Sādhanā Vol. 31, Part 2, April 2006, pp. 81–104. © Printed in India

On the Markov Chain Monte Carlo (MCMC) method

RAJEEVA L KARANDIKAR

7, SJS Sansanwal Marg, Indian Statistical Institute, New Delhi 110 016, India
e-mail: rlk@isid.ac.in

Abstract. Markov Chain Monte Carlo (MCMC) is a popular method used to generate samples from arbitrary distributions, which may be specified indirectly. In this article, we give an introduction to this method along with some examples.

Keywords. Markov chains; Monte Carlo method; random number generator; simulation.

1. Introduction

In this article, we give an introduction to Monte Carlo techniques with special emphasis on Markov Chain Monte Carlo (MCMC). Since the latter needs Markov chains with state space that is $\mathbb{R}$ or $\mathbb{R}^d$ and most text books on Markov chains do not discuss such chains, we have included a short appendix that gives basic definitions and results in this case.

Suppose X is a random variable (with known distribution, say with density f) and we are interested in computing the expected value

$$E[g(X)] = \int g(x) f(x) \mathrm{d}x \tag{1}$$

for a given function g. If the functions f, g are such that the integral in (1) cannot be computed explicitly (as a formula for the indefinite integral may not be available in closed form) then we can do as follows.

Assuming that we can generate a random sample from the distribution of X, generate a random sample of size n:

$$x_1, x_2 \ldots x_n,$$

from this distribution and compute

$$a_n = \frac{1}{n} \sum_{i=1}^{n} g(x_i).$$

Then by the law of large numbers, a_n approximates $E[g(X)]$.

Moreover, the central limit theorem gives the order of error; the error here is of the order of

$$O(n^{-\frac{1}{2}}).$$

As of now we have not said anything about the random variable X – it could be taking values in $\mathbb{R}$ or $\mathbb{R}^d$ for any dimension d. The important thing to note is that *the order of error does not depend upon the dimension.* This is very crucial if d is high as most of the numerical analysis techniques do not fare well in higher dimension. This technique of generating $x_1, x_2, \ldots x_n$ to approximate quantities associated with the distribution of X is called *Monte Carlo simulation* or just *Simulation*.

The most crucial part of the procedure described above is the generation of a random sample from the distribution of X. We deal here with the case where the state space is a subset of $\mathbb{R}$ or $\mathbb{R}^d$ and the distribution is given by its density f.

We assume that we have access to a "good" random number generator which gives us a way of generating a random sample from Uniform (0,1). For example, one could use the Mersenne Twister random number generator (see http://www.math.keio.ac.jp/∼matumoto/emt.html).

There is a canonical way of generating a univariate random variable from any distribution F: Let F^{-1} be the "inverse" of F and let U be a sample from Uniform (0,1). Then $X = F^{-1}(U)$ is a random sample from F.

Often, the distribution is described by a density and a closed form for the distribution function F is not available, and so the method described above fails. Another method is "transformation" of variables: Thus, if we have a method to generate $N(0, 1)$ random variables, then to generate a sample from t distribution with k degrees of freedom, we can generate independent samples $X, Y_1, \ldots, Y_k$ from $N(0, 1)$ and then take

$$t = X \Big/ \left(\sum_{j=1}^{k} Y_j^2 / k \right)^{1/2}$$

Indeed, the common method to generate samples from $N(0, 1)$ also uses the idea of transformation of variables: Generate U, V from Unifrom (0,1), U, V independent) and define

$$X = [-2\log(U)]^{1/2} \cos(2\pi V)$$
$$Y = [-2\log(U)]^{1/2} \sin(2\pi V).$$

Then X, Y are independent samples from $N(0, 1)$.

Of course, transformation of variables is a powerful method, but given a distribution, it is not clear how to use this method. So even when density is known we may have difficulty in generating samples from the distribution corresponding to it.

There are many situations where f may not be explicitly known but is described indirectly. For example, f may be known only upto a normalizing constant. Another possibility is that the distribution of interest is a multivariate distribution that is not known, but all the conditional distributions are specified.

Suppose we know $f_1(x) = Kf(x)$ but do not explicitly know K. Of course, K equals the integral of f_1. Numerically computing K and then proceeding to numerically compute $\int g(x) f(x) \mathrm{d}x$ can inflate the error. It would be much better if just knowing f_1, we can device a scheme to generate random sample from $f = (1/K) f_1$ and thereby compute $\int g(x) f(x) \mathrm{d}x$ approximately.

Bayesian framework: Suppose that given θ, X has a density $p(x \mid \theta)$ and the *prior* on θ is given by a density $\pi(\theta)$. Then the *posterior* density $\pi(\theta \mid x)$ of θ given an observation $X = x$ is given by

$$\pi(\theta \mid x) = [p(x \mid \theta)\pi(\theta)] \Big/ \left[\int p(x \mid \theta)\pi(\theta) \mathrm{d}\theta \right].$$

Often it is difficult to obtain exact expression for

$$\int p(x \mid \theta)\pi(\theta)\mathrm{d}\theta,$$

but given $p(x \mid \theta), \pi(\theta)$ we know $\pi(\theta \mid x)$ upto a normalizing constant!

Note that once there is a method to generate samples from the posterior density, there is no need for a practitioner to restrict the choice of prior to a conjugate prior (roughly, these are priors for which the the posterior can be computed in closed form). Even though the posterior density of θ may not be available in closed form, all quantities of interest could be obtained by simulation.

2. Rejection sampling

In 1951 von Neumann gave a method to generate samples from a density $f = (1/K)f_1$ knowing only f_1 if there is a density h such that (it is possible to generate samples from h and)

$$f_1(x) \leq Mh(x) \quad \forall x.$$

The algorithm, given below, is known as *rejection sampling*.

(1) Generate a random sample from the distribution with density h. Let it be y.
(2) Accept y as the sample with probability $[f_1(y)/Mh(y)]$.
(3) If step (2) is not a success, then go to step (1).

Here f (or f_1) is called the target density and $h(x)$ is called a majorizing function or an envelope or in some contexts the proposal density. We can repeat the steps (1)–(3) several times to generate i.i.d. samples from f.

The algorithm can be described as follows (as pseudo-code). (Here and in the sequel, successive "calls" to the random number generator are assumed to yield independent observations.) The following algorithm generates a random sample $z_1, \ldots, z_N$ from the distribution f.

Rejection sampler algorithm:

```
i = 0
do
{
      i = i + 1
      k = 0
      do
        k = k + 1;
        Generate u_k from Uniform (0,1) and x_k from h
      } while (u_k > f_1(x_k)/(Mh(x_k)))
      z_i = x_k
} while i ≤ N
```

We will now prove that the algorithm described above yields a random sample from f.

Theorem 1 (Rejection sampling). *Suppose we are given f_1, such that $f_1(x) = Kf(x)$ for a density f. Suppose there exists a density $h(x)$ and a constant M such that*

$$f_1(x) \leq Mh(x) \ \ \forall x. \tag{2}$$

Let X_k be i.i.d. with common density h, U_k be i.i.d. Uniform (0,1). Let B be given by

$$B = \{(x, u) : \ u \leq f_1(x)/Mh(x)\}$$

and τ be the first m such that $(X_m, U_m) \in B$ and let $W = X_\tau$. Then W has density f.

Proof. Take $Z_k = (X_k, U_k)$. Note that

$$\begin{aligned} P(\tau = m) &= P(Z_1 \notin B, \ldots Z_{m-1} \notin B, Z_m \in B) \\ &= P(Z_1 \notin B)^{m-1} P(Z_m \in B) \\ &= (1 - P(Z_1 \in B))^{m-1} P(Z_1 \in B), \end{aligned}$$

and hence $P(\tau < \infty) = 1$. Now

$$\begin{aligned} P(Z_m \in A \mid \tau = m) &= P(Z_m \in A \mid Z_1 \notin B, \ldots Z_{m-1} \notin B, Z_m \in B) \\ &= P(Z_m \in A \mid Z_m \in B) \\ &= P(Z_1 \in A \mid Z_1 \in B), \end{aligned}$$

and hence

$$\begin{aligned} P(Z_\tau \in A) &= \sum_m P(Z_m \in A \mid \tau = m) P(\tau = m) \\ &= \sum_m P(Z_1 \in A \mid Z_1 \in B) P(\tau = m) \\ &= P(Z_1 \in A \mid Z_1 \in B). \end{aligned}$$

Taking $A = (-\infty, a] \times [0, 1]$ for $a \in \mathbb{R}$, we have (using $\{Z_\tau \in A\} = \{W \leq a\}$),

$$\begin{aligned} P(W \leq a) &= P(X_1 \leq a \mid Z_1 \in B) \\ &= \frac{P(X_1 \leq a, Z_1 \in B)}{P(Z_1 \in B)} \\ &= \left\{ \int_{-\infty}^{a} \int_0^1 1_B(x, u) h(x) \mathrm{d}u \mathrm{d}x \right\} \Big/ \left\{ \int_{-\infty}^{\infty} \int_0^1 1_B(x, u) h(x) \mathrm{d}u \mathrm{d}x \right\} \\ &= \left\{ \int_{-\infty}^{a} \frac{f_1(x)}{Mh(x)} h(x) \mathrm{d}x \right\} \Big/ \left\{ \int_{-\infty}^{\infty} \frac{f_1(x)}{Mh(x)} h(x) \mathrm{d}x \right\} \\ &= \left\{ \int_{-\infty}^{a} f_1(x) \mathrm{d}x \right\} \Big/ \left\{ \int_{-\infty}^{\infty} f_1(x) \mathrm{d}x \right\} \\ &= \int_{-\infty}^{a} f(x) \mathrm{d}x. \end{aligned}$$

We have used $\int_0^1 1_B(x,u)\mathrm{d}u = f_1(x)/Mh(x)$ and also that f_1 is proportional to the density f. This completes the proof.

Let us examine what happens if we use the rejection sampling algorithm when the envelope condition (2) is not true:

The integral $\int_{-\infty}^{a}\int_0^1 1_B(x,u)h(x)\mathrm{d}u\mathrm{d}x$ now equals,

$$\int_{-\infty}^{a} \min\left(\frac{f_1(x)}{Mh(x)}, 1\right) h(x)\mathrm{d}x,$$

which simplifies to

$$\int_{-\infty}^{a} \frac{1}{M}\min(f_1(x), Mh(x))\mathrm{d}x.$$

Thus the density of the output W is proportional to,

$$f_1^*(x) = \min(f_1(x), Mh(x)),$$

rather than to $f_1(x)$.

Rejection method is a good method if a suitable envelope can be found for the target density. Suppose the target density is $f(x)$ and $f_1(x) = Kf(x)$ is known. Suppose $g(x)$ is the proposal or majorizing density and suppose that

$$f(x) \le Mg(x),$$

(so that $f_1(x) \le KMg(x)$). The probability of accepting a sample is $(1/M)$ and the distribution of number of trials needed for one acceptance is geometric. Thus, on the average M trials would be needed for accepting one sample. This can be a problem if M is large.

3. Markov Chain Monte Carlo

The Monte Carlo methods discussed above were based on generating independent samples from the specified distribution. Metropolis and others in a paper published in *Journal of Chemical Physics* in 1953 use a very different approach for simulation. The paper deals with computation of certain properties of chemical substances, and uses Monte Carlo techniques for the same – but in a novel way:

For the distribution of interest whose density is π, they construct a Markov Chain $\{X_n\}$ in such a way that the given distribution π is the stationary distribution for the chain. The chain constructed is aperiodic and irreducible so that the stationary distribution is unique. Then the ergodic theorem ensures that

$$\frac{1}{N}\sum_{n=1}^{N} g(X_n) \to \int g(x)\pi(x)\mathrm{d}x,$$

as $N \to \infty$. This can be used to estimate $\int g(x)\pi(x)\mathrm{d}x$.

Given a distribution π how does one construct a Markov Chain with π as the stationary distribution?

The answer to this question is surprisingly simple. We begin with a simple example. The 3-dimensional analogue of this example was introduced in statistical physics to study behaviour of a gas whose particles have non-negligible radii and thus cannot overlap.

Consider an $N \times N$ "chessboard". Each square is assigned a **1** or **0**.

1 means the square is occupied and **0** means that the square is unoccupied. Each such assignment is called a *configuration*. A configuration is said to be feasible if all the neighbours of every square that is occupied are unoccupied. (Every square that is not in the first or last row or column has 8 neighbours.)

Thus a configuration is feasible if for every pair of adjacent squares, at most one square has a **1**.

For a feasible configuration (denoted by Γ), let $n(\Gamma)$ denote the number of **1**'s in Γ. The quantity of interest to physicists is the average of $n(\Gamma)$ where the average is taken over the uniform distribution over all the feasible configurations.

The total number of configurations is 2^{N*N} and even when $N = 25$, this number is 2^{625}, thus it is not computationally feasible to scan all configurations. Hence, count the feasible configurations and take the average of $n(\Gamma)$.

Assume that a powerful computer can sequentially scan the configurations Γ, decide if it is feasible and, if so, count $n(\Gamma)$ in one "cycle", and suppose the clock speed is 1000 GHz. Suppose there are a million such machines working in parallel. Then in one second, $2^{10+30+20} = 2^{60}$ configurations will be scanned. In one year, there are $24 \times 3600 \times 365 = 31536000$ which is approximately $2^{25} = 33554432$ seconds. Thus, we can scan 2^{85} configurations in one year, when we have a million computers working at 1000 GHz speed. It will still take 2^{540} years. Even if the size is 10, the number of configurations is 2^{100} and it would take $2^{15} = 32768$ years.

It is easy to see that when $N = 25$, the total number of feasible configurations is at least 2^{169}. To see this, assign **0** to all squares that have one of the coordinates even (the squares are indexed from 1 to 25). In the remaining 169 positions, we can assign a **1** or **0**. It is clear that each such configuration is feasible and the total number of such configurations is 2^{169}.

Let π denote the discrete uniform distribution on the set of feasible configurations:

$$\pi(\Gamma) = 1/M,$$

where M is the total number of feasible configurations. We construct a Markov Chain $\{X_k\}$ on the set of feasible configurations in such a way that it is aperiodic and irreducible and π is the stationary distribution for the chain.

Then as $N \to \infty$

$$\frac{1}{N}\sum_{k=1}^{N} n(X_k) \to \sum n(\Gamma)\pi(\Gamma).$$

The transition function $p(\Gamma, \Lambda)$ is described as follows: Fix $0 < p < 1$. Given a feasible configuration Γ, choose a square **s** (out of the N^2 squares) with equal probability. If any of the neighbours of **s** is occupied (has **1**) then $\Lambda = \Gamma$; if all the neighbours of **s** are unoccupied (have **0**) then with probability p, flip the "state" of the square **s** and otherwise do nothing. (Since the chain is slow moving, it would be better to take p close to 1.)

Let us observe that the chain is irreducible. First note that the transition function is symmetric:

$$p(\Gamma, \Lambda) = p(\Lambda, \Gamma).$$

If Γ, Λ differ at more than one square, then the above equality holds as both the probabilities are zero. The same is true if they differ at one square, then all the adjacent squares must have **0** and then both the terms above are

$$p/(N \times N).$$

Thus the transition function is symmetric.

Thus to prove that the chain is irreducible suffices to prove that the null configuration (where every square has a **0**) leads to any other square. If a configuration has exactly one **1** then it is clear that it can be reached from the null configuration in one step. It follows that any feasible configuration Γ can be reached from null configuration in $n(\Gamma)$ steps. Hence, the chain is irreducible.

Since $p(\Gamma, \Gamma) > 0$ for every feasible Γ, it follows that the chain is aperiodic. Thus the chain is a finite state Markov Chain that is irreducible and aperiodic. Hence it is positive recurrent and admits a unique stationary distribution. Note that

$$\begin{aligned}\sum_{\Gamma} \pi(\Gamma) p(\Gamma, \Lambda) &= \sum_{\Gamma} \frac{1}{M} p(\Gamma, \Lambda) \\ &= \sum_{\Gamma} \frac{1}{M} p(\Lambda, \Gamma) \\ &= \frac{1}{M} = \pi(\Lambda).\end{aligned}$$

Thus π is the unique stationary distribution.

Hence as $L \to \infty$,

$$\frac{1}{L} \sum_{k=1}^{L} n(X_k) \to \sum n(\Gamma)\pi(\Gamma), \tag{3}$$

where $\{X_k\}$ is the Markov Chain described above. Thus to approximate $\sum n(\Gamma)\pi(\Gamma)$, we can choose a large L and take $(1/L) \sum_{k=1}^{L} n(X_k)$ as an approximation. Even better, to reduce dependence on initial state X_0,we can first choose J, L integers, and then take

$$\frac{1}{L} \sum_{k=J+1}^{J+L} n(X_k)$$

as an approximation for $\sum n(\Gamma)\pi(\Gamma)$.

Thus by generating the Markov Chain as described above, we can estimate the "average number" of occupied sites. This is an example of the MCMC technique.

How large should J, L be for

$$\frac{1}{L} \sum_{k=J+1}^{J+L} n(X_k)$$

to give a good approximation to $\sum n(\Gamma)\pi(\Gamma)$?

Consider a simple random walk on $N = 2^{100}$ points placed on a (large) circle, so that

$$p_{ij} = 0{\cdot}5 \quad \text{if} \quad j = i + 1 \bmod (N) \text{ or } j = i - 1 \bmod (N),$$

and zero otherwise. Here also, the chain is irreducible and the transition probability matrix is doubly stochastic and thus the unique invariant probability distribution is the uniform distribution on the N points. Let h be a function on $\{0, 1, 2, \ldots, N-1\}$ and X_n be the Markov Chain. Since in L steps, this Markov Chain will at most move L steps to the right and L steps to the left, (and with very high probability, does not go more than $10 \times \sqrt{L}$ steps away from X_0), it is clear that L must be much larger than N for the "ergodic average"

$$\frac{1}{L}\sum_{k=J+1}^{J+L} h(X_k),$$

to be a good approximation of

$$\frac{1}{N}\sum_{j=0}^{N-1} h(j).$$

Therefore, in this case, the MCMC technique does not yield a good answer.

One has to be careful in choosing an appropriate L. As a thumb rule, let M be the smallest integrer such that

$$P(X_M = j \mid X_0 = i) > 0 \; \forall \text{ states } i, j.$$

Then J should be of the order of M and L should be much larger. In the "chessboard example" with $N = 25$, we can see that $M \leq 2 \times 169$.

In the "chessboard example" with $N = 25$, what J, L would suffice? To see this, we can generate the Markov Chain and compute the approximation several times, say 1000 times, and compute the variance of the estimate for various choices of J, L (table 1).
It can be seen that $J = 1000$ and $L = 100000$ gives a very good approximation.

Table 1. Monte Carlo results for Uniform distribution.

J	L	Mean	Variance
1000	1000	89·618	7·17764
1000	2000	89·8109	3·6315
1000	4000	89·9856	2·31912
1000	5000	90·0753	1·79497
1000	10000	90·2991	0·918608
1000	20000	90·3284	0·475608
1000	40000	90·389	0·243295
1000	50000	90·4042	0·21296
1000	100000	90·4365	0·0999061
1000	200000	90·4486	0·0527094
1000	400000	90·4464	0·0261528
1000	500000	90·4449	0·0207095
1000	1000000	90·4499	0·00986929
1000	2000000	90·4519	0·00535309
1000	4000000	90·4486	0·00245182
1000	5000000	90·4506	0·00195814
1000	10000000	90·4515	0·00104744

What if for the "chessboard example" we were interested in computing

$$\sum n(\Gamma)\pi(\Gamma),$$

with stationary invariant distribution $\pi(\Gamma)$ that is no longer the uniform distribution, but another distribution – say

$$\pi(\Gamma) = c\exp\{-Kn(\Gamma)\}$$

where K is a constant and c is normalising constant.

One possibility is to estimate c^{-1} by (for suitable J, L),

$$\sum_{k=J+1}^{J+L} \exp\{-Kn(X_k)\},$$

and then estimate $\sum n(\Gamma)\exp\{-Kn(\Gamma)\}$ by

$$\sum_{k=J+1}^{J+L} n(X_k)\exp\{-Kn(X_k)\},$$

so that the required approximation is

$$\left[\sum_{k=J+1}^{J+L} n(X_k)\exp\{-Kn(X_k)\}\right] \Big/ \left[\sum_{k=J+1}^{J+L} \exp\{-Kn(X_k)\}\right].$$

Here again, we can generate the estimate for J, L several times and compute the variance of the estimate (table 2).

Table 2. Monte Carlo results for Gibbs distribution: Ratio Method.

J	L	Mean	Variance
1000	1000	82·2744	12·4061
1000	2000	80·6866	8·93338
1000	4000	79·4026	7·99263
1000	5000	78·9912	7·55298
1000	10000	78·055	6·26682
1000	20000	76·8683	5·64003
1000	40000	75·9593	4·79803
1000	50000	75·6392	4·37511
1000	100000	74·6713	4·81159
1000	200000	73·9778	3·99164
1000	400000	73·1445	3·85673
1000	500000	72·9314	3·8598
1000	1000000	72·2698	3·81293
1000	2000000	71·6584	3·49804
1000	4000000	71·0077	3·54857
1000	5000000	70·8849	3·3355
1000	10000000	70·304	3·60938

Here, we can see that the variance of the estimate does not go down as expected, even when we take $L = 1000000$ and more.

Instead, can we construct a Markov Chain $\{X_n\}$ whose invariant distribution is $\pi(\Gamma)$ so that (3) is valid?

Let $p(\Gamma, \Lambda)$ denote the transition function described in the earlier discussion. Recall that it is symmetric.

Let

$$\alpha(\Gamma, \Lambda) = \min\left\{1, \frac{\pi(\Lambda)}{\pi(\Gamma)}\right\}.$$

Define

$$q(\Gamma, \Lambda) = p(\Gamma, \Lambda)\alpha(\Gamma, \Lambda),$$

if $\Gamma \neq \Lambda$ and

$$q(\Gamma, \Gamma) = 1 - \sum_{\Gamma \neq \Lambda} q(\Gamma, \Lambda).$$

For configurations Γ, Λ, if $\pi(\Gamma) \leq \pi(\Lambda)$,

$$q(\Gamma, \Lambda) = p(\Gamma, \Lambda),$$

and

$$q(\Lambda, \Gamma) = p(\Lambda, \Gamma)[\pi(\Gamma)/\pi(\Lambda)],$$

and hence

$$\pi(\Gamma)q(\Gamma, \Lambda) = \pi(\Lambda)q(\Lambda, \Gamma). \tag{4}$$

By interchanging roles of Γ, Λ, it follows that (4) is true in the other case: $\pi(\Lambda) \leq \pi(\Gamma)$ as well. As a consequence of (4), it follows that π is an invariant distribution for the transition function q. (Equation (4) is known as the "detailed balance equation".) Since p is irreducible, aperiodic, it follows that so is q and hence that π is the unique invariant measure and that the q chain is ergodic (table 3).

Observe that here the variance reduces as expected and the mean is very stable for $L = 100000$ as in the uniform distribution case. Thus we have reason to believe that this method gives a good approximation while the earlier method is way off the mark even with $L = 10000000$.

This construction shows that given any symmetric transition kernel $p(\Gamma, \Lambda)$ such that the underlying Markov Chain is an irreducible aperiodic chain which is easy to simulate from $p(\Gamma, \cdot)$, we can create a transition kernel q for which the stationary invariant distribution is π. As we will see, it is easy to simulate from $q(\Gamma, \cdot)$ - first we simulate a move from the distribution $p(\Gamma, \cdot)$ (to say Λ) and then accept the move with probability $\alpha(\Gamma, \Lambda)$, otherwise we stay put at Γ. As in rejection sampling, "the move with probability $\alpha(\Gamma, \Lambda)$" is implemented by simulating an observation, say u, from uniform (0,1) distribution and then accepting the move if $u < \alpha(\Gamma, \Lambda)$, otherwise, not to move from Γ in that step.

Let us now move to continuous case (see Robert & Casella 1999; Roberts & Rosenthal 2004). For now, let us look at real valued random variables. Again, we are given a target

Table 3. Monte Carlo results for Gibbs distribution: MCMC technique.

J	L	Mean	Variance
1000	1000	66·3028	9·1541
1000	2000	66·4575	5·76939
1000	4000	66·4792	2·93947
1000	5000	66·625	2·56194
1000	10000	66·5286	1·36421
1000	20000	66·6623	0·678447
1000	40000	66·6348	0·326311
1000	50000	66·6676	0·267248
1000	100000	66·6442	0·127588
1000	200000	66·6476	0·0629443
1000	400000	66·6632	0·0309451
1000	500000	66·66	0·0256965
1000	1000000	66·6581	0·0138524
1000	2000000	66·6539	0·00687151
1000	4000000	66·655	0·00344363
1000	5000000	66·6546	0·00289359
1000	10000000	66·6586	0·00133429

function $f_1(x) = Kf(x)$ with f being a density, K is not known and we want to generate samples from f. The starting point is to get a Markov Chain with good properties (irreducible, aperiodic) with the probability transition density function $q(x, y)$ (assumed to be symmetric, and such that it is possible to simulate from $q(x, \cdot)$ for every x. q is called the "proposal"). Then define (as in the finite case)

$$\alpha(x, y) = \min\{1, f_1(y)/f_1(x)\},$$

(with the usual convention: $\alpha(x, y) = 0$ if $f_1(y) = 0$ and $\alpha(x, y) = 1$ if $f_1(y) > 0$ but $f_1(x) = 0$) and then

$$p(x, y) = q(x, y)\alpha(x, y).$$

It is easy to check that

$$\alpha(x, y)/\alpha(y, x) = f_1(y)/f_1(x),$$

and hence that the "detailed balance equation" holds:

$$f(x)p(x, y) = f(y)p(y, x) \; \forall \, x, y. \tag{5}$$

We can now define a Markov Chain $\{X_n\}$ that has f as its stationary distribution as follows: Given that $X_n = x$, the chain does not move (*i.e.*, $X_{n+1} = x$) with probability $1 - \beta(x)$ where

$$\beta(x) = \int p(x, y)\mathrm{d}y$$

and given that it is going to move, it moves to a point y chosen according to the density

$$p(x, y)/\beta(x).$$

The transition kernel $P(x, A)$ for this chain is given by, for a bounded measurable function g

$$\int g(z)P(x, \mathrm{d}z) = (1 - \beta(x))g(x) + \int g(z)p(x, z)\mathrm{d}z. \tag{6}$$

This can be implemented as follows: given $X_k = x$, we first "propose" a move to a point y chosen according to the law $q(x, \cdot)$ and then choose u according to the Uniform distribution on (0, 1) and then set $X_{k+1} = y$ if $u < \alpha(x, y)$ and $X_{k+1} = x$ if $u \geq \alpha(x, y)$. Once again we can verify the "detailed balance equation"

$$f(x)p(x, y) = f(y)p(y, x) \ \forall\, x, y,$$

and hence (on integration w.r.t. x) it follows that

$$\int f(x)p(x, y)\mathrm{d}x = \beta(y)f(y), \tag{7}$$

and hence using (6), (7) and Fubini's theorem we can verify that

$$\begin{aligned}
\int \left(\int g(z)P(x, \mathrm{d}z)\right) f(x)\mathrm{d}x &= \int (1 - \beta(x))g(x)f(x)\mathrm{d}x \\
&\quad + \int \left(\int g(y)p(x, y)\mathrm{d}y\right) f(x)\mathrm{d}x \\
&= \int (1 - \beta(x))g(x)f(x)\mathrm{d}x \\
&\quad + \int g(y)\beta(y)f(y)\mathrm{d}y \\
&= \int g(y)f(y)\mathrm{d}y.
\end{aligned}$$

Thus, $f(x)$ is the density of a stationary invariant distribution of the constructed Markov Chain. Note that here the transition probability function is a mixture of a point mass and a density w.r.t. the Lebesgue measure.

Let us note that in the procedure described above, if $f_1(X_k) > 0$ then $f_1(X_{k+1}) > 0$ and hence if we choose the starting point carefully (so that $f_1(X_0) > 0$), we move only in the set $\{y : f_1(y) > 0\}$. Usually, the starting point X_0 is chosen according to a suitable initial distribution.

We can choose the "proposal" chain in many ways. One simple choice: Take a continuous symmetric density q_0 on $\mathbb{R}$ with $q_0(0) > 0$ and then define

$$q(x, y) = q_0(y - x).$$

The chain $\{W_n\}$ corresponding to this is simply the random walk where each step is chosen according to the density q_0, which is chosen so that it has a finite mean (which then has to be zero since q_0 is symmetric) and such that efficient algorithm is available to generate samples from q_0. We also need to specify a starting point, which could be chosen according to a specified density g_0.

The resulting procedure is known as the Metropolis–Hastings Random Walk MCMC. Here is the algorithm as a psuedo-code: to simulate $\{X_k : 0 \leq k \leq N\}$ –

(1) Generate X_0 from the distribution with density q_0 and set $n = 0$;
(2) $n = n + 1$;
(3) generate W_n from the distribution with density q_0;
(4) $Y_n = X_n + W_n$ proposed move;
(5) generate U_n from uniform (0,1);
(6) if $U_n * f_1(X_n) \leq f_1(Y_n)$, then $X_{n+1} = Y_n$, otherwise $X_{n+1} = X_n$;
(7) if $n < N$, go to (2) else stop.

Then the generated Markov Chain $\{X_k : 0 \leq k \leq N\}$ has f as its stationary distribution.

Like in the discrete case, here too for a function g such that $\int |g(x)| f(x)\mathrm{d}x$

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} g(X_i) = \int g(x) f(x)\mathrm{d}x.$$

However, since the aim is to approximate the integral based on a finite sample, we ignore an initial segment of the change with the hope that the distribution of the chain may be closer to the limiting distribution. Also, in order to reduce "dependence" between successive values (we are going to have lots of instances where the chain does not move), one records the chain after suitable "Gap" G. Thus we record

$$Z_k = X_{B+kG},$$

for suitable "Burn In" B, "Gap" G for $k = 1, 2, \ldots L$ and then use

$$\frac{1}{N} \sum_{i=1}^{N} g(Z_i),$$

as an approximation to

$$\int g(x) f(x)\mathrm{d}x.$$

Example: Target is a mixture of two normals, with equal weights,

$$f_1(x) = \exp\{-(x-4)^2/8\} + \exp\{-(x-16)^2/8\}.$$

This is bi-modal, with the two distributions having almost disjoint supports. The mean of f is 10 and variance is 40 (so that standard deviation is 6·324).

The "random walk proposal" distribution is taken as double exponential with parameters 0,1 (we must ensure that the mean of the proposal is 0), the "burn in" is taken as 5000, "gap" as 50. We generate 10 samples of size 10000 and the mean, standard deviation and variance in each of the sample is given in table 4.

The above algorithm is an adaptation of the Metropolis algorithm. Hastings in 1970 suggested a modification that does not require the "proposal" kernel to be symmetric. This allows us to consider the "proposal" chain to be an i.i.d sequence. Thus the chain is just a sequence

Table 4. Monte Carlo results for Mixture of Normal distributions.

Mean	SD	VAR
10·173955	6·283718	39·485116
9·719518	6·317257	39·90773
9·783166	6·326743	40·02768
9·799898	6·300697	39·698783
10·124491	6·308687	39·799526
10·100052	6·326426	40·023664
9·309102	6·297279	39·655727
10·258374	6·322546	39·974584
9·819395	6·3241	39·994238
10·204787	6·303478	39·733837

of independent random variables W_n with common distribution having a density q_1 and take the transition function as

$$q(x, y) = q_1(y).$$

The multiplier $\alpha(x, y)$ is now given by the formula

$$\alpha(x, y) = \min\{1, [f_1(y)q_1(x)]/[f_1(x)q_1(y)]\},$$

and the transition kernel $p(x, y)$ is given by

$$p(x, y) = q(x, y)\alpha(x, y).$$

As in the random walk case, we can define a Markov Chain $\{X_n\}$ that has f as its stationary distribution as follows.

Given that $X_n = x$, the chain does not move (*i.e.* $X_{n+1} = x$) with probability $1 - \beta(x)$ where

$$\beta(x) = \int p(x, y)\mathrm{d}y,$$

and given that it is going to move, it moves to a point y chosen according to the density

$$p(x, y)/\beta(x).$$

This can be implemented as follows: given $X_k = x$, we first "propose" a move to a point y chosen according to the law $q(x, \cdot)$ and then choose u according to the uniform distribution on (0, 1) and then set $X_{k+1} = y$ if $u < \alpha(x, y)$ and $X_{k+1} = x$ if $u \geq \alpha(x, y)$. Once again we can verify the "detailed balance equation"

$$f(x)p(x, y) = f(y)p(y, x) \; \forall \, x, y,$$

and as in the random walk case, it follows that $f(x)$ is the density of a stationary invariant distribution of the constructed Markov Chain. Note that here the transition probability function is a mixture of a point mass and an absolutely continuous density.

Metropolis–Hastings independence chain: Here is the algorithm as a psuedo-code: to simulate $\{X_k : 0 \leq k \leq N\}$ –

(1) Generate X_0 from the distribution with density q_0 and set $n = 0$;
(2) $n = n + 1$;
(3) generate W_n from the distribution with density q_0;
(4) $Y_n = W_n$ proposed move;
(5) generate U_n from uniform (0,1);
(6) if $U_n * f_1(X_n)q_1(Y_n) \leq f_1(Y_n)q_1(X_n)$, then $X_{n+1} = Y_n$ otherwise $X_{n+1} = X_n$;
(7) if $n < N$, go to (2) else stop.

Then the generated Markov Chain $\{X_k : 0 \leq k \leq N\}$ has f as its stationary distribution.

When we have two Markov chains with the same stationary distribution, we can generate yet another chain where at each step we move according to one chain say with probability 0·5 and the other chain with probability 0·5.

This has an advantage that if for the given target, even if one of the two chains is well behaved then the "hybrid" chain is also well behaved.

How does the algorithm behave for *fat-tailed* distributions?

We ran the programme (Hybrid version) for Cauchy and found that even with 50000 sample size, burn in of 50000 and Gap of 20; the results were not encouraging. And this when the Cauchy distribution is taken with median 0, and the proposal and RW proposal also have mean 0 (double exponential (0,8) and Uniform $(-5, 5)$ respectively).

To see if burn in and gap (same sample size) improves the situation, we ran the hybrid algorithm with a burn in of 50,000,000 and a gap of 10000. Five samples each of size 50000 meant a total of 2,750,000,000. This took 7293 seconds (little more than 2 hours).

Also, with burn in of 50,000,000 and gap of 20000, total samples generated were 5,250,000,000 (over 5 billion). The time taken was a little under 4 hours.

With both these runs, the outputs seem to be stable. It appears that for fat-tailed distribution, we need large burn in and large gap.

Gibbs sampler

Suppose it is given that X, Y are real-valued random variables such that the conditional distribution of Y given X is normal with mean $0{\cdot}3Y$ and variance 4 and the conditional distribution of X given Y is normal with mean $0{\cdot}3X$ and variance 4. Does this determine the joint distribution of X, Y uniquely?

More general question: Let $\pi(x, y)$ be the joint density of X, Y; $f(y; x)$ be the conditional density of Y given $X = x$ and $g(x; y)$ be the conditional density of X given $Y = y$. Do $f(y; x)$, $g(x; y)$ determine $\pi(x, y)$?

Consider the one-step transition function $P((x, y), A)$ with density

$$h((u, v); (x, y)) = f(v; x)g(u; v).$$

This corresponds to the following: starting from (x, y), first update the second component from y to v by sampling from the distribution with density $f(v; x)$ and then update the first component from x to u by sampling from the distribution with density $g(u; v)$.

Let us note that if f^*, g^* denote the marginal densities of X, Y respectively, then

$$f(y; x) = [\pi(x, y)]/[f^*(x)],\ g(x; y) = [\pi(x, y)]/[g^*(y)],$$

and hence

$$\begin{aligned}\int\int h((u,v);(x,y))\pi(x,y)\mathrm{d}y\mathrm{d}x &= \int\int f(v;x)g(u;v)\pi(x,y)\mathrm{d}y\mathrm{d}x\\ &= \int f(v;x)g(u;v)\left[\int \pi(x,y)\mathrm{d}y\right]\mathrm{d}x\\ &= \int f(v;x)g(u;v)f^*(x)\mathrm{d}x\\ &= \int g(u;v)\pi(x,v)\mathrm{d}x\\ &= g(u;v)g^*(v)\\ &= \pi(u,v).\end{aligned}$$

Now if $f(y;x)$ and $g(x;y)$ are continuous and (strictly) positive for all x, y, then this chain is ψ irreducible and aperiodic (see appendix for definition) and has a stationary distribution $\pi(x,y)$ which must then be unique.

This answers the question posed above in the affirmative. Further, if we have algorithms to generate samples from the univariate densities $f(y;x)$ and $g(x;y)$, this gives an algorithm to (approximately) generate samples from $\pi(x,y)$ – run the chain for a sufficiently long time. This is an MCMC algorithm.

Note that we could have instead taken

$$h((u,v);(x,y)) = f(v;u)g(u;y),$$

or

$$h((u,v);(x,y)) = 0{\cdot}5(f(v;x)g(u;v) + f(v;u)g(u;y)).$$

In either case, the resulting Markov Chain would have $\pi(x,y)$ as its stationary invariant distribution.

This can be easily generalized to higher dimensions. The resulting MCMC algorithm is known is Gibbs sampler that is useful in situations where we want to sample from a multivariate distribution which is indirectly specified- the distribution of interest π is a distribution on $\mathbb{R}^d$ (for $d > 1$) and it is prescribed via its full conditional distributions.

Let $X = (X_1, X_2, \ldots, X_d)$ have distribution π and $x = (x_1, x_2, \ldots x_n)$. Let

$$\begin{aligned}X_{-i} &= (X_1, X_2, \ldots X_{i-1}, X_{i+1} \ldots ,, X_d)\\ x_{-i} &= (x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n).\end{aligned}$$

The conditional density of X_i given $X_{-i} = x_{-i}$ is denoted by

$$f_i(x_i; x_{-i}).$$

As in the case when $n = 2$, the collection $\{f_i : 1 \le i \le d\}$ completely determines π if each f_i is a strictly positive continuous function. It should be noted that if instead of the full conditional densities f_i (conditional density of ith component given all the rest), the

conditional densities of the ith component, given all the preceding components, is available for $i = 2, 3, \ldots d$ along with the density of the first component, then it is easy to simulate a sample: first we simulate X_1, then X_2 and so on till X_d.

If we only know f_i, $1 \leq i \leq n$, Gibbs sampler is an algorithm to generate a sample from π. In d-dimensions, we can either update the d components sequentially in some fixed order or at each step choose one component (drawing from uniform distribution on $\{1, 2, \ldots d\}$).

Let $x = (x_1, x_2, \ldots x_d)$ be a point from the support of the joint distribution. Set $X^0 = x$. Having simulated $X^1, X^2, \ldots X^n$, do the following to obtain X^{n+1} –

(1) choose i from the discrete uniform distribution on $\{1, 2, \ldots, d\}$;
(2) simulate w from the conditional density $f_i(x_i; X^n_{-i})$;
(3) set $X^{n+1}_i = w$ and $X^{n+1}_j = X^n_j$, for $j \neq i$.

Note that at each step, only one component is updated. It can be shown that the Markov Chain has π as its unique invariant measure and hence for large n, X_n can be taken to be a sample from π.

For more details on MCMC, see Robert & Casella (1999) and references therein.

4. Perfect sampling

We have seen some algorithms to simulate Markov chains in order to estimate quantities associated with their limiting distribution. One of the difficulties in this is to decide when to stop, *i.e.* what sample size to use so as to achieve close approximation.

Propp & Wilson (1996) proposed a refinement of the MCMC yielding an algorithm that generates samples exactly from the stationary distribution.

This algorithm is called *Perfect Sampling* or *Exact Sampling*. The algorithm is based on the idea of coupling of Markov chains. Propp & Wilson (1996) called this algorithm *Coupling from the past*. It consists of simulating several copies of the Markov Chain with different starting points, all of them coupled with each other.

Let us now focus on finite state Markov chains. Given a transition matrix $P(i, j)$, we can construct a function ψ such that $X_0 = i_0$, and

$$X_{n+1} = \psi(X_n, U_n),$$

where $\{U_k\}$ is a sequence of independent simulations from uniform (0,1) yields a Markov Chain with transition matrix P and starting at i_0. This gives us an algorithm for simulating a Markov Chain.

There is no unique choice of the function given the matrix P. For example, given one function ψ, one can define $\phi(x, u) = \psi(x, 1 - u)$ and then

$$Y_{n+1} = \phi(Y_n, U_n),$$

also yields a Markov Chain with the same transition probabilities.

For the case of a finite state Markov Chain, one choice is: We can assume that the state space is $E = \{1, 2, \ldots, N\}$. Let us define

$$q(i, j) = \sum_{k=1}^{j} p(i, k).$$

Define

$$\psi(i,u) = 1 \quad \text{if } u \le q(i,1),$$
$$\psi(i,u) = 2 \quad \text{if } q(i,1) < u \le q(i,2),$$
$$\cdots$$
$$\psi(i,u) = j \quad \text{if } q(i,j-1) < u \le q(i,j),$$
$$\cdots$$
$$\psi(i,u) = N \quad \text{if } q(i,N-1) < u.$$

Then it follows that for a uniform (0,1) random variable U, the distribution of $\psi(i,U)$ is $\{p(i,j),\ 1 \le j \le N\}$.

Now fix i_0 and let $\{U_n : n \ge 1\}$ be i.i.d. uniform (0,1). Let $\{X_n\}$ be defined by

$$X_{n+1} = \psi(X_n, U_n).$$

Then $\{X_n\}$ is a Markov Chain with transition probability matrix P and initial state i_0.

Let $\{U_n : n \ge 1\}$ be i.i.d. uniform (0,1), and $\{V_n : n \ge 1\}$ also be i.i.d. uniform (0,1). Let i_0 and j_0 be fixed.

Define $\{X_n : n \ge 0\}$, $\{Y_n : n \ge 0\}$ and $\{Z_n : n \ge 0\}$ as follows: $X_0 = i_0$, $Y_0 = j_0$, $Z_0 = j_0$

$$X_{n+1} = \psi(X_n, U_n),\ n \ge 1,$$
$$Y_{n+1} = \psi(Y_n, V_n),\ n \ge 1,$$
$$Z_{n+1} = \psi(Z_n, U_n),\ n \ge 1.$$

All the three processes are Markov chains with transition probability matrix P and X starts at i_0 while Y and Z both start at j_0.

$\{X_n : n \ge 0\}$ and $\{Y_n : n \ge 0\}$ are independent chains, while $\{Y_n : n \ge 0\}$, $\{Z_n : n \ge 0\}$ have the same distribution. Hence, if we were required to simulate a chain starting at j_0 we can use either $\{Y_n : n \ge 0\}$ or $\{Z_n : n \ge 0\}$.

Since the same sequence $\{U_n : n \ge 1\}$ is used in generating the chains X and Z, they are obviously correlated.

X and Z above are said to be coupled.

Now let us fix a transition probability matrix P. Instead of starting the chain at $n = 0$, we can start the chain at $n = -10000$ or $n = -100000000$!

If the chain begins at $n = -\infty$ (in the infinite past, this can be made precise) with the stationary distribution π, then at each step its marginal distribution is π.

The Propp–Wilson algorithm: We will to generate samples from uniform (0,1) and for reasons to be made clear later, number them as $U_{-0}, U_{-1}, \ldots, U_{-m}, \ldots$. Let $m = 1$.

(1) For each starting point $i \in E$, generate a chain $X_n^{i,m}$, $-m \le n \le 0$:

$$X_{-m}^{i,m} = i$$
$$X_{n+1}^{i,m} = \psi(X_n^{i,m}, U_n)\ \ 0 \le -m < n \le 0.$$

(2) if $X_0^{i,m} = X_0^{j,m}$ for all i, j (*i.e.*, if all the N chains meet) then stop and return $W = X_0^{1,m}$. Otherwise, set m=m+1 and goto 1.

Note that the chain $\{X_{-k}^{i,m} : -m \leq -k \leq 0\}$ uses the random variables $\{U_{-m}, \ldots U_{-1}, U_0\}$. Thus as we go from m to $m+1$, only one new uniform (0,1) is generated and we reuse the m samples generated earlier.

If the Propp–Wilson algorithm terminates with probability one, then the sample returned has exact distribution π.

To see this, suppose that for the given realization of $U_{-0}, U_{-1}, \ldots, U_{-m}, \ldots$., the algorithm has terminated with $m = 17600$ and has returned a sample W.

Let us examine what would happen if we do not stop, but keep generating the N chains.

Take $m = 17601$. We will argue that $X_0^{1,17601}$ is still W. The chain $X^{1,17601}$ now starts at 1: $(X_{-17601}^{1,17601} = 1)$ and goes to some state i, $X_{-17600}^{1,17601} = i$. From then on, it follows the trajectory of the chain $X^{i,17600}$ which was initialized at $m = 17600$ at the state i (since both the chains begin at i and use the same set of uniform variables $U_{-17600}, \ldots, U_0$). Hence $X_0^{1,17601} = W$.

The same argument can be repeated and we can conclude that if we ran the algorithm for any $m > 17600$, all the N chains $X^{i,m}$ will meet at time 0 and the common value will be W.

Thus,

$$\lim_{m\to\infty} X_0^{i,m} = W \ \forall\, i.$$

It can be shown that the distribution of W is the stationary distribution π.

Generating N chains in order to generate one sample seems tedious. Suppose that P is such that

$$q(i, j) \geq q(i+1, j) \ \ \forall j,\ 1 \leq j < N, \forall i.$$

This means, conditional distribution of X_{n+1}, given $X_n = i+1$, stochastically dominates the conditional distribution of X_{n+1}, given $X_n = i$. The chain is then called stochastically monotone.

Under this condition, it can be checked that for the canonical choice of the function ψ described earlier,

$$\psi(i, u) \leq \psi(i+1, u),$$

and hence that for $i \leq k$

$$\psi(i, u) \leq \psi(k, u)$$

Thus, by induction it follows that if $i_0 < j_0$ and X, Z are defined by

$$X_{n+1} = \psi(X_n, U_n), \quad n \geq 1,$$

$$Z_{n+1} = \psi(Z_n, U_n), \quad n \geq 1,$$

then

$$X_n \leq Z_n \ \ \forall n \geq 1.$$

If the Markov Chain is stochastically monotone, then instead of generating all the chains and checking if they meet, we can generate only two chains

$$X_n^1, \quad X_n^N,$$

since

$$X_n^1 \leq X_n^{i,m} \leq X_n^N \; \forall \, i.$$

(This is true becasue we are generating coupled chain via the special function ψ.) So if X^1 and X^N meet, all the chains meet.

Thus even for a large state space, it is feasible to run the Propp–Wilson algorithm to generate an exact sample from the target stationary distribution. See Propp & Wilson (1996).

The natural ordering on the state space has no specific role. If there exists an ordering with respect to which the chain is stochastically monotone, we can generate chains starting at minimum and maximum and then stop when they meet.

A great deal of research is going on on this theme.

Appendix A

Markov chains on a general state space: For more details on material in this appendix including proofs etc. see (Meyn & Tweedie 1993). Suppose E is a locally compact separable metric space and suppose $P(x, A)$ is a "probability transition function" on E:

- for each x, $P(x, \cdot)$ is a probability measure on E (equipped with its Borel sigma field $\mathcal{B}(E)$);
- for each $A \in \mathcal{B}(E)$, $P(\cdot, A)$ is a Borel measurable function on E.

Example. $E = \mathbb{R}$ and for $x \in \mathbb{R}$, $A \in \mathcal{B}(\mathbb{R})$

$$P(x, A) = \int_A [1/\sqrt{2\pi}] \exp\{-[1/2](y - x)^2\} \mathrm{d}x.$$

Let $\{X_n\}$ be the Markov Chain with P as the transition probability kernel. Let $\mathbb{P}_x$ be the distribution of the chain when $X_0 = x$. The n-step transition probability function is defined by

$$P^{n(x,A)} = \mathbb{P}_x(X_n \in A).$$

Also for $x \in E$, $A \in \mathcal{B}(E)$ let,

$$L(x, A) = \mathbb{P}_x(X_n \in A \text{ for some } n \geq 1).$$

$$U(x, A) = \sum_{n=1}^{\infty} \mathbb{P}_x(X_n \in A).$$

$$Q(x, A) = \mathbb{P}_x(X_n \in A \textit{ infinitely often}).$$

Let

$$\eta_A = \sum_{n=1}^{\infty} I_A(X_n).$$

Then

$$U(x, A) = \mathbb{E}_x(\eta_A)$$

and

$$Q(x, A) = \mathbb{P}_x(\eta_A = \infty).$$

$L(x, A)$ is the probability of reaching the set A starting from x, $U(x, A)$ is the average number of visits to the set A starting from x and $Q(x, A)$ is the probability of infinitely many visits to the set A starting from x.

The Markov Chain is said to be "ϕ-irreducible" if there exists a positive measure λ on $(E, \mathcal{B}(E))$ such that for all $A \in \mathcal{B}(E)$, with $\lambda(A) > 0$

$$P(x, A) > 0 \; \forall \; x \in E.$$

λ is said to be a "irreducibility measure".

For a ϕ-irreducible Markov Chain there exists a "maximal irreducibility" measure ψ such that ψ dominates every other "irreducibility measure" of the chain. The phrase "The Markov Chain is ϕ-irreducible with maximal irreducibility measure ψ" is often written as "The Markov Chain is ψ-irreducible".

If E is a finite state space and P is a transition probability matrix such that there is one communicating class F and the rest of the states (belonging to $E \cap F^C$) are transient. Then the chain is ψ-irreducible with maximal irreducibility measure ψ being the uniform measure on F (or any measure equivalent to it).

A ψ-irreducible Markov Chain is said to be "recurrent" if for all $A \in \mathcal{B}(E)$ with $\psi(A) > 0$,

$$U(x, A) = \infty \; \forall \; x \in E.$$

(Recall: $U(x, A)$ is the average number of visits to the set A starting from x.) A ψ-irreducible Markov Chain is said to be "transient" if $\exists A_n \in \mathcal{B}(E), n \geq 1$ such that $E = \cup_n A_n$ and $M_n < \infty$,

$$U(x, A_n) \leq M_n \; \forall x \in A_n.$$

As in the countable state space case, we have a dichotomy: A ψ-irreducible Markov Chain is either recurrent or transient.

A ψ-irreducible Markov Chain is said to be "Harris recurrent", if for all $A \in \mathcal{B}(E)$ with $\psi(A) > 0$,

$$Q(x, A) = 1 \;\; \forall x \in A.$$

Every recurrent chain is essentially Harris recurrent. We now make this precise.

A set $H \subseteq E$ is said to be absorbing if

$$P(x, H) = 1 \; \forall \; x \in H.$$

A set $H \subseteq E$ is said to be full if

$$\psi(H^C) = 0.$$

If H is a full absorbing set, we can restrict the chain to H retaining all its properties.

Theorem A1. *If a ψ-irreducible chain is recurrent, then it admits a full absorbing set H such that restricting to H, the chain is "Harris recurrent".*

A set A is said to be an "atom" if $\psi(A) > 0$ and

$$P(x, B) = P(y, B) \ \forall\ B \in \mathcal{B}(E),\ x, y \in A.$$

In this case, we can lump all the states in A together and treat the set A as a singleton, retaining the Markov property for the reduced chain.

If the chain has an atom A, then everytime the chain reaches A, the chain "regenerates" itself and thus we can mimic the usual arguments in countable state space case for recurrence, ergodicity etc.

Athreya *et al* (1996) showed how to create a "pseudo atom" for a large class of chains and use it to study the chain. We outline the underlying idea in a special case.

A set C is said to be "small", if there exists $m \geq 1$ and a positive measure ν such that

$$P^m(x, A) \geq \nu(A) \ \ \forall x \in C,\ \forall A \in \mathcal{B}(E). \tag{A1}$$

Let C be a small set with m, ν satisfying (A1). Let d be the g.c.d. of the set of integers k such that there exists $\delta_k > 0$ with

$$P^k(x, A) \geq \delta_k \nu(A)\ \forall\ x \in C,\ \forall\ A \in \mathcal{B}(E).$$

It can be shown that d does not depend on the small set C, or the measure ν. The chain is said to be "aperiodic" if $d = 1$.

It is said to be "strongly aperiodic" if (A1) holds for $m = 1$ and some C, ν. Consider a "strongly aperiodic" ψ-irreducible chain with a small set C. Thus we have a probability measure ν such that for $\delta > 0$

$$P(x, A) \geq \delta \nu(A) \ \ \forall x \in C,\ \forall A \in \mathcal{B}(E). \tag{A2}$$

The Athreya–Ney and Nummelin idea is as follows (Athreya & Ney 1978; Nummelin 1978):

Define a chain (Y_n, η_n) taking values in $E \times \{0, 1\}$ as follows:

If $\eta_k = 1$, we draw a sample from ν; if $Y_k \in C$, $\eta_k = 0$, we draw a sample from

$$[P(x, \cdot) - \delta \nu(\cdot)]/(1 - \delta),$$

and if $Y_k \notin C,\ \eta_k = 0$, we draw a sample from $P(x, \cdot)$ and set it as Y_{k+1}. Further, if $Y_{k+1} \notin C$, we set $\eta_{k+1} = 0$ and if $Y_{k+1} \in C$, we set $\eta_{k+1} = 0$ with probability $1 - \delta$ and equal to 1 with probability δ.

It can be seen that the event $Y_k \notin C,\ \eta_k = 1$ will never occur. Clearly $E \times \{1\}$ is an atom of the "split chain" (Y_n, η_n). A little calculation shows that the marginal chain Y_n is also a Markov Chain with transition probability function P. As a result, successive hitting times of $E \times \{1\}$ are regeneration times for the chain Y_n. Note that since $\psi(A) > 0$ and the chain is ψ-irreducible,

$$P(x, A) > 0,\ \forall\ x \in E.$$

The Athreya-Ney-Neumalin idea also works if we have a set C with $\psi(C) > 0$ and a positive measure ν such that

$$\sum_{m=1}^{\infty} (1/2^m) P^m(x, A) \geq \nu(A) \ \ \forall x \in C,\ \forall A \in \mathcal{B}(E).$$

The Markov Chain is said to be "Feller" (or "weak Feller") if for all bounded continuous f, the function,

$$h(x) = \int f(y)P(x, \mathrm{d}y),$$

is continuous. The chain is said to be "strong Feller" if for all bounded measurable f, h defined above is continuous.

A probability measure π is said to be invariant or stationary if

$$\int P(x, A)\, \pi(\mathrm{d}x) = \pi(A)\ \forall\ A \in A \in \mathcal{B}(E).$$

A bounded function f on E is said to be "harmonic" if

$$\int P(x, \mathrm{d}y) f(y) = f(x).$$

If the Markov Chain is ψ-irreducible aperiodic and if a invariant probability measure π exists, then it is recurrent and every bounded harmonic function is constant π a.s. Further, in this case the chain is Harris recurrent if and only if every bounded harmonic function is constant (*everywhere*).

Ergodic theorem: If the Markov Chain is ψ-irreducible aperiodic and if a invariant probability measure π exists, then for a bounded measurable function g on E, for all x outside a π-null set,

$$\frac{1}{N}\sum_{j=1}^{N} g(X_j) \to \int g\, \mathrm{d}\pi, \quad \mathbb{P}_x - a.s.$$

Further if the chain is Harris recurrent, then the relation above holds for all x.

Suppose $E = \mathbb{R}^d$ or a connected subset of $\mathbb{R}^d$. Assume that there exists a continuous function u on $E \times E$ and a probability measure λ on $(E, \mathcal{B}(E))$ such that

$$P(x, A) = \int_A u(x, y)\ \lambda(\mathrm{d}y) + (1 - \beta(x))\delta_{\{x\}}(A),$$

for all $x \in E,\ A \in \mathcal{B}(E)$ with

$$\beta(x) = \int u(x, y)\lambda(\mathrm{d}y).$$

Suppose that

$$u(x, y) > 0\ \forall x \in E,\ y \in E.$$

Then the Markov Chain $\{X_n\}$ is ψ-irreducible aperiodic. For such a chain, if there exists a positive function f such that

$$f(y) = \int u(x, y) f(x)\lambda(\mathrm{d}x),$$

then (upto normalization) f is the density of the unique stationary invariant distribution, the chain is Harris positive recurrent and for any bounded g

$$\frac{1}{N}\sum_{i=1}^{N} g(X_i) \to \int g(x)f(x)\lambda(\mathrm{d}x).$$

The Markov chains appearing in the Metropolis–Hastings algorithm often satisfy these conditions.

References

Athreya K B, Ney P 1978 A new approach to the limit theory of recurrent Markov chains. *Trans. Am. Math. Soc.* 245: 493–501

Athreya K B, Doss H, Sethuraman J 1996 On the convergence of the Markov Chain simulation method. *Ann. Stat.* 24: 69–100

Meyn S P, Tweedie R L 1993 *Markov chains and stochastic stability* (Berlin: Springer-Verlag)

Nummelin E 1978 A splitting technique for Harris recurrent Markov chains. *Z. Wahrsch. Verw. Gebiete* 43: 309–318

Propp J G, Wilson D B 1996 Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Struct. Algorithms* 9: 223–252

Robert C P, Casella G 1999 *Monte Carlo statistical methods* (Berlin: Springer-Verlag)

Roberts G O, Rosenthal J S 2004 General state space Markov chains and MCMC algorithms. *Probab. Surv.* 1: 20–71