A Minimum Variance Sampling Technique for Simulation Models

A. J. BAYES

IBM Systems Development Institute, Canberra, Australia

ABSTRACT. In a normal simulation run, the states of the model are sampled in proportion to their natural frequency of occurrence. For a given sampling effort, this does not in general estimate a given statistic of the model with maximum precision. A sampling theory of Markov chains is developed which allows some statistics of the Markov state frequencies to be estimated with minimum variance for a given sampling effort. A technique is presented to allow the sampling frequency of the states of the simulation to be independent of their natural frequency. By representing a simulation model as a Markov chain, the theory is applied to estimate some statistics of the simulation model with minimum variance; for instance, the frequency of overload of a teleprocessing computer system. A numerical case is presented in which the sampling effort is reduced by a factor of sixty compared to a normal simulation run.

KEY WORDS AND PHRASES: minimum variance sampling, maximum precision sampling, simulation, Markov chains, Markov processes, queues, importance sampling, stratified sampling, probability, teleprocessing

CR CATEGORIES: 3.65, 5.5

1. *A Sampling Problem for a Markov Chain*

1.1. THE PROBLEM. Consider a finite Markov chain with n states labeled 1, 2, $\cdots$, n. Let the matrix P of transition probabilities be defined by

$$\begin{aligned} P_{i,i+1} &= p_i, & 1 \leq i < n, \\ P_{i+1,i} &= 1 - p_{i+1}, & 1 \leq i < n, \\ P_{i,j} &= 0, & \text{otherwise,} \end{aligned}$$

where $p_1 = 1$, $p_n = 0$, $0 < p_i < 1$, $1 < i < n$. Thus transition probabilities are zero except between adjacent states. Let t_i be the interval spent in state i. The expectation $E(t_i)$ and the variance var (t_i) are known.

We divide the states into two subsets by choosing an integer s where $1 < s < n$. The states 1, 2, $\cdots$, $s - 1$ are said to be in the lower set and the states s, $s + 1$, $\cdots$, n in the upper set. Let λ be the proportion of time spent in the upper set in the steady-state solution to the Markov chain. Then λ is a function of the values p_i and $E(t_i)$.

In the notation we represent matrices by uppercase letters, vectors by boldface lowercase letters, and variables by lightface lowercase letters. We append the sub-

Author's address: A. J. Bayes, IBM Systems Development Institute, IBM Australia Limited, 80 Northbourne Avenue, Canberra, A.C.T. 2600, Australia.

script 1 or 2 to the left to denote attributes of the upper and lower set, respectively. A subscript to the right defines an attribute of a state. For example, $\mathbf{t}$ is the vector $t_1, t_2, \cdots, t_n$, $_1\mathbf{t}$ *is the vector* $t_s, t_{s+1}, \cdots, t_n$, *and* $_2\mathbf{t}$ *is the vector* $t_1, t_2, \cdots, t_{s-1}$.

A sample of a state i is defined to be a sample of the random variable t_i and a ample of the transition out of the state. Suppose that state i is sampled m_i times. et m be defined by

$$\sum_i m_i = m. \tag{1}$$

et $\bar{t}_i$ be the estimate of $E(t_i)$ calculated as the mean of the m_i values. $\bar{p}_i$ is the stimate of p_i calculated as (number of transitions to state $i + 1$)$/m_i$. $\bar{\lambda}$ is the stimate of λ calculated from the steady-state solution of the Markov chain using and $\bar{p}_i$. Then the problem to be solved is as follows: For given s and fixed m, hoose m_i satisfying (1) so that var $(\bar{\lambda})$ is as small as possible.

1.2. The Duration in the Upper Set. Let l_i, $s \leq i \leq n$, be the remaining uration in the upper set when state i is entered. $\bar{l}_i$ is an estimate of l_i, calculated om recurrence relations using $\bar{t}_i$ and $\bar{p}_i$. In this subsection we calculate var $(\bar{l}_i)$, $\leq i \leq n$, or, in different notation, var $(_1\mathbf{l})$, as a function of $_1\bar{p}$, $_1\bar{t}$, and $_1\bar{m}$. Since the pper set is always entered at state s, the expected duration in the upper set is (l_s). The variance of the estimate of the expected duration is var $(\bar{l}_s)$.

It is easily seen that the l_i are connected by the following equations:

$$\left.\begin{aligned} l_s &= t_s + p_s l_{s+1}, \\ l_i &= t_i + p_i l_{i+1} + (1 - p_i) l_{i-1}, \qquad s < i < n, \\ l_n &= t_n + l_{n-1}. \end{aligned}\right\} \tag{2}$$

We put the terms involving l_i to the left-hand side and rewrite eqs. (2) in matrix orm as

$$_1D\ _1\mathbf{l} = {}_1\mathbf{t}, \tag{3}$$

where $_1D$ is defined appropriately as a function of the p_i. Equation (3) remains true when the variables are replaced by their expectations. Thus we can write

$$_1DE(_1\mathbf{l}) = E(_1\mathbf{t}). \tag{4}$$

Solving (4), we obtain

$$E(_1\mathbf{l}) = (_1D)^{-1}E(_1\mathbf{t}). \tag{5}$$

We now consider the effect on $E(_1\mathbf{l})$ of small changes in $_1\mathbf{p}$ and $_1\mathbf{t}$. The effect of mall changes in $_1\mathbf{t}$ is immediately obvious from (5). Suppose that p_i is changed by δp_i, $s \leq i < n$. Let $_1\mathbf{l}'$ be the variables, corresponding to $_1\mathbf{l}$, associated with the ew p_i. We can write the following equations which are similar to (2):

$$\left.\begin{aligned} l_s' &= t_s + (p_s + \delta p_s) l_{s+1}', \\ l_i' &= t_i + (p_i + \delta p_i) l_{i+1}' + (1 - p_i - \delta p_i) l_{i-1}', \quad s < i < n, \\ l_n' &= t_n + l_{n-1}. \end{aligned}\right\} \tag{6}$$

We now take expectations of (2) and (6) and subtract each equation in (2) from he corresponding equation in (6). We ignore second-order and higher terms in δp_i

and $(E(l_i') - E(l_i))$, since these are small. This yields

$$\left.\begin{aligned} \delta l_s &= l_{s+1}\delta p_s + p_s\delta l_{s+1}, \\ \delta l_i &= (l_{i+1} - l_{i-1})\delta p_i + p_i\delta l_{i+1} + (1 - p_i)\delta l_{i-1}, \qquad s < i < n, \\ \delta l_n &= \delta l_{n-1}. \end{aligned}\right\} \tag{7}$$

In eqs. (7) δl_i is written as a convenient shorthand for $E(l_i') - E(l_i)$.

Equations (7) have a very similar structure to (2). We define ${}_1\mathbf{u}$ by the equations

$$\left.\begin{aligned} u_s &= l_{s+1}\delta p_i, \\ u_i &= (l_{i+1} - l_{i-1})\delta p_i, \qquad s < i < n, \\ u_n &= 0. \end{aligned}\right\} \tag{8}$$

Equations (7) can be solved for $\delta\,{}_1\mathbf{l}$ in terms of $\delta\,{}_1\mathbf{p}$ giving

$$\delta\,{}_1\mathbf{l} = ({}_1D)^{-1}\,{}_1\mathbf{u}. \tag{9}$$

The samples of the ${}_1\mathbf{t}$ and the transitions are all independent, so the errors are independent. Equation (5) expresses $E({}_1\mathbf{l})$ linearly in $E({}_1\mathbf{t})$. Equation (9) expresses $\delta\,{}_1\mathbf{l}$ linearly in terms of $\delta\,{}_1\mathbf{p}$ (in ${}_1\mathbf{u}$). Hence var $({}_1\mathbf{l})$ can be found by adding the components of variance due to individual errors in ${}_1\bar{\mathbf{t}}$ and ${}_1\bar{\mathbf{p}}$. The variance of the error in estimating the ith term of $E({}_1\mathbf{t})$ is var $(t_i)/m_i$ and the variance of the error of estimation of p_i is $p_i(1 - p_i)/m_i$, $s \leq i < n$. If the m_i are large, then the estimation errors are small and the above equations can be applied. Thus we define ${}_1\mathbf{v}$ by the equations

$$\left.\begin{aligned} v_s &= (\text{var } (t_s) + p_s(1 - p_s)E(l_{s+1})^2)/m_s, \\ v_i &= (\text{var } (t_i) + p_i(1 - p_i)(E(l_{i+1}) - E(l_{i-1}))^2)/m_i, \quad s < i < n, \\ v_n &= \text{var } (t_n)/m_n. \end{aligned}\right\} \tag{10}$$

Then from (5) and (9),

$$\text{var } ({}_1\bar{\mathbf{l}}) = ({}_1D)^{-1}\,{}_1\mathbf{v}. \tag{11}$$

This is the result we require.

A similar result can be obtained for the lower set, using an almost identical argument.

1.3. Choice of m_i To Minimize Var $(\bar{\lambda})$. We first state a result which we require later. Let x and y be independent random variables whose s.d. is small compared to $E(x) + E(y)$. Then the variance of $x/(x + y)$ is

$$[E(y)^2 \text{ var } (x) + E(x)^2 \text{ var } (y)]/(E(x) + E(y))^4. \tag{12}$$

This is merely the expression

$$\text{var } (x)\left(\frac{\partial}{\partial x}\left(\frac{x}{x + y}\right)\right)^2 + \text{var } (y)\left(\frac{\partial}{\partial y}\left(\frac{x}{x + y}\right)\right)^2$$

evaluated at $x = E(x)$, $y = E(y)$. It follows that for fixed $E(x)$ and $E(y)$, the variance of $x/(x + y)$ minimizes with

$$E(y)^2 \text{ var } (x) + E(x^2) \text{ var } (y). \tag{13}$$

The value λ can be calculated as (expected interval in the upper set)/(expected interval in the upper set plus expected interval in the lower set). That is,

$$\lambda = E(l_s)/(E(l_s) + E(l_{s-1})). \tag{14}$$

Here l_{s-1} is the expected remaining interval in the lower set on entering state $s - 1$. Using expression (13), var $(\bar{\lambda})$ is minimized if

$$E(l_s)^2 \text{ var } (\bar{l}_{s-1}) + E(l_{s-1})^2 \text{ var } (\bar{l}_s) \tag{15}$$

is minimized. The values in this expression can be calculated from eqs. (5) and (11) and corresponding equations for the lower set.

The expectations in (15) are independent of m_i. The variances involve v_i and are linear in the reciprocals of m_i. We define f_i to be the coefficient of $1/m_i$ in (15). Then (15) can be rewritten as

$$\sum_i (f_i/m_i). \tag{16}$$

It is easily shown, using Lagrange's undetermined multipliers, that the minimum of (16) with respect to (1) occurs when

$$m_i = k\sqrt{f_i}, \tag{17}$$

where

$$k = m/(\sum_i \sqrt{f_i}). \tag{18}$$

The minimum value of (16) is

$$(\sum_i \sqrt{f_i})^2/m. \tag{19}$$

Using (12) and (17) the minimum of var $(\bar{\lambda})$ is

$$(\sum_i \sqrt{(f_i)})^2/((E(l_s) + E(l_{s-1}))^4 m). \tag{20}$$

Equations (17) and (18) give the answer to the problem posed at the start of this section.

2. *Application to a Problem in Simulation*

The purpose of a simulation model is to answer a question or questions about the modeled system. The most natural and frequently used method is to write a program representing the system to be modeled, to choose an initial state for the system, to run the program for a period of simulated time from the initial state, to print statistics and data from the simulation run, and to answer questions about the modeled system from this output. This method is almost always statistically wasteful because the states of the modeled system are sampled in proportion to their natural frequency. It is highly unlikely that this sampling frequency answers a question to a given precision with minimum sampling effort. We would expect the sampling frequency to vary with the question that is being asked.

We develop the discussion in terms of a teleprocessing computer system, although it also applies to many other simulation subjects. Suppose a teleprocessing system is simulated to find the frequency of overload, defined to occur when a queue of messages in core exceeds the space which has been allocated for it. Since overloading is rare, a long duration must be simulated to obtain an accurate estimate of its

frequency. Most of the time the queue is small, and very little information is being generated about conditions of overload. We need a method to force the simulation to spend more time investigating the behavior of large queues. One way of doing this has been called importance sampling by J. E. Flanagan. Its application to simulation is discussed by Bayes [1] and may be illustrated by the following simple example:

Suppose we wish to estimate the proportion of time that the size q of a queue is equal to or greater than 15. Consider the following scheme:

1. Set variables $V1$ and $V2$ to zero. Start the simulation with some value of q less than 10.
2. Run the simulation until $q = 10$. Add the elapsed time to $V1$. Store the variables which define the state of the system.
3. Simulate until $q < 10$. Add $\frac{1}{4}$ of the elapsed time to $V1$, and $\frac{1}{4}$ of the time that $q \geq 15$ to $V2$.
4. Restore the state variables which were stored in step 2 and repeat step 3 to a total of four times. Then go to step 2.
5. At the end of the run, use $V2/V1$ as the estimate.

In this example, $q = 10$ is an importance boundary and 4 is the importance factor. An immediate generalization is to construct a number of importance boundaries, each with its own importance factor. The boundaries must not intersect. Statistics collected in a region influenced by several boundaries are scaled down by the product of the associated importance factors.

Importance sampling appears similar to stratified sampling [2]. The difference is that in stratified sampling the population is allocated to classes of known relative size. Weighted samples of the classes are used to estimate an attribute of the parent population. In importance sampling the population is allocated to classes of unknown relative size. Weighted samples of the population are used to estimate the relative size of the classes.

To apply the foregoing theory, we approximate the simulation by a Markov chain as follows: (1) State i occurs when i messages are held in core; (2) the transition probabilities from state i to state $i - 1$ or $i + 1$ are the same as in the simulation model; (3) the mean and variance of the interval spent in state i, for each i, are the same as in the simulation model.

It seems likely that the statistical properties of a Markov chain thus defined will be similar to the statistical properties of the simulation. In any case, we assume that optimum sampling frequencies, calculated from the Markov chain, will be near optimal for the simulation and that the "natural" sampling frequencies of the Markov chain are close to the natural sampling frequencies of the simulation.

To find the importance factors corresponding to the optimum sampling frequencies, it is necessary to know the relative frequency with which the states are entered in the steady-state solutions of the Markov chain. It is known [4] that, if G is the transition matrix of an irreducible ergodic chain, then $\lim_{n\to\infty} G^n$ exists and all the rows are equal. The sum of the elements in any row is unity. The elements in a row are the relative frequency of entering the corresponding state.

The transition matrix P is not ergodic since it is periodic. To pass from a state back to itself requires an even number of transitions. The matrix

$$G \equiv (P + I_n), \tag{21}$$

where I_n is the identity matrix, is ergodic and has the required property. Thus if we define $\mathbf{g}$ as the top row of $\lim_{n\to\infty} G^n$, then $\mathbf{g}$ is the natural sampling frequency.

The ratio of optimum to natural sampling frequency is thus m_i/g_i. The importance boundaries and factors are chosen so that, over i, the product of the importance factors influencing state i is proportional to m_i/g_i.

In a real life simulation the values required for estimating the importance factors are not known in advance. It might be argued that this invalidates the method as a practical procedure. There are two possible solutions. One possibility is to estimate the factors either intuitively or from an analytic solution of a simplified model. If the estimates are badly chosen, λ will have a somewhat larger variance than the minimum for a given sampling effort. However, the result will almost certainly be better than using no importance sampling. Another possibility is to perform the simulation twice and use the results of the first simulation to guide the choice of importance factors in the second simulation. This is similar to two-stage sampling as used in stratified sampling.

3. *Optimum Sampling for Mean Queue Length*

In this section we generalize the notation established so far by appending to the right of a variable a subscript (or another subscript) to define the lower boundary of the upper set. For instance, $l_{i,j}$, $i \geq j$, is the remaining time, on entering state i, in the upper set whose lower boundary is j.

Consider the sampling variance of

$$\sum_{i=1}^{n} \alpha_i \bar{\lambda}_i , \tag{22}$$

where the α_i are arbitrary constants. From (12), (14), (15), and (16) we can write

$$\begin{aligned} \operatorname{var}\left(\sum_{i=1}^{n} \alpha_i \bar{\lambda}_i\right) &= \sum_{i=1}^{n} \left(\alpha_i \sum_j [f_{j,i}/(E(l_{i,i}) + E(l_{i-1,i}))^4 m_j]\right), \\ &= \sum_j m_j^{-1} \left(\sum_{i=1}^{n} [\alpha_i f_{j,i}/(E(l_{i,i}) + E(l_{i-1,i}))^4]\right). \end{aligned} \tag{23}$$

Thus the sampling variance of (22) is a linear function of $1/m_i$ and can be minimized by a suitable choice of m_i by using the method described in Section 2.

This result has an application to the teleprocessing simulation discussed earlier. For instance, if we set $\alpha_i = 1$, $1 \leq i \leq n$, then (22) is the mean queue length. As another example, if $\alpha_i = 2 \cdot i - 1$, $1 \leq i \leq n$, then (22) is the second moment of the queue length. Other values of α might be of interest in particular circumstances.

4. *A Numerical Example*

We consider a simple queue with a service utilization of $\frac{2}{3}$ and with negative exponential distributions for the service time and interarrival time. The maximum permitted size of the queue is 20, and any arrivals when the queue size is 20 are lost. We wish to estimate the proportion of the time that the queue size is 16 or more.

The model is represented accurately by a Markov chain with the following parameters:

$p_1 = 1,$

$$
\begin{aligned}
p_{20} &= 0, \\
p_i &= 0.4, \quad 1 < i < 20, \\
E(t_1) &= \text{var } (t_1) = 2.5, \\
E(t_{20}) &= \text{var } (t_{20}) = 1.666667, \\
E(t_i) &= \text{var } (t_i) = 1, \quad 1 < i < 20, \\
s &= 16.
\end{aligned}
$$

Table I shows some statistics for the states in the Markov chain. Column (1) is the state number i. Column (2) is the "natural" sampling frequency g_i. Column (3) is the optimum sampling frequency derived from eq. (18) and normalized by setting $m = 1$. Column (4) is the ratio of column (3) to column (2), scaled to set the first value to unity. This column is used to guide the choice of importance factors. Column (5) is a set of importance factors chosen by hand so that the product of the factors up to state i, shown in column (6), is an approximation to column (4).

The s.d.of the estimate of $\bar{\lambda}$, based on $m = 1$, is, for "natural" sampling, 0.161; for optimized sampling using column (3), 0.0193; using weights from column (6), 0.020. The use of importance sampling has reduced the standard deviation of the estimate by a factor of about 8 or for fixed precision of the estimate, has reduced the sample size by a factor of about 60. Numerical studies show that, in general, the rarer the event being estimated, the more worthwhile is importance sampling.

The peak in the optimized frequency occurs at the boundary between the upper and the lower set. This has been observed in many numerical examples and is intuitively reasonable.

Although column (6) appears to be a poor approximation to column (4), the effect of the approximation on the precision of the estimate is negligible.

The optimized frequencies in column (3) provide data for a crude sensitivity

TABLE I. STATISTICS FOR NUMERICAL EXAMPLE

(1) *State number*	(2) *"Natural" frequency*	(3) *Optimized frequency*	(4) *Col. 3 : Col. 2*	(5) *Importance factors*	(6) *Weights using (5)*
1	0.1760	0.0013	1		1
2	0.2779	0.0046	2	2	2
3	0.1852	0.0076	5	3	6
4	0.1235	0.0109	12	2	12
5	0.0823	0.0147	24	2	24
6	0.0549	0.0191	46	2	48
7	0.0366	0.0244	88	2	96
8	0.0242	0.0306	165		96
9	0.0163	0.0381	308	3	288
10	0.0108	0.0472	572	2	576
11	0.0072	0.0582	1058	2	1152
12	0.0048	0.0716	1952		1152
13	0.0032	0.0880	3598	3	3456
14	0.0021	0.1080	6624	2	6912
15	0.0014	0.1324	12185		6912
16	0.0010	0.1286	17745	3	20736
17	0.0006	0.0953	19719		20736
18	0.0004	0.0660	20487		20736
19	0.0003	0.0398	18540		20736
20	0.0001	0.0136	15796		20736

analysis on the raw data for the Markov chain. Let r_i be the number of samples of ith state. The s.d. of the sampling estimates of the data for state i is of the order of $r_i^{-\frac{1}{2}}$. The reduction in the s.d. due to one extra sample is the differential of this, that is $r_i^{-\frac{3}{2}}$. If the states have been sampled optimally, it is a matter of indifference which state is sampled next. Thus the relative importance of small errors of equal size in the data for state i and state j is of the order of $(r_i/r_j)^{\frac{3}{2}}$.

Sensitivity analyses are important in simulation, since frequently a greater accuracy in the data can be obtained, although at a greater cost. The figures presented can be used to help minimize the cost of data collection for a given accuracy of simulated answer. Future work is planned in this area.

REFERENCES

(*Note.* Reference [3] is not cited in the text.)

1. BAYES, A. J. Statistical techniques for simulation models. *Australian Comput. J. 2* (1970), 180–184.
2. COCHRAN, W. G. *Sampling Techniques.* Wiley, New York, 1963, Ch. 5.
3. CRAMER, H. *Mathematical Methods of Statistics.* Princeton U. Press, Princeton, N. J., 1946.
4. FELLER, W. *An Introduction to Probability Theory and Its Applications.* Wiley, New York, 1950.

RECEIVED APRIL 1971; REVISED FEBRUARY 1972