Some "Simple" Theory of Markov Chain Monte Carlo

It is sometimes the case that one has a "known" but unwieldy (joint) distribution for a large dimensional random vector $Y$, and would like to know something about the distribution (like, for example, the marginal distribution of $Y_1$). While in theory it might be straightforward to compute the quantity of interest, in practice the necessary computation can be impossible. With the advent of cheap computing, given appropriate theory the possibility exists of simulating a sample of $n$ realizations of $Y$ and computing an empirical version of the quantity of interest to approximate the (intractable) theoretical one.

This possibility is especially important to Bayesians, where $X|\theta \sim P_\theta$ and $\theta \sim G$ produce a joint distribution for $(X, \theta)$ and then (in theory) a posterior distribution $G(\theta|X)$, the object of primary Bayesian interest. Now (in the usual kind of notation)

\[ g(\theta|X = x) \propto f_\theta(x)g(\theta) \]

so that in this sense the posterior of the possibly high dimensional $\theta$ is "known." But it may well be analytically intractable, with (for example) no analytical way to compute $\mathbb{E} \theta_1$. Simulation is a way of getting (approximate) properties of an intractable/high dimensional posterior distribution.

"Straightforward" simulation from an arbitrary multivariate distribution is, however, typically not (straightforward). Markov Chain Monte Carlo methods have recently become very popular as a solution to this problem.

Discrete Cases

The following is some simple theory covering the case where the (joint) distribution of $Y$ is discrete.

**Definition 1** A (discrete time/discrete state space) Markov Chain is a sequence of random quantities $\{X_k\}$, each taking values in a (finite or) countable set $\mathcal{X}$, with the property that

\[ P[X_n = x_n|X_1 = x_1, ..., X_{n-1} = x_{n-1}] = P[X_n = x_n|X_{n-1} = x_{n-1}] \]

**Definition 2** A Markov Chain is stationary provided $P[X_n = x|X_{n-1} = x']$ is independent of $n$.

WOLOG we will henceforth name the elements of $\mathcal{X}$ with the integers $1, 2, 3, ...$ and call them "states."

**Definition 3** With $p_{ij} = P[X_n = j|X_{n-1} = i]$, the square matrix $P = (p_{ij})$ is called the transition matrix for a stationary Markov Chain and the $p_{ij}$ are called transition probabilities.

---

1 The following exposition is based on old ISU lecture notes obtained from Noel Cressie and on Luke Tierney’s December 1994 *Annals of Statistics* paper.
Note that a transition matrix has nonnegative entries and row sums of 1. Such matrices are often called “stochastic” matrices. As a matter of further notation for a stationary Markov Chain, let
\[ p^k_{ij} = P[X_{n+k} = j | X_n = i] \]
and
\[ f^k_{ij} = P[X_{n+k} = j, X_{n+k-1} \neq j, \ldots, X_{n+1} \neq j | X_n = i] \] .
(These are respectively the probabilities of moving from \( i \) to \( j \) in \( k \) steps and first moving from \( i \) to \( j \) in \( k \) steps.)

**Definition 4** We say that a MC is irreducible if for each \( i \) and \( j \) \( \exists k \) (possibly depending upon \( i \) and \( j \)) such that \( p^k_{ij} > 0 \).

(A chain is irreducible if it is possible to eventually get from any state \( i \) to any other state \( j \).)

**Definition 5** We say that the \( i \)th state of a MC is transient if \( \sum_{k=1}^{\infty} f^k_{ii} < 1 \) and say that the state is persistent if \( \sum_{k=1}^{\infty} f^k_{ii} = 1 \). A chain is called persistent if all of its states are persistent.

(A state is transient if once in it, there is some possibility that the chain will never return. A state is persistent if once in it, the chain will with certainty be in it again.)

**Definition 6** We say that state \( i \) of a MC has period \( t \) if \( p^k_{ii} = 0 \) unless \( k = \nu t \) (\( k \) is a multiple of \( t \)) and \( t \) is the largest integer with this property. The state is aperiodic if no such \( t > 1 \) exists. And a MC is called aperiodic if all of its states are aperiodic.

Many sources (including Chapter 15 of the 3rd Edition of Feller Volume 1) present a number of useful simple results about MC’s. Among them are the following.

**Theorem 7** All states of an irreducible MC are of the same type (with regard to persistence and periodicity).

**Theorem 8** A finite state space irreducible MC is persistent.

**Theorem 9** Suppose that a MC is irreducible, aperiodic and persistent. Suppose further that for each state \( i \) the mean recurrence time is finite, i.e.
\[ \sum_{k=1}^{\infty} k f^k_{ii} < \infty \] .

Then an invariant/stationary distribution for the MC exists, i.e. \( \exists \{u_j\} \) with \( u_j > 0 \) and \( \sum u_j = 1 \) such that
\[ u_j = \sum_i u_i p_{ij} \] .
(If the chain is started with distribution \(\{u_j\}\), after one transition it is in states 1, 2, 3,... with probabilities \(\{u_j\}\).) Further, this distribution \(\{u_j\}\) satisfies

\[ u_j = \lim_{k \to \infty} p_{ij}^k \forall i , \]

and

\[ u_j = \frac{1}{\sum_{k=1}^{\infty} kf_{jj}^k} . \]

There is a converse of this theorem.

**Theorem 10** An irreducible, aperiodic MC for which \(\exists \{u_j\}\) with \(u_j > 0\) and \(\sum u_j = 1\) such that \(u_j = \sum_i u_ip_{ij}\) must be persistent with \(u_j = \frac{1}{\sum_{k=1}^{\infty} kf_{jj}^k}\).

And there is an important “ergodic” result that guarantees that “time averages” have the right limits.

**Theorem 11** Under the hypotheses of Theorem 9, if \(g\) is a real-valued function such that \(\sum_j |g(j)|u_j < \infty\)

then for any \(j\), if \(X_0 = j\)

\[ \frac{1}{n} \sum_{k=1}^{n} g(X_k) \xrightarrow{a.s.} \sum_j g(j)u_j \]

(Note that the choice of \(g\) as an indicator provides approximations for stationary probabilities.)

With this background, the basic idea of MCMC is the following. If we wish to simulate from a distribution \(\{u_j\}\) or approximate properties of the distribution that can be expressed as moments of some function \(g\), we find a convenient MC \(\{X_k\}\) whose invariant distribution is \(\{u_j\}\). From a starting state \(X_0 = i\), one uses \(P\) to simulate \(X_1\). Using the realization \(X_1 = x_1\) and \(P\), one simulates \(X_2\), etc. One applies Theorem 11 to approximate the quantity of interest. Actually, it is common practice to use a “burn in” of \(m\) periods before starting the kind of time average indicated in Theorem 11. Two important questions about this plan (only one of which we’ll address) are:

1. How does one set up an appropriate/convenient \(P\)?

2. How big should \(m\) be? (We’ll not touch this matter.)

In answer to question 1), there are presumably a multitude of chains that would do the job. But there is the following useful sufficient condition (that has application in the original motivating problem of simulating from high dimensional distributions) for a chain to have \(\{u_j\}\) for an invariant distribution.
Lemma 12 If \( \{X_k\} \) is a MC with transition probabilities satisfying
\[
u_ip_{ij} = u_jp_{ji},
\]
then it has invariant distribution \( \{u_j\} \).

Note then that if a candidate \( P \) satisfies (1) and is irreducible and aperiodic, Theorem 10 shows that it is persistent and Theorem 9 then shows that any arbitrary starting value can be used and yields approximate realizations from \( \{u_j\} \) and Theorem 11 implies that “time averages” can be used to approximate properties of \( \{u_j\} \).

Several useful MCMC schemes can be shown to have the “correct” invariant distributions by noting that they satisfy (1).

For example, Lemma 12 can be applied to the “Metropolis-Hastings Algorithm.” That is, let \( T = (t_{ij}) \) be any stochastic matrix corresponding to an irreducible aperiodic MC. (Presumably, some choices will be better that others in terms of providing quick convergence to a target invariant distribution. But that issue is beyond the scope of the present exposition.) (Note that in a finite case, one can take \( t_{ij} = 1/(\text{the number of states}) \).) The Metropolis-Hastings algorithm simulates as follows:

- Supposing that \( X_{n-1} = i \), generate \( J \) (at random) according to the distribution over the state space specified by row \( i \) of \( T \), that is according to \( \{t_{ij}\} \).
- Then generate \( X_n \) based on \( i \) and (the randomly generated) \( J \) according to
\[
X_n = \begin{cases} 
J & \text{with probability } \min \left( 1, \frac{u_J t_{ji}}{u_i t_{ij}} \right) \\
i & \text{with probability } \max \left( 1, 1 - \frac{u_J t_{ji}}{u_i t_{ij}} \right)
\end{cases}
\]

Note that for \( \{X_k\} \) so generated, for \( j \neq i \)
\[
p_{ij} = P[X_n = j | X_{n-1} = i] = \min \left( 1, \frac{u_J t_{ji}}{u_i t_{ij}} \right) t_{ij}
\]
and
\[
p_{ii} = P[X_n = i | X_{n-1} = i] = t_{ii} + \sum_{j \neq i} \max \left( 0, 1 - \frac{u_J t_{ji}}{u_i t_{ij}} \right) t_{ij}
\]
So, for \( i \neq j \)
\[
u_i p_{ij} = \min(u_i t_{ij}, u_j t_{ji}) = u_j p_{ji}
\]
That is, (1) holds and the MC \( \{X_k\} \) has stationary distribution \( \{u_j\} \). (Further, the assumption that \( T \) corresponds to an irreducible aperiodic chain implies that \( \{X_k\} \) is irreducible and aperiodic.)

Notice, by the way, that in order to use the Metropolis-Hastings Algorithm one only has to have the \( u_j \)'s up to a multiplicative constant. That is, if we have
we may compute the ratios $u_j/u_i$ as $v_j/v_i$ and are in business as far as simulation goes. (This fact is especially attractive in Bayesian contexts where it can be convenient to know a posterior only up to a multiplicative constant.) Notice also that if $T$ is symmetric, (i.e. $t_{ij} = t_{ji}$), step (2) of the Metropolis algorithm becomes

$$X_n = \begin{cases} J & \text{with probability } \min \left(1, \frac{v_j}{v_i} \right) \\ i & \text{with probability } \max \left(1, 1 - \frac{v_i}{v_j} \right) \end{cases}$$

An algorithm similar to the Metropolis-Hastings algorithm is the “Barker Algorithm.” The Barker algorithm modifies the above by replacing $\min \left(1, \frac{u_j t_{ji}}{u_i t_{ij}} \right)$ with $\frac{u_j t_{ji}}{u_i t_{ij} + u_j t_{ji}}$ in (2). Note that for this modification of the Metropolis algorithm, for $j \neq i$

$$p_{ij} = \left( \frac{u_j t_{ji}}{u_i t_{ij} + u_j t_{ji}} \right) t_{ij}.$$  

so

$$u_i p_{ij} = \frac{(u_i t_{ij})(u_j t_{ji})}{u_i t_{ij} + u_j t_{ji}} = u_j p_{ji}.$$ 

That is, (1) holds and thus Lemma 12 guarantees that $\{X_k\}$ has invariant distribution $\{u_j\}$. (And $T$ irreducible and aperiodic continues to imply that $\{X_k\}$ is also.)

Note also that since

$$\frac{u_j t_{ji}}{u_i t_{ij} + u_j t_{ji}} = \frac{u_i t_{ji} + u_j t_{ji}}{u_i t_{ij} + u_j t_{ji}} \frac{u_j t_{ji}}{u_i t_{ij} + u_j t_{ji}}$$

once again it suffices to know the $u_j$ up to a multiplicative constant in order to be able to implement Barker’s algorithm. (This is again of comfort to Bayesians.) Finally, note that if $T$ is symmetric, step (2) of the Barker algorithm becomes

$$X_n = \begin{cases} J & \text{with probability } \frac{u_j}{u_i + u_j} \\ i & \text{with probability } \frac{u_i}{u_i + u_j} \end{cases}$$

Finally, consider now the “Gibbs Sampler,” or “Successive Substitution Sampling,” for generating an observation from a high dimensional distribution. For sake of concreteness, consider the situation where the distribution of a discrete 3-dimensional random vector $(Y, V, W)$ with probability mass function $f(y, v, w)$ is at issue. In the general MC setup, a possible outcome $(y, v, w)$ represents a “state” and the distribution $f(y, v, w)$ is the “$\{u_j\}$” from which one wants to simulate. One defines a MC $\{X_k\}$ as follows. For an arbitrary starting state $(y_0, v_0, w_0)$:
the chain is not stationary. The transition mechanism is not the same for a
this problem is to note that if 
P is irreducible and aperiodic, Theorem 10 says that the chain
is invariant. So one is in the position to apply Theorems 10 and 9. If 

\[ P(Y, V, W) \]

where \( a \) describes an entire cycle of the SSS algorithm. Then

\[ P(Y, V, W) \]

• Generate \( X_1 = (Y_1, V_1, W_1) \) by generating \( Y_1 \) from the conditional distribution of \( V = Y_1 \) and \( W = W_0 \), i.e. from the (conditional) distribution with probability function

\[ f_{Y|V,W}(y_1, w_0) = \frac{f(y_1, w_0)}{\sum_l f(l, y_1, w_0)} \]

Let \( y_1 \) be the realized value of \( Y_1 \).

• Generate \( X_2 = (y_1, V_2, W_2) \) by generating \( V_1 \) from the conditional distribution of \( V = Y_1 \) and \( W = W_0 \), i.e. from the (conditional) distribution with probability function

\[ f_{V|Y,W}(v_1, w_0) = \frac{f(v_1, w_0)}{\sum_l f(l, v_1, w_0)} \]

Let \( v_1 \) be the realized value of \( V_1 \).

• Generate \( X_3 = (y_1, v_1, W_3) \) by generating \( W_1 \) from the conditional distribution of \( W = Y_1 \) and \( V = V_1 \), i.e. from the (conditional) distribution with probability function

\[ f_{W|Y,V}(w_1, v_1) = \frac{f(w_1, v_1)}{\sum_l f(l, w_1, v_1)} \]

Let \( w_1 \) be the realized value of \( W_1 \).

• Generate \( X_4 = (Y_2, v_1, W_4) \) by generating \( Y_2 \) from the conditional distribution of \( V = V_1 \) and \( W = W_1 \), i.e. from the (conditional) distribution with probability function

\[ f_{Y|V,W}(y_2, w_1) = \frac{f(y_2, w_1)}{\sum_l f(l, y_2, w_1)} \]

Let \( y_2 \) be the realized value of \( Y_2 \).

• And so on, for some appropriate number of cycles.

Note that with this algorithm, a typical transition probability (for a step where a \( Y \) realization is going to be generated) is

\[ P[X_n = (y', v, w) | X_{n-1} = (y, v, w)] = \frac{f(y', v, w)}{\sum_l f(l, v, w)} \]

so if \( X_{n-1} \) has distribution \( f \), the probability that \( X_n = (y, v, w) \) is

\[ \sum_{l'} f(l', v, w) \frac{f(y, v, w)}{\sum_l f(l, v, w)} = f(y, v, w) \]

that is, \( X_n \) also has distribution \( f \). And analogous results hold for transitions of all 3 types (\( Y, V \) and \( U \) realizations).

But it is not absolutely obvious how to apply the results 9 and 10 here since the chain is not stationary. The transition mechanism is not the same for a “\( Y \)” transition as it is for a “\( V \)” transition. The key to extricating oneself from this problem is to note that if \( P_Y, P_V \) and \( P_W \) are respectively the transition matrices for \( Y, V \) and \( W \) exchanges, then

\[ P = P_Y P_V P_W \]

describes an entire cycle of the SSS algorithm. Then \( \{X'_k\} \) defined by \( X'_n = X_{3n} \) is a stationary Markov Chain with transition matrix \( P \). The fact that \( P_Y, P_V \) and \( P_W \) all leave the distribution \( f \) invariant then implies that \( P \) also leaves \( f \) invariant. So one is in the position to apply Theorems 10 and 9. If \( P \) is irreducible and aperiodic, Theorem 10 says that the chain \( \{X'_k\} \) is persistent and then Theorems 9 and 11 say that \( f \) can be simulated using an arbitrary starting state.
General Cases

The non discrete $Y$ version of all this is somewhat more subtle, but in many ways parallel to the exposition given above. The following may not be completely understandable to students without a Stat 642 background.

Consider now a (possibly non discrete) state space $X$ and a target distribution $\pi$ on $(X, \mathcal{B})$ (for which one wishes to approximate some property via simulation). Suppose that $P(x, A) : X \times \mathcal{B} \rightarrow [0,1]$ is a “transition kernel” (a regular conditional probability, a function such that for each $x$, $P(x, \cdot)$ is a probability measure, and for each $A$, $P(\cdot, A)$ is $\mathcal{B}$ measurable). Consider a stochastic process $\{X_k\}$ with the property that given $X_1 = x_1, ..., X_{n-1} = x_{n-1}$, the variable $X_n$ has distribution $P(x_{n-1}, \cdot)$. $\{X_k\}$ is a general stationary discrete time MC with transition kernel $P(x, A)$, and provided one can simulate from $P(x, A)$, one can simulate realizations of $\{X_k\}$. If $P(x, A)$ is properly chosen, empirical properties of a realization of $\{X_k\}$ can approximate theoretical properties of $\pi$.

Notice that the $n$-step transition kernels can be computed recursively beginning with $P(x, A)$ using

$$P^n(x, A) = \int P^{n-1}(y, A)dP(x, y)$$

**Definition 13** $P$ has invariant distribution $\pi$ means that $\forall A \in \mathcal{B}$

$$\pi(A) = \int P(x, A)d\pi(x)$$

**Definition 14** $P$ is called $\pi$-irreducible if for each $x$ and each $A$ with $\pi(A) > 0$, $\exists n(x, A) \geq 1$ such that $P^{n(x, A)}(x, A) > 0$.

**Definition 15** A $\pi$-irreducible transition kernel $P$ is called periodic if $\exists$ an integer $d \geq 2$ and a sequence of sets $A_0, A_1, ..., A_{d-1}, A_d = A_0$ in $\mathcal{B}$ such that for $i = 0, 1, ..., d-1$ and all $x \in A_i$, $P(x, A_{i+1}) = 1$. If a $\pi$-irreducible transition kernel $P$ is not periodic it is called aperiodic.

**Definition 16** A $\pi$-irreducible transition kernel $P$ with invariant distribution $\pi$ is called Harris recurrent if the corresponding MC, $\{X_k\}$, has the property that $\forall A \in \mathcal{B}$ with $\pi(A) > 0$ and all $x \in X$

$$P_{X_0=x}[X_n \in A \text{ infinitely often}] = 1$$

**Theorem 17** Suppose that transition kernel $P$ has invariant distribution $\pi$, is $\pi$-irreducible, aperiodic and Harris recurrent. If $g$ is such that

$$\int |g(x)|d\pi(x) < \infty$$

then for $X_0 = x$ any element of $X$

$$\frac{1}{n} \sum_{k=1}^{n} g(X_k) \xrightarrow{a.s.} \int g(x)d\pi(x)$$
Theorem 17 says that integrals of \( \pi \) (like, for example, probabilities and moments) can be approximated by long run sample averages of a sample path of the MC \( \{X_k\} \) beginning from any starting value, provided a suitable \( P \) can be identified. “Suitable” means, of course, that it has the properties hypothesized in Theorem 17 and one can simulate from it knowing no more about \( \pi \) than is known in practice. The game of MCMC is then to identify usable \( P \)’s. As Tierney’s article lays out, the general versions of SSS, the Barker algorithm and the Metropolis-Hastings algorithm (and others) can often be shown to satisfy the hypotheses of Theorem 17.

The meaning of “SSS” in the general case is obvious (one generates in turn from each conditional of a coordinate of a multidimensional \( Y \) given all other current coordinates and considers what one has at the end of each full cycle of these simulations). To illustrate what another version of MCMC means in the general case, consider the Metropolis-Hastings algorithm.

Suppose that \( Q(x, A) \) is some convenient transition kernel with the property that for some \( \sigma \)-finite measure \( \mu \) on \( (\mathcal{X}, \mathcal{B}) \) (dominating \( \pi \)) and a nonnegative function \( q(x, y) \) on \( \mathcal{X} \times \mathcal{X} \) (that is \( \mathcal{B} \times \mathcal{B} \) measurable) with \( \int_X q(x, y) d\mu(y) = 1 \) for all \( x \),

\[
Q(x, A) = \int_A q(x, y) d\mu(y)
\]

Let

\[
\frac{d\pi}{d\mu} = p
\]

and suppose that \( \pi \) is not concentrated on a single point. Then with

\[
\alpha(x, y) = \begin{cases} 
\min \left( \frac{p(y)q(y,x)}{p(x)q(x,y)}, 1 \right) & \text{if } p(x)q(x,y) > 0 \\
1 & \text{if } p(x)q(x,y) = 0
\end{cases}
\]

and

\[
p(x, y) = \begin{cases} 
q(x, y)\alpha(x, y) & \text{if } x \neq y \\
0 & \text{if } x = y
\end{cases}
\]

a Metropolis-Hastings transition kernel is

\[
P(x, A) = \int_A p(x, y) d\mu(y) + \left( 1 - \int p(x, y) d\mu(y) \right) \Delta_x(A)
\]

where \( \Delta_x \) is a probability measure degenerate at \( x \). The “simulation interpretation” of this is that if \( X_{n-1} = x \), one generates \( y \) from \( Q(x, \cdot) \) and computes \( \alpha(x, y) \). Then with probability \( \alpha(x, y) \), one sets \( X_n = y \) and otherwise takes \( X_n = x \). Note that as in the discrete version, the algorithm depends on \( p \) only through the ratios \( p(y)/p(x) \) so it suffices to know \( p \) only up to a multiplicative constant, a fact of great comfort to Bayesians wishing to study analytically intractable posteriors.

It is straightforward to verify that the kernel defined by (3) is \( \pi \)-invariant. It turns out that sufficient conditions for the kernel \( P \) to be \( \pi \)-irreducible are
that $Q$ is $\pi$-irreducible and either that i) $q(x, y) > 0 \forall x$ and $y$, or that ii) $q(x, y) = q(y, x) \forall x$ and $y$. In turn, $\pi$-irreducibility of $P$ is sufficient to imply Harris recurrence for $P$. And, if $P$ is $\pi$-irreducible, a sufficient condition for aperiodicity of $P$ is that

$$\pi \left( \{ x \mid \int p(x, y) d\mu(y) < 1 \} \right) > 0$$

so there are some fairly concrete conditions that can be checked to verify the hypotheses of Theorem 17.