How Markov's Little Idea Transformed Statistics

Jeffrey S. Rosenthal

Abstract We discuss Andrey Andreyevich Markov's early 20th century idea of a Markov chain, which generalized the traditional notion of independent random variables to a model that was more general but still mathematical tractable. We then describe how that led to the hugely popular modern statistical approach of using Markov chain Monte Carlo algorithms to estimate complicated quantities such as Bayesian posterior distributions.

Introduction

It was the beginning of the 20th century, and Andrey Andreyevich Markov was not happy. The field of mathematical probability was flourishing, with seminal contributions from de Moivre and Fermat through to Gauss, Laplace, Cauchy, Poisson, Chebyshev, and many others. And yet, most of the key results – laws of large numbers, the central limit theorem, and so on – were always about independent random variables, i.e. random quantities which have no effect on each other. This assumption seemed very limiting and unrealistic. But without this assumption, the random variables could behave however they wanted, and it seemed impossible to make mathematical progress. How could the independence assumption be generalized, while still allowing for valid theorems to be proven?

Markov eventually settled on a model in which each random variable depended on the others, but only through the adjacent ones. That is, a sequence of jointly-defined random variables $X_0, X_1, X_2, \ldots$ form a Markov chain if the conditional distribution of X_n, conditional on all the other variables, is completely determined by X_{n-1} and X_{n+1}. This setup was more general than independence, but still tractable enough to analyze mathematically.

Jeffrey S. Rosenthal
University of Toronto, Canada, e-mail: jeff@math.toronto.edu

Markov published his results in 1906 [6], adding a new model to the world of probability theory. He proved a version of the law of large numbers, and later of the central limit theorem, for his new concept of Markov chains. In the decades following, many mathematical papers about Markov chains were published. And Markov chains had some success as models of card shuffling, diffusions, branching processes, and later to stock prices. And yet, genuine applications remained somewhat limited. In fact, Markov's original example had been the pattern of consonants and vowels in written language, but they surely does *not* really follow a Markov chain. For example, a vowel is much less likely to occur of the last three letters were all vowels, than if just the last one letter were a vowel. It was challenging to find real-world examples of random sequences in which X_n was indeed influenced by X_{n-1}, but completely conditionally independent of X_{n-2}.

And then, in the latter half of the 20th century, a whole new application of Markov chains emerged, in the burgeoning field of computer simulation. In 1953, Metropolis et al. [7] used a simple iterative computer simulation to study the behavior of interacting molecular systems, for up to 224 particles. This was an early application of a Monte Carlo algorithm which uses random simulation to estimate complicated quantities. In doing so, they created the Metropolis algorithm, the world's first Markov chain Monte Carlo (MCMC) algorithm. Their algorithm was later generalized to the Metropolis-Hastings algorithm [5], the Gibbs sampler [3, 11, 2], and many other variations, see e.g. [12]. The term "Markov chain Monte Carlo (MCMC)", emphasizing that these new algorithms applied Markov's model to the Monte Carlo paradigm, originated in [4]. MCMC algorithms are now hugely popular in many branches of statistics, especially in Bayesian inference, as well as many other fields from molecular dynamics to engineering to machine learning to financial modeling. The volume [1] contains thousands of references, and a casual Google search for "Markov chain Monte Carlo" (in quotation marks) returns over two million web pages, indicating just how far Markov's little idea has come.

In this paper, we review Markov chains and MCMC algorithms, and explain how they have been so useful in modern statistical analysis.

Random Variables

A random variable is a mathematical model of an uncertain quantity. It is defined on some underlying probability space $(\Omega, \mathscr{F}, \mathbf{P})$, where Ω is an underlying sample space, $\mathscr{F}$ is a σ-algebra of subsets of Ω, and $\mathbf{P}$ is a probability measure on $(\Omega, \mathscr{F})$. A random variable is then a measurable function from $(\Omega, \mathscr{F})$ to a measurable range space $(\mathscr{X}, \mathscr{G})$, most commonly to $(\mathbf{R}, \mathscr{B})$ (the real numbers with the Borel subsets). The probability that a random variable X takes a value within a subset $A \in \mathscr{G}$ is then given by

$$\mathbf{P}[X \in A] \; := \; \mathbf{P}\{\omega \in \Omega : X(\omega) \in A\}. \tag{1}$$

(If Ω is finite or countable, then we can let $\mathscr{F}$ be the collection of all subsets of Ω, and simply take $\mathbf{P}[X \in A] = \sum\{\mathbf{P}(\omega) : X(\omega) \in A\}$, i.e. define probabilities by a simple sum over individual points.)

Individual random variables can be studied in terms of their probability distributions $\{\mathbf{P}[X \in A] : A \in \mathscr{G}\}$, including such familiar notions as binomial, normal, exponential, Poisson, and other univariate distributions, perhaps extended to distributions of functions h of random variables as $\{\mathbf{P}[h(X) \in A] : A \in \mathscr{G}\}$.

The situation gets more intricate when there are multiple random variables defined jointly on the same underlying space $(\Omega, \mathscr{F}, \mathbf{P})$. If X_1 and X_2 are two such random variables, then we can define their joint probability distribution by

$$\mathbf{P}[X_1 \in A,\ X_2 \in B] \ = \ \sum\{\mathbf{P}(\omega) : X_1(\omega) \in A \text{ and } X_2(\omega) \in B\}, \tag{2}$$

for any $A, B \in \mathscr{G}$. This leads to questions about the relationship between the random variables.

The simplest case is independent random variables, meaning random quantities which do not affect each other, or formally for which

$$\mathbf{P}[X_1 \in A,\ X_2 \in B] \ = \ \mathbf{P}[X_1 \in A]\ \mathbf{P}[X_2 \in B] \tag{3}$$

for all $A, B \in \mathscr{G}$. More generally, a collection $\{X_i\}_{i \in I}$ of random variables are all independent if for all subsets $I_0 \subseteq I$, and all choices $\{A_i\}_{i \in I_0}$ of elements of $\mathscr{G}$, we have

$$\mathbf{P}[X_i \in A_i \text{ for all } i \in I_0] \ = \ \prod_{i \in I_0} \mathbf{P}[X_i \in A_i]. \tag{4}$$

Related, a collection $\{X_i\}_{i \in I}$ is identically distributed if for each fixed $A \in \mathscr{F}$, the value $\mathbf{P}[X_i \in A]$ is the same for all $i \in I$.

If a sequence $X_1, X_2, \ldots$ of random variables is both independent and identically distributed (i.i.d.), then many results are known. For example, the law of large numbers says that the sample averages $\frac{1}{n}\sum_{i=1}^{n} h(X_i)$ converge to their common mean $\mu := \mathbf{E}[h(X_i)]$, and the central limit theorem says $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[h(X_i) - \mu]$ converges in distribution to a Normal$(0, \sigma^2)$ distribution if $\sigma^2 := \mathbf{E}[(h(X_i) - \mu)^2] < \infty$. These two fundamental theorems, and many others, give a fairly rich and detailed picture of the behavior of independent random variables.

Markov Chains

As mentioned, Markov provided a generalization of the independence assumption. In particular, a sequence $X_0, X_1, X_2, \ldots$ forms a time-homogeneous Markov chain if there is a fixed transition kernel $\{P(x, A)\}_{x \in \mathscr{X},\ A \in \mathscr{G}}$ with the property that

$$\mathbf{P}[X_n \in A \,|\, X_0, X_1, \ldots, X_{n-1}] \ = \ P(X_{n-1}, A), \qquad A \in \mathscr{G}. \tag{5}$$

That is, given the history up to time $n-1$, the probabilities for the value at time n depend only on the most recent value X_{n-1}, in an explicit way given by the transition kernel P.

For a concrete example, suppose $X = \{1,2,3,\ldots\}$, with $\mathscr{G}$ the collection of all subsets of $\mathscr{X}$, and with transition kernel P defined by $P(x,\{x+1\}) = 1/4$ for all $x \in \mathscr{X}$, $P(x,\{x-1\}) = 1/2$ for all $x \geq 2$, $P(x,\{x\}) = 1/4$ for all $x \geq 2$, and $P(1,\{1\}) = 3/4$, together with additivity. This Markov chain has, at each time, a probability 1/4 of increasing by 1, a probability 1/4 of staying the same, and a 1/2 probability of decreasing by 1 (unless it is already at the state 1, in which case the probability 1/2 is added to staying at 1); see Figure 1.

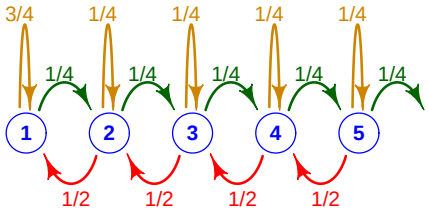

Fig. 1 The transition probabilities of the Running Example Markov chain.

Over time, a Markov chain might settle into a "stable pattern", or more precisely a stationary distribution. A probability distribution π on $(\mathscr{X},\mathscr{G})$ is called *stationary* for the Markov chain $\{X_n\}$ if it has the property that whenever X_n has the distribution π, then X_{n+1} will also have the distribution π. That is,

$$\int_{x\in\mathscr{X}} \pi(dx)\,P(x,A) \;=\; \pi(A). \tag{6}$$

On a discrete space, this can be written as

$$\sum_{x\in\mathscr{X}} \pi\{x\}\,P(x,A) \;=\; \pi(A). \tag{7}$$

Intuitive, the probability distribution π is invariant for the chain, i.e. its probabilities do not change while the chain runs.

The importance of stationary distributions is the Markov Chain Convergence Theorem which says that, under the mild assumptions of irreducibility and aperiodicity (nearly always satisfied, and not considered further here), a chain will eventually *converge* to its stationary distribution. That is, if a Markov chain has a stationary distribution π, and is irreducible and aperiodic, then for all $A \in \mathscr{G}$,

$$\lim_{n\to\infty} \mathbf{P}[X_n \in A] \;=\; \pi(A). \tag{8}$$

Informally, the chain will settle down into its stationary distribution in the long run.

To see this more concretely, consider the above Running Example. Figure 2 shows a typical realization of that Markov chain. Although it starts with larger values, from approximately time $n = 40$ onwards it appears to be in a stable pattern, visiting approximately the same states with approximately the same probabilities.

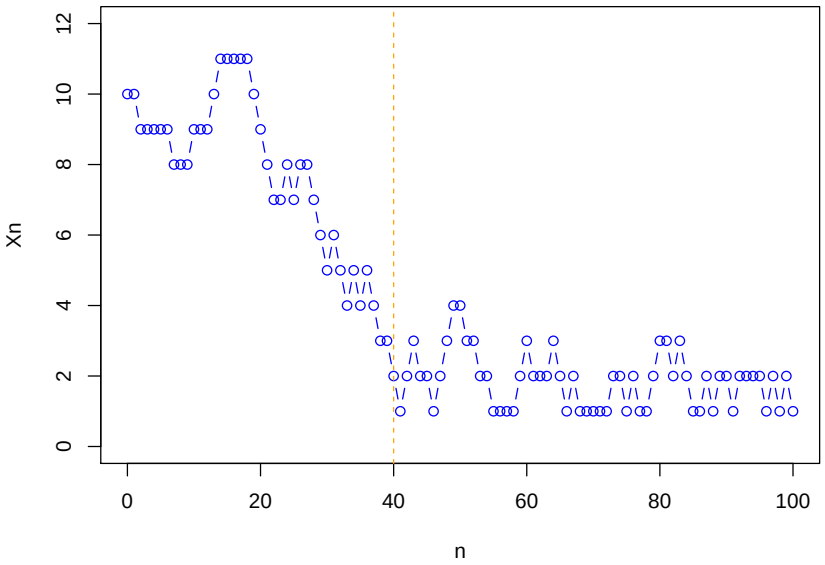

Fig. 2 A typical realization of the Running Example Markov chain.

This "settling down" Markov chain convergence property is enough to prove a law of large numbers and central limit theorem for Markov chains, similar to the i.i.d. case but under stronger assumptions; for a modern treatment see e.g. the monograph [8]. Thus, the time-homogeneous Markov assumption helps to re-capture some of the power and utility of i.i.d. sequences, but in a more general context.

Markov Chain Monte Carlo (MCMC)

When they were first introduced, Markov chains served as models of certain natural scientific (and even social) phenomenon. But in the latter half of the 20th century, it was slowly realized that Markov chains could be used in an entirely different way, as computational algorithms.

The basic idea of a Monte Carlo algorithm is simple. Suppose π is a "target" probability distribution on $(\mathscr{X},\mathscr{G})$, and $h : \mathscr{X} \to \mathbf{R}$ is some measurable function, and you want to estimate the expected value

$$\mu := \mathbf{E}_\pi(h) := \int_{x\in\mathscr{X}} h(x)\,\pi(dx) \tag{9}$$

(so, on a discrete space, $\mu = \sum_{x\in\mathscr{X}} h(x)\,\pi(x)$). In principle, μ could be calculated directly by an integral or sum, but in a complicated application on a high-dimensional state space that might be quite infeasible. However, if you could simulate random variables $X_1, X_2, \ldots, X_M$ which are i.i.d. with distribution π, then you could estimate μ by

$$E := \frac{1}{M}\sum_{i=1}^{M} h(X_i). \tag{10}$$

Indeed, the law of large numbers says that as $M \to \infty$, the estimator E will converge to μ.

The problem is that, if $\mathscr{X}$ is high-dimensional and π is complicated, there is no feasible computer program to simulate independent random variables $X_1, X_2, \ldots$ which each have the distribution π. However, it turns out to be surprisingly easy to simulate a *Markov chain* $X_0, X_1, X_2, \ldots$ which has π as a stationary distribution, and will thus converge to π in distribution. This gives rise to Markov chain Monte Carlo (MCMC) algorithms which simulate a Markov chain $X_0, X_1, X_2, \ldots$ which converges to π, and then estimate μ by

$$E := \frac{1}{M-B}\sum_{i=B+1}^{M} h(X_i). \tag{11}$$

If the burn-in period B is large enough that the distribution of X_B is approximately π, and $M - B$ is large enough that the Markov chain law of large numbers approximately holds, then this estimator E will again provide a good approximation to μ.

The Metropolis Algorithm

While there are many ways to construct Markov chains which leave π stationary, the oldest and simplest one is the Metropolis algorithm [7]. Assume the state space $(\mathscr{X}, \mathscr{G})$ has a reference measure ρ (e.g. counting measure on a discrete state space, or Lebesgue measure on $\mathbf{R}$), with respect to which π has a density function f so that $\pi(dx) = f(x)\,\rho(dx)$. The algorithm uses proposal distributions $Q(x, dy) = q(x, y)\,\rho(dy)$ for $x, y \in \mathscr{X}$, which are symmetric (i.e., $q(x, y) = q(y, x)$). (For example, on a discrete state space, $Q(x, \cdot)$ might give probability 1/2 to each of $x+1$ and $x-1$. Or, on the real line, $Q(x, \cdot)$ might correspond to a Normal(x, v) distribution for some fixed variance parameter $v > 0$. Importantly, $Q(x, \cdot)$ should be easy to sample from on a computer.)

The algorithm then proceeds as follows. First choose some initial state $X_0 \in \mathscr{X}$. Then, iteratively for $n = 1, 2, \ldots$, given $X_{n-1} \in \mathscr{X}$, generate X_n as follows. First, generate a *proposal* state Y_n sampled from the probability distribution $Q(X_{n-1}, \cdot)$, and

independently generate a random variable U_n sampled from the uniform (Lebesgue) measure on the interval [0,1]. Then define X_n by:

$$X_n \;=\; \begin{cases} Y_n & U_n < f(Y_n)/f(X_{n-1}) \\ X_{n-1} & \text{otherwise} \end{cases} \tag{12}$$

That is, if $U_n < f(Y_n)/f(X_n)$ then the proposed new state Y_n is "accepted" so we set $X_n = Y_n$, otherwise Y_n is "rejected" so we keep $X_n = X_{n-1}$. Incredibly, this simple algorithm, which only barely mentions π via its density f in the accept/reject rule, actually makes π a stationary distribution to which it will converge:

Theorem 1. *The above Metropolis algorithm defines a Markov chain $\{X_n\}$ which has π as a stationary distribution. Furthermore, if the chain is irreducible and aperiodic, then it will converge in distribution to π.*

Proof. For $y \neq x$, the only way for the chain to move from x to y (or to a neighborhood dy of y) is to first propose such a move, and then accept it. Thus, for $y \neq x$,

$$\begin{aligned} P(x,dy) \;&:=\; \mathbf{P}[X_n \in dy \,|\, X_{n-1} = x] \\ &=\; \mathbf{P}[Y_n \in dy \,|\, X_{n-1} = x]\; \mathbf{P}[U_n < f(Y_n)/f(x)] \\ &=\; Q(x,dy)\;\min[1, f(Y_n)/f(x)] \\ &=\; q(x,y)\;\rho(dy)\;\min[1, f(y)/f(x)]. \end{aligned} \tag{13}$$

It follows that the bivariate measure

$$\begin{aligned} \pi(dx)\;P(x,dy) \;&=\; \pi(dx)\;\mathbf{P}[X_n \in dy \,|\, X_{n-1} = x] \\ &=\; f(x)\,\rho(dx)\;q(x,y)\;\rho(dy)\;\min[1, f(y)/f(x)] \\ &=\; q(x,y)\;\min[f(x), f(y)]\;\rho(dx)\;\rho(dy). \end{aligned} \tag{14}$$

Since $q(x,y) = q(y,x)$, this measure is symmetric upon swapping x and y, i.e. $\pi(dx)\,P(x,dy) = \pi(dy)\,P(y,dx)$, which says that $\{X_n\}$ is a reversible Markov chain. (The above argument only shows this for $y \neq x$, but it is immediately true for $y = x$ too.) It follows that, if X_{n-1} has the distribution π, then

$$\begin{aligned} \mathbf{P}[X_n \in A] \;&=\; \int_{x \in \mathscr{X}} \pi(dx)\;\mathbf{P}[X_n \in A \,|\, X_{n-1} = x] \\ &=\; \int_{x \in \mathscr{X}} \int_{y \in A} \pi(dx)\;P(x,dy) \\ &=\; \int_{x \in \mathscr{X}} \int_{y \in A} \pi(dy)\;P(y,dx) \\ &=\; \int_{y \in A} \pi(dy)\;P(y,\mathscr{X}) \\ &=\; \int_{y \in A} \pi(dy)\;(1) \;=\; \pi(A), \end{aligned} \tag{15}$$

so X_n also has the distribution π. Therefore, π is a stationary distribution.

The second statement follows immediately from the first one and the Markov Chain Convergence Theorem. □

In summary, the Metropolis algorithm provides a simple method – propose a new state symmetrically, then accept or reject it according to a simple rule – which guarantees that π is a stationary distribution, to which (under very mild conditions) the chain will converge. This is what makes MCMC possible. For a graphical simulation of the Metropolis algorithm in action, see the web page at [10].

MCMC in Practice

To see how the Metropolis algorithm actually works, suppose the state space is $\mathscr{X} = \{1,2,3,\ldots\}$, so ρ is counting measure, and suppose the target probability distribution given by $\pi(\{x\}) = 2^{-x}$ for $x \in \mathscr{X}$. Suppose we choose the simple proposal distributions $Q(x,\{x-1\}) = Q(x,\{x+1\}) = 1/2$. Then, at each iteration, our algorithm proposes a new state $Y = X_{n-1} \pm 1$ which increases or decreases the currently state by one (with probability 1/2 each), and then with probability $\min[1, \pi(Y)/\pi(X_{n-1})] = \min[1, 2^{-Y}/2^{-X_{n-1}}]$ it accepts the proposal and sets $X_n = Y$, otherwise it rejects the proposal and sets $X_n = X_{n-1}$.

So what are the resulting transition probabilities? Well, since $\pi(\{0\}) = 0$, proposed moves from state 1 to state 0 are always rejected. But for $x \geq 2$, since $\pi(\{x-1\}) > \pi(\{x\})$, proposed moves from $x-1$ to x are always accepted. Meanwhile, for any $x \in \mathscr{X}$, proposed moves from x to $x+1$ are accepted with probability $\min[1, 2^{-(x+1)}/2^{-x}] = 1/2$. And whenever a move is rejected, then the chain stays at x. A careful inspection shows that these transition rules are exactly the same as those in the diagram of Figure 1. That is, this Metropolis algorithm exactly coincides with our Running Example, which therefore by Theorem 1 has stationary distribution π, given by $\pi(\{x\}) = 2^{-x}$ for $x \in \mathscr{X}$ as above.

Hence, a typical realization of this Metropolis algorithm will be something like that in Figure 2. In particular, it might start at some larger state (e.g. $X_0 = 10$), but after some burn-in period B, it will send to settle into a pattern of sampling the different states with probabilities roughly equal to the target stationary distribution π. So, as above, if the burn-in period B and total run length M are both large enough, then $\frac{1}{M-B}\sum_{i=B+1}^{M} h(X_i)$ will be a good estimate of the expected value $\mu := \mathbf{E}_\pi(h)$.

For MCMC algorithms to be successful, we need to choose the burn-in value B large enough that the distribution of X_B is approximately equal to π. While there have been some attempts to provide precise mathematical bounds on this burn-in time (see e.g. [9] and the references therein), such bounds cannot usually be obtained. Instead, convergence is usually assessed empirically, by monitoring the output from the Markov chain and using statistical or time-series approaches to decide if the chain seems to have reached a steady state. For example, for chain output as in Figure 2, it might be determined empirically that roughly $B = 40$ iterations (the orange dotted line) are sufficient to approximately reach stationarity.

Of course, the distribution π in our Running Example is so simple that it can easily be sampled without using MCMC. But MCMC is indeed very commonly required in more complicated situations. In particular, it is very often needed for Bayesian inference. The basic Bayesian paradigm is as follows. We wish to estimate some unknown parameters θ, given some observed data $D = d_0$, a prior distribution $\{\mathbf{P}[\theta \in A]\}_A$, and a statistical likelihood model $\{\mathbf{P}[D \in A \,|\, \theta]\}_{\theta,A}$. The all-important posterior distribution, which tells us the final probability distribution of the unknown parameters θ given the observed data and the model, is then given (using Bayes' Rule) by

$$\pi(d\theta) \;:=\; \mathbf{P}[\theta \in d\theta \,|\, D = d_0] \;=\; \frac{\mathbf{P}[\theta \in d\theta]\;\mathbf{P}[D = d_0 \,|\, \theta]}{\mathbf{P}[D = d_0]}, \tag{16}$$

which is proportional to the prior $\mathbf{P}[\theta \in d\theta]$ times the likelihood $\mathbf{P}[D = d_0 \,|\, \theta]$.

In principle, the formula (16) gives precise information about the nature of the posterior distribution, usually expressible as a posterior density function. That should be sufficient to compute posterior probabilities $\pi(A)$, or more generally posterior expectations $\mathbf{E}_\pi(h)$, by computing an integral (or sum) with respect to π as in (9). However, in practical statistical applications, the posterior distribution π is often very complicated and high-dimensional, so that these integrals cannot be computed directly, nor even well approximated by numerical integration, so their computation was thought to be infeasible. This challenge presented a major obstacle to the use of Bayesian inference in practical statistical problems, and limited it for most of the 20th century primarily to theoretical and philosophical and foundational interest rather than to genuine application. However, it was realized around 1990 [11, 2, 12] that MCMC algorithms such as the Metropolis algorithm can provide good estimates of $\mathbf{E}_\pi(h)$ as in (11), often quite easily with good accuracy and limited computational effort.

These MCMC algorithms are now routinely used in Bayesian statistics, which accounts for their extreme popularity. They have transformed the entire field of Bayesian statistics, from a theoretical curiosity to one of the most influential areas of statistical analysis, with significant applications to physics and finance and machine learning and so much more.

Conclusion

This paper has explained how Markov chains, introduced in 1906 as a tractable generalization of independent random variables, found tremendous application many years later as a tool for statistical computation, which completely transformed statistics by making the application of Bayesian inference computationally feasible, thus allowing its application to so many modern fields of study.

Andrey Andreyevich Markov died one hundred years ago, in 1922. If he were alive today, he would surely be astonished and thrilled by the incredible amount of

progress that has been achieved in statistical computation and so many other areas, as a direct result of that little idea that he proposed so long ago.

References

1. S. Brooks, A. Gelman, G. Jones, and X.-L. Meng, eds. (2011), Handbook of Markov chain Monte Carlo. Chapman & Hall, New York.
2. A.E. Gelfand and A.F.M. Smith (1990), Sampling based approaches to calculating marginal densities. J. Amer. Stat. Assoc. **85**, 398–409.
3. S. Geman and D. Geman (1984), Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. IEEE Trans. on pattern analysis and machine intelligence **6**, 721–741.
4. C. Geyer (1992), Practical Markov chain Monte Carlo. Stat. Sci., Vol. **7**, No. **4**, 473-483.
5. W.K. Hastings (1970), Monte Carlo sampling methods using Markov chains and their applications. Biometrika **57**, 97–109.
6. A.A. Markov (1906), Extension of the law of large numbers to dependent quantities (in Russian). Izv. Fiz.-Matem. Obsch. Kazan Univ. (2nd Ser) **15**, 135–156.
7. N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller (1953), Equations of state calculations by fast computing machines. J. Chem. Phys. **21**, 1087–1091.
8. S.P. Meyn and R.L. Tweedie (1993), Markov chains and stochastic stability. Springer-Verlag, London. Available at: http://probability.ca/MT/
9. J.S. Rosenthal (2002), Quantitative convergence rates of Markov chains: A simple account. Elec. Comm. Prob. **7**, No. 13, 123–128.
10. J.S. Rosenthal (2020), Metropolis Algorithm javascript program. Available at: http://probability.ca/jeff/js/metropolis.html
11. M.A. Tanner and W.H. Wong (1987), The calculation of posterior distributions by data augmentation (with discussion). J. Amer. Stat. Assoc. **82**, 528-550.
12. L. Tierney (1994), Markov chains for exploring posterior distributions (with discussion). Ann. Stat. **22**, 1701–1762.