Biometrika (1970), 57, 1, *p.* 97
Printed in Great Britain

Monte Carlo sampling methods using Markov chains and their applications

BY W. K. HASTINGS
University of Toronto

SUMMARY

A generalization of the sampling method introduced by Metropolis *et al.* (1953) is presented along with an exposition of the relevant theory, techniques of application and methods and difficulties of assessing the error in Monte Carlo estimates. Examples of the methods, including the generation of random orthogonal matrices and potential applications of the methods to numerical problems arising in statistics, are discussed.

1. INTRODUCTION

For numerical problems in a large number of dimensions, Monte Carlo methods are often more efficient than conventional numerical methods. However, implementation of the Monte Carlo methods requires sampling from high dimensional probability distributions and this may be very difficult and expensive in analysis and computer time. General methods for sampling from, or estimating expectations with respect to, such distributions are as follows.

(i) If possible, factorize the distribution into the product of one-dimensional conditional distributions from which samples may be obtained.

(ii) Use importance sampling, which may also be used for variance reduction. That is, in order to evaluate the integral

$$J = \int f(x)p(x)dx = E_p(f),$$

where $p(x)$ is a probability density function, instead of obtaining independent samples $x_1, \ldots, x_N$ from $p(x)$ and using the estimate $\hat{J}_1 = \Sigma f(x_i)/N$, we instead obtain the sample from a distribution with density $q(x)$ and use the estimate $\hat{J}_2 = \Sigma\{f(x_i)p(x_i)\}/\{q(x_i)N\}$. This may be advantageous if it is easier to sample from $q(x)$ than $p(x)$, but it is a difficult method to use in a large number of dimensions, since the values of the weights $w(x_i) = p(x_i)/q(x_i)$ for reasonable values of N may all be extremely small, or a few may be extremely large. In estimating the probability of an event A, however, these difficulties may not be as serious since the only values of $w(x)$ which are important are those for which $x \in A$. Since the methods proposed by Trotter & Tukey (1956) for the estimation of conditional expectations require the use of importance sampling, the same difficulties may be encountered in their use.

(iii) Use a simulation technique; that is, if it is difficult to sample directly from $p(x)$ or if $p(x)$ is unknown, sample from some distribution $q(y)$ and obtain the sample x values as some function of the corresponding y values. If we want samples from the conditional distribution of $x = g(y)$, given $h(y) = h_0$, then the simulation technique will not be satisfactory if $\text{pr}\{h(y) = h_0\}$ is small, since the condition $h(y) = h_0$ will be rarely if ever satisfied even when the sample size from $q(y)$ is large.

In this paper, we shall consider Markov chain methods of sampling that are generalizations of a method proposed by Metropolis *et al.* (1953), which has been used extensively for numerical problems in statistical mechanics. An introduction to Metropolis's method and its applications in statistical mechanics is given by Hammersley & Handscomb (1964, p. 117). The main features of these methods for sampling from a distribution with density $p(x)$ are:

(*a*) The computations depend on $p(x)$ only through ratios of the form $p(x')/p(x)$, where x' and x are sample points. Thus, the normalizing constant need not be known, no factorization of $p(x)$ is necessary, and the methods are very easily implemented on a computer. Also, conditional distributions do not require special treatment and therefore the methods provide a convenient means for obtaining correlated samples from, for example, the conditional distributions given by Fraser (1968) or the distributions of the elements in a multiway table given the marginal totals.

(*b*) A sequence of samples is obtained by simulating a Markov chain. The resulting samples are therefore correlated and estimation of the standard deviation of an estimate and assessment of the error of an estimate may require more care than with independent samples.

2. Dependent samples using Markov chains

2·1. *Basic formulation of the method*

Let $\mathbf{P} = \{p_{ij}\}$ be the transition matrix of an irreducible Markov chain with states $0, 1, \ldots, S$. Then, if $X(t)$ denotes the state occupied by the process at time t, we have

$$\text{pr}\{X(t+1) = j | X(t) = i\} = p_{ij}.$$

If $\boldsymbol{\pi} = (\pi_0, \pi_1, \ldots, \pi_S)$ is a probability distribution with $\pi_i > 0$ for all i, and if $f(\cdot)$ is a function defined on the states, and we wish to estimate

$$I = E_\pi(f) = \sum_{i=0}^{S} f(i)\pi_i,$$

we may do this in the following way. Choose $\mathbf{P}$ so that $\boldsymbol{\pi}$ is its unique stationary distribution, i.e. $\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P}$. Simulate this Markov chain for times $t = 1, \ldots, N$ and use the estimate

$$\hat{I} = \sum_{t=1}^{N} f\{X(t)\}/N.$$

For finite irreducible Markov chains we know that $\hat{I}$ is asymptotically normally distributed and that $\hat{I} \to I$ in mean square as $N \to \infty$ (Chung, 1960, p. 99).

In order to estimate the variance of $\hat{I}$, we observe that the process $X(t)$ is asymptotically stationary and hence so is the process $Y(t) = f\{X(t)\}$. The asymptotic variance of the mean of such a process is independent of the initial distribution of $X(0)$, which may, for example, attach probability 1 to a single state, or may be $\boldsymbol{\pi}$ itself, in which case the process is stationary. Thus, if N is large enough, we may estimate $\text{var}(\hat{I})$, using results appropriate for estimating the variance of the mean of a stationary process.

Let ρ_j be the correlation of $Y(t)$ and $Y(t+j)$ and let $\sigma^2 = \text{var}\{Y(t)\}$. It is well known (Bartlett, 1966, p. 284) that for a stationary process

$$\text{var}(\bar{Y}) = \frac{\sigma^2}{N} \sum_{j=-N+1}^{N-1} \left(1 - \frac{|j|}{N}\right) \rho_j$$

and that as $N \to \infty$ $\qquad \mathrm{var}(\overline{Y}) \simeq 2\pi g(0)/N,$

where $g(\omega)$ is the spectral density function at frequency ω. If the ρ_j are negligible for $j \geqslant j_0$, then we may use Hannan's (1957) modification of an estimate of $\mathrm{var}(\overline{Y})$ proposed by Jowett (1955), namely

$$v^2_{\overline{Y}} = \frac{N}{(N-j_0)(N-j_0+1)} \left\{ \sum_{j=-j_0+1}^{j_0-1} \left(1 - \frac{|j|}{N}\right)(c_j - \overline{Y}^2) \right\}, \tag{1}$$

where

$$c_j = \sum_{t=1}^{N-j} Y(t)\, Y(t+j)/(N-j) \quad \text{for} \quad j \geqslant 0 \quad \text{and} \quad c_{-j} = c_j.$$

A satisfactory alternative which is less expensive to compute is obtained by making use of the pilot estimate, corrected for the mean, for the spectral density function at zero frequency suggested by Blackman & Tukey (1958, p. 136) and Blackman (1965). We divide our observations into L groups of K consecutive observations each. Denoting the mean of the ith block by

$$\overline{Y}_i = \sum_{t=1}^{K} Y\{(i-1)K + t\}/K,$$

we use the estimate

$$s^2_{\overline{Y}} = \sum_{i=1}^{L} (\overline{Y}_i - \overline{Y})^2/\{L(L-1)\}. \tag{2}$$

This estimate has approximately the stability of a chi-squared distribution on $(L-1)$ degrees of freedom. Similarly, the covariance of the means of two jointly stationary processes $Y(t)$ and $Z(t)$ may be estimated by

$$s_{\overline{Y}\overline{Z}} = \sum_{i=1}^{L} (\overline{Y}_i - \overline{Y})(\overline{Z}_i - \overline{Z})/\{L(L-1)\}. \tag{3}$$

2·2. *Construction of the transition matrix*

In order to use this method for a given distribution $\boldsymbol{\pi}$, we must construct a Markov chain $\mathbf{P}$ with $\boldsymbol{\pi}$ as its stationary distribution. We now describe a general procedure for doing this which contains as special cases the methods which have been used for problems in statistical mechanics, in those cases where the matrix $\mathbf{P}$ was made to satisfy the reversibility condition that for all i and j

$$\pi_i p_{ij} = \pi_j p_{ji}. \tag{4}$$

The property ensures that $\Sigma \pi_i p_{ij} = \pi_j$, for all j, and hence that $\boldsymbol{\pi}$ is a stationary distribution of $\mathbf{P}$. The irreducibility of $\mathbf{P}$ must be checked in each specific application. It is only necessary to check that there is a positive probability of going from state i to state j in some finite number of transitions, for all pairs of states i and j.

We assume that p_{ij} has the form

$$p_{ij} = q_{ij}\alpha_{ij} \quad (i \neq j), \tag{5}$$

with

$$p_{ii} = 1 - \sum_{j \neq i} p_{ij},$$

where $\mathbf{Q} = \{q_{ij}\}$ is the transition matrix of an arbitrary Markov chain on the states $0, 1, \ldots, S$ and α_{ij} is given by

$$\alpha_{ij} = \frac{s_{ij}}{1 + \dfrac{\pi_i q_{ij}}{\pi_j q_{ji}}}, \tag{6}$$

where s_{ij} is a symmetric function of i and j chosen so that $0 \leqslant \alpha_{ij} \leqslant 1$ for all i and j. With this form for p_{ij} it is readily verified that $\pi_i p_{ij} = \pi_j p_{ji}$, as required. In order to simulate this process we carry out the following steps for each time t:

(i) assume that $X(t) = i$ and select a state j using the distribution given by the ith row of $\mathbf{Q}$;

(ii) take $X(t+1) = j$ with probability α_{ij} and $X(t+1) = i$ with probability $1 - \alpha_{ij}$.

For the choices of s_{ij} we will consider, only the quantity $(\pi_j q_{ji})/(\pi_i q_{ij})$ enters into the simulation and we will henceforth refer to it as the test ratio.

Two simple choices for s_{ij} are given for all i and j by

$$s_{ij}^{(M)} = \begin{cases} 1 + \dfrac{\pi_i q_{ij}}{\pi_j q_{ji}} & \left(\dfrac{\pi_j q_{ji}}{\pi_i q_{ij}} \geqslant 1\right), \\ 1 + \dfrac{\pi_j q_{ji}}{\pi_i q_{ij}} & \left(\dfrac{\pi_j q_{ji}}{\pi_i q_{ij}} \leqslant 1\right), \end{cases}$$

$$s_{ij}^{(B)} = 1.$$

With $q_{ij} = q_{ji}$ and $s_{ij} = s_{ij}^{(M)}$ we have the method devised by Metropolis *et al.* (1953) and with $q_{ij} = q_{ji}$ and $s_{ij} = s_{ij}^{(B)}$ we have Barker's (1965) method.

Little is known about the relative merits of these two choices for s_{ij}, but when $q_{ij} = q_{ji}$, we have

$$\alpha_{ij}^{(M)} = \begin{cases} 1 & (\pi_j/\pi_i \geqslant 1), \\ \pi_j/\pi_i & (\pi_j/\pi_i < 1), \end{cases}$$

$$\alpha_{ij}^{(B)} = \pi_j/(\pi_i + \pi_j).$$

Thus we see that if $\pi_j = \pi_i$, we will take $X(t+1) = j$ with probability 1 with Metropolis's method and with probability $\frac{1}{2}$ with Barker's method. This suggests that Metropolis's method may be preferable since it seems to encourage a better sampling of the states.

More generally, we may choose

$$s_{ij} = g[\min\{(\pi_i q_{ij})/(\pi_j q_{ji}), (\pi_j q_{ji})/(\pi_i q_{ij})\}],$$

where the function $g(x)$ is chosen so that $0 \leqslant g(x) \leqslant 1 + x$ for $0 \leqslant x \leqslant 1$, and $g(x)$ may itself be symmetric in i and j. For example, we may choose $g(x) = 1 + 2(\frac{1}{2}x)^\gamma$ with the constant $\gamma \geqslant 1$, obtaining $s_{ij}^{(M)}$ with $\gamma = 1$ and $s_{ij}^{(B)}$ with $\gamma = \infty$.

We may define a *rejection rate* as the proportion of times t for which $X(t+1) = X(t)$. Clearly, in choosing $\mathbf{Q}$, high rejection rates are to be avoided. For example, if $X(t) = i$ and i is near the mode of an unimodal distribution, then $\mathbf{Q}$ should be chosen so that j is not too far from i, otherwise, π_j/π_i will be small and it is likely that $X(t+1) = i$. For each simulation it is useful to record the rejection rate since a high rejection rate may be indicative of a poor choice of initial state or transition matrix, or of a 'bug' in the computer program.

We shall apply these methods to distributions $\boldsymbol{\pi}$ defined mathematically on an infinite sample space although, when we actually simulate the Markov chains on a digital computer, we will have a large but finite number of states. When $\boldsymbol{\pi}$ is continuous, we will have a discrete approximation to $\boldsymbol{\pi}$. Let $\pi(x)\,d\mu(x)$, $p(x, x')\,d\mu(x')$ and $q(x, x')\,d\mu(x')$ be the probability elements for the distribution $\boldsymbol{\pi}$ and the Markov processes analogous to $\mathbf{P}$ and $\mathbf{Q}$, respectively. Let the possible values for x and x' on the computer be $x_0, \ldots, x_s$. These values depend on the word length and on the representation in the computer, float-point, fixed-

point, etc. The probability elements may be approximated by $\pi(x_i)\,\delta\mu(x_i)$, $p(x_i, x_j)\,\delta\mu(x_j)$ and $q(x_i, x_j)\,\delta\mu(x_j)$, respectively. Substituting $\pi(x_i)\,\delta\mu(x_i)$ for π_i etc., we have

$$p(x_i, x_j) = q(x_i, x_j)\alpha_{ij}$$

and $(\pi_i q_{ij})/(\pi_j q_{ji})$ is replaced by $\{\pi(x_i)q(x_i, x_j)\}/\{\pi(x_j)q(x_j, x_i)\}$. Therefore, so long as α_{ij} is chosen so that it depends on $\boldsymbol{\pi}$ and $\mathbf{Q}$ only through the quantity $(\pi_i q_{ij})/(\pi_j q_{ji})$, we may use the densities in place of the corresponding probability mass functions, and we may think in terms of continuous distributions ignoring the underlying approximation which will be more than adequate in most applications.

For simplicity we have considered reversible Markov chains only. An example of an irreversible $\mathbf{P}$ is given in Handscomb (1962), where the states may be subdivided into finite subsets of states, where the states within a given subset are equally probable. Transitions amongst states within such a subset are made in a cyclic fashion; all other transitions are reversible.

2·3. *Elementary examples*

Example 1. Let $\boldsymbol{\pi}$ be the Poisson distribution with $\pi_i = \lambda^i e^{-\lambda}/i!$ $(i = 0, 1, \ldots)$. For λ small we may use the following choice of $\mathbf{Q}$ to generate samples from $\boldsymbol{\pi}$:

$$q_{ij} = \tfrac{1}{2} \quad (j = i-1, i+1;\ i \neq 0), \quad q_{00} = q_{01} = \tfrac{1}{2}.$$

Note that $\pi_{i+1}/\pi_i = \lambda/(i+1)$ and $\pi_{i-1}/\pi_i = i/\lambda$ so that the computations may be performed rapidly. For λ large $\mathbf{Q}$ must be chosen so that step sizes greater than unity are permitted or else very long realizations will be required in order to ensure adequate coverage of the sample space.

Example 2. To sample from the standard normal distribution we define $\mathbf{Q}_k$ $(k = 1, 2)$ in the following ways. Let $X(t)$ be the state at time t and choose $X'(t)$ to be uniformly distributed on the interval $[\epsilon_k X(t)-\Delta, \epsilon_k X(t)+\Delta]$, where $\Delta > 0$ is a constant, $\epsilon_1 = +1$ and $\epsilon_2 = -1$. To use Metropolis's method with either of these choices of $\mathbf{Q}$, both of which are symmetric, at each time t compute the test ratio $\beta(t) = \exp(\frac{1}{2}[X^2(t)-\{X'(t)\}^2])$. If $\beta(t) \geqslant 1$, set $X(t+1) = X'(t)$; if $\beta(t) < 1$, $X(t+1) = X'(t)$ with probability $\beta(t)$ and $X(t+1) = X(t)$ with probability $1-\beta(t)$. Computing time can be saved if we compare $X^2(t)$ with $\{X'(t)\}^2$ instead of $\beta(t)$ with 1 and compute the exponential only when $\{X'(t)\}^2 > X^2(t)$.

Simulations carried out on an IBM 7094 II using the above transition matrices yielded the following estimates $\bar{X}$ of the mean of the distribution where, in all cases, we chose $N = 1000$ and $L = 25$: with $X(0) = 0$ and $\Delta = 1$, we obtained $\bar{X} = -0{\cdot}12$ and $s_{\bar{x}} = 0{\cdot}11$ with $\epsilon_k = +1$, and $\bar{X} = -0{\cdot}013$ and $s_{\bar{x}} = 0{\cdot}02$ with $\epsilon_k = -1$. The estimated standard deviation of the estimate is smaller with $\epsilon_k = -1$, and is comparable to the theoretical standard deviation of the mean with 1000 independent observations, 0·031. Here we have an ideal situation with a symmetric distribution and known mean, but a similar device for variance reduction may sometimes be applicable in other problems of more practical interest. Other results indicated that there is little to choose between moderate values of Δ in the range 0·2–1·8, but extreme values of Δ led to poor estimates. For example, with $X(0) = 1{\cdot}0$ and $\Delta = 0{\cdot}001$ we obtained $\bar{X} = 1{\cdot}001$ and $s_{\bar{x}} = 0{\cdot}002$.

2·4. *Multidimensional distributions*

If the distribution $\boldsymbol{\pi}$ is d-dimensional and the simulated process is $\mathbf{X}(t) = \{X_1(t), \ldots, X_d(t)\}$, there are many additional techniques which may be used to construct $\mathbf{P}$:

(1) In the transition from time t to $(t+1)$ all co-ordinates of $\mathbf{X}(t)$ may be changed.

(2) In the transition from time t to $t+1$ only one co-ordinate of $\mathbf{X}(t)$ may be changed, that selection being made at random from amongst the d co-ordinates.

(3) Only one co-ordinate may change in each transition, the co-ordinates being selected in a fixed rather than a random sequence.

Method (3) was used by Ehrman, Fosdick & Handscomb (1960), who justified the method for their particular application. A general justification for the method may be obtained in the following way. Let the transition matrix when co-ordinate k is to be moved by $\mathbf{P}_k$ $(k = 1, \ldots, d)$. Assume that the co-ordinates are moved in the order $1, 2, \ldots$ and that the process is observed only at times $0, d, \ldots$. The resulting process is a Markov process with transition matrix $\mathbf{P} = \mathbf{P}_1 \ldots \mathbf{P}_d$. If, for each k, $\mathbf{P}_k$ is constructed so that $\boldsymbol{\pi}\mathbf{P}_k = \boldsymbol{\pi}$, then $\boldsymbol{\pi}$ will be a stationary distribution of $\mathbf{P}$ since $\boldsymbol{\pi}\mathbf{P} = \boldsymbol{\pi}\mathbf{P}_1 \ldots \mathbf{P}_d = \pi\mathbf{P}_2 \ldots \mathbf{P}_d = \ldots = \boldsymbol{\pi}$. In practice, we must check the irreducibility of $\mathbf{P}$ to ensure uniqueness of the stationary distribution. Note that for the validity of this proof it is not necessary that each $\mathbf{P}_k$ satisfy the reversibility conditions (4). Also, in our estimates we may average observed values of the function $f(.)$ at all times t of the original process although it would not be desirable to do so if the function values only change every d steps.

Example 3. To illustrate method (3) we consider sampling from a distribution with probability density function $p(\mathbf{x}) = p(x_1, \ldots, x_d)$ defined over the domain

$$0 \leqslant x_1 \leqslant \ldots \leqslant x_d < \infty.$$

Let the co-ordinates at time t be $\mathbf{x}(t) = \{x_1(t), \ldots, x_d(t)\}$ and assume co-ordinate k $(k \neq d)$ is to be changed. Let $x'_k(t)$ be chosen from the uniform distribution on the interval $\{x_{k-1}(t), x_{k+1}(t)\}$, where we assume that $x_0(t) \equiv 0$, and define $x'_i(t) = x_i(t)$ $(i \neq k)$. Using Metropolis's method, we find that the test ratio is $p\{\mathbf{x}'(t)\}/p\{\mathbf{x}(t)\}$. When $k = d$, we may choose $x'_d(t)$ from the uniform distribution on the interval

$$[\tfrac{1}{2}\{x_{d-1}(t) + x_d(t)\},\ 2x_d(t) - x_{d-1}(t)]$$

and use the test ratio

$$[p\{\mathbf{x}'(t)\}\{x_d(t) - x_{d-1}(t)\}]/[p\{\mathbf{x}(t)\}\{x'_d(t) - x_{d-1}(t)\}].$$

A few computer runs using this form of transition matrix have yielded satisfactory results for (i) sampling from the distribution of eigenvalues of a Wishart matrix, and (ii) estimating the conditional expected values of order statistics needed for the probability plotting method proposed by Wilk & Gnanadesikan (1968). The latter problem requires estimation of $E_p(x_k)$ $(k = 1, \ldots, d)$, where

$$p(\mathbf{x}) = c \prod_{i=1}^{d} (\beta_i)^{\beta_i} x_i^{\beta_i - 1} \exp(-\beta_i x_i)/\Gamma(\beta_i) \quad (0 \leqslant x_i \leqslant \ldots \leqslant x_d),$$

the β_i's are constants and c is a normalizing constant.

Generalizations of these methods are readily obtained by noting that, for example, the co-ordinates need not be moved equally often and that the co-ordinates may be moved in groups instead of one at a time. For example, in statistical mechanics applications, one particle and hence three co-ordinates are moved at a time.

When only a few of the co-ordinates are moved at one time, computing effort can usually be saved by employing a recurrence formula for obtaining $f\{\mathbf{X}(t+1)\}$ from $f\{\mathbf{X}(t)\}$. However it is then necessary to guard against excessive build-up of error. One way to accomplish this is periodically to calculate $f\{\mathbf{X}(t)\}$ without the aid of the recursion, although an error

analysis in some cases may indicate that this is unnecssary. If, however, it is expensive to compute either of π_i or $f(i)$, it may be preferable to attempt to move all of the co-ordinates a small distance each at time t.

2·5. *Importance sampling*

Usually, **P** is constructed so that it depends only on the ratios π_j/π_i as in the methods of Metropolis and Barker. In this case we need only know the π_i's up to a constant of proportionality, and if $\pi_0 + \ldots + \pi_S \neq 1$, we are estimating

$$E(f) = \sum_{i=0}^{S} f(i)\pi_i \Big/ \sum_{i=0}^{S} \pi_i.$$

This expression may be rewritten in the form

$$\frac{\sum_{i=0}^{S} \{f(i)\pi_i/\pi'_i\}\pi'_i \Big/ \sum_{i=0}^{S} \pi'_i}{\sum_{i=0}^{S} (\pi_i/\pi'_i)\pi'_i \Big/ \sum_{i=0}^{S} \pi'_i},$$

which we recognize as $E_{\pi'}\{f(i)\pi_i/\pi'_i\}/E_{\pi'}(\pi_i/\pi'_i)$. Hence, if we set up the Markov chain using the distribution $\boldsymbol{\pi}'$ instead of $\boldsymbol{\pi}$ we can estimate $I = E_\pi(f)$ by the ratio

$$\hat{I} = \frac{\sum_{t=1}^{N} [f\{X(t)\}\pi_{X(t)}/\pi'_{X(t)}]/N}{\sum_{t=1}^{N} \{\pi_{X(t)}/\pi'_{X(t)}\}/N}. \tag{7}$$

This device permits us to use importance sampling with the Markov chain methods as suggested by Fosdick (1963).

If $\pi_0 + \ldots + \pi_S = 1$ and $\pi'_0 + \ldots + \pi'_S = 1$, then we may replace the denominator in $\hat{I}$ by 1, simplifying the estimate. Otherwise we have a ratio estimate and we can estimate its variance using the usual approximation. Thus, if we denote $\hat{I}$ by $\overline{Y}/\overline{Z}$, the variance is given approximately by

$$\operatorname{var}(\overline{Y}/\overline{Z}) = \{\operatorname{var}(\overline{Y}) - 2I \operatorname{cov}(\overline{Y}, \overline{Z}) + I^2 \operatorname{var}(Z)\}/\{E(\overline{Z})\}^2.$$

This may be estimated by

$$(s^2_{\overline{Y}} - 2\hat{I}s_{\overline{Y}\overline{Z}} + \hat{I}^2 s_{\overline{Z}})/\overline{Z}^2,$$

where $s^2_{\overline{Y}}$, $s_{\overline{Y}\overline{Z}}$ and $s^2_{\overline{Z}}$ are obtained as in (2) and (3). For a discussion of the validity of this approximation, see Hansen, Hurwitz & Madow (1953, p. 164).

Importance sampling for variance reduction is more easily implemented with the Markov chain methods than with methods using independent samples, since with the Markov chain methods it is not necessary to construct the distribution $\boldsymbol{\pi}'$ so that independent samples can be obtained from it. If, however, we can obtain independent samples from the distribution $\boldsymbol{\pi}'$, we may obtain estimates of $E_\pi(f)$ using the Markov chain methods, which have the advantage that they do not involve the weights $\pi_{X(t)}/\pi'_{X(t)}$; see the discussion of importance sampling in §1. To accomplish this we set up the Markov chain **P** so that it has $\boldsymbol{\pi}$, not $\boldsymbol{\pi}'$, as its stationary distribution and choose $q_{ij} = \pi'_j$ for all i and j.

2·6. *Assessment of the error*

In using Monte Carlo methods to estimate some quantity we usually use the standard deviation of the estimate to obtain some indication of the magnitude of the error of the estimate. There are, of course, many sources of error common to all Monte Carlo methods whose magnitude cannot be assessed using the standard deviation alone. These include: (i) the source of uniform random numbers used to generate the sample, which should be of as high a quality as possible; (ii) the nonnormality of the distribution of the estimate; (iii) computational errors which arise in computing the estimate; (iv) computation errors which arise in generating the samples (including discretization and truncation of the distribution); and (v) errors induced because the sample size is too small, which are often best overcome by methods other than increasing the sample size; for example, by the use of importance sampling. In what follows we shall concentrate upon categories (iv) and (v) above.

In generating successive samples using the Markov chain methods, errors will arise in computing the new state $X(t+1)$ and in computing the test ratio. An error analysis may sometimes be useful for the computation of $X(t+1)$ (see, for example, § 3), but it is difficult to assess the effects of using inaccurate test ratios. The situation is also difficult to analyze, in general, when successive samples are independent and are generated using a factorization of the probability density function. To see this, let

$$p(\mathbf{x}) = p(x_1, x_2, \ldots, x_d) = \prod_{i=1}^{d} p_i(x_i)$$

be the joint density function, where $p_i(x_i)$ is the conditional density function for x_i given $x_1, \ldots, x_{i-1}$. When we attempt to sample from each of the one-dimensional distributions $p_i(x_i)$ in turn, errors will be introduced and we will instead be sampling from some distribution $\overline{p}_i(x_i) = p_i(x_i)(1+y_i)$, where y_i is a function of $x_1, \ldots, x_i$ and will not, in general, be small for all sample points. Consequently, we will be generating samples from a distribution

$$\overline{p}(\mathbf{x}) = p(\mathbf{x}) \prod_{i=1}^{d} (1+y_i)$$

and, especially when d is large, it will be difficult to assess the error in our estimate induced by the y_i's. A similar, but more involved, analysis might also be applied to the Markov chain methods if we consider a single realization of length N as a single sample from a distribution of dimension Nd; sampling at each step of the Markov chain would correspond to sampling from a single factor above.

If the sample size is not large enough, important regions of the sample space may be inadequately represented. For example, if we are estimating the integral $\int f(x)p(x)dx$ by sampling from $p(x)$ and if the major contribution to the value of the integral comes from a small region of low probability in which $f(x)$ has very large values, then we may obtain a very poor estimate and a deceptively small standard deviation even with seemingly large sample sizes. This difficulty may be encountered with any method of numerical quadrature and there is no substitute for a thorough study of the integrand and a consequent adjustment of the method if gross errors are to be avoided. With the Markov chain methods the influence on the result of the choice of initial state and the correlation of the samples, which may be considerable if the sample size is small, may be minimized by adopting the following procedures:

(i) Choose a transition matrix $\mathbf{Q}$ so that the sample point in one step may move as large a distance as possible in the sample space, consistent with a low rejection rate. This offers some protection against the possibility that the sample points for the whole realization remain near one mode of the distribution which is separated from other modes by a deep trough, or that the initial state is in a region of low probability.

(ii) If possible, choose the initial state so that it is in a region of high probability, by sampling from $\boldsymbol{\pi}$ if this is possible, or set the co-ordinates of the initial state equal to their expected values, or asymptotic approximations of these.

In view of the many sources of error one or more of the following techniques, depending upon the application, should be used in practice to aid in assessing the suitability of the Markov chain methods, the magnitude of the error and the adequacy of the length of the realization and of the standard deviation as a measure of error:

(*a*) Test the method on problems which bear a close resemblance to the problem under study and for which the results are known analytically.

(*b*) If the expected value of some function with respect to $\boldsymbol{\pi}$ is known, this may be estimated for checking purposes, and possibly for variance reduction, while other aspects of $\boldsymbol{\pi}$ are under study.

(*c*) Compare estimates obtained from different segments of the same realization to see if there is evidence of nonconvergence.

(*d*) Compare results obtained with and without the use of importance sampling, or using different choices of $\mathbf{P}$, or using different random numbers, or using the Markov chain method and some other numerical method, for which adequate assessment of the error may also be difficult as is often the case with asymptotic results, for example.

The illustrations given by Fox & Mayers (1968), for example, show how even the simplest of numerical methods may yield spurious results if insufficient care is taken in their use, and how difficult it often is to assess the magnitude of the errors. The discussion above indicates that the situation is certainly no better for the Markov chain methods and that they should be used with appropriate caution.

3. Random orthogonal matrices

We now consider methods for generating random orthogonal and unitary matrices and their application to the evaluation of averages over the orthogonal group with respect to invariant measure, a problem considered analytically by James (1955), and to other related problems.

Let $\mathbf{H}$ be an orthogonal $m \times m$ matrix with $|\mathbf{H}| = 1$ and let $\mathbf{E}_{ij}(\theta)$ be an elementary orthogonal matrix with elements given by

$$e_{ii} = \cos\theta, \quad e_{ij} = \sin\theta, \quad e_{ji} = -\sin\theta, \quad e_{jj} = \cos\theta,$$
$$e_{\alpha\alpha} = 1 \quad (\alpha \neq i, j), \qquad e_{\alpha\beta} = 0 \quad \text{otherwise},$$

for some angle θ. We may generate a sequence of orthogonal matrices $\mathbf{H}(t)$ $(t = 1, 2, \ldots)$ using Metropolis's method, which is suitable for estimating averages over the orthogonal group with respect to invariant measure $\{d\mathbf{H}\}$, in the following way:

(i) Let $\mathbf{H}(0) = \mathbf{H}_0$, where $\mathbf{H}_0$ is an orthogonal matrix.

(ii) For each t, select i and j at random from the set $\{1, \ldots, m\}$ with $i \neq j$. Select θ from the uniform distribution on $[0, 2\pi]$.

(iii) Let $\mathbf{H}'(t) = \mathbf{E}_{ij}(\theta)\,\mathbf{H}(t)$.

Since the measure $\{d\mathbf{H}\}$ is left invariant, $\{d\mathbf{H}'(t)\}$ and $\{d\mathbf{H}(t)\}$ are equal and hence, in Metropolis's method, the new state is always the newly computed value, i.e. $\mathbf{H}(t+1) = \mathbf{H}'(t)$. To estimate the integral

$$J = \int_{O(m)} f(\mathbf{H})\{d\mathbf{H}\},$$

where $O(m)$ denotes the group of orthogonal matrices, we use the estimate

$$J = \sum_{t=1}^{N} f\{\mathbf{H}(t)\}/N.$$

To remcve the restriction that $|\mathbf{H}| = 1$ we may use the following procedure. After i and j are selected as above, select one of i and j at random, each with probability $\frac{1}{2}$, and call the result i'. Multiply the i'th row of $\mathbf{E}_{ij}(\theta)$ by ± 1, the sign being selected at random, to form a new matrix which we denote by $\mathbf{E}'_{ij}(\theta)$; $\mathbf{E}'_{ij}(\theta)$ is then used in step (iii) above, in place of $\mathbf{E}_{ij}(\theta)$.

To show that the Markov chain is irreducible we need only show that every orthogonal matrix $\mathbf{H}$ may be represented as the product of matrices of the form $\mathbf{E}'_{ij}(\theta)$. This, in turn, may be done by a simple annihilation argument. Choose θ so that $\mathbf{E}_{12}(\theta)\mathbf{H} = \mathbf{H}^{(1)}$ has the element $h_{21}^{(1)} = 0$. Choose θ so that $\mathbf{E}_{13}(\theta)\mathbf{H}^{(1)} = \mathbf{H}^{(2)}$ has elements $h_{31}^{(2)} = h_{21}^{(2)} = 0$. Continue in this way until we have the matrix $\mathbf{H}^{(m-1)}$ with the first column annihilated. Since $\mathbf{H}^{(m-1)}$ is orthogonal, we must have $h_{12}^{(m-1)} = \ldots = h_{1m}^{(m-1)} = 0$ and $h_{11}^{(m-1)} = \pm 1$. Continuing this procedure we may annihilate the remaining off diagonal elements and obtain a diagonal matrix with elements ± 1 on the diagonal. The desired factorization is easily deduced from this.

We now show that the process, by which the sequence of orthogonal matrices is generated, is numerically stable. Denote the computed values of $\mathbf{H}(t)$ and $\mathbf{E}_{ij}(\theta)$ by $\mathbf{H}_c(t)$ and $\mathbf{E}_c(t)$. Let $\mathbf{H}_c(t) = \mathbf{H}(t) + \mathbf{H}_e(t)$, $\mathbf{E}_c(t) = \mathbf{E}_{ij}(\theta) + \mathbf{E}_e(t)$ and $\mathbf{H}_c(t+1) = \mathbf{E}_c(t)\mathbf{H}_c(t) + \mathbf{F}(t)$. We are interested in how far $\mathbf{H}_e(t)$ is from zero, which we measure by $\|\mathbf{H}_e(t)\|$, where the Euclidean norm $\|\mathbf{A}\|$ of a matrix $\mathbf{A}$ is defined by

$$\|\mathbf{A}\| = \left(\sum_{i=1}^{m} \sum_{j=1}^{m} a_{ij}^2 \right)^{\frac{1}{2}}.$$

Using the fact that $\|\mathbf{A}\|$ is preserved under orthogonal transformations and the inequalities $\|\mathbf{AB}\| \leqslant \|\mathbf{A}\|\,\|\mathbf{B}\|$ and $\|\mathbf{A}+\mathbf{B}\| \leqslant \|\mathbf{A}\| + \|\mathbf{B}\|$, we can easily show that

$$\|\mathbf{H}_e(t+1)\| \leqslant (1+k)\,\|\mathbf{H}_e(t)\| + k,$$

where k is chosen such that $\|\mathbf{E}_c(t)\| + \|\mathbf{F}(t)\| \leqslant k$ for all t, and by induction that $\|\mathbf{H}_c(t)\| \leqslant U_t$, where U_t is the solution of the difference equation $U_{t+1} = (1+k)U_t + k$ with $U_0 = \|\mathbf{H}_e(0)\|$. Solving the difference equation, we arrive at the bound

$$\|\mathbf{H}_e(t)\| \leqslant (1+k)^t - 1 + \|\mathbf{H}_e(0)\|.$$

Therefore, unless the word length of the computer is short or t is very large, the accumulation of error may be neglected since, for all t, $\|\mathbf{F}(t)\|$ will be small in view of the few arithmetic operations involved in forming the elements of the product $\mathbf{E}_c(t)\mathbf{H}_c(t)$, and $\|\mathbf{E}_e(t)\|$ will also be small if the method below is used for the generation of the four nontrivial elements of $\mathbf{E}_c(t)$. For short computer word length or very large t, orthonormalization of the columns of $\mathbf{H}_c(t)$ at regular intervals will prevent excessive accumulation of error.

One method for computing $\cos\theta$ and $\sin\theta$ required in $\mathbf{E}_{ij}(\theta)$ is that of von Neumann. Let U_1 and U_2 be independent and uniformly distributed on $[0, 1]$. If $U_1^2+U_2^2 \leqslant 1$, compute

$$\cos\theta = (U_1^2-U_2^2)/(U_1^2+U_2^2), \quad \sin\theta = \pm 2U_1U_2/(U_1^2+U_2^2),$$

where the sign of $\sin\theta$ is chosen at random. If $U_1^2+U_2^2 > 1$, obtain new values of U_1 and U_2 until $U_1^2+U_2^2 \leqslant 1$.

Example 4. When $f(\mathbf{H}) = h_{11}^2+\ldots+h_{mm}^2$, then $J = 1$. Using $N = 1000$, $L = 25$, $m = 50$ and $\mathbf{H}(0) = \mathbf{I}$, we obtained the estimate $\hat{J} = 3{\cdot}5$ with standard deviation $1{\cdot}5$. This estimate is poor because of the very poor choice for $\mathbf{H}(0)$ but the estimated standard deviation gives us adequate warning. It would be natural to increase the sample size N in order to obtain greater precision but this may also be achieved by a better choice of $\mathbf{H}(0)$ with a saving in computer time. For problems of the kind being considered here the following choice of $\mathbf{H}(0)$ is generally useful and better than that considered above. For $j = 1, \ldots, m$ and for m even, set

$$h_{1j} = 1/\sqrt{m}, \quad h_{2j} = (1/\sqrt{m})\cos\{(j-1)\pi\},$$

$$h_{rj} = \sqrt{(2/m)}\cos\{(j-1)(r-2)\,2\pi/m\} \quad (r = 3, 4, \ldots, \tfrac{1}{2}m+1),$$

$$h_{sj} = \sqrt{(2/\pi)}\sin\{(j-1)(s-\tfrac{1}{2}m-1)\,2\pi/m\} \quad (s = \tfrac{1}{2}m+2, \ldots, m).$$

For m odd a similar choice for $\mathbf{H}(0)$ may be made. Using this matrix for $\mathbf{H}(0)$, we obtained the estimate $\hat{J} = 0{\cdot}96$ with standard deviation $0{\cdot}03$; the computer time required was $0{\cdot}08$ minutes.

The techniques discussed above may be applied to the following problems:

(i) To generate random permutation matrices we set $\theta = \frac{1}{2}\pi$. To generate random permutations, of m elements, at each time t, we interchange a randomly selected pair of elements in the latest permutation in order to generate a new permutation.

(ii) To generate k orthogonal vectors in m dimensions we replace $\mathbf{H}(t)$ by an $m \times k$ matrix where columns will be the desired vectors.

(iii) To generate random unitary matrices the above procedure for orthogonal matrices is modified as follows:

In place of $\mathbf{E}_{ij}(\theta)$ we use an elementary unitary matrix $\mathbf{U}_{ij}(\theta, \phi)$ with elements given by

$$U_{ii} = \cos\theta, \quad U_{ij} = e^{-i\phi}\sin\theta, \quad U_{ji} = -e^{i\phi}\sin\theta, \quad U_{jj} = \cos\theta,$$

$$U_{\alpha\alpha} = 1 \quad (\alpha \neq i, j), \qquad U_{\alpha\beta} = 0 \quad \text{otherwise}.$$

The matrix $\mathbf{U}'_{ij}(\theta, \phi)$ is obtained by multiplying the ith row of $\mathbf{U}'_{ij}(\theta, \phi)$ by $e^{i\gamma}$. Here, θ, ϕ and γ are chosen to be independent and uniformly distributed on $[0, 2\pi]$. Irreducibility may again be established by an annihilation argument and the proof of numerical stability given above requires only minor modifications.

(iv) To sample from a distribution defined on a group $G = \{g\}$, the transition matrix $\mathbf{Q}$ must generate $g'(t)$ from $g(t)$, and this may be done by arranging that $g'(t) = hg(t)$, where $h \in G$ and h is chosen from some distribution on G. If the probability element for g is $p(g)\{d\mu(g)\}$, where μ is the left invariant measure on G, then the ratio corresponding to π_j/π_i is $p\{g'(t)\}/p\{g(t)\}$. As above we must ensure that $g(t)$ is close to being a group element.

(v) To generate samples from a distribution with probability element

$$p(\mathbf{x})\,d\mathbf{x} = \int_{O(m)} q(\mathbf{x}, \mathbf{H})\{d\mathbf{H}\}\,d\mathbf{x}$$

we may remove the integral and sample from $q(\mathbf{x}, \mathbf{H})\{d\mathbf{H}\}d\mathbf{x}$ with, unfortunately, a consequent increase in the dimension of the space being sampled. The sampling may be carried out by combining the methods of § 2·4, alternating the changes of $\mathbf{x}$ and $\mathbf{H}$, with the techniques given above for orthogonal matrices. Note that the value of the test ratio for changing $\mathbf{H}$ will no longer be unity. More generally, sampling from any distribution whose density may be expressed as an average may be accomplished in a similar way. Many of the normal theory distributions of multivariate analysis given by James (1964) have this form.

In the method described above for orthogonal matrices the successive matrices are statistically dependent and we now consider a method in which successive matrices are independent. Let x_{ij} $(i = 1, \ldots, k; j = 1, \ldots, m)$ be independent standard normal variates and form the k vectors

$$x_i = (x_{i1}, \ldots, x_{im}) \quad (i = 1, 2, \ldots, k).$$

If we now apply a Gram–Schmidt orthonormalization to these vectors, we will obtain k vectors $\mathbf{y}_i$ with the desired properties, since it is easy to see that the joint distribution of the $\mathbf{y}_i$'s will be unchanged if they are subjected to an arbitrary orthogonal transformation. For the case $k = 1$, see Tocher (1963). For each set of k orthogonal vectors generated by this procedure we require mk standard normal deviates, k square roots and, approximately, $2mk^2$ arithmetic operations. The Markov chain method, on the other hand, requires only $6k$ arithmetic operations and the cosine and sine of a uniformly distributed angle and, if the function $f(\cdot)$ being averaged is evaluated only at times $T, 2T, \ldots$, we must have T as large as $\frac{1}{3}mk$ before the computing times of the two procedures are comparable. The degree of correlation amongst the sample values for the Markov chain method will depend upon the particular function $f(\cdot)$ and will determine, ignoring error analysis considerations, which of the two methods is better.

It is a pleasure to thank Messrs Damon Card, Stuart Whittington and Professor John P. Valleau, who generously shared their considerable knowledge of and experience with the Markov chain methods in statistical mechanics applications, Professor John C. Ogilvie, who made several helpful suggestions and Mr Ross D. MacBride, who capably prepared the computer programs. I am grateful to the referees whose comments were very helpful in the revision of this paper. The work was supported by the National Research Council of Canada.

REFERENCES

BARKER, A. A. (1965). Monte Carlo calculations of the radial distribution functions for a proton-electron plasma. *Aust. J. Phys.* **18**, 119–33.

BARTLETT, M. S. (1966). *An Introduction to Stochastic Processes*, 2nd edition. Cambridge University Press.

BLACKMAN, R. B. (1965). *Data Smoothing and Prediction*. Reading, Mass.: Addison-Wesley.

BLACKMAN, R. B. & TUKEY, J. W. (1958). *The Measurement of Power Spectra*. New York: Dover.

CHUNG, K. L. (1960). *Markov Processes with Stationary Transition Probabilities*. Heidelberg: Springer-Verlag.

ERHMAN, J. R., FOSDICK, L. D. & HANDSCOMB, D. C. (1960). Computation of order parameters in an Ising lattice by the Monte Carlo method. *J. Math. Phys.* **1**, 547–58.

FOSDICK, L. D. (1963). Monte Carlo calculations on the Ising lattice. *Methods Computational Phys.* **1**, 245–80.

FOX, L. & MAYERS, D. F. (1968). *Computing Methods for Scientists and Engineers*. Oxford: Clarendon Press.

FRASER, D. A. S. (1968). *The Structure of Inference*. New York: Wiley.

HAMMERSLEY, J. M. & HANDSCOMB, D. C. (1964). *Monte Carlo Methods*. London: Methuen.
HANDSCOMB, D. C. (1962). The Monte Carlo method in quantum statistical mechanics. *Proc. Camb. Phil. Soc.* **58**, 594–8.
HANNAN, E. J. (1957). The variance of the mean of a stationary process. *J. R. Statist. Soc.* B **19**, 282–5.
HANSEN, M. H., HURWITZ, W. N. & MADOW, W. G. (1953). *Sample Survey Methods and Theory*. New York: Wiley.
JAMES, A. T. (1955). A generating function for averages over the orthogonal group. *Proc. R. Soc.* A **229**, 367–75.
JAMES, A. T. (1964). Distributions of matrix variates and latent roots derived from normal samples. *Ann. Math. Statist.* **35**, 475–501.
JOWETT, G. H. (1955). The comparison of means of sets of observations from sections of independent stochastic series. *J. R. Statist. Soc.* B **17**, 208–27.
METROPOLIS, N., ROSENBLUTH, A. W., ROSENBLUTH, M. N., TELLER, A. H. & TELLER, E. (1953). Equations of state calculations by fast computing machines. *J. Chem. Phys.* **21**, 1087–92.
TOCHER, K. D. (1963). *The Art of Simulation*. London: English Universities Press.
TROTTER, H. F. & TUKEY, J. W. (1956). Conditional Monte Carlo for normal samples. In *Symposium on Monte Carlo Methods*, ed. H. A. Meyer, pp. 64–79. New York: Wiley.
WILK, M. B. & GNANADESIKAN, R. (1968). Probability plotting methods for the analysis of data. *Biometrika* **55**, 1–17.

[*Received February* 1969. *Revised July* 1969]