REPLICATION SPLITTING AND VARIANCE FOR SIMULATING DISCRETE-PARAMETER STOCHASTIC PROCESSES

W. David KELTON

Department of Industrial and Operations Engineering, The University of Michigan, Ann Arbor, MI 48109-2117, USA

Received July 1985
Revised November 1985

Previous results for stationary continuous-time processes concerning allocation of a fixed amount of simulation effort across independent replications are extended both to stationary and certain non-stationary discrete-time processes. In particular, in the presence of positive autocorrelation, variance is reduced if more short replications are designed. The magnitude, however, of the variance reduction is not great as long as the computation budget is not tight, suggesting that a good strategy is to design for a moderate number of replications in any case, which also mitigates potential bias problems.

simulation * variance * replication * budget constraint

1. Introduction

One of the principal drawbacks of using simulation to study the behavior of complex stochastic systems is that we obtain only estimates (as opposed to exact values) of desired system characteristics. Such estimators are properly regarded as random variables (r.v.'s), whose degree of imprecision or uncertainty is typically measured by their variance. Accordingly, considerable effort has been devoted to finding techniques to reduce the variance of such output r.v. estimators, at litttle or no additional cost, in which case more precise results are obtained for the same simulation effort, or (equivalently) less effort is required to attain a desired precision. Many of these *variance reduction techniques* (common random numbers, antithetic variates, and control variates, for example) manipulate the random number generator to induce certain correlations in the simulation output which then enter the variance formula, with appropriately signed coefficients, to reduce the variance of the final estimator. Thus, some amount of internal modification of the simulation code itself is usually required to use such techniques.

This paper examines a different method of variance reduction that is entirely external to the simulation program, affecting only the duration and number of independent replications of the simulation through the experimental design of the simulation study. Assuming a budget constraint given in terms of the total amount of simulation possible (expressed either as simulation clock time or as the number of discretely-indexed observations), a decision must be made before the simulations are run as to the number of independent, identically initialized and terminated replications to make, and the duration of each. Gafarian and Ancker [2] considered monitoring a stationary continuous-time process with a positive, exponentially declining autocorrelation function during a simulation, and showed that it is better (in terms of variance of the time-integral output estimator) to break up the simulation effort into 'many short' runs, rather than 'a few long' runs. This paper establishes similar results for discrete-time processes, which are often more informative and easier to observe in simulation, that are either (a) stationary with any form of positive autocorrelation function, or (b) non-stationary first-order autoregressive (AR(1)) with positive multiplicative factor; a counterexample, however, demonstrates that

0167-6377/86/$3.50

the result does not hold for arbitrary non-stationary discrete-time processes, even with positive autocorrelation.

Section 2 treats stationary processes with arbitrary positive autocorrelation function, and Section 3 shows that, while similar results do not in general hold for non-stationary processes, they are valid for positively correlated non-stationary AR(1) processes. In Section 4 some numerical quantification of the results is presented, and conclusions and observations appear in Section 5.

2. Stationary processes

Let $\{X_1, X_2, \ldots\}$ be a covariance stationary process with $E(X_i) = \mu$ and $\gamma_p = \text{cov}(X_i, X_{i+p})$. From a simulated realization $X_1, X_2, \ldots, X_m$ of m consecutive observations from the process, $\bar{X}_m = \sum_{i=1}^{m} X_i/m$ is an unbiased estimator of μ, with

$$\text{var}(\bar{X}_m) = \left(\gamma_0 + 2\sum_{p=1}^{m-1}(1-p/m)\gamma_p\right)/m. \qquad (1)$$

(Empty sums, such as (1) in the case $m = 1$, are taken throughout as 0.) Thus, if we make k independent replications of length m observations each, resulting in k independent realizations of $\bar{X}_m$, our final unbiased point estimator is $\bar{\bar{X}}_{km}$, the sample average of the k independent $\bar{X}_m$'s, which has variance equal to the expression in (1), divided by k.

Suppose a budget constraint is imposed in the form of a limit n on the *total* number of X_i's that can be simulated, regardless of how these n observations are allocated to replications. We must then decide, before simulating, on how many replications k to make, each of length $m = n/k$, under the budget constraint. (It is assumed that n is divisible by k, a mild restriction since n will probably be relatively large. Also, we require the replications to be of equal length m to preserve the identically distributed nature of the within-replication averages, in order to allow application of statistical methods based on an identical-distribution assumption.) Since $\bar{\bar{X}}_{km}$ is unbiased for μ regardless of the choice of k and m, it is reasonable to focus on the effect of the splitting of n into k times m on $\text{var}(\bar{\bar{X}}_{km})$; the following result shows that in the common case of positive autocovariance, choosing many short replications is preferable to choosing a few long replications.

Proposition 1. *For $j = 1, 2$ let k_j be a positive integer dividing n, let $m_j = n/k_j$, and assume that $k_1 < k_2$. For a covariance stationary process $\{X_1, \ldots, X_n\}$ with $\gamma_p \geq 0$ for all p, we have* $\text{var}(\bar{\bar{X}}_{k_1 m_1}) \geq \text{var}(\bar{\bar{X}}_{k_2 m_2})$.

Proof. For $j = 1, 2$ let

$$g_j = \sum_{p=1}^{m_j - 1}(1 - p/m_j)\gamma_p,$$

so that $\text{var}(\bar{\bar{X}}_{k_j m_j}) = (\gamma_0 + 2g_j)/n$. Thus, it is enough to show that $g_1 \geq g_2$. Since $m_1 > m_2$,

$$g_1 - g_2 = \left(\frac{1}{m_2} - \frac{1}{m_1}\right)\sum_{p=1}^{m_2-1} p\gamma_p + \sum_{p=m_2}^{m_1-1}\left(1 - \frac{p}{m_1}\right)\gamma_p, \qquad (2)$$

which is clearly non-negative since the autocovariance function is non-negative. □

Note that the assumption of non-negative autocorrelation was used only in the final step of the proof, and that non-negativity of (2) is necessary and sufficient for the result of the proposition to hold. If the γ_p's could be negative (as can arise, for example, in inventory systems) then (2) could be negative, resulting in the opposition conclusion, i.e., that a few long replications are preferable to may short ones. Thus, some knowledge of the sign of the autocorrelations would appear to be helpful in designing the simulation experiment.

3. Non-stationary processes

Simulation of complex systems typically results in output stochastic processes which are non-stationary, due to the often artificial nature of the initial conditions needed to start the simulation. This section establishes a result similar to Proposition 1 for one useful class of such processes; the proof, however, is entirely different.

Before establishing this, we demonstrate that the result of Proposition 1 does not hold in general for any discrete-parameter non-stationary process, even if positive autocovariance is assumed. For a general (possibly non-stationary) process $\{X_1, X_2, \ldots\}$ let $\gamma_{ij} = \text{cov}(X_i, X_j)$ and define $\bar{\bar{X}}_{km}$ formally as in Section 2 (except now

using the X_i's from the non-stationary process). In this case,

$$\operatorname{var}\left(\bar{\bar{X}}_{km}\right) = \frac{1}{n^2} k \sum_{i=1}^{m} \sum_{j=1}^{m} \gamma_{ij}, \tag{3}$$

where $n = km$. As a counterexample, let $\{X_1, X_2, \ldots, X_6\}$ be multivariate normal with covariance matrix

$$\Sigma = \begin{pmatrix} 4 & 3 & 1 & 1 & 1 & 1 \\ 3 & 3 & 1 & 1 & 1 & 1 \\ 1 & 1 & 2 & 1 & 1 & 1 \\ 1 & 1 & 1 & 2 & 1 & 1 \\ 1 & 1 & 1 & 1 & 2 & 1 \\ 1 & 1 & 1 & 1 & 1 & 2 \end{pmatrix}.$$

Note that Σ is positive definite (so such a process exists), and that the correlations are all positive (but not stationary). Choosing $n = 6$ and, as before, setting $m = n/k$, (3) is equal to 1.056 if $k = 2$, but is 1.083 when $k = 3$; thus, increasing k led to an increase in $\operatorname{var}(\bar{\bar{X}}_{km})$, contrary to Proposition 1.

One general class of non-stationary processes, however, does yield the result of Proposition 1. The AR(1) process is defined by the recursion

$$X_i = \mu + \phi(X_{i-1} - \mu) + \epsilon_i,$$

for $i = 1,2,\ldots,$ and the ϵi's are a sequence of uncorrelated r.v.'s with mean 0 and common variance σ^2; we assume throughout that $|\phi| < 1$. This class of processes was introduced as a model for simulation output processes by Fishman [1] and investigated further by Turnquist and Sussman [4], and by Kelton and Law [3]. While making such an assumption certainly entails some amount of approximation, the relatively simple form of the AR(1) process enables a more intensive analysis. Further, this process shares many important features with actual output processes from simulation, such as having an autocorrelation function that (at least asymptotically) declines exponentially; depending on the initial specification of X_0, the process may or may not be stationary.

If we make the additional assumption that the ϵ_i's are normally distributed and that X_0 is drawn from a normal distribution with mean μ and variance $\sigma^2/(1-\phi^2)$, then the $\{X_i\}$ process is stationary with mean 0, variance $\sigma^2/(1-\phi^2)$, and lag-p autocovariance $\gamma_p = \phi^p \sigma^2/(1-\phi^2)$; thus, the result of Proposition 1 would apply. If, however, we do not make these additional assumptions and let X_0 be deterministically specified, then the $\{X_i\}$ process is neither first- nor second-order stationary; for the rest of this section, we will assume that this is the case, and so Proposition 1 would not apply.

The conclusion of Proposition 1, however, still holds, as we show in the remainder of this section. Defining $\bar{X}_m$ and $\bar{\bar{X}}_{km}$ formally as in Section 2 (except now using the X_i's from the non-stationary AR(1) process), note first that from eq. (4) of [3],

$$\operatorname{var}(\bar{X}_m) = \frac{\sigma^2}{m(1-\phi)^2} \times \left(1 - \frac{\phi(1-\phi^m)}{m(1-\phi^2)}(2 + \phi(1-\phi^m))\right). \tag{4}$$

The following Lemma is stated for use in evaluating the key expression appearing below in Proposition 2.

Lemma. *For q a non-negative integer and any real y,*

(i) $$\sum_{p=0}^{q} y^p = (1 - y^{q+1})/(1-y),$$

(ii) $$\sum_{p=0}^{q} p y^{p-1} = \left(1 - (q+1)y^q + q y^{q+1}\right)/(1-y)^2,$$

(iii) $$\sum_{p=0}^{q} p^2 y^{p-2} = \left(1 + y - (q+1)^2 y^q + (2q^2 + 2q - 1) y^{q+1} - q^2 y^{q+2}\right)/\left(y(1-y)^3\right).$$

Proof (i) follows from induction on q, and (ii) and (iii) are obtained by successively differentiating through (i) with respect to y. □

Proposition 2. *Let n, k_j, and m_j be as in Proposition 1. Then for any AR(1) process with $\phi > 0$, we have* $\operatorname{var}(\bar{\bar{X}}_{k_1 m_1}) \geq \operatorname{var}(\bar{\bar{X}}_{k_2 m_2})$.

Proof. Dividing (4) by k_j and rewriting yields

$$\operatorname{var}\left(\bar{\bar{X}}_{k_j m_j}\right) = c_1\left(1 - c_2 g_{m_j}\right),$$

with $c_1 = \sigma^2/(n(1-\phi)^2)$, $c_2 = \phi/(1-\phi^2)$, and g_m

$= (2(1-\phi^m) + \phi(1-\phi^m)^2)/m$ for any positive integer m; note that neither c_1 nor c_2 depends on m_j. Thus, it is enough to show that $g_{m_1} \le g_{m_2}$, which would follow establishing that g_m is non-increasing in m (since $m_1 \ge m_2$). Noting that $g_m - g_{m+1} \ge 0$ if and only if

$$h_m = 2 + \phi - 2(m+1)\phi^m - 2\phi^{m+1} + 2m\phi^{m+2} + (m+1)\phi^{2m+1} - m\phi^{2m+3}$$

is non-negative, we use the relations in the Lemma to rewrite

$$h_m = \frac{(1-\phi)^3}{2}\left(\sum_{p=0}^{m-1} (3p+4)(p+1)\phi^p + \phi^m \sum_{p=0}^{m} (-p^2 - (2m+1)p + 3m(m+1))\phi^p \right). \quad (5)$$

Since $\phi > 0$ and the coefficients in (5) of ϕ^p in the summations are always positive over the sums' respective ranges, we see that $h_m \ge 0$, and the proof is complete. □

Again, the assumption of non-negative autocovariance ($\phi > 0$ here) is critical for the result, and the opposite conclusion could be reached otherwise. In principle, one could investigate whether the conclusion of Proposition 2 holds for higher-order autoregressive processes, as well as for more general ARMA processes, by using methods similar to those above; it seems, however, that the complexity involved would be formidable.

4. Numerical illustration

Propositions 1 and 2 establish inequalities about the variances resulting from alternative splitting of the simulation budget, but say nothing about the magnitude of the variance reduction obtained from choosing a larger value of k. In this section we use the AR(1) model (both stationary and non-stationary) to quantify the nature of the decrease in $\text{var}(\bar{\bar{X}}_{km})$ as k increases, with n fixed.

With the AR(1) process (with normal ϵ_i's) initialized by drawing X_0 from a normal distribution with mean μ and variance $\sigma^2/(1-\phi^2)$, the process is stationary with lag-p autocovariance $\gamma_p = \phi^p\sigma^2/(1-\phi^2)$. Combining this with (1) results in

$$\text{var}\left(\bar{\bar{X}}_{km}^{(S)}\right) = \sigma^2 \frac{m - 2\phi - m\phi^2 + 2\phi^{m+1}}{km^2(1+\phi)(1-\phi)^3},$$

where the superscript (S) denotes stationary.

On the other hand, if the AR(1) (with possibly non-normal ϵ_i's) is initialized via a deterministic choice for X_0, we get from the expression for $\text{var}(\bar{\bar{X}}_{km})$ in the proof of Proposition 2 that

$$\text{var}\left(\bar{\bar{X}}_{km}^{(NS)}\right) = \text{var}\left(\bar{\bar{X}}_{km}^{(S)}\right) - \sigma^2 \frac{(\phi^{m+1} - \phi)^2}{km^2(1+\phi)(1-\phi)^3},$$

where the superscript (NS) denotes non-stationarity. As an aside, note that $\text{var}(\bar{\bar{X}}_{km}^{(NS)}) \le \text{var}(\bar{\bar{X}}_{km}^{(S)})$, evidently reflecting the lower variability induced by the deterministic initialization in the non-stationary case.

With $\sigma = 1$, n fixed at 1000, 2000, 4000 and 8000, and setting $m = n/k$, Figure 1 plots the standard deviations $\sqrt{\text{var}\left(\bar{\bar{X}}_{km}^{(S)}\right)}$ (solid lines) and $\sqrt{\text{var}\left(\bar{\bar{X}}_{km}^{(NS)}\right)}$ (dashed lines) as functions of k, for $k \le 125$ and dividing n; ϕ is taken as 0.90. The decrease in variance with k is evident in all cases, but is less marked for a more generous budget. In fact, for n not too small, the quantitative advantage of splitting into many (high k) replications appears to be only minor. (From plots not shown for other values of ϕ, the size of n required for the preceding statement to hold is smaller for small ϕ, and vice versa.) This is significant in the non-stationary case if we are using $\bar{\bar{X}}_{km}^{(NS)}$ as an estimator of μ; this estimator is biased for μ, and the bias increases with k (see [3]), so that choosing k small is attractive from this standpoint. In the stationary case, however, $\bar{\bar{X}}_{km}^{(S)}$ is unbiased for μ, and it appears advantageous to make a number of replications provided that it is neither expensive nor inconvenient to implement the multiple simulation initializations this would entail.

Finally, Figure 1 displays a curious crossing of the dashed lines, for example, $\text{var}(\bar{\bar{X}}_{100,10}^{(NS)})$ ($n = 1000$) is less than $\text{var}(\bar{\bar{X}}_{100,20}^{(NS)})$ ($n = 2000$), so that the variance is larger with *more* total data. To explain this apparent paradox, recall that these simulations are initialized deterministically, and allowing them to run for $m = 20$ points (rather than $m = 10$) extends further away from the deterministic initialization, providing greater opportunity for variation that in some cases more than offsets the increased information in the additional data.

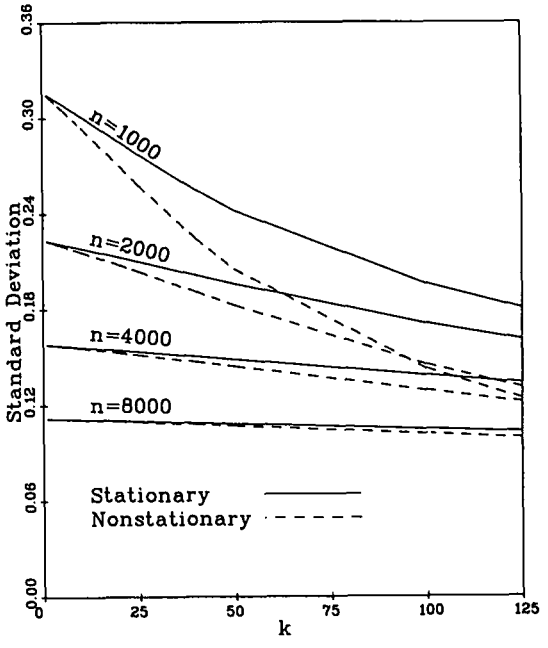

Fig. 1. $\sqrt{\text{var}(\bar{\bar{X}}^{(S)}_{k,n/k})}$ (solid lines) and $\sqrt{\text{var}(\bar{\bar{X}}^{(NS)}_{k,n/k})}$ (dashed lines) as functions of k for AR(1) processes with $\sigma = 1$ and $\phi = 0.90$.

5. Conclusions

This paper has focused on variability of estimators of means of positively autocorrelated discrete-parameter processes (or asymptotic means in the non-stationary case), and showed that splitting the budget into multiple replications is always preferable in the stationary case, may not be in the general non-stationary case, but is still preferable for the non-stationary AR(1) model considered. Looking at the actual magnitudes, however, of the variance reductions obtained for both stationary and non-stationary AR(1) processes, it appears that relatively little is to be gained by splitting into many short replications unless the budget is tight. Thus, especially in the non-stationary case where bias may also be a concern, a moderate number (no more than, say, 25) of replications should result in reasonably stable estimators and adequate degrees of freedom for efficient application of various statistical procedures, such as hypothesis testing about μ or forming confidence intervals for μ. A cost model incorporating both the cost of variance and the cost of eliminating the biasing effects of such non-stationarity on point estimators could quantify the tradeoffs involved in the splitting of the budget into replications. In any case, it appears that in automated procedures for run length determination, preference should be given to run length increase over increases beyond about 25 in the number of replications to attain a desired precision in the output.

Acknowledgements

The author would like to express his thanks to Averill M. Law for his valuable input, and to Diane P. Bischak for her careful reading of the manuscript. The comments of two referees were also quite helpful, including pointing out an error in the expression for $\text{var}(\bar{\bar{X}}^{(S)}_{km})$.

References

[1] G.S. Fishman, "Bias considerations in simulation experiments", *Operations Research* **20**, 785–790 (1972).

[2] A.V. Gafarian and C.J. Ancker, Jr., "Mean value estimation from digital computer simulation", *Operations Research* **14**, 25–44 (1966).

[3] W.D. Kelton and A.M. Law, "An analytical evaluation of alternative Strategies in steady-state simulation", *Operations Research* **32**, 169–184 (1984).

[4] M.A. Turnquist and J.M. Sussman, "Toward guidelines for designing experiments in queueing simulation", *Simulation* **28**, 137–144 (1977).