Online Supplement to "Performance of a Wavelet-based Spectral Procedure for Steady-state Simulation Analysis"

Emily K. Lada
Statistical and Applied Mathematical Sciences Institute, Research Triangle Park, North Carolina, 27709-4006, USA, eklada@eos.ncsu.edu

James R. Wilson
Department of Industrial Engineering, North Carolina State University, Campus Box 7906, Raleigh, North Carolina 27695-7906, USA, jwilson@ncsu.edu

Natalie M. Steiger
Maine Business School, University of Maine, Orono, ME 04469-5723, USA, nsteiger@maine.edu

Jeffrey A. Joines
Department of Textile Engineering, Chemistry, & Science, North Carolina State University, Raleigh, North Carolina 27695-7913, USA, jeffjoines@ncsu.edu

We summarize the results of an experimental performance evaluation of $\mathcal{WASSP}$, an automated wavelet-based spectral method for constructing an approximate confidence interval on the steady-state mean of a simulation output process so that the delivered confidence interval satisfies user-specified requirements on absolute or relative precision as well as coverage probability. We selected three difficult test problems whose output processes exhibit long warm-up periods, persistent autocorrelation structures, and highly nonnormal marginal distributions; and we used these problems to compare the performance of $\mathcal{WASSP}$ with that of the Heidelberger-Welch algorithm and ASAP3, two sequential procedures based respectively on the methods of spectral analysis and nonoverlapping batch means. Concerning efficiency (required sample sizes) and robustness against the statistical anomalies commonly encountered in simulation studies, $\mathcal{WASSP}$ outperformed the Heidelberger-Welch procedure and compared favorably with ASAP3.

Key words: simulation, statistical analysis; spectral analysis; steady-state analysis; wavelet analysis

---

1. Introduction

A nonterminating simulation is one in which we are interested in long-run (steady-state) average performance measures. Usually in a nonterminating probabilistic simulation, we seek to compute point and confidence-interval estimators for some parameter, or characteristic, of the steady-state cumulative distribution function (c.d.f.) of a particular simulation-generated response. In Lada and Wilson (2004), we develop an automated wavelet-based spectral method for constructing an approximate confidence interval (CI) on the steady-state mean of a simulation output process. This

procedure, called $\mathcal{WASSP}$, determines first a batch size and a truncation point (end of the warm-up period, statistics clearing time) beyond which successive batch means form an approximately stationary Gaussian process. For this purpose we use the randomness test of von Neumann (1941) to determine the size of the spacer (number of ignored observations) preceding each batch that is sufficiently large to ensure the resulting spaced batch means are approximately independent and identically distributed (i.i.d.). We then conclude that the spacer preceding the first batch must contain the warm-up period and hence defines the truncation point; and we use the univariate normality test of Shapiro and Wilk (1965) to determine a batch size that is sufficiently large to ensure the spaced batch means are approximately normal.

Next $\mathcal{WASSP}$ computes the discrete wavelet transform of the bias-corrected log-smoothed-periodogram of the truncated, nonspaced batch means (that is, the batch means which are computed from adjacent nonoverlapping batches observed beyond the truncation point); and the resulting wavelet coefficients are denoised by applying a soft-thresholding scheme. Then by computing the inverse discrete wavelet transform of the thresholded wavelet coefficients, $\mathcal{WASSP}$ delivers an estimator of the batch means log-spectrum and ultimately the steady-state variance constant (SSVC) of the original (unbatched) process—that is, the sum of the covariances at all lags for the original process. Finally $\mathcal{WASSP}$ combines the estimator of the SSVC with the grand average of the truncated batch means in a sequential procedure for constructing a CI estimator of the steady-state mean that satisfies user-specified requirements on absolute or relative precision as well as coverage probability.

In this paper, we summarize some experimental results that are representative of the performance we observed in applying $\mathcal{WASSP}$ and other selected procedures for steady-state simulation output analysis to a suite of particularly difficult test problems. We designed these test problems to explore the following characteristics of the selected output analysis procedures:

- the efficiency of each procedure in terms of the sample size required to deliver a CI that is supposed to attain the user-specified levels of precision and coverage probability; and
- the robustness of each procedure against the statistical anomalies such as initialization bias, correlation, and nonnormality that are commonly encountered in the analysis of outputs generated by large-scale steady-state simulation experiments.

Our experimental performance evaluation is focused on the following test problems:

1. the $M/M/1$ queue waiting time process for which the underlying system has an arrival rate of 0.90, a service rate of 1, and an empty-and-idle initial condition;

2. the first-order autoregressive (AR(1)) process that has a lag-one correlation of 0.995, a white noise variance of 1, a steady-state mean of 100, and an initial value of 0; and

3. the "AR(1)-to-Pareto" (ARTOP) process that has marginals given by a Pareto distribution with lower limit and shape parameter equal to 1 and 2.1, respectively (implying the marginal mean and variance are both finite while the marginal skewness and kurtosis are both infinite), and that is obtained by applying to a standardized version of process 2 above the composite of (*a*) the inverse of the specified Pareto c.d.f., and (*b*) the standard normal c.d.f.

For each of the above test problems, we used the following measures to evaluate the performance of WASSP and its competitors: (i) the empirical coverage probability of the delivered CIs; (ii) the mean and variance of the half-lengths of the delivered CIs; and (iii) the mean and maximum of the total required sample sizes. We performed independent replications of each simulation analysis procedure to construct nominal 90% and 95% CIs that satisfy a given precision requirement, specified as either a maximum percentage of the magnitude of the final point estimator (for a relative precision requirement), or as a maximum absolute half-length (for an absolute precision requirement). We used the following three precision requirements:

- no precision—that is, there was no upper bound on the CI half-length so the final CI delivered by WASSP was based on the batch count and batch size required to pass the randomness and normality tests;

- $\pm$ 15% precision—that is, the half-length of the final CI delivered by WASSP was less than or equal to 15% of the magnitude of the midpoint of the CI; and

- $\pm$ 7.5% precision—that is, the half-length of the final CI delivered by WASSP was less than or equal to 7.5% of the magnitude of the midpoint of the CI.

For each test problem, the theoretical steady-state mean response is available analytically; thus we were able to evaluate the performance of WASSP and its competitors in terms of actual versus nominal CI coverage probabilities as well as sample sizes and half-lengths for the CIs delivered by each procedure. For comparison, we also applied the spectral method of Heidelberger and Welch (1983) and the batch means procedure ASAP3 (Steiger et al. 2005) to each test problem.

The rest of this paper is organized as follows. In §2 we introduce the notation required for our performance evaluation; and we provide brief summaries of the Heidelberger-Welch (H&W) and ASAP3 procedures. In §3 we describe the test problems 1–3 in more detail; and we also

discuss the results of applying $\mathcal{WASSP}$ and its competitors to these test problems. Finally in §4 we summarize the main findings of this research. Lada (2003) provides a complete discussion of the experimental performance evaluation of $\mathcal{WASSP}$. We present some preliminary results on the formulation and evaluation of $\mathcal{WASSP}$ in Lada et al. (2003, 2004a). A stand-alone Windows-based version of $\mathcal{WASSP}$ and a user's manual are available online via Lada et al. (2004b).

2. Simulation Analysis Methods to Be Compared with $\mathcal{WASSP}$

Although we use the notation and terminology of Lada and Wilson (2004) throughout the paper, in this section we summarize the most frequently used notation for completeness before giving overviews of the H&W method and ASAP3. If $\{X_u : u = 1, \ldots, n\}$ is a covariance stationary simulation output process for which we seek to compute point and CI estimators of the mean $\mu_X = \mathrm{E}[X_u]$, then the covariance at lag ℓ for this process is $\gamma_X(\ell) = \mathrm{E}[(X_u - \mu_X)(X_{u+\ell} - \mu_X)]$ for $\ell = 0, \pm 1, \pm 2, \ldots$ and $u = 1, 2, \ldots$; and the SSVC (or variance parameter) of the process is

$$\gamma_X = \sum_{\ell=-\infty}^{\infty} \gamma_X(\ell), \tag{1}$$

where we assume (1) is absolutely convergent so that γ_X is well defined. Moreover, we let $\overline{X}_j(m) = m^{-1} \sum_{u=(j-1)m+1}^{jm} X_u$ denote the jth batch mean for batches of size m computed from the process $\{X_u : u = 1, \ldots, n\}$ for $j = 1, \ldots, k = \lfloor n/m \rfloor$; and we let $\overline{\overline{X}} = \overline{\overline{X}}(m, k) = k^{-1} \sum_{j=1}^{k} \overline{X}_j(m)$ denote the grand mean computed over all k batches of size m.

If the process $\{X_u\}$ is covariance stationary, then the power spectrum $p_X(\omega)$ of this process is given by the cosine transform of the covariance function $\gamma_X(\ell)$,

$$p_X(\omega) = \sum_{\ell=-\infty}^{\infty} \gamma_X(\ell) \cos(2\pi\omega\ell) \quad \text{for} \quad -\tfrac{1}{2} \le \omega \le \tfrac{1}{2}. \tag{2}$$

At frequency $\omega = 0$, we have $p_X(0) = \sum_{\ell=-\infty}^{\infty} \gamma_X(\ell) = \gamma_X$. In using a spectral method to analyze the time series $\{X_u : u = 1, \ldots, n\}$ of length n, first we compute the periodogram

$$I\left(\tfrac{\ell}{n}\right) = \frac{1}{n}\left\{\left[\sum_{u=1}^{n} X_u \cos\left(\tfrac{2\pi(u-1)\ell}{n}\right)\right]^2 + \left[\sum_{u=1}^{n} X_u \sin\left(\tfrac{2\pi(u-1)\ell}{n}\right)\right]^2\right\} \quad \text{for} \quad \ell = 1, \ldots, n-1 \tag{3}$$

as an estimator of $p_X\left(\tfrac{\ell}{n}\right)$ at the Fourier frequency $\tfrac{\ell}{n}$ cycles per time unit for $\ell = 1, \ldots, n-1$; and then by an appropriate extrapolation of (3) to zero frequency, we obtain an estimator of $p_X(0)$.

Both the H&W procedure and $\mathcal{WASSP}$ are designed to deliver a spectral estimator $\widehat{\gamma}_X$ of $\gamma_X = p_X(0)$ that can be used effectively in constructing a $100(1-\beta)\%$ CI for μ_X having the form

$$\overline{\overline{X}} \pm H, \quad \text{with half-length} \quad H = t_{1-\beta/2,\nu} \sqrt{\widehat{\gamma}_X / n'}, \tag{4}$$

where: (i) n' is the length of the truncated output process after deleting (if necessary) a warm-up period containing initialization bias; (ii) the grand mean $\overline{\overline{X}}$ and the SSVC estimator $\widehat{\gamma}_X$ are computed from the truncated output process; (iii) ν denotes the "effective" degrees of freedom (d.f.) associated with $\widehat{\gamma}_X$; and (iv) $t_{1-\beta/2,\nu}$ denotes the $1-\beta/2$ quantile of Student's t-distribution with ν d.f., provided $0 < \beta < 1$.

In $\mathcal{WASSP}$ the user may specify that the CI (4) must satisfy a precision requirement expressed in terms of either of the following quantities:

- a maximum acceptable half-length H^* (for an absolute precision requirement) so that if the latest CI (4) computed by $\mathcal{WASSP}$ satisfies the stopping rule

$$H \le H^*, \tag{5}$$

  then $\mathcal{WASSP}$ delivers (4) as the final CI estimator for μ_X and terminates; or

- a maximum acceptable fraction r^* of the magnitude of the CI midpoint (for a relative precision requirement) so that if the latest CI (4) computed by $\mathcal{WASSP}$ satisfies the stopping rule

$$H \le r^* \left| \overline{\overline{X}} \right|, \tag{6}$$

  then $\mathcal{WASSP}$ delivers (4) as the final CI estimator for μ_X and terminates.

Because the H&W procedure was apparently designed only for use with a relative precision specification, in this paper we base all our experiments on a stopping rule of the form (6).

2.1. Overview of Heidelberger and Welch's Spectral Method

Heidelberger and Welch (1981a, 1981b, 1983) develop a spectral method for steady-state simulation analysis in which they use standard regression techniques to estimate the power spectrum (2) of the given output process at zero frequency. Heidelberger and Welch estimate γ_X by fitting a quadratic polynomial to the logarithm of a smoothed version of the periodogram (3) for the given output process over the frequency range between 0 and $\frac{1}{2}$ cycles per time unit (excluding the endpoints),

where the smoothing operation consists of averaging nonoverlapping pairs of periodogram values. The resulting SSVC estimator is then used to compute a CI of the form (4) for μ_X.

Comparing the performance of $\mathcal{WASSP}$ with that of the H&W procedure is complicated since the latter procedure requires the user to specify an upper limit $t_{\max}$ on the allowable length of a given test process. To make a fair comparison, first we apply $\mathcal{WASSP}$ to the test process so as to obtain not only the corresponding $\mathcal{WASSP}$-generated CI of the form (4) but also a complete (untruncated) time series $\{X_u : n = 1, \ldots, n\}$ to which we can apply the (partially) sequential version of the H&W procedure; and we set $t_{\max} = n$, the length of the simulation-generated time series, for that replication of the H&W procedure.

Heidelberger and Welch (1983) describe a scheme for batching data prior to applying their spectral method, and we employ this scheme in our implementation of the H&W procedure. The batch count k for the H&W procedure is always in the range $L \le k \le 2L$, where we take $L = 200$ as recommended by Heidelberger and Welch (1983). Within each replication of the H&W procedure, we let t_i denote the "time"—that is, the current (untruncated) sample size—at the ith checkpoint in the analysis of a given output process, where $t_1 = \lceil 0.15\, t_{\max} \rceil$ and $t_i = \min\left\{ \lceil 1.5\, t_{i-1} \rceil, t_{\max} \right\}$ for $i = 2, 3, \ldots$. If $t_i \ge L$ and we take $b_i = \lfloor \log_2\{(t_i - 1)/L\} \rfloor$, then at the ith checkpoint the batch size m_i and the number of batches k_i are given by $m_i = 2^{b_i}$ and $k_i = \lfloor t_i/m_i \rfloor$, respectively.

We also employ the method for detecting and eliminating initialization bias described in Heidelberger and Welch (1983). At the ith checkpoint in the H&W procedure (for $i = 1, 2, \ldots$), we test the null hypothesis that the untruncated batch means process $\left\{ \overline{X}_j(m_i) : j = 1, \ldots, k_i \right\}$ is covariance stationary by computing the Cramér–von Mises (CVM) test statistic, $\mathrm{CVM}(m_i, k_i) = \left[\sum_{j=0}^{k_i - 1} D_{ij}^2 \right] \Big/ \left[k_i^2 \widehat{p}_{\overline{X}(m_i)}(0) \right]$, where: (*a*) for each frequency ω in a neighborhood of zero, we let $\widehat{p}_{\overline{X}(m_i)}(\omega)$ denote the H&W estimator of the power spectrum $p_{\overline{X}(m_i)}(\omega)$ of the untruncated batch means process, with $p_{\overline{X}(m_i)}(\omega)$ defined similarly to the power spectrum (2) of the original (unbatched) process; and (*b*) we take $D_{ij} = \sum_{u=1}^{j} \left[\overline{X}_u(m_i) - \overline{\overline{X}}(m_i, k_i) \right]$ for $j = 1, \ldots, k_i$ and $D_{i0} = 0$.

If the untruncated batch means process $\left\{ \overline{X}_j(m_i) : j = 1, \ldots, k_i \right\}$ is covariance stationary, then under widely applicable conditions as we let $m_i \to \infty$ and $k_i \to \infty$, the asymptotic distribution of $\mathrm{CVM}(m_i, k_i)$ is equal to that of $\mathrm{CVM}(\mathcal{B}) = \int_0^1 \mathcal{B}^2(u)\, \mathrm{d}u$, where $\{\mathcal{B}(u) : u \in [0, 1]\}$ is a Brownian bridge process so that the asymptotic 0.9 quantile of the Cramér–von Mises test statistic is $\mathrm{CVM}(\mathcal{B})_{0.9} = 0.3473$; see Table 1 of Anderson and Darling (1952). If $\mathrm{CVM}(m_i, k_i) > 0.3473$, then we conclude that the CVM test has detected nonstationarity (initialization bias) in the untruncated sequence of batch means; and we delete the initial 10% of this sequence and recompute the

CVM test statistic from the truncated sequence of batch means.

After each repetition of the CVM test that detects nonstationarity at the ith checkpoint, the H&W procedure tries to delete an additional 10% of the current untruncated sequence of batch means before repeating the CVM test on the truncated batch means. If the CVM test is failed six times, then the H&W procedure tries to advance to the next checkpoint so that the current (untruncated) sample size is increased by 50% before the batch size, batch count, and untruncated batch means sequence are all updated. The CVM test is repeated at successive checkpoints with warm-up periods (truncation points) ranging from 0% to 50% of the untruncated batch means sequence until either (i) the CVM test is passed and a CI of the form (4) satisfying (6) is computed from the truncated batch means; or (ii) the untruncated sample size required by the H&W procedure reaches the upper limit $t_{\max}$. If case (ii) holds, then the CVM test is performed one last time. If that final CVM test is failed, then the H&W procedure terminates without delivering a CI; otherwise the H&W procedure terminates after delivering a CI of the form (4) that might not satisfy (6). As in Heidelberger and Welch (1981a, 1981b, 1983), we estimate the batch means log-spectrum by fitting a quadratic polynomial to the first 25 points on the log-smoothed-periodogram of the batch means so that in the H&W-generated CI of the form (4), we have $\nu = 7$ d.f.

2.2. Overview of ASAP3

Steiger et al. (2005) formulated ASAP3 as an improved variant of the batch means algorithms ASAP (Steiger and Wilson 2002) and ASAP2 (Steiger et al. 2002) for steady-state simulation analysis. ASAP3 operates as follows: the batch size is progressively increased until spaced groups of four adjacent batch means pass the Shapiro-Wilk test for four-dimensional normality, where the spacer preceding each group also consists of four adjacent batch means; and then after skipping the first spacer as the warm-up period, ASAP3 fits a first-order autoregressive (AR(1)) time series model to the truncated, nonspaced batch means. If necessary, the batch size is further increased until the autoregressive parameter in the AR(1) model does not significantly exceed 0.8. Next ASAP3 computes the terms of an inverse Cornish-Fisher expansion for the classical batch means t-ratio based on the AR(1) parameter estimates; and finally ASAP3 delivers a correlation-adjusted CI based on this expansion. ASAP3 is a sequential procedure designed to deliver a CI that satisfies a user-specified precision requirement of the form (5) or (6).

3. Test Problems Used in the Performance Evaluation

3.1. The $M/M/1$ Queue Waiting Time Process

For the first test problem, we let X_u denote the waiting time in the queue for the uth customer ($u = 1, 2, \ldots$) in a single-server queueing system with i.i.d. exponential interarrival times having mean 10/9 (so that the arrival rate $\lambda = 0.9$); i.i.d. exponential service times having mean 1 (so that the service rate $\mu = 1$); steady-state server utilization $\tau = \lambda/\mu = 0.9$; and an empty-and-idle initial condition (so that $X_1 = 0$). The steady-state mean for this process is $\mu_X = 9.0$.

The selected $M/M/1$ queue waiting time process is a particularly difficult test case for the following reasons. (i) Because the system starts empty and idle, both the magnitude and duration of the initial transient in the process $\{X_u : u = 1, 2, \ldots\}$ are pronounced. (ii) Once the system has reached steady-state operation, the autocorrelation function of the appropriately truncated process $\{X_u\}$ decays very slowly with increasing lags. (iii) The steady-state marginal distribution of waiting times is markedly nonnormal, having an atom at zero and an exponential tail. Hence, this process will also allow us to evaluate the effectiveness of $\mathcal{WASSP}$'s independence and normality tests in determining both an appropriate batch size and an appropriate truncation point beyond which successive batch means approximately constitute a covariance stationary Gaussian process.

If $\{X_u\}$ is in steady-state operation, then the associated power spectrum is given by

$$p_X(\omega) = \frac{\tau^3(2-\tau)}{(1-\tau)^2\mu^2} + \frac{1-\tau^2}{\pi\mu^2}\int_0^r \frac{t^{5/2}(r-t)^{1/2}[\cos(2\pi\omega) - t]}{(1-t)^3[1 - 2t\cos(2\pi\omega) + t^2]}\,\mathrm{d}t \tag{7}$$

for $\omega \in \left[-\frac{1}{2}, \frac{1}{2}\right]$, where $r = 4\tau/(1+\tau)^2$. Since the literature seems to lack readily available computing formulas for $p_X(\omega)$, the result (7) is derived in Appendix A.

Figure 1 displays the log-spectrum $\ln\left[p_X(\omega)\right]$ for $\omega \in \left[-\frac{1}{2}, \frac{1}{2}\right]$. In $\mathcal{WASSP}$, however, we estimate $\ln\left[p_{\overline{X}(m)}(\omega)\right]$, the log-spectrum of the batch means process $\left\{\overline{X}_j(m)\right\}$. From Figure 1, we can get a general idea of the shape of $\ln\left[p_{\overline{X}(m)}(\omega)\right]$ since the peakedness of this function at zero frequency depends on the peakedness of $\ln\left[p_X(\omega)\right]$ at that point. While $\ln\left[p_{\overline{X}(m)}(\omega)\right]$ will be less peaked than $\ln\left[p_X(\omega)\right]$ because of the averaging operation performed on each batch, the batch means log-spectrum will still be sharply peaked; and this characteristic will enable us to gauge the robustness of $\mathcal{WASSP}$'s wavelet-based technique for estimating γ_X.

From the batch means $\left\{\overline{X}_j(m) : j = 1, \ldots, k\right\}$, we compute the associated periodogram $I_{\overline{X}(m)}\left(\frac{\ell}{k}\right)$ at frequency $\frac{\ell}{k}$ (for $\ell = 1, \ldots, k-1$) in the same way that we compute the periodogram (3) from the original (unbatched) process; and in $\mathcal{WASSP}$ the batch means periodogram is smoothed

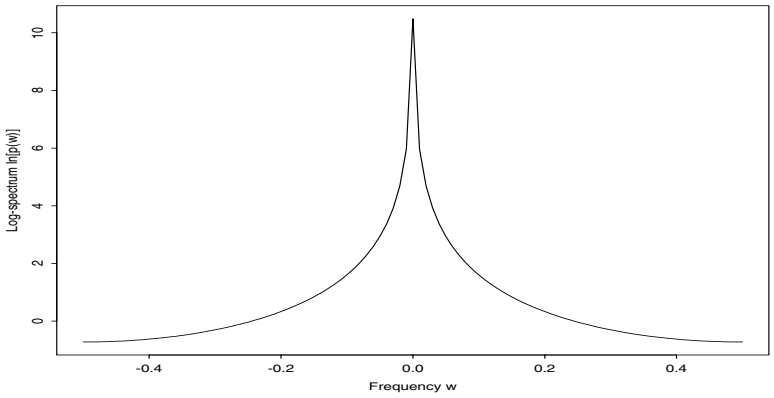


Figure 1: Log-spectrum $\ln[p_X(\omega)]$ of the Steady-state $M/M/1$ Queue Waiting Time Process for Frequency $\omega \in \left[-\frac{1}{2}, \frac{1}{2}\right]$

by computing a moving average of $A = 2a + 1$ points, where $a \in \{2, 3, 4, 5\}$. As explained in §3.3.3 of Lada (2003) and in §4.4.1 of Lada and Wilson (2004), at zero frequency the resulting smoothed periodogram value turns out to be equal to $a^{-1} \sum_{\ell=1}^{a} I_{\overline{X}(m)}\left(\frac{\ell}{k}\right)$. As the overall sample size $n \to \infty$ with a fixed batch size m so that the batch count $k \to \infty$, the smoothed periodogram value at zero frequency is (i) asymptotically independent of the grand average of the batch means; and (ii) approximately a chi-squared random variable with $2a$ d.f. that has been scaled by the multiplier $p_{\overline{X}(m)}(0)\big/(2a)$, where $p_{\overline{X}(m)}(0) = p_X(0)\big/m = \gamma_X\big/m$. Thus $\mathcal{WASSP}$'s $100(1-\beta)\%$ CI for μ_X of the form (4) is based on the $1 - \beta/2$ quantile of Student's t-distribution with $\nu = 2a$ d.f. The user selects the smoothing parameter $A \in \{5, 7, 9, 11\}$, with the default being $A = 7$ so that $\mathcal{WASSP}$'s CIs are based on $\nu = 6$ d.f. by default.

Table 1 shows the performance of $\mathcal{WASSP}$ for the $M/M/1$ queue waiting time process using the smoothing parameter values $A = 5, 7, 9$, and 11. The results are based on 1,000 independent replications of nominal 90% CIs. For each coverage estimator in Table 1, the standard error was less than 1%. From this table, we concluded that the coverage probability decreased in general as A increased. This was due to the target process having a power spectrum with a sharp peak at zero frequency. As A was increased, $\mathcal{WASSP}$'s estimate of the batch means power spectrum near zero frequency became flatter than it should have been; and this behavior ultimately resulted in underestimation of the SSVC. For the no precision case, we found that $A = 5$ yielded the best results in terms of CI coverage probability for each test process used in our performance evaluation. In general, if one were only interested in generating an initial, or pilot, CI for the steady-state mean of a particular process without imposing a precision requirement, then it might be desirable to change the smoothing parameter from the default value $A = 7$ to $A = 5$. Similarly, for the

$\pm 15\%$ precision case, the results for $A = 5$ appeared to be better than those for $A = 7$. However, for the $\pm 7.5\%$ precision case, there was significant CI overcoverage for $A = 5$; and we concluded that for progressively smaller values of r^*, the default value $A = 7$ produced slightly better results than $A = 5$ for this test process. In summary, from Table 1 we concluded that while there might be small differences in the results for the allowable values of A, setting $A = 5, 7, 9$, or 11 yielded acceptable results for this test process.

Table 1: Performance of $\mathcal{WASSP}$ Using Different Values of A in the $M/M/1$ Queue Waiting Time Process Based on 1,000 Replications of 90% CIs

Precision Requirement	Performance Measure	Smoothing Parameter			
		$A = 5$	$A = 7$	$A = 9$	$A = 11$
None	CI coverage	88.8%	87.7%	86.1%	84.2%
	Avg. sample size	18,369	18,090	17,696	18,369
	Max. sample size	318,920	241,152	318,920	318,920
	Avg. CI half-length	3.3957	3.0715	2.9116	2.6684
	Var. CI half-length	2.6495	2.0026	1.6165	1.2476
$\pm 15\%$	CI coverage	89.6%	87.2%	83.5%	82.8%
	Avg. sample size	114,710	92,049	79,824	68,533
	Max. sample size	1,024,096	688,256	694,970	500,934
	Avg. CI half-length	1.1000	1.1103	1.1223	1.1451
	Var. CI half-length	0.0414	0.0387	0.0381	0.0340
$\pm 7.5\%$	CI coverage	93.6%	90.4%	88.5%	91.5%
	Avg. sample size	467,370	388,000	341,380	322,990
	Max. sample size	2,640,173	2,614,458	2,478,192	2,342,976
	Avg. CI half-length	0.5846	0.5866	0.5855	0.5911
	Var. CI half-length	0.0072	0.0072	0.0067	0.0060

3.1.1. Validation of Student's t-Ratio Assumptions for $M/M/1$ Queue Waiting Time Process

The validity of a $\mathcal{WASSP}$-generated CI for μ_X having the form (4) depends on the ratio

$$\mathcal{T} = \frac{\overline{\overline{X}}(m,k) - \mu_X}{\sqrt{\widehat{\gamma}_X / n'}} = \left\{ \left[\overline{\overline{X}}(m,k) - \mu_X \right] \Big/ \sqrt{\frac{\gamma_X}{n'}} \right\} \Big/ \sqrt{\left(\frac{2a\widehat{\gamma}_X}{\gamma_X} \right) \Big/ (2a)} = \mathcal{Z} \Big/ \sqrt{\mathcal{Q}/(2a)} \qquad (8)$$

having Student's t-distribution with $2a$ d.f.; and in general this distributional requirement is met if the following assumptions hold—

Assumption $\mathbf{A_1}$: The numerator of (8) satisfies $\mathcal{Z} = \left[\overline{\overline{X}}(m,k) - \mu_X \right] \Big/ \sqrt{\gamma_X / n'} \sim N(0,1)$, where in general $N(\alpha, \sigma^2)$ denotes a normal distribution with mean α and variance σ^2.

Assumption $\mathbf{A_2}$: The squared denominator of (8) satisfies $\mathcal{Q} = \left(2a\widehat{\gamma}_X\right) \big/ \gamma_X \sim \chi^2(2a)$, where in general $\chi^2(\nu)$ denotes a chi-squared random variable with ν d.f.

Assumption $\mathbf{A_3}$: The numerator $\mathcal{Z}$ and the squared denominator $\mathcal{Q}$ are independent.

In the following experimental study of the probabilistic behavior of $\mathcal{T}$, $\mathcal{Z}$, and $\mathcal{Q}$ when $\mathcal{WASSP}$ was applied to the waiting time process in the $M/M/1$ queue, we used the default value $A = 7$ so that $\widehat{\gamma}_X = m\widehat{p}_{\overline{X}(m)}(0)$, the $\mathcal{WASSP}$-based estimator of γ_X, had 6 d.f. To determine if the $\mathcal{T}$-values generated by $\mathcal{WASSP}$ possessed Student's t-distribution with 6 d.f. to an acceptable degree of approximation, we generated the following plots, each based on 400 independent replications of $\mathcal{WASSP}$ with the CI specifications of no precision, $\pm 15\%$ precision, and $\pm 7.5\%$ precision and with nominal confidence levels of 90% and 95%:

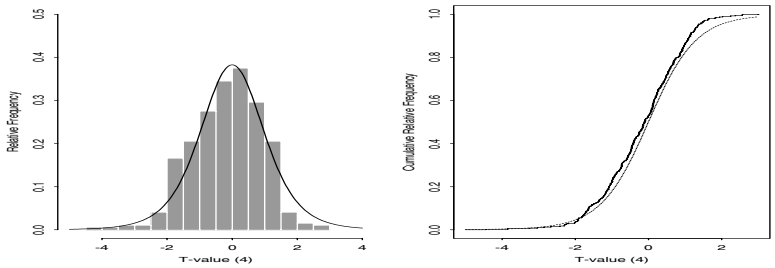
1. A plot showing the histogram of $\mathcal{T}$-values superimposed on the probability density function (p.d.f.) of Student's t-distribution with 6 d.f., together with a plot showing the empirical c.d.f. of the $\mathcal{T}$-values superimposed on the c.d.f. of Student's t-distribution with 6 d.f.

2. A plot showing the histogram of $\mathcal{Z}$-values superimposed on the p.d.f. of the $N(0, 1)$ distribution, together with a plot showing the empirical c.d.f. of the $\mathcal{Z}$-values superimposed on the c.d.f. of the $N(0, 1)$ distribution.

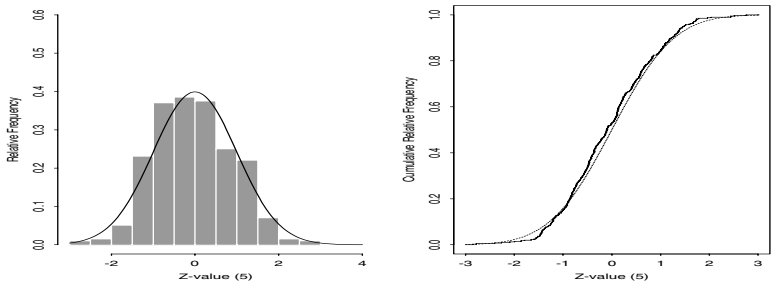
3. A plot showing the histogram of $\mathcal{Q}$-values superimposed on the p.d.f. of the $\chi^2(6)$ distribution, together with a plot showing the empirical c.d.f. of the $\mathcal{Q}$-values superimposed on the c.d.f. of the $\chi^2(6)$ distribution.

Some of the plots described in items 1–3 above are displayed in Figures 2(a)–2(c); see §4.2.1 of Lada (2003) for all the plots describing the results of applying $\mathcal{WASSP}$ to the given $M/M/1$ queue waiting time process. To generate these plots, first we computed the theoretical SSVC γ_X of this process. From Daley (1968) we have $\gamma_X = \left[\tau^3(\tau^3 - 4\tau^2 + 5\tau + 2)\right] / \left[(1 - \tau)^4\lambda^2\right]$; and substituting $\tau = 0.9$ and $\lambda = 0.9$ into this expression, we obtain $\gamma_X = 35{,}901$.

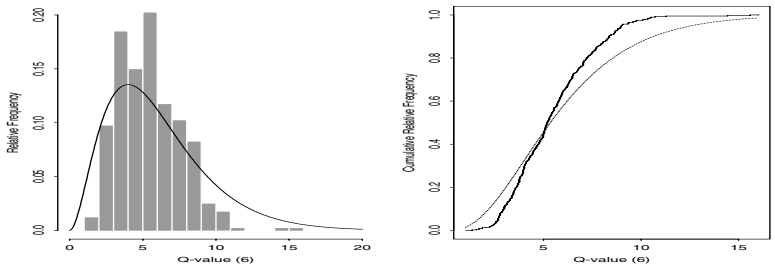
From the plot in Figure 2(b) where the empirical distribution of $\mathcal{Z}$ is compared with the $N(0, 1)$ distribution, we concluded that at the $\pm 7.5\%$ precision level the distribution of $\mathcal{Z}$-values closely followed a $N(0, 1)$ distribution and that any bias effects had been largely eliminated. Moreover we judged that to a fair approximation, the $\mathcal{Q}$-values followed the $\chi^2(6)$ distribution. Since $\mathcal{Q}$ is a random variable based on the sample covariance structure (that is, second-degree sample moments) of the process $\{X_u : u = 1, \ldots, n\}$, we anticipated slower convergence of $\mathcal{Q}$ to its limiting $\chi^2(6)$ distribution as $r^* \to 0$ compared with the rate of convergence of $\mathcal{Z}$ to its limiting $N(0, 1)$ distribution. From Figures 2(b) and 2(c), we concluded that Assumptions A_1 and A_2 appeared to hold (at least to a fair approximation) for the selected cases of the $M/M/1$ queue



(a) Empirical Distribution of $\mathcal{T}$-values and Student's t-Distribution with 6 d.f.



(b) Empirical Distribution of $\mathcal{Z}$-values and the $N(0, 1)$ Distribution



(c) Empirical Distribution of $\mathcal{Q}$-values and the $\chi^2(6)$ Distribution

Figure 2: Empirical Distributions (Step Functions) of $\mathcal{T}$-, $\mathcal{Z}$-, and $\mathcal{Q}$-values and Their Assumed Theoretical Distributions (Smooth Curves) Based on 400 Runs of $\mathcal{WASSP}$ in $M/M/1$ Queue Using 90% CIs with $\pm 7.5\%$ Precision

waiting time process. A final observation is that overall, the $\mathcal{T}$-values closely followed Student's t-distribution with 6 d.f., as can been seen in Figure 2(a).

To verify Assumption A_3, we estimated the correlation between $\overline{\overline{X}}(m, k)$ and $\widehat{\gamma}_X$ for 400 replications of $\mathcal{WASSP}$ with confidence levels of 90% and 95% and with $A = 7$. In Tables 2(a) and 2(b), we see that there was significant correlation between $\overline{\overline{X}}(m, k)$ and $\widehat{\gamma}_X$, suggesting that in this application of $\mathcal{WASSP}$ with the specified levels of r^*, Assumption A_3 did not hold so that $\mathcal{Z}$ and $\mathcal{Q}$ were not independent. For progressively smaller values of r^*, the correlation between $\overline{\overline{X}}(m, k)$ and $\widehat{\gamma}_X$ appeared to decrease very slowly. This may be a partial explanation for the CI undercoverage seen in some of the small-sample cases (that is, in the no precision and the $\pm 15\%$ precision cases)

for this test problem. However, it is unclear that Assumption A_3 is a necessary condition for $\mathcal{T}$ to have Student's t-distribution with 6 d.f.

Table 2: Correlation between $\overline{\overline{X}}(m,k)$ and $\widehat{\gamma}_X$ Based on 400 Runs of $\mathcal{WASSP}$ in the $M/M/1$ Queue Waiting Time Process

(a) Using 90% CIs

Precision Reqt.	Corr$\left[\overline{\overline{X}}(m,k), \widehat{\gamma}_X\right]$
None	0.5829
±15%	0.5778
±7.5%	0.4922

(b) Using 95% CIs

Precision Reqt.	Corr$\left[\overline{\overline{X}}(m,k), \widehat{\gamma}_X\right]$
None	0.6540
±15%	0.5605
±7.5%	0.4531

3.1.2. Summary of Experimental Results for the $M/M/1$ Queue Waiting Time Process

For the $M/M/1$ queue waiting time process, Table 3 shows a comparison of the performance of the following procedures: $\mathcal{WASSP}$ (using $A = 7$); ASAP3; the H&W procedure as specified in Heidelberger and Welch (1983); and the H&W procedure without the Cramér-von Mises test (labeled HW−CV) as originally formulated in Heidelberger and Welch (1981a) with the recommendation to take $L = 100$ in the batching scheme outlined in the third paragraph of §2.1. We included the results for the HW−CV procedure to evaluate the effectiveness of the Cramér-von Mises test for detecting and eliminating initialization bias. Because of the extensive disk space requirements for the ASAP3 software, we were limited to 400 replications of ASAP3; and thus the reported coverage probabilities for ASAP3 had standard errors of approximately 1.5% and 1% for nominal coverages of 90% and 95%, respectively. Because we performed 1,000 replications of the $\mathcal{WASSP}$, H&W, and HW−CV procedures, the coverage probabilities for these three procedures had standard errors of approximately 0.95% and 0.69% for nominal coverages of 90% and 95%, respectively.

Since the H&W and HW−CV procedures could run out of data before satisfying the precision requirement, we included in Table 3 the overall CI coverage (for all 1,000 replications, whether or not the precision requirement (6) was met) as well as the coverage for those CIs that satisfied (6). Although in theory both the H&W and HW−CV procedures could terminate without delivering a CI, in practice we never observed this behavior. From Table 3 we see that in the no precision case, $\mathcal{WASSP}$ and ASAP3 yielded similar results in terms of CI coverage; however, in this case $\mathcal{WASSP}$'s average required sample size was 42% smaller than that of ASAP3.

Table 3 shows that the empirical coverage probabilities of the CIs delivered by the H&W and HW−CV methods were significantly below not only their specified nominal levels but also the empirical coverage probabilities of the CIs delivered by $\mathcal{WASSP}$ and ASAP3. For example in the

Table 3: Performance of $\mathcal{WASSP}$ (Using $A = 7$), ASAP3, H&W, and HW−CV in the $M/M/1$ Queue Waiting Time Process

Prec. Reqt.	Performance Measure	$\mathcal{WASSP}$	H&W	HW−CV	ASAP3
Results for Nominal 90% Confidence Intervals					
None	# replications	1,000	1,000	1,000	400
	CI coverage	87.7%	67.8%	67.9%	87.5%
	Avg. sample size	18,090	2,714	2,756	31,181
	Max. sample size	241,152	36,173	47,838	185,344
	Avg. CI half-length	3.0715	4.0535	3.9994	2.0719
	Var. CI half-length	2.0026	4.4582	4.8842	0.3478
	# replications satisfying prec. reqt.	1,000	1,000	1,000	400
	Coverage of CIs satisfying prec. reqt.	87.7%	67.8%	67.9%	87.5%
±15%	# replications	1,000	1,000	1,000	400
	CI coverage	87.2%	81.3%	79.6%	91%
	Avg. sample size	92,049	62,112	65,282	103,742
	Max. sample size	688,256	348,434	314,464	424,536
	Avg. CI half-length	1.1103	1.1486	1.3154	1.1820
	Var. CI half-length	0.0387	0.0406	0.3765	0.0259
	# replications satisfying prec. reqt.	1,000	939	767	400
	Coverage of CIs satisfying prec. reqt.	87.2%	80.9%	75%	91%
±7.5%	# replications	1,000	1,000	1,000	400
	CI coverage	90.4%	85%	84.1%	89.5%
	Avg. sample size	388,000	275,610	298,860	287,568
	Max. sample size	2,614,458	1,323,572	1,216,420	700,700
	Avg. CI half-length	0.5866	0.5899	0.6852	0.6273
	Var. CI half-length	0.0072	0.0072	0.0599	0.0023
	# replications satisfying prec. reqt.	1,000	918	673	400
	Coverage of CIs satisfying prec. reqt.	90.4%	83.9 %	79.35%	89.5%
Results for Nominal 95% Confidence Intervals					
None	# replications	1,000	1,000	1,000	400
	CI coverage	93.4%	76.2%	77.2%	91.5%
	Avg. sample size	17,971	2,696	2,595	31,181
	Max. sample size	171,456	25,719	47,838	185,344
	Avg. CI half-length	3.9987	5.1817	5.0668	2.5209
	Var. CI half-length	3.6999	7.9996	6.5633	0.5350
	# replications satisfying prec. reqt.	1,000	1,000	1,000	400
	Coverage of CIs satisfying prec. reqt.	93.4%	76.2%	77.2%	91.5%
±15%	# replications	1,000	1,000	1,000	400
	CI coverage	93%	88.6%	87%	95.5%
	Avg. sample size	143,920	98,838	104,290	140,052
	Max. sample size	953,424	482,673	482,306	418,263
	Avg. CI half-length	1.1342	1.1550	1.3521	1.2059
	Var. CI half-length	0.0314	0.0347	0.3791	0.0205
	# replications satisfying prec. reqt.	1,000	944	717	400
	Coverage of CIs satisfying prec. reqt.	93%	88.4%	83.7%	95.5%
±7.5%	# replications	1,000	1,000	1,000	400
	CI coverage	97%	91.8%	92.6%	94%
	Avg. sample size	598,020	431,590	458,310	382,958
	Max. sample size	3,408,016	1,517,616	1,841,421	956,610
	Avg. CI half-length	0.5950	0.5903	0.6910	0.6324
	Var. CI half-length	0.0056	0.0078	0.0523	0.0020
	# replications satisfying prec. reqt.	1,000	917	673	400
	Coverage of CIs satisfying prec. reqt.	97%	91.1 %	89.6%	94%

case of nominal 95% CIs with a required precision of ±7.5%, the H&W procedure delivered 917 CIs of acceptable precision. Since 91.1% of those CIs covered μ_X, only $917 \times 0.911 \approx 835$ out of 1,000 replications of the H&W procedure yielded correct results in this case. Similarly, only

$673 \times 0.896 \approx 603$ out of 1,000 replications of the HW−CV procedure yielded correct results in this case.

To investigate the causes of the poor performance of the H&W and HW−CV procedures relative to $\mathcal{WASSP}$, we estimated the bias, variance, and mean squared error of the final point estimator $\overline{\overline{X}}(m, k)$ delivered by each of these procedures for each selected combination of CI nominal coverage and required precision. (We omitted ASAP3 from this analysis because its performance is thoroughly examined in Steiger et al. 2005; moreover since the other three procedures operated on exactly the same data sets as explained in the second paragraph of §2.1, it was natural to limit our comparison to those procedures). We used the following statistics:

$$\widehat{\text{Bias}}\big[\,\overline{\overline{X}}(m,k)\big] = \left[\frac{1}{R}\sum_{u=1}^{R}\overline{\overline{X}}_u(m_u,k_u)\right] - \mu_X, \quad \widehat{\text{MSE}}\big[\,\overline{\overline{X}}(m,k)\,\big] = \frac{1}{R}\sum_{u=1}^{R}\big[\,\overline{\overline{X}}_u(m_u,k_u) - \mu_X\,\big]^2, \qquad (9)$$

where: (i) we let R denote the number of replications of the procedure with the selected coverage probability that satisfied the given precision requirement; (ii) for the uth such replication ($u = 1, \ldots, R$), we let $\overline{\overline{X}}_u(m_u, k_u)$ denote the associated grand average of the truncated batch means based on k_u batches of size m_u; and (iii) we let $\widehat{\text{Var}}\big[\,\overline{\overline{X}}(m,k)\,\big]$ denote the sample variance of the truncated batch means $\big\{\overline{\overline{X}}_u(m_u, k_u) : u = 1, \ldots, R\big\}$.

Table 4 shows the estimated absolute bias, variance, and mean squared error for the final point estimators delivered by the $\mathcal{WASSP}$, H&W, and HW−CV procedures in the $M/M/1$ queue waiting time process. In the case of no precision requirement, all three performance measures for the final point estimator $\overline{\overline{X}}(m, k)$ delivered by the H&W and HW−CV methods were substantially larger than the corresponding quantities for $\mathcal{WASSP}$. We concluded that in the no precision case, the Cramér–von Mises test failed to yield significant reductions in initialization bias. Moreover, Table 3 shows that in the no precision case, the H&W and HW−CV procedures required much smaller final sample sizes than $\mathcal{WASSP}$ required; and in Table 4, this behavior is reflected in much larger point-estimator variances for the H&W and HW−CV procedures compared with that for $\mathcal{WASSP}$. Essentially, the Cramér–von Mises test was often passed at relatively small values of both the total (untruncated) sample size t_i and the associated truncation point; and as a result, the truncated time series used to construct the delivered CI of the form (4) was neither sufficiently long nor sufficiently free of initialization bias to yield an accurate estimate of μ_X.

From Table 4, we see also that once a precision requirement was imposed on each procedure and the sample size began to increase, the bias, variance, and mean squared error of $\overline{\overline{X}}(m, k)$ began to decrease for all three procedures. Unfortunately Tables 3 and 4 reveal that even with a nontrivial

Table 4: Mean Squared Error, Variance, and Absolute Bias of $\overline{\overline{X}}(m,k)$ Based on 1,000 Runs the $M/M/1$ Queue Waiting Process

Precision Requirement	Performance Measure	Nominal 90% CIs $\mathcal{WASSP}$	H&W	HW−CV	Nominal 95% CIs $\mathcal{WASSP}$	H&W	HW−CV
None	$\widehat{\text{MSE}}\left[\overline{\overline{X}}(m,k)\right]$	2.5834	10.6874	11.4094	2.9743	12.3617	10.9297
	$\widehat{\text{Var}}\left[\overline{\overline{X}}(m,k)\right]$	2.5302	10.3261	11.1897	2.969	12.2787	10.7499
	$\left\lvert\widehat{\text{Bias}}\left[\overline{\overline{X}}(m,k)\right]\right\rvert$	0.2307	0.6011	0.4687	0.0728	0.2881	0.4240
±15%	$\widehat{\text{MSE}}\left[\overline{\overline{X}}(m,k)\right]$	0.6288	0.8504	0.8176	0.4165	0.5502	0.4838
	$\widehat{\text{Var}}\left[\overline{\overline{X}}(m,k)\right]$	0.5694	0.7676	0.7246	0.382	0.5052	0.4555
	$\left\lvert\widehat{\text{Bias}}\left[\overline{\overline{X}}(m,k)\right]\right\rvert$	0.2437	0.2877	0.3050	0.1857	0.2121	0.1682
±7.5%	$\widehat{\text{MSE}}\left[\overline{\overline{X}}(m,k)\right]$	0.1158	0.1871	0.1861	0.0703	0.3144	0.1066
	$\widehat{\text{Var}}\left[\overline{\overline{X}}(m,k)\right]$	0.1125	0.1795	0.1802	0.0694	0.3077	0.1038
	$\left\lvert\widehat{\text{Bias}}\left[\overline{\overline{X}}(m,k)\right]\right\rvert$	0.0574	0.0872	0.0768	0.0300	0.0819	0.0529

precision requirement, incorporating the Cramér–von Mises test did not significantly improve the performance of the H&W procedure in terms of any of the following: CI coverage; CI half-length; and bias, variance, and mean squared error of the final point estimator. There appeared to be some improvement in the stability and precision of the CIs delivered by the H&W procedure relative to the HW−CV procedure, as evidenced by the reduced variance of the CI half-length and the increased number of replications satisfying the precision requirement. Finally from Tables 3 and 4 we concluded that $\mathcal{WASSP}$ outperformed the H&W and HW−CV procedures with respect to point-estimator accuracy and precision as well as CI coverage, precision, and stability.

Although we found that the overall performance of the H&W procedure was only slightly better than that of the HW−CV procedure in some experiments with the $M/M/1$ waiting time process, we recognize that the H&W procedure is referenced and used far more frequently than the HW−CV procedure (Pawlikowski 1990). Thus in the rest of our performance evaluation, we dropped the HW−CV procedure and narrowed our comparison to the H&W procedure, ASAP3, and $\mathcal{WASSP}$.

3.2. The First-order Autoregressive (AR(1)) Process

If $\{\delta_u : u = 1, 2, \ldots\} \overset{\text{i.i.d.}}{\sim} N\left(0, \sigma_\delta^2\right)$ is a white noise process, then a first-order autoregressive (AR(1)) process with the starting value X_0 can be generated as follows,

$$X_u = \mu_X + \rho(X_{u-1} - \mu_X) + \delta_u, \quad \text{for } u = 1, 2, \ldots, \tag{10}$$

where μ_X is the mean and ρ is the lag-one correlation of the process in steady-state operation. We set the mean $\mu_X = 100$, the autoregressive parameter $\rho = 0.995$, and the white noise variance $\sigma_\delta^2 = 1$;

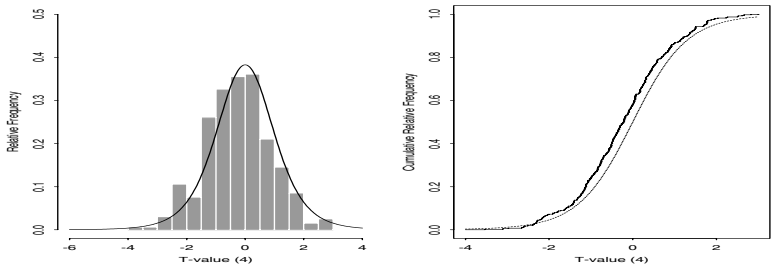
and we set $X_0 = 0$ to obtain the analogue for the AR(1) process of an "empty-and-idle" initial condition for the $M/M/1$ queue. The most difficult aspects of this test process are its exceptionally long initial transient period and its persistent autocorrelation structure. On the other hand, for every batch size the batch means computed from this process are multivariate normal.

The spectrum of the steady-state AR(1) process (10) is $p_X(\omega) = \sigma_\delta^2 \big/ \big[1 - 2\rho\cos(2\pi\omega) + \rho^2\big]$ for $\omega \in \big[-\frac{1}{2}, \frac{1}{2}\big]$; see §4.3 of Lada (2003). The variance of the process is $\sigma_X^2 = \sigma_\delta^2 \big/ \big(1 - \rho^2\big) = 100.25$ while the SSVC is $\gamma_X = p_X(0) = \sigma_\delta^2 \big/ (1-\rho)^2 = 40{,}000$. For $\omega \in \big[-\frac{1}{2}, \frac{1}{2}\big]$, the log-spectrum $\ln\big[p_X(\omega)\big]$ exhibits peakedness at zero frequency similar to that exhibited by its counterpart for the $M/M/1$ waiting time process; and this property resulted in more pronounced underestimation of the SSVC with increasing values of $\mathcal{WASSP}$'s smoothing parameter A. However, as detailed in §4.3 of Lada (2003), for this test process we found that setting $A = 5, 7, 9$, or 11 yielded acceptable results in terms of the coverage probabilities of the CIs delivered by $\mathcal{WASSP}$.

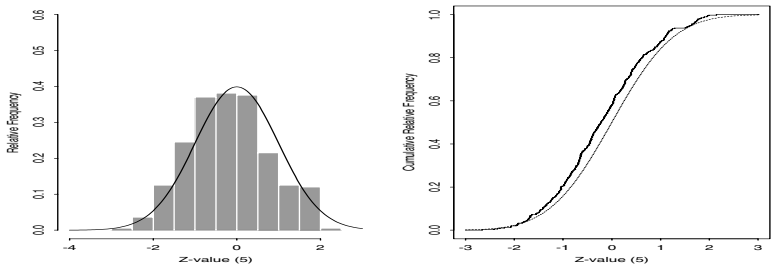
3.2.1. Validation of Student's t-Ratio Assumptions for the AR(1) Process

In this section we summarize the experiments used to validate the assumptions A_1–A_3 on which we based the conclusion that the auxiliary random variable $\mathcal{T}$ defined in (8) would possess Student's t-distribution with $2a = 6$ d.f. when we set $A = 7$ and applied $\mathcal{WASSP}$ to the AR(1) process. As detailed below, $\mathcal{WASSP}$ required nearly the same average sample sizes and generated not only CIs but also $\mathcal{T}$-, $\mathcal{Z}$-, and $\mathcal{Q}$-values with nearly the same performance characteristics at all three previously selected levels of precision—namely, no precision, ±15% precision, and ±7.5% precision. To provide a more comprehensive experimental validation of Assumptions A_1–A_3 for this test process, we also included the case of ±3.75% precision; and the resulting plots for nominal 95% CIs at the ±3.75% precision level are displayed in Figures 3(a)–3(c).

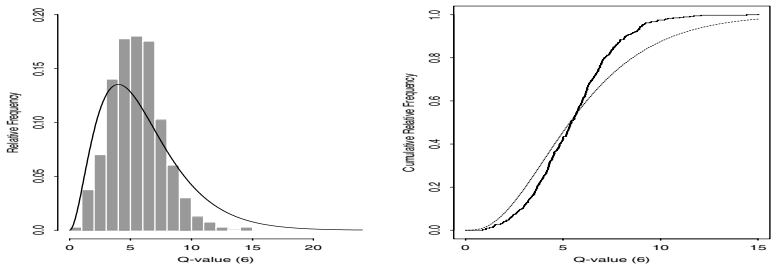
From Figures 3(b)–3(c) and the other plots presented in §4.3.1 of Lada (2003), we concluded that Assumptions A_1 and A_2 were satisfied to a good approximation in the given AR(1) process. Finally, to validate Assumption A_3, we computed the sample correlation between $\overline{\overline{X}}(m, k)$ and $\widehat{\gamma}_X$ for 400 independent replications of $\mathcal{WASSP}$ applied to the given AR(1) process. As shown in Tables 5(a) and 5(b), the correlations between $\overline{\overline{X}}(m, k)$ and $\widehat{\gamma}_X$ were small, indicating approximate validity of Assumption A_3 in the test cases we considered. Finally from Figure 3(a) we concluded that Student's t-distribution with 6 d.f. provided a reasonably good approximation to the empirical distribution of $\mathcal{T}$-values.



(a) Empirical Distribution of $\mathcal{T}$-values and Student's t-Distribution with 6 d.f.



(b) Empirical Distribution of $\mathcal{Z}$-values and the $N(0, 1)$ Distribution



(c) Empirical Distribution of $\mathcal{Q}$-values and the $\chi^2(6)$ Distribution

Figure 3: Empirical Distributions (Step Functions) of $\mathcal{T}$-, $\mathcal{Z}$-, and $\mathcal{Q}$-values and Their Assumed Theoretical Distributions (Smooth Curves) Based on 400 Runs of $\mathcal{WASSP}$ in AR(1) Process Using 95% CIs with ±3.75% Precision

Table 5: Correlation between $\overline{\overline{X}}(m, k)$ and $\widehat{\gamma}_X$ Based on 400 Runs of $\mathcal{WASSP}$ in the AR(1) Process

(a) Using 90% CIs

Precision Reqt.	$\text{Corr}\big[\overline{\overline{X}}(m, k), \widehat{\gamma}_X\big]$
None	−0.0041
±15%	0.0209
±7.5%	0.2768
±3.75%	0.1006

(b) Using 95% CIs

Precision Reqt.	$\text{Corr}\big[\overline{\overline{X}}(m, k), \widehat{\gamma}_X\big]$
None	0.1115
±15%	0.0189
±7.5%	0.0844
±3.75%	0.0647

3.2.2. Summary of Experimental Results for the AR(1) Process

Tables 6 and 7 show a comparison of the performance of $\mathcal{WASSP}$ (using $A = 7$); ASAP3; and the H&W spectral method for the given AR(1) process. In many cases, the actual precision of the CIs

delivered by ASAP3 was significantly smaller (more precise) than the requested level. For example, in the case of nominal 90% CIs with no precision requirement, ASAP3 delivered CIs that could satisfy a precision requirement of $\pm 2.325\%$ while exceeding the required coverage probability by 5.5%. To show a meaningful side-by-side comparison of the performance of $\mathcal{WASSP}$ with that of ASAP3 in this test process, we chose to display in Tables 6 and 7 the results for the following levels of precision: no precision; $\pm 3.75\%$; $\pm 1.875\%$; and $\pm 0.9375\%$. Regarding conformance to the precision and coverage-probability requirements for the delivered CIs, $\mathcal{WASSP}$ outperformed ASAP3 in the cases of no precision and $\pm 3.75\%$ precision while requiring substantially smaller sample sizes than ASAP3 required. For the precision levels $\pm 1.875\%$ and $\pm 0.9375\%$, both $\mathcal{WASSP}$ and ASAP3 achieved close conformance to the requested precision but exhibited significant CI overcoverage while requiring nearly the same sample sizes. The cause of this overcoverage is the subject of ongoing research.

Table 6: Performance of $\mathcal{WASSP}$ (Using $A = 7$), ASAP3, and H&W in the AR(1) Process Based on Independent Replications of 90% CIs

Prec. Reqt.	Performance Measure	$\mathcal{WASSP}$	H&W	ASAP3
None	# replications	1,000	1,000	400
	CI coverage	90.9%	46.9%	95.5%
	Avg. sample size	9,866	1,480	41,076
	Max. sample size	30,208	4,532	68,864
	Avg. CI half-length	5.3048	13.4345	2.325
	Var. CI half-length	1.8280	3.0321	0.170
	# replications satisfying prec. reqt.	1,000	1,000	400
	Coverage of CIs satisfying prec. reqt.	90.9%	46.9%	95.5%
$\pm 3.75\%$	# replications	1,000	1,000	400
	CI coverage	87%	94.9%	95.5%
	Avg. sample size	13,535	13,281	41,076
	Max. sample size	48,288	36,674	68,864
	Avg. CI half-length	3.2133	5.0865	2.325
	Var. CI half-length	0.1420	1.6660	0.170
	# replications satisfying prec. reqt.	1,000	149	400
	Coverage of CIs satisfying prec. reqt.	87%	91.95%	95.5%
$\pm 1.875\%$	# replications	1,000	1,000	400
	CI coverage	93.5%	90.1%	93.5%
	Avg. sample size	57,449	50,152	68,474
	Max. sample size	166,308	131,607	147,877
	Avg. CI half-length	1.6481	1.7659	1.7603
	Var. CI half-length	0.0423	0.1044	0.0134
	# replications satisfying prec. reqt.	1,000	697	400
	Coverage of CIs satisfying prec. reqt.	93.5%	87.2%	93.5%
$\pm 0.9375\%$	# replications	1,000	1,000	400
	CI coverage	94%	87.9%	94.25%
	Avg. sample size	229,730	173,700	213,826
	Max. sample size	717,388	429,686	381,184
	Avg. CI half-length	0.8297	0.8384	0.8941
	Var. CI half-length	0.0105	0.0201	0.0026
	# replications satisfying prec. reqt.	1,000	841	400
	Coverage of CIs satisfying prec. reqt.	94%	86.2%	94.25%

Table 7: Performance of $\mathcal{WASSP}$ (Using $A = 7$), ASAP3, and H&W in the AR(1) Process Based on Independent Replications of 95% CIs

Prec. Reqt.	Performance Measure	$\mathcal{WASSP}$	H&W	ASAP3
None	# replications	1,000	1,000	400
	CI coverage	94.5%	65.9%	98.8%
	Avg. sample size	9,824	1,474	41,076
	Max. sample size	22,592	3,389	68,864
	Avg. CI half-length	6.7331	16.7078	2.825
	Var. CI half-length	2.8826	4.5517	0.270
	# replications satisfying prec. reqt.	1,000	1,000	400
	Coverage of CIs satisfying prec. reqt.	94.5%	65.9%	98.8%
±3.75%	# replications	1,000	1,000	400
	CI coverage	95%	96.8%	98.8%
	Avg. sample size	21,099	20,176	41,208
	Max. sample size	59,360	48,036	68,864
	Avg. CI half-length	3.2800	4.5724	2.817
	Var. CI half-length	0.1529	1.9042	0.257
	# replications satisfying prec. reqt.	1,000	310	400
	Coverage of CIs satisfying prec. reqt.	95%	93.87%	98.8%
±1.875%	# replications	1,000	1,000	400
	CI coverage	97.7%	94%	99.25%
	Avg. sample size	90,371	73,249	101,526
	Max. sample size	328,160	168,236	223,173
	Avg. CI half-length	1.6584	1.6881	1.7703
	Var. CI half-length	0.0429	0.0784	0.0120
	# replications satisfying prec. reqt.	1,000	816	400
	Coverage of CIs satisfying prec. reqt.	97.7%	92.65%	99.25%
±0.9375%	# replications	1,000	1,000	400
	CI coverage	98%	92.9%	97.25%
	Avg. sample size	333,050	255,180	254,920
	Max. sample size	967,136	637,907	384,512
	Avg. CI half-length	0.8667	0.8539	0.8959
	Var. CI half-length	0.0115	0.0253	0.0021
	# replications satisfying prec. reqt.	1,000	817	400
	Coverage of CIs satisfying prec. reqt.	98%	91.3%	97.25%

The results for the H&W procedure were obtained in the same way as described in §3.1.2. For the no precision case, H&W-based CIs with nominal coverage probabilities of 90% and 95% had empirical coverage probabilities of 46.9% and 65.9%, respectively. For the case of nominal 90% CIs with a required precision of ±3.75%, the H&W procedure delivered 149 CIs with acceptable precision; and since 91.95% of those CIs actually covered μ_X, only $149 \times 0.9195 \approx 137$ out of 1,000 replications of the H&W procedure yielded correct results in this case. Overall, we judged the H&W spectral method to have broken down completely in the given AR(1) process.

By examining the absolute bias, variance, and mean squared error statistics in Table 8, we concluded that in the no precision case the H&W method had significant point-estimator bias and that the Cramér–von Mises test was not effective in detecting and eliminating that bias. For both the $\mathcal{WASSP}$ and H&W procedures, the bias of $\overline{\overline{X}}(m, k)$ shown in Table 8 represents a combination of two different effects. First, $\overline{\overline{X}}(m, k)$ is influenced in general by residual initialization bias—after

all, there is no unique, well-defined end of the warm-up period for the AR(1) process. Second, the truncation point $\mathcal{S}$ and the truncated simulation run length $n' = mk$ (as determined by $\mathcal{WASSP}$ or the H&W procedure) are random variables so that $\overline{\overline{X}}(m,k) = \left(\sum_{u=\mathcal{S}+1}^{\mathcal{S}+n'} X_u\right)\Big/ n'$ is a ratio of two random variables; and thus for the reasons detailed in §2.1 of Lada et al. (2004a) and in §4.2.1 of Lada (2003), $\overline{\overline{X}}(m,k)$ can also exhibit significant ratio-estimator bias due to (i) randomness of $\mathcal{S}$ and n'; or (ii) an insufficiently large value of n'.

Table 8: Mean Squared Error, Variance, and Absolute Bias of $\overline{\overline{X}}(m,k)$ Based on 1,000 Runs of $\mathcal{WASSP}$ and H&W in the AR(1) Process

Precision Requirement	Performance Measure	Nominal 90% CIs $\mathcal{WASSP}$	Nominal 90% CIs H&W	Nominal 95% CIs $\mathcal{WASSP}$	Nominal 95% CIs H&W
None	$\widehat{\text{MSE}}\left[\overline{\overline{X}}(m,k)\right]$	8.7493	224.033	8.0646	225.2592
	$\widehat{\text{Var}}\left[\overline{\overline{X}}(m,k)\right]$	7.7565	22.5305	7.2625	23.2921
	$\left\lvert\widehat{\text{Bias}}\left[\overline{\overline{X}}(m,k)\right]\right\rvert$	0.9964	14.1952	0.8956	14.2115
±3.75%	$\widehat{\text{MSE}}\left[\overline{\overline{X}}(m,k)\right]$	4.3745	4.2114	2.6315	2.549
	$\widehat{\text{Var}}\left[\overline{\overline{X}}(m,k)\right]$	4.1525	2.1591	2.5556	1.9094
	$\left\lvert\widehat{\text{Bias}}\left[\overline{\overline{X}}(m,k)\right]\right\rvert$	0.4712	1.4326	0.2755	0.7997
±1.875%	$\widehat{\text{MSE}}\left[\overline{\overline{X}}(m,k)\right]$	0.8698	1.1539	0.5177	0.7466
	$\widehat{\text{Var}}\left[\overline{\overline{X}}(m,k)\right]$	0.861	0.9242	0.5166	0.6627
	$\left\lvert\widehat{\text{Bias}}\left[\overline{\overline{X}}(m,k)\right]\right\rvert$	0.0938	0.4793	0.0332	0.2897
±0.9375%	$\widehat{\text{MSE}}\left[\overline{\overline{X}}(m,k)\right]$	0.1864	0.2966	0.1097	0.1793
	$\widehat{\text{Var}}\left[\overline{\overline{X}}(m,k)\right]$	0.1860	0.2785	0.1096	0.1709
	$\left\lvert\widehat{\text{Bias}}\left[\overline{\overline{X}}(m,k)\right]\right\rvert$	0.0200	0.1345	0.0100	0.0917

In summary from Tables 6, 7, and 8, we found that in the given AR(1) process, the performance of $\mathcal{WASSP}$ and ASAP3 was acceptable but the performance of the H&W procedure was unacceptable with respect to point-estimator accuracy and precision as well as CI coverage, precision, and stability.

3.3. The AR(1)-to-Pareto (ARTOP) Process

The "AR(1)-to-Pareto," or ARTOP process, is defined as follows. Let $\{Z_u : u = 1, 2, \ldots\}$ be a stationary AR(1) process with $N(0, 1)$ marginals and lag-one correlation ρ, which can be generated by the relation $Z_u = \rho Z_{u-1} + \delta_u$, where $Z_0 \sim N(0, 1)$ and $\{\delta_u : u = 1, 2, \ldots\} \overset{\text{i.i.d.}}{\sim} N\left(0, \sigma_\delta^2\right)$ is a white noise process with variance $\sigma_\delta^2 = 1 - \rho^2$. If $\{X_u : u = 1, 2, \ldots\}$ is an ARTOP process with marginal c.d.f.

$$F_X(x) \equiv \Pr\{X \le x\} = \begin{cases} 1 - (\xi/x)^\vartheta, & x \ge \xi, \\ 0, & x < \xi, \end{cases} \tag{11}$$

where $\xi > 0$ is a location parameter and $\vartheta > 0$ is a shape parameter, then we generate $\{X_u\}$ from the "base process" $\{Z_u\}$ as follows. For all real z, let $\Phi(z) = (2\pi)^{-1/2} \int_{-\infty}^{z} \exp\left(-\frac{1}{2}w^2\right) \mathrm{d}w$ denote the c.d.f. of the $N(0, 1)$ distribution. Taking $R_u = \Phi(Z_u)$ for $u = 1, 2, \ldots$, we obtain a sequence of correlated random numbers $\{R_u : u = 1, 2, \ldots\}$; and then applying the inverse of the Pareto c.d.f. (11), we finally obtain $\{X_u : u = 1, 2, \ldots\}$ so that

$$X_u = F_X^{-1}(R_u) = F_X^{-1}\left[\Phi(Z_u)\right] = \xi \Big/ [1 - \Phi(Z_u)]^{1/\vartheta}, \quad u = 1, 2, \ldots . \tag{12}$$

The mean and the variance of the ARTOP process (12) are given by $\mu_X = \mathrm{E}[X_u] = \vartheta\xi(\vartheta - 1)^{-1}$ (for $\vartheta > 1$) and $\sigma_X^2 = \xi^2\vartheta(\vartheta - 1)^{-2}(\vartheta - 2)^{-1}$ (for $\vartheta > 2$), respectively (Lada 2003).

We set the parameters of the Pareto distribution (11) according to $\vartheta = 2.1$ and $\xi = 1$; and we set the lag-one correlation in the base process $\{Z_u\}$ to $\rho = 0.995$. This yields a test process whose marginal distribution has mean, variance, skewness, and kurtosis respectively given by $\mu_X = 1.9091$; $\sigma_X^2 = 17.3554$; $\mathrm{E}\left\{[(X_u - \mu_X)/\sigma_X]^3\right\} = \infty$; and $\mathrm{E}\left\{[(X_u - \mu_X)/\sigma_X]^4\right\} = \infty$. The most difficult aspects of this test process are its highly nonnormal marginals and persistent autocorrelation structure. We took $Z_0 \sim N(0, 1)$ so that the process started in steady-state operation and therefore had no warm-up period.

Table 9 shows the performance of $\mathcal{WASSP}$ for the given ARTOP process using the smoothing parameter values $A = 5$, 7, and 9. The results are based on 400 independent replications of nominal 90% CIs. From this table, we concluded that the coverage decreased in general as A increased. For nominal 90% CIs with $A = 7$ and $A = 9$, we judged the resulting coverage probabilities to be unacceptable at all three precision levels. In this application of $\mathcal{WASSP}$, the value $A = 5$ appeared to yield the best results, especially in the $\pm 7.5\%$ precision case. While there was significant undercoverage when we used $A = 5$ in the small-sample cases, clearly as the sample size increased the coverage probabilities approached the nominal level. It is unclear at this point why $A = 5$ produced the best results for this process. Nonetheless, we decided to override the default value of the smoothing parameter and take $A = 5$ for this ARTOP process.

3.3.1. Validation of Student's t-Ratio Assumptions for the ARTOP Process

In this section we summarize the experiments used to validate the Assumptions A_1–A_3 on which we based the conclusion that the random variable $\mathcal{T}$ defined in (8) would have Student's t-distribution with $2a = 4$ d.f. when we set $A = 5$ and applied $\mathcal{WASSP}$ to the ARTOP process (12). Using the method detailed in Appendix B, first we computed the SSVC for the ARTOP process; and thus we obtained $\gamma_X \approx 1612.8$ with maximum relative error $\varepsilon_{\mathrm{rel}} = 10^{-6}$. Figures 4(a)–4(c) show the

Table 9: Performance of $\mathcal{WASSP}$ Using Different Values of A in the ARTOP Process Based on 400 Replications of 90% CIs

Precision Requirement	Performance Measure	Smoothing Parameter		
		$A = 5$	$A = 7$	$A = 9$
None	CI coverage	84.2%	79%	78%
	Avg. sample size	19,880	22,512	22,512
	Max. sample size	955,008	1,906,880	1,906,880
	Avg. CI half-length	0.5183	0.4475	0.4155
	Var. CI half-length	0.0774	0.0544	0.0441
±15%	CI coverage	77.3%	71.5%	72.3%
	Avg. sample size	79,095	66,158	54,551
	Max. sample size	1,674,368	1,177,712	1,126,512
	Avg. CI half-length	0.2198	0.2231	0.2296
	Var. CI half-length	0.0020	0.0018	0.0018
±7.5%	CI coverage	89%	85.3%	82.5%
	Avg. sample size	430,430	345,870	272,670
	Max. sample size	8,466,630	6,633,568	5,365,936
	Avg. CI half-length	0.1152	0.1159	0.1177
	Var. CI half-length	0.0005	0.0005	0.0005

histogram and c.d.f. plots of the $\mathcal{T}$-, $\mathcal{Z}$-, and $\mathcal{Q}$-values generated by 400 replications of the given ARTOP process using $A = 5$ with the precision level ±7.5% for nominal 95% CIs.

From the plot in Figure 4(b), we concluded that the distribution of $\mathcal{Z}$-values was skewed and that the mean $\overline{\overline{X}}(m, k)$ was biased. Since the given ARTOP process was started in steady-state operation, there was no initialization bias in the point estimator $\overline{\overline{X}}(m, k)$ delivered by $\mathcal{WASSP}$. Therefore, the bias must have been caused entirely by the randomness of the truncation point $\mathcal{S}$ and the overall simulation run length $n = \mathcal{S} + n'$ as explained in the next-to-last paragraph of §3.2.2. From Figures 4(b) and 4(c) and the other plots presented in §4.4.1 of Lada (2003), we concluded that Assumptions A_1 and A_2 were satisfied at least to a fair approximation in this test process.

To validate Assumption A_3, we computed the correlation between $\overline{\overline{X}}(m, k)$ and $\widehat{\gamma}_X$ for 400 independent replications of $\mathcal{WASSP}$ applied to the ARTOP process. From Tables 10(a) and 10(b), we concluded that the correlation between $\overline{\overline{X}}(m, k)$ and $\widehat{\gamma}_X$ was significant at all three levels of precision and that $\mathcal{Z}$ and $\mathcal{Q}$ were not independent in this test problem. This phenomenon may partially explain the undercoverage seen in the no precision and ±15% precision cases in Table 9.

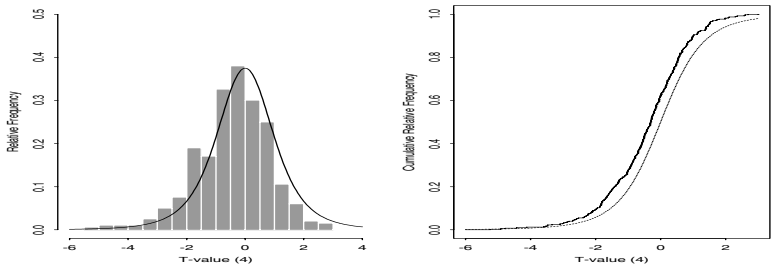
Table 10: Correlation between $\overline{\overline{X}}(m, k)$ and $\widehat{\gamma}_X$ Based on 400 Runs of $\mathcal{WASSP}$ in the ARTOP Process

(a) Using 90% CIs

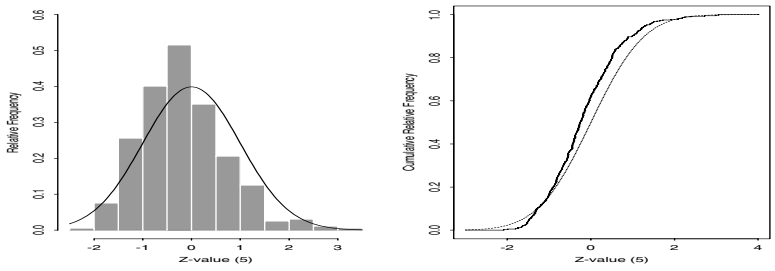
Precision Reqt.	$\text{Corr}\big[\overline{\overline{X}}(m, k), \widehat{\gamma}_X\big]$
None	0.5906
±15%	0.5587
±7.5%	0.5462

(b) Using 95% CIs

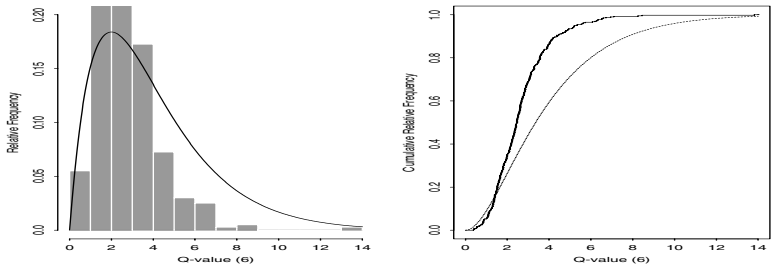
Precision Reqt.	$\text{Corr}\big[\overline{\overline{X}}(m, k), \widehat{\gamma}_X\big]$
None	0.6203
±15%	0.5882
±7.5%	0.5034



(a) Empirical Distribution of $\mathcal{T}$-values and Student's t-Distribution with 4 d.f.



(b) Empirical Distribution of $\mathcal{Z}$-values and the $N(0, 1)$ Distribution



(c) Empirical Distribution of $\mathcal{Q}$-values and the $\chi^2(4)$ Distribution

Figure 4: Empirical Distributions (Step Functions) of $\mathcal{T}$-, $\mathcal{Z}$-, and $\mathcal{Q}$-values and Their Assumed Theoretical Distributions (Smooth Curves) Based on 400 Runs of $\mathcal{WASSP}$ in ARTOP Process Using 95% CIs with $\pm$7.5% Precision

3.3.2. Summary of Experimental Results for the ARTOP Process

Table 11 shows a comparison of the performance of $\mathcal{WASSP}$ (using $A = 5$); ASAP3; and the H&W spectral method for the ARTOP process (12). For the no precision case, ASAP3 and $\mathcal{WASSP}$ yielded somewhat similar results in terms of CI coverage. In this case, however, ASAP3 required significantly larger sample sizes on the average than $\mathcal{WASSP}$ required. For nominal 90% CIs in the $\pm$15% precision case, ASAP3 outperformed $\mathcal{WASSP}$; but for nominal 95% CIs with $\pm$15% precision, the two methods produced comparable results. In the case of nominal 90% CIs with $\pm$7.5% precision, the coverage probability for $\mathcal{WASSP}$ was close to the nominal level, while the coverage probability for ASAP3 was significantly below the nominal level; however, in this case

$\mathcal{WASSP}$'s average required sample size was 130% larger than that of ASAP3.

Table 11: Performance of $\mathcal{WASSP}$ (Using $A = 5$), ASAP3, and H&W in the ARTOP Process

Results for Nominal 90% Confidence Intervals				
Prec. Reqt.	Performance Measure	$\mathcal{WASSP}$	H&W	ASAP3
None	# replications	400	400	400
	CI coverage	84.2%	67%	85.5%
	Avg. sample size	19,880	2,982	114,053
	Max. sample size	955,008	143,252	524,288
	Avg. CI half-length	0.5183	0.7124	0.0909
	Var. CI half-length	0.0774	0.6841	0.0019
	# replications satisfying prec. reqt.	400	400	400
	Coverage of CIs satisfying prec. reqt.	84.2%	67%	85.5%
±15%	# replications	400	400	400
	CI coverage	77.3%	72.8%	85.5%
	Avg. sample size	79,095	39,781	117,092
	Max. sample size	1,674,368	673,442	722,944
	Avg. CI half-length	0.2198	0.2341	0.0867
	Var. CI half-length	0.0022	0.0024	0.0006
	# replications satisfying prec. reqt.	400	389	400
	Coverage of CIs satisfying prec. reqt.	77.3%	72%	85.5%
±7.5%	# reps.	400	400	400
	CI coverage	89%	81.8%	84%
	Avg. sample size	430,430	208,570	186,517
	Max. sample size	8,466,630	1,311,863	2,873,344
	Avg. CI half-length	0.1152	0.1194	0.0676
	Var. CI half-length	0.0005	0.0003	0.00005
	# replications satisfying prec. reqt.	400	395	400
	Coverage of CIs satisfying prec. reqt.	89%	82%	84%
Results for Nominal 95% Confidence Intervals				
None	# replications	400	400	400
	CI coverage	89%	75.5%	90.75%
	Avg. sample size	17,028	2,555	114,053
	Max. sample size	454,410	10,741	524,288
	Avg. CI half-length	0.7042	0.9391	0.1089
	Var. CI half-length	0.1802	1.8882	0.0029
	# replications satisfying prec. reqt.	400	400	400
	Coverage of CIs satisfying prec. reqt.	89%	75.5%	90.75%
±15%	# replications	400	400	400
	CI coverage	87%	84%	90.75%
	Avg. sample size	151,190	72,093	120,660
	Max. sample size	3,059,792	688,454	1,028,096
	Avg. CI half-length	0.2232	0.2406	0.1008
	Var. CI half-length	0.0021	0.0187	0.0006
	# replications satisfying prec. reqt.	400	394	400
	Coverage of CIs satisfying prec. reqt.	87%	84%	90.75%
±7.5%	# replications	400	400	400
	CI coverage	95%	90%	90.25%
	Avg. sample size	747,640	348,470	255,512
	Max. sample size	6,975,686	2,354,295	2,097,152
	Avg. CI half-length	0.1190	0.1200	0.0696
	Var. CI half-length	0.0005	0.0003	0.00003
	# replications satisfying prec. reqt.	400	334	400
	Coverage of CIs satisfying prec. reqt.	95%	88.1%	90.25%

We also concluded from Table 11 that $\mathcal{WASSP}$-generated CIs had much better coverage probabilities than the H&W-generated CIs across all levels of the precision requirement. Even though

$\mathcal{WASSP}$ produced significant undercoverage in the small-sample cases (especially for nominal 90% CIs), in the case of $\pm 7.5\%$ precision it produced coverages that were close to the nominal confidence level. For the case of nominal 95% CIs with a required precision of $\pm 7.5\%$, the H&W procedure delivered $R = 334$ CIs with acceptable precision; and since 88.1% of those CIs actually covered μ_X, we see that only $334 \times 0.881 \approx 294$ out of 1,000 replications of the H&W procedure yielded correct results in this case. Because this test process was started in steady-state operation, we did not tabulate the mean-squared error and absolute bias of $\overline{\overline{X}}(m, k)$ for any of the output analysis procedures.

In summary from Table 11 we found that in the given ARTOP process, the performance of $\mathcal{WASSP}$ and ASAP3 was acceptable but the performance of the H&W procedure was unacceptable with respect to CI coverage, precision, and stability.

4. Conclusions and Recommendations

In the experimental performance evaluation summarized in §3, we presented three test processes that were specifically designed to compare $\mathcal{WASSP}$, the H&W spectral method, and ASAP3 with respect to their efficiency and the robustness of the CIs delivered by these procedures. From the experimental results presented in §3, we concluded that $\mathcal{WASSP}$ outperformed the H&W method in every respect; and we believe $\mathcal{WASSP}$ represents an advance in spectral methods for simulation output analysis. The comparison of $\mathcal{WASSP}$ and ASAP3 is less clear-cut, with neither procedure dominating the other in the experiments we performed. Both $\mathcal{WASSP}$ and ASAP3 were designed to deliver point and confidence-interval estimators for the steady-state mean of a simulation output process. $\mathcal{WASSP}$, however, also provides an estimator of the SSVC with reasonably stable behavior as well as an estimator of the entire power spectrum of the delivered set of batch means. This additional information can be useful in planning follow-up experiments.

The experimental results detailed in §3 provide substantial evidence of $\mathcal{WASSP}$'s ability to deliver approximately valid CIs for the steady-state mean of a simulation-generated process with relative precision levels and nominal coverage probabilities that often arise in practical applications. However, we will continue our experimental investigation of the efficiency and robustness of $\mathcal{WASSP}$ when it is applied to interesting test problems—including processes with long-range dependence as well as queuing network models with multiple customer classes, probabilistic routing, subnetwork capacity constraints, and workstation utilizations that are commonly encountered in certain application domains.

Appendix A: Spectrum of the $M/M/1$ Waiting Time Process

In terms of the setup in §3.1, the waiting time process has mean $\mu_X = \tau/[\mu(1-\tau)] = \tau^2/[\lambda(1-\tau)]$, variance $\sigma_X^2 = \tau^3(2-\tau)\big/\big[\lambda^2(1-\tau)^2\big]$, and steady-state variance constant γ_X as given in §3.1.1. We seek to calculate and plot the power spectrum (2) of $\{X_u\}$, which can be expressed as $p_X(\omega) = \sigma_X^2\big[1 + 2\sum_{\ell=1}^{\infty}\rho_X(\ell)\cos(2\pi\omega\ell)\big]$, where $-\frac{1}{2} \le \omega \le \frac{1}{2}$ and $\rho_X(\ell) = \mathrm{Corr}(X_u, X_{u+\ell}) = \gamma_X(\ell)\big/\sigma_X^2$ for $\ell = 0, \pm1, \pm2, \ldots$ denotes the corresponding autocorrelation function. Let $q_X(\omega)$ denote the cosine transform of the autocorrelation function. From equation (34) of Daley (1968) as corrected on p. 117 of Song and Schmeiser (1995), we have

$$\rho_X(\ell) = \frac{(1-\tau)^3(1+\tau)}{2\pi\,\tau^3(2-\tau)} \int_0^r t^{|\ell|}\,\frac{t^{3/2}(r-t)^{1/2}}{(1-t)^3}\,\mathrm{d}t \quad \text{for } \ell = \pm1, \pm2, \ldots,$$

and hence

$$q_X(\omega) = \frac{p_X(\omega)}{\sigma_X^2} = 1 + \frac{(1-\tau)^3(1+\tau)}{\pi\,\tau^3(2-\tau)} \int_0^r \left[\sum_{\ell=1}^{\infty} t^\ell \cos(2\pi\omega\ell)\right] \frac{t^{3/2}(r-t)^{1/2}}{(1-t)^3}\,\mathrm{d}t, \tag{13}$$

where $r = 4\tau\big/\big[(1+\tau)^2\big] \in (0, 1)$ since $\tau \in (0, 1)$.

The interchange of the integration and summation operations in (13) is justified by observing that the definition of r implies $0 \le t \le r < 1$ and $\big|\sum_{\ell=1}^{\infty} t^\ell \cos(2\pi\omega\ell)\big| \le \sum_{\ell=1}^{\infty} t^\ell = t/(1-t)$; and since the integrand in (13) is majorized by the function $t^{5/2}(r-t)^{1/2}\big/(1-t)^4$, which is continuous and hence integrable on $[0, r]$, we see that (13) follows by Lebesgue's dominated convergence theorem. Moreover, formula 1.447 2 of Gradshteyn and Ryzhik (2000) implies that

$$\sum_{\ell=1}^{\infty} t^\ell \cos(2\pi\omega\ell) = \frac{t[\cos(2\pi\omega) - t]}{1 - 2t\cos(2\pi\omega) + t^2}. \tag{14}$$

Inserting (14) and the expression for the process variance σ_X^2 into (13), we obtain the formula (7).

Appendix B: SSVC for an "AR(1)-to-Anything" Process

In terms of the setup in §3.3, we seek to calculate the SSVC for the process $\{X_u\}$. Notice that

$$\gamma_X(\ell) \;=\; \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \Big\{F_X^{-1}[\Phi(z_1)] - \mu_X\Big\}\Big\{F_X^{-1}[\Phi(z_2)] - \mu_X\Big\}\,\varphi_2\Big(z_1, z_2; \rho^{|\ell|}\Big)\,\mathrm{d}z_1\,\mathrm{d}z_2,$$

where $\quad \varphi_2(z_1, z_2; \vartheta) \equiv \dfrac{\exp\big\{-\big(z_1^2 - 2\vartheta z_1 z_2 + z_2^2\big)\,/\,\big[2(1-\vartheta^2)\big]\big\}}{2\pi\sqrt{1-\vartheta^2}}$ for $-\infty < z_1, z_2 < \infty$

is the bivariate standard normal p.d.f. with correlation $\vartheta \in (-1, 1)$. To simplify the notation, we let $\Delta(z_u) \equiv F_X^{-1}[\Phi(z_u)] - \mu_X$ for $u = 1, 2$, so that provided $\rho \in (-1, 1)$ we have

$$\gamma_X(\ell) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \Delta(z_1)\, \Delta(z_2)\, \varphi_2\Big(z_1, z_2; \rho^{|\ell|}\Big)\, \mathrm{d}z_1\, \mathrm{d}z_2 \;\text{ for }\; \ell = \pm 1, \pm 2, \ldots . \tag{15}$$

It follows from equation (12.6.8) of Cramér (1946) that for $\vartheta \in (-1, +1)$, we have

$$\varphi_2(z_1, z_2; \vartheta) = \sum_{u=0}^{\infty} \frac{1}{u!} H_u(z_1)\, H_u(z_2)\, \varphi(z_1)\, \varphi(z_2)\, \vartheta^u \;\text{ for }\; -\infty < z_1, z_2 < \infty, \tag{16}$$

where $\varphi(z) = \exp\big(-\frac{1}{2}z^2\big)\big/\sqrt{2\pi}$ is the standard normal p.d.f. and for $u = 0, 1, 2, \ldots$, the uth Hermite polynomial is defined by $\frac{d^u}{dz^u}\varphi(z) = (-1)^u H_u(z)\varphi(z)$. Inserting (16) into (15) and then inserting the result into (1), we have

$$\begin{aligned} \gamma_X &= \sigma_X^2 + 2\sum_{\ell=1}^{\infty} \int_{-\infty}^{+\infty} \prod_{i=1}^{2} \Delta(z_i) \left[\sum_{u=0}^{\infty} \frac{\rho^{|\ell|u}}{u!} \prod_{j=1}^{2} H_u(z_j)\, \varphi(z_j)\right] \mathrm{d}z_1\, \mathrm{d}z_2 \\ &= \sigma_X^2 + 2\sum_{\ell=1}^{\infty} \sum_{u=0}^{\infty} \frac{\rho^{u\ell}}{u!} \left[\int_{-\infty}^{+\infty} \Delta(z)\, H_u(z)\, \varphi(z)\, \mathrm{d}z\right]^2 . \end{aligned} \tag{17}$$

The interchanges of integration and summation operations in (17) are justified as follows. If we let $\Xi_u \equiv \int_{-\infty}^{+\infty} \Delta(z)\, H_u(z)\, \varphi(z)\, \mathrm{d}z$ for $u = 0, 1, \ldots$, then we have $\Xi_0 = 0$ and $\Xi_u = \int_{-\infty}^{\infty} F_X^{-1}[\Phi(z)]\, H_u(z)\, \varphi(z)\, \mathrm{d}z$ for $u = 1, 2, \ldots$, by standard properties of Hermite polynomials (specifically, equation (12.6.5) of Cramér); and thus we see that

$$\Xi_u^2 \;\le\; \left(\int_{-\infty}^{+\infty} \Big\{F_X^{-1}[\Phi(z)]\Big\}^2 \varphi(z)\, \mathrm{d}z\right) \left(\int_{-\infty}^{+\infty} [H_u(z)]^2\, \varphi(z)\, \mathrm{d}z\right) = \big(\sigma_X^2 + \mu_X^2\big)\, u! \tag{18}$$

by the Cauchy-Schwarz inequality (display (9.5.1) of Cramér). Inserting the final right-hand side of (18) into (17), we see that (17) is absolutely convergent so that the desired result holds. Moreover, it also follows from (18) that

$$\gamma_X = \sigma_X^2 + 2\sum_{u=1}^{\infty} \frac{\Xi_u^2\, \rho^u}{u!(1 - \rho^u)}.$$

To estimate γ_X with maximum relative error $\varepsilon_{\mathrm{rel}}$, we compute the partial sum $Q_v = \sigma_X^2 + 2\sum_{u=1}^{v} \Xi_u^2\, \rho^u \big/ \big[u!(1 - \rho^u)\big]$ for $v = 1, 2, \ldots$, where $Q_0 \equiv \sigma_X^2$ and for $u \ge 1$, we evaluate Ξ_u numerically as $\Xi_u \approx \int_{-8}^{8} F_X^{-1}\,[\Phi(z)]\; H_u(z)\, \varphi(z)\, \mathrm{d}z$. We stop evaluating Ξ_u in this way when $|(Q_v - Q_{v-1})/Q_{v-1}| \le \varepsilon_{\mathrm{rel}}$; and then we deliver Q_v as the estimate of γ_X with maximum relative error $\varepsilon_{\mathrm{rel}}$.

Acknowledgements

The authors thank Stephen D. Roberts and Charles E. Smith (NC State University); and David Goldsman (Georgia Tech) for many enlightening discussions on this paper. This research was partially supported by NSF grant DMI-9900164 and by the American Association of University Women (AAUW) through an AAUW Educational Foundation Engineering Dissertation Fellowship.

References

Anderson, T.W., D.A. Darling. 1952. Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. *The Annals of Mathematical Statistics* **23** (2) 193–212.

Cramér, H. 1946. *Mathematical Methods of Statistics*. Princeton University Press, Princeton, NJ.

Daley, D.J. 1968. The serial correlation coefficients of waiting times in a stationary single server queue. *Journal of the Australian Mathematical Society* **8** 683–699.

Gradshteyn, I.S., I.M. Ryzhik. 2000. *Table of Integrals, Series, and Products, Sixth Edition*. Alan Jeffrey, ed. Academic Press, San Diego.

Heidelberger, P., P.D. Welch. 1981a. A spectral method for confidence interval generation and run length control in simulations. *Communications of the ACM* **24** (4) 233–245.

Heidelberger, P., P.D. Welch. 1981b. Adaptive spectral methods for simulation output analysis. *IBM Journal of Research and Development* **25** (6) 860–876.

Heidelberger, P., P.D. Welch. 1983. Simulation run length control in the presence of an initial transient. *Operations Research* **31** (6) 1109–1144.

Lada, E.K. 2003. A wavelet-based procedure for steady-state simulation output analysis. Ph.D. Dissertation, Operations Research Program, NC State University, Raleigh, NC. `www.lib.ncsu.edu/theses/available/etd-04032003-141616/unrestricted/etd.pdf` [accessed March 22, 20042].

Lada, E.K., J.R. Wilson. 2004. A wavelet-based spectral procedure for steady-state simulation analysis. *European Journal of Operational Research* in review. `ftp.eos.ncsu.edu/pub/jwilson/lada04ejor.pdf` [accessed December 2, 2004].

Lada, E.K., J.R. Wilson, N.M. Steiger. 2003. A wavelet-based spectral method for steady-state simulation analysis. S. Chick, P. J. Sánchez, D. Ferrin, and D. J. Morrice, eds. *Proceedings of the 2003 Winter Simulation Conference*. Institute of Electrical and Electronics Engineers, Piscataway, NJ. 422–430.

Lada, E.K., J.R. Wilson, N.M. Steiger, and J.A. Joines. 2004a. Experimental performance evaluation of a wavelet-based spectral analysis procedure for steady-state simulation. R. Ingalls, M. Rossetti, J. Smith, and B. Peters, eds. *Proceedings of the 2004 Winter Simulation Conference*. Institute of Electrical and Electronics Engineers, Piscataway, NJ. 692–702.

Lada, E.K., J.R. Wilson, N.M. Steiger, and J.A. Joines. 2004b. *WASSP* software and user's manual

[online]. Department of Industrial Engineering, NC State University, Raleigh, NC. `ftp.eos.ncsu.edu/pub/jwilson/installwassp.exe`. [accessed December 2, 2004].

Pawlikowski, K. 1990. Steady-state simulation of queueing processes: A survey of problems and solutions. *ACM Computing Surveys* **22** (2) 123–170.

Shapiro, S.S., M.B. Wilk. 1965. An analysis of variance test for normality (complete samples). *Biometrika* **52** (3–4) 591–611.

Song, W.T., B.W. Schmeiser. 1995. Optimal mean-squared-error batch sizes. *Management Science* **41** (1) 110–123.

Steiger, N.M., E.K. Lada, J.R. Wilson, C. Alexopoulos, D. Goldsman, F. Zouaoui. 2002. ASAP2: An improved batch means procedure for simulation output analysis. E. Yücesan, C.-H. Chen, J. L. Snowdon, and J. M. Charnes, eds. *Proceedings of the 2002 Winter Simulation Conference*, Institute of Electrical and Electronics Engineers, Piscataway, NJ. 336–344.

Steiger, N.M., E.K. Lada, J.R. Wilson, J.A. Joines, C. Alexopoulos, D. Goldsman. 2005. ASAP3: A batch means procedure for steady-state simulation analysis. *ACM Transactions on Modeling and Computer Simulation* to appear.

Steiger, N.M., J.R. Wilson. 2002. An improved batch means procedure for simulation output analysis. *Management Science* **48** (12) 1569–1586.

von Neumann, J. 1941. Distribution of the ratio of the mean square successive difference to the variance. *Annals of Mathematical Statistics* **12** 367–395.