Markov Chain Monte Carlo

2. A Method of Simulation

Arnab Chakraborty

Arnab Chakraborty did his BStat (Hons.) and MStat degrees from the Indian Statistical Institute (Calcutta). He is now a PhD student in the Department of Statistics at Stanford University, California. His research interest lies in areas of statistics and computational algebra. Apart from that he is interested in playing the piano and wishes he could really play well.

Part 1. Examples, *Resonance*, Vol.7, No.3, pp.25-34, 2002.

Keywords
Gibbs sampling, Markov Chain Monte Carlo, Bayesian inference, stationary distribution, convergence, image restoration.

We describe the mathematics behind the Markov Chain Monte Carlo method of simulation.

Markov Chain Sampling

'Martians on Earth' (Part 1) is an example of a situation in which MCMC is used. We want random samples from a distribution from which either it is difficult to sample directly or which is not specified completely, but which we know is the limiting distribution of a Markov chain (with certain properties to be made precise later) with known transition probabilities. Typically in Bayesian inference, the (joint) distribution from which random samples are needed, is cumbersome, but can be specified (uniquely) with the help of a collection of marginal and conditional distributions, which are easy to sample from. Then a random sample from the target distribution can be obtained as a limiting value of a sample path of a suitable Markov chain defined with the known marginal and conditional distributions.

In our day to day life we have to encounter chance at every moment. We have to predict tomorrow's weather before we plan a picnic, we have to estimate the condition of the pitch before we play cricket. In short, we are to make decisions under situations where the outcome is random. There are two major approaches to cope with such a situation. The first one is the *analytic method* – makes a mathematical model of chance and deals with the problem by mathematical means. While very appealing, this approach cannot handle many complicated situations arising in practice. In such situations we resort to a second approach – the method of *simulation*.

This latter method simply says that before you gamble in a casino, play the same game in your home a large number of times and assess how lucrative the game is. In a sense, this is the simplest and oldest way of dealing with randomness. The article by Kunte [1] motivates this method using simple examples, where independent samples are drawn from a given distribution. A variant of this situation is where resampling is done of a given set of data; an instance of such a resampling procedure is the method of bootstrap (see [2]). With the advent of tremendous computational power, it is a simple thing to run a simulation program a large number of times, and analyze the results. Did I say it is simple? Yes, it *is* simple, that is, once you know how to write the simulation program. It so happens that *designing* an algorithm to simulate a real life random problem is often not that easy. Often due to the limitations of a problem, such as lack of complete knowledge of the distribution to be sampled from or the complicated nature of the distribution to be sampled from, independent and identically distributed samples cannot be generated, but samples can be generated in a Markovian dependent sequence such that its limiting distribution is our target distribution. The 'Martians on Earth' problem is one such case. And this is the essence of the Markov chain Monte Carlo method.

Samples can be generated in a Markovian dependent sequence such that its limiting distribution is our target distribution.

Basics of Markov Chain

In order to understand the mathematics of MCMC, recapitulate your Markov chain. The two-part series article by Athreya [3] covers much of the basics on MC (with countable state space). So we assume that you know transition probabilities (stationary or otherwise), Chapman–Kolmogorov relation, transient, recurrent and periodic states, irreducible chains, and conditions under which the n-step transition probabilities will converge to the chain's stationary (an invariant or equilibrium) distribution.

Here is a brief summary. A (discrete time) Markov chain (MC) consists of a non-empty set, $\mathcal{S}$, called the *state space*, a sequence of random variables X_t for $t = 0, 1, 2, 3,$ taking values in $\mathcal{S}$, governed by the following probability laws:

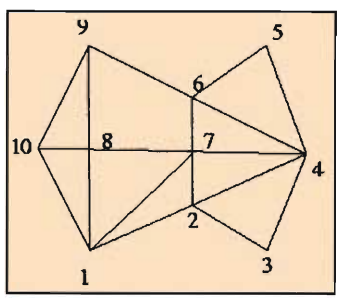
1. X_0 has some given distribution $\mathbf{v_0}$.
2. $P(X_{t+1}|X_t, \quad , X_0) = P(X_{t+1}|X_t)$.

We call the MC *homogeneous* if $P(X_{t+1}|X_t)$ is free of t. As you should already know, under certain conditions the law of X_t converges to some *stationary distribution* as t goes to infinity, and this limit is the same irrespective of the initial distribution $\mathbf{v_0}$. Here is one version of the result for MC's with finite state spaces.

> **Theorem :** A (finite state space) irreducible, aperiodic MC always converges to its unique stationary distribution.

There are similar results that guarantee convergence (though under stricter conditions) for MC's with infinite state spaces. MCMC procedures use such results to simulate from difficult distributions as follows. Let π be our target distribution on some set $\mathcal{S}$. Then we first construct some MC with state space $\mathcal{S}$ such that the above theorem (or one of its analogues) applies to it, and such that the MC has π as its stationary distribution. Then we start the chain from some arbitrary point X_0 in $\mathcal{S}$, and run it for long, say upto n steps. If n is large, we have X_n as a random variable with distribution approximately π.

Figure 1.



Another MCMC Example

Let us look at an example. Consider the graph shown in *Figure* 1. You may consider this as a city road map

where the edges denote roads, and the vertices denote important points in the city.

Our state space is the set of vertices. Our target distribution is proportional to the degree of a vertex. Here is our MC. Start from vertex 1, *i.e.*, $X_0 = 1$. Then, if you are at vertex i at time t, (*i.e.*, if $X_t = i$), pick a road at random and follow it to the other end. You should be able to check for yourself that this MC is irreducible and aperiodic, and has the target distribution as its stationary distribution. So by the above theorem the MC will converge to this stationary distribution. Can you relate the condition of irreducibility to any property of the graph? What about aperiodicity?

"What's the big deal about it?" you might say. "Do we not already know (say from Kunte [1]) a much simpler method to do *exact* simulation from π? We can just write each vertex number on as many chits of paper as its degree, and then pick one chit at random." Yes, you can do that, and that *will* give you an exact simulation scheme. But to do that you need to know *all* the vertex degrees before you begin simulation; further, it will be practicable only if the number of required chits is small. There are many situations where one of these is not true. In [1], we have seen how randomization is used to start a game of cricket. Another game where randomization is much more important is any game of cards. You surely do not want the players to know beforehand which card is going to turn up when. So we need to start the game from a random permutation of the deck. For a deck of 52 cards, there are $52! \cong 8.065817 \times 10^{67}$ permutations.

You may wish to write these permutations on as many chits, put the chits in a hat, shake the hat and draw one chit at random. Excellent idea, but even if each chit is 5 inches long you will need about 6×10^{63} miles of paper! And the hat will weigh much more than 10^{60} kgs (even if you *can* procure such a hat)!

'Do we not already know a much simpler method to do *exact* simulation from π?

Shuffling is just a MCMC scheme.

But in practice do we at all try to do such a cumbersome thing? No! We simply shuffle the deck a number of times. Shuffling is just a MCMC scheme. The state space S consists of all the permutations, and is of size 52!. Our target distribution is uniform on S. You start from some initial permutation. Your move is as follows: Cut the deck into two (possibly unequal) halves. Riffle them together. Starting from the original deck permutation, this procedure takes you to a random deck permutation. So we have a Markov chain. Is it irreducible? I leave it for you to check that it is. Is the Markov chain aperiodic? It is! Here is how you can see it. Suppose that after you do the cut, you hold the top half in your right hand, and the bottom half in your left. Then there is a (very small) chance that all the cards in your left hand drop first, followed by those in your right hand. This will give you back your original deck. So now you know that if you go on doing this again and again, then eventually the deck will be random (a sample from uniform distribution).

Well, if that is all there is to it, why not settle for something simpler? Simply pick a card at random, and insert it into a random place in the remaining deck! Well, this will also work *in the limit*. But the problem here is that this scheme is very slow. Indeed, you may like to know that our usual shuffling scheme approaches the limit pretty fast (one calls it *rapidly mixing*), and for a deck of 52 cards 7 shuffles suffice. But beware! Less than 7 do not suffice in general. In other words, if the initial configuration had some hidden pattern just 5 or 6 shuffles are not enough to wash them off. I once read about a magic trick based on this. The magician starts with a specially prepared deck, lets the audience shuffle it 4 times. While the naive audience thinks that the deck is now pretty random, the magician knows that it is far from so, and utilizes the pattern to fool the audience.

Notice that the MCMC method does not need any in-

formation about how large the state space is. It also does not really need to know the exact numerical value of target distribution probabilities.

So far we have remained silent as to how to *devise* a MC on a given state space so that it will be irreducible, aperiodic and have a given stationary distribution. There are many ways to achieve that. The most popular one is called the *Metropolis–Hastings Scheme* (MH scheme), which we discuss below.

So far we have remained silent as to how to devise a MC on a given state space so that it will be irreducible, aperiodic and have a given stationary distribution.

Let $\mathcal{S}$ and π be given to us. Suppose that $\mathcal{S}$ is finite. To use the MH scheme we already need to have some irreducible MC on $\mathcal{S}$. This MC, often called the *proposal chain*, need not be aperiodic or have π as its stationary distribution. It is *any irreducible* MC. Let its transition matrix be $\mathbf{Q} = ((\mathbf{Q_{ij}}))$. Then the MH scheme is as follows: Start at any point. At each step make a tentative move using the proposal chain. If this proposes to take you from i to j compute

$$p = \frac{\pi_i \mathbf{Q_{ij}}}{\pi_j \mathbf{Q_{ji}}}.$$

Then move to j with probability $\min\{1, p\}$, else stay at i. Using arguments that we shall not present here, it can be shown that the resulting MC converges to π in the limit.

Note that π enters in the entire computation only as the ratio π_i/π_j. So it is enough to know π only up to a constant. Now we shall take a look at an example where MCMC is almost indispensable: Bayesian image analysis. In this example all the complications occur simultaneously: huge state space, complicated distribution, target distribution known only up to a constant. But be forewarned. The example, while appearing very realistic, is actually quite remote from real life. These methods are still not widely used in practice for doing image restoration.

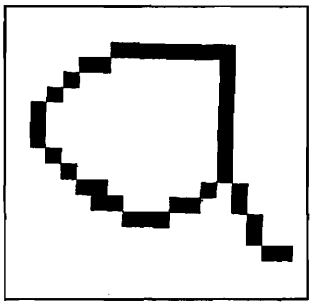


Figure 2.

MCMC and Gibbs Sampling

I assume that from [4], you have got an idea of how Bayesian inference works. Consider a digital black and white image like the one shown in *Figure* 2.

We know that it is a line drawing where the line is of 1 pixel width. So it is highly likely that most of the black pixels will have exactly 2 black neighbours, some will have 3 black neighbours, and very few will have other numbers of black neighbours. Similarly, most white pixels will have all white neighbours, some will have 2 or 3 black neighbours. We shall quantify this in terms of a *prior distribution* as follows. Let the number of black pixels with i black neighbours be B_i, and the number of white pixels with i black neighbours be W_i. Then we shall define our prior as

$$f(A) = \text{ constant} \times \exp(B_2 + B_3 + W_0 + W_2 + W_3).$$

Now suppose that the picture is contaminated with noise. The pixels change colour with probability 0.01 independently. Such things might happen if you are transmitting the picture as a series of electric pulses (high voltage $\equiv$ black, low voltage $\equiv$ white) and the receiver occasionally fails to detect the pulse voltages properly. We express this through the *model*

$$h(B|A) = \text{constant} \times (1/99)^{d(A,B)},$$

where $d(A, B)$ counts the number of pixel mismatches between the images A and B. What is the *posterior distribution* of the underlying *exact* image? It is

$$g(A) = f(A) \times h(B|A).$$

Note that this is *all* that we know about the underlying exact image. So the best that we can do is to generate a random image from this distribution. This distribution sits on the space of all images, a very large set. The

probability is known only up to a constant and that too is pretty complicated.

But we can still do MCMC as follows. Order the pixels rowwise. And start updating as follows. For each pixel find out the conditional probability that it should be black given the others. This is easily computed since it depends only on its neighbours. Generate the pixel from this distribution. Do this again and again. You will hope that after some iterations the noises will start to vanish, and the original image will show up (along with some distortion). A word of caution: With our over-simplified prior and model you may not see any convergence at all; however, with more polished priors and bigger pictures, we have a greater chance of success.

This version of MCMC is called Gibbs sampling (though the physicist Gibbs had absolutely nothing to do with it). It was so named because the method was first used to simulate from a distribution called the Gibbs distribution.

This version of MCMC is called Gibbs sampling (though the physicist Gibbs had absolutely nothing to do with it).

Next we mention something about the rate of convergence, *i.e.*, how long we should run the MC in order to get close enough to the target distribution. We shall do this only for homogeneous irreducible, aperiodic, reversible MC's on finite state spaces. Let our initial distribution be $\mathbf{v_0}$ and the transition matrix be $\mathbf{P}$. We know that

$$\mathbf{v_t} = \mathbf{v_{t-1}}\mathbf{P},$$

and that

$$\pi = \pi\mathbf{P}.$$

Thus π is a (left) eigenvector of $\mathbf{P}$ corresponding to eigenvalue 1. We know that all the eigenvalues of a real symmetric matrix are real. In the case of our transition matrix we can also show that they all lie in $(-1, 1]$, with exactly one eigenvalue equal to 1. The eigenvector

corresponding to that is π. Let the eigenvalues of $\mathbf{P}$ be

$$1 > \lambda_2 \geq \quad \lambda_k,$$

with corresponding (left) eigenvectors

$$\pi, \mathbf{u_2}, \quad , \mathbf{u_k}.$$

Now let us start from any initial vector $\mathbf{v_0}$, which we can write as

$$\mathbf{v_0} = \pi + \lambda_2 \mathbf{u_2} + \quad + \lambda_k \mathbf{u_k},$$

since $\{\pi, \quad , \mathbf{u_k}\}$ form a basis. Then

$$\mathbf{v_1} = \mathbf{v_0 P} = \pi + \lambda_2^2 \mathbf{u_2} + \quad + \lambda_k^2 \mathbf{u_k},$$

and, in general,

$$\mathbf{v_t} = \pi + \lambda_2^t \mathbf{u_2} + \quad + \lambda_k^t \mathbf{u_k}.$$

So we see that apart from π all the other terms are decaying away. The convergence rate of the MC depends on the rate of this decay. Different terms decay at different rates. The most die-hard term is either the $\mathbf{u_2}$ term or the $\mathbf{u_k}$ term depending on which of $|\lambda_2|$ and λ_k is bigger. So the number

$$1/(1 - \max(|\lambda_2|, |\lambda_k|))$$

measures how fast the MCMC converges. We often call it as the *mixing rate* of the MCMC algorithm.

Computing the eigenvalues is often not very easy. So people try to estimate them, or find bounds for them. Indeed, study of convergence of MCMC algorithms is a hot research topic.

Let us conclude this article by mentioning a close kin of MCMC that has been developed recently. These schemes are called *perfect sampling methods* owing to their ability to simulate from a distribution *exactly*. So

one does not have to worry about convergence. However, the price you pay for that is in algorithmic complexity. These algorithms tend to be much more complicated to design and to implement than MCMC algorithms. MCMC can be applied to a wide variety of problems. But unfortunately few of its properties are yet well known for MC's with infinite state spaces. But its wide applicability and ease of implementation have made it extremely popular.

Suggested Reading

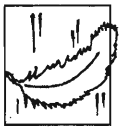
[1] S Kunte, Statistical Computing, Part 1, *Resonance*, Vol.4, No.10, pp.16-24, 1999; Part 2, Vol. 5, No.4, pp.19-27, 2000.

[2] P Hall and A Sarkar, Bootstrap Methods in Statistics, *Resonance*, Vol.5, No.9, pp.41-48, 2000.

[3] K B Athreya, The vacillating mathematician, *Resonance*, Part 1, Vol.2, No.1, pp.16-24, 1997; Part 2, Vol.2, No. 2, pp.34-40, 1997.

[4] Mohan Delampady and T Krishnan, Bayesian Statistics, *Resonance*, Vol.7, No.4, pp.27-39, 2002.

Address for Correspondence
Arnab Chakraborty
Department of Statistics
Stanford University
California, USA.



English language as a medium for higher education. - Japan benefited greatly by maintaining the English language in her universities from the very start. English helped Japan to maintain close touch in all departments of progress with two of the world's foremost nations, namely, great Britain and the United States of America. India should not throw away the advantages that she now possesses by retaining English, unless any great change in world conditions justifies the abandonment of its use.

It is one of the basic duties of Government to find occupations for the people, no matter to what community or party they may belong, who are willing to work but who are not able to find employment for themselves.

Nepotism and class preferences are a common fault in official life. The practice of appointing men of the same caste or the same region without regard to qualifications is often noticed. If such practices are not rooted out, the chances of India ever rising above the level of a second class State are slender.

By Sir M Visvesvaraya