Markov Chain Monte Carlo Methods

2. The Markov Chain Case

K B Athreya, Mohan Delampady and T Krishnan

K 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 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.

1. Introduction

In Part 1[1] we considered Monte Carlo methods based on independent identically distributed random draws from specified distributions, which we called the IID case. In this Part 2, we carry the methodology further by considering random draws following the Markovian regime. In Part 4, we discuss some applications of the Markov chain Monte Carlo (MCMC) method in some statistical problems wherein the IID Monte Carlo is not applicable. In Part 3, we discuss some statistical preliminaries required for an understanding of the statistical issues involved.

2. Monte Carlo Methods: the Markov Chain Case

The main theoretical basis for the IID Monte Carlo method is the law of large numbers (LLN). It turns out LLN remains valid even if we drop the assumption of $\{X_i\}$ being i.i.d. but have some weak dependence. An example of such a sequence is the case of a Markov chain. A sequence of random variables $\{X_n\}_{n \ge 0}$ is called a Markov chain if for any n given the current value, i.e., X_n, the past, i.e., $\{X_j : j \le n-1\}$ and the future, i.e., $\{X_j : j \ge n+1\}$ are statistically independent in the sense

$$\text{Prob}\,(A \cap B|X_n) = \text{Prob}\,(A|X_n)\ \text{Prob}\,(B|X_n); \quad (1)$$

where A and B are events defined respectively in terms of the past and the future, and $P(A|X_n)$ is the conditional probability of A given X_n. Thus the sequence $\{X_n\}_{n \ge 0}$ has no memory. Given the present, the prob-

[1] Simple Monte Carlo, *Resonance*, Vol.8, No.4, p.17, 2003.

abilistic nature of future evolution does not depend on the past. Among Markov chains there is a subclass that has wide applicability. This is the case of Markov chains with time homogeneous or stationary transition probabilities. This means that the probability distribution of X_{n+1} given $X_n = x$, and the past $X_j : j \le n-1$ depends only on x and does not depend on the values of $X_j : j \le n-1$ or n. In particular, if the set S of values $\{X_n\}$ can take, known as the state space is finite or countable, this reduces to specifying a matrix known as the transition probability matrix $P \equiv ((p_{ij}))$ where for any two values i, j in S, p_{ij} is the probability that $X_{n+1} = j$ given $X_n = i$, i.e., of moving from state i to state j in one time unit. If the state space S is not countable then one has to specify a transition kernel or transition function $P(x, \cdot)$, where $P(x, A)$ is the probability of moving from x into A in one step, i.e., $P(X_{n+1} \in A | X_n = x)$. Given the transition probability and the probability distribution of the initial value X_0 one can construct the joint probability distribution of $\{X_j : 0 \le j \le n\}$ for any finite n. For example, in the countable state space case

$$\begin{aligned} P(X_0 &= i_0, X_1 = i_1, \ldots, X_n = i_n) \\ &= P(X_0 = i_0) p_{i_0 i_1} p_{i_1 i_2} \ldots p_{i_{n-1} i_n} \end{aligned} \qquad (2)$$

as can be shown by using the Markov property and induction. Indeed

$$\begin{aligned} & P(X_0 = i_0, X_1 = i_1, \ldots, X_{n-1} = i_{n-1}, X_n = i_n) \\ = \; & P(X_n = i_n | X_0 = i_0, \ldots, X_{n-1} = i_{n-1}) \\ & \quad \times P(X_0 = i_0, X_1 = i_1, \ldots X_{n-1} = i_{n-1}) \\ = \; & p_{i_{n-1} i_n} P(X_0 = i_0, \ldots, X_{n-1} = i_{n-1}). \end{aligned}$$

A probability distribution π is called stationary or invariant for a transition probability P or the associated Markov chain $\{X_n\}$ if the probability distribution of X_0 is π then the same is true for all $n \ge 1$. Thus in the countable state space case a probability distribution

Keywords
Markov chain, state space, stationary transition probability, stationary distribution, irreducibility, aperiodicity, stationarity, M-H algorithm, proposal distribution, acceptance probability, image processing, Gibbs sampler.

$\pi = \{\pi_i : i \in S\}$ is stationary for a transition probability matrix P if for each j in S,

$$P(X_1 = j) = \sum_i P(X_1 = j|X_0 = i)P(X_0 = i)$$
$$= \sum_i \pi_i p_{ij} = P(X_0 = j) = \pi_j; \qquad (3)$$

i.e., for each j, $\pi_j = \sum_i \pi_i p_{ij}$. In vector notation it says $\pi = (\pi_1, \pi_2, \ldots)$ is a left eigenvector of the matrix P with eigenvalue 1 and

$$\pi = \pi P. \qquad (4)$$

Similarly, if S is a continuum a probability distribution π with density $p(x)$ is stationary for the transition kernel $P(\cdot,\cdot)$ if

$$\pi(A) = \int_A p(x)dx = \int_S P(x, A)p(x)dx$$

for all $A \subset S$.

In order to state LLN for Markov chains one more notion is needed. A Markov chain $\{X_n\}$ with a countable state space S and transition probability matrix $P \equiv ((p_{ij}))$ is said to be irreducible if for any two states i and j the probability of the Markov chain visiting j starting from i is positive. That is, for some $n \geq 1, p_{ij}^{(n)} \equiv P(X_n = j|X_0 = i)$ is strictly greater than zero. There is a similar notion of irreducibility for the general state space case. It is known as Harris or Doeblin irreducibility. For details on this somewhat advanced topic see Robert and Casella [1].

LLN for Markov Chains: Let $\{X_n\}_{n \geq 0}$ be a Markov chain with a countable state space S and a transition probability matrix P. Suppose it is irreducible and there is a stationary probability distribution $\pi \equiv (\pi_i : i \in S)$. Then, for any bounded function $h : S \to \mathbb{R}$ and for any initial distribution of X_0

$$\frac{1}{n}\sum_{i=0}^{n-1} h(X_i) \to \sum_j h(j)\pi_j \qquad (5)$$

in probability as $n \to \infty$.

A similar LLN holds when the state space S is not countable. The limit value in (5) will be the integral of h with respect to the stationary distribution π. A sufficient condition for the validity of this LLN is that the Markov chain $\{X_n\}$ be Harris irreducible and have a stationary distribution π. There are other sufficient conditions as well.

Notice that the limit is the expected value of $h(X_0)$ when X_0 has the distribution π. That is, the 'time average' converges to the space average (in steady state) as $n \to \infty$. The same law of large numbers holds for the general state space case assuming Harris irreducibility.

How does this LLN get used in applications?

Here is the essence of the Markov Chain Monte Carlo (MCMC) method.

Given a probability distribution π on a set S, and a function h on S suppose it is desired to compute the 'integral of f with respect to π', which reduces to $\sum_j f(j)\pi_j$ in the countable case. One looks for an irreducible Markov chain $\{X_n\}$ with S as its state space and π as its stationary distribution. Then, starting from some initial value X_0, run the Markov chain $\{X_j\}$ for a period of time, say $0, 1, 2, \ldots n-1$ and offer as an estimate

$$\mu_n = \frac{1}{n}\sum_{0}^{n-1} f(X_j). \qquad (6)$$

By the LLN (5), this estimate μ_n will be close to $\sum_j h(j)\pi_j$ for large n.

This method is known as Markov Chain Monte Carlo (MCMC). In particular, if one is interested in $\pi(A) \equiv \sum_{j \in A} \pi_j$ for some subset $A \subset S$ then by LLN (5) this reduces to

$$\hat{\pi}_n(A) \equiv \frac{1}{n}\sum_{0}^{n-1} I_A(X_j) \to \pi(A)$$

in probability as $n \to \infty$; where $I_A(X_j) = 1$ if $X_j \in A$ and 0 otherwise. In other words $\hat{\pi}_n(A)$ is the sample proportion of visits to A during $\{0, 1, 2, \ldots, n-1\}$ by the Markov chain.

An irreducible Markov chain $\{X_n\}$ with a countable state space S is called aperiodic if for some $i \in S$ the greatest common divisor, g.c.d.$\{n : p_{ii}^{(n)} > 0\} = 1$. Then, in addition to the LLN (5), the following result on the convergence of $P(X_n = j)$ holds, namely, that:

$$\sum_j |P(X_n = j) - \pi_j| \to 0 \qquad (7)$$

as $n \to \infty$, for any initial distribution of X_0. This means that for large n the probability distribution of X_n is close to π

There is a result similar to (7) for the general state space case that asserts that under suitable conditions, the probability distribution of X_n is close to π as $n \to \infty$.

This suggests that instead of doing one run of length n, one could do N independent runs each of length m so that $n \simeq Nm$ and then from the i^{th} run use only the m^{th} observation, say, $X_{m,i}$ and offer the estimate

$$\hat{\mu}_{N,m} \equiv \frac{1}{N}\sum_{i=1}^{N} f(X_{m,i}). \qquad (8)$$

There are other variations as well.

3. Special Markov Chains Used in MCMC

3.1 Metropolis–Hastings Algorithm

Let S be a finite or countable set. Let π be a probability distribution on S. (π is called the target distribution.) Let $Q \equiv ((q_{ij}))$ be a transition probability matrix such that for each i, it is computationally easy to generate a sample from the distribution $\{q_{ij} : j \in S\}$. Then generate a Markov chain $\{X_n\}$ as follows.

Step 1: If $X_n = i$, first sample from the distribution $\{q_{ij} : j \in S\}$ and denote that observation Y_n.

Step 2: Choose X_{n+1} from the two values X_n and Y_n according to the probability distribution

$$P(X_{n+1} = Y_n | X_n, Y_n) = \rho(X_n, Y_n)$$
$$P(X_{n+1} = X_n | X_n, Y_n) = 1 - \rho(X_n, Y_n); \tag{9}$$

where the 'acceptance probability' $\rho(\cdot, \cdot)$ is given by

$$\rho_{ij} = \rho(i, j) = \rho(x_i, x_j) = \min\left\{ \frac{\pi_j q_{ji}}{\pi_i q_{ij}}, 1 \right\} \tag{10}$$

for all (i, j) such that $\pi_i q_{ij} > 0$. It is not difficult to verify that $\{X_n\}$ is a Markov chain with transition probability matrix $P = ((p_{ij}))$ given by

$$p_{ij} = \begin{cases} q_{ij}\rho_{ij} & j \neq i \\ 1 - \sum_{k \neq i} p_{ik}, & j = i. \end{cases} \tag{11}$$

Q is called the 'proposal transition probability' and ρ the 'acceptance probability'. A most useful feature of this transition mechanism P is that P and π satisfy the so called detailed balance condition:

$$\pi_i p_{ij} = \pi_j p_{ji} \quad \text{for all } i, j. \tag{12}$$

This implies that for any j

$$\sum_i \pi_i p_{ij} = \pi_j \sum_i p_{ji} = \pi_j. \tag{13}$$

That is, π is a stationary probability distribution for P.

Now assume that S is irreducible w.r.t. Q and $\pi_i > 0$ for all i in S. Then it can be shown that P is irreducible and since it has a stationary distribution π and LLN (5) is available. This algorithm is thus a very flexible and useful one. The choice of Q is subject only to the condition that S is irreducible w.r.t. Q. Clearly, it is no loss of generality to assume that $\pi_i > 0$ for all i in S. A sufficient condition for the aperiodicity of P is that $p_{ii} > 0$ for some i or equivalently

$$\sum_{j \neq i} q_{ij} \rho_{ij} < 1.$$

A sufficient condition for this is that there exists a pair (i, j) such that $\pi_i q_{ij} > 0$ and $\pi_j q_{ji} < \pi_i q_{ij}$.

Recall that if P is aperiodic then both the LLN (5) and (7) hold. If S is not finite or countable but is a continuum and the target distribution $\pi(\cdot)$ has a density $p(\cdot)$ then one proceeds as follows: Let Q be a transition function such that for each x, $Q(x, \cdot)$ has a density $q(x, y)$. Then proceed as in the discrete case but set the 'acceptance probability' $\rho(x, y)$ to be

$$\rho(x, y) = \min\left\{ \frac{p(y)q(y, x)}{p(x)q(x, y)}, 1 \right\}$$

for all (x, y) such that $p(x)q(x, y) > 0$. Another useful feature of the above algorithm is that it is enough to know $\{\pi_i\}$ upto a multiplicative constant as in the definition of 'the acceptance probability' $\rho(\cdot, \cdot)$, only the ratio $\frac{\pi_j}{\pi_i}$ need to be calculated. This is useful in (i) Bayesian statistical applications of MCMC for calculating the moments of the posterior distribution of the parameters given the data; and (ii) in image processing and statistical mechanics where the set S consists of configurations over a multidimensional grid of pixels where in each pixel there is a fixed number of levels and the probability distribution is specified via a Gibbs potential function whose normalising constant ('partition function') is not easy to compute as shown below in a simple example.

Example 1 (Statistical Mechanics/Image Processing) Consider a 3×3 square of nine pixels and suppose in each pixel the colour level is either a 0 or 1. A typical configuration w is a vector $w = (w_{ij} : 1 \leq i, j \leq 3)$ with 9 coordinates and each w_{ij} is 0 or 1.

Thus, the set S of all configurations has $2^9 = 512$ elements. Suppose there is a 'potential function' $V : S \to (0, \infty)$ and for each w

$$\pi(w) = c\, e^{-\beta V(w)},$$

where $\beta > 0$ is a known constant and c is the normalising constant $(Z(\beta))^{-1}$ with $Z(\beta)$ equalling $\left(\sum_{w \in S} e^{-\beta V(w)}\right)$, known as the 'partition function'. For each configuration w define its neighbourhood $N(w)$ as the set consisting of all configurations w' such that w and w' differ exactly at one pixel. Thus in our example each w will have 9 neighbours. Now suppose in the Metropolis–Hastings algorithm for this context 'proposal transition Q' is the 'random walk' transition, i.e.,

$$q_{w,w'} = \frac{1}{9} \quad \text{if } w' \in N(w)$$

and 0 otherwise.

Here the 'acceptance probability' $\pi(w, w')$ reduces to

$$\pi(w, w') = \min\{e^{-\beta(V(w') - V(w))}, 1\} \quad \text{for } w' \in N(w).$$

Thus, for the actual transition P, at each step, given the present configuration w, choose one of the neighbours w' of w at random with uniform distribution and move there if $V(w') > V(w)$, otherwise stay at w. That is, the configuration at the next step is w' if $V(w') > V(w)$

or w itself. If $V(\cdot)$ is not a constant function then one can show that P is aperiodic. Since 'random walk Q' is irreducible, P is also so. So P is irreducible, aperiodic and has π as its stationary distribution.

Starting from some initial configuration, say, $w_{ij} = 0$ for all i, j, one can run this Markov chain $\{X_n\}$ by the easy-to-run Markov chain and estimate $\sum_{w \in S} h(w)\pi(w)$ for any function $h(\cdot)$ on S by $\frac{1}{n}\sum_{1}^{n} h(X_j)$.

Gibbs Sampler

In many applications such as statistical mechanics, image processing, Bayesian statistics the target distribution π is a high-dimensional multivariate distribution. For example, in the simple example of the last section each configuration has nine coordinates. In image processing typically one has $N \times N$ square with $N = 256$ and each pixel has $k \geq 2$ possible values. Thus each configuration has $(256)^2$ components and S has $k^{(256)^2}$ configurations. To simulate a random configuration from a target distribution π over such a large S is not an easy task. The Gibbs sampler generates an irreducible aperiodic Markov chain with π as its stationary distribution. To run this Markov chain it suffices to generate observations from univariate distributions. We now describe the Gibbs sampler in the context of a bivariate probability distribution. Let π be a probability distribution of a bivariate random vector (X, Y). For each x, let $P(x, \cdot)$ be the conditional probability distribution of Y given $X = x$. Thus, if π is a discrete probability distribution, $P(x_i, y_j)$ is the conditional probability that $Y = y_j$ given $X = x_i$. Similarly, let $Q(y, \cdot)$ be the conditional probability distribution of X given $Y = y$. Notice that for each x, $P(x, \cdot)$ is a univariate distribution and for each y, $Q(y, \cdot)$ is also a univariate distribution. Now generate a bivariate Markov chain $Z_n = (X_n, Y_n)$ as follows:

Start with some $X_0 = x_0$. Then generate an observation Y_0 from the distribution $P(x_0;\cdot)$. Next generate an observation X_1 from $Q(Y_0;\cdot)$. Next generate an observation Y_1 from $P(X_1;\cdot)$ and so on. At stage n if $Z_n = (X_n;Y_n)$ is known then generate X_{n+1} from $Q(Y_n;\cdot)$ and Y_{n+1} from $P(X_{n+1};\cdot)$. If π is a discrete distribution concentrated on $\{(x_i;y_j) : 1 \le i \le K; 1 \le j \le L\}$ and if $\pi_{ij} = \pi(x_i;y_j)$ then $P(x_i;y_j) = \frac{\pi_{ij}}{\pi_{i\cdot}}$ and

$$Q(y_j;x_i) = \frac{\pi_{ij}}{\pi_{\cdot j}};$$

where $\pi_{i\cdot} = \sum_j \pi_{ij}$; $\pi_{\cdot j} = \sum_i \pi_{ij}$. Thus the transition probability R for the $\{Z_n\}$ chain is

$$\begin{aligned} r_{(i;j);(k;\ell)} &= Q(y_j;x_k)P(x_k;y_\ell) \\ &= \frac{\pi_{kj}}{\pi_{\cdot j}}\frac{\pi_{k\ell}}{\pi_{k\cdot}}. \end{aligned}$$

It can be verified that this chain is irreducible, aperiodic and has π as its stationary distribution. Thus LLN (5) and (7) hold in this case. Thus for large n, Z_n can be viewed as a sample from a distribution that is close to π and one can approximate $\sum_{i;j} h(i;j)\pi_{ij}$ by $\frac{1}{n}\sum_{1=1}^{n} h(X_i;Y_i)$. The multivariate extension of the above is very straightforward. If π is a probability distribution of a k-dimensional random vector $(X_1;X_2;\ldots;X_k)$ let $\pi_i(\cdot|\underset{\sim}{x}_{-i} \equiv (x_1;x_2;x_{i-1};x_{i+1}\ldots x_k))$ denote the univariate conditional distribution of X_i given that $\underset{\sim}{X}_{-i} \equiv (X_1;X_2;X_{i-1};X_{i+1};\ldots;X_k) = \underset{\sim}{x}_{-i}$. Now starting with some initial value for $\underset{\sim}{X}_0 = (x_{01};x_{02};\ldots;x_{0k})$ generate $\underset{\sim}{X}_1 = (X_{11};X_{12};\ldots;X_{1k})$ sequentially by generating X_{11} according to the univariate distribution $\pi_1(\cdot|\underset{\sim}{x}_{0-1})$ and then generating X_{12} according to the univariate distribution $\pi_2(\cdot|X_{11};x_{03};x_{04};\ldots;x_{0k})$ and so on.

This method also works for continuous distributions. We illustrate this by considering the case of a bivariate normal distribution. Let $\begin{pmatrix} X \\ Y \end{pmatrix}$ be a bivariate (normal) Gaussian random vector with 0 means and unit variances and correlation coefficient ρ. We write this as

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}; \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right).$$

It is a property of the bivariate normal distribution that the conditional distribution X given $Y = y$ is univariate normal with mean ρy and variance $(1 - \rho^2)$, i.e.,

$$(X|Y = y) \sim N(\rho y; 1 - \rho^2)$$

and similarly

$$(Y|X = x) \sim N(\rho x; 1 - \rho^2).$$

Using this property, Gibbs sampling proceeds as follows to generate $(X_n; Y_n); n = 0; 1; 2; \ldots$.

1. Start from a arbitrary value x_0 for X_0.

   Repeat the following steps for $i = 0; 1; \ldots; n$.

2. Given x_i for X, draw a random sample from $N(\rho x_i; 1 - \rho^2)$ and denote it Y_i.

3. Given y_i for Y, draw a random sample from $N(\rho y_i; 1 - \rho^2)$ and denote it X_{i+1}.

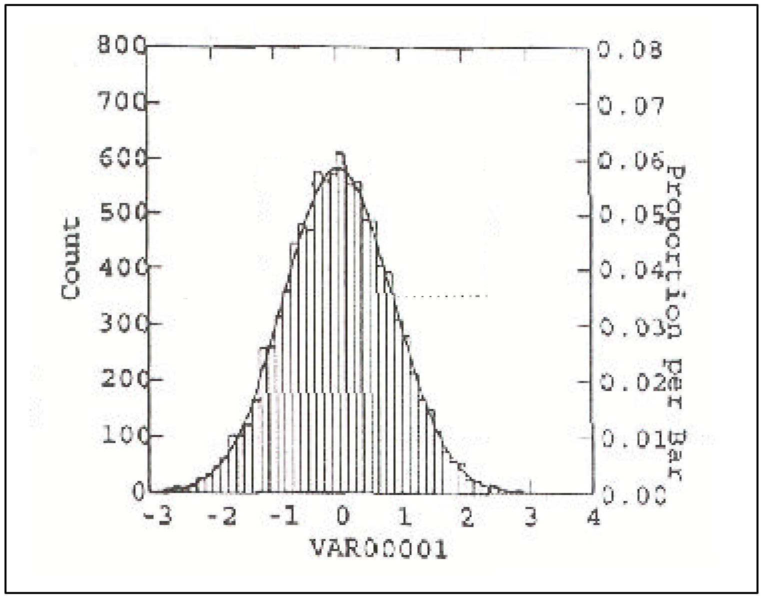
Thus we have $(x_i; y_i); i = 0; 1; 2; \ldots; n$. The theory of Gibbs Sampling tells us that if n is large then $(x_n; y_n)$ can be considered to be a random sample whose probability distribution is close to

$$N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}; \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right).$$

We used this Gibbs sampling to generate a random sample of 1000 from

$$N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}; \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}\right);$$

Figure 1.



(i.e., $\rho = 0.5$) using $n = 1000$. We present the histogram of these 1000 values of x, representing the marginal distribution of X, overlaid with the standard normal curve, which is the theoretical marginal distribution of X in Figure 1. We notice that the histogram is a very good approximation to the standard normal curve.

Why does Gibbs sampling work? Recall that a sufficient condition for the LLN (5) and the limit result (7) is that an appropriate irreducibility condition holds and a stationary distribution exists. We shall verify the latter for the chain $\{X_n\}$ in our example. From steps 2 and 3 of the procedure outlined above, it is not difficult to see that one can write

$$Y_i = \rho X_i + \sqrt{1-\rho^2}\,\eta_i$$

and

$$X_{i+1} = \rho Y_i + \sqrt{1-\rho^2}\,\xi_i;$$

where η_i and ξ_i are independent $N(0;1)$ (standard normal) random variables and independent of X_i. Thus the sequence $\{X_i\}$ satisfies the stochastic difference equation

$$X_{i+1} = \rho^2 X_i + U_{i+1};$$

where

$$U_{i+1} = \rho\sqrt{1-\rho^2}\,\epsilon_i + \sqrt{1-\rho^2}\,\xi_i.$$

Since ϵ_i, ξ_i are independent $N(0,1)$ random variables, U_{i+1} being a linear combination of ϵ_i and ξ_i is also a normally distributed random variable with mean 0 and variance $\rho^2(1-\rho^2) + (1-\rho^2) = 1-\rho^4$. Also $\{U_i\}_{i\geq 1}$ being i.i.d., makes $\{X_i\}_{i\geq 0}$ a Markov chain. It turns out that the irreducibility condition holds here. Turning to stationarity, note that if X_0 is an $N(0,1)$ random variable, then $X_1 = \rho^2 X_0 + U_1$ is also a $N(0,1)$ random variable, since the variance of X_1 is $\rho^4 + 1 - \rho^4 = 1$ and the mean of X_1 is 0. This makes the standard $N(0,1)$ distribution a stationary distribution for $\{X_n\}$. Thus both the LLN and the convergence of the distribution of X_n, i.e., the analogues of (5) and (7) are valid here. Hence one can expect the histogram of the sample $\{x_i\}_1^n$ to be close to the standard $N(0,1)$ density and also regard the observation x_n as a sample from a distribution close to $N(0,1)$.

A good reference for MCMC methods is the recent book of Robert and Casella [1]. For Markov chains on discrete spaces, see Hoel, Port and Stone [2]. For Markov chains on general state spaces, see Meyn and Tweedie [3]. For some interesting examples of Markov chains, see Athreya [4].

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

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 Ltd.
Floor 5, 'C' Tower
Golden Enclave, Airport Road
Bangalore 560 017, India.

Suggested Reading

[1] C P Robert and G Casella, *Monte Carlo Statistical Methods* (New York, Springer-Verlag), 1999.

[2] P G Hoel, S C Port and C J Stone, *Introduction to Stochastic Processes* (Upper Saddle River, N.J., Prentice-Hall) Eastern Edition, 1972.

[3] S Meyn and R L Tweedie, *Markov Chains and Stochastic Stability* (New York, Springer-Verlag), 1993.

[4] K B Athreya, The vacillating mathematician: A stochastic version, *Resonance*, Vol.2, No.2, February, pp.34-40, 1997. (Also in M Delampady, T Krishnan and S Ramasubramanian (Eds.) *Echoes from Resonance: Probability and Statistics* (Indian Academy of Sciences and Universities Press) pp. 83-91, 2001.