Introduction to MCMC

Rajeeva L Karandikar
Distinguished Professor
Chennai Mathematcial Institute
Chennai
rlk@cmi.ac.in
http://www.cmi.ac.in/~rlk

1 Introduction

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

$$E[g(X)] = \int g(x)f(x)dx \tag{1.1}$$

for a given function g. If the functions f, g are such that the integral in (1.1) 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

$$u_n = \tfrac{1}{n} \textstyle\sum_{i=1}^{N} g(x_i)$$

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

Moreover, the Central limit theorem gives the order of error: the error 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 will be dealing 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: In order to generate sample from Normal distribution, mean 0, variance 1:
Generate U, V from Unifrom (0,1), U, V independent) and define

$$X = \sqrt{-2\log(U)}\cos(2\pi V)$$

$$Y = \sqrt{-2\log(U)}\sin(2\pi V).$$

Then it can be shown that 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

$$\int g(x)f(x)dx$$

can inflate the error.

It would be much better if just knowing f_1, we can device a scheme to generate random sample from

$$f = \frac{f_1}{K}$$

and thereby compute $\int g(x)f(x)dx$ 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

$$\pi(\theta \mid x) = \frac{p(x \mid \theta)\pi(\theta)}{\int p(x \mid \theta)\pi(\theta)d\theta}$$

Often it is difficult to obtain exact expression for

$$\int p(x \mid \theta)\pi(\theta)d\theta$$

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.

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

von Neumann in 1951 proposed an algorithm known as Rejection Sampling

To understand this method, consider the following problem. We are given a fair coin and are supposed to devise a scheme for choosing one colour out of

Green, Red and Purple

with equal probabilities.

von Neumann's algorithm in this situation would give the following procedure:

(i) Toss the coin two times.

(ii) If the outcome is 'HH', then choose Green, if the outcome is 'HT', then choose Red and if the outcome is 'TH', then choose Purple (and stop in all the 3 cases).

(iii) If the outcome is 'TT', continue (GOTO (i)).

The event that the procedure will continue beyond k steps is the same as getting 'TT' in each of the first k tosses and hence its probability equals $\frac{1}{4^k}$. So the procedure will terminate in finite number of steps with probability 1.

Given that the procedure has terminated on the m^{th} step, the (conditional) probability of getting green is:
$P(HH \mid \{HH, HT, TH\}) = \frac{1}{3}$

Simillarly, it follows that when the procedure terminates, the probability of choosing each of the three colours is $\frac{1}{3}$.

Here is the Rejection Sampling algorithm of von Neuman to generate samples from a density $f = \frac{1}{K} f_1$ knowing only f_1 if there is a density h such that (it is possible to generate samples from h and)

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

1. Generate a random sample from the distribution with density h. Let it be y.

2. Accept y as the sample with probability $\dfrac{f_1(y)}{Mh(y)}$.

3. If step 2 is not a success, then goto 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.

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

Theorem **2.1 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

$$f_1(x) \leq Mh(x) \;\; \forall x. \tag{2.2}$$

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

$$B = \left\{(x,u):\ u \leq \frac{f_1(x)}{Mh(x)}\right\}$$

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

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

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 $\{Z_\tau \in A\} = \{W \leq a\}$)

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

We have used $\int_0^1 1_B(x,u) du = \frac{f_1(x)}{Mh(x)}$ 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 envolope condition (2.2) is not true:

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

$$\int_{-\infty}^{a}\min(\frac{f_1(x)}{Mh(x)},1)h(x)dx$$

which simplifies to

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

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 target density and suppose that

$$f(x)\leq Mg(x)$$

(so that $f_1(x)\leq Mg(x)/K$). The probability of accepting a sample is $\frac{1}{M}$ and the distribution of no of trials needed for one acceptance is geometric. Thus, on the avarage M trials would be needed for accepting one sample. This can be a problem if M is large.

3 Markov Chain Monte Carlo (MCMC)

The Monte Carlo method discussed above were based on generating independent samples from the specified distribution. Metropolis et al in a paper published in Journal of Chemical Physics in 1953 used a very different approach for simulation. The paper deals with computation of certain properties of chemical substances. The paper used Monte Carlo techniques for the same - but in a novel way:

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

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

as $N\to\infty$. This can be used to estimate $\int g(x)\pi(x)dx$.

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 will begin with a simple example.

The 3-dimensional analogue of this example was introduced in statistical physics to study behavior of a gas whose particles have non-negligible radii and thus cannot overlap.

Consider a $N \times N$ 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 neighbors of every square that is occupied are unoccupied. (Every square that is not in the first or last row or column has 8 neighbors.)

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} so it is not computationally feasible to scan all configurations, 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 a 1000 GHz. Suppose there are 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 will be able to scan through 2^{85} configurations in one year, when we have a million computers working at a 1000 GHz speed. It will still take 2^{540} 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 whose one of the coordinates is even (the squares are indexed from 1 to 15). In the remaining 64 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:

$$\pi(\Gamma) = \frac{1}{M}$$

where M is the total number of feasible configurations. We wish to evaluate

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

where the sum is over all feasible configurations.

To achieve this, we will 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 neighbors of s is occupied (has 1) then $\Lambda = \Gamma$; if all the neighbors of s are unoccupied (have 0) then with probability p, flip the state of the square $\mathbf{s}$ and otherwise do nothing. (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 and if one of the neighbor is occupied. If they differ at one square and all the adjacent squares have 0 and then both the terms above are

$$\frac{p}{N \times N}.$$

Thus the transition function is symmetric.

So 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. So 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. Since the transition probability matrix is symmetric, it is doubly stochastic and hence the uniform distribution on the state space- namely π is the unique stationary distribution.

Hence as $L \to \infty$

$$\frac{1}{L}\sum_{k=1}^{L} n(X_k) \to \sum n(\Gamma)\pi(\Gamma) \tag{3.3}$$

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 $\frac{1}{L}\sum_{k=1}^{L} n(X_k)$ as an approximation. Even better, to reduce dependence on intial 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 $\sum n(\Gamma)\pi(\Gamma)$?

To examine this question, let us look at an example where even L equal to a trillion is not enough!

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

$$p_{ij} = 0.5 \quad \text{if } j = i + 1 \bmod (N) \text{ or } j = i - 1 \bmod (N)$$

and zero otherwise. Here also, the chain is irreducible and the transition probability matrix is doubly stochastic and so 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 larger than N for the ergodic average (and should be much larger)

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

to be a good approximation of

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

Thus, here even $L = 2^{80}$ wont suffice.

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

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

J	L	Mean	Variance
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

It can be seen that $J = 1000$ and $L = 100000$ gives a very good approximation. What if for the chessboard example we were interested in computing

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

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 normalizing constant. One possibility is to estimate $\frac{1}{c}$ by (for suitable J, L)

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

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

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

so that the required approximation is

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

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

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

J	L	Mean	Variance
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

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

is a good approximation to $\sum n(\Gamma)\pi(\Gamma)$ for suitable choice of J, L.

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

Let

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

For $\Gamma \neq \Lambda$ define

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

and

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

Note that for configurations Γ, Λ, if $\pi(\Gamma) \leq \pi(\Lambda)$,

$$q(\Gamma, \Lambda) = p(\Gamma, \Lambda) \tag{3.4}$$

and

$$q(\Lambda, \Gamma) = p(\Lambda, \Gamma)\frac{\pi(\Gamma)}{\pi(\Lambda)} \tag{3.5}$$

and hence

$$\begin{aligned} \pi(\Lambda)q(\Lambda, \Gamma) &= p(\Lambda, \Gamma)\pi(\Gamma) \\ &= p(\Gamma, \Lambda)\pi(\Gamma) \\ &= q(\Gamma, \Lambda)\pi(\Gamma) \end{aligned}$$

Here we have used (3.4), (3.5) and the fact thet $p(\Lambda, \Gamma)$ is symmetric.

Thus for $\pi(\Gamma) \leq \pi(\Lambda)$, we have

$$\pi(\Lambda)q(\Lambda, \Gamma) = \pi(\Gamma)q(\Gamma, \Lambda). \tag{3.6}$$

By interchanging roles of Γ, Λ, it follows that (3.6) is true in the other case $\pi(\Lambda) \leq \pi(\Gamma)$ as well.

Summing over Γ in (3.6) we get

$$\sum_{\Gamma} \pi(\Gamma) q(\Gamma, \Lambda) = \pi(\Lambda)$$

Hence π is an invariant distribution for the transition function q. The equation (3.6) 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.

Once again we present simulation results by generating the Markov Chain and computing the approximation several times, say 1000 times, and computing the variance of the estimate for various choices of J, L.

Simulation results:

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

J	L	Mean	Variance
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

Observe that here, the variance reduces as expected and 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 irreducible aperiodic chain and such that it 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.

There are many other ways of generating transition functions q with the required properties.

Let us now move to continuous case. For now, let us look at real valued random variables. Again, we are given a target 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. The transition function q is called the proposal).

Then define (as in the finite case)

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

(with the 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

$$\frac{\alpha(x,y)}{\alpha(y,x)} = \frac{f_1(y)}{f_1(x)}$$

and hence that the detailed balance equation holds:

$$f(x)p(x,y) = f(y)p(y,x) \ \ \forall x,y. \tag{3.7}$$

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

$$\beta(x) = \int p(x,y)dy$$

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

$$\frac{p(x,y)}{\beta(x)}.$$

We will see that we actually do not need to compute $\beta(x)$ in order to simulate the chain.

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

$$\int g(z)P(x,dz) = (1-\beta(x))g(x) + \int g(z)p(x,z)dz. \tag{3.8}$$

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

$$\int f(x)p(x,y)dx = \beta(y)f(y). \tag{3.9}$$

and hence using (3.8), (3.9) and Fubini's theorem we can verify that

$$\begin{aligned}\int (\int g(z)P(x,dz))f(x)dx &= \int (1-\beta(x))g(x)f(x)dx \\ &\quad + \int (\int g(y)p(x,y)dy)f(x)dx \\ &= \int (1-\beta(x))g(x)f(x)dx \\ &\quad + \int g(y)\beta(y)f(y)dy \\ &= \int g(y)f(y)dy\end{aligned}$$

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 an absolutely continuous density.

Let us note that in the procedure described above, $f_1(X_k) > 0$ implies that $f_1(X_{k+1}) > 0$ and hence if we choose the starting point carefully (so that $f_1(X_0) > 0$), we will 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 Metropolis-Hastings Random Walk MCMC. Here is the algorithm as a psuedo-code: to simulate $\{X_k : 0 \leq k \leq N\}$. We have taken the density g_0 to be q_0.

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 (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 $n < N$, goto (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 $\int \mid g(x) \mid f(x)dx$

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

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 B, G for $k = 1, 2, \ldots L$ and then use

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

as an approximation to

$$\int g(x)f(x)dx.$$

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 and 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 below

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

The above algorithm was a adaption of Metropolis algorithm. Hastings in 1970 had suggested a modification that does not require the proposal kernel to be symmetric. This allows us to consider the proposal chain to be Independent chain. Thus the chain is just a sequence 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

$$\alpha(x, y) = \min\Big\{1, \frac{f_1(y)q_1(x)}{f_1(x)q_1(y)}\Big\}$$

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

$$\beta(x) = \int p(x, y)dy$$

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

$$\frac{p(x, y)}{\beta(x)}.$$

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 (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 $n < N$, goto (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 with say 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. So it is recommended to use the hybrid algorithm.

4 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 $0.3Y$ and variance 4 and the conditional distribution of X given Y is Normal with mean $0.3X$ 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 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

$$f(y; x) = \frac{\pi(x, y)}{f^*(x)}, \quad g(x; y) = \frac{\pi(x, y)}{g^*(y)}$$

and hence

$$\begin{aligned}
& \int\int h((u, v); (x, y))\, \pi(x, y)\, dydx \\
= & \int\int f(v; x)g(u; v)\, \pi(x, y)\, dydx \\
= & \int f(v; x)g(u; v)[\int \pi(x, y)\, dy]dx \\
= & \int f(v; x)g(u; v)f^*(x)dx \\
= & \int g(u; v)\pi(x, v)dx \\
= & \; g(u; v)g^*(v) \\
= & \; \pi(u, v).
\end{aligned}$$

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

This answers the question posed above in 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 the 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

$$h((u, v); (x, y)) = 0.5(f(v;x)g(u;v) + f(v;u)g(u;y)).$$

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 knwon 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

$$X_{-i} = (X_1, X_2, \ldots X_{i-1}, X_{i+1} \ldots, , X_d)$$

$$x_{-i} = (x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n).$$

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 $\{f_i : \ 1 \leq i \leq d\}$ 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 i^{th} component given all the rest), the conditional densities of i^{th} 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 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.