (2.2.3) satisfies conditions (i,ii) above. (Technically speaking, we assume that all measures on X are σ-finite and second countable and defined on the same sigma algebra of subsets of X. This includes all measures that arise in practice.)

Similarly, we say that the transition function $p(x, dy)$ has $\pi(dx)$ as a stationary measure if

$$\int_X p(x, dy)\, \pi(dx) \;=\; \pi(dy) \tag{2.2.4}$$

considered as measure in y on X, or equivalently if

$$\int_X \int_X f(y)p(x, dy)\, \pi(dx) \;=\; \int_X f(y)\pi(dy) \tag{2.2.5}$$

for all measurable functions $f(y) \geq 0$ on X. The condition (2.2.5) is usually easier to work with than (2.2.4). Recall that $\int f(y)\pi(dy) = \infty$ is allowed in measure theory if $f(y) \geq 0$ and $\pi(dy) \geq 0$ is a nonnegative measure. In general, most theorems in measure theory for nonegative σ-finite measures are valid if the real numbers are extended in this way, so that there is no need to require that the integrals in (2.2.5) be finite.

(*Exercises:* (1) Prove that (2.2.5) for all measurable $f(y) \geq 0$ is equivalent to (2.2.4).

(2) Show that $p(x, dy)\pi(dx)$ in (2.2.4) defines a probability measure on the product space $X \times X$.)

If the measures $\pi(dx)$, $g(dx)$, and $q(x, y)$ have densities with respect to a measure dx on X, then

$$\pi(dx) = \pi_1(x)dx, \qquad g(dx) = g_1(x)dx, \qquad q(x, dy) = q_1(x, y)dy \tag{2.2.6}$$

where $\pi_1(x)$, $g_1(x)$, and $q_1(x, y)$ are measurable functions on X or $X \times X$. The larger family of transition functions $q(x, dy)$ includes not only examples of ergodic Markov chains on X that cannot be expressed as $q(x, y)dy$ for Lebesgue measure dy on $X \subseteq R^d$, but also Markov chains on linear or nonlinear subspaces of R^d whose transition function does not have a density with respect to d-dimensional Lebesgue measure. We will see examples of this in the next chapter.

2.3. Reversibility and the "detailed balance" condition. If $\pi(dx)$ is a probability measure and $p(x, dy)$ a Markov transition function on X, then

$$\iint_{X\times X} \pi(dx)p(x, dy) \;=\; \int_{x\in X} \left(\int_{y\in X} p(x, dy) \right) \pi(dx) \;=\; 1$$

by Fubini's theorem. Thus $\pi(dx)p(x, dy)$ is a probability measure on the product space $X \times X$. It follows that one can define random variables X_1 and Y_1 with joint distribution

$$P(X_1 \in dx, Y_1 \in dy) = \pi(dx)p(x, dy) \tag{2.3.1}$$

on $X \times X$. It follows from Fubini's theorem that $P(X_1 \in dx) = \int_y \pi(dx)p(x, dy) = \pi(dx)$ and hence $E\big(f(X_1)\big) = \int_X f(x)\pi(dx)$ for measurable functions $f(x) \geq 0$. Similarly $E\big(f(Y_1)\big) = \int_X f(x)\pi(dx)$ if the stationarity condition (2.2.4) holds. In general, (2.2.4) is equivalent to

$$E\big(f(X_1)\big) = E\big(f(Y_1)\big) \tag{2.3.2}$$

for measurable functions $f(y) \geq 0$.

In many Markov chains for which stationarity holds, the relation

$$\pi(dx)p(x, dy) \;=\; \pi(dy)p(y, dx) \tag{2.3.3}$$

also holds. Since then

$$\int_{x\in X} \pi(dx)p(x, dy) = \int_{x\in X} \pi(dy)p(y, dx) = \pi(dy)$$

it follows that (2.3.3) implies stationarity (2.2.4), although the reverse may not hold. The relation (2.3.3) is called the "detailed balance condition" by Chib and Greenberg (1995).

An equivalent form of (2.3.3) is

$$\int_X \int_X f(x, y)\pi(dx)p(x, dy) \;=\; \int_X \int_X f(x, y)\pi(dy)p(y, dx) \tag{2.3.4}$$

or

$$E\big(f(X, Y)\big) = E\big(f(Y, X)\big)$$

for measurable functions $f(x, y) \geq 0$ on the product space $X \times X$ for the random variables X, Y in (2.3.1).

(*Exercise:* Prove that (2.3.4) for all measurable $f(x, y) \geq 0$ is equivalent to (2.3.3), and that (2.3.3) implies (2.2.4).)

Since (2.3.1) can be viewed as the probability that X starts at x and then goes to $Y = y$, and the right-hand side of (2.3.3) can be viewed as the probability that X starts at y and then goes to $Y = x$, the relation (2.3.3) can be described as saying that the transition function $p(x, dy)$ is *time reversible* with respect to the probability measure $\pi(dx)$. (Usually time reversible is abbreviated to just *reversible*.)

A useful necessary and sufficient condition for (2.3.3) is the following.

Lemma 2.3.1 (Tierney 1994). Suppose that $p(x, dy)$ in (2.2.3) is derived from the transition function $q(x, dy)$, the acceptance function $a(x, y)$ with $0 \leq a(x, y) \leq 1$, and the probability measure $\pi(dx)$. In particular

$$p(x, dy) = a(x, y)q(x, dy) + A(x)\delta_x(dy), \tag{2.3.5}$$
$$A(x) = 1 - \int_{z\in X} a(x, z)q(x, dz)$$

Then the detailed balanced condition (2.3.3) holds if and only if

$$\pi(dx)a(x, y)q(x, dy) \;=\; \pi(dy)a(y, x)q(y, dx) \tag{2.3.6}$$

in the sense that the two measures on $X \times X$ are the same.

Corollary 2.3.1. If (2.3.6) holds, then $\pi(dx)$ is a stationary measure for $p(x, dy)$ in (2.2.5)

Proof of Lemma 2.3.1. By (2.3.5),

$$\pi(dx)p(x, dy) \;=\; \pi(dx)a(x, y)q(x, dy) + \pi(dx)A(x)\delta_x(dy) \tag{2.3.7}$$

and

$$\pi(dy)p(y, dx) \;=\; \pi(dy)a(y, x)q(y, dx) + \pi(dy)A(y)\delta_y(dx) \tag{2.3.8}$$

The two measures on the right-hand side of (2.3.7)–(2.3.8) are the same by Tonelli's theorem since, by arguing as in (2.3.4),

$$\int_x \int_y f(x, y)\delta_x(dy)\pi(dx)A(x) \;=\; \int_x f(x, x)A(x)\pi(dx)$$

and

$$\int_y \int_x f(x, y)\delta_y(dx)A(y)\pi(dy) \;=\; \int_x f(y, y)A(y)\pi(dy)$$

for measurable $f(x, y) \geq 0$. The two measures in the middle of the lines (2.3.7)–(2.3.8) are exactly the two measures in (2.3.6). If they are the same, the two measures in the detailed balance condition (2.3.3) are the same, and vice versa. This completes the proof of Lemma 2.3.1.

2.4. The Metropolis-Hastings theorem. The purpose here is to prove

Theorem 2.4.1. (Metropolis-Hastings) Assume that the probability measure $pi(dx)$ and transition function $q(x, dy)$ satisfy

$$q(x, dy) = q(x, y)dy \quad \text{and} \quad \pi(dx) = \pi(x)dx \tag{2.4.1}$$

for a nonnegative measure dx on X. Define $p(x, dy)$ by (2.3.5) for the acceptance function

$$a(x, y) = \min\left\{1,\; \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}\right\} \tag{2.4.2}$$

where the right-hand size of (2.4.2) is replaced by 1 if $\pi(x)q(x, y) = 0$. Then $p(x, dy)$ satisfies the detailed balance condition (2.3.3). In particular, $\pi(x)dx$ is a stationary measure for $p(x, dy)$.

Proof. By (2.4.2),

$$\begin{aligned} \pi(x)a(x,y)q(x,y) &= \min\{\,\pi(x)q(x,y),\ \pi(y)q(y,x)\,\} \\ &= \pi(y)a(y,x)q(y,x) \end{aligned} \tag{2.4.3}$$

since the right-hand side of the first equation in (2.4.3) is a symmetric function of x and y. This implies (2.3.6) in Lemma 2.3.1 and hence the reversibility condition (2.3.3).

Since all that is required for the reversibility condition (2.3.3) is

$$\pi(x)a(x,y)q(x,y) \;=\; \pi(y)a(y,x)q(y,x) \tag{2.4.4}$$

one can ask whether or not there exists a modification X_n^* of X_n as in Section 2.1 with a minimum number of wait states. The answer is yes, as the following Corollary indicates.

Corollary 2.4.1. The function $a(x,y)$ in (2.4.2) is the pointwise maximum value of all functions with $0 \le a(x,y) \le 1$ that satisfy the reversibility condition (2.3.3) for $p(x,dy)$ in (2.3.5). That is, among all acceptance functions $a(x,y)$ satisfying (2.4.4), the function $a(x,y)$ in (2.4.2) has the smallest probability of wait states at all $x \in X$.

Proof. Let Q be the set of all functions $b(x,y)$ such that $0 \le b(x,y) \le 1$ and

$$\pi(x)b(x,y)q(x,y) = \pi(y)b(y,x)q(y,x) \tag{2.4.5}$$

for all $x, y \in X$. Note that if $b_1(x,y)$ and $b_2(x,y)$ both satisfy (2.4.5), then so does $b_3(x,y) = \max\{b_1(x,y), b_2(x,y)\}$, and similarly so does $a(x,y) = \max_{b\in Q} b(x,y)$.

If $\pi(x)q(x,y) = 0$, then $a(x,y) = 1$ by (2.4.5), and (2.4.2) holds by definition. If $\pi(x)q(x,y) > 0$ and $0 \le \pi(y)q(y,x) \le \pi(x)q(x,y)$, then $a(y,x) = 1$ and $a(x,y) \le 1$ by (2.4.5). Then

$$a(x,y) \;=\; \frac{\pi(y)q(y,x)}{\pi(x)q(x,y)} \;<\; 1$$

by (2.4.5), which implies (2.4.2). If $0 < \pi(x)q(x,y) < \pi(y)q(y,x)$, then $a(x,y) = 1$ and the right-hand side of (2.4.2) is also 1. This completes the proof of Corollary 2.4.1.

2.5. Consequences of the Metropolis-Hastings theorem. It follows in general that

Theorem 2.5.1. *Let X_n^* be an ergodic Markov chain with stationary distribution $\pi(dx)$. Assume $F(x) \geq 0$. Then*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} F(X_k^*) \;=\; \int_X F(x)\,\pi(x)dx \quad \text{a.s.} \tag{2.5.1}$$

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} F(X_k^*)^2 \;=\; \int_X F(x)^2\,\pi(x)dx \quad \text{a.s.}$$

and for integers $m \geq 0$

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} F(X_k^*)F(X_{k+m}^*)$$
$$= \int_X F(x) \int_X F(y)\,\pi(y)p^m(x,dy)\,\pi(x)\,dx \quad \text{a.s.} \tag{2.5.2}$$

The function $p^m(x,dy) = \Pr(X_m \in dy \mid X_0 = x)$ in (2.5.2) is the m^{th} power of the transition density $p(x,dy)$ in (2.3.5).

The relations (2.5.1) imply that, in principle, it does not matter what proposal function $q(x,y)$ we use as long as we use the correct acceptance function for $\pi(x)$. The random variables $F(X_k^*)$ will have the same asymptotic mean and variance in all cases.

However, the asymptotic variance of the average $(1/n)\sum_{k=1}^{n} F(X_k^*)$ in (2.5.1) (as opposed to the asymptotic variance of $F(X_k^*)$) depends on the limiting autocovariances in (2.5.2). Proposal functions that minimize these autocovariances will lead to more accurate estimators of the right-hand side of (2.5.1) for finite n.

Similarly, different ergodic Markov chains X_n^* take longer to converge or "mix", so that the minimum n for which the left-hand side of (2.5.1) is a reasonable approximation of the integral side can vary even if the asymptotic covariances are the same. Thus, in practice, particular choices of the proposal functions $q(x,dy)$ do matter.

3. Important examples of MCMC Markov chains. Some particular forms of MCMC algorithms or proposal functions are used often enough to have special names:

3.1. Random walk sampling. In general if $q(x,y) = q(y,x)$ is a symmetric density, then the acceptance function (2.4.2) simplifies to

$$a(x,y) = \min\left\{1,\; \frac{\pi(y)}{\pi(x)}\right\} \tag{3.1.1}$$

This is Metropolis' (1953) original form of the algorithm. The usual definition of random walk for $q(x, y)$ in R^d assumes $q(x, y) = q(y - x)$, or equivalently that the proposal random variable Y satisfies

$$\{ Y \mid X = x \} \approx x + W \tag{3.1.2}$$

for a random variable W whose density is $f_W(y) = q(y)$. Metropolis' condition $q(x, y) = q(y, x)$ is equivalent to $f_W(x) = f_W(-x)$, which is equivalent to $W \approx -W$ or that the distribution of W be symmetric about 0.

3.1.1. Step sizes for random-walk sampling. While the behavior of the Markov chain generated by $f_W(x)$ and $\pi(x)$ is less sensitive to the shape of the distribution $f_W(x)$, it is sensitive to the size of the steps. Usually (3.1.2) is written in the form

$$\{ Y \mid X = x \} \approx x + hW \tag{3.1.3}$$

and then the value of $h > 0$ is chosen (or "tuned") so that the Markov chain has good convergence properties. While Theorem 2.5.1 says that the chain will eventually approximate the stationary distribution $\pi(x)$ for any $h > 0$, this may take a very long time if the value of h is chosen inappropriately.

If h in (3.1.3) is too large, then proposals will rarely be accepted and the Markov chain in (2.5.1) will rarely move, and the ratios in (2.5.1) will take a long time to converge. If h is too small, then two problems can arise. First, the chain X_n will take only small steps and may take a long time before it can cover a significant portion of X, let alone before its sample statistics approximate $\pi(x)$. A second, less obvious, problem is that proposals may be accepted too often. This means that the process X_n may be more determined in the short run by the proposal distribution $q(x, y)$ (and its stationary distribution) then by $\pi(x)$, which is the stationary distribution of $p(x, y)$. Again, the chain X_n^* may take a long time before its sample statistics approximate $\pi(x)$.

As a rule of thumb, conventional wisdom is to adjust h empirically so that the acceptance ratio (which is one minus the proportion of wait states) is in the range 30–50%, with 30-40% tending to work better than 40-50%. Metropolis random walks with a smaller acceptance ratio (generally due to a larger value of h) are said to take "fewer but higher-quality steps".

3.1.2. Random walks with x-dependent step size. It can happen that values of h in (3.1.3) are too large in certain parts of the domain, where perhaps $\pi(x)$ is highly variable, and too small in other parts of the domain, where $\pi(x)$ is relatively flat. The latter problem can happen if, for example, the density $\pi(x) \sim C/x^a$ for large x. If $X_n = x$ for large x, then $\pi(x)$ is

essentially constant in the range $x_1 = x \pm h$, and nearly all proposals are accepted. The result is similar to an unbiased coin-tossing random walk on the tail of the distribution. In general, if you continually toss an unbiased coin, subtract one head for each tail received, and start with $X_0 = 20$ (for example), then eventually you will have $X_n = 0$, but the number of tosses until this happens has an infinite mean and a very heavy-tailed distribution. For a Metropolis random walk, this can show up as rare long excursions to improbably large values of x if h is too small on the tail of the distribution.

In either case, a possible solution is to have the step size depend on x, either in a regular way or else as different values in different regions in X. In general, define the proposal $q(x, y)$ as the distribution of

$$\{\, Y \mid X = x \,\} \;\approx\; x + h_x W \tag{3.1.4}$$

where W is a random variable with density $f_W(x)$. Then for any measurable function $\phi(y) \geq 0$

$$\begin{aligned} E_x\big(\phi(Y)\big) \;&=\; \int_X \phi(x + h_x y) f_W(y)\, dy \;=\; \left(\frac{1}{h_x}\right)^d \int_X \phi(x+y) f_W\left(\frac{y}{h_x}\right) dy \\ &=\; \left(\frac{1}{h_x}\right)^d \int_X \phi(y) f_W\left(\frac{y-x}{h_x}\right) dy \end{aligned}$$

This means that the proposal density $q(x, y)$ for (3.1.4) is

$$q(x,y) = \left(\frac{1}{h_x}\right)^d f_W\left(\frac{y-x}{h_x}\right) \tag{3.1.5}$$

with respect to dy. The acceptance function is then

$$a(x,y) = \min\left\{1,\; \frac{\pi(y)}{\pi(x)} \left(\frac{h_x}{h_y}\right)^d \frac{f_W\big((x-y)/h_y\big)}{f_W\big((y-x)/h_x\big)}\right\} \tag{3.1.6}$$

by Theorem 2.4.1. If W is symmetric ($f_W(x) = f_W(-x)$), then

$$a(x,y) = \min\left\{1,\; \frac{\pi(y)}{\pi(x)} \left(\frac{h_x}{h_y}\right)^d\right\}$$

As a check, note that this reduces to (3.1.1) if $h_x = h_y$.

For example, suppose that W in (3.1.4) is multivariate normal with mean zero and positive definite covariance matrix Σ. Then

$$f_W(x) = \frac{1}{\sqrt{(2\pi)^d \det(\Sigma|)}} \exp\big(-(1/2) x' \Sigma^{-1} x\big) \tag{3.1.7}$$

Since Σ is symmetric and positive definite, $\Sigma = RDR'$ where R is an orthogonal matrix and D is diagonal with positive entries. A way of simulating the random variable W is to set

$$W = RD^{1/2}Z \tag{3.1.8}$$

where $Z = (Z_1, Z_2, \ldots, Z_d)$ for independent mean-one variance-one normal random variables Z_i. Then

$$\mathrm{Cov}(W) = \big(RD^{1/2}\big)\,\mathrm{Cov}(Z)\big(RD^{1/2}\big)' = RDR' = \Sigma$$

so that W in (3.1.8) has the correct distribution. This means that Y in (3.1.4) can be simulated by

$$Y \;=\; X + h_X W \;=\; X + h_X RD^{1/2}Z \tag{3.1.9}$$

By (3.1.7)

$$\begin{aligned}\frac{f_W\big((x-y)/h_y\big)}{f_W\big((y-x)/h_x\big)} &= \exp\left(-\frac{1}{2}A(x-y)\left(\frac{1}{h_y^2}-\frac{1}{h_x^2}\right)\right) \\ &= \exp\left(-\frac{1}{2h_x^2}A(x-y)\left(\left(\frac{h_x^2}{h_y^2}\right)-1\right)\right)\end{aligned} \tag{3.1.10}$$

where $A(x) = x'\Sigma^{-1}x$. By (3.1.9)

$$\begin{aligned}A(X-Y) = A(h_X RD^{1/2}Z) &= h_X^2\big(RD^{1/2}Z\big)'\Sigma^{-1}\big(RD^{1/2}Z\big) \\ &= h_X^2 Z'D^{1/2}R'\Sigma^{-1}RD^{1/2}Z = h_X^2 Z'Z\end{aligned}$$

Thus, ignoring the minimum in (3.1.6), the logarithm of the acceptance function can be written

$$\log a(X,Y) = \log\left(\frac{\pi(Y)}{\pi(X)}\right) + d\log\left(\frac{h_X}{h_Y}\right) - \frac{1}{2}Z'Z\left(\left(\frac{h_X}{h_Y}\right)^2 - 1\right)$$

for the vector Z in (3.1.9).

3.1.3. Multiplicative random walks. It is sometimes useful to have multiplicative updates

$$\{\, Y \mid X = x \,\} \;\approx\; xW \tag{3.1.11}$$

instead of additive updates (3.1.2). This allows proposed values Y to be larger when X is larger and smaller when X is smaller. It follows from (3.1.11) that

$$E\big(\phi(Y) \mid X = x\big) = E\big(\phi(xW)\big) = \int \phi(xy) f_W(y)\, dy$$
$$= \int \phi(y) q(x,y)\, dy = (1/x) \int \phi(y) f_W(y/x)\, dy$$

for $\phi(y) \geq 0$, where $f_W(y)$ is the density for W. Thus the proposal density is

$$q(x,y) = \frac{1}{x}\, f_W\Big(\frac{y}{x}\Big)$$

The proposal function $q(x,y)$ is said to define a *symmetric multiplicative random walk* if $W \approx 1/W$, which is equivalent to $f_W(y) = (1/y)^2 f_W(1/y)$. (*Exercise*: Prove this.) In that case

$$\frac{\pi(y)q(y,x)}{\pi(x)q(x,y)} = \frac{\pi(y)(1/y)f_W(x/y)}{\pi(x)(1/x)f_W(y/x)} = \frac{\pi(y)(1/y)(y/x)^2 f_W(y/x)}{\pi(x)(1/x)f_W(y/x)} = \frac{\pi(y)y}{\pi(x)x}$$

Thus the acceptance function for an arbitrary symmetric multiplicative random walk is

$$a(x,y) = \min\left\{1,\ \frac{\pi(y)\, y}{\pi(x)\, x}\right\} \tag{3.1.12}$$

3.2. Independence sampling. If the value proposed for X_{n+1} is *independent* of X_n, that is, if $q(x,y) = q(y)$, then the MH algorithm is called an *independence sampler*. The acceptance function becomes

$$a(x,y) \ = \ \min\left\{1,\ \frac{\pi(y)q(x)}{\pi(x)q(y)}\right\} \ = \ \min\left\{1,\ \frac{\pi(y)/q(y)}{\pi(x)/q(x)}\right\} \tag{3.2.1}$$

That is, given $X_n = x$ and $k(y) = \pi(y)/q(y)$, the proposed value y is accepted if $k(y) \geq k(x)$ and accepted with probability $k(y)/k(x)$ if $k(y) < k(x)$. Again, we do not need to know the normalizing constants of either $\pi(y)$ or $q(y)$.

It is very important for independence samplers that $q(y)$ not be lighter-tailed than $\pi(y)$ for large values of y if X is noncompact, nor than $q(y)$ be lighter tailed at a singularity of $\pi(y)$. That is, we should NOT have $k(y) = \pi(y)/q(y) \gg 1$ either for large y or at a singularity of $\pi(y)$. In that case, large values of Y (as measured by $\pi(y)$) are rarely proposed and it is easy to find examples in which the independence sampler has truly horrible convergence properties. That is, while

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} F(X_k^*) \ = \ \int_X F(x)\pi(x)dx \quad \text{a.s.} \tag{3.2.2}$$

the left-hand side of (3.2.2) is a reasonable approximation of $\int F(x)\pi(x)dx$ only for extremely large values of n. Even worse, the averages in (3.2.2) may appear to converge but to an incorrect value, even for $n \approx 10^6$ or $n \approx 10^7$

One way to make this less likely for heavy-tailed $\pi(y)$ for large y is to use a proposal distribtion $q(y)$ that has a power law ($q(y) = \max\{1, 1/y^a\}$) or a Student-t distribution, but one would have to make sure that $\pi(y)$ is not more heavy-tailed yet.

If we know how to generate random variables X_{n+1} whose distribution is exactly equal to $\pi(x) = Cg(x)$ and set $q(x) = \pi(x)$, then $a(x, y) = 1$. In this case, there are no wait states and the independence sampler is the same as classical Monte Carlo sampling.

3.2.1. Von Neumann's rejection sampling. The independence sampler is similar in spirit to the "rejection method" of von Neumann (1951), which can be used to generate random variables with an arbitrary distribution $\pi(x)$. Von Neumann's rejection method assumes

$$\pi(x) = cA(x)q(x) \tag{3.2.3}$$

where $0 \le A(x) \le 1$ and that we know how to generate random variables with distribution $q(x)$. The algorithm is to sample values Y from the proposal distribution $q(y)$ and accept Y with probability $A(y)$. If the value is rejected, the algorithm continues to sample from $q(y)$ until a value is accepted. The final accepted value of Y has distribution that is exactly $\pi(x)$. (*Exercise:* Prove this.)

In contrast with von Neumann's method, the independence sampler does not have retrials. If a value is rejected, the previous value is used. The Markov chain values X_n have $\pi(dx)$ as a stationary distribution, but in general do not have the distribution $\pi(dx)$ themselves.

If we know how to generate random variables X_{n+1} with distribution exactly equal to $\pi(x) = q(x)$, then we can take $A(x) = 1$. In this case, there are no rejections and the method is again the same as classical Monte Carlo sampling.

3.3. Random mixtures of update proposals. A *random mixture* of proposals $q_a(x, dy)$ with *mixture weights* $w_a > 0$ is defined by, first, choosing an index a with probability w_a, and second, choosing a proposed value $y \in X$ with probability $q_a(x, dy)$. Random mixture updates on the same set of coordinates could be useful if no one proposal function has good properties for all of x. Choosing the proposal based on $x \in X$ as in Section 3.1.2 might be more efficient in this case, but the random mixture model would have a simpler acceptance function and would be easier to implement.

Given a random mixture, we have two possible ways of defining the acceptance function and the combined Metropolis-Hastings Markov chain $p(x, dy)$. The first (and generally simplest) way is to forget that the index a came from a random choice and use the acceptance function

$$a_a(x, y) = \min\left\{1,\ \frac{\pi(y)q_a(y, x)}{\pi(x)q_a(x, y)}\right\} \tag{3.3.1}$$

to define the chain $p_a(x, dy)$ as in Theorem 2.4.1. Then

$$\pi(dx)p_a(x, dy) \ = \ \pi(dy)p_a(y, dx) \tag{3.3.2}$$

for $p_a(x, dy)$ defined by

$$p_a(x, dy) = a_a(x, y)q_a(x, dy) + A_a(x)\delta_x(dy),$$
$$A_a(x) = 1 - \int_{z\in X} a_a(x, z)q_a(x, dz)$$

Since a was chosen randomly with probability w_a, the unmixed (or marginal) Metropolis-Hastings Markov chain has transition function

$$p(x, dy) = \sum_a w_a p_a(x, dy)$$

Thus by linearity $p(x, dy)$ also satisfies the detailed balance condition

$$\pi(dx)p(x, dy) \ = \ \pi(dy)p(y, dx) \tag{3.3.3}$$

and $\pi(x)dx$ is a stationary measure for $p(x, dy)$.

The second method is to use the combined proposal function

$$q(x, dy) = \sum_a w_a q_a(x, dy)$$

with the corresponding acceptance function

$$a(x, y) = \min\left\{1,\ \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}\right\} \tag{3.3.4}$$

This also leads to the detailed balance condition (3.3.3) and hence $p(x, dy)$ also preserves $\pi(x)dx$. Due to the nonlinearity of the acceptance functions (3.3.1) and (3.3.4), there is no simple relationship between the update transition functions $p_a(x, dy)$ in (3.3.2) and $p(x, dy)$ from (3.3.4).

3.4. Metropolis-Hastings and importance sampling. A useful technique to improve the efficiency of any Monte Carlo technique is the following. Suppose in the limiting approximation (2.5.1) in Theorem 2.5.1

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} F(X_k^*) \;=\; \int_X F(x)\pi(x)\,dx \tag{3.4.1}$$

that $F(x)$ is small on a portion of the state space X that has significant weight under $\pi(x)$.

Then the Markov chain X_n (with stationary distribution $\pi(x)dx$) may spend most of its time in parts of the state space X that contribute little to the sum and hence little to the estimation of the integral. In general, *importance sampling* refers to changing any Monte Carlo method so that, first, the sampling is biased towards parts of the state space that are more important for the sum or integral being estimated and, second, corrections are made for the resulting biased sampling.

Let $H(x) \geq 0$ be a function on X that is small where $F(x)$ is small and relatively large where $F(x)$ is large. Of course, $H(x) = F(x)$ satisfies this criterion, but we assume that $H(x)$ easier to work with than $F(x)$. Let X_n be the Markov chain defined in Section 2 with transition density $q(x, y)$, and let Y_n^* be the corresponding Metropolis-Hastings Markov chain (see Theorem 2.4.1) with the acceptance function defined with $H(x)g(x)$ (or $H(x)\pi(x)$) instead of $g(x)$ or $\pi(x)$. (Recall that $\int \pi(x)dx = 1$ but that $g(x)$ is unnormalized.) Specifically, we use the acceptance function

$$a(x, y) = \min\left\{1,\; \frac{H(y)g(y)q(y, x)}{H(x)g(x)q(x, y)}\right\} \tag{3.4.2}$$

instead of the same acceptance function with $H(x) = 1$. If $q(x, y)$ corresponds to random-walk or independence sampling, then (3.4.2) is of the same form with $H(x)g(x)$ in place of $g(x)$. Then Theorem 2.4.1 implies that Y_n^* has the stationary probability density

$$g_H(x) = C_H\, H(x)g(x) \qquad \text{where} \qquad \int_X g_H(x)\,dx = 1 \tag{3.4.3}$$

instead of $\pi(x)$. Since Y_n^* does not have $g(x)\,dx$ as a stationary distribution, it cannot be used to estimate integrals of the form $\int F(x)\pi(x)\,dx$ directly, but by the ergodic theorem for Y_n^* we do have

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \frac{F(Y_k^*)}{H(Y_k^*)} \;=\; \int_X \frac{F(x)}{H(x)} g_H(x)\,dx \tag{3.4.4}$$

$$=\; C_H \int_X F(x)g(x)\,dx \;=\; \frac{\int_X F(x)g(x)\,dx}{\int_X H(x)g(x)\,dx} \tag{3.4.5}$$

since $C_H = 1 \,/\, \int H(x)g(x)dx$. The ratios $F(Y_n^*)/H(Y_n^*)$ in (3.4.4) should spend less time in the parts of X where $F(x)$ is small since Y_n^* has stationary distribution $c_H H(x)g(x)dx$ instead of $\pi(x)dx$. Similarly, since $H(x)$ resembles $F(x)$ more than a constant function, the terms in the sum in (3.4.4) should have smaller variance than the corresponding terms in (3.4.1).

If the denominator of (3.4.5) is known or is easy to calculate, this should give a more efficient way to estimate $\int_X F(x)\pi(x)\,dx$.

3.5. Component-wise Metropolis-Hastings sampling. In many typical MCMC applications, the vectors $x \in X$ and density $\pi(x)$ are multidimensional and often highly multidimensional. However, it is usually possible to find Markov chain update X_n^* where each update of X_n^* consists of several substeps, each of which updates one component or one group of components of x.

If $x = (x_1, \ldots, x_d)$ and we have d potential updates, where the i^{th} update changes only the i^{th} coordinate of x and is ergodic in the sense of Section 2.1 on the line through x parallel to the i^{th} coordinate axis, then the overall Metropolis-Hastings Markov chain X_n will be ergodic on X. (See Theorem A.3 in Appendix A.) This means that we can use long-term averages of X_n^* to estimate components of X_n^* or functions of components of X_n^* as in Section 2.5.

We call changing or updating one coordinate or group of the coordinates of $x \in X$ a *component move*. By *components-in-sequence* MCMC we mean updating a Markov chain X_n on X by a sequence of substeps that each update one or more coordinates of X.

Assume for definiteness that $x, X_n \in X \subseteq R^d$ and that we have d one-dimensional proposal functions of the form

$$q_i(x_i, y) = q_i(x_i, x_{-i}, y), \quad x \in R^d,\ x_i, y \in R,\ x_{-i} \in R^{d-1} \tag{3.5.1}$$

Here $x_{-i} \in R^{d-1}$ are the components of $x \in R^d$ other than the i^{th} component x_i, which are viewed in (3.5.1) as parameters of the one-dimensional proposal density $q_i(x_i, x_{-i}, y)$.

Given (3.5.1), we update the Markov chain X_n in such a way that each full step of X_n (that is, $X_n \to X_{n+1}$) consists of d consecutive one-dimensional substeps in turn. The acceptance function for the i^{th} substep is

$$a_i(x_i, x_{-i}, y_i) = \min\left\{1,\ \frac{\pi_i(y_i \mid x_{-i})\, q_i(y_i, x_{-i}, x_i)}{\pi_i(x_i \mid x_{-i})\, q_i(x_i, x_{-i}, y_i)}\right\} \tag{3.5.2}$$

where

$$\pi_i(x_i \mid x_{-i}) = \pi(x \mid x_{-i}) = \frac{\pi(x_i, x_{-i})}{\int \pi(z, x_{-i})dz}$$

is the conditional distribution of x_i given x_{-i}. The denominators of $\pi_i(y_i \mid x_{-i})$ and $\pi_i(x_i \mid x_{-i})$ in (3.5.2) cancel out, so that in fact

$$a_i(x_i, x_{-i}, y_i) = \min\left\{1, \frac{\pi(y_i, x_{-i})\, q(y_i, x_{-i}, x_i)}{\pi(x_i, x_{-i})\, q(x_i, x_{-i}, y_i)}\right\} \tag{3.5.3}$$

By convention, $\pi(y \mid x_{-i}) = q_i(x, x_{-i}, y) = 0$ unless $y_j = x_j$ for $j \neq i$, so that the extended measures $\pi(dx \mid x_{-i})$ and $q_i(x, dy)$ on R^d are concentrated on the line through x parallel to the i^{th} coordinate axis.

The corresponding one-dimensional Metropolis-Hastings transition measure

$$p_i(x_i, dy) = a(x_i, y_i)q_i(x_i, y)dy + A_i(x)\delta_{x_i}(dy) \quad \text{where} \tag{3.5.4}$$
$$A(x_i) = 1 - \int_{z \in X} a(x_i, z)q_i(x_i, z)dz, \quad x_i, y, y_i \in R^1$$

(with x_{-i} suppressed) has $\pi_i(x_i \mid x_{-i})$ as a stationary measure on R^1. Each $p_i(x_i, x_{-i}, dy)$ on R^1 extends to a transition measure $p_i(x, dy)$ on X that has $\pi(x)$ as a stationary measure (see Appendix A). Since each of the d transition measures $p_i(x, dy)$ preserve $\pi(x)$, so does the Markov chain X_n. If each of the proposal functions (3.5.4) is ergodic in one dimension in the sense of Section 2.1 for each fixed x_{-i}, then X_n is also ergodic and can be used to estimate components of $\pi(x)$ (see Theorem A.3 in Appendix A).

There is no reason why each of the component substeps (3.5.4) that make up a full step in X_n has to be one dimensional, although that is the most common case. If coordinates of $\pi(x)$ are highly correlated, it may be more advantageous to update groups of coordinates simultaneously, or overlapping groups of components. See Chapter 3 for some examples.

In the more usual case of d one-dimensional substeps in sequence, we can use *oversampling* for any components of x for which the trajectories of the Markov chain are unusually autocorrelated, have an unusually low acceptance rate, or perhaps are just less expensive to compute in terms of computation time. By oversampling, we mean that proposals (3.5.1) with the acceptance functions (3.5.2) are repeated a fixed number of times in each Markov chain iteration (for example, 5 or 10 or 100 times) and only the last value is used. Oversampling can improve the efficiency of Metropolis and Metropolis-Hastings updates.

3.5.1. Random component updates. There is no requirement that the component steps in the MH algorithm should always be done in the same sequence. We could carry out the updates (3.5.4) in a random order, so that each update of X_n updates a single randomly-chosen component of x. This

is a special case of a random mixture of updates in Section 3.3, which we can call *random-component-update* MCMC (see Appendix A below).

In this case, the updates do not satisfy the assumptions of the Metropolis-Hastings theorem (Section 2.4), since the proposal density

$$q(x, dy) = \sum_a w_a q_a(x, dy)$$

is concentrated on a one-dimensional subset of R^d, specifically the union of the coordinate axes through x and so does not have a density with respect to d-dimensional Lebesgue measure. However, the detailed-balance condition still holds (Section 3.3).

3.6. Gibbs samplers. In many cases, the individual conditional distributions $\pi(x_i \mid x_{-i})$ are easy to describe even when the structure of $\pi(x)$ is complex.

If we know how to generate random variables with the conditional distribution $\pi(y_i \mid x_{-i})$, then we can use

$$q_i(x_i, y_i) = q_i(x_i, x_{-i}, y_i) = \pi(y_i \mid x_{-i}) \tag{3.6.1}$$

as a proposal function in (3.5.1). Given x_{-i}, this is equivalent to independence sampling from the conditional distribution of x_i. Since then $a_i(x_{-i}, x_i, y_i) = 1$ in (3.5.2), there are no wait states or rejection steps. This is called a *Gibbs sampler* or a *Gibbs sampler step* for $\pi(x \mid x_{-i})$. If $d = 1$, there is no dependence on x_{-i} and this is the same as classical Monte Carlo sampling.

(*Exercise*: Explain carefully why there is no benefit for oversampling a Gibbs sampler substep.)

In practice, Gibbs sampler updates tend to be much more efficient than other Metropolis or Metropolis-Hastings updates. The rate of convergence of a components-in-sequence or random-component-update MCMC is often a function of the proportion of updates that are Gibbs samplers.

3.7. Knowing when to stop. Often MH trajectories that start in different parts of the parameter space may stay diverged for a very long period of time. One way of measuring the extent to which this may happen is the Gelman scale reduction factor (Gelman *et al.* 2003, Gilks *et al.* 1996, specifically Section 8.4 in Chapter 8, p136–139).

In some cases, computation of the function values $f(X_n)$ is more expensive than generating the values X_n, or we want to assume that the sampled

values $f(X_n)$ are approximately independent, or we want to store fewer values for later estimation of medians and credible intervals. In these situations, we can sample values $f(X_n)$ only every K^{th} step. This is equivalent to oversampling the full Markov chain X_n. In this case, we distinguish between samples or sampled values and the steps or iterations of the MC Markov chain.

For definiteness, suppose that we run J trajectories each of (iteration) length nK, so that each trajectory has n samples. We assume $1 < J \ll n$ and $K \geq 1$. These could be either different runs with different starting positions or else consecutive blocks of values in a single run of length nKJ iterations.

For a particular parameter of interest, let ψ_{ij} be the $(iK)^{\text{th}}$ iteration (or the i^{th} sampled value) in the j^{th} trajectory, where $1 \leq j \leq J$ for separate runs and $1 \leq i \leq n$ within each trajectory. (That is, $\psi_{ij} = F(X_{iK,j})$ for a function $F(x)$ and J copies $X_j(i) = X_{i,j}$ of the MH process in Section 2.) Let

$$a_j = \frac{1}{n}\sum_{i=1}^{n} \psi_{ij} = \overline{\psi}_{+j}, \qquad 1 \leq j \leq J$$

be the sample mean of the j^{th} trajectory. Let $b_j = E(\psi_{ij})$ be the corresponding theoretical mean. Then

$$B = \frac{n}{J-1}\sum_{j=1}^{J}(a_j - \overline{a})^2 \qquad \text{for} \qquad \overline{a} = \frac{1}{J}\sum_{j=1}^{J} a_j$$

is the numerator of the one-way ANOVA F-test of $H_0 : b_j = b_0$, which we can call $B = \text{MSMod}$. In particular

$$E(B) \;=\; \sigma^2 \;+\; \frac{n}{J-1}\sum_{j=1}^{J}(b_j - \overline{b})^2 \qquad \text{for} \qquad \overline{b} = \frac{1}{J}\sum_{j=1}^{J} b_j$$

if each set of sampled values ψ_{ij} $(1 \leq i \leq n)$ are approximately uncorrelated with variance σ^2. We call B the between-sequence variance. Similarly

$$W = \frac{1}{J}\sum_{j=1}^{J} s_j^2 \qquad \text{for} \quad s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\psi_{ij} - a_i)^2$$

is the denominator MSE of the same one-way ANOVA test. It satisfies $E(W) = \sigma^2$ if each trajectory ψ_{ij} is uncorrelated for fixed j. This is called the within-sequence variance. Then

$$\frac{E(B) - E(W)}{n} \;=\; \frac{1}{J-1}\sum_{j=1}^{J}(b_j - \overline{b})^2 \;=\; s_b^2 \tag{3.7.1}$$

is the sample variance of the J theoretical trajectory means b_j. Gelman's scale-reduction factor is

$$\widehat{GR} = \frac{W + \frac{B-W}{n}}{W} = 1 + \frac{(B-W)}{nW} = 1 + \frac{(F-1)}{n} \qquad (3.7.2)$$
$$= \frac{n-1}{n} + \frac{F}{n}$$

where $F = \text{MSMod} / \text{MSE}$ is the one-way ANOVA test statistic. This can be viewed an estimator of $(\sigma^2 + s_b^2)/\sigma^2 = 1 + s_b^2/\sigma^2$ in (3.7.1).

Note that one can have $\widehat{GR} < 1$ or $B - W < 0$ due to sampling variation. In fact, if the MH Markov chains converges very strongly, then $\psi_{i,j}$ for $i = 1, 2, \ldots$ will be negatively correlated in i due to their being bound to this stationary distribution. In this case, B can understate the variance σ^2 resulting in $B < W$.

If $\widehat{GR}$ is close to one, this may be an indication that all J trajectories are sampling from the stationary distribution in an unbiased manner. Gelman suggests, "In practice, we generally run the simulations until the values of $[GR]$ are all less than 1.1 or 1.2" (Gelman 1996, p138). Of course, the same calculation must be done for every sampled function $f(x)$ of interest, since some parameters in an MCMC simulation may converge much faster than others.

A related measure is

$$R^2 = \frac{n \sum_{j=1}^{J} (a_j - \overline{a})^2}{\sum_{j=1}^{J} \sum_{i=1}^{n} (\psi_{ij} - \overline{a})^2} = \frac{(J-1)B}{J(n-1)W + (J-1)B} \qquad (3.7.3)$$

For statistical regressions, this is called the "proportion of the total variability of the ψ_{ij}" that is "explained" by the subchain or trajectory means. Typically R^2 is small if and only if $\widehat{GR}$ is close to one. Often $\widehat{GR} < 1$ and R^2 is tiny when an MCMC Markov chain converges strongly.

4. Skew Transformations.

4.1. Introduction. In some cases the components of x are highly correlated and $\pi(dx)$ is *stiff* (that is, rapidly changing) as a function of x. In that case, the Metropolis-Hastings algorithm is likely to accept only small changes in individual components of x. This can cause the MC Markov chain to take an extremely long time to converge.

If this happens, one way of improving convergence is to use a proposal function that is, for example, jointly normal in x with a covariance matrix that is estimated from a preliminary run. Another way (which is the point of this section) is to update key components of x and then make an parallel deterministic changes in other components in an attempt to preserve the values of $\pi(x)$.

For definiteness, assume $x = (x_1, x_2)$ where $x_1 \in R^d$, $x_2 \in R^m$, and $x \in R^n$ for $n = d + m$. Consider a Markov transition proposal function $q(x, dy)$ defined in terms of random variables by

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \to \begin{pmatrix} Y_1 \\ h(X_1, X_2, Y_1) \end{pmatrix} \tag{4.1.1}$$

where $h(x_1, x_2, y_1)$ is a deterministic function of x_1, x_2, and y_1. We assume that the random motion $X_1 \to Y_1$ depends only on X_1 and not on X_2. Specifically,

$$P\big(Y_1 \in dy_1 \;\big|\; X_0 = (x_1, x_2)\big) = q(x_1, y_1)dy_1$$

where $q(x_1, y_1)$ is a d-dimensional density. We call (4.1.1) a *skew transformation* of $X \in R^n$, since a change from X_1 to Y_1 in R^d is accompanied by a deterministic change in X_2 that is a function of Y_1.

The Markov chain block updates that we have considered so far have been fairly simple: That is, the proposal distribution maps a subinterval of a one-dimensional subspace of R^n into the same subinterval with the remaining coordinates viewed as a parameter. In contrast, (4.1.1) maps a d-dimensional subspace of R^n into a possibly different d-dimensional submanifold of R^n.

The mapping (4.1.1) is very similar to a transformation that changes only X_1 in a different coordinate system, and then changes coordinates back to the original coordinates. These *reparametrization* updates are a subset of the skew transformations (4.1.1) (see Section 4.3 below). While it is not known whether all transformations (4.1.1) can be obtained by reparametrization in this manner, reparametrization updates are often the easiest to find and work with.

The cases $d = 0$ and $m = 0$ are not excluded in (4.1.1): The former is a purely deterministic move in R^n and the latter a mapping of full rank in R^n. Note that the transformations of X_1 and X_2 in (4.1.1) cannot in general be done independently or in sequence: Update proposals must be Markov, X_1 would not be available after updating X_1, and Y_1 is not available until X_1 is updated.

For smooth functions $h(x_1, x_2, y_1)$ in (4.1.1), we now find general sufficient conditions for the existence of an acceptance function $a(x, y)$ such that

the corresponding update transition function $p(x, dy)$ satisfies the detailed balance condition (2.3.3).

Let $J_2h(x, y) = J_2h(x_1, x_2, y_1)$ be the absolute value of the $m \times m$ Jacobian matrix of $h(x_1, x_2, y_1)$ with respect to x_2 and let $\pi(x)$ be a nonnegative integrable function on R^n. Then

Theorem 4.1.1. Let $p(x, dy)$ be the transition function of the Markov process corresponding to the proposal distribution $q(x, y)$ defined by $q(x_1, y_1)$ and (4.1.1) and the acceptance function

$$a(x, y) = \min\left\{1,\ \frac{\pi(y)q(y_1, x_1)}{\pi(x)q(x_1, y_1)} J_2h(x_1, x_2, y_1)\right\} \tag{4.1.2}$$

Then a sufficient condition for $p(x, dy)$ to satisfy the detailed balance condition (2.3.3) in Section 2.3, or equivalently

$$\pi(x)dx\, p(x, dy) \ = \ \pi(y)dy\, p(y, dx) \tag{4.1.3}$$

is that, for $(x_1, x_2, y_1, y_2) \in R^{2n}$,

$$y_2 = h(x_1, x_2, y_1) \quad \text{if and only if} \quad x_2 = h(y_1, y_2, x_1) \tag{4.1.4}$$

We will prove Theorem 4.1.1 in Section 4.2. A useful sufficient condition for (4.1.4) is

Corollary 4.1.1. A sufficient condition for (4.1.4) is that there exists a function $R(x) = R(x_1, x_2)$ on R^n such that

$$y_2 = h(x_1, x_2, y_1) \quad \text{if and only if} \quad R(x) = R(y) \tag{4.1.5}$$

Exercise: Prove Corollary 4.1.1 given Theorem 4.1.1.

We will see in the next section that the class (4.1.4)–(4.1.5) contains a large class of updates defined by temporary reparametrization.

Given (4.1.5), the Jacobian $J_2h(x_1, x_2, y_1)$ in the acceptance function (4.1.2) can be expressed directly in terms of $R(x)$. For fixed (x_1, y_1) and $y_2 = h(x_1, x_2, y_1)$, (4.1.4) and (4.1.5) imply

$$R(x_1, x_2) \ = \ R\big(y_1,\, h(x_1, x_2, y_1)\,\big)$$

By the chain rule for Jacobians

$$J_2R(x_1, x_2) \ = \ J_2R(y_1, y_2)\, J_2h(x_1, x_2, y_1)$$

and thus

$$J_2 h(x_1, x_2, y_1) \;=\; \frac{J_2 R(x_1, x_2)}{J_2 R(y_1, y_2)} \qquad \text{for} \quad y_2 = h(x_1, x_2, y_1) \tag{4.1.6}$$

In this case, the acceptance function (4.1.2) can be replaced by the more symmetric relation

$$a(x, y) = \min\left\{1,\; \frac{\pi(y) q(y_1, x_1) J_2 R(x_1, x_2)}{\pi(x) q(x_1, y_1) J_2 R(y_1, y_2)}\right\} \tag{4.1.7}$$

It is possible that, for any smooth function $h(x_1, x_2, y_1)$ satisfying (4.1.4), there exists a function $R(x)$ satisfying (4.1.5). As far as I know, this is an open question.

We give two examples before proving Theorem 4.1.1.

4.1.1. Examples. (1) Assume $d = 1$, $n = m + 1$, and $x = (x_1, x_2)$ for $x_1 \in R^1$ and $x_2 \in R^m$. Define a random walk in R^n with an arbitrary step in x_1 and

$$h(x_1, x_2, y_1) = x_2 + (y_1 - x_1) \tag{4.1.8}$$

in (4.1.1). This corresponds to a random motion $X_1 \to Y_1$ in R^1 followed by a parallel shift by the same amount in each component of X_2. Note that by (4.1.8)

$$\begin{aligned} y_2 - h(x_1, x_2, y_1) \;&=\; y_2 - y_1 - (x_2 - x_1) \\ &=\; -\big(x_2 - h(y_1, y_2, x_1)\big) \\ &=\; R(y) - R(x) \end{aligned}$$

where $R(x_1, x_2) = x_2 - x_1$ shifts $x_2 \in R^m$ by x_1. Thus $h(x_1, x_2, y_1)$ satisfies (4.1.4) and is also of the form (4.1.5). Since

$$J_2 h(x_1, x_2, y_1) \;=\; J_2 R(x_1, x_2) \;=\; 1$$

the Jacobian does not appear in the acceptance function (4.1.2).

(2) Assume $d = 2$ and $n = m + 2$ and write

$$x = (x_1, x_2) = (x_{11}, x_{12}, x_2), \quad x_2 \in R^m$$

Define a random walk by an arbitrary step in $x_1 = (x_{11}, x_{12})$ and

$$y_2 = h(x_1, x_2, y_1) = y_{11} + (y_{12}/x_{12})(x_2 - x_{11}) \tag{4.1.9}$$

in (4.1.1). This views x_{11} and x_{12} as like a mean and standard deviation for the components x_{2i} of x_2, and updates $x_2 \to y_2$ accordingly after changes in x_{11} and x_{12}. In this case

$$\begin{aligned} y_2 - h(x_1, x_2, y_1) &= y_2 - y_{11} - (y_{12}/x_{12})(x_2 - x_{11}) \\ &= -(y_{12}/x_{12})\big(x_2 - h(y_1, y_2, x_1)\big) \\ &= y_{12}\big(R(y) - R(x)\big) \end{aligned}$$

where

$$R(x) = (x_2 - x_{11})/x_{12}$$

is the normalizing transformation for $x_2 \in R^m$. Thus $h(x_1, x_2, y_1)$ is also of the form (4.1.4) and (4.1.5). In this case

$$J_2 h(x_1, x_2, y_1) = (y_{12}/x_{12})^m$$

and the acceptance function (4.1.2) is

$$a(x, y) = \min\left\{1,\ \frac{\pi(y)q(y, x_1)(y_{12}/x_{12})^m}{\pi(x)q(x, y_1)}\right\} \tag{4.1.10}$$

Since $J_2 R(x_1, x_2) = (1/x_{12})^m$, we obtain the same result from (4.1.6).

4.1.2. Domain-dependent steps in skew transformations.
Suppose that the update $X_1 \to Y_1$ in (4.1.1) is defined by a random walk with a position-dependent step size as in Section 3.1.2, so that

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \to \begin{pmatrix} Y_1 \\ k(X_1, X_2, Y_1) \end{pmatrix} = \begin{pmatrix} X_1 + h_X W \\ k(X_1, X_2, X_1 + h_X W) \end{pmatrix} \tag{4.1.11}$$

where $h_X > 0$ and W is a random variable with distribution $f_W(x)$. The transformation $X_1 \to Y_1$ has transition function

$$q(x_1, y_1) = \left(\frac{1}{h_x}\right)^d f_W\left(\frac{y_1 - x_1}{h_x}\right) \tag{4.1.12}$$

as in (3.1.5). Since $k(x_1, x_2, y_1)$ in (4.1.11) does not depend explicitly on h_x, the Jacobian factor in (4.1.2) is unaffected. Thus the random walk satisfies the detailed balance condition with acceptance function

$$\begin{aligned} a(x, y) &= \min\left\{1,\ \frac{\pi(y)q(y_1, x_1)J_2 k(x_1, x_2, y_1)}{\pi(x)q(x_1, y_1)}\right\} \qquad (4.1.13) \\ &= \min\left\{1,\ \frac{\pi(y)}{\pi(x)}\left(\frac{h_X}{h_Y}\right)^d \frac{f_W\big((x_1 - y_2)/h_Y\big)}{f_W\big((y_1 - x_2)/h_X\big)} J_2 k(x_1, x_2, y_1)\right\} \end{aligned}$$

as in (3.1.6).

4.2. Proof of Theorem 4.1.1. The proposal transition function defined by (4.1.1) can be written

$$q(x, dy) = q(x_1, y_1)dy_1\, \delta_{[h(x_1,x_2,y)]}(dy_2) \tag{4.2.1}$$

where (4.2.1) means that the measure $q(x, dy)$ on R^n satisfies

$$\int_{y\in X} \phi(y)q(x, dy) = \int \phi\big(y_1, h(x_1, x_2, y_1)\big)dy_1$$

for measurable functions $\phi(y) \geq 0$.

The Markov chain with transition function $p(x, dy)$ defined in Theorem 2.1.1 for $q(x, dy)$ and $a(x, y)$ is the chain that, for a given value of x, "proposes" a value y with distribution (4.2.1) (which implies $y_2 = h(x_1, x_2, y_1)$) and then either "accepts" the value y and moves to that point, which takes place with probability $a(x, y)$, or else "rejects" the value y and remains at the point x, which takes place with probability $1 - a(x, y)$. This means

$$p(x, dy) = a(x, y)q(x_1, y_1)dy_1\, \delta_{[h(x,y_1)]}(dy_2) + A(x)\delta_x(dy)$$

where

$$A(x) = 1 - \int a(x, y)q(x, dy) = 1 - \int a\big(x, y_1, h(x, y_1)\big)q(x_1, y_1)dy_1$$

We now find sufficient conditions on $h(x_1, x_2, y_1)$ in (4.1.1) and $a(x, y)$ in (4.1.2) for the detailed balance condition (4.1.3).

First, note that the two measures $\pi(x)dx\, p(x, dy)$ and $\pi(y)dy\, p(y, dx)$ in (4.1.3) are measures that are concentrated in $(n+d)$-dimensional submanifolds of R^{2n}. We first require that these two manifolds be the same. The first measure is concentrated on the set of points

$$\mathcal{D}_1 = \{\, (x_1, x_2, y_1, y_2) \,:\, y_2 = h(x_1, x_2, y_1) \,\} \tag{4.2.2}$$

while the second measure is concentrated on

$$\mathcal{D}_2 = \{\, (x_1, x_2, y_1, y_2) \,:\, x_2 = h(y_1, y_2, x_1) \,\} \tag{4.2.3}$$

The condition $\mathcal{D}_1 = \mathcal{D}_2$ is exactly the symmetry condition (4.1.4).

As in the proof of Lemma 2.4.1, the detailed balance condition (4.1.3) follows from the relation

$$\pi(x)dx\; a(x, y)q(x, dy) \;=\; \pi(y)dy\; a(y, x)q(y, dx) \tag{4.2.4}$$

In general, two measures $\mu_1(dxdy)$ and $\mu_2(dxdy)$ are the same in R^{2n} if and only if

$$\iint \phi(x,y)\mu_1(dxdy) = \iint \phi(x,y)\mu_2(dxdy)$$

for all measurable functions $\phi(x,y) \geq 0$. The integral of $\phi(x,y)$ with respect to the left-hand measure in (4.2.4) is

$$\begin{aligned}
&\iint \phi(x,y)\pi(x)a(x,y)q(x,dy)dx \\
&= \iiint \phi\big(x_1,x_2,\, y_1,h(x_1,x_2,y_1)\big)\pi(x_1,x_2) \\
&\qquad \times a\big(x_1,x_2,\, y_1,h(x_1,x_2,y_1)\big)q(x_1,y_1)\,dx_1dx_2dy_1
\end{aligned} \tag{4.2.5}$$

The integral with respect to the second measure in (4.2.4) is

$$\begin{aligned}
&\iint \phi(x,y)\pi(y)a(y,x)q(y,dx)dy \\
&= \iiint \phi\big(x_1,h(y_1,y_2,x_1),\, y_1,y_2\big)\pi(y_1,y_2) \\
&\qquad \times a\big(y_1,y_2,\, x_1,h(y_1,y_2,x_1)\big)q(y_1,x_1)\,dy_1dy_2dx_1
\end{aligned} \tag{4.2.6}$$

The substitution $x_2 = h(y_1,y_2,x_1)$ ($y_2 = h(x_1,x_2,y_1)$) in (4.2.6), viewing (x_1,y_1) as fixed, implies

$$\begin{aligned}
&\iint \phi(x,y)\pi(y)a(y,x)q(y,dx)dy \\
&= \iiint \phi\big(x_1,x_2,\, y_1,h(x_1,x_2,y_1)\big)\pi\big(y_1,h(x_1,x_2,y_1)\big) \\
&\qquad \times a\big(y_1,h(x_1,x_2,y_1),\, x_1,x_2\big)q\big(y_1,x_1\big)\,J_2h(x_1,x_2,y_1)\,dx_1dx_2dy_1 \\
&= \int_{\mathcal{D}_1} \phi(x,y)\pi(y)a(y,x)q(y_1,x_1)J_2h(x_1,x_2,y_1)\,dx_1dx_2dy_1
\end{aligned} \tag{4.2.7}$$

where $J_2h(x_1,x_2,y_1)$ is the Jacobian function in (4.1.2). It follows from (4.2.5) and (4.2.7) that the two measures in (4.2.4) are the same if

$$\pi(x)a(x,y)q(x_1,y_1) \;=\; \pi(y)a(y,x)q(y_1,x_1)J_2h(x_1,x_2,y_1) \tag{4.2.8}$$

for all $(x,y) \in \mathcal{D}$ for $\mathcal{D} = \mathcal{D}_1 = \mathcal{D}_2$ in (4.2.2)–(4.2.3). I now claim that

$$a(x,y) = \min\left\{1,\; \frac{\pi(y)q(y_1,x_1)J_2h(x_1,x_2,y_1)}{\pi(x)q(x_1,y_1)}\right\} \tag{4.2.9}$$

in (4.1.2) satisfies (4.2.8). By (4.1.4)

$$h\big(x_1,\, h(y_1, y_2, x_1),\, y_1\big) = y_2$$

and the chain rule for Jacobians with $x_2 = h(y_1, y_2, x_1)$ implies

$$J_2h(x_1, x_2, y_1)\, J_2h(y_1, y_2, x_1) = 1 \tag{4.2.10}$$

It follows from (4.2.9) that

$$\pi(x)a(x,y)q(x_1,y_1) = \min\{\,\pi(x)q(x_1,y_1),\ \pi(y)q(y_1,x_1)J_2h(x_1,x_2,y_1)\,\}$$

and hence

$$\pi(y)a(y,x)q(y_1,x_1) = \min\{\,\pi(y)q(y_1,x_1),\ \pi(x)q(x_1,y_1)J_2h(y_1,y_2,x_1)\,\}$$

Hence by (4.2.10)

$$\begin{aligned}&\pi(y)a(y,x)q(y_1,x_1)J_2h(x_1,x_2,y_1)\\ &\qquad = \min\{\pi(y)q(y_1,x_1)J_2h(x_1,x_2,y_1),\ \pi(x)q(x_1,y_1)\}\end{aligned}$$

The relation (4.2.8) follows from the fact that the right-hand sides of the first and third of the three equations above are the same.

It follows as in the proof of Theorem 2.4.1 that the maximal pointwise solution of (4.2.8) for functions $0 \le a(x,y) \le 1$ is the acceptance function (4.2.9).

Remark. The notion of skew transformation is similar to the ideas of *partial resampling* and *generalized multigrid methods* discussed in Liu and Sabatti (2000) and in Sections 8.1 and 8.3 of Liu (2001). See also the remarks at the end of the next section.

4.3. Temporary reparametrization as a skew transformation.
Let $x = (x_1, x_2)$ for $x_1 \in R^d$, $x_2 \in R^m$, and $x \in R^n$ as in the previous section. In principle, a natural way to carry out a transformation

$$X \;=\; \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \;\to\; \begin{pmatrix} Y_1 \\ h(X_1, X_2, Y_1) \end{pmatrix} \tag{4.3.1}$$

where $P(Y_1 \in dy_1 \mid X_0 = x) = q(x, y_1)dy_1$ is to change coordinates in R^n in such a way to fix the first d components, then carry out a component update of the first d coordinates only, and then change coordinates back. Specifically, let

$$S(x) \;=\; S\left[\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}\right] \;=\; \begin{pmatrix} x_1 \\ S_2(x_1, x_2) \end{pmatrix}$$

be a one-one continuously-differentiable nonsingular mapping of R^n into itself that fixes the first d coordinates. Suppose that we make a proposal $X_1 \to Y_1$ that changes only the first d new coordinates:

$$S\left[\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}\right] = \begin{pmatrix} X_1 \\ S_2(X) \end{pmatrix} \to \begin{pmatrix} Y_1 \\ S_2(X) \end{pmatrix} \tag{4.3.2}$$

Set $S_{22}(x_2; x_1) = S_2(x_1, x_2)$ when we view $S_2(x_1, x_2)$ as a transformation of $x \in R^m$ with $x_1 \in R^d$ as a parameter. Then the mapping (4.3.1) in the original coordinates is

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \to S^{-1}\left[\begin{pmatrix} Y_1 \\ S_2(X) \end{pmatrix}\right] = \begin{pmatrix} Y_1 \\ S_{22}^{-1}\big(S_{22}(X_2; X_1)\,; Y_1\big) \end{pmatrix} \tag{4.3.3}$$

This is a skew transformation of the form (4.1.1) for

$$h(X_1, X_2, Y_1) = S_{22}^{-1}\big(S_{22}(X_2; X_1);\, Y_1\big) \tag{4.3.4}$$

If $Y_2 = h(X_1, X_2, Y_1)$, then by (4.3.4) and (4.3.3)

$$S_2(Y_1, Y_2) = S_{22}\big(h(X, Y_1); Y_1\big) = S_{22}(X_2, X_1) = S_2(X_1, X_2)$$

Since $S(X)$ is one-one, this implies that if $Y_2 = h(X, Y_1)$ for $Y = (Y_1, Y_2)$ if and only $S_2(Y) = S_2(X)$. In particular, the function $h(x_1, x_2, y_1)$ in the reparametrization transformation (4.3.3) satisfies the symmetry conditions (4.1.4)–(4.1.5) for $R(x) = S_2(x)$. It then follows from Theorem 2.1.1 that

Theorem 4.3.1. Let $q(x, dy)$ be the proposal distribution defined by (4.3.3) for the reparametrization (4.3.2). Then the associated Markov transition function $p(x, dy)$ for a density $\pi(x)$ and the acceptance function

$$a(x, y) = \min\left\{1,\ \frac{\pi(y)q(y, x_1)J_2S(x_1, x_2)}{\pi(x)q(x, y_1)J_2S(y_1, y_2)}\right\}$$

satisfies the detailed balanced condition (4.1.3). In particular, the Markov chain associated with (4.3.1) has $\pi(x)$ as a stationary measure.

Remarks. (1) If $h(X, Y_1)$ in (4.3.3) is replaced by a Gibbs sampler step from the density $\pi(x)$ restricted to the "fiber"

$$\{\, Y : S_2(Y) = S_2(X) \,\} \tag{4.3.5}$$

then (4.3.1)–(4.3.3) is essentially the same as the *covariance-adjusted* Markov chain of Liu (1998) (referenced in Chen *et al.* 2000).

(2) Liu and Sabatti (2000) define a similar procedure called *Grouped Move Multi-Grid Monte Carlo* by Chen *et al.* (2000). Here (4.3.5) are the orbits of a locally compact group G acting on X. The transformation $h(X, Y_1)$ in (4.3.3) is either a Gibbs-sampler step from the measure induced on (4.3.5) by $\pi(x)dx$ and Haar measure on G or else related Metropolis-Hastings updates (Chen *et al.* 2000, Liu 2001). See Liu and Sabatti (2000) for examples. Possible choices of the group G for $X = R^n$ would be the Lie group R^m itself or the nilpotent Lie group of shift and scale transformations acting on R^m. This in fact is the same as the examples (4.1.8)–(4.1.10) except that the transformations in Section 3.1 are deterministic within fibers rather than random.

5. Bayesian models in statistics.

5.1. Introduction. The Metropolis-Hastings algorithm is often used in Bayesian analyses in statistics. In fact, the MH algorithm is large part of the reason for the increased popularity of Bayesian methods in the last 50 years.

For definiteness, assume that one has an observation of a vector-valued random variable X (which might, for example, be repeated observations of vector-vectored random variables of lower dimension). Assume that the distribution of X depends on a parameter θ that we want to estimate. Both X and θ may be vector valued. Assume for definiteness that the theoretical distribution of X is given by the density $G(X, \theta)$, so that

$$E\big(f(X)\big) = \int f(x)\, G(x, \theta)\, dx \tag{5.1.1}$$

for functions $f(x) \geq 0$. "Likelihood methods" in statistics for estimating θ are based on the idea that, for given observed data X, those values of θ for which $G(X, \theta)$ is relatively large are more likely to be close to the "true" value of θ that generated the random data X. This approach assumes that x in $G(x, \theta)$ is held constant at the observed value X and that the parameter θ is the true variable. To emphasize this difference in viewpoint, we use the term "likelihood" instead of "probability density" and generally use a different letter for G, for example

$$L(X \mid \theta) = G(X, \theta) \tag{5.1.2}$$

Note that, strictly speaking, (5.1.2) is an abuse of notation, since "$\mid$" in (5.1.2) usually denotes conditioning with respect to a random variable and θ in (5.1.1) is a parameter, not a random variable.

The Bayesian approach is to go one step further and, in fact, set up a structure in which θ can be treated a random variable instead of a parameter, and more specifically in such a way that θ and X are random variables on the same probability space. To do this, we first introduce an arbitrary "prior distribution" or "prior density" $\pi_0(\theta)$ for θ and write

$$L(X, \theta) = \pi_0(\theta)\, L(X \mid \theta) \tag{5.1.3}$$

Then

(i) $$\int L(x, \theta)\, dx \;=\; \pi_0(\theta) \int L(x \mid \theta)\, dx \;=\; \pi_0(\theta) \int G(x, \theta)\, dx \;=\; \pi_0(\theta)$$

(ii) $$\iint L(x, \theta)\, dx d\theta \;=\; \int \pi_0(\theta)\, d\theta \;=\; 1$$

Thus (i) $L(x, \theta)$ in (5.1.3) is a joint probability density for two random variables X and θ (more commonly X and θ are random vectors) and (ii) for the joint density $L(x, \theta)$, the marginal density of θ is $L(\theta) = \pi_0(\theta)$. By the usual formula for conditional density

$$L(X \mid \theta) \;=\; \frac{L(X, \theta)}{L(\theta)} \;=\; \frac{\pi_0(\theta)\, L(X \mid \theta)}{\pi_0(\theta)} \;=\; L(X \mid \theta)$$

Thus, in this framework, $L(X \mid \theta)$ in (5.1.2) is indeed a conditional density.

Given $L(X, \theta)$, it is natural to consider the conditional distribution of the random variable θ given the observed data X, which is

$$\pi_1(\theta \mid X) \;=\; L(\theta \mid X) \;=\; \frac{L(X, \theta)}{\int L(s, X)\, ds} \tag{5.1.4}$$

$$= \; \frac{\pi_0(\theta)\, L(X \mid \theta)}{\int \pi_0(s) L(X \mid s)\, ds} \;=\; C_X\, \pi_0(\theta) L(X \mid \theta)$$

where C_X depends only on X.

The function $\pi_1(\theta \mid X) = L(\theta \mid X)$ in (5.1.4) is called the *posterior density* of θ given X. The basic idea of Bayesian statistics is to make inferences about the unknown value of θ given data X based on this conditional density. For example, given (5.1.4), the *Bayes estimator* of θ is the average of θ over $\pi_1(\theta \mid X)$, or equivalently

$$\widehat{\theta} \;=\; \widehat{\theta}_B \;=\; E(\theta \mid X) \;=\; \int \theta\, \pi_1(\theta \mid X)\, d\theta \tag{5.1.5}$$

$$= \; C_X \int \theta\, \pi_0(\theta) L(\theta, X)\, d\theta$$

An important advantage of Bayesian methods is that recipes like (5.1.5) give you well-defined estimators of unknown paraments almost without thinking in situations in which a classical estimator is not clear, or even worse if there are multiple conceivable classical estimates that give widely different answers.

The main disadvantage of Bayesian methods is that all inferences depend on the prior $\pi_0(\theta)$. Only in rare cases is there a natural candidate for $\pi_0(\theta)$. In principle, inferences should always be done for more than one choice for $\pi_0(\theta)$. If the resulting inferences are not close, then the problem should be rethought, or else you should gather more data.

The degree of confidence that one might have in a Bayesian estimator such as $\widehat{\theta}_B$ can be measured by the distribution of $\pi_1(\theta \mid X)$ about $\widehat{\theta}_B$. For example, if Q is a set of values of θ such that

$$\widehat{\theta}_B \in Q \quad \text{and} \quad \int_Q \pi_1(\theta)\, d\theta \geq 0.95$$

then we can say that we are "95% posterior sure" that Q contains the true value of θ. A set Q with these properties is called a "95% credible region" for θ. This is the Bayesian analog of the classical 95% confidence interval or confidence region, in which θ is treated as a parameter and not as an unobserved value of a random variable. However, the terms "credible region" and "confidence region" are sometimes used interchangeably.

5.2. "Improper" priors and quasi-stability. An alternative justification of Bayesian methods can be given as follows. Given a statistical model and data X, the likelihood $L(X \mid \theta)$ should give most of our information about θ. Classical (non-Bayesian) statistical methods are often based on the *Maximum Likelihood Estimator* (MLE) of θ, which is that value $\widehat{\theta} = \widehat{\theta}(X)$ at which $L(\theta, X)$ attains its maximum value over θ.

An alternative approach might be to consider the measure $L(\theta, X)d\theta$ instead and ask where most of the mass of this measure is concentrated. This leads to the measure

$$\pi_2(\theta)d\theta \;=\; \pi_2(\theta \mid X)\, d\theta \;=\; L(X \mid \theta)\, d\theta \tag{5.2.1}$$

instead of $\pi_1(\theta) = \pi_1(\theta \mid X)$. The measure $\pi_2(\theta)d\theta$ is called the "Bayesian posterior with uniform improper prior" if $\int_X d\theta = \infty$ and "with uniform (proper) prior" if $\int_X d\theta < \infty$ (within the normalization constant $\int_X d\theta$). In general, MCMC in the sense of

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} F(X_k) \;=\; \int_X F(\theta)\pi_2(\theta)\, d\theta \quad \text{a.s.}$$

will fail if

$$\int \pi_2(\theta)\, d\theta = \int L(X \mid \theta)\, d\theta = \infty \tag{5.2.2}$$

If X_n is ergodic, (5.2.2) implies that X_n is null recurrent with

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} F(X_k) \;=\; 0 \quad \text{a.s.} \tag{5.2.3}$$

whenever $\int |F(\theta)|\, \pi_2(\theta) d\theta < \infty$. When this happens, the Markov chain X_n is said to "converge to infinity".

If in fact (5.2.2) is the case, then *any* estimator of a function of θ based on Bayesian methods can be considered to be an artifact of the prior $\pi_0(\theta)$, since the closer that the prior $\pi_0(\theta)$ is to the "uniform improper prior" $\pi_0(\theta) = 1$, the closer that the estimator is to the situation (5.2.3). Note that (5.2.3) implies that median estimators of θ are infinite as well as mean-based estimators. However, it is still possible for a well-behaved classical MLE $\widehat{\theta}(X)$ to exist.

An argument against the improper prior (5.2.1) or against uniform priors in general is that "$d\theta$" in (5.2.1) is not invariant under changes of variable unless it is a discrete measure. For example, if $\psi = \theta^2$, then $\theta = \sqrt{\psi}$ and $L(\theta)d\theta = L(\psi)d\psi/2\sqrt{\psi}$. Then the "natural" measure $d\theta$ has been changed to $d\psi/2\sqrt{\psi}$. This problem does not arise with the posterior distribution $\pi_1(\theta) = \pi_0(\theta)L(X \mid \theta)$ as long as one views the prior and posterior distributions $\pi_0(d\theta)$ and $\pi_1(d\theta)$ as measures instead of as functions.

Of course, the strongest argument against an improper prior (5.2.1) with (5.2.2) is (5.2.3). In that case, if n is sufficiently large, all parameter values estimated by the sample-path averages (5.2.3) will be zero.

It often happens in practice that (5.2.3) appears to converges to reasonable values for most components of a high-dimensional X_n with an improper prior. This is because the components of X_n often converge on different time scales. That is, some component or components of X_n converge to infinity, but other components are nearly independent of the first set of components, and adjust themselves to the first set of components in a fast enough time scale that the averages (5.2.3) give stable reasonable estimates for those components. Phenomena of this sort are called *quasi-stability*.

A related problem is that an improper prior for one component, for which the likelihood is obviously integrable without a normalizable prior distribution, can mysteriously lead to unstable behavior in other components even for other components with normalized priors. This is another aspect of quasi-stability.

5.3. Sampling from the prior. Given the form of the distribution $g(\theta) = \pi_0(\theta)L(X \mid \theta)$ in (5.1.4), a plausible choice for updating θ in the MH algorithm might be the proposal function

$$q(\theta, y) = \pi_0(y)$$

This is independence sampling from the prior distribution for θ (where we now assume $\int \pi_0(\theta)d\theta = 1$). The acceptance function (5.1.4) becomes

$$a(\theta, y) = \min\left\{1,\ \frac{g(y)/q(y)}{g(\theta)/q(\theta)}\right\} = \min\left\{1,\ \frac{L(X \mid y)}{L(X \mid \theta)}\right\}$$

This nicely separates the effect of the prior distribution $\pi_0(\theta)$ and likelihood $L(X \mid \theta)$ on the Markov chain X_n.

5.4. Conjugate priors. The joint density $\pi_0(\theta)L(X \mid \theta)$ takes a simple form in many important cases in statistics. For example, suppose that X has a Poisson distribution with mean θ, so that the likelihood is

$$L(X \mid \theta) = e^{-\theta}\,\frac{\theta^X}{X!}\,, \qquad X = 0, 1, 2, 3, \ldots \tag{5.4.1}$$

Suppose that we choose a gamma density with parameters (α, β) for the prior density $\pi_0(\theta)$:

$$\pi_0(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta}\,, \qquad 0 \le \theta < \infty \tag{5.4.2}$$

We can write $\pi_0(\theta) \approx \mathcal{G}(\alpha, \beta)$ symbolically. Then the posterior density

$$\begin{aligned}\pi_1(\theta \mid X) &= C_X \pi_0(\theta) L(X \mid \theta) = C_X\, \theta^{\alpha-1}e^{-\beta\theta}e^{-\theta}\theta^X \\ &= C_X(\alpha, \beta)\, \theta^{X+\alpha-1}e^{-(1+\beta)\theta}\end{aligned}$$

is also a gamma distribution. Symbolically

$$\pi_1(\theta) = \pi(\theta \mid X) \ \approx\ \mathcal{G}(\alpha + X, \beta + 1)$$

In particular, if $L(X \mid \theta)$ is the Poisson likelihood (5.4.1) and $\pi_0(\theta)$ is the gamma density (5.4.2), then the posterior density $\pi_1(\theta)$ is a gamma density with different parameters.

When this happens for an arbitrary family of densities (here, the gamma densities (5.4.2)), we say that family of densities is a *conjugate prior* for the

likelihood, here for the Poisson likelihood (5.4.1). Colloquially, we say that the family of gamma densities is a conjugate prior for Poisson sampling. There are only a few cases where conjugate priors for likelihoods are known, but they cover many of the most important distributions in statistics (see e.g. DeGroot 1989, Chapter 6).

As a second example, suppose that $\pi_0(\theta)$ is the beta density

$$\pi_0(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\,\theta^{\alpha-1}(1-\theta)^{\beta-1} \qquad \text{for} \qquad 0 \le \theta \le 1 \tag{5.4.3}$$

(Symbolically, $\pi_0(\theta) \approx \mathcal{B}(\alpha,\beta)$.) If we toss a biased coin n times with probability of heads $\Pr(H) = \theta$ and observe $X = k$ heads in a particular order, then the joint density of (θ, X) is

$$\pi_0(\theta)\binom{n}{X}\theta^X(1-\theta)^{n-X} = C(\alpha,\beta,n,X)\,\theta^{X+\alpha-1}(1-\theta)^{n-X+\beta-1}$$

Thus $\pi_1(\theta \mid X) \approx \mathcal{B}(\alpha+X,\ \beta+n-X)$. This means that the family of beta densities is a conjugate prior for binomial sampling.

It follows from a similar argument that the family of *Dirichlet densities*

$$\pi_0(\theta) = \frac{\Gamma(\alpha)}{\prod_{i=1}^d \Gamma(\alpha_i)}\prod_{i=1}^d \theta_i^{\alpha_i-1}, \qquad \alpha_i > 0, \quad \sum_{i=1}\alpha_i = \alpha \tag{5.4.4}$$

where $\theta = (\theta_1,\ldots,\theta_d)$ with $\theta_i > 0$ and $\sum_{i=1}^d \theta_i = 1$ is a conjugate prior for *multinomial sampling*:

$$\Pr(X = i \mid \theta) = \theta_i, \qquad X \in \{\,1,2,\ldots,d\,\} \tag{5.4.5}$$

(*Exercise*: Prove that (5.4.4) is a conjugate prior for the likelihood (5.4.5)).

The advantage of conjugate priors in the MH algorithm is that if one knows how to generate independent random variates θ_n efficiently from the prior familiy (for example, the gamma density (5.4.1) or the beta density (5.4.3)), then one can do Monte Carlo sampling from the posterior density for any observed X.

This means that we always have available at least one numerically efficient candidate for MH sampling. If θ is one part of a larger parameter vector and the prior family is the conditional distribution, this would be a Gibbs sampler step. Efficient methods are available for simulating independent random variables with arbitrary gamma, beta, normal, uniform, and exponential distributions. See for example Devroye (1986), Press *et al.* (1992), and Fishman (1995) in the references.

6. Hidden Variables.

In many applications, the likelihood $L(X \mid \theta)$ is complex but would be much simpler if an unobserved random variable Z could be observed or treated as data. More generally, assume

$$L(X \mid \theta) = \int_z G(X, z, \theta)\, dz \tag{6.1.1}$$

for some function $G(x, z, \theta) \geq 0$. Then

$$\int_x L(x \mid \theta)\, dx \;=\; \iint G(x, z, \theta)\, dz dx \;=\; 1$$

so that $G(x, z, \theta)$ is a probability density in (x, z) for fixed θ, and can be viewed as the joint probability density of two random variables, X and Z. Note that we might not have thought of Z as a potential random variable before noticing the relationship (6.1.1.)

If we use the same prior density $\pi_0(\theta)$ for θ as for $L(X \mid \theta)$ in equation (5.1.3) in Section 4.1, then

$$L(x, z, \theta) \;=\; \pi_0(\theta) G(x, z, \theta) \tag{6.1.2}$$

is a joint probability of three random variables X, Z, and θ. With respect to this joint density,

$$\text{(i)} \quad L(X, Z \mid \theta) \;=\; \frac{L(X, Z, \theta)}{L(\theta)} \;=\; \frac{\pi_0(\theta)\, G(X, Z, \theta)}{\pi_0(\theta)} \;=\; G(X, Z, \theta)$$

$$\text{(ii)} \quad \int_z L(X, z \mid \theta)\, dz \;=\; \int_z G(X, z, \theta)\, dz \;=\; L(X \mid \theta)$$

$$\text{(iii)} \quad L(Z \mid \theta) \;=\; \int_x L(x, Z \mid \theta)\, dx \;=\; \int_x G(x, Z, \theta)\, dx$$

Thus (i) $G(X, Z, \theta)$ is the likelihood of (X, Z) as a function of θ, (ii) the marginal density of X with respect to $L(X, Z \mid \theta)$ is $L(X \mid \theta)$, and (iii) the marginal $L(Z \mid \theta)$ can be calculated in terms of $G(x, z, \theta)$.

The variables Z in (6.1.1) and (6.1.2) can be viewed as either unobserved data or else as additional parameters, although (6.1.1) may be more suggestive of unobserved data. The process of going from data X with likelihood $L(X \mid \theta)$ to (X, Z) with likelihood $L(X, Z \mid \theta)$ is often called *data augmentation*. A formula that connects two senses of the hidden variables Z, as

possible parameters or as unobserved data, can be found in the following result.

Lemma 6.1.1. Let the likelihood $L(X \mid \theta)$ and joint density $L(x, z, \theta)$ be defined by (6.1.1) and (6.1.2), and use the same prior density $\pi_0(\theta)$ for both X and (X, Z). Then the posterior density

$$\pi_1(\theta, Z \mid X) \;=\; L(\theta, Z \mid X) \;=\; C_X\, L(X, Z \mid \theta)\, \pi_0(\theta) \tag{6.1.3}$$

Remark. This formula connects the posterior density $\pi_1(\theta, Z \mid X)$, in which Z is viewed as a parameter, with the likelihood $L(X, Z \mid \theta)$, in which Z appears as unobserved data.

Proof.

$$\begin{aligned}\pi_1(\theta, Z \mid X) \;&=\; L(\theta, Z \mid X) \;=\; \frac{L(\theta, X, Z)}{P(X)} \;=\; \frac{L(X, Z \mid \theta)\, \pi_0(\theta)}{P(X)} \\ &=\; C_X\, L(X, Z \mid \theta)\, \pi_0(\theta)\end{aligned}$$

Remarks. **(1)** The Metropolis-Hastings Markov chain $X_n = (\theta_n, Z_n)$ generated from (6.1.2) provides not only estimates of θ but also provides estimates of the conditional distribution of Z given X. That is, any MCMC that treats a "hidden variable" Z as a parameter also provides estimates of Z. If the hidden variable Z is of high dimension, as is often the case, this can provide a great deal of additional information.

(2) The right-hand side of (6.1.3) can be found either directly or through the identity

$$\begin{aligned}P(X, Z \mid \theta) \;&=\; \frac{P(X, Z, \theta)}{P(\theta)} \;=\; \frac{P(X \mid Z, \theta)\, P(Z, \theta)}{P(\theta)} \\ &=\; P(X \mid Z, \theta)\, P(Z \mid \theta)\end{aligned} \tag{6.1.4}$$

(3) The Metropolis-Hastings Markov chains $W_n^{(1)} = (\theta_n, Z_n)$ and $W_n^{(2)} = \theta_n$ can both be used to estimate the posterior density of θ. An important difference is that $W_n^{(2)} = \theta_n$ requires us to carry out the integral in (6.1.1) for each evaluation of the likelihood $L(X \mid \theta)$ while $W_n = (\theta_n, Z_n)$ does not.

By (6.1.1), each value Z_n in $W_n^{(1)} = (\theta_n, Z_n)$ is a Monte Carlo simulation of the integral (6.1.1) for one value of z for the current value θ_n of θ. When we would not normally evaluate an integral by a single Monte Carlo simulation, the fact that we are averaging over a long trajectory $\{(\theta_n, Z_n)\}$ usually corrects for this. In practice, hidden-variable Markov chain $Z_n = (\theta_n, Z_n)$

are nearly always almost as efficient as the "integrated" chain $W_n^{(2)} = \theta_n$ *without* the hidden variables Z_n, and it is difficult to find examples in which the chain (θ_n, Z_n) provides a verifiably less efficient estimator of θ.

(4) Note the we do not need a prior for Z. This is not surprising if we view Z as unobserved data. If instead we view Z as additional parameter(s), one could say that its prior is implicit in (6.1.2). However, in any event, supplying an (additional) prior for Z will lead to incorrect formulas.

(5) We still need an updating procedure for $W = (\theta, Z)$. If the function $G(x, z, \theta)$ is simple, this can often be done as a simple component update of θ for fixed Z and then of Z for fixed θ.

Example. Suppose that we are given DNA sequence data $\mathfrak{D}$ from n individuals from the same species. Assume that $\mathfrak{D}$ depends only on the DNA sequence of the most recent common ancestor (MRCA) of the n individuals and the mutation rate on the pedigree that connects the n individuals with their MRCA. Assume that there are no repeat mutations at the same site in the pedigree and let X be the number of DNA sites in the sample that are *polymorphic*; that is, at which there is more than one base in the n sequences.

Then, under reasonable assumptions about DNA mutation, including the assumption that the DNA sequences are not subject to Darwinian selection at the polymorphic sites, the likelihood $L(X \mid \theta)$ for the number of polymorphic sites X is Poisson with mean $\theta \operatorname{Len}(Z)$, where Z is the unobserved pedigree of the n individuals since their MRCA, $\operatorname{Len}(Z)$ is the total length of the pedigree Z, and θ is a scaled mutation rate. Given these assumptions and the pedigree Z, the number of polymorphic sites is Poisson with a mean proportion to the total length of the pedigree, so that

$$L(X \mid \theta) = \int_z L_P\big(X, \theta \operatorname{Len}(z)\big)\, g(z)\, dz \tag{6.1.5}$$

where $L_P(X, A)$ in (6.1.5) is the Poisson likelihood (5.4.1) with mean A. The integral $g(z)\, dz$ is the integral over all possible pedigrees of a sample of n individuals. This is very complex, but is easy to simulate using Kingman's coalescent algorithm.

If we could observe Z, then we could use the much simpler likelihood $L_P\big(X, \theta \operatorname{Len}(Z)\big)$. While Z is not known, we can consider it as a hidden variable with joint likelihood

$$L(X, Z \mid \theta) = L_P\big(X, \theta \operatorname{Len}(Z)\big)\, g(Z)$$

and joint posterior density

$$\pi_1(\theta, Z \mid X) = L_P\big(X, \theta \operatorname{Len}(Z)\big)\, g(Z)\, \pi_0(\theta) \tag{6.1.6}$$

by Theorem 6.1.1. Since Poisson random variables are easy to simulate, we have Gibbs-sampler updates for both θ and Z and hence an estimation procedure for θ given the number of polymorphic sites X.

Appendix A. Ergodic theory and random mixtures.

A.1. Ergodic theorems. As before, let X_n be a Markov chain with state space $X \subseteq R^d$ defined on a state space X. Assume (i) $\pi(x)dx$ is a stationary probability measure for a transition function $p(x, dy)$ of X_n, (ii) $X_n = X_n(\omega)$ are X-valued random variables on a probability space $(\Omega, \mathcal{F}, P)$, and (iii) X_0 has distribution $\pi(x)dx$. Then by induction

$$E\big(f(X_n)\big) = \int_X f(y) \int_X \pi(x)p^n(x, dy)dx = \int_X f(y)\pi(y)dy = E\big(f(X_0)\big) \tag{A.1}$$

for measurable $f(y) \geq 0$, so that X_n also has distribution $\pi(x)dx$. The Birkhoff ergodic theorem (see below for the exact statement and references) implies that if $\int_X |f(y)|\, \pi(y)dy < \infty$

$$\lim_{n\to\infty} \frac{f(X_0(\omega)) + f(X_1(\omega)) + \cdots + f(X_{n-1}(\omega))}{n} \;=\; g(\omega) \quad \text{a.s. } P(d\omega) \tag{A.2}$$

Here in the following, a.s. ("almost surely") and a.e. ("almost everywhere") mean with the exception of a set of probability or measure zero.

The Markov chain X_n is called *ergodic* if the limiting random variable $g(\omega)$ in (A.2) is always constant. In that case, $g(\omega) = E(f) = \int_X f(y)\pi(y)dy$ a.s.

To obtain more information about the limiting random variable $g(\omega)$, we need more information about the probability space $(\Omega, \mathcal{F}\, P)$. By the Kolmogorov consistency theorem (Kolmogorov 1950, Chung 2001), it is sufficient to assume that that Ω is the infinite product space

$$\Omega = X^\infty = \{\, \omega : \omega = (x_0, x_1, x_2, \ldots) \,\} \tag{A.3}$$

with $X_n(\omega) = x_n \in X$ and $\mathcal{F} = \mathcal{B}(X_1, X_2, \ldots)$. That is, X_n are the coordinate functions on Ω and $\mathcal{F}$ is the smallest σ-algebra of subsets of Ω for which the coordinate functions are measurable.

Given Ω in (A.3), we define the *unilateral shift*

$$\theta(\, (x_0, x_1, x_2, \ldots)\,) = (x_1, x_2, x_3, \ldots)$$

for $\omega \in \Omega$. Note that this implies $X_1(\omega) = X_0(\theta(\omega))$ and, by induction, $X_n(\omega) = X_0(\theta^n(\omega))$ for $n \geq 0$.

The random variables $X_0(\omega)$ and $X_n(\omega)$ have the same probability distribution by (A.1). More generally, for any measurable set $B \in \mathcal{F}$,

$$\theta^{-1}(B) = \{\, \omega : \theta(\omega) \in B \,\} \in \mathcal{F} \quad \text{and} \quad P\big(\theta^{-1}(B)\big) = P(B) \qquad (A.4)$$

In this case, this is equivalent to the condition

$$(X_0, X_1, \ldots, X_m) \approx (X_n, X_{n+1}, \ldots, X_{n+m}) \quad \text{for all } m, n \geq 0 \qquad (A.5)$$

for the $(m+1)$-dimensional joint distributions, which follows from the fact that X_n is a Markov chain with a stationary distribution $\pi(x)dx$.

A mapping $\theta : \Omega \to \Omega$ for a probability space $(\Omega, \mathcal{F}, P)$ is called a *measure-preserving transformation* (m.p.t.) $\theta^{-1}(B) \in \mathcal{F}$ for any $B \in \mathcal{F}$ and (A.4) holds. The m.p.t. $\theta(\omega)$ is *ergodic* if $B = \theta^{-1}(B)$ implies either $P(B) = 0$ or $P(B) = 1$. This is equivalent to the condition that if $f(\theta(\omega)) = f(\omega)$ a.s., then $f(\omega) = \text{const.}$ a.s. (*Exercise*: Prove the equivalence.)

In general, let $Tf(\omega) = f(\theta(\omega))$ for $\mathcal{F}$-measurable functions $f(\omega)$ on Ω. The full statement of the Birkhoff ergodic theorem (Halmos 1956, Garsia 2001) is

Theorem A.1. (Birkhoff Ergodic Theorem) Define $Tf(\omega) = f(\theta(\omega))$ as above where $\theta(\omega)$ is an arbitrary measure-preserving transformation on a probability space $(\Omega, \mathcal{F}, P)$. Then

$$\lim_{n\to\infty} \frac{f(\omega) + Tf(\omega) + \cdots + T^{n-1}f(\omega)}{n} \;=\; g(\omega) \quad \text{a.s. } P(d\omega) \qquad (A.6)$$

for any measurable function $f(\omega)$ with $E(|f|) < \infty$.

Corollary A.1. The limiting function $g(\omega)$ in (A.6) satisfies $Tg(\omega) = g(\omega)$ a.s. The limit is a.s. constant for all $f \in L^1(\Omega, \mathcal{F}, P)$ if and only if $\theta(\omega)$ is ergodic, in which case the limit is $E(f)$.

The operator $Tf(\omega) = f(\theta(\omega))$ in Theorem A.1 satisfies

$$\begin{aligned} &\text{(i)} \;\; f(\omega) \geq 0 \text{ implies } Tf(\omega) \geq 0 \\ &\text{(ii)} \;\; T1(\omega) \;=\; 1 \quad \text{for } 1(\omega) \equiv 1 \\ &\text{(iii)} \;\; \int_\Omega |Tf(\omega)|\, P(d\omega) \leq \int_\Omega |f(\omega)|\, P(d\omega) \end{aligned} \qquad (A.7)$$

for all $f \in L^1(\Omega, \mathcal{F}, P)$.

If Tf is an arbitrary linear operator on $L^1(\Omega, \mathcal{F}, P)$ that satisfies (A.7), we say that T is ergodic if $Tf(\omega) = f(\omega)$ a.e. implies $f(\omega)$ is a.s. constant on Ω. If $Tf(\omega) = f(\theta(\omega))$ for a m.p.t. $\theta(\omega)$, this is equivalent to $\theta(\omega)$ being ergodic.

The *Hopf ergodic theorem* (Garsia 2001) states that if (A.7) holds for an arbitrary linear operator T on $L^1(\Omega, \mathcal{F}, P)$, then the conclusions of Theorem A.1 also hold.

A.2. The von Neumann ergodic theorem. It follows from interpolation theorems in harmonic analysis (Zygmund 1959) that (A.7) implies

$$\int_\Omega Tf(\omega)^2\, P(d\omega) \le \int_\Omega f(\omega)^2\, P(d\omega) \tag{A.8}$$

If $g(\omega)$ is the limiting random variable in (A.6), it follows from Minkowski's inequality and Fatou's theorem that

$$\int_\Omega g(\omega)^2 P(d\omega) \le \int_\Omega f(\omega)^2 P(d\omega) \tag{A.9}$$

If $Tf = f$, then $T^n f = f$ for all $n \ge 0$ by induction and the left-hand side of (A.6) is constant. This implies $f = g$, which of course implies equality in (A.8). A very useful converse is contained in the following:

Theorem A.2. (von Neumann Ergodic Theorem) Let T be a linear operator on a Hilbert space $\mathcal{H}$ such that $\|Tf\| \le \|f\|$ for all $f \in \mathcal{H}$. Then, for any $f \in \mathcal{H}$,

$$\lim_{n\to\infty} \left\| \frac{f + Tf + \cdots + T^{n-1} f}{n} - g \right\| = 0 \tag{A.10}$$

for some $g \in \mathcal{H}$. If $\|f\| = \|g\|$, then $Tf = f$ and $f = g$.

(Riesz-Nagy 1955, Garsia 2001, Halmos 1956).

If T satisfies the Hopf conditions (A.7), then it follows from the same arguments that one has both almost-sure convergence in (A.6) as well as quadratic-mean convergence (A.10) for $\mathcal{H} = L^2(\Omega, \mathcal{F}, P)$. The importance of Theorem A.2 is the second conclusion: Namely, that if

$$\int_\Omega g(\omega)^2 P(d\omega) = \int_\Omega f(\omega)^2 P(d\omega)$$

in (A.6), then one must have had $g(\omega) = f(\omega)$ almost surely and hence $T^n f = f$ for all $n \ge 0$.

A.3. Ergodic families and products of contractions. A set of measure-preserving transformations $\{\, \theta_a(\omega) \,\}$ is called an *ergodic family* on a measure space $(\Omega, \mathcal{F}, P)$ if, whenever $\theta_a(B) = B$ within null sets for all a and some $B \in \mathcal{F}$, then either $P(B) = 0$ or $P(B) = 1$ (Sawyer 1966; the idea also appears in Riesz-Nagy 1956). A generalization of the Birkhoff ergodic theorem is

Theorem A.3. Let $T = T_1 T_2 \dots T_m$ where $T_a f(\omega) = f(\theta_a(\omega))$ where $\{ \theta_a : 1 \le a \le m \}$ is an ergodic family on $(\Omega, \mathcal{F}, P)$. Then, for any $f \in L^1(\Omega, \mathcal{F}, P)$,

$$\lim_{n\to\infty} \frac{f(\omega) + Tf(\omega) + \cdots + T^{n-1} f(\omega)}{n} \;=\; E(f) \quad \text{a.s. } P(d\omega) \qquad (A.11)$$

Proof. Since $T_a f(\omega) = f(\theta_a(\omega))$ for measure-preserving transformations $\theta_a(\omega)$, $Tf(\omega) = f(\theta(\omega))$ is of the same form for

$$\theta(\omega) = \theta_m(\theta_{m-1}(\dots \theta_1(\omega) \dots))$$

Hence by the Birkhoff ergodic theorem

$$\lim_{n\to\infty} \frac{f(\omega) + Tf(\omega) + \cdots + T^{n-1} f(\omega)}{n} \;=\; g(\omega) \quad \text{a.s.}$$

for a random variable $g(\omega)$ that satisfies $g(\theta(\omega)) = g(\omega)$ a.s. It only remains to show that $g(\omega)$ is a.s. constant.

Since $g(\theta(\omega)) = g(\omega)$ a.s., $h(\omega) = \phi(g(\omega))$ for a bounded measurable function $\phi(x)$ also satisfies $h(\theta(\omega)) = h(\omega)$ a.s. and hence $Th(\omega) = h(\theta(\omega)) = h(\omega)$ a.s. Since $\phi(x)$ is bounded, $h(\omega)$ is also bounded. Hence

$$\begin{aligned} E(h^2) \;&=\; E\big((Th)^2\big) \;=\; E\Big(\big((T_1 T_2 \dots T_{m-1}) T_m h\big)^2\Big) \\ &\le\; E\big((T_m h)^2\big) \;\le\; E(h^2) \end{aligned}$$

by multiple applications of (A.8). This implies $E((T_m h)^2) = E(h^2)$ and hence $T_m h = h$ a.s. by the von Neumann ergodic theorem (Theorem A.2). By induction, $T_m h = T_{m-1} h = \dots = T_1 h = h$ a.s. Since $\{ T_a \}$ is an ergodic family, $h(\omega) = \phi(g(\omega)) = C_\phi$ a.s. for all bounded measurable functions $\phi(x)$.

In particular this holds if $\phi(x) = I_{(-\infty,\lambda]}(x)$ for all real λ, which implies that $I_{(-\infty,\lambda]}(g(\omega)) = C(\lambda)$ a.s. For each λ, either $C(\lambda) = 0$ and $g(\omega) > \lambda$ a.s., or else $C(\lambda) = 1$ and $g(\omega) \le \lambda$ a.s. Let $\lambda_0 = \sup\{ r : C(r) = 0 \}$ for rational r. Then $g(\omega) = \lambda_0$ a.s., which completes the proof of Theorem A.3.

A.4. Ergodic families and components-in-sequence MCMC. We next show how Theorem A.3 applies to the components-in-sequence MCMC defined in Section 12. In Section 12, $T_a f(\omega)$ is the result of a one-dimensional Metropolis-Hastings Markov-chain update of the a^{th} component of $X_n \in R^m$, treating the remaining components $x_{-a} \in R^{m-1}$ as parameters. The update is defined in such a way that the one-dimensional conditional density $\pi_i(x_i, x_{-i})$ is a stationary measure of the update.

We first show that T_a acting on R^m has $\pi(x)dx$ as a stationary measure. By definition, $T_a f(x)$ is the Markov operator on R^1 corresponding to the one-dimensional transition function $p_i(x_i, x_{-i}, dy)$ in (3.5.4). The corresponding transition function on R^m is

$$p_i(x, x_i, dy) \;=\; p_i(x_i, x_{-i}, y_i)dy_i \;\delta(x_{-i}, dy_{-i}) \tag{A.12}$$

for $x \in R^m$. Then for measurable $g(x) \geq 0$ on R^m

$$\begin{aligned}
&\iint g(y)p_i(x, x_{-i}, dy)\pi(dx) \\
&\quad = \iint\iint g(y_i, y_{-i})p_i(x_i, x_{-i}, y_i)dy_i\;\delta(x_{-i}, dy_{-i})\pi(dx)dx_{-i}dx_i \\
&\quad = \int\iint g(y_i, x_{-i})p_i(x_i, x_{-i}, y_i)\pi_i(x_i \mid x_{-i})dx_i\;\pi_i(x_{-i})dx_{-i}\;dy_i \\
&\quad = \int\int g(y_i, x_{-i})\pi_i(y_i \mid x_{-i})\pi_i(x_{-i})dx_{-i}\;dy_i \\
&\quad = \int_x g(x)\pi(x)dx
\end{aligned}$$

This implies that T_a is $\pi(x)dx$-preserving on R^m.

Hence by induction $T = T_1T_2 \ldots T_m$ also has $\pi(x)dx$ as a stationary measure. This implies that the limits in (A.11) exist a.s., but does not yet prove that the limits are a.s. constant.

For this, we need to show that the T_a $(1 \leq a \leq m)$ form an ergodic family. By assumption, each operator $T_a f(x)$ is ergodic in the single variable x_a for fixed $x_{-a} \in R^{m-1}$. (We assumed in Section 12 that the acceptance functions $A_a(x) > 0$ a.s., so that the one-dimensional Metropolis updates are ergodic if the one-dimensional proposals are ergodic.)

If $T_a f = f$, it follows from Fubini's theorem, for a.e. $x_{-a} \in R^{m-1}$, that $f(x)$ is a.s. constant in x_a. It follows by induction that if $T_a f = f$ a.s. for $1 \leq a \leq m$, then $f(x)$ is a.s. constant. This completes the proof that T_a $(1 \leq a \leq m)$ form an ergodic family on R^m, and hence completes the proof that Theorem A.3 applies to the components-in-sequence updates in Section 12.

A.5. Ergodic families and random mixtures. Let $\{\, p_a(x, dy) : a \in I \,\}$ be set of Markov transition functions on a state space X that have the same stationary measure $\pi(x)dx$. Define a random process X_n on X by, at each n, choosing a value of $a \in I$ at random with probability $w_a > 0$ and then X_n by

$$P(X_{n+1} \in dy \mid X_n = x, a) \;=\; p_a(x, dy) \tag{A.13}$$

Since the values of a are chosen independently with probabilities w_a at each n, it follows that

$$P(X_{n+1} \in dy \mid X_n = x) \;=\; \sum_a w_a p_a(x, dy) \tag{A.14}$$

The sample paths $\{ X_n(\omega) \}$ of the Markov chains defined by (A.14) and implicityly by (A.13) are the same, but (A.13) can be used to define a Markov chain $\{ (X_n(\omega), A_n(\omega) \}$ that also has information about the states $a \in I$.

We consider the Markov chain X_n modeled by (A.14) first. Since the right-hand side of (A.14) also has $\pi(x)dx$ as a stationary measure, it follows from Theorem A.1 that, if $\Pr(X_0 \in dx) = \pi(x)dx$, then

$$\lim_{n\to\infty} \frac{f(X_0(\omega)) + f(X_1(\omega)) + \cdots + f(X_{n-1}(\omega))}{n} \;=\; g(\omega) \tag{A.15}$$

almost surely for any $f(x)$ with $\int f(y)\pi(y)dy < \infty$. Then

Theorem A.4. Suppose that $\{ p_a(x, dy) \}$ is an ergodic family in the sense of Theorem A.3. Then the limit in (A.15) is a.s. constant, and in particular equal to the constant $E(f) = \sum_a w_a \int f(y)\pi(y)dy$.

Proof. As in Section A.1, it is sufficient to assume that $\Omega = X^\infty$ is the infinite product in (A.3) and $X_n(\omega) = X_0(\theta^n(\omega))$ where $\theta(\omega)$ is the unilateral shift on Ω. Also, by Corollary A.1, $g\big(\theta(\omega)\big) = g(\omega)$ a.s.

Similarly, $\phi(g(\theta(\omega))) = \phi(g(\omega))$ a.s. for any bounded measurable function $\phi(x)$. Since X_n is Markov with the transition function $p(x, dy)$ in (A.14),

$$h(x) = E\big(\phi(g) \mid X_0 = x\big) = \sum_a w_a \int_y h(y) p_a(x, dy) \tag{A.16}$$

Now $h(x)$ is bounded because $\phi(g)$ is bounded, and by Cauchy's inequality

$$\int h(x)^2 \pi(x)dx \;=\; \int \bigg(\sum_a w_a \int h(y) p_a(x, dy) \bigg)^2 \pi(x)dx \tag{A.17}$$

$$\le \int \sum_a w_a \bigg(\int h(y) p_a(x, dy) \bigg)^2 \pi(x)dx \;\le\; \sum_a w_a \iint h(y)^2 p_a(x, dy)\pi(x)dx$$

$$= \sum_a w_a \int h(y)^2 \pi(y)dy \;=\; \int h(y)^2 \pi(y)dy \;<\; \infty$$

since $\int p_a(x, dy)\pi(x)dx = \pi(y)dy$ for all a. Thus the inequalities in (A.17) are all equalities, which implies

$$T_a h(x) = \int h(y) p_a(x, dy) = h(x) \quad \text{for } \pi(x)dx \text{ a.e. } x, \text{ all } a \tag{A.18}$$

Hence $h(x) = E\big(\phi(g) \mid X_0 = x\big) = C(\phi)$ a.e. $\pi(x)dx$, since $p_a(x, dy)$ is an ergodic family.

By a similar argument, if $\psi(x_0, x_1, \ldots, x_n)$ is bounded,

$$E\big(\psi(X_0, \ldots, X_n)\phi(g(\omega))\big) \;=\; E\big(\psi(X_0, \ldots, X_n)\phi(g(\theta^n(\omega))\big) \qquad (A.19)$$
$$=\; E\big(\psi(X_0, \ldots, X_n)E\big(\phi(g) \mid X_0 = X_n\big)\big) \;=\; E\big(\psi(X_0, \ldots, X_n)\big)\, C(\phi)$$

Since functions of the form $\phi(X_0, X_1, \ldots, X_n)$ for all n are dense in $L^2(\Omega, \mathcal{F}, P)$, it follows from (A.19) that

$$E\big(\phi(g)^2\big) = E(\phi(g))C(\phi) = E\big(\phi(g)\big)^2 \qquad (A.20)$$

and $\phi(g(\omega)) = C(\phi)$ a.s. for all bounded measurable functions $\phi(x)$. Since we can take $\phi(x) = I_{(-\infty,\lambda]}(x)$ for arbitrary real λ, we conclude $g(\omega) = C(g) = E(g) = E(f)$ a.s. as in the proof of Theorem A.3. This completes the proof of Theorem A.4.

We can also model the Markov chain $Z_n = (X_n, A_n)$ formed by the Markov chain X_n together with the choice of transition function A_n in (A.13). First, we extend the probability space Ω in (A.3) by setting

$$\Omega^e = \{\, w : w = (z_0, z_1, z_2, \ldots), \quad z_i = (x_i, a_i) \,\} \qquad (A.21)$$

where $x_i \in X$ and $a_i \in I$ (Halmos 1956). Define $X_n(w) = x_n$ and $A_n(w) = a_n$ as before. The transition function $p(z_1, dz_2)$ corresponding to independent choices $A_n = a \in I$ in (A.13) is

$$\Pr\big(X_{n+1} \in dy, A_{n+1} = b \mid X_n = x, A_n = a\big) \;=\; p_a(x, dy)w_b \qquad (A.22)$$

Then $A_1, A_2, \ldots$ are independent with $\Pr(A_i = a) = w_a$, and the conditional distribution of X_{n+1} given X_n and $A_n = a$ is given by (A.13) or (A.22).

It is easy to check that $\pi(x)dx\, w_a$ is a stationary measure for (A.22), and thus by Theorem A.1

$$\lim_{n\to\infty} \frac{f(X_0, A_0)(w) + f(X_1, A_1)(w) + \cdots + f(X_{n-1}, A_{n-1})(w)}{n} \;=\; g(w) \qquad (A.23)$$

converges a.s. whenever

$$E\big(|f(X_0, A_0)|\big) \;=\; \sum_a w_a E\big(|f(X_0, a)|\big) \;<\; \infty$$

Note that the left-hand side of (A.23) depends on A even if $f(x, a)$ does not depend on a, since the components A_i of A for $i < n$ determine the current

value of X_n. By using similar arguments as in the proof of Theorem A.4, we can show

Theorem A.5. Suppose that $\{ p_a(x, dy) \}$ is an ergodic family in the sense of Theorem A.3 or Theorem A.4. Then the limit in (A.23) is a.s. constant, and in particular equal to the constant $E(f) = \sum_a w_a C_a$ for $C_a = \int f(y, a)\pi(y)dy$.

Proof. If we define $\theta(w)$ on Ω^e by $\theta(\,(z_0, z_1, \ldots)\,) = (z_1, z_2, \ldots)$, then by Corollary A.1 the limit $g(w)$ in (A.23) satisfies $g(w) = g(\theta(w))$ a.s. Define

$$h(x, a) \;=\; E\big(\phi\big(g(w)\big) \mid X_0 = x, A_0 = a\big)$$

as in (A.16) for any bounded real function $\phi(r) \geq 0$. Then $h(x, a)$ is bounded since $\phi(y)$ is bounded, and, since $\phi(g(w)) = \phi(g(\theta(w)))$ a.s.,

$$\begin{aligned} h(x, a) \;&=\; \sum_b \int_X h(y, b) p_a(x, dy) w_b \\ &=\; \int_X \overline{h}(y) p_a(x, dy) \qquad (A.24) \end{aligned}$$

for $\overline{h}(x) = \sum_a w_a h(x, a)$. Since (A.24) is the same equation as (A.16), we conclude $\overline{h}(x) = C(g, \phi)$ a.s. by arguing as in (A.17)–(A.18), and by arguing as in (A.19)–(A.20) that $g(w) = C(g) = E(g) = E(f)$ a.s. This completes the proof of Theorem A.5.

A.6. Random-components-update MCMC. In Section 12, $T_a f(\omega) = f(\theta_a(\omega))$ is the result of a one-dimensional Metropolis-Hastings Markov-chain update of the a^{th} component of $X_n \in R^m$, treating the remaining components $x_{-a} \in R^{m-1}$ as parameters. The update is defined in such a way that the one-dimensional conditional density $\pi_i(x_i, x_{-i})$ is a stationary measure of the update. Random-component-update MCMC is defined by, at each time step, choosing a at random with probability $w_a > 0$ and applying that update.

This is exactly of the form described in the previous section, so that Theorem A.4 applies if the set $\{ \theta_a(\omega) \}$ is an ergodic family.

If we generalize (A.22) to

$$\Pr\big(X_{n+1} \in dy, A_{n+1} = b \mid X_n = x, A_n = a\big) \;=\; p_a(x, dy) w_b(y) \qquad (A.19)$$

then the proof of Theorem A.4 does not carry over, since in general

$$\sum_a \int \pi(x) w_a(x) p_a(x, dy) dx \neq \pi(y) dy$$

The arguments in the previous subsection would apply if we chose the a^{th} block of coordinates (or the a^{th} coordinate) with probability $w_a > 0$ and then a particular update of those coordinates with probability $w_{ab}(x_{-a})$, since then the update in the a^{th} block of coordinates would be ergodic by Theorem A.4.

References.

1. Chib, Siddhartha, and Edward Greenberg (1995) Understanding the Metropolis-Hastings algorithm. *American Statistician* **49,** 327–335.
2. Chen, Ming-Hui, Qi-Man Shao, and J. G. Ibrahim (2000) Monte Carlo methods in Bayesian computation. Springer Series in Statistics, Springer-Verlag.
3. Chung, Kai Lai (2001) A course in probability theory, 3rd edition. Academic Press.
4. DeGroot, M. (1989) Probability and statistics. Addison-Wesley.
5. Devroye, L. (1986) Non-uniform random variate generation. Springer-Verlag, New York.
6. Fishman, George S. (1995) Monte Carlo: Concepts, algorithms, and applications. Springer Series in Operations Research, Springer-Verlag.
7. Garsia, Adriano (1970) Topics in almost everywhere convergence. Markham Publishing.
8. Gelman, A. (1996) Inference and monitoring convergence. Chapter 8 in Gilks *et al.* (1996).
9. Gelman, A, J. Carlin, H. Stern, and D. Rubin (2003) Bayesian data analysis, 2nd edition. Chapman & Hall/CRC, Boca Raton.
10. Gilks, W. R., S. Richardson, and D. J. Spiegelhalter (1996) Markov chain Monte Carlo in practice. Chapman & Hall/CRC, Boca Raton.
11. Halmos, P. R. (1956) Lectures on ergodic theory. Chelsea, New York.
12. Hastings, W. K. (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57,** 97–109.
13. Kolmogorov, A. N. (1950) Foundations of the theory of probability. Chelsea Publishing, New York.
14. Liu, Jun S. (1998) Covariance adjustment for Markov chain Monte Carlo — A general framework and the covariance-adjusted data augmentation algorithm. *Technical Report*, Bell Laboratories, Lucent Technologies, reference by Chen *et al.* 2000.

15. Liu, Jun S. (2001) Monte Carlo strategies in scientific computing. Springer Series in Statistics, Springer-Verlag.

16. Liu, J., and C. Sabatti (2000) Generalized Gibbs sampler and multigrid Monte Carlo for Bayesian computation. *Biometrika* **87 (2),** 353–369.

17. Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller (1953) Equations of state calculations by fast computing machines. *J. Chem. Phys.* **21,** 1087–1092.

18. Von Neumann, J. (1951) Various techniques used in connection with random digits. *Monte Carlo Method*, Applied Mathematics Series 12, National Bureau of Standards, Washington, D.C.

19. Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery (1992) Numerical recipes in C: the art of scientific computing, 2nd edition. Cambridge University Press, Cambridge, England.

20. Riesz, F., and B. Sz.-Nagy (1955) Functional Analysis. Frederick Ungar Publishing, New York.

21. Sawyer, S. A. (1966) Maximal inequalities of weak type. *Annals of Math.* **84,** 157–174.

22. Tierney, L (1994) Markov chains for exploring posterior distributions. *Annals of Statistics* **22,** 1701–1728, with discussion 1728–1762.

23. Zygmund, A. (1959) Trigonometric series. Cambridge University Press.