18.465, Revised November 14, 2012

Rejection methods, Markov chain Monte Carlo, the Gibbs sampler, and ergodicity

In one dimension, to generate random variables $X_1, X_2, ...$ with a given distribution function F, one method is to define the function $F^{\leftarrow}(y) = \inf\{x : \ F(x) \geq y\}$ for $0 < y < 1$ and generate $U_1, ..., U_n$ i.i.d. $U[0,1]$. Then $X_j = F^{\leftarrow}(U_j)$ will be i.i.d. (F), as shown in the handout on the Kolmogorov and Kolmogorov–Smirnov tests.

1. The rejection method

In case $F^{\leftarrow}$ is hard to compute, or in dimension $d > 1$, we need another approach. The rejection method is a well-known way of generating random variables with a density f where it may be difficult to generate X_j with density f directly, but we know a density g such that for some $a < +\infty$, $f(x) \leq ag(x)$ for all x, and we can generate random variables $Y_1, Y_2, ...$ i.i.d. (g) relatively easily. The method is given in some beginning probability textbooks. It actually extends without difficulty to cases where we have a density $f(x)/c_0$ and don't know the normalizing constant c_0 (because it's hard to compute), as is pointed out e.g. in Devroye, 1986, Section II.3. We have the following, which holds in any dimension d. First to define a notation,

$$\int f(x)dx \ := \ \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, \ldots, x_d) dx_1 \cdots dx_d.$$

Theorem 1. *Let g be a probability density and let $f \geq 0$ be a function such that for some $a < \infty$, $f(x) \leq ag(x)$ for all x. Let $Y_1, ..., Y_n, ...$ be i.i.d. (g) and let $c_0 = \int f(x)dx > 0$ (the value of c_0 need not be known). Let $U_1, ..., U_n, ...$ be i.i.d. $U[0,1]$ variables independent of the Y_j. Given Y_j and U_j, accept Y_j as the next X_i if $U_j \leq f(Y_j)/(ag(Y_j))$, otherwise reject Y_j and go on to (Y_{j+1}, U_{j+1}). Then $X_1, \ldots, X_m, \ldots$ will be i.i.d. with density f/c_0.*

Proof. Clearly X_i are i.i.d. and $c_0 \leq a$. Let $A := \{y : \ g(y) = 0\}$. Then $P(Y_j \in A) = \int_A g(y)dy = 0$ for all j. So in the following we can assume $g(Y_j) > 0$ for all j, as this holds with probability 1.

Let $J = J(\omega)$ be the least j such that Y_j is not rejected. Then for any event (measurable set) B,

$$P(X_1 \in B) = P(Y_J \in B) = P(Y_j \in B | U_j \leq f(Y_j)/(ag(Y_j)))$$

(for any j, since the pairs (Y_j, U_j) are i.i.d.)

$$= \frac{P(Y_j \in B,\ U_j \leq f(Y_j)/(ag(Y_j)))}{P(U_j \leq f(Y_j)/(ag(Y_j)))}. \tag{1}$$

By independence of Y_j and U_j, the joint density of Y_j and U_j is

$$g(Y_j)1_{[0,1]}(U_j),$$

so the numerator of (1) equals

$$\int_B g(y) \cdot \frac{f(y)}{ag(y)} dy = \int_B f(y)dy/a$$

and the denominator

$$P\left(U_j \leq \frac{f(Y_j)}{ag(Y_j)}\right) = \int g(y) \cdot \frac{f(y)}{ag(y)} dy = \int f(y)dy/a = \frac{c_0}{a}, \tag{2}$$

so $P(X_1 \in B) = \int_B f(y)dy/c_0$ as desired. □

If m_n is the number of $j = 1, ..., n$ such that $U_j \leq f(Y_j)/(ag(Y_j))$, then by (2), m_n/n gives us an estimate of c_0/a, so am_n/n estimates c_0.

Often, however, we may have an f where we not only don't know the normalizing constant c_0 but may also not know any g we can use as above. For example, f may be defined on a rather high-dimensional space and/or we can evaluate it by some algorithm, but not by such a nice formula as would make it possible to see good densities g such that $f \leq ag$ for some a. Then there's another method as follows.

2. The Metropolis–Hastings algorithm

2.1. **The discrete case.** The basic paper by W. K. Hastings (1970) describes a method in some detail for discrete state spaces. In the title of his paper one can see what are now buzz words, "Markov chain[s]...Monte Carlo."

Hastings (1970, p. 99) gives the following. Suppose given a countable state space indexed by positive integers, say, on which we have some $\pi_j \geq 0$ such that $0 < S := \sum_j \pi_j < \infty$ but where it may be hard to compute S, or at any rate we don't need to assume S known to generate random variables with distribution having approximately the probability mass function π_j/S. We can and do assume all $\pi_j > 0$, otherwise eliminate j with $\pi_j = 0$ from the state space. Let $q_{ij} \geq 0$ be a Markov transition matrix, i.e. $\sum_j q_{ij} = 1$ for each i, and let it be chosen so that $q_{ii} \equiv 0$ and $q_{ij} > 0$ for all $i \neq j$. The statistician can choose $\{q_{ij}\}$. It should be chosen so that for each i, one can easily generate a random variable J such that $\Pr(J = j) = q_{ij}$ for all $j \neq i$. We want to find a Markov transition matrix p_{ij} for which $\{\pi_j\}_{j=1}^{\infty}$, or equivalently $\{\pi_j/S\}_{j=1}^{\infty}$, is an invariant measure for p_{ij}, namely for each j,

$$\sum_i \pi_i p_{ij} = \pi_j. \tag{3}$$

Hastings defines $p_{ij} = q_{ij}\alpha_{ij}$ for $i \neq j$ where

$$\alpha_{ij} \;=\; \frac{s_{ij}}{1 + (\pi_i q_{ij})/(\pi_j q_{ji})} \tag{4}$$

and s_{ij} are numbers such that $s_{ij} \equiv s_{ji} > 0$ and $0 < \alpha_{ij} \leq 1$ for all $i \neq j$. Such s_{ij} always exist, for example Hastings points out that one could take simply $s_{ij} \equiv 1$, although he doesn't recommend that choice and the "Metropolis–Hastings" choice is different.

Once p_{ij} are defined for $i \neq j$ then we define $p_{ii} := 1 - \sum_{j \neq i} p_{ij}$, so we get a Markov transition matrix. It can happen that for some i, $\alpha_{ij} = 1$ for all j, and then $p_{ii} = 0$. For example, if the state space is finite and we let π_i decrease down toward 0 while other π_j remain the same, eventually all α_{ij} will equal 1, if s_{ij} are chosen to maximize α_{ij}.

The following fact may be called "detailed balance" in the discrete case.

Claim 1: With definitions as above, we have $\pi_i p_{ij} = \pi_j p_{ji}$ for all i and j.

Proof. This is clear for $i = j$. For $i \neq j$, multiply α_{ij} from (4) by $\pi_i q_{ij}$. Multiply the numerator and denominator of the resulting expression by $\pi_j q_{ji}$. Doing the same with i and j interchanged gives the same result. In more detail, we get

$$\pi_i p_{ij} = \alpha_{ij}\pi_i q_{ij} = \frac{s_{ij}\pi_i q_{ij}}{1 + \frac{\pi_i q_{ij}}{\pi_j q_{ji}}} = \frac{s_{ij}\pi_j q_{ji}\pi_i q_{ij}}{\pi_i q_{ij} + \pi_j q_{ji}}$$

where the last expression is symmetric under interchanging i and j because $s_{ij} \equiv s_{ji}$. Thus, it equals the preceding expressions with i and j interchanged, giving

$$\pi_i p_{ij} = \pi_i q_{ij}\alpha_{ij} = \pi_j q_{ji}\alpha_{ji} = \pi_j p_{ji},$$

proving the claim. □

Now, (3) does hold because $\sum_i \pi_i p_{ij} = \sum_i \pi_j p_{ji} = \pi_j \sum_i p_{ji} = \pi_j$. Since all $q_{ij} > 0$ for $i \neq j$ and we chose s_{ij} such that all $\alpha_{ij} > 0$ for $i \neq j$, we will have $p_{ij} > 0$ for all $i \neq j$.

One can generate $X_1, X_2, ...,$ as follows: let $X_1 = Y_1$ be arbitrary in the state space. Let $U_1, U_2, ...,$ be i.i.d. $U[0,1]$. Given $X_n = i$, generate a Y_{n+1} such that $\Pr(Y_{n+1} = j) = q_{ij}$ (recall that q_{ij} are chosen so that this is easy). Suppose $Y_{n+1} = j$. If $U_{n+1} \leq \alpha_{ij}$ then set $X_{n+1} = Y_{n+1}$, otherwise reject Y_{n+1} and set $X_{n+1} = X_n$. Then $\{X_n\}_{n\geq 1}$ form a Markov chain with transition probabilities p_{ij}.

To choose α_{ij} specifically in (4), Hastings (1970) gave some options for defining s_{ij}. One he gave with a superscript (M), referring to Metropolis, is as follows: for $i \neq j$,

$$s_{ij} = 1 + \min\left(\frac{\pi_i q_{ij}}{\pi_j q_{ji}}, \frac{\pi_j q_{ji}}{\pi_i q_{ij}}\right). \tag{5}$$

(For $i = j$, s_{ii} and α_{ii} need not be defined, as in defining p_{ij}, s_{ij} and α_{ij} are only used for $i \neq j$.) This choice results in a relatively simple form of the acceptance probability α_{ij}. From (5), (4) and some algebra, one can see that $\alpha_{ij} = 1$ if $\pi_i q_{ij} \leq \pi_j q_{ji}$ and otherwise $\alpha_{ij} = \pi_j q_{ji}/(\pi_i q_{ij})$. Here are details. We have for any $i \neq j$, $\pi_i q_{ij} \leq \pi_j q_{ji}$ if and only if

$$\pi_i q_{ij}/(\pi_j q_{ji}) \leq 1 \leq \pi_j q_{ji}/(\pi_i q_{ij}).$$

It follows that $\alpha_{ij} = 1$ if $\pi_i q_{ij} \leq \pi_j q_{ji}$. Otherwise it equals

$$\left(1 + \frac{\pi_j q_{ji}}{\pi_i q_{ij}}\right) \Big/ \left(1 + \frac{\pi_i q_{ij}}{\pi_j q_{ji}}\right) = \frac{\pi_i q_{ij} + \pi_j q_{ji}}{\pi_i q_{ij}} \cdot \frac{\pi_j q_{ji}}{\pi_j q_{ji} + \pi_i q_{ij}} = \frac{\pi_j q_{ji}}{\pi_i q_{ij}}$$

as claimed. In other words we have

$$\alpha_{ij} \ = \ \min\left(1, \frac{\pi_j q_{ji}}{\pi_i q_{ij}}\right). \tag{6}$$

If $X_n = i$, one makes the move from i to j if and only if it is proposed according to q_{ij} and for a $U[0,1]$ variable U_k independent of those used previously, $U_k \leq \pi_j q_{ji}/(\pi_i q_{ij})$. (Thus when the right side is ≥ 1, the move is accepted with probability 1.) For the choice (5) of s_{ij} one only needs to know the π_j and q_{ij} to find α_{ij} directly by (6). The Metropolis–Hastings choice (5) makes s_{ij} and thus α_{ij} as large as possible subject to the given requirements, because we need $s_{ij} \leq 1 + \pi_i q_{ij}/(\pi_j q_{ji})$ in order that $\alpha_{ij} \leq 1$, and since $s_{ij} \equiv s_{ji}$ we also need that $s_{ij} \leq 1 + \pi_j q_{ji}/(\pi_i q_{ij})$. So this choice maximizes the probability that moves are accepted, under the requirements.

2.2. **The continuous case.** Here the state space will be a non-discrete subset U of $\mathbb{R}^d$. Hastings considers the continuous case only briefly, in terms of approximations from the discrete case. For the continuous case, this exposition is based on other references, especially Tierney (1998).

Suppose given $P(x)$ with $P(x) \geq 0$ for all $x \in \mathbb{R}^d$ and $0 < c_0 = \int_{\mathbb{R}^d} P(x)dx < \infty$ where again we don't know c_0. Let $U = \{x : P(x) > 0\}$. Let $Q(x; y) \geq 0$ be a transition density, i.e. for each x,

$\int_{\mathbb{R}^d} Q(x;y)dy = 1$, with for each $x \in U$, $Q(x;y) > 0$ for all $y \in U$. Although sets U such as half-spaces can be of interest, for simplicity the rest of this exposition will assume that U is all of $\mathbb{R}^d$.

We want to be able to sample easily from $Q(x;\cdot)$. For example, $Q(x;\cdot)$ may be the multivariate normal $N(x, \sigma^2 I)$ for some $\sigma > 0$, which may be adjusted as one goes along (see suggestions in the Wikipedia article). We want to find a function $\alpha(x,y) \geq 0$, the analogue of α_{ij} in the discrete case (4), such that

$$P(x)Q(x;y)\alpha(x,y) \ \equiv \ P(y)Q(y;x)\alpha(y,x). \tag{7}$$

Let

$$\alpha(x,y) \ := \ \frac{s(x,y)}{1 + (P(x)Q(x;y))/(P(y)Q(y;x))} \tag{8}$$

where $s(x,y) \equiv s(y,x) > 0$ is chosen so that $0 < \alpha(x,y) \leq 1$. Then (7) will hold by the same proof as for Claim 1.

For each $x \in \mathbb{R}^d$, define a measure with respect to y which has a continuous part and a point mass at x, namely

$$P(x;dy) \ := \ Q(x;y)\alpha(x,y)dy + \delta_x(y) \int (1 - \alpha(x,u))Q(x;u)du. \tag{9}$$

Then what is called "detailed balance," in our notation

$$P(x)dx\, P(x;dy) \ = \ P(y)dy\, P(y;dx) \tag{10}$$

as measures on $\mathbb{R}^d \times \mathbb{R}^d = \mathbb{R}^{2d}$, holds because each side has a continuous part with density given by either side of (7) with respect to $dx\, dy$ on $\mathbb{R}^{2d}$, and another part concentrated in the "diagonal" d-dimensional hyperplane $x = y$, or $\{(x,x) : \ x \in \mathbb{R}^d\}$, having density there $P(x) \int (1 - \alpha(x,u))Q(x;u)du$ with respect to dx, or equivalently replacing x by y.

To show the continuous analogue of (3), namely that $P(x)dx$ is invariant under the transition probability $P(x;dy)$, with $f = 1_B$ the indicator function of an event B, by (10),

$$\int \int_{y \in B} P(x;dy)P(x)dx \ = \ \int_{y \in B} \int P(y;dx)P(y)dy \ = \ \int_B P(y)dy$$

as desired.

For a Markov chain with values in $\mathbb{R}^d$ and transition probability $P(x;dy)$, take any X_1 and, given X_n, choose Y_{n+1} at random from $Q(X_n;\cdot)$, recalling that we've chosen Q to make this easy, and set $X_{n+1} = Y_{n+1}$ if an independent $U[0,1]$ variable $U_{n+1} \leq \alpha(X_n, Y_{n+1})$, otherwise $X_{n+1} = X_n$, and so on.

To make the specific Metropolis–Hastings choice of $s(x, y)$ and thus $\alpha(x, y)$ in the continuous case, there is a natural analogue of (5),

$$s(x, y) = 1 + \min\left(\frac{P(x)Q(x; y)}{P(y)Q(y; x)}, \frac{P(y)Q(y; x)}{P(x)Q(x; y)}\right). \tag{11}$$

This implies

$$\alpha(x, y) = \min\left(1, \frac{P(y)Q(y; x)}{P(x)Q(x; y)}\right), \tag{12}$$

just analogous to (6) in the discrete case.

3. Ergodicity of some Markov chains

3.1. The discrete case.

Let V be a countable state space, represented by positive integers. The case $V = \{1, 2, ..., M\}$ with $M < \infty$ is included, where we'll assume $M \geq 3$, but we'll be mainly interested in the case that V is infinite and consists of all the positive integers. Sums $\sum_j := \sum_{j \in V}$ will then naturally mean $\sum_{j=1}^{M}$ in the finite case and $\sum_{j=1}^{\infty}$ in the infinite case.

Any array $P = \{P_{ij}\}_{i,j \in V}$ of numbers $P_{ij} \geq 0$ will be called a *Markov transition matrix* if for each $i \in V$, $\sum_{j=1}^{\infty} P_{ij} = 1$. For two such matrices P and Q, the *product* is defined, as for finite matrices, by $PQ \equiv P \cdot Q$ where $(PQ)_{ik} = \sum_j P_{ij} Q_{jk}$ for all $i, k \in V$. Then it's easily checked that PQ is itself a Markov transition matrix and that this multiplication is associative. Let $P^1 := P$ and recursively for $n \geq 1$ define $P^{n+1} := P^n \cdot P$, which equals $P \cdot P^n$ by associativity applied as many times as needed. Let $P_{ij}^n := (P^n)_{ij}$ for any $n = 1, 2, ...$ and $i, j \in V$. (Powers $(P_{ij})^n$ of individual entries are not usually of interest.)

A Markov transition matrix P is called *irreducible* iff for any i and j in V there is some $n = 1, 2, ...$ such that $P_{ij}^n > 0$. This clearly holds if $P_{ij} > 0$ for all $i \neq j$, with $n = 2$.

A finite signed measure μ on V is given by a sequence $\{\mu_j\}_{j \in V}$ such that $\sum_j |\mu_j| < \infty$. Then μ is called *invariant* for a Markov transition matrix P if for each $j \in V$, $\sum_{i \in V} \mu_i P_{ij} = \mu_j$. If $\mu_j \geq 0$ for all $j \in V$ then μ is called a *measure*, and a *probability measure* if also $\sum_j \mu_j = 1$. The following will be rather easy to prove:

Theorem 2. *If an irreducible Markov transition matrix P has an invariant probability measure μ then μ is unique.*

Proof. Suppose $\mu \neq \nu$ are two invariant probability measure measures for P. Then both are also invariant for any power P^n of P, and so is the nonzero finite signed measure $\mu - \nu$. Let $\rho_j := \max(0, \mu_j - \nu_j) \geq 0, > 0$

if and only if j is in a non-empty set J, and $\tau_i := -\min(0, \mu_i - \nu_i) \geq 0$, > 0 if and only if i is in a non-empty set I, disjoint from J. Then ρ and τ are non-zero finite measures with $\mu - \nu = \rho - \tau$ (called the Jordan decomposition of $\mu - \nu$). Let $T := \sum_j \rho_j = \sum_i \tau_i$. Then $0 < T \leq 1$. For each j let $\rho'_j := \sum_i \rho_i P_{ij} \geq 0$ and $\tau'_j := \sum_i \tau_i P_{ij} \geq 0$. Then $\sum_j \rho'_j = T = \sum_j \tau'_j$ and $\mu - \nu = \rho - \tau = \rho' - \tau'$. If there is any i with $\rho'_i > 0$ and $\tau'_i > 0$ then $|\rho'_i - \tau'_i| < \rho'_i + \tau'_i$, whereas $|\rho_i - \tau_i| \equiv \rho_i + \tau_i$, so $\sum_i |\rho'_i - \tau'_i| < \sum_i |\rho_i - \tau_i| = 2T$, a contradiction. Thus $\rho = \rho'$ and $\tau = \tau'$, i.e. ρ and τ are invariant for P and thus for P^n for any n. But for any $i \in I$ and $j \in J$ there is an n with $P^n_{ij} > 0$, so $\rho_j > 0$, another contradiction, and μ is unique. □

Corollary 1. *For a discrete state space indexed by positive integers j and numbers $\pi_j > 0$ such that $S = \sum_j \pi_j < \infty$, and any Markov transition matrix $P = \{p_{ij}\}_{i,j\geq 1}$ satisfying Hastings' conditions for $\{\pi_j\}$, the unique stationary probability measure for P is π_j/S.*

Proof. π_j/S is an invariant probability for P by equation (3), which follows from Claim 1 and the paragraph after its proof, and P is irreducible because $p_{ij} > 0$ for all $i \neq j$. □

A state $i \in V$ will be called *periodic of period* k for a Markov transition matrix P and integer $k \geq 2$ iff $P^n_{ii} > 0$ for some $n \geq 1$ and all such n are divisible by k. P is called *aperiodic* iff it has no periodic states of any period ≥ 2. Clearly P is aperiodic if $P_{ij} > 0$ for all i and j, or if $M \geq 3$ as we assume and $P_{ij} > 0$ for all $i \neq j$. Thus the Markov transition matrices $\{p_{ij}\}$ satisfying Hastings' conditions will be aperiodic. For them, we can see directly that $N = 2$ in the following.

Lemma 1. *Let P be an aperiodic Markov transition matrix and $i \in V$ any state. Let $A := \{n \geq 1 : \ P^n_{ii} > 0\}$. Then for some N, A contains all $n \geq N$.*

Proof. By associativity, for any $m \in A$ and $n \in A$ we have $m + n \in A$, i.e. A is an *additive semigroup* of positive integers. By the aperiodicity, the greatest common divisor of all $n \in V$ is 1, so by the Euclidean algorithm, known since antiquity, there exist some $r < \infty$, some $m_1, ..., m_r$ in A, and some integers $k_i \in \mathbb{Z}$ (some of which may be negative, and some must be positive) such that $\sum_{i=1}^r k_i m_i = 1$. Let $K := \sum_{i=1}^r |k_i| m_i \in A$ and also $K^2 \in A$. Let $s = K^2 + tK + j$ for any integers $t \geq 0$ and $j = 0, 1, ..., K-1$. Then

$$s = \sum_{i=1}^{r} (jk_i + (K+t)|k_i|) m_i \in A$$

since $K+t \geq K > j$ so for each i, $(K+t)|k_i| + jk_i > 0$ and the lemma holds with $N = K^2$. □

Next is a further theorem, whose proof is not so elementary. It is given in Kallenberg (1997), Theorem 7.18. It holds for the p_{ij} satisfying Hastings' conditions, applied to the invariant probability measure with probability at j equal to π_j/S.

Theorem 3. *Let P be an irreducible, aperiodic Markov transition matrix with an invariant probability measure π. Then for any $i \in V$, $\sum_{j\in V} |P^n_{ij} - \pi_j| \to 0$ as $n \to \infty$. Further, for any probability measure μ on V, $\sum_j |\sum_i \mu_i P^n_{ij} - \pi_j| \to 0$ as $n \to \infty$.*

3.2. **Ergodicity in the continuous case.** Now let U be a continuous state space, specifically the Euclidean space $\mathbb{R}^d$ (or a non-discrete subset of it). Let $\delta_x(A) = 1_A(x) = 1$ for $x \in A$ and 0 otherwise. A *Markov transition kernel* will be a function $x \mapsto M_x$ from U into probability measures on the Borel sets of (events included in) U, where M_x is of the form

$$M_x(dy) \;=\; M(x;y)dy + \left(1 - \int M(x;u)du\right)\delta_x(dy),$$

$M(x;y) \geq 0$ for all x and y in U, $M_{(x)} := \int M(x;u)du \leq 1$ for all x, and $M(\cdot,\cdot)$ is a jointly Borel measurable function of its two arguments (this hypothesis is made just so that all integrals to be written are defined). Let $M_{\{x\}} := 1 - M_{(x)}$. Thus M_x is a continuous measure with a density, except for a point mass of size $M_{\{x\}}$ at x. The probability of a Borel set A is given by

$$(13) \qquad M_x(A) \;=\; \int_A M(x;y)dy + M_{\{x\}}\delta_x(A).$$

The Metropolis–Hastings definition (9) of $P(x;dy)$ does give a Markov transition kernel $P_x(dy)$ with $P(x;y) = Q(x;y)\alpha(x,y)$, where in (9), $\int Q(x;u)du = 1$, so the coefficient of $\delta_x(y)$ there is indeed equal to $1 - P_{(x)} = P_{\{x\}}$ as desired. In this case we will have typically $P_{\{x\}} > 0$ so that a non-zero point mass is present. Whereas, in the Gibbs sampling case to be treated in Section 4, we will have $\int Q(x;u)du = 1$ and the kernel has a density, $Q_x(dy) = Q(x;y)dy$.

The analogue of the product of two Markov transition matrices is the following product operation: let P_x, $x \in U$, and Q_y, $y \in U$, be two Markov transition kernels. Let

$$(P \cdot Q)_x(dz) \;:=\; \int Q_y(dz)P_x(dy),$$

in other words, for any Borel set $B \subset \mathbb{R}^d$ and $x \in \mathbb{R}^d$, define

$$\begin{aligned}
(P \cdot Q)_x(B) &:= \int Q_y(B) P_x(dy) \\
&= P_{\{x\}} Q_x(B) + \int Q_y(B) P(x;y) dy \\
&= P_{\{x\}} Q_{\{x\}} \delta_x(B) + P_{\{x\}} \int_B Q(x;z) dz \\
&\quad + \int_B P(x;y) Q_{\{y\}} dy + \int \int_B P(x;y) Q(y;z) dy dz,
\end{aligned}$$

using that $\int \delta_y(B) \cdots dy = \int_B \cdots dy$. Letting $R := P \cdot Q$, we get that R_x is indeed a Markov transition kernel, with $R_{\{x\}} = P_{\{x\}} Q_{\{x\}}$ and

$$R(x;z) := P_{\{x\}} Q(x;z) + P(x;z) Q_{\{z\}} + \int P(x;y) Q(y;z) dy.$$

Thus, starting at x and using first P, then Q for transitions, one either stays at x with probability $P_{\{x\}} Q_{\{x\}}$, or one moves to a point $z \neq x$ with a sub-probability density $R(x;z)$, which can be done in just one of three mutually exclusive ways (except for probability 0 events): either one first stays at x with probability $P_{\{x\}}$, then transitions to z with sub-probability density $Q(x;z)$; or, one goes directly to z with sub-probability density $P(x;z)$ and stays there with probability $Q_{\{z\}}$; or one goes first to a point y different from x or z with subprobability density $P(x;y)$, then from y to z with subprobability density $Q(y;z)$, and we integrate $\int P(x;y) Q(y;z) dy$ to get the total contribution to $R(x;z)$ from going via such y.

If $Q = P$ then we get the kernel $P^2 := P \cdot P$ and so by recursion any power of P. A probability measure μ given by a density $\pi(x) \geq 0$ on $\mathbb{R}^d$ will be called *invariant* for the Markov transition kernel P if for any Borel set B, $\int_B \pi(x) dx = \int \pi(x) P_x(B) dx$.

A Markov transition kernel M_x defined for $x \in U \subset \mathbb{R}^d$ will be called *U-transitive* if U is a Borel subset of $\mathbb{R}^d$ with volume $\int_U dx > 0$ and for each $x \in U$, $M(x;y) > 0$ for all $y \in U$ and $M(x;y) = 0$ for all $y \notin U$. The kernels defined by the Metropolis–Hastings method (in the continuous case) are U-transitive, but those in the Gibbs sampler (Section 4) in general may not be. We have the following:

Theorem 4. *Let P_x be a Markov transition kernel such that for some Borel set $U \subset \mathbb{R}^d$ P_x is U-transitive. Suppose P has an invariant probability measure μ on U having a density. Then μ is unique.*

The proof is very similar to that of Theorem 2. It follows from the facts and proof in Hernández-Lerma and Lasserre (2003), Lemma 2.2.3 pp. 26-27 and Proposition 4.2.2 p. 48.

Now, we'd like a continuous analogue of Theorem 3. Let ϕ be a σ-finite measure, which in our case will just be d-dimensional Lebesgue measure (volume) λ^d on $\mathbb{R}^d$ or a subset U of it with $\phi(U) > 0$, so we can write

$$d\phi(x) \ = \ \phi(dx) \ = \ d\lambda^d(x) \ = \ dx \ = \ dx_1 dx_2 \cdots dx_d.$$

A Markov transition kernel P on U, where $P_x(U) = 1$ for all $x \in U$, is called *ϕ-irreducible* if for each $x \in U$ and each $A \subset U$ with volume $\phi(A) > 0$ we have for some n $P_x^n(A) > 0$. For a U-transitive kernel we can take $n = 1$ and get $P_x(A) \geq \int_A P(x; y)dy > 0$ since $P(x; y) > 0$ for all $y \in U$, where "$\geq$" becomes equality if $x \notin A$ and usually $>$ if $x \in A$. Thus all Markov transition kernels defined by the Metropolis–Hastings method for the continuous state space U are ϕ-irreducible for $\phi = \lambda^d$ restricted to U.

A Markov transition kernel P is called *periodic* if for some $k \geq 2$ there exist disjoint sets $A_1, ..., A_k$ such that for $j = 1, ..., k-1$, $P_x(A_{j+1}) = 1$ for all $x \in A_j$, and $P_x(A_1) = 1$ for all $x \in A_k$. Otherwise P is called *aperiodic*. If P_x is U-transitive and $k \geq 2$, each set A_j would need to have positive volume, in order for $P_x(A_j)$ to equal 1 for some $x \notin A_j$, but then since $P_x(A_j) > 0$ for all x and all j, we'd get $P_x(A_j) < 1$ for all x, a contradiction, so $k = 1$ and P is aperiodic.

In preparation for the next theorem, the total variation distance will be defined. If μ and ν are two probability measures, then we have the Jordan decomposition $\mu - \nu = (\mu - \nu)^+ - (\mu - \nu)^-$ where $(\mu - \nu)^+$ and $(\mu - \nu)^-$ are nonnegative finite measures, such that there exists a set A with $(\mu - \nu)^+(A) = 0$ and $(\mu - \nu)^-(\mathbb{R}^d \setminus A) = 0$, and the *total variation distance* $\|\mu - \nu\|_{TV}$ between μ and ν is defined as $2(\mu - \nu)^+(\mathbb{R}^d)$. If μ and ν each have densities f and g respectively, then $\|\mu - \nu\|_{TV} = \int |f - g| d\phi$. In the following theorem, however, P_x^n doesn't quite have a density in general because it has a point mass at x, with a coefficient $P_{\{x\}}^n$ which becomes small geometrically as n becomes large.

Theorem 5. *If the Markov transition kernel P_x, $x \in U$ on U is λ^d-irreducible and aperiodic, it has a unique invariant probability μ on U, and for μ-almost all x, $\|P_x^n - \mu\|_{TV} \to 0$ as $n \to \infty$.*

This follows from Theorem 4 above and Theorem 1 of Tierney (1994, p. 1712). Tierney says it results from facts proved in Nummelin (1984) and had been stated previously by Athreya, Doss and Sethuraman (1992).

Next is a "pathwise ergodic theorem."

Theorem 6. *Let P_x, $x \in U$, be a Markov transition kernel having a unique invariant probability μ, meaning that*

$$\mu(A) = \int P_x(A) d\mu(x)$$

for each Borel set (event) $A \subset U$ and no other probability measure ν in place of μ has this property. Let $(x_1, x_2, ...)$ be a Markov chain with transition probabilities P_x, in other words, choose any x_1 in U, or let x_1 have any given probability distribution in U, and for each $n = 1, 2, ...,$, given x_n, let x_{n+1} have distribution P_{x_n}. Let f be an integrable function for μ, i.e. $\int f(x)d\mu(x)$ is defined and finite. Then for almost all choices of x_1 with respect to μ, and then $x_2, x_3, ...$ with their given distributions,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} f(x_j) = \int f(x) d\mu(x). \tag{14}$$

This theorem is essentially Corollary 2.5.2, p. 38, of Hernández-Lerma and Lasserre (2003). Tierney (1994, Theorem 3) also mentions this fact and gives as a reference Theorem 3.6 of Chapter 4 of Revuz (1975).

For P_x of the Metropolis–Hastings form, there are no possible bad choices of x_1. In fact, for any m and choices of $x_1, ..., x_m$, (14) will still hold with probability 1 for choices of $x_{m+1}, x_{m+2}, ...$, with their given distributions.

Tierney (1994, Corollary 2, p. 1713) points out that Theorem 5 applies to Markov chains generated by the Metropolis–Hastings method. In other words, starting at any $x = x_0$, the probability distribution of the point x_n reached after n steps with the Metropolis–Hastings algorithm converges in total variation to μ, so that for n large, we can generate x_n with distribution approximately μ. If we want a sequence of random variables with distribution approximately μ (although not independent), we can take $x_m, x_{m+1}, ...$ for some large enough m, and not use x_j for $j < m$ which form the so-called "burn-in" stage (as mentioned e.g. in the Wikipedia article).

4. The Gibbs sampler

For $x = (x_1, ..., x_d) \in \mathbb{R}^d$ and $j = 1, ..., d$ let $x_{(j)} := \{x_i\}_{1 \le i \le d,\ i \ne j}$ $= (x_1, ..., x_{j-1}, x_{j+1}, ..., x_d) \in \mathbb{R}^{d-1}$. In Gibbs sampling, we assume there is a probability density f on $\mathbb{R}^d$ such that for each $j = 1, ..., d$, we know the conditional density $f(x_j | x_{(j)})$ and are able to generate samples with this density. In fact $f(x_j | x_{(j)}) = f(x_1, ..., x_d)/f_{(j)}(x_{(j)})$ where $f_{(j)}$

is the marginal density of $x_{(j)}$, namely $f_{(j)}(x_{(j)}) = \int f(x)dx_j$, but we may know the density f only to within a multiplicative constant, which divides out in forming the conditional densities.

One version of the Gibbs sampling procedure is as follows. Suppose we know a set $U \subset \mathbb{R}^d$ of positive volume such that $f(x) > 0$ if and only if $x \in U$. Choose an $x^{(0)} \in U$ such that the set (typically an interval) of x_1 such that $(x_1, x^{(0)}_{(1)}) \in U$ has length (Lebesgue measure) > 0, as will always be true if U is open. Generate a new value $x_1^{(1)}$ from the conditional distribution of x_1 given $x^{(0)}_{(1)}$. Then successively generate $x_j^{(1)}$ for $j = 2, ..., d$ from the conditional distribution of x_j given $x_{(j)} = y$ where $y_i = x_i^{(1)}$ for $i < j$ and $y_i = x_i^{(0)}$ for $i > j$. After finishing the step with $j = d$, return to $j = 1$ and repeat the process to generate $x^{(2)}$ from $x^{(1)}$, and so on.

The probability distribution of $x^{(1)}$ given $x^{(0)} = x$ is then a Markov transition kernel P_x, moreover one with a density, since in Gibbs sampling moves are always accepted, so there is 0 probability of remaining at the same point. Moreover, the conditional distributions each have densities by the assumptions. We have the following:

Theorem 7. *The probability distribution on $\mathbb{R}^d$ with density f is invariant under the kernel $P_x,\ x \in U$, defined by Gibbs sampling.*

Proof. Suppose $x^{(0)}$ has density f. Then $x^{(0)}_{(1)}$ has the marginal density given by f, and $x_1^{(1)}$ has the correct conditional distribution given $x^{(0)}_{(1)}$, so $(x_1^{(1)}, x^{(0)}_{(1)})$ has density f. Likewise for each $j = 2, ..., d$, inductively,

$$y := (x_1^{(1)}, ..., x_{j-1}^{(1)}, x_{j+1}^{(0)}, x_d^{(0)})$$

has the marginal density given by f, and $x_j^{(1)}$ the f conditional density given y, so $(\{x_i^{(1)}\}_{i \leq j}, \{x^{(0)}_{(i)}\}_{i>j})$ will have density f. For $j = d$ this gives the conclusion. □

We'd like the Gibbs Markov chain to be ϕ-irreducible where ϕ is d-dimensional Lebesgue measure λ^d (volume) on $U \subset \mathbb{R}^d$. This is not true for arbitrary sets U with positive volume. For example, let $U = ([0, 1] \times [0, 1]) \cup ([1, 2] \times [1, 2]) \subset \mathbb{R}^2$ where $\times$ is the Cartesian product, $A \times B := \{(x, y) :\ x \in A, y \in B\}$. [According to context, (x, y) can mean an ordered pair as in this last definition, or (a, b) can mean the open interval $\{x :\ a < x < b\}$.] Then whichever of the two squares the Gibbs Markov chain starts in, it will remain in it for all n with probability 1. One assumption mentioned somewhere is that U

should be open and connected. Actually connectedness is not necessary, e.g. if $U = ((0,1) \times (0,1)) \cup ((0,1) \times (2,3))$ then U is not connected but one can access either square from the other when choosing y for any $x \in (0,1)$.

Theorem 4 says that if a Markov transition kernel is U-transitive for some U, it has a unique invariant probability measure. For the Metropolis–Hastings procedure, since $Q(x;y)$ was relatively arbitrary, it was easy to make it > 0 everywhere on U, but that isn't the case with Gibbs sampling, where we may need more iterations to reach all parts of U.

We get using Theorem 7:

Theorem 8. *If the Markov transition kernel P_x, $x \in U$, given by the Gibbs sampler is λ^d-irreducible, then the distribution with density f is the unique invariant probability measure for P_x.*

It follows by Theorem 6 that the pathwise ergodic theorem holds for Markov chains generated by the Gibbs sampler under the given condition. Also, we'd like the distribution of $x^{(n)}$ given by the Gibbs sampler to approach that given by f in total variation. For this we need that the P_x of the Gibbs sampler is aperiodic. Supposing that U is open, then for any $x \in U$, there is some $r > 0$ small enough such that all y with $|y-x| < r$ are also in U, and the one-step density $P(x;y) > 0$ for all such y. For the definition of periodic in the continuous case as given in subsection 3.2 if the Gibbs chain were periodic and $x \in A_1$, then almost all $y \neq x$ with $|x-y| < r/2$ would have to be in A_2, where "almost all" means except for a set with 0 volume. Moreover, for each such y there must be a $\delta_y > 0$ with $\delta_y < r/2$ such that almost all z with $|z-y| < \delta_y$ must be in A_3 if $k \geq 3$ or in A_1 if $k = 2$, either of which would be a contradiction since also $z \in A_2$ but $A_1 \cap A_2 = \emptyset$ and if A_3 is defined, $A_2 \cap A_3 = \emptyset$. So the Gibbs P_x is aperiodic and by Theorem 5, the distribution of $x^{(n)}$ for Gibbs sampling converges in total variation to the distribution with density f.

Suppose for example that $d = 2$, so we're given conditional densities $f(y|x)$ and $f(x|y)$. Then we have $f(x,y) \equiv f(y|x)g(x)$ where g is the marginal density of x, and $f(x,y) \equiv f(x|y)h(y)$ where h is the marginal density of y. It seems to follow that $f(y|x)/f(x|y) \equiv h(y)/g(x)$, but the conclusion doesn't make sense for x and y such that $f(x|y) = 0$, which can happen even though $h(y) > 0$ and $g(x) > 0$. Supposing U is a rectangle $(a,b) \times (u,v)$ in $\mathbb{R}^2$ with $a < b$ and $u < v$, then if $f(y|x)$ and $f(x|y)$ are given explicitly, we should be able to factor $f(y|x)/f(x|y)$ in the form $h(y)/g(x)$ and thus determine the marginal densities, up to a multiplicative constant in each.

In dimension $d = 2$ for non-rectangular U, or for $d \geq 3$, however, there seems not to be such a direct approach to finding the form of the joint marginal densities of $d - 1$ variables.

5. Posterior densities

The handouts bayestopix.pdf and morebayes.pdf from the Spring 2012 18.443 website www-math.mit.edu/~rmd/443S12 give material on Bayesian statistics. Posterior densities provide a large and important class of densities whose normalizing constants may be difficult to compute and yet for which it's desirable to sample from the distributions with these densities.

Let's recall some basics of Bayesian statistics and estimation. Suppose given a family of probability distributions parametrized by a continuous finite-dimensional parameter θ, and given by a likelihood function $f(\theta, x)$ for one observation x, where $f(\theta, \cdot)$ is a probability mass function in case x is discrete, or a density function if x is continuous.

There are cases where the parameters of distributions are discrete, such as hypergeometric distributions, but often even for discrete distributions, the parameters are continuous, such as the probability p for the binomial or geometric distribution or the parameter λ for the Poisson distribution.

Let θ range over a space Θ of possible values. Let $\pi(\theta)$ be a probability density defined for $\theta \in \Theta$, called the *prior* density.

Given $X_1, \ldots, X_n$ assumed to be i.i.d. $f(\theta, \cdot)$ for some unknown θ, let X be the vector $(X_1, \ldots, X_n)$, so that the likelihood function based on X is $f(\theta, X) = \prod_{j=1}^n f(\theta, X_j)$. Then the *posterior* density of θ is

$$\pi_X(\theta) = \frac{\pi(\theta) f_X(\theta)}{\int \pi(\phi) f_X(\phi) d\phi}. \tag{15}$$

Integrating the numerator first with respect to each X_j and last with respect to θ we have

$$\int_\Theta \pi(\theta) \int f(\theta, x_1) dx_1 \cdots \int f(\theta, x_n) dx_n d\theta = \int_\Theta \pi(\theta) \cdot 1 \cdot 1 \cdots 1 \, d\theta = 1,$$

where the integrals dx_j are replaced by sums over x_j if the x_j are discrete. It follows, interchanging integrals (for nonnegative functions) that with probability 1, the integral in the denominator of (15) is finite and strictly positive, so the posterior density π_X exists as a probability density.

If the prior π happens to be in what's called a "conjugate prior" family for θ and the family $f(\theta, \cdot)$ of distributions, such as beta priors for the binomial p or gamma priors for the Poisson λ, then the normalizing constants of posterior densities are easy to evaluate and the posterior densities belong to the same conjugate prior family. Such cases are emphasized in first courses in statistics. But in general, we may not be so lucky as to have a conjugate prior and there is not such an easy way to find the normalizing constants.

A class of problems in Bayesian statistics is to find Bayes estimators. Let $g(\theta)$ be a real-valued function of the unknown θ. A Bayes estimator of $g(\theta)$ is a statistic, namely a real-valued function $T(X)$ of the vector X of observations, which minimizes the mean-square error

$$\int_\Theta E_\theta(T(X) - g(\theta))^2 \pi(\theta) d\theta,$$

where E_θ denotes expectation when θ is the true value of the parameter, so that $X_1, \ldots, X_n$ are i.i.d. $f(\theta, \cdot)$. A theorem (proved in the 18.443 handout bayestopix.pdf) says that whenever a Bayes estimator exists, it is unique and given by the integral of $g(\theta)$ with respect to the posterior density,

$$T(X) = \int g(\theta)\pi_X(\theta)d\theta. \tag{16}$$

Many books say that $T(X)$ is the conditional expectation $E(g(\theta)|X)$ of $g(\theta)$ given X, which is correct if the unconditional expectation $\int g(\theta)d\pi(\theta)$ is defined and finite, but it may not be and (16) can still apply. For example, suppose we have a family $N(\mu, 1)$ of normal distributions with μ unknown and and are very uncertain about μ, so we assume a Cauchy prior distribution for it allowing possibly large values. The expectation of μ is undefined for a Cauchy distribution. We may naturally want to estimate $g(\mu) = \mu$ after observing $X = (X_1, ..., X_n)$. Because of the normal form of the likelihood function, μ will have a finite integral with respect to the posterior distribution π_X for any X.

Now, suppose we are unable to evaluate the normalizing constant for π_X. Still, using MCMC (Markov chain Monte Carlo), i.e. the Metropolis–Hastings algorithm in the continuous case, as long as $\pi(\theta)$ and $f(\theta, X)$ are reasonably easy to evaluate for each θ and our observed X, we can generate $\theta_1, \ldots, \theta_m$ whose distribution for m large is approximately π_X, even if θ_j are not even approximately i.i.d., but by the pathwise ergodic theorem, Theorem 6, if g is integrable for π_X, we can approximately evaluate $T(X)$ by the average $\frac{1}{m}\sum_{i=1}^m g(\theta_i)$.

NOTES

The paper by Tierney (1994), although its title emphasizes posterior distributions as in Section 5, gives in addition a general exposition of the application of Markov chain theory to Markov chain Monte Carlo. The paper has been very influential, having been cited over 2700 times according to Google Scholar (11/2012).

I have not seen the preprint by Athreya, Doss and Sethuraman (1992), but very possibly from the coincidence of the three authors and similarity of titles, the final form of the paper appeared in 1996 as listed below.

REFERENCES

Athreya, K. B., Doss, H., and Sethuraman, J. (1992), A proof of convergence of the Markov chain simulation method, Technical Report 868, Dept. of Statistics, Florida State Univ. (reference from Tierney, 1994).

Athreya, K. B., Doss, H., and Sethuraman, J. (1996), On the convergence of the Markov chain simulation method, *Ann. Statist.* **24**, 69-100.

Devroye, Luc, *Non-Uniform Random Variate Generation*, Springer-Verlag, 1986, available free of charge at
http://luc.devroye.org/rnbookindex.html
and also advertised on Amazon.

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

Hernández-Lerma, Onésimo, and Lasserre, Jean Bernard (2003), *Markov Chains and Invariant Probabilities*, Birkhäuser, Basel.

Kallenberg, O. (1997), *Foundations of Modern Probability*, Springer, London.

Nummelin, E. (1984), *General Irreducible Markov Chains and Non-Negative Operators*, Cambridge University Press.

Revuz, D. (1975), *Markov Chains*, North–Holland, Amsterdam. Cited by Tierney (1994).

Tierney, Luke (1994), Markov chains for exploring posterior distributions, *Ann. Statist.* **22**, 1701-1728.

Tierney, Luke (1998), A note on Metropolis–Hastings kernels for general state spaces, *Ann. Applied Probability* **8**, 1-9.