Markov Chain Monte Carlo Methods

1. Simple Monte Carlo

K B Athreya, Mohan Delampady and T Krishnan

1. Introduction

In earlier articles in *Resonance* mentioned in the Suggested Reading, various authors discussed Monte Carlo simulation methods and their applications. In [1, 2], Markov Chain Monte Carlo (MCMC) was introduced with examples and its method of simulation explained. In this series of articles we describe MCMC methods in general and their rationale, discuss special cases of the MCMC algorithms and work out examples and applications in statistics, especially Bayesian statistics. In Part 1, we discuss the independent identically distributed (IID) Monte Carlo procedure (i.e., without a real Markov chain structure) with applications to integration including integration in a Bayesian context. In Part 2, we describe the algorithms of MCMC and explain how they work. In Part 3, we discuss some statistical preliminaries which are required to understand statistical applications of MCMC described in Part 4.

The most significant applications of MCMC are in Bayesian inference, an introduction to which was given in [3]. Bayesian inference is an alternative paradigm to classical frequentist method of inductive inference and derives its name from the application of Bayes theorem to derive inverse (or posterior) probability using data and prior probability. Thus it is significant that this series of articles on MCMC begins in this issue devoted to Thomas Bayes. The applications of MCMC to Bayesian inference will have to wait for the concluding part of this series.

Krishan B Athreya is a Professor at Cornell University. His research interests include mathematical analysis, probability theory and its application and statistics. He enjoys writing for *Resonance*. His spare time is spent listening to Indian classical music.

Mohan Delampady is at the Indian Statistical Institute, Bangalore. His research interests include robustness, nonparametric inference and computing in Bayesian statistics.

T Krishnan is now a full-time Technical Consultant to Systat Software Asia-Pacific (P) Ltd., in Bangalore, where the technical work for the development of the statistical software Systat takes place. His research interests have been in statistical pattern recognition and biostatistics.

Keywords
Markov chain, Monte Carlo sampling, Markov chain Monte Carlo, Bayesian statistics, numerical integration, law of large numbers, statistical simulation, importance sampling.

2. Numerical Integration

Evaluating integrals is an interesting exercise in calculus. Various tricks involving substitutions, trigonometric identities, transformations, etc. play an important role in this. Vast tables of well-known integral formulae are available. See for instance, [4] and website www.cs.berkeley.edu/fateman/htest.html. Nevertheless it is not difficult to come up with a function whose integral over an interval is very difficult to evaluate. And in many areas of applications of science and engineering one encounters functions over complicated domains whose integrals have great practical importance but are very difficult to evaluate analytically.

In such cases one has to resort to numerical integration. Here one uses the basic definition of integrals as limits of Riemann sums. One divides the given domain D into a grid of small subdomains $\{D_i\}_1^n$, evaluates the given function $f(.)$ at some point x_i in each of the subdomains D_i and then approximates the integral of f over $D, I \equiv \int_D f$ by $I_n \equiv \sum_{i=1}^{n} f(x_i)m(D_i)$, where $m(D_i)$ is the length, area or volume of D_i. Depending on the smoothness of the function and the geometry of the domain D one finds appropriate ways of subdividing D into subdomains $\{D_i\}_1^n$ so that the error in approximation, i.e., $I - I_n$ is not too large nor is the computational effort. This area of mathematics is called numerical analysis and is quite well developed. (See for instance [5].)

3. Monte Carlo Methods, the IID Case

An alternative to the above-mentioned method is the use of probability theory and in particular the method of *statistical sampling*. Public opinion polls, market surveys, etc. are based on this. In order to determine the proportion p of individuals in a given population that support a given party it is not necessary and is often not feasible to survey the entire population. The method of *statistical sampling* is the following: Select a sample n of individuals from the full population (of size N), determine the proportion p_n in the sample that supports the given party and use this as an *estimate* or *approximation* of the population proportion p. One needs to know how reliable this estimate p_n of p is. This is possible if the sample is drawn according to well-established statistical procedures and one can quantify the probabilities of the approximation error $(p_n - p)$ exceeding some prescribed levels. This is based on the famous *Laws of Large Numbers (LLN):* Let $X_1, X_2, X_3,$ be a sequence of independent identically distributed random variables with a finite expected value μ_X then the *sample mean*

$$\bar{X}_n = \frac{1}{n}\sum_1^n X_i \tag{1}$$

converges to μ_X as n becomes large. In particular, if $h(\cdot)$ is a bounded function and $Y_i \equiv h(X_i)$ then

$$\bar{Y}_n \equiv \frac{1}{n}\sum_1^n h(X_i) \tag{2}$$

converges to the expected value μ_Y of Y_1.

Here '*independent*' means that for any finite k, the probability of the event $\{X_i \leq x_i, i = 1, 2, \quad , k\}$ is equal to the product of the probabilities of the events $\{X_i \leq x_i\}, i = 1, 2, \quad k$, for each set $\{x_i\}_1^k$ of numerical values; '*identically distributed*' means that for any x, the probability $X_i \leq x$ does not change with i; the expected value μ_X of a random variable X is defined as $\sum_i a_i P(X = a_i)$ if X is a *discrete random variable* with values $\{a_i\}$ and $P(X = a_i)$ is the probability that $X = a_i$ and as $\int x f_X(x) dx$, if X is a *continuous random variable* with *probability density* $f_X(x)$, i.e., $P(x < X < x + h)$ is approximately equal to $f_X(x)h$ for h small. For a bounded function $h(\cdot)$ the expected value μ_Y of $Y = h(X)$ can be computed via the formula $\mu_Y = \sum_i h(a_i) P(X = a_i)$

in the discrete case and $= \int h(x) f_X(x) dx$ in the continuous case. Finally, the notion of convergence of $\bar{X}_n$ to μ_X is that of 'in probability', that is, for any $\epsilon > 0$, the probability that $|\bar{X}_n - \mu_X| > \epsilon$ goes to zero as $n \to \infty$.

The articles in *Resonance* [6, 7] have more details on these matters.

As another application of the LLN one gets the following procedure to approximate the integral I of a function f over a domain D. Let D be a domain in some Euclidean space $\mathbb{R}^k$. Let $f : D \to \mathbb{R}$ be a bounded function. Let $X_1, X_2,$ be a sequence of independent random variables that are *uniformly distributed over D*. That is, the probability X_i falls in region A_i equals the ratio of the volume of $A_i \cap D$ to the volume of D (here volume refers to the k-dimensional volume: for $k = 1$ it is length, for $k = 2$ it is area). Then by LLN

$$\frac{1}{n}\sum_1^n f(X_i) \to \frac{1}{\text{Vol}(D)} \int_D f(x) dx \tag{3}$$

in probability as $n \to \infty$. Thus an estimate I_n of $I \equiv \int_D f(x) dx$ is simply

$$I_n \equiv (\text{Vol}(D)) \frac{1}{n} \sum_1^n f(X_i). \tag{4}$$

If one wants to evaluate the integral of f with respect to a *mass distribution* $m(\cdot)$ over D, say, $J = \int_D f(x)\, m(x)\, dx$ then again by LLN

$$\frac{1}{n}\sum_1^n f(X_i) m(X_i) \to \frac{\int_D f(x)\, m(x)\, dx}{\text{Vol } D}$$

in probability as $n \to \infty$. Thus an estimate J_n of J is simply

$$J_n \equiv (\text{Vol } (D)) \frac{1}{n} \sum_1^n f(X_i)\, m(X_i). \tag{5}$$

This method depends on being able to generate a sample $\{X_i\}_1^n$ of i.i.d. uniform r.v. (independent, identically distributed uniform random variables over D) and the computation of $f(X_i)\ m(X_i)$ for each $i = 1, 2, . \ . n$.

Sometimes it may be easier to generate $\{X_i\}$ i.i.d with some nonuniform distribution, say, with density $g(\cdot)$, i.e., the probability of X_i falling in O_x is approximately $g(x)V(O_x)$, where $V(O_x)$ is the volume of O_x, a small neighborhood of x. In this case an estimate J_n of J is

$$J_n = \frac{1}{n}\sum_1^n \frac{f(X_i)m(X_i)}{g(X_i)} \tag{6}$$

since by the LLN J_n converges in probability to the expected value of $\frac{f(X_1)m(X_1)}{g(X_1)}$ which turns out to be

$$\int_D \frac{f(x)m(x)}{g(x)} g(x)\ dx = \int_D f(x)\ m(x)\ dx.$$

This reduces to (5) in the uniform case since, then $g(x) \equiv (\mathrm{Vol}\, D)^{-1}$. An added advantage with this method is that if $g(x)$ is approximately proportional to $f(x)m(x)$ for all $x \in D$, then the random variable $\frac{f(X_i)m(X_i)}{g(X_i)}$ will have a small variance σ^2 leading in turn to the variance of J_n to be $\frac{\sigma^2}{n}$.

The above procedures are known collectively as IID Monte Carlo. IID refers to the use of i.i.d. random variables. Monte Carlo is a city in Monaco, famous for its casinos offering games of chance. Games of chance exhibit random behaviour, much like the random variables generated for the statistical simulation exercises. Early ideas of probability and simulation were developed in the context of gambling here and hence these simulation techniques are known as Monte Carlo techniques. The theoretical basis for this method is the LLN and a further refinement of it is known as the central limit theorem (CLT). It says that if $X_1, X_2, .$ are i.i.d.r.v. with expectation μ and variance σ^2 (i.e. $\mathbb{E}(X - \mu)^2 = \sigma^2$

where $\mathbb{E}$ stands for expectation) then for each $a < b$, Prob.$(\mu + a\sigma/\sqrt{n} \leq \bar{X}_n \leq \mu + b\sigma/\sqrt{n})$ converges to $\int\limits_a^b \frac{1}{\sqrt{2}} e^{-\frac{x^2}{2}} dx = \Phi(b) - \Phi(a)$ as $n \to \infty$, where $\Phi(\cdot)$ is the standard normal distribution function

$$\Phi(t) \equiv \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{t} e^{-\frac{x^2}{2}} dx.$$

Thus, one can quantify the probability that the error $(\bar{X}_n - \mu)$ is of the magnitude $(\sqrt{n})^{-1}$ and use this to obtain confidence intervals of the form $CI \equiv (\bar{X}_n - Z_\alpha(\sqrt{n})^{-1}, \bar{X}_n + Z_\alpha(\sqrt{n})^{-1})$ where Z_α is defined by $\Phi(Z_\alpha) = 1 - \frac{\alpha}{2}$ for any $0 < \alpha < 1$. Then one can claim that the true value μ falls in CI with a probability that is approximately $1 - \alpha$. Typically one chooses α as 0.05.

4. Examples

Example 1: Consider the evaluation of the integral

$$J = \int_0^1 \cos(\frac{\pi x}{2}) dx.$$

Actually this is easily evaluated to be $\frac{2}{} \approx \frac{7}{11} \approx 0.63636$. Suppose we wish to evaluate the integral by Monte Carlo methods. Following the above discussion, we can use the estimate

$$J_n = \frac{1}{n} \sum_{i=1}^{n} \cos(\frac{\pi X_i}{2}),$$

where X_i are drawn from the uniform distribution on [0, 1]. We did that and the results and statistics are in *Table* 1 and in *Figure* 1.

To enhance the appeal of Monte Carlo integration methods, various techniques are used for making J_n more accurate. As remarked in Section 3, one such technique is choosing $g(x)$ to be approximately proportional to $f(x)m(x)$ for all x in the set D. In our example since $\cos\frac{x}{2}$ is approximately $1 - \frac{2x^2}{8}$ (two-term Taylor exp-

	U	Theoretical Values	V	Theoretical Values
No. of cases	10000	—	10000	—
Minimum	0.00006	0	-0.00055	0
Maximum	0.99	1	1.00000	1
Median	0.50459	0.5	0.70176	
Mean	0.50210	0.5	0.63374	0.63636
Standard Dev	0.28962	0.28868	0.30940	0.30822
Variance	0.08388	0.08333	0.09573	0.095

Table 1. Descriptive Statistics of Uniform = [0,1] and V=cos(π U/2) from 10000 samples of U.

ansion) and $\frac{2}{8}$ being nearly one, a reasonable $g(x)$ is a probability density proportional to $(1 - x^2)$, i.e., $\frac{3}{2}(1 - x^2)$. Results of such a simulation are given in *Table* 2 and *Figure* 2. This method is often called *importance sampling*, the function $g(x)$ being called the *importance function.* Here sampling is made efficient by drawing from regions of higher density using the importance function.

Example 2: As another example, consider the problem that Bayes addressed in his essay, and which has been explained in the last part of the 'Article in a Box' of this issue (p.5). Also recall Example 1 of [3]. In this problem, we are given data x from a binomial distribution with

Figure 1. (a) Histogram of U= Uniform [0,1] and (b) Histogram of V= cos(π U/2) from 10000 samples of U.

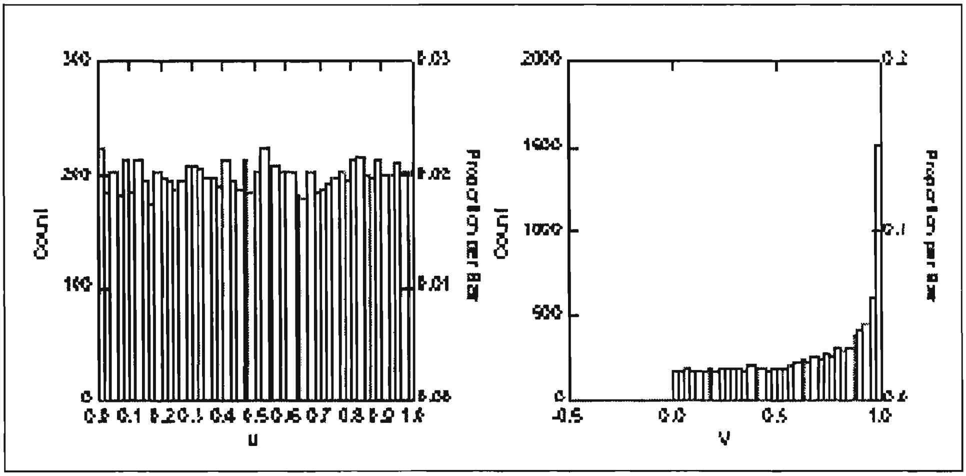

Table 2. Descriptive statistics of Y with density $3(1-y^2)/2$ and $Z = \cos(\pi Y/2)$ using density of Y based on 10000 samples.

	Y	Z
No. of cases	10000	10000
Minimum	0.00011	0.48907
Maximum	0.99417	0.66667
Median	0.34258	0.64848
Mean	0.37245	0.63674
Standard Dev	0.24234	0.03180
Variance	0.05873	0.00101

parameter θ, namely

$$P(X = x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}, \quad x = 0, 1, \quad , n.$$

The prior distribution for θ is given to be the uniform distribution in the interval [0, 1]. What is required in this situation is the posterior probability that the success rate (parameter) θ lies in a given interval (a, b):

$$\begin{aligned} P(a < \theta < b|x) &= \frac{\int_a^b \pi(\theta|x)\, d\theta}{\int_0^1 \pi(\theta|x)\, d\theta} \\ &= \frac{\int_a^b \theta^x(1-\theta)^{n-x}\, d\theta}{\int_0^1 \theta^x(1-\theta)^{n-x}\, d\theta}. \end{aligned}$$

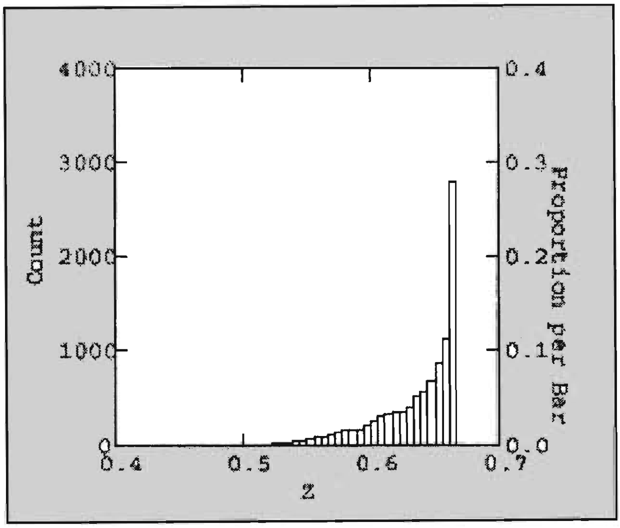

Figure 2. Histogram of $Z = \cos(\pi Y/2)$ using density $3(1-y^2)/2$ based on 10000 samples.

Suppose $n = 20$ and $x = 5$. We need an algorithm to (approximately) compute the above mentioned posterior probability, say, for the interval from 0.1 to 0.5. Note that this is just the probability of the interval $(0.1, 0.5)$ under the Beta$(6, 16)$ distribution. Thus this exercise involves the computation of what is known as the 'incomplete' beta integral, which has for a long time engaged the attention of numerical analysts. As an alternative to numerical analysis, we could employ the IID Monte Carlo technique for this purpose. For this, simply simulate i.i.d. random variables from the Beta$(6, 16)$ distribution, and take the sample proportion of these random variables which fall in the interval $(0.1, 0.5)$. This could be regarded as Monte Carlo integration in the sense of the previous sections, the integrand being just the indicator function of the interval $(0.1, 0.5)$. How does one simulate from the Beta$(6, 16)$ distribution? A well-known result in probability theory says that, if $Y_1 \sim \chi^2_6$, $Y_2 \sim \chi^2_{16}$ and they are independently distributed, then $Y_1/(Y_1 + Y_2) \sim$ Beta$(6, 16)$. Then the question arises as to how does one generate random samples from χ^2_{2k} distributions. It is easy to show that if $U_1, U_2, \quad, U_k$ are i.i.d. $\mathcal{U}[0, 1]$ (uniform distribution on the interval [0, 1]) variables then

$$-2\sum_{i=1}^{k} \ln U_i \sim \chi^2_{2k}.$$

This gives us a method of generating random samples from Beta$(6, 16)$ and carrying out the Monte Carlo integration, starting from draws from the uniform distribution, which most computer software help you do, using the well-known random number generation routines. We shall leave this as an exercise to the readers.

The computations cited in this article were carried out using Systat Statistical Software.

Suggested Reading

[1] Arnab Chakraborty, Markov Chain Monte Carlo: Examples, *Resonance*, Vol. 7, No. 3, pp. 25-34, 2002.

[2] Arnab Chakraborty, Markov Chain Monte Carlo: a method of simulation, *Resonance*, Vol.7 No. 5, pp. 66-75, 2002.

[3] Mohan Delampady and T Krishnan, Bayesian Statistics, *Resonance*, Vol. 7, No.4, pp. 27-38, 2002.

[4] I S Gradshteyn and I M Ryzhik, *Table of Integrals, Series, and Products*, Sixth Edition, San Diego: Academic Press, 2000.

[5] A R Krommer and C W Ueberhuber, *Computational Integration*, SIAM, 1998.

[6] R L Karandikar, On randomness and probability: How to model uncertain events mathematically, Vol. 1, No. 2, pp. 55-68, 1996. (Also in M Delampady, T Krishnan and S Ramasubramanian (Eds.) *Echoes from Resonance: Probability and Statistics*, Indian Academy of Sciences & Universities Press, pp. 12-26, 2001.)

[7] M Delampady and V R Padmawar, Sampling, Probability Models and Statistical Reasoning, *Resonance*, Vol. 1, No.5, pp. 49-58, 1996. (Also in M Delampady, T Krishnan and S Ramasubramanian (Eds.) *Echoes from Resonance: Probability and Statistics*, Indian Academy of Sciences & Universities Press, pp. 27-36, 2001.)

[8] Peter Hall and Abhinanda Sarkar, Bootstrap methods in statistics: resampling, *Resonance*, Vol. 5, No. 9, pp.41-48, 2002.

[9] Sudhakar Kunte, Statistical Computing: 1. Understanding Randomness and Random Numbers, *Resonance*, Vol.4, No.10, pp.16-21, 1999.

[10] Sudhakar Kunte, Statistical Computing: 2. Technique of Statistical Simulation, *Resonance*, Vol. 5, No.4, pp. 18-27, 2000.

Address for Correspondence
K B Athreya
School of ORIE
Rhodes Hall
Cornell University, Ithaca
New York 14853, USA.

Mohan Delampady
Indian Statistical Institute
8th Mile, Mysore Road
Bangalore 560 059, India.

T Krishnan
Systat Software Asia-Pacific
(P) Ltd., Floor 5, 'C' Tower
Golden Enclave, Airport Road
Bangalore 560 017, India.

The most beautiful experience we can have is the mysterious. It is the fundamental emotion which stands at the cradle of true art and true science.

Albert Einstein
(1879-1955)
US Physics, born in Germany