Proceedings of the 2013 Winter Simulation Conference
R. Pasupathy, S.-H. Kim, A. Tolk, R. Hill, and M. E. Kuhl, eds.

ARD: AN AUTOMATED REPLICATION-DELETION METHOD FOR SIMULATION ANALYSIS

Emily K. Lada
Anup C. Mokashi

SAS Institute Inc.
100 SAS Campus Drive,
Cary, NC 27513-8617, USA

James R. Wilson

North Carolina State University
Edward P. Fitts Department of
Industrial and Systems Engineering
Raleigh, NC 27695-7906, USA

ABSTRACT

ARD is an automated replication-deletion procedure for computing point and confidence interval (CI) estimators for the steady-state mean of a simulation-generated output process. The CI can have user-specified values for its absolute or relative precision and its coverage probability. To compensate for skewness in the truncated sample mean for each replication, the CI incorporates a skewness adjustment. With increasingly stringent precision requirements, ARD's sampling plan increases the run length and number of runs so as to minimize a weighted average of the mean squared errors of the following: (i) the grand mean of the truncated sample means for all runs; and (ii) the conventional replication-deletion estimator of the standard error of (i). We explain the operation of ARD, and we summarize an experimental performance evaluation of ARD. Although ARD's CIs closely conformed to given coverage and precision requirements, ARD generally required a larger computing budget than single-run procedures.

1 INTRODUCTION

In a nonterminating simulation, we are often interested in long-run (steady-state) average performance measures. Let $\{X_{i,j} : j = 1,2,...,n_i\}$ denote a stochastic process representing the sequence of outputs generated by the ith run of a nonterminating probabilistic simulation of length n_i for $i = 1,\ldots,q$. If the simulation is in steady-state operation, then the random variables $\{X_{i,j}\}$ will have the same steady-state cumulative distribution function (c.d.f.) $F_X(x) = \Pr\{X_{i,j} \le x\}$ for $i = 1,\ldots,q$, for $j = 1,\ldots,n_i$, and for all real x.

Usually in a nonterminating simulation, we are interested in constructing point and confidence interval (CI) estimators for some parameter of the steady-state c.d.f. $F_X(\cdot)$. In this work, we are primarily interested in estimating the steady-state mean, $\mu_X = \mathrm{E}[X] = \int_{-\infty}^{\infty} x\,\mathrm{d}F_X(x)$; and we limit the discussion to output processes for which $\mathrm{E}[X_{i,j}^2] < \infty$ so that the process mean μ_X and process variance $\sigma_X^2 = \mathrm{Var}[X_{i,j}] = \mathrm{E}[(X_{i,j} - \mu_X)^2]$ are well defined. In terms of the sample mean $\overline{X}_i = n_i^{-1}\sum_{j=1}^{n_i} X_{i,j}$ computed on run i, we also assume that the steady-state variance parameter $\gamma_X = \lim_{n_i\to\infty} n_i \mathrm{Var}\left[\overline{X}_i\right]$ is nonzero and finite.

One fundamental problem associated with analyzing stochastic output from a nonterminating simulation is that we usually do not possess sufficient information to start a simulation in steady-state operation; and thus it is necessary to determine an adequate length for the initial "warm-up" period so that for each simulation output generated after the end of the warm-up period, the corresponding expected value is sufficiently close to the steady-state mean. If observations generated prior to the end of the warm-up period are included in the analysis, then the resulting point estimator of the steady-state mean may be biased. This phenomenon is known as the start-up (or initialization-bias) problem, and any steady-state analysis method should include steps for determining a suitable data truncation point to eliminate bias in the point estimator that is due to the simulation's initial condition.

A number of methods for steady-state analysis have been developed that are based on a single simulation replication (so that $q = 1$). One class of these single-run methods includes procedures that are based on the use of nonoverlapping batch means (NBM), where the sequence of outputs from a single simulation run is divided into adjacent nonoverlapping batches of sufficiently large size so that the resulting batch means are approximately independent and identically distributed (i.i.d.) observations from a normal distribution centered on the steady-state mean μ_X. A CI for μ_X can then be based on the classical Student t-ratio involving the grand average and sample variance of the batch means. Some of the more recent single-run methods that are based on NBM include ASAP3 (Steiger et al. 2005), SBatch (Lada, Steiger, and Wilson 2008), and Skart (Tafazzoli and Wilson 2011). All three of these sequential methods are completely automatable, include steps to determine a suitable truncation point to address the start-up problem, and ultimately return a CI estimator for the steady-state mean μ_X.

Another class of single-run methods for steady-state simulation analysis include those procedures based on a spectral approach, in which we seek to estimate the power spectrum of a given output process as well as the steady-state mean. The resulting estimator of the power spectrum at zero frequency provides an estimator of the steady-state variance parameter γ_X; and from the latter statistic, a CI for the steady-state mean can be computed. The WASSP algorithm (Lada and Wilson 2006) is a completely automated sequential spectral method that addresses the warm-up problem and returns a CI for the steady-state mean as well as a wavelet-based estimator of the entire power spectrum of a given output process.

Although steady-state analysis methods based on a single simulation replication are convenient and efficient in the sense that data from only one warm-up period must be eliminated, there are typically pronounced stochastic dependencies between successive responses generated within a single simulation run. This phenomenon, sometimes called the correlation problem, complicates the construction of a CI for the steady-state mean in single-run methods because standard statistical methods require i.i.d. normal observations to yield a valid CI. The ASAP3, SBatch, Skart, and WASSP algorithms do incorporate steps to address the correlation problem; and comprehensize experimental results indicate that all four methods are effective in delivering approximately valid CIs for a steady-state mean.

As an alternative to single-run methods for simulation analysis, the replication/deletion method is popular with many pracitioners for the reasons elaborated in Section 9.5.2 of Law (2007). This is especially true when excessive sample sizes are required to attenuate the correlation between successive responses within a single run; moreover, in simulation-generated processes exhibiting long-range dependence, the classical single-run methods are generally inapplicable. In the replication/deletion method of simulation analysis, q independent replications of length n_i are executed, and generally $q \geq 3$. For each run i, the observations $\{X_{i,j} : j = 1, \ldots, n_i\}$ are first analyzed to determine a suitable truncation point w_i for eliminating initialization bias; and then the remaining observations $\{X_{i,j} : j = w_i + 1, \ldots, n_i\}$ beyond the truncation point are used to compute a truncated sample mean $\overline{X}_i = (n_i - w_i)^{-1} \sum_{j=w_i+1}^{n_i} X_{i,j}$ for that replication. The mean and variance of the set of truncated sample means $\{\overline{X}_i : i = 1, \ldots, q\}$ over all replications are then computed and used to construct a CI for the steady-state mean. Because the set of truncated sample means $\{\overline{X}_i\}$ are generated from independent replications of the simulation, they are not correlated; and consequently, methods based on this type of replication/deletion approach do not have to be concerned with the correlation problem.

In this paper we describe a sequential method for steady-state simulation analysis that is based on the replication/deletion approach. Called ARD, our automated replication/deletion procedure includes steps for eliminating initialization bias that are based on the warm-up algorithm first used in WASSP; and ARD returns a skewness-adjusted CI for the steady-state mean, similar to that in Skart. Based on our experimentation with a variety of test processes, we have concluded that ARD performs reasonably well in terms of delivering CIs that conform closely to the desired coverage probability and level of precision. However, we also found during the development of ARD that applying a replication/deletion approach in practice is complicated and not nearly as straightforward as is often implied in the simulation literature. Moreover, the total sample sizes required by ARD are larger than those require by efficient single-run

procedures such as Skart. We believe that much of ARD's sampling inefficiency is due to the warm-up period that must be deleted on each replication of the simulation.

The rest of this paper is organized as follows. In Section 2 we provide a high-level overview of ARD, and in Section 3 we explain the steps of ARD in more detail. Section 4 contains a summary of the results of applying ARD to test problems that are specifically designed to provide a "stress test" for the procedure. Conclusions and recommendations for follow-up work are given in Section 5.

2 OVERVIEW OF ARD

Figure 1 depicts a high-level flowchart of the ARD algorithm. The following user-supplied inputs are required:

1. The desired CI coverage probability $1-\alpha$, where $0 < \alpha < 1$; and
2. An absolute or relative precision requirement specifying the final CI half-length in terms of a maximum acceptable half-length h^* (for an absolute precision requirement) or a maximum acceptable fraction r^* of the magnitude of the CI midpoint (for a relative precision requirement).

To construct point and CI estimators for μ_X, ARD begins by performing $q_0 = 8$ pilot runs (replications) of the simulation each with a run length of 2048 observations. For each run i, the simulation-generated output process of size $n_i = 2048$ observations is first divided into $k = 256$ adjacent batches of size $m_i = 8$, with a spacer of initial size $s_i = 0$ observations preceding each batch. The randomness test of von Neumann (1941) is then applied to the initial set of batch means calculated for each batch within run i. The primary purpose of the randomness test is to determine an appropriate data-truncation point for each replication in the pilot study beyond which all computed batch means are approximately independent of the simulation model's initial conditions.

Each time the randomness test is failed for replication i, an additional batch is added to each spacer (up to a limit of 3 batches), and the randomness test is reperformed on the new reduced set of spaced batch means (SBMs). If the randomness test is failed with spacers each consisting of 3 batches so that only 64 spaced batch means are used in the randomness test, then the spacer size s_i is reset to zero and both the batch size m_i and the total sample size n_i are increased by the factor $\sqrt{2}$; the required additional observations are obtained; the augmented sample is rebatched into $k = 256$ nonspaced batches of the new batch size m_i; and a new set of k batch means is computed and tested for randomness with a spacer of size $s_i = 0$. This process of testing the SBMs for randomness with increasing sizes for s_i or m_i continues until the randomness test is finally passed so that acceptable values for the spacer size s_i and the batch size m_i have finally been determined for run i.

Once the randomness test has been passed for all q_0 replications, we compute the maximum final spacer size s^*, the maximum final sample size n^*, and the maximum final batch size m^* over all q_0 pilot runs. These values are then used to compute the final number of spaced batch means k^*. The q_0 pilot runs are then re-executed using the run length n^*. For each run i (where $1 \le i \le q_0$), the first $w^* = s^*$ observations $\{X_{i,j} : j = 1, \ldots, w^*\}$ are discarded to eliminate any warm-up effects and the truncated sample mean $\overline{X}_i$ is computed using the remaining observations $\{X_{i,j} : j = w^* + 1, \ldots n^*\}$,

$$\overline{X}_i = \frac{1}{n^* - w^*} \sum_{u=w^*+1}^{n^*} X_{i,u} \text{ for } i = 1, \ldots, q_0. \tag{1}$$

A point estimator $\overline{\overline{X}}$ of μ_X is computed as the grand mean of the truncated sample means over all q_0 runs in the pilot study,

$$\overline{\overline{X}} = \frac{1}{q_0} \sum_{i=1}^{q_0} \overline{X}_i. \tag{2}$$

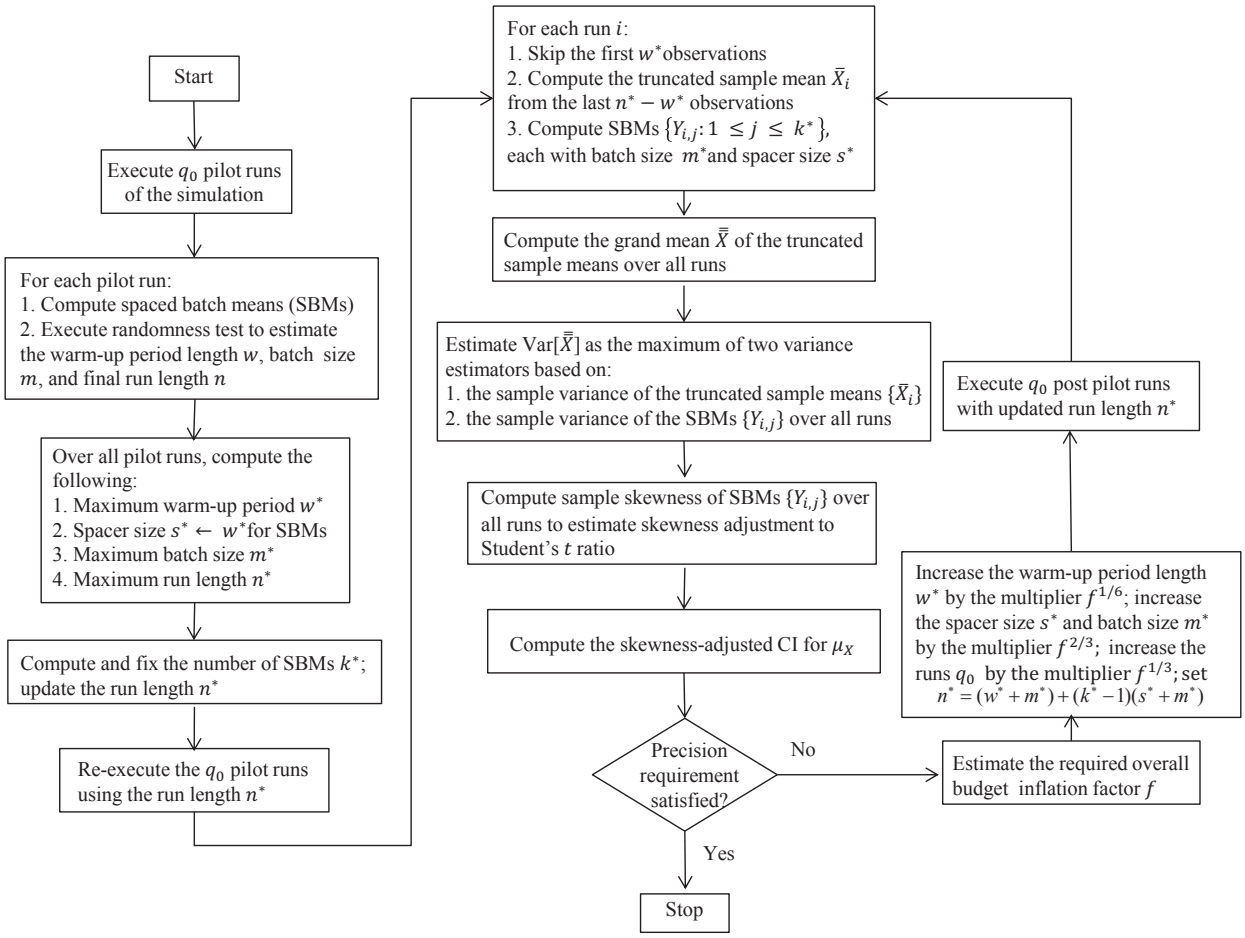

Figure 1: Flow chart of ARD

In terms of the sample variance of the truncated sample means,

$$S^2_{\bar{X}} = \frac{1}{q_0 - 1} \sum_{i=1}^{q_0} \left(\bar{X}_i - \bar{\bar{X}} \right)^2, \tag{3}$$

we see that the "ordinary" (conventional) replication-deletion estimator of the standard error of $\bar{\bar{X}}$ has the form

$$\widehat{\mathrm{SE}}_{\mathrm{ORD}}\left[\bar{\bar{X}}\right] = \frac{S_{\bar{X}}}{\sqrt{q_0}}. \tag{4}$$

ARD also makes extensive use of the SBMs $\{Y_{i,j} : j = 1, \ldots k^*\}$ with warm-up period w^*, spacer size s^*, and batch size m^* for each run in the pilot study,

$$Y_{i,j} = \frac{1}{m^*} \sum_{u=1}^{m^*} X_{i,(j-1)(s^*+m^*)+w^*+u} \text{ for } j = 1, \ldots, k^*. \tag{5}$$

(Note that we distinguish the warm-up period length w^* from the spacer size s^* because these two quantities will grow at different rates with progressively more stringent precision requirements on the final CI estimator

for μ_X as specified in Equation (15) below.) In terms of the grand average $\overline{Y}$ and the sample variance S_Y^2 of the SBMs $\{Y_{i,j}\}$ over all q_0 runs,

$$\overline{Y} = \frac{1}{q_0 k^*} \sum_{i=1}^{q_0} \sum_{j=1}^{k^*} Y_{i,j} \text{ and} \tag{6}$$

$$S_Y^2 = \frac{1}{q_0 k^* - 1} \sum_{i=1}^{q_0} \sum_{j=1}^{k^*} \left(Y_{i,j} - \overline{Y} \right)^2, \tag{7}$$

we see that the SBM-based estimator of the steady-state variance parameter is $\widehat{\gamma}_X = m^* S_Y^2$ and the SBM-based estimator of the standard error of $\overline{\overline{X}}$ has the form

$$\widehat{\mathrm{SE}}_{\mathrm{SBM}}\left[\overline{\overline{X}}\right] = \sqrt{\frac{m^* S_Y^2}{q_0 (n^* - w^*)}}. \tag{8}$$

To compute a CI estimator for μ_X, ARD estimates the standard error of $\overline{\overline{X}}$ according to

$$\widehat{\mathrm{SE}}\left[\overline{\overline{X}}\right] = \max\left\{ \widehat{\mathrm{SE}}_{\mathrm{ORD}}\left[\overline{\overline{X}}\right],\ \widehat{\mathrm{SE}}_{\mathrm{SBM}}\left[\overline{\overline{X}}\right] \right\}. \tag{9}$$

Then ARD computes the approximate $100(1-\alpha)\%$ skewness-adjusted CI for μ_X as follows,

$$\left[\overline{\overline{X}} - G(t_{1-\alpha/2,q_0-1}) \widehat{\mathrm{SE}}\left[\overline{\overline{X}}\right],\ \overline{\overline{X}} - G(t_{\alpha/2,q_0-1}) \widehat{\mathrm{SE}}\left[\overline{\overline{X}}\right] \right], \tag{10}$$

where the skewness adjustment $G(\cdot)$ is defined in terms of the function

$$G(\zeta) \equiv \left\{ \begin{array}{ll} \left[\sqrt[3]{1 + 6\beta(\zeta - \beta)} - 1 \right] / (2\beta), & \text{if } \beta \neq 0, \\ \zeta, & \text{if } \beta = 0, \end{array} \right\} \text{ with } \beta = \frac{\widehat{\mathscr{B}}_Y}{6\sqrt{q_0 k^*}}, \tag{11}$$

and $\widehat{\mathscr{B}}_Y$ is the sample skewness of the SBMs taken over all q_0 pilot runs. A further explanation of equations (8)–(11) is given in Section 3.

If the user has not specified a precision requirement to be satisfied by the final CI estimator of μ_X, then ARD delivers the CI (10) and stops. Otherwise, the CI (10) is tested to determine if it satisfies the user-specified absolute or relative precision requirement. Between the left- and right-hand subintervals of (10) for which $\overline{\overline{X}}$ is their respective upper and lower endpoint, let $\widetilde{H}$ denote the larger subinterval length. If the user has specified a relative precision requirement r^*, then we set

$$H^* = r^* \left| \overline{\overline{X}} \right|; \tag{12}$$

and if the user has specified an absolute precision requirement h^*, then we set

$$H^* = h^*. \tag{13}$$

If $\widetilde{H} \leq H^*$, then ARD delivers the CI (10) and stops; otherwise ARD computes the overall budget-inflation factor

$$f = \mathrm{mid}\left\{ 1.05, \left(\widetilde{H} / H^* \right)^2, 2 \right\} \tag{14}$$

required to satisfy the precision requirement. Then the warm-up period length w^*, the spacer size s^*, the batch size m^*, and the run length n^* are increased as follows:

$$\left.\begin{aligned} w^* &\leftarrow \lceil f^{1/6} w^* \rceil \\ s^* &\leftarrow \lceil f^{2/3} s^* \rceil \\ m^* &\leftarrow \lceil f^{2/3} m^* \rceil \\ n^* &\leftarrow (w^* + m^*) + (k^* - 1)(s^* + m^*) \end{aligned}\right\}; \tag{15}$$

and the number of runs is increased according to

$$q_0 \leftarrow \lceil f^{1/3} q_0 \rceil. \tag{16}$$

A postpilot experiment is executed consisting of q_0 runs, each with updated run length n^*. The CI (10) is recomputed and the precision requirement is retested. Successive iterations of Equations (1)–(16) are repeated until the precision requirement is finally satisfied and the final CI of the form (10) can be delivered.

3 DETAILED OPERATIONAL STEPS OF ARD

3.1 Eliminating Initialization Bias

As described briefly in Section 2, ARD's procedure for identifying the end of the warm-up period is based on a pilot study consisting of $q_0 = 8$ independent runs each of length 2048 observations. For $i = 1, \ldots, q_0$, run i is initially subdivided into $k = 256$ adjacent (nonspaced) batches of size $m_i = 8$. Using initial spacers of size $s_i = 0$, we computed the "spaced" batch means

$$Y_{i,j} = \frac{1}{m_i} \sum_{u=(j-1)(s_i+m_i)+1}^{jm_i} X_{i,u} \quad \text{for} \quad j = 1, \ldots, k. \tag{17}$$

If the SBMs with the current spacer size s_i pass the von Neumann randomness test at the level of significance $\alpha_{\text{ran}} = 0.25$, then we set the warm-up length $w_i = s_i$ and proceed to the computation of the skewness-adjusted CI. Otherwise, we update the spacer size and number of (remaining) batches of size m_i according to

$$s_i = s_i + m_i \quad \text{and} \quad k = \left\lfloor \frac{n_i}{m_i + s_i} \right\rfloor. \tag{18}$$

If $k \geq 64$ (so that $s_i \leq 3m_i$), then the current set of SBMs is tested for randomness; and if the randomness test is passed, then we set $w_i = s_i$ as the length of the warm-up period on run i. Otherwise, the SBMs are updated according to (18) and retested for randomness until the randomness test is passed or $k < 64$. If the condition $k < 64$ is satisfied, then the batch size m_i, the batch count k, the overall sample size n_i, and the spacer size s_i are updated according to

$$m_i \leftarrow \lfloor m_i \sqrt{2} \rfloor, \quad k \leftarrow 256, \quad n_i \leftarrow k m_i, \text{ and } \quad s_i \leftarrow 0, \tag{19}$$

respectively; the required additional observations are obtained for run i (by restarting the simulation if necessary) to complete the overall sample $\{X_{i,j} : j = 1, \ldots, n_i\}$ and then k adjacent (nonspaced) batch means are computed from the overall sample according to (17).

If the step (19) is reached, then ARD reperforms the entire randomness-testing procedure, starting with the current set of $k = 256$ adjacent batch means of the current batch size m_i with spacer size $s_i = 0$. The steps outlined in the preceding two paragraphs are repeated until the randomness test is passed for run i. Once the randomness test is passed, we set the warm-up truncation length for run i to the final spacer size,

$w_i = s_i$. The entire randomness test is then repeated for the remaining runs in the pilot study to obtain a final warm-up length w_i, a final sample size n_i, and a final batch size m_i for each run i where $i = 1, \ldots, q_0$.

This approach to handling the simulation start-up problem is very similar to the approach first used in WASSP and later in SBatch and Skart, all of which are single-run methods for computing point and CI-estimators for the steady-state mean. Through extensive experimentation with the WASSP, SBatch, and Skart algorithms, we found this approach to be effective in determining an appropriate spacer size s_i so that the observations $\{X_{i,j} : j = 1, \ldots, s_i\}$ constituting the first spacer in run i can be regarded as containing the warm-up period because the spaced batch means beyond the first spacer do not exhibit significant departures from randomness—that is, they do not exhibit a deterministic trend or any type of stochastic dependence on the simulation's initial conditions.

There are a few key differences between the original method used in WASSP and the method used in ARD for eliminating initialization bias. WASSP requires an initial sample of size $n = 4096$ observations, an initial batch size of $m = 16$ observations, a significance level $\alpha_{\text{ran}} = 0.2$, and allows a maximum of 9 batches per spacer, resulting in a final spaced batch count in the range $25 \leq k' \leq 256$. By contrast, ARD requires an initial sample of size $n_i = 2048$ observations, an initial batch size of $m_i = 8$ observations, a significance level of $\alpha_{\text{ran}} = 0.25$, and allows up to 3 batches per spacer, resulting in a final spaced batch count in the range $64 \leq k' \leq 256$. The choice to start the ARD algorithm with $n_i = 2048$ observations for each run is designed so that the initial total sample size required for the algorithm is reasonable. As a result of decreasing the initial sample size, it is therefore also necessary to decrease the starting batch size used in ARD so that the initial randomness test has 256 observations (batch means). Furthermore, increasing the significance level α_{ran} from 0.2 to 0.25 and the minimum batch count from $k = 25$ to $k = 64$ are both choices designed to increase the sensitivity of the randomness test, which is critical to the effectiveness of ARD since the algorithm uses multiple runs, each of which contains initialization bias. The WASSP algorithm, however, is less sensitive to initialization bias since it is a single run method and as the run length increases, the effect of the warm-up bias on the final point estimate of the steady-state mean decreases.

3.2 Computing the Skewness-Adjusted Confidence Interval

Once the randomness test has been passed for each replication $i : i = 1, \ldots, q_0$ in the pilot phase, the ARD algorithm proceeds by computing the following maximum values over all q_0 pilot runs,

$$w^* = \max\{w_i : 1 \leq i \leq q_0\}, \quad m^* = \max\{m_i : 1 \leq i \leq q_0\}, \quad n^* = \max\{n_i : 1 \leq i \leq q_0\}. \tag{20}$$

To ensure the warm-up period always consists of at least one batch, we update the final size w^* of the warm-up period and the final spacer size s^* for the pilot study according to the relations

$$w^* \leftarrow \max\{w^*, m^*\} \quad \text{and} \quad s^* \leftarrow w^* \tag{21}$$

so that the final number of spaced batch means k^* and the updated run length n^* are updated according to the relation

$$k^* \leftarrow \left\lceil \frac{n^*}{w^* + m^*} \right\rceil \text{ and } n^* \leftarrow k^*(w^* + m^*). \tag{22}$$

At this point, we reperform the pilot study of $q_0 = 8$ runs having length n^* observations and truncate the first w^* observations from each run so that the truncated sample mean on run i can be computed according to (1). A skewness-adjusted $100(1-\alpha)\%$ CI for μ_X is then computed according to (5)–(11), where the skewness-adjustment (11) is defined in terms of the sample skewness of the spaced batch means,

$$\widehat{\mathscr{B}}_Y = \tau_Y / S_Y^3, \tag{23}$$

where τ_Y is an approximately unbiased estimator of the third central moment of the SBMs taken over all q_0 runs,

$$\tau_Y = \frac{q_0 k^*}{(q_0 k^* - 1)(q_0 k^* - 2)} \sum_{i=1}^{q_0} \sum_{j=1}^{k^*} \left(Y_{i,j} - \overline{Y}\right)^3. \tag{24}$$

The rationale for the skewness adjustment (11) parallels that of the skewness adjustment in the single-run CIs delivered by Skart (Tafazzoli and Wilson 2011). In many applications, the truncated sample means $\{\overline{X}_i : i = 1, \ldots, q_0\}$ exhibit substantial skewness; and this skewness can seriously degrade the coverage of conventional CIs for μ_X based merely on the grand mean (2) and the sample variance (3) of the $\{\overline{X}_i\}$.

ARD's estimator (9) of the standard error of the grand mean (2) is based on our observation that in many applications with relatively loose precision requirements, there is some loss of coverage in skewness-adjusted CIs for μ_X that are based on the conventional standard error estimator (4). Furthermore, this loss of coverage seems to be effectively counteracted by the supplemental use of the SBM-based standard error estimator (8).

3.3 Fulfilling the Precision Requirement

The way in which q_0, w^*, and n^* increase as functions of the budget-inflation factor f is specifically designed to minimize a weighted average of the mean squared errors of $\overline{\overline{X}}$ and $\widehat{\mathrm{SE}}_{\mathrm{ORD}}\left[\overline{\overline{X}}\right]$ as estimators of μ_X and $\sqrt{\gamma_X/[q_0(n^* - w^*)]}$, respectively, where the variance parameter γ_X is defined in the second paragraph of Section 1. This objective ultimately leads to the conclusion that we should take

$$n^* \propto f^{2/3} \quad \text{and} \quad q_0 \propto f^{1/3}. \tag{25}$$

Combining the sampling plan (25) with Theorem 3, part (i) of Glynn and Heidelberger (1991), we took

$$m^* \propto f^{2/3}, \quad s^* \propto f^{2/3}, \quad \text{and} \quad w^* \propto f^{1/6}. \tag{26}$$

Notice that with the sampling plan specified by (25) and (26), the number of SBMs k^* remains fixed on each iteration of ARD.

4 EXPERIMENTAL RESULTS

To evaluate the performance ARD, we selected a suite of test problems that includes some output processes typically used to stress-test steady-state simulation analysis procedures and some output processes whose main characteristics more closely resemble those of real-world applications. Because of space constraints, we limit the present discussion to the four test processes described below; but complete experimental results are provided in Lada, Mokashi, and Wilson (2013). We conducted 1,000 independent replications of ARD for each test process to construct nominal 90% and 95% CIs that satisfied certain precision requirements. For each test process, we include results in the case of no precision requirement, wherein ARD merely delivers the CI computed after the initial pilot study.

The first test process is the queue-waiting-time process for an $M/M/1$ queueing system with a first-in-first-out (FIFO) queueing discipline, an empty-and-idle initial condition, an arrival rate of 0.9, and a service rate of 1.0. In this system the steady-state server utilization is 0.9, and the steady-state expected waiting time in the queue is 9. This process has a relatively short warm-up period; however its autocorrelation function decays slowly with increasing lags. Also, the marginal distribution of the waiting times has a nonzero probability mass at zero and an exponential tail and is therefore markedly nonnormal. These characteristics result in slow convergence to the classical requirement that the truncated sample means $\left\{\overline{X}_i : i = 1, \ldots, q_0\right\}$ are i.i.d. normal random variables. ARD's performance in this test process is summarized in Table 1.

The second test process is the queue-waiting-time process for an $M/M/1$ queueing system with a first-in-first-out (FIFO) queueing discipline, an initial condition of 113 customers in the queue, an arrival rate of 0.9, and a service rate of 1.0. The steady-state expected waiting time in the queue is 9, where waiting times are only accumulated for arrivals after time 0. This system has a pronounced warm-up period and tests ARD's ability to remove severe initialization bias. ARD's performance in this test process is summarized in Table 2.

The third test process, referred to as the Central Server Model 3 (Law and Carson 1979), consists of a central processing unit (CPU or workcenter 1) and $M-1$ peripheral units, referred to as workcenters 2

Table 1: $M/M/1$ FIFO queue-waiting-time process with 90% server utilization, empty-and-idle initial condition, and $\mu_X = 9$.

Rel. Prec. Reqmt.	NONE		7.5%		3.75%	
Nominal CI Cov. Prob.	90%	95%	90%	95%	90%	95%
CI Coverage	88.8%	94.8%	87.8%	94.3%	87.7%	94.1%
Avg. CI Half-length $\widetilde{H}$	0.632466	0.804820	0.531849	0.566327	0.298526	0.301212
Var. CI Half-length	0.044199	0.072068	0.009478	0.005236	0.000695	0.000684
Avg. warm-up length w^*	479	479	489	507	568	606
Avg. # of runs q_0	8	8	9	10	14	16
Avg. sample size/run n^*	47907	47907	52418	60439	96784	124579
Mean Estimator $\overline{\overline{X}}$	8.953485	8.953485	8.958279	8.964682	8.984336	8.988620
Abs. Avg. Bias	0.046515	0.046515	0.041721	0.035318	0.015663	0.011380
Var. of Mean Estimator	0.107834	0.107834	0.089951	0.070791	0.032412	0.021817
Mean Squared Error (MSE)	0.109997	0.109997	0.091691	0.072038	0.032657	0.021947

Table 2: $M/M/1$ FIFO queue-waiting-time process with 90% server utilization, 113 initial customers, and $\mu_X = 9$.

Rel. Prec. Reqmt.	NONE		3.75%		1.875%	
Nominal CI Cov. Prob.	90%	95%	90%	95%	90%	95%
CI Coverage	92.7%	96.6%	89.8%	95.9%	92.7%	96.4%
Avg. CI Half-length $\widetilde{H}$	0.412994	0.527723	0.293640	0.299334	0.154096	0.154918
Var. CI Half-length	0.014386	0.023808	0.001027	0.000760	0.000129	0.000119
Avg. warm-up length w^*	1342	1342	1419	1500	1712	1825
Avg. # of runs q_0	8	8	10	12	16	18
Avg. sample size/run n^*	129084	129084	163779	204376	345311	444507
Mean Estimator $\overline{\overline{X}}$	9.018339	9.018339	9.005641	9.002196	8.998361	8.998238
Abs. Avg. Bias	0.018339	0.018339	0.005641	0.002196	0.001639	0.001762
Var. of Mean Estimator	0.039210	0.039210	0.024694	0.016849	0.006955	0.004710
Mean Squared Error (MSE)	0.039546	0.039546	0.024725	0.016854	0.006957	0.004713

through M. The system has a fixed number of jobs J that cycle through it; and when a job is finished at the CPU, it leaves the system with probability p_1 and is replaced immediately with another job from the CPU queue. If the job does not leave the system, then it is routed to peripheral unit ℓ with probability p_ℓ for $\ell = 2,\ldots,M$. After receiving service at one of the peripheral units, the job leaves the system and is immediately replaced by a job joining the CPU queue. The output process being monitored is each job's *response time*, which is the time between the job's arrival at the CPU queue and its departure from the system. Specifically, we take $J = 8$ jobs and $M = 3$ workcenters with respective service rates 1.0, 0.45, and 0.05. Initially there are five jobs at the CPU, one job at peripheral unit 2, and two jobs at peripheral unit 3. In this system the steady-state utilizations at the three workcenters are 0.44, 0.88, and 0.88, respectively; and the steady-state expected response time for each job is 18.279. Although this test process has a relatively short warm-up period, it exhibits pronounced positive skewness that is often encountered in real-world applications. As a result, this system serves as a challenging test for ARD's skewness-adjusted CIs. ARD's performance in this test process is summarized in Table 3.

The fourth test process is a single-server queueing system with a last-in-first-out (LIFO) queueing discipline, an empty-and-idle initial condition, a mean interarrival time of 1.0, and a mean service time of 0.8. The steady-state server utilization in this system is 0.8, and the steady-state mean queue waiting time is $\mu_X = 3.20$. Although this test process has a relatively short warm-up period, it has highly nonnormal marginals. Moreover, unlike all the other test processes we studied, the queue-waiting-time process in the

Table 3: Central Server 3 Model with $\mu_X = 18.279$.

Rel. Prec. Reqmt.	NONE		1.875%		0.9375%	
Nominal CI Cov. Prob.	90%	95%	90%	95%	90%	95%
CI Coverage	90.2%	95.9%	86.6%	93.3%	89.7%	95.8%
Avg. CI Half-length $\widetilde{H}$	0.624197	0.790914	0.305990	0.307637	0.157993	0.159004
Var. CI Half-length	0.027016	0.043905	0.000649	0.000601	0.000101	0.000078
Avg. warm-up length w^*	22	22	26	28	33	36
Avg. # of runs q_0	8	8	13	15	20	23
Avg. sample size/run n^*	2399	2399	4892	6301	11485	14683
Mean Estimator $\overline{\overline{X}}$	18.202970	18.202970	18.225377	18.239819	18.260944	18.268874
Abs. Avg. Bias	0.076030	0.076030	0.053623	0.039181	0.018056	0.010125
Var. of Mean Estimator	0.104032	0.104032	0.033975	0.023250	0.009081	0.005650
Mean Squared Error (MSE)	0.109812	0.109812	0.036850	0.024785	0.009407	0.005753

$M/M/1$/LIFO queueing system has an autocorrelation function that does not decay geometrically fast with increasing lags. As a consequence of these anomalous properties, some single-run procedures deliver CIs for μ_X with substantially degraded coverage probabilities (Tafazzoli et al. 2011, §4.6). ARD's performance in this test process is summarized in Table 4.

Table 4: $M/M/1$ LIFO queue-waiting-time process with 80% server utilization, empty-and-idle initial condition, and $\mu_X = 3.2$.

Rel. Prec. Reqmt.	NONE		3.75%		1.875%	
Nominal CI Cov. Prob.	90%	95%	90%	95%	90%	95%
CI Coverage	90.6%	94.9%	88.8%	95.0%	89.5%	96.2%
Avg. CI Half-length $\widetilde{H}$	0.636377	0.908514	0.108752	0.109162	0.055814	0.055985
Var. CI Half-length	0.232694	0.529745	0.000136	0.000128	0.000028	0.000026
Avg. warm-up length w^*	26	26	42	45	53	56
Avg. # of runs q_0	8	8	23	26	36	40
Avg. sample size/run n^*	2531	2531	15164	19658	36480	46444
Mean Estimator $\overline{\overline{X}}$	3.188057	3.188057	3.195352	3.198762	3.198850	3.198891
Abs. Avg. Bias	0.011943	0.011943	0.004648	0.001238	0.001150	0.001109
Var. of Mean Estimator	0.067732	0.067732	0.004076	0.002723	0.000998	0.000674
Mean Squared Error (MSE)	0.067875	0.067875	0.004097	0.002724	0.000999	0.000675

From Tables 1–4 we see that for each of the four test processes in the no-precision case, ARD delivered a skewness-adjusted CI with half-length between 2% and 20% of the magnitude of the steady-state mean and the actual CI coverages were reasonably close to the nominal CI coverage probabilities. The no-precision case also demonstrates ARD's ability to determine the length of the warm-up period with a relatively small initial sample size. From the reported point-estimator performance measures (namely, the absolute average bias, variance, and mean squared error of the grand mean $\overline{\overline{X}}$), we concluded that ARD could successfully remove the effects of initialization bias in each of the test processes. In particular, the absolute average bias was in the range of $0.02\% - 0.5\%$ of the steady-state mean for each test process. The warm-up period length (w^*) was $\approx 1\%$ of the average run length (n^*). Relatively larger average run lengths were observed for processes with a substantial initial transient, thus resulting in a larger average overall simulation budget $(q_0 n^*)$.

The precision levels for each test process are chosen so as to stress-test ARD's CI estimator, based on its performance in the no-precision case. Subsequent higher precision levels chosen were half the original chosen levels. This enabled us to observe the effects on the various performance statistics resulting from a 50% reduction in the CI half-length. In particular, it would be desirable to investigate any drop in

CI coverages, or excessive inflation in the overall simulation budget requirement or the warm-up period length resulting from more stringent precision requirements. From Tables 1–4 we observed a gradual increase in q_0 and a more rapid increase in n^* with tighter precision requirements. The actual CI coverages exhibited reasonably close conformance to the nominal coverage probabilities for all the test processes and for all levels of precision that we studied. The point-estimator performance measures showed substantial improvement with progressively tighter precision requirements. In particular, we saw order-of-magnitude reductions in the values of the absolute average bias and the MSE of $\overline{\overline{X}}$ with each successive halving of the precision requirement. In many cases we observed that the total simulation budget required by ARD was substantially larger than the final sample sizes required by some single-run procedures—namely, SBatch (Lada, Steiger, and Wilson 2008) and Skart (Tafazzoli and Wilson 2011; Tafazzoli et al. 2011). We observed that for test processes resulting in highly nonnormal SBMs (namely, the Central Server Model 3 and the M/M/1 LIFO queueing system) ARD recommended running more replications with shorter run-lengths.

5 CONCLUSIONS AND RECOMMENDATIONS

As an approach to automating the replication-deletion method of steady-state simulation analysis, we believe that ARD shows some promise. In particular, ARD delivers point and CI estimators of the steady-state mean that are largely free of the undesirable effects of initialization bias, nonnormality, or correlation in the associated output process. On the other hand, areas for improvement include ARD's sampling efficiency and the ease with which a procedure like ARD can be applied in large-scale practical applications. These issues are the subject of ongoing work.

REFERENCES

Glynn, P. W., and P. Heidelberger. 1991. "Analysis of Initial Transient Deletion for Replicated Steady-State Simulations." *Operations Research Letters* 10: 437–443.

Lada, E. K., A. C. Mokashi, and J. R. Wilson. 2013. "ARD: An Automated Replication-Deletion Method for Simulation Analysis." Technical Report, Edward P. Fitts Department of Industrial and Systems Engineering, North Carolina State University, Raleigh, North Carolina. Available via http://www.ise.ncsu.edu/jwilson/files/ardtr413.pdf [accessed April 13, 2013].

Lada, E. K., N. M. Steiger, and J. R. Wilson. 2008. "SBatch: A Spaced Batch Means Procedure for Steady-State Simulation Analysis." *Journal of Simulation* 2: 170–185.

Lada, E. K., and J. R. Wilson. 2006. "A Wavelet-Based Spectral Procedure for Steady-State Simulation Analysis." *European Journal of Operational Research* 174: 1769–1901.

Law, A. M. 2007. *Simulation Modeling and Analysis*. 4th ed. New York: McGraw-Hill.

Law, A. M., and J. S. Carson. 1979. "A Sequential Procedure for Determining the Length of a Steady-State Simulation." *Operations Research* 27 (5): 1011–1025.

Steiger, N. M., E. K. Lada, J. R. Wilson, J. A. Joines, C. Alexopoulos, and D. Goldsman. 2005. "ASAP3: A Batch Means Procedure for Steady-State Simulation Output Analysis." *ACM Transactions on Modeling and Computer Simulation* 15 (1): 39–73.

Tafazzoli, A., and J. R. Wilson. 2011. "Skart: A Skewness- and Autoregression-Adjusted Batch-Means Procedure for Simulation Analysis." *IIE Transactions* 43 (2):110–128.

Tafazzoli, A., J. R. Wilson, E. K. Lada, and N. M. Steiger. 2011. "Performance of Skart: A Skewness- and Autoregression-Adjusted Batch-Means Procedure for Simulation Analysis." *INFORMS Journal on Computing* 23 (2): 297–314.

von Neumann, J. 1941. "Distribution of the Ratio of the Mean Square Successive Difference to the Variance." *The Annals of Mathematical Statistics* 12 (4): 367–395.

AUTHOR BIOGRAPHIES

EMILY K. LADA is a senior operations research specialist and team lead for SAS Simulation Studio at the SAS Institute. She is a member of INFORMS, and her e-mail address is Emily.Lada@sas.com.

ANUP C. MOKASHI is an operations research development tester for SAS Simulation Studio at the SAS Institute. He is a member of IIE and INFORMS. His e-mail address is Anup.Mokashi@sas.com.

JAMES R. WILSON is a professor in the Edward P. Fitts Department of Industrial and Systems Engineering at North Carolina State University. He is a member of ACM, ASA, ASEE, and SCS; and he is a Fellow of IIE and INFORMS. His e-mail address is jwilson@ncsu.edu, and his web address is http://www.ise.ncsu.edu/jwilson.