On the Marginal Standard Error Rule and the Testing of Initial Transient Deletion Methods

ROB J. WANG and PETER W. GLYNN, Stanford University

In the planning of steady-state simulations, a central issue is the initial transient problem, in which an initial segment of the simulation output is adversely contaminated by initialization bias. Our article makes several contributions toward the analysis of this computational challenge. To begin, we introduce useful ways for measuring the magnitude of the initial transient effect in the single replication setting. We then analyze the marginal standard error rule (MSER) and prove that MSER's deletion point is determined, as the simulation time horizon tends to infinity, by the minimizer of a certain random walk. We use this insight, together with fluid limit intuition associated with queueing models, to generate two nonpathological examples in which at least one variant of MSER fails to accurately predict the duration of the initial transient. Our results suggest that the efficacy of a deletion procedure is sensitive to the choice of performance measure, and that the set of standard test problems on which initial transient procedures are tested should be significantly broadened.

Categories and Subject Descriptors: I.6.6 [**Simulation and Modeling**]: Simulation Output Analysis

General Terms: Algorithms, Performance, Theory

Additional Key Words and Phrases: Initial transient problem, MSER, truncation procedures, queueing theory, fluid limits

ACM Reference Format:
Rob J. Wang and Peter W. Glynn. 2016. On the marginal standard error rule and the testing of initial transient deletion methods. ACM Trans. Model. Comput. Simul. 27, 1, Article 1 (August 2016), 30 pages.
DOI: http://dx.doi.org/10.1145/2961052

1. INTRODUCTION

Let $Y = (Y_j : j \geq 0)$ be a real-valued sequence representing the output of a simulation for which there exists a (deterministic) real-valued constant α with

$$\overline{Y}_n \triangleq \frac{1}{n}\sum_{j=0}^{n-1} Y_j \xrightarrow{p} \alpha \tag{1}$$

as $n \to \infty$, where $\xrightarrow{p}$ denotes convergence in probability. The quantity α is known as the steady-state mean of Y, and computing α accurately is the goal of a steady-state simulation of Y.

In the typical performance simulation context, Y is initialized according to a distribution atypical of equilibrium, thereby inducing an initial transient in the simulation of Y in which the observations collected are biased as estimators of the steady-state mean. Since one has no a priori sense of how much bias the initial transient is generating, it

R. J. Wang was supported by an Arvanitidis Stanford Graduate Fellowship (SGF) in memory of William K. Linvill and an NSERC Postgraduate Scholarship (PGS D).
Authors' address: R. J. Wang and P. W. Glynn, Department of Management Science and Engineering, Stanford University, 475 Via Ortega, Stanford, CA. 94305; emails: {robjwang, glynn}@stanford.edu.

DOI: http://dx.doi.org/10.1145/2961052

is of great interest to develop automatic procedures for determining the duration of the initial transient so that the initial transient (determined to be of length $\hat{d}$, say) can be deleted, thereby reducing the bias of the estimator that averages over the remaining data. Naturally, one would like to choose $\hat{d}$ that strikes the best bias-variance trade-off: truncating too little fails to sufficiently reduce initialization bias, whereas truncating too much leads to an increase in variance. As a result, in the steady-state simulation community, a popular criterion for measuring the quality of a steady-state estimator is mean square error (MSE); in turn, the effectiveness of an initial transient rule is often measured by the ability to minimize MSE.

The initial transient problem has been studied extensively in recent decades. An early survey of the subject dates back to Wilson and Pritsker [1978]. Robinson [2002] contains a list of truncation procedures that are popular in practice. Other works, such as Schruben [1982] and Schruben et al. [1983], consider initialization bias detection from a hypothesis testing perspective. In addition to estimating steady-state means, one sometimes also wishes to deliver confidence intervals (CIs) in the presence of an initial transient. Attempts at these constructions may be found, for example, in Lada et al. [2006], Tafazzoli et al. [2011], and Mokashi et al. [2010]. A key strategy in these works is the employment of von Neumann's randomness test to batch means obtained from nonoverlapping batches, deleting the first batch of data once von Neumann's test fails to reject the hypothesis of independence. Such a heuristic procedure often leads to good computational results.

Our article is organized as follows. In Section 2, we propose various ways of measuring the strength of the transient effect in a given simulation, emphasizing the difference between what we call the "distributional initial transient" and the "functional initial transient." In the course of this discussion, we provide a complete description (Theorem 1) of the MSE for steady-state estimators associated with a geometrically ergodic Markov chain, as a function of the initial distribution and the truncation point. This result complements the MSE analysis in Pasupathy and Schmeiser [2010] and Grassmann [2011] by exposing the second-order terms in the MSE that are influenced by initial transient effects. We therefore expect this result to be a useful theoretical device in future research on the initial transient problem.

Section 3 is focused on MSER and its variants. We show, through Theorem 2, that the truncation point specified by MSER converges almost surely (a.s.) to a finite-valued quantity as the simulation time horizon tends to infinity, and that the behavior of MSER is (very) closely connected to that of a certain random walk. In Section 4, we argue that fluid limits in queueing theory provide a valuable theoretical tool for generating test problems in which the duration of the initial transient can be unambiguously determined. We verify that MSER truncates an appropriate amount in the setting of a $G/G/1$ waiting time sequence as the simulation time horizon and initial condition simultaneously tend to infinity. However, the combination of fluid level intuition with the random walk structure of MSER allows us to generate, in Section 5, two practically relevant models in which at least one variant of MSER fails to find an appropriate truncation point. A common feature of these examples is that the simulation output results from a particular performance measure, and thus does not take into account the full state space information of the underlying stochastic process.

In Section 6, we summarize the conclusions of our article. The Appendix contains all proofs of theorems and propositions.

2. THE INITIAL TRANSIENT PROBLEM: PERSPECTIVES AND CHALLENGES

Analysis of the initial transient effect differs depending on whether one uses a single long replication of Y or multiple (independent) shorter replications. In the single replication context, the extraction of the initial transient "signal" from the surrounding stochastic "noise" associated with the simulation is challenging because one has only a single sample path from which to infer the effect of the initial transient. Nevertheless, a substantial proportion of the literature focuses on the single replication setting, largely because the effect of the initial transient is mitigated by using a single long run (as opposed to multiple short runs, in which the initial transient is replicated with each run). Our work focuses only on the single replication setting; the theory for multiple replication exhibits different behavior and will appear elsewhere.

In the rest of this article, we presume that the simulation output Y admits the representation $Y_j = f(X_j)$, where $X = (X_j : j \geq 0)$ is an $\mathbb{S}$-valued Markov chain (with state space $\mathbb{S}$) and $f : \mathbb{S} \to \mathbb{R}$ is a real-valued performance measure. Indeed, this assumption is quite natural, as a large subset of discrete-event simulations can be modeled by Markov or generalizations of Markov chains. To see this, recall that in a typical discrete-event simulation, the simulation program maintains a set of state variables as well as a future event list (ordered from most imminent to least imminent). At each event time, random variables (which, without loss of generality, can be taken as uniformly distributed) are generated so as to determine new future events. These are then appropriately inserted into the future event list, and the state variables are updated in accord with the most imminent event in the list (after which this event is removed). Thus, the simulation updating dynamics are Markovian in nature. For example, see Glynn [1989] for a deeper discussion of the relationship between discrete-event simulations, generalized semi-Markov processes (GSMPs), and general state space Markov chains.

We start with a simple observation that is nevertheless (perhaps) surprising. Suppose that we estimate the deletion time, based on a simulation of X up to time $n-1$, via a random variable $\hat{d}_n$ that is a function of $(X_j : 0 \leq j \leq n-1)$. If π is the stationary distribution of X, then

$$P_\pi(\hat{d}_n \in \cdot) = \int_{\mathbb{S}} \pi(dx) P_x(\hat{d}_n \in \cdot),$$

where $P_x(\cdot) \overset{\Delta}{=} P(\cdot \,|\, X_0 = x)$ and

$$P_\pi(\cdot) = \int_{\mathbb{S}} \pi(dx) P_x(\cdot).$$

It follows that $\hat{d}_n = 0$ a.s. under P_π if and only if $P_x(\hat{d}_n = 0)$ for almost every x under π. Therefore, in this single replication setting, any initial transient deletion rule that deletes a nonzero portion of the simulation run in the presence of the transient induced by starting at state x will necessarily also need to delete a nonzero portion of the simulated data when the system is started in equilibrium.

Our remaining discussion assumes that there exists a subset $A \subseteq \mathbb{S}$, $\lambda > 0$, $\kappa > 0$, $c > 0$, a probability measure $\phi(\cdot)$ concentrated on A, and a function $v : \mathbb{S} \to [1, \infty)$ for which

$$\text{(A1)} \qquad \text{(i)} \quad P_x(X_1 \in B) \geq \lambda\phi(B), \;\; \forall x \in A, \;\; \text{(measurable) } B \subseteq \mathbb{S}$$

$$\text{(ii)} \quad E_x v(X_1) \leq (1-\kappa)v(x) + cI(x \in A), x \in \mathbb{S}.$$

Here, $E_x(\cdot) \overset{\Delta}{=} E(\cdot \,|X_0 = x)$. The function v is called a *Lyapunov function*. Condition (i) guarantees that A is a so-called small set, and that X is strongly aperiodic; see Meyn and Tweedie [2009] for these definitions. Condition (ii) implies that X is geometrically ergodic. At a practical level, (A1) can be viewed as being a condition that covers many steady-state simulations. For example, all stable first-order autoregressive processes for which the noise terms admit an everywhere positive density satisfy (A1) (see p. 389

of Meyn and Tweedie [2009]), as does the $G/G/1$ waiting time sequence (i.e., a random walk on the half-line) with light-tailed input distributions (see pp. 399 and 400 of Meyn and Tweedie [2009]). (Here, light tailed means that the distributions have moment-generating functions that converge in a neighborhood of the origin.) Any such Markov chain automatically has a unique stationary distribution π. Furthermore, if

$$\|f\|_v \triangleq \sup_{x\in\mathbb{S}} \left\{ \frac{|f(x)|}{v(x)} \right\} < \infty,$$

then f is π-integrable, and

$$\begin{aligned} \alpha &= \pi f \\ &\triangleq \int_{\mathbb{S}} f(x)\pi(dx); \end{aligned}$$

see Chapter 15 of Meyn and Tweedie [2009].

For any such f, put $f_c(x) = f(x) - \alpha$. Then $E_x f_c(X_n)$ converges to zero geometrically fast as $n \to \infty$ and $\sum_{j=0}^{\infty} |E_x f_c(X_j)|$ is bounded by a multiple of $v(x)$; see p. 363 of Meyn and Tweedie [2009]. Letting

$$g(x) \triangleq \sum_{j=0}^{\infty} E_x f_c(X_j),$$

it is then easily verified that, under P_x, the process $M = (M_j : j \geq 0)$ for which

$$M_j \triangleq g(X_j) + \sum_{l=0}^{n-1} f_c(X_l)$$

is a martingale adapted to $(\mathcal{F}_j : j \geq 0)$ (where $\mathcal{F}_j$ is the σ-algebra generated by $X_0, \ldots, X_j$, and $M_0 \triangleq g(X_0)$).

2.1. The Distributional Initial Transient

There are two different perspectives that one can take in assessing the initial transient. The first is a *distributional* perspective, in which one seeks to identify a time d^* at which the Markov chain X is in appropriate equilibrium. The second is a *functional* perspective (treated in the next section), in which one wishes to find a time d_f^* at which Y is in approximate equilibrium. Clearly, we should expect to find that d_f^* is less than or equal to d^*.

The relation

$$\frac{g(x)}{n} = E_x \frac{1}{n} \sum_{j=0}^{n-1} f_c(X_j) + \frac{1}{n} \sum_{j=n}^{\infty} E_x f_c(X_j)$$

implies that

$$E_x \overline{Y}_n = \alpha + \frac{1}{n} g(x) + o\left(\frac{1}{n}\right)$$

as $n \to \infty$, where $o(a_n)$ is a deterministic sequence b_n for which $b_n/a_n \to 0$ as $n \to \infty$. The initialization bias is therefore determined by $g(x)$ to order n^{-1}. (In the simulation literature, this quantity is sometimes known as the asymptotic bias; see p. 391 of Whitt [2006].)

Suppose now that we initialize X with initial distribution μ so that $\mu(\cdot) \triangleq P(X_0 \in \cdot)$. Let

$$P_\mu(\cdot) \triangleq \int_{\mathbb{S}} \mu(dx) P_x(\cdot)$$

and

$$E_\mu(\cdot) \triangleq \int_{\mathbb{S}} \mu(dx) E_x(\cdot).$$

Assuming that $v(\cdot)$ is integrable with respect to μ, the dominated convergence theorem ensures that

$$E_\mu \overline{Y}_n = \alpha + \frac{1}{n} E_\mu g(X_0) + o\left(\frac{1}{n}\right)$$

as $n \to \infty$ and

$$E_\mu g(X_0) = \sum_{j=0}^{\infty} (E_\mu f(X_j) - \pi f),$$

suggesting that a natural metric for assessing the distributional effect of the initial transient associated with initialization under μ is to use the quantity

$$\beta(\mu) \triangleq \sup_{|\tilde{f}| \leq h} \left| \sum_{j=0}^{\infty} (E_\mu \tilde{f}(X_j) - \pi \tilde{f}) \right|$$

for some "envelope" function h. We refer to $\beta(\mu)$ as the (distributional) unconditional initial transient effect (UNITE) measure. In fact, $\beta(\mu)$ can be re-expressed in terms of the h-total variation norm on the space of measures defined on $\mathbb{S}$. In particular, for any (possibly signed) measure ν on $\mathbb{S}$, set

$$\|\nu\|_h = \sup_{\|\tilde{f}\|_h \leq 1} \left\{ \left| \int_{\mathbb{S}} \tilde{f}(x) \nu(dx) \right| \right\}.$$

Then, clearly

$$\beta(\mu) = \left\| \sum_{j=0}^{\infty} (P_\mu(X_j \in \cdot) - P_\pi(X_j \in \cdot)) \right\|_h .$$

As is to be expected intuitively, the measure μ that minimizes the distributional UNITE norm is $\mu = \pi$. To theoretically determine the value d^* discussed earlier, note that if the transient is assumed to end at time k, then it is natural to delete the data associated with periods 0 through $k - 1$. Starting data collection at time k is effectively equivalent to initializing the simulation with distribution μP^k, where

$$(\mu P^k)(\cdot) \triangleq P_\mu(X_k \in \cdot).$$

This suggests that one means of theoretically determining d^* is to select $d^* = d^*(\epsilon)$ so that $\beta(\mu P^{d^*}) < \epsilon$ for some prescribed error tolerance $\epsilon > 0$.

Remark. Set $e(x) = 1$ for all $x \in \mathbb{S}$. When $h = e$, it is easily seen that $\beta(\mu P^n)$ is nonincreasing in n. For other choices of h, there is no guarantee of monotonicity. This makes using the e-total variation norm especially natural.

Of course, if one initializes the simulation with π, one could be "unlucky" and randomly select an initial state X_0 that is far from stationarity, in which case the resulting measure of the initial transient effect is $\beta(\delta_{X_0})$, where δ_x is the unit point mass probability at x given by $\delta_x(B) = 1$ or 0 depending on whether or not $x \in B$. Thus, an alternative measure of the effect of the initialization with distribution μ is

$$\tilde{\beta}(\mu) = \int_{\mathbb{S}} \mu(dx)\beta(\delta_x).$$

Because it averages over the effect conditional on X_0, we call $\tilde{\beta}(\mu)$ the (distributional) conditional initial transient effect (CITE) measure. In contrast to UNITE, the initial distribution μ that minimizes $\tilde{\beta}(\mu)$ is δ_{x^*}, where x^* is the minimizer (assumed to exist) of $\beta(\delta_x)$ over $x \in \mathbb{S}$. We take the view (without proof) that any distribution μ for which $\tilde{\beta}(\mu)$ is within a (modest) constant multiple of $\tilde{\beta}(\pi)$ is likely to be a reasonable choice as an initial distribution, and probably offers performance comparable to that under initialization π, even in the absence of truncation.

2.2. The Functional Initial Transient

Turning now to the functional perspective, suppose that f is the performance measure underlying Y. By analogy with the preceding distributional discussion, it is natural to let the functional UNITE measure of the initial transient effect for a given μ be given by

$$\beta_f(\mu) = \left| \sum_{j=0}^{\infty} (E_\mu f(X_j) - \pi f) \right|,$$

and to let the functional CITE measure be determined as

$$\tilde{\beta}_f(\mu) = \int_{\mathbb{S}} \mu(dx)\beta_f(\delta_x).$$

The theoretical value d_f^* can then be selected as the smallest k for which $\beta_f(\mu P^k) < \epsilon$ for some given error tolerance $\epsilon > 0$. It is trivial to see that $d_f^* \leq d^*$.

In contrast to the distributional UNITE measure, there are no general guarantees that $\beta_f(\mu P^k)$ or $\tilde{\beta}_f(\mu P^k)$ are monotone in k (unless one is dealing with a stochastically monotone Markov chain and f is suitably monotone; see Bhattacharya et al. [2010]). From a mathematical standpoint, the distributional view of the initial transient problem is more natural, as it is better aligned with the recognition that the underlying dynamics of the system are governed by the distribution of the Markov chain (as opposed to that of a single functional f). However, most of the existing literature on the initial transient problem deals with the functional version of the initial transient problem. This likely has to do with the relative simplicity of building initial transient detection algorithms that focus on real-valued simulation output rather than on the much more complex underlying state of the associated GSMP.

Remark. In recent work, Wang and Glynn [2014] compute both the functional and distributional UNITE/CITE measures in the setting of one-dimensional reflected Brownian motion (RBM) and use these computations to identify "good" starting states for simulations of this process. Moreover, Wang and Glynn [2016] provide a rigorous characterization of the ϵ-mixing time of RBM. Indeed, RBM is mathematically tractable and describes the limiting behavior of many queueing systems in heavy traffic (for which the initial transient problem is relevant), and is therefore a natural theoretical model for analysis.

Remark. In many large-scale simulations, it is of interest to analyze the problem of initialization bias for multiple performance measures $\{f_l : l = 1, \ldots, L\}$. In particular, one might wish to estimate the corresponding steady state means

$$\alpha_l \overset{\Delta}{=} E_\pi f_l(X_0) = \int_{\mathbb{S}} f_l(x)\pi(dx)$$

for $l = 1, \ldots, L$. Given corresponding error tolerances $\{\epsilon_l\}$, a natural choice of theoretical truncation point would be the maximum of $d^*_{f_l}, l = 1, \ldots, L$. Alternatively, one could adopt (as an approximation) the distributional perspective and set the "envelope function" as the maximum of the $|f_l|$'s. This is also a potential way of identifying suitable envelope functions in practice (other than $e(x) \equiv 1$, say).

2.3. MSE: A Decomposition

In the functional setting, one popular approach to the theoretical analysis of the initial transient starts from the MSE of the truncated estimator $\overline{Y}_{n,k}$, given by $E(\overline{Y}_{n,k} - \alpha)^2$, where

$$\overline{Y}_{n,k} \overset{\Delta}{=} \frac{1}{n-k}\sum_{j=k}^{n-1} Y_j.$$

Under P_x, set

$$\begin{aligned} D_j &\overset{\Delta}{=} g(X_j) - E(g(X_j) \mid \mathcal{F}_{j-1}) \\ &= g(X_j) - E_x(g(X_j) \mid X_{j-1}) \\ &= M_j - M_{j-1} \end{aligned}$$

for $j \geq 1$, where the second equality follows from the Markov property.

THEOREM 1. *Assume (A1) and suppose that $\|f\|_{v^{1/2}} < \infty$. Then*

$$E_x(\overline{Y}_{n,k} - \alpha)^2 = \frac{\eta^2}{n} + \frac{w}{n^2} + \frac{k\eta^2}{n^2} + \frac{E_x r(X_k)}{n^2} + o\left(\frac{1}{n^2}\right)$$

as $n \to \infty$ for all $x \in \mathbb{S}$, where $\eta^2 = E_\pi D_1^2$,

$$w = E_\pi g^2(X_0) - 2\sum_{j=1}^{\infty} E_\pi D_1 g(X_j),$$

and

$$r(x) = g^2(x) + \sum_{j=1}^{\infty}\left(E_x D_j^2 - E_\pi D_1^2\right)$$

for all $x \in \mathbb{S}$. Furthermore, $E_x r(X_n) \to E_\pi r(X_0)$ for all $x \in \mathbb{S}$ as $n \to \infty$.

If the goal is to select the initial transient index k^* so as to minimize $E_\mu(\overline{Y}_{n,k} - \alpha)^2$, then Theorem 1 asserts that one can instead choose k^* so as to minimize $k\eta^2 + E_\mu r(X_k)$ over k (at least when n is large). Given that $E_\mu r(X_k)$ converges to $E_\pi r(X_0)$ and $k\eta^2$ is increasing linearly in k, this suggests that k^* will typically be small (unless the initial transient is strong). Furthermore, the exact value of k^* depends on determining $E_\mu r(X_k)$. Even if $r(\cdot)$ is known in closed form, this would be difficult. (However, see Franklin and White [2010] for a calculation of k^* when $r(\cdot)$ takes a specific parametric form.) Of course, in a simulation context, $r(\cdot)$ would need to be estimated from the simulated data, making the problem extremely challenging, particularly in view of the fact that in a single replication context, one has only a single realization available from which to estimate $r(\cdot)$. In fact, unless the Markov chain is such that x is visited infinitely often by X, $r(x)$ cannot be estimated consistently from a single run.

Theorem 1 also makes clear that, as for the conditional initial transient measure discussed earlier, the use of the MSE criterion in the single run initial transient setting has the characteristic that initializing the system in equilibrium is not the choice of initialization that minimizes the MSE of $\overline{Y}_n$. In particular, the initial distribution μ^* that minimizes the MSE of $\overline{Y}_n$ is (at least asymptotically in n) δ_{x^*}, where x^* is the minimizer (assumed to exist) of $r(\cdot)$. Again, it is reasonable to take the view that any initial distribution μ that leads to a MSE that is within a factor of 1 of that associated with π is likely to be an acceptable initialization, even in the absence of truncation.

2.4. When Does the Initial Transient Matter?

One immediate implication of Theorem 1 is that when n is large (as is needed for reasonable accuracy in most problems), the $r(x)$ term that carries the influence of the initial condition must be of order n in order that the $r(x)/n^2$ term be of the same magnitude as the variance term η^2/n. This suggests that it is only when the effect of the initial transient is large that deletion will have a substantial impact on MSE.

For a more complete argument, we consider the case in which X is an m-state reversible aperiodic irreducible Markov chain (with $m < \infty$) having transition matrix $P = (P(x, y) : x, y \in \mathbb{S})$. In such a setting, the stationary distribution π is positive, and the matrix A in which

$$A(x, y) = \sqrt{\frac{\pi(x)}{\pi(y)}} P(x, y)$$

is real and symmetric. Thus, A is diagonalizable with real eigenvalues $\lambda_1, \ldots, \lambda_m$ and corresponding orthogonal column eigenvectors $h_1, \ldots, h_m$ (e.g., see p. 136 of Axler [1997]). It is easy to see that $\lambda_1, \ldots, \lambda_m$ are eigenvalues of P, with corresponding column eigenvectors $w_1, \ldots, w_m$, where

$$w_j(x) \triangleq \frac{h_j(x)}{\sqrt{\pi(x)}}.$$

For any two column vectors ζ_1, ζ_2, consider the inner product defined by

$$\langle \zeta_1, \zeta_2 \rangle \triangleq \sum_{x \in \mathbb{S}} \pi(x) \zeta_1(x) \zeta_2(x)$$

and note that the orthogonality of $h_1, \ldots, h_m$ implies that $\langle w_i, w_j \rangle = 0$ for all $i \neq j$. It follows that any column vector f can be expressed as

$$f = \sum_{j=1}^{m} \frac{\langle f, w_j \rangle}{\langle w_j, w_j \rangle} w_j,$$

so

$$\begin{aligned} P^n f &= \sum_{j=1}^{m} \frac{\langle f, w_j \rangle}{\langle w_j, w_j \rangle} P^n w_j \\ &= \sum_{j=1}^{m} \lambda_j^n \frac{\langle f, w_j \rangle}{\langle w_j, w_j \rangle} w_j. \end{aligned}$$

Without loss of generality, one may assume that $1 = \lambda_1 > |\lambda_2| \geq \cdots \geq |\lambda_m|$. (Irreducibility implies that λ_2 cannot equal 1; aperiodicity implies that the (complex) modulus of λ_2 cannot equal 1; e.g., see p. 376 of Cinlar [1975].) Hence,

$$(P^n f)(x) - \alpha = \sum_{j=2}^{m} \lambda_j^n \frac{\langle f,\ w_j \rangle}{\langle w_j,\ w_j \rangle} w_j(x).$$

Under the additional assumption that $|\lambda_2| > |\lambda_3|$, if f has a nonzero projection onto w_2 with respect to the inner product $\langle \cdot\,,\ \cdot \rangle$, then

$$E_x f_c(X_n) \sim b(x)\lambda_2^n \tag{2}$$

as $n \to \infty$, where

$$b(x) = \frac{\langle f,\ w_2 \rangle}{\langle w_2,\ w_2 \rangle} w_2(x)$$

and $f_c(x) = f(x) - \pi f$ as before. Here, we write $a_n \sim b_n$ as $n \to \infty$ to mean that

$$\frac{a_n}{b_n} \to 1$$

as $n \to \infty$. If, in addition, $E_\pi f^2(X_0) > 0$, then

$$\frac{E_\pi f_c(X_0) f_c(X_n)}{E_\pi f_c(X_0)^2} \sim a\lambda_2^n \tag{3}$$

as $n \to \infty$, where

$$a \triangleq \frac{\langle f_c,\ b \rangle}{\langle f_c,\ f_c \rangle}.$$

Whenever $\lambda_2 \neq 0$, set $\tau = \log(1/|\lambda_2|)$. Relation (3) asserts that for the correlation between $f(X_j)$ and $f(X_{j+k})$ to be less than ϵ in equilibrium, k must be such that it is roughly equal to $(1/\tau)\log(|a|/\epsilon)$. Hence, for the simulation run to be long enough that it involves at least l ϵ-uncorrelated observations, $n \approx (l/\tau)\log(|a|/\epsilon)$ (where $\approx$ means "is approximately equal to"). Obviously, a high-accuracy simulation will require that l be large. On the other hand, (2) asserts that for the functional UNITE measure to be smaller than ϵ, we must set d_f^* to be of the order $(1/\tau)\log(|b(x)|/\epsilon)$ when the simulation is initialized with $\mu = \delta_x$. Thus, unless $|b(x)|$ is large relative to $|a|$, the influence of the initial transient (even if undeleted) tends to be small (at least for run-lengths n that are appropriate).

Remark. In some simulation studies, it will be the case that although initialization bias is small (or moderate), the duration of the initial transient is nonetheless significant. This would occur in "slowly mixing" systems in which $|\lambda_2|$ is close to 1 (e.g., a queue in heavy traffic). In such settings in which the transient is "persistent," the integrated bias over time (as reflected by the UNITE/CITE measures) can again be large, as would the function r appearing in the statement of Theorem 1. Therefore, we view a "large" initial transient both as one in which the instantaneous bias $|E_x f_c(X_n)|$ is large for n small, and one in which the transient is relatively smaller, but persistent.

3. THE MSER RULE AS THE SIMULATION TIME HORIZON TENDS TO INFINITY

A particular family of truncation rules known as marginal standard error rules (MSERs) has become popular in recent years. The literature on MSERs is fairly extensive. Letting $\hat{d}(n)$ denote the amount of truncation after having collected $Y_0, \ldots, Y_{n-1}$, the most prominent variants of MSER are listed next:

(1) The original MSER as proposed by White [1997]:

$$\hat{d}(n) = \arg\min_{0 \le k \le n-2} g_n(k),$$

where

$$g_n(k) \triangleq \frac{1}{(n-k)^2} \sum_{j=k}^{n-1} (Y_j - \overline{Y}_{n,k})^2.$$

(2) MSER-*m*, which deals with batches of simulation output (see White et al. [2000]). In particular, it implements MSER on the series given by

$$\hat{Z}_j = \frac{1}{m} \sum_{l=0}^{m-1} Y_{m(j-1)+l}, \ j = 1, \ldots, \left\lfloor \frac{n}{m} \right\rfloor,$$

where m is user specified (here, $\lfloor \cdot \rfloor$ denotes the floor function). A typical choice is $m = 5$; see Hoad and Robinson [2011] and Franklin and White [2008].
(3) MSER-*m* Overlapping, which is identical to MSER-*m* but with *overlapping* batches, as suggested in Pasupathy and Schmeiser [2010].
(4) MSER-LLM, which identifies the truncation point as the first local minimizer of $g_n(k)$.
(5) MSER-LLM2, which identifies the first local minimizer among local minimizers of $g_n(k)$; see Pasupathy and Schmeiser [2010] for a discussion.

If the Y_j's are discrete rv's, then the event

$$\{Y_{n-l} = Y_{n-l+1} = \cdots = Y_{n-1}\}$$

typically has positive probability (for every fixed value of l) so that $g_n(n-l) = 0$, in which case $n-l$ is, with positive probability, a minimizer of $g_n(k)$ as a function of k. When this occurs, $\hat{d}(n)$ provides no information about the duration of the initial transient. In view of this, works such as White et al. [2000] and Hoad et al. [2010] let the arg min range over $k = 0, \ldots, \lfloor n/2 \rfloor - 1$. A truncation value of $\lfloor n/2 \rfloor - 1$ is taken to mean that a longer simulation time horizon is required before making any conclusions.

Set $W_j \triangleq (Y_j - \alpha)^2$ for all $j \ge 0$. Additionally, to provide maximum algorithmic flexibility, we modify (slightly) the definition of MSER into

$$\hat{d}_\gamma(n) = \max \arg\min_{0 \le k \le \lfloor n - n^\gamma \rfloor} g_n(k)$$

for $0 < \gamma < 1$. In other words, we do not allow the arg min to occur in the final n^γ observations of the simulation output. This means that, asymptotically, almost the entire series of simulation output plays a role in the arg min (since $\gamma < 1$), and the anomalous behavior when the Y_j's are discrete is avoided (since $\gamma > 0$). If we wish to restrict to a range $0, 1, \ldots, \lfloor \delta^* n \rfloor$, where $0 < \delta^* < 1$, then the proof of Theorem 2 actually simplifies.

Our approximation to MSER for large n depends on the following assumption:

(A2) There exist $p > 4$ and deterministic constants α and σ^2 such that

$$C \triangleq \sup_{\substack{0 \le k \le n-1 \\ n \ge 1}} E\left(\left| \frac{1}{\sqrt{n-k}} \sum_{j=k}^{n-1} (Y_j - \alpha) \right|^p + \left| \frac{1}{\sqrt{n-k}} \sum_{j=k}^{n-1} (W_j - \sigma^2) \right|^p \right) < \infty.$$

Assumption (A2) is a modest condition in practice, as illustrated by the following proposition.

PROPOSITION 3.1. *Assume condition (A1) and* $p > 4$. *If*

$$\sup_{x\in\mathbb{S}} \frac{|f(x)|^{2p}}{v(x)} < \infty,$$

then (A2) holds under P_x *for each* $x \in \mathbb{S}$, *with*

$$\alpha = E_\pi f(X_0) \ \ \textit{and} \ \ \sigma^2 = E_\pi(f(X_0) - \alpha)^2.$$

We are now poised to state the main theorem of our article.

THEOREM 2. *Suppose that the stochastic process* Y *satisfies (A2). If* $\gamma \in (4/p, 1)$, *then*

$$\hat{d}_\gamma(n) \overset{a.s.}{\to} \arg\min_{k\geq 0} \left(\sum_{j=0}^{k-1} (2\sigma^2 - W_j) \right)$$

as $n \to \infty$, *provided that the* arg min *is a.s. unique.*

Remark. A sufficient condition for the arg min appearing on the right-hand side of Theorem 2 to be almost surely unique is that the W_i's be continuous rv's. Even if the arg min is not a.s. unique, the uniform convergence established in the proof of Theorem 2 guarantees that the a.s. limit points of $(\hat{d}_\gamma(n) : n \geq 0)$ will be contained in the set of minimizers of $S = (S_k : k \geq 0)$, where

$$S_k \triangleq 2k\sigma^2 - \sum_{j=0}^{k-1} W_j.$$

Remark. In the proof of Theorem 2, we argue that

$$\frac{1}{k} \sum_{j=0}^{k-1} W_j \overset{a.s.}{\to} \sigma^2$$

as $k \to \infty$. It thus follows that

$$\arg\min_{k\geq 0} \left(\sum_{j=0}^{k-1} (2\sigma^2 - W_j) \right) < \infty$$

a.s. This guarantees that MSER does not delete more and more data as the run-length n gets larger.

Remark: Our theorem also makes clear that one may need to assume more about Y (i.e. a larger value of p) if one wishes to inspect more of the simulated output sequence (i.e. a smaller value of γ).

While Theorem 2 is stated only for MSER, it also extends to some of the other variants. For example, if assumptions (A1) and (A2) hold for Y, then they also hold for the process $\hat{Z}$ that appears in MSER-*m*, as well as the process $\tilde{Z}$ defined by

$$\tilde{Z}_i = \frac{1}{m} \sum_{j=i}^{i+m-1} Y_j$$

that can be associated with MSER-m Overlapping. Theorem 2 then asserts, for example, that the MSER-m Overlapping deletion point converges a.s., as $n \to \infty$, to

$$\arg\min_{k\geq 0}\left(\sum_{j=0}^{k-1}(2E_\pi(\overline{Y}_m - E_\pi Y_0)^2 - (\tilde{Z}_j - E_\pi Y_0)^2\right)$$

(provided that the arg min is a.s. unique). The key point of this theoretical analysis is that it establishes that MSER is determined by the minimizer (assuming uniqueness) of the positive drift random walk $(S_j : j \geq 0)$. As we shall see later, this gives us insight into what settings will lead to failures to correctly identify the appropriate deletion points.

Remark. The focus of our article is on the behavior of MSER (and its variants) in asymptotic regimes for which n is large. Analyses of the small-sample regime may be found, for example, in Mokashi et al. [2010] and pages 520 through 522 of Law [2015]. When the simulation time horizon n is relatively small, failure often occurs in one of two noteworthy ways:

(1) MSER-5 fails to delete a significant amount of biased data;
(2) the minimizer of the usual MSER statistic exceeds $\lfloor n/2 \rfloor - 1$ so that the procedure declares that further simulation is needed.

From a practical standpoint, understanding the small-sample properties of MSER is as important as understanding its large-sample properties, especially in the early stages of a simulation study. However, these issues lie outside the scope of our article.

As a final point, we note that computing MSER from a simulation run of length n requires an additional number of floating point operations that is linear in n (say, an for some constant $a > 0$). Thus, for a given computational budget, one has a choice: implement MSER at a cost an or simulate $a'n$ additional steps of the Markov chain. It is clear, from Theorem 1, that simulating the additional steps is always a superior solution for (very) large n. However, as argued earlier, for any given value of n, we cannot a priori know the magnitude of the constants appearing in Theorem 1. In the presence of an unexpectedly large transient, the constants could be such that initial bias deletion, even at computing cost an, is a sensible strategy. Thus, the transient needs to be large enough that deleting bias reduces (for example) the MSE enough to "pay" for the additional computing cost of magnitude an. This is a further argument in favor of the view that initial bias deletion, in the single replication setting, is primarily a protective measure to deal with potentially large transients (rather than as a means of reducing MSE in the presence of small transients).

4. THE MSER RULE AND FLUID LIMITS FOR QUEUES

Models in which queues arise occur naturally in the discrete-event simulation setting. As such, they are useful vehicles to use as test beds for assessing the behavior of competing initial transient deletion methods.

One of the major developments within the queueing community over the past 20 years has been the introduction of fluid limits as a means of analyzing the stability of various complex queueing systems, particularly in the network setting (e.g., see Dai [1996]). Roughly speaking, for a stable queue, a fluid limit describes the behavior of the system as it empties out from an initial condition corresponding to having a substantial amount of work initially present in the system.

The time at which the fluid limit has drained all of the work for the system is then roughly the time at which one would expect the system to reach equilibrium. Thus, queueing networks for which fluid limits have been computed provide a rich class of models having the property that they exhibit substantial initial transients and the time at which they reach equilibrium is approximately computable. Of course, in view of our discussion in Section 2, focusing on the ability of deletion procedures to satisfactorily deal with strong transients seems appropriate.

We start by rigorously proving that the waiting time sequence of the $G/G/1$ single-server queue does indeed reach equilibrium at the emptying time predicted by the fluid limit. For this model, the waiting time sequence associated with initial condition x satisfies the recursion

$$W_{j+1}(x) = [W_j(x) + V_j - \chi_{j+1}]^+$$

for $j \geq 0$, where $V = (V_j : j \geq 0)$ is the sequence of independent and identically distributed service times, $\chi = (\chi_j : j \geq 1)$ is the sequence of independent and identically distributed interarrival times (independent of V), and $W_0(x) = x$. When $EV_0 < E\chi_1$, it is well known that $W_n(x) \Rightarrow W_\infty$ as $n \to \infty$ for each fixed $x \geq 0$. However, we can induce a strong transient in this model by sending $x \to \infty$.

The fluid limit associated with $W(x) = (W_j(x) : j \geq 0)$ starts from level x time units (say hours) of waiting in the queue and decreases linearly (and deterministically) at rate $\lambda^{-1} \triangleq E(\chi_1 - V_0)$ hours per customer departing the queue, emptying when the customer whose index is approximately λx departs the queue (e.g., see Anantharam [1988]). Thus, the conjecture discussed earlier, when specialized to the $G/G/1$ setting, asserts that $(W_j(x) : j \geq 0)$ reaches equilibrium roughly at time λx when x is large. Proposition 4.1 makes this claim rigorous.

PROPOSITION 4.1. *Suppose that $EV_0 < E\chi_1 < \infty$. Then*

$$\sup_B |P(W_{\lfloor tx \rfloor}(x) \in B) - P(W_\infty \in B)| \to \begin{cases} 1, & 0 < t < \lambda \\ 0, & t > \lambda \end{cases} \tag{4}$$

as $x \to \infty$.

We next proceed to show that when $Y_j = f(W_j)$ with $f(w) = w$, MSER identifies a truncation point that is also roughly of order λx when x is big, verifying that for this choice of f, MSER behaves sensibly.

PROPOSITION 4.2. *Suppose that $EV_0 < E\chi_1 < \infty$, and let V_0 be a light-tailed rv. Adopting the notation of Section 3, put*

$$g_n(k, x) = \frac{1}{(n-k)^2} \sum_{j=k}^{n-1} (W_j(x) - \overline{W}_{n,k}(x))^2$$

and

$$\hat{d}_\gamma(n, x) = \operatorname*{arg\,min}_{0 \leq k \leq \lfloor n - n^\gamma \rfloor} g_n(k, x)$$

(with $\gamma \in (0, 1)$). Moreover, suppose that the simulated time horizon $n = n(x)$ is such that

$$\frac{n(x)}{x} \to \beta > \lambda$$

as $x \to \infty$. Then

$$\frac{\hat{d}_\gamma(n(x), x)}{x} \xrightarrow{p} \lambda$$

as $x \to \infty$.

Table I. Exceedance Probabilities for $M/M/1$ Waiting Time Sequence: $n = 8{,}000$, $\mu_A = 0.8$, $\mu_S = 1$, $x = 100$, and $c = 2$

Opt. MSE	Est. Trunc.
0.00082	538

Remark. Since V_0 is assumed to be light tailed, $W(x)$ satisfies assumption (A1). In particular, $v(w) = \exp(\theta w)$ satisfies (A1) for some suitably chosen $\theta > 0$ (e.g., see p. 400 of Meyn and Tweedie [2009]). Consequently, Proposition 3.1 implies that (A2) holds for any function f that is polynomial in z. In particular, Proposition 3.1 applies to $W(x)$, and therefore Theorem 2 also applies to $W(x)$ for any fixed $x \geq 0$.

Remark. Although the use of fluid limits for testing initial transient procedures seems natural, we believe that the current article is the first to suggest this approach.

In the next section, we shall see that one or more variants of MSER can fail to identify the correct deletion point, even for a transient as "simple" as that associated with the $G/G/1$ queue initialized from a state x that is large. The problems arise when one chooses performance measures other than $f(w) = w$.

5. NUMERICAL RESULTS: EXAMPLES IN WHICH AT LEAST ONE VARIANT OF MSER FAILS

In this section, we combine the asymptotic insights of Sections 3 and 4 to construct two nonpathological and illustrative examples in which at least one variant of MSER fails to appropriately identify the duration of the initial transient. In the first, all variants of MSER fail; in the second, MSER-LLM and MSER-LLM2 fail, whereas the other variants still perform decently. The chief reason for failure in both cases is the observation that f (the real-valued performance measure) "masks" one or more key characteristics of the stochastic process X containing crucial information about the nature of the initial transient.

Our empirical investigations proceed as follows. We choose a simulation time horizon n sufficiently large (so that the initial transient has essentially washed out by time $n - 1$), and replicate 10,000 independent and identically distributed sample paths. We denote by $\overline{Y}^{(i)}_{n,k}$ the sample mean from the ith replication averaged over the observations collected from time k to time $n - 1$. For each value of k, we compute

$$\frac{1}{10{,}000} \sum_{i=1}^{10{,}000} \left(\overline{Y}^{(i)}_{n,k} - \alpha\right)^2.$$

This gives us an estimate of the MSE if we had chosen k as the deletion point. Next, we find the value $\tilde{d}$ that minimizes the empirical MSE as a function of k. With MSE as our criterion, this is an estimate of the best (deterministic) truncation value for k; we denote this MSE as "Opt. MSE" in Tables I and III. We choose to use MSE not necessarily because we view this as the best measure of a procedure's quality, but because it has been commonly used in the initial transient literature.

Similarly, we look at

$$\frac{1}{10{,}000} \sum_{i=1}^{10{,}000} \left(\overline{Y}^{(i)}_{n,\hat{d}^{(i)}} - \alpha\right)^2, \tag{5}$$

where the truncation index $\hat{d}^{(i)}$ is itself a function of the ith sample path and is computed via a particular variant of MSER (for MSER, the parameter γ is chosen to be 2/3; a similar convention is adopted for our implementations of MSER-5 and MSER-5

Table II. Exceedance Probabilities for $M/M/1$ Waiting Time Sequence: $n = 8{,}000$, $\mu_A = 0.8$, $\mu_S = 1$, $x = 100$, and $c = 2$

	Est. MSE	95% CI for MSE	Est. Trunc.
Zero Trunc.	0.00125	0.00125 ± 0.00003	0
MSER	0.00133	0.00133 ± 0.00009	3.03
MSER-LLM	0.00125	0.00125 ± 0.00003	0
MSER-LLM2	0.00439	0.00439 ± 0.00047	146
MSER-5	0.00125	0.00125 ± 0.00003	0
MSER-5 Overlapping	0.00133	0.00133 ± 0.00009	3.03

Overlapping). This is an estimate of the MSE

$$E(\overline{Y}_{n,\hat{d}} - \alpha)^2$$

achieved by the particular truncation method. The natural question becomes, how much larger is this estimated MSE compared to the estimated Opt. MSE for each of the different examples? Obviously, as n tends to infinity, we expect the ratio to converge to 1. However, for n that is only moderately larger than the "time to equilibrium," the ratio could be substantial, and no better (possibly worse) than the estimated zero truncation MSE (i.e., when one uses the entire simulation output; see row "Zero Trunc." in Tables II and IV), which we also include as a comparison point.

Remark. In Tables II and IV, MSE estimates for the various truncation procedures are displayed in the second column. The third column reports approximate 95% CIs for these MSEs, based on the normal approximation (i.e., the midpoint is the sample mean (namely, (5)) and the half-width is given by 1.96 (from the normal table) multiplied by the sample standard deviation (of the independent and identically distributed observations $((\overline{Y}^{(i)}_{n,\hat{d}^{(i)}} - \alpha)^2 : 1 \le i \le 10{,}000))$ divided by the square root of the number of simulations (in this case, 10,000)). Finally, the fourth column displays the average truncation values.

5.1. Example: $G/G/1$ Exceedance Probabilities

Recall that according to Theorem 2, MSER approximately truncates according to the location of the minimizer of the random walk $(S_j : j \ge 0)$. Suppose that we consider the sequence $(W_j(x) : j \ge 0)$ with $f(w) = I(w > c)$ for some appropriately chosen value of c (where $I(B)$ denotes the indicator of a set B). Then, $\alpha = P_\pi(W_0 > c)$ and $\sigma^2 = \alpha(1-\alpha)$. If we start from a value of x much greater than c, then $(W_j(x) : j \ge 0)$ will experience a long period of time before entering the interval $[0, c]$, during which $Y_i \equiv 1$. For any d less than this entry time, it follows that

$$S_j = j(2\alpha(1-\alpha) - (1-\alpha)^2).$$

Thus, if we select c so that $2\alpha(1-\alpha) - (1-\alpha)^2 > 0$ (i.e., $\alpha > 1/3$), the simulation will therefore start with a long period of time during which $(S_j : j \ge 0)$ increases linearly, creating a situation where MSER and its variants will often identify 0 as the minimizer of the random walk (and hence 0 as the deletion time). Since this type of performance measure f arises in many practical settings, this example is a potential concern for MSER.

For our $G/G/1$ example, we specialize to the case where the interarrival and service times are exponentially distributed with rate parameters μ_A and μ_S, respectively (so that the queue is an $M/M/1$ system), in which case

$$P_\pi(W_0 > c) = \frac{\mu_A}{\mu_S} \exp((\mu_A - \mu_S)c).$$

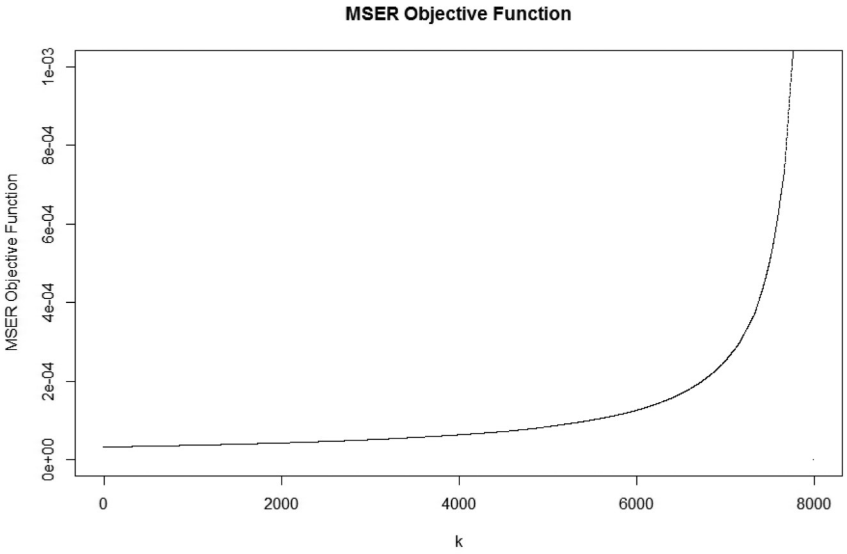

Fig. 1. Exceedance probabilities for $M/M/1$ waiting time sequence: plot of MSER objective function for a particular sample path.

We specifically set $\mu_A = 0.8$, $\mu_S = 1$, $x = 100$, and $c = 2$ so that $\alpha > 1/3$. Our fluid scale theory suggests that the theoretically optimal truncation value should be the order of 400 (although for any finite value of x, the transient is expected to persist somewhat longer).

For the realized sample paths in our simulation, indeed most truncation values (for all MSER variants) were equal to 0 (which is far smaller than the estimated theoretically optimal truncation value), in agreement with the aforementioned theory inspired by random walk insights. The few nonzero deletion points occurred near the maximum allowable index for the arg min (in the definition of each MSER variant), leading to nonzero estimates for the truncations. This alternative form of failure is attributed to the fact that (potentially a significant number of) consecutive observations near the end of the simulation output can be identically equal to 1, resulting in consecutive indices at which the MSER objective function is identically 0. It is worth noting, however, that all MSER-LLM and MSER-5 truncation values were equal to 0.

The results of the numerical study are not surprising, given the typical behavior of $g_n(k)$ (MSER objective function) over $0 \le k \le n-1$, a plot of which (for one particular sample path) is given in Figure 1.

In sum, we recall from Section 4 that when initialized at a high initial condition, the $G/G/1$ waiting time sequence (for $\rho < 1$) decreases linearly (in fluid scale) until the first time it reaches state 0 and equilibrates afterward. In particular, it exemplifies "strictly monotone emptying" as the initial transient is washed out. Due to the fact that initial transient is so distinctly differently from steady-state behavior, MSER behaves appropriately (whenever $f(w) = w$) and asymptotically truncates the theoretically correct amount. However, when f is chosen to be an indicator function, the attractive monotonicity property is lost, and thus MSER fails to distinguish the initial transient from steady-state behavior over a wide range of parameter values.

5.2. Example: A Polling System

Our first example was such that the stochastic system was both one dimensional and stochastically monotone as the initial transient is washed out. It is natural to

Table III. Polling System with Two Stations and Performance Measure $f(w_1, w_2) = w_1$: $n = 8{,}000$, $\mu_A = 0.8$, $\mu_S = 2$, and $x = 100$ at Both Stations

Opt. MSE	Est. Trunc.
0.00904	348

Table IV. Polling System with Two Stations and Performance Measure $f(w_1, w_2) = w_1$: $n = 8{,}000$, $\mu_A = 0.8$, $\mu_S = 2$, and $x = 100$ at Both Stations

	Est. MSE	95% CI for MSE	Est. Trunc.
Zero Trunc.	1.55	1.55 ± 0.0151	0
MSER	0.00933	0.00933 ± 0.00028	277
MSER-LLM	0.406	0.406 ± 0.00617	96.4
MSER-LLM2	0.0365	0.0365 ± 0.00049	229
MSER-5	0.00915	0.00915 ± 0.00030	308
MSER-5 Overlapping	0.00922	0.00922 ± 0.00029	303

subsequently investigate how MSER (and its variants) perform on a high-dimensional system not exhibiting any monotonicity during the course of the initial transient.

To this end, consider a polling network in which one has a single server that visits two stations in cyclic order. Customers are arriving to each of the stations according to independent Poisson processes (for simplicity, assume the same arrival rate μ_A for each). When the server (with service rate $\mu_S > 2\mu_A$) arrives at a given station, the server serves all customers waiting at that station plus arriving customers until the server has no remaining work ("exhaustive service" queueing discipline) at that station. The server subsequently travels to the other station and follows the same protocol (assuming zero traveling time and that the server will not switch again on arriving at an empty station until it has conducted some service). An analysis of such systems (including formulas for the steady-state expected waiting times) may be found, for example, in Avi-Itzhak et al. [1965] and Winands et al. [2006].

Now, suppose that both stations are equipped with the same enormous initial workload x, and suppose that we wish to estimate the steady-state waiting time (exclusive of service) at the first station. Note that the sequence of waiting times at station one depends on the number of customers x_1 and x_2 present in the system at the start of the simulation; we make clear this dependence by writing $W_j(x_1, x_2)$ for the waiting time of customer j to arrive at station one. The fluid limit for the waiting time sequence $(W_j(x, x) : j \geq 0)$ at the first station will then involve piecewise linear oscillatory behavior as the system empties. In particular, if d' indexes the last customer in station one to enter service before station one empties for the first time, then the waiting time for customer $d' + 1$ at station one will be at least as large as the time it takes for station two to empty for the first time (which will be long because x is large). It is therefore clear that the optimal truncation point will be far to the right of d'.

We now again utilize the insight associated with Theorem 2. In particular, we observe that for indices $j < d'$ and sufficiently close to d', $2\sigma^2 - (Y_j - \alpha)^2 > 0$ with high probability. On the other hand, $2\sigma^2 - (Y_{d'+1} - \alpha)^2$ is almost inevitably negative. This suggests that the random walk $(S_j : j \geq 0)$ will, with high probability, have a local minimizer to the left of d', leading to a failure of MSER-LLM.

For our empirical investigation, we choose $\mu_A = 0.8$, $\mu_S = 2$, and $x = 100$ at both stations. The corresponding value of $\alpha = E_\pi W_0$ is 2.5. In Table IV, we see that MSER-LLM indeed truncates less than the desired order of magnitude due to the presence of local minimizers in $g_n(k)$, a plot of which for a particular sample path is also provided

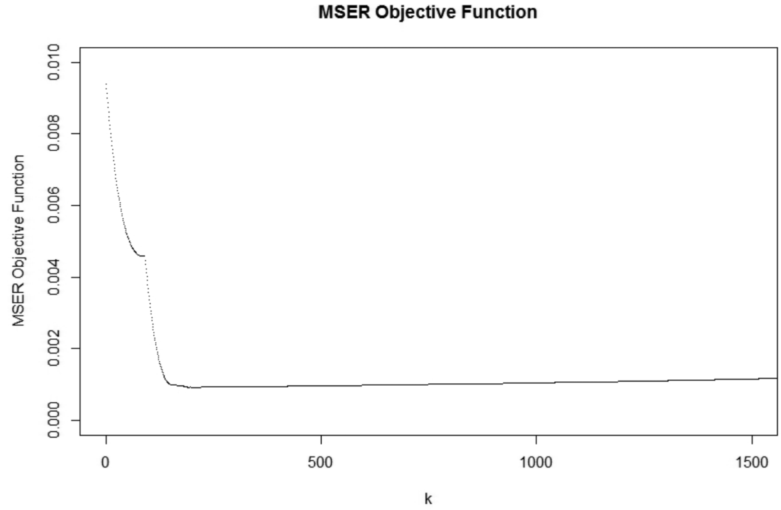

Fig. 2. Polling system with two stations and performance measure $f(w_1,\ w_2) = w_1$: plot of MSER objective function for a particular sample path.

in Figure 2. MSER-LLM2 performs better but also fails to remove a small portion of the initial transient (for the same reason as MSER-LLM).

As expected, MSER, MSER-5, and MSER-5 Overlapping continue to truncate an appropriate amount, as despite the lack of monotonicity, the initial transient is still sufficiently distinct from steady-state behavior. In sum, this example reveals that the local versions of MSER are less reliable whenever the dynamics of the initial transient induce local minimizers in the MSER objective function during early portions of the simulation.

5.3. Discussion

Much more complex fluid behavior than that exhibited in the polling context can occur in the setting of multiclass queueing networks; see Bramson [2008] for an extensive overview. As noted earlier, these models have the nice analytical property that there is an easily computable (and relatively unambiguous) time at which such systems reach equilibrium (assuming our conjecture is correct). Thus, these multiclass models (and their related transients) are excellent candidates for the testing of initial transient deletion procedures.

6. CONCLUSION

The theoretical and computational investigation provided in this article support the following conclusions:

(1) There are significant differences between the distributional initial transient and functional initial transient problems. These differences are clearly “masked” when one simulates one-dimensional Markov chains and selects the performance measure $f(x) = x$.
(2) Because typical discrete-event simulations usually induce high-dimensional Markovian state spaces, with associated functions f that are not bijections, the set of test problems used to compare deletion procedures should be broadened to include many examples of output series Y that are not themselves Markov chains.

(3) In the single replication setting, the implementation of an initial transient deletion procedure is largely supported by a desire to protect oneself against the possible existence of an unexpectedly large transient. As such, it is reasonable to test single replication procedures against test problems in which the transient is large.
(4) When the simulation time horizon n is large, the behavior of the MSER initial transient deletion procedure (and its variants) is closely related to that of an associated random walk process. Understanding the behavior of a truncation procedure as $n \to \infty$ can give rise to new insights regarding algorithmic performance/robustness.
(5) Queueing models, in which fluid limits have been computed, form an attractive class of high-dimensional models, exhibiting potentially complex behavior, for which we can reasonably conjecture the "correct" time at which equilibrium occurs.

APPENDIX: PROOFS

Proof of Theorem 1. In view of condition (A1), we can apply Theorem 15.0.1 Equation (15.4) of Meyn and Tweedie [2009] to conclude that there exists $c_1 < \infty$ and $r_1 > 1$ such that

$$\sum_{j=0}^{\infty} r_1^j \sup_{\|\tilde{u}\|_v \leq 1} |E_x \tilde{u}(X_j) - \pi \tilde{u}| \leq c_1 v(x) \tag{6}$$

for $x \in \mathbb{S}$. Note that if $\|u\|_v < \infty$, then $\tilde{u} = u / \|u\|_v$ satisfies $\|\tilde{u}\|_v \leq 1$. Consequently, (6) implies that

$$\sum_{j=0}^{\infty} r_1^j |E_x u(X_j) - \pi u| \leq c_1 v(x) \|u\|_v \tag{7}$$

whenever $\|u\|_v < \infty$.

We now invoke the Cauchy-Schwarz inequality to conclude that

$$\left(E_x v^{1/2}(X_1)\right)^2 \leq E_x v(X_1).$$

Thus, condition (A1) ensures that

$$\begin{aligned} E_x v^{1/2}(X_1) &\leq \sqrt{(1-\kappa)v(x) + cI(x \in A)} \\ &\leq (1-\kappa)^{1/2} v^{1/2}(x) + \sqrt{c} I(x \in A) \end{aligned} \tag{8}$$

for $x \in \mathbb{S}$, where we have used the fact that $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for $a,\ b \geq 0$. In view of (8), we can now separately apply Theorem 15.0.1 of Meyn and Tweedie [2009] with $v^{1/2}$ playing the role of v, thereby yielding the existence of $r_2 > 1$ and $c_2 < \infty$ such that

$$\sum_{j=0}^{\infty} r_2^j |E_x u(X_j) - \pi u| \leq c_2 v^{1/2}(x) \|u\|_{v^{1/2}} \tag{9}$$

whenever $\|u\|_{v^{1/2}} < \infty$.

Given (7), we find that if $\|u\|_v < \infty$, then

$$|E_x u(X_n)| \leq |\pi u| + r_1^{-n} c_1 v(x) \|u\|_v . \tag{10}$$

Furthermore, if we set

$$(\Gamma u)(x) \triangleq \sum_{j=0}^{\infty} (E_x u(X_j) - \pi u),$$

we note that (7) implies that Γu is finite valued when $\|u\|_v < \infty$, and

$$\begin{aligned}|E_x(\Gamma u)(X_n)| &= \left|\sum_{j=n}^{\infty}(E_x u(X_j) - \pi u)\right| \\ &\le \sum_{j=n}^{\infty} |E_x u(X_j) - \pi u|.\end{aligned}$$

Since $r_1^j / r_1^n \ge 1$ for $j \ge n$, it follows that

$$\begin{aligned}|E_x(\Gamma u)(X_n)| &\le r_1^{-n} \sum_{j=n}^{\infty} r_1^j |E_x u(X_j) - \pi u| \\ &\le r_1^{-n} \sum_{j=0}^{\infty} r_1^j |E_x u(X_j) - \pi u| \\ &\le r_1^{-n} c_1 v(x) \|u\|_v . \qquad (11)\end{aligned}$$

Similarly, if $\|u\|_{v^{1/2}} < \infty$, then

$$|E_x(\Gamma u)(X_n)| \le r_2^{-n} c_2 v(x)^{1/2} \|u\|_{v^{1/2}} . \qquad (12)$$

Because $\|f\|_{v^{1/2}} < \infty$, it follows that

$$\begin{aligned}|E_x g(X_n)| &= |E_x(\Gamma f)(X_n)| \\ &\le r_2^{-n} c_2 v(x)^{1/2} \|f\|_{v^{1/2}} . \qquad (13)\end{aligned}$$

Setting $n = 0$ in (13), we conclude that $\|g\|_{v^{1/2}} < \infty$. This immediately also implies that $\|g^2\|_v < \infty$. The inequality (13) therefore yields the conclusion that $E_x g^2(X_n) < \infty$ (and $E_x f^2(X_n) < \infty$) for all $x \in \mathbb{S}$ so that the process $(M_j : j \ge 0)$ for which

$$M_j = g(X_j) + \sum_{l=0}^{n-1} f_c(X_l)$$

is a square-integrable martingale under P_x. Since $E_x D_i D_j = 0$ for $i \ne j$ and $E_x g(X_k) D_j = 0$ for all $j > k$,

$$\begin{aligned}E_x\left(\sum_{j=k}^{n-1} f_c(X_j)\right)^2 &= E_x\left(\sum_{j=k+1}^{n} D_j + g(X_k) - g(X_n)\right)^2 \\ &= (n-k)\eta^2 + \sum_{j=k+1}^{n}\left(E_x D_j^2 - E_\pi D_1^2\right) + E_x g^2(X_k) + E_x g^2(X_n) \\ &\quad - 2E_x g(X_n) \sum_{j=k+1}^{n} D_j - 2E_x g(X_k) g(X_n). \qquad (14)\end{aligned}$$

Starting with the last term in (14), we observe that (13) with $n = 0$ yields the inequality

$$\begin{aligned}|E_x g(X_k) g(X_n)| &= |E_x g(X_k) E(g(X_n)| X_k)| \\ &\le E_x |g(X_k)| |E(g(X_n| X_k)| \\ &\le r_2^{-(n-k)} c_2 E_x |g(X_k)| v(X_k)^{1/2} \|f\|_{v^{1/2}} \\ &\le r_2^{-(n-k)} c_2^2 E_x v(X_k) \|f\|_{v^{1/2}}^2 .\end{aligned}$$

Applying (10) with $u = v$ (and recalling that $v \geq 1$), we conclude that there exists $c_3 < \infty$ such that

$$|E_x g(X_k) g(X_n)| \leq r_2^{-n} c_3 v(x). \tag{15}$$

To handle the term involving the D_j's, we set

$$\begin{aligned} u_1(x) &= E_x D_1^2 \\ &= \mathrm{var}(D_1 | \, X_0 = x) \\ &= E_x g^2(X_1) - (E_x g(X_1))^2 \end{aligned}$$

so that (13) yields the bound

$$\begin{aligned} |u_1(x)| &\leq E_x g^2(X_1) \\ &\leq c_2^2 E_x v(X_1) \, \| f \|^2_{v^{1/2}} \, . \end{aligned} \tag{16}$$

Again, (10) implies that there exists a constant $c_4 < \infty$ for which

$$|u_1(x)| \leq c_4 v(x) \tag{17}$$

for $x \in \mathbb{S}$. Then,

$$\begin{aligned} \sum_{j=k+1}^{n} \left(E_x D_j^2 - E_\pi D_1^2 \right) + E_x g^2(X_k) &= \sum_{j=k}^{n-1} \left(E_x u_1(X_j) - \pi u_1 \right) + E_x g^2(X_k) \\ &= E_x (\Gamma u_1)(X_k) + E_x g^2(X_k) - E_x (\Gamma u_1)(X_n) \\ &= E_x r(X_k) - E_x (\Gamma u_1)(X_n). \end{aligned} \tag{18}$$

Inequalities (12) and (16) then yield

$$|E_x (\Gamma u_1)(X_n)| \leq r_1^{-n} c_1 c_4 v(x) \tag{19}$$

for $x \in \mathbb{S}$.

For the term involving the $g(X_n) D_j$'s, note that for $j \leq n$,

$$\begin{aligned} E_x D_j u(X_n) &= E_x D_j E(u(X_n) | \, X_{j-1}, \, X_j) \\ &= E_x D_j E(u(X_n) | \, X_j) \\ &= E_x u'_{n-j}(X_{j-1}), \end{aligned}$$

where $u'_j(x) \overset{\triangle}{=} E_x D_1 g(X_j)$ for $j \geq 1$. But results (13) and (17) together with the Cauchy-Schwarz inequality imply that

$$\begin{aligned} |u'_j(x)| &= |E_x D_1 g(X_j)| \\ &= |E_x D_1 E(g(X_j) | \, X_1)| \\ &\leq r_2^{-(j-1)} c_2 \, \| f \|_{v^{1/2}} \, E_x |D_1| v(X_1)^{1/2} \\ &\leq r_2^{-(j-1)} c_2 \, \| f \|_{v^{1/2}} \sqrt{E_x D_1^2} \sqrt{E_x v(X_1)} \\ &= r_2^{-(j-1)} c_2 \, \| f \|_{v^{1/2}} \sqrt{u_1(x)} \sqrt{E_x v(X_1)} \\ &\leq r_2^{-(j-1)} c_2^2 \, \| f \|^2_{v^{1/2}} \, E_x v(X_1). \end{aligned}$$

Inequality (10) therefore proves that there exists a constant $c_5 < \infty$ independent of j such that

$$|u'_j(x)| \leq c_5 r_2^{-j} v(x) \tag{20}$$

for $j \geq 1$. In addition, Inequality (10), as applied to $u'_j(x) - \pi u_j$, establishes that

$$\begin{aligned}|E_x u'_j(X_i) - \pi u'_j| &\leq r_1^{-i} c_1 v(x) \left\| u'_j \right\|_v \\ &\leq r_1^{-i} c_1 c_5 v(x) r_2^{-j}.\end{aligned} \tag{21}$$

We now observe that

$$\begin{aligned}\sum_{j=k+1}^{n} E_x D_j g(X_n) &= \sum_{j=k+1}^{n} E_x u'_{n-j+1}(X_{j-1}) \\ &= \sum_{l=1}^{n-k} E_x u'_l(X_{n-l})\end{aligned}$$

so that (21) yields the inequality

$$\begin{aligned}\left| \sum_{j=k+1}^{n} E_x D_j g(X_n) - \sum_{l=1}^{n-k} \pi u'_l \right| &\leq \sum_{l=1}^{n-k} r_1^{-n+l} c_1 c_5 v(x) r_2^{-l} \\ &= r_1^{-n} c_1 c_5 v(x) \sum_{l=1}^{n-k} \left(\frac{r_1}{r_2} \right)^l .\end{aligned} \tag{22}$$

If $r_1 < r_2$, then the right-hand side of (22) is of order r_1^{-n}, whereas if $r_1 > r_2$, then the right-hand side is of order r_2^{-n}. If $r_1 = r_2$, then the right-hand side is of order nr_1^{-n}. In all cases, the right-hand side converges to 0 geometrically in n.

On the other hand, result (20) implies that $|\pi u'_j| \leq c_5 r_2^{-j} \pi v$ so that

$$\sum_{l=n-k+1}^{\infty} \pi u'_l = o\left(\frac{1}{n^2} \right) \tag{23}$$

as $n \to \infty$. Additionally, in view of (10) and (13), we see that

$$\begin{aligned}|E_x g^2(X_k) - \pi g| &\leq r_1^{-k} c_1 \|g^2\|_v v(x) \\ &\leq r_1^{-k} c_1 c_2^2 \, \| f \|_{v^{1/2}}^2 \, v(x).\end{aligned} \tag{24}$$

Thus, results (22), (23), and (24) together imply that

$$\begin{aligned}E_x g^2(X_n) - 2E_x g(X_n) \sum_{j=k+1}^{n} D_j &= \pi g^2 - 2 \sum_{l=1}^{\infty} \pi u'_l + o\left(\frac{1}{n^2} \right) \\ &= w + o\left(\frac{1}{n^2} \right)\end{aligned} \tag{25}$$

as $n \to \infty$.

Combining results (14), (15), (18), (19), and (25) yields the theorem. □

Proof of Proposition 3.1. Let v be as in assumption (A1), and observe that Lyapunov's inequality implies that

$$\begin{aligned}E_x v(X_1)^{1/p} &\leq (E_x((v(X_1)^{1/p})^p))^{1/p} \\ &= (E_x v(X_1))^{1/p}.\end{aligned} \tag{26}$$

Furthermore, it is easily verified that

$$w(z) \overset{\Delta}{=} (1 + z)^{1/p} - 1 - z^{1/p}$$

is nonincreasing on $[0, \infty)$ (since $w'(z) \le 0$ for all $z \ge 0$) so that $(1+z)^{1/p} \le 1 + z^{1/p}$ for $z \ge 0$. Using (A1) and (26) then yields

$$\begin{aligned} E_x v^{1/p}(X_1) &\le ((1-\kappa)v(x) + cI(x \in A))^{1/p} \\ &= (1-\kappa)^{1/p} v^{1/p}(x) \left(1 + \frac{cI(x \in A)}{(1-\kappa)v(x)}\right)^{\frac{1}{p}} \\ &\le (1-\kappa)^{1/p} v^{1/p}(x) + c^{1/p} I(x \in A). \end{aligned}$$

We now apply Theorem 15.0.1 of Meyn and Tweedie [2009] with $v^{1/p}$ playing the role of v. Since $\|f\|_{v^{1/p}} = (\|f^p\|_v)^{1/p} < \infty$, we can apply the same argument as in the proof of Theorem 1 to conclude that $g(x) = \sum_{j=0}^{\infty} E_x f_c(X_j)$ is such that

$$|E_x g(X_n)| \le c_6 r_3^{-n} \|f\|_{v^{1/p}} v^{1/p}(x) \tag{27}$$

for $x \in \mathbb{S}$ for some $c_6 < \infty$ and $r_3 > 1$. As in the proof of Theorem 1, the sequence $(M_j : j \ge 0)$ is then a P_x square-integrable martingale, and

$$E_x \left| \frac{1}{\sqrt{n-k}} \sum_{j=k}^{n-1} f_c(X_j) \right|^p = \frac{1}{(n-k)^{\frac{p}{2}}} E_x \left| \sum_{j=k+1}^{n} D_j + g(X_k) - g(X_n) \right|^p . \tag{28}$$

We now recall that, since the mapping $y \mapsto y^r$ is convex on $[0, \infty)$ for $r \ge 1$, it is evident that

$$\left(\frac{\xi_1 + \cdots + \xi_m}{m} \right)^r \le \frac{1}{m} \sum_{j=1}^{m} \xi_i^r$$

for any nonnegative rv's $\xi_1, \ldots, \xi_m$. Hence,

$$E_x(\xi_1 + \cdots + \xi_m)^r \le m^{r-1} \sum_{j=1}^{m} E_x \xi_j^m \tag{29}$$

for $r \ge 1$. It follows from (28) that

$$E_x \left| \frac{1}{\sqrt{n-k}} \sum_{j=k}^{n-1} f_c(X_j) \right|^p \le \frac{3^{p-1}}{(n-k)^{\frac{p}{2}}} \left(E_x \left| \sum_{j=k+1}^{n} D_j \right|^p + E_x|g(X_k)|^p + E_x|g(X_n)|^p \right). \tag{30}$$

We now apply the Burkholder-Davis-Gundy inequality (e.g., see p. 482 of DasGupta [2011]) to assert the existence of a constant $c_7 < \infty$ for which

$$E_x \left| \sum_{j=k+1}^{n} D_j \right|^p \le c_7 E_x \left(\sum_{j=k+1}^{n} D_j^2 \right)^{p/2} . \tag{31}$$

Another application of (29) yields

$$\begin{aligned} \frac{1}{(n-k)^{\frac{p}{2}}} E_x \left(\sum_{j=k+1}^{n} D_j^2 \right)^{p/2} &\le \frac{1}{(n-k)^{\frac{p}{2}}} \cdot (n-k)^{\frac{p}{2}-1} \sum_{j=k+1}^{n} E_x \left(D_j^2\right)^{p/2} \\ &= \frac{1}{n-k} \sum_{j=k+1}^{n} E_x |D_j|^p. \end{aligned} \tag{32}$$

Combining results (28), (30), (31), and (32) and recalling the bound (27), it is clear that

$$E_x \left| \frac{1}{\sqrt{n-k}} \sum_{j=k}^{n-1} f_c(X_j) \right|^p$$

is uniformly bounded in $k < n < \infty$ provided that $\sup_{j \geq 0} E_x v(X_j) < \infty$. But this follows from applying Theorem 15.0.1 of Meyn and Tweedie [2009] to v itself (and setting $u = v$ in (10)).

To prove the comparable result for

$$E_x \left| \frac{1}{\sqrt{n-k}} \sum_{j=k}^{n-1} (f_c(X_j)^2 - \sigma^2) \right|^p,$$

we note that $\|h^p\|_v < \infty$ for $h = f_c^2$ and apply the preceding argument for f_c to h. □

PROOF OF THEOREM 2. We begin by observing that

$$\begin{aligned} g_n(k) &= \frac{1}{(n-k)^2} \sum_{j=k}^{n-1} (Y_j - \alpha + \alpha - \overline{Y}_{n,k})^2 \\ &= \frac{1}{(n-k)^2} \sum_{j=k}^{n-1} W_j - \frac{(\overline{Y}_{n,k} - \alpha)^2}{n-k}. \end{aligned} \quad (33)$$

Set $m = m_n = \lfloor n^\gamma \rfloor$, where $\lfloor x \rfloor$ is the greatest integer less than or equal to x. We further put $q = (\gamma p - 4)/(4p)$ and note that since $p > 4$ and $\gamma \in (4/p, 1)$, it follows that $q \in (0, 1)$. We next recognize that (A2) and Markov's inequality imply that

$$\begin{aligned} \sum_{n=1}^{\infty} \sum_{k=0}^{n-m} P\big(n^q |\overline{Y}_{n,k} - \alpha| > \epsilon\big) &\leq \sum_{n=1}^{\infty} \sum_{k=0}^{n-m} \epsilon^{-p} n^{qp} (n-k)^{-p/2} E \left| \frac{\sum_{j=k}^{n-1} (Y_j - \alpha)}{\sqrt{n-k}} \right|^p \\ &\leq C\epsilon^{-p} \sum_{n=1}^{\infty} \sum_{k=0}^{n-m} n^{qp} n^{-\gamma p/2} \\ &\leq C\epsilon^{-p} \sum_{n=1}^{\infty} n^{1+p(q-\gamma/2)} \\ &= C\epsilon^{-p} \sum_{n=1}^{\infty} n^{-p\gamma/4} \\ &< \infty, \end{aligned}$$

due to our choice of q and the fact that $\gamma > 4/p$. It follows that for each rational value $\epsilon > 0$, the event

$$\max_{0 \leq k \leq n-m} |\overline{Y}_{n,k} - \alpha| > \epsilon n^{-q}$$

occurs only finitely often (in n) a.s., and hence

$$n^q \max_{0 \leq k \leq n-m} |\overline{Y}_{n,k} - \alpha| \overset{a.s.}{\to} 0$$

under (A2). In other words,

$$\max_{0\le k\le n-m} |\overline{Y}_{n,k} - \alpha| = o(n^{-q}) \tag{34}$$

a.s. as $n \to \infty$, where $o(f(n))$ denotes a sequence of rv's $(\beta_j : j \ge 0)$ such that $\beta_n/f(n) \overset{a.s.}{\to} 0$ as $n \to \infty$. Similarly,

$$\max_{0\le k\le n-m} \left| \frac{1}{n-k} \sum_{j=k}^{n-1} W_j - \sigma^2 \right| = o(n^{-q}) \tag{35}$$

a.s. as $n \to \infty$.

Relations (33), (34), and (35) imply that, uniformly in $k \in \{\lfloor n^{q/2} \rfloor, \dots, n-m\}$,

$$\begin{aligned} g_n(k) &= \frac{1}{n-k}(\sigma^2 + o(n^{-q})) - \frac{o(n^{-2q})}{n-k} \\ &= \frac{1}{n-k}(\sigma^2 + o(n^{-q})) \end{aligned}$$

a.s. so that

$$\frac{g_n(k)}{g_n(0)} = \frac{(\sigma^2 + o(n^{-q}))}{(\sigma^2 + o(n^{-q}))} \left(\frac{n}{n-k} \right).$$

Then,

$$\begin{aligned} \frac{g_n(k)}{g_n(0)} &\ge \frac{(\sigma^2 - (\sigma^2/4)n^{-q})}{(\sigma^2 + (\sigma^2/4)n^{-q})} \left(\frac{1}{1-k/n} \right) \\ &= \left(1 - \frac{1}{2} n^{-q} + o(n^{-q}) \right) \left(\frac{1}{1-k/n} \right) \end{aligned}$$

for n sufficiently large so that a.s.

$$\min_{n^{q/2}\le k\le n-m} \frac{g_n(k)}{g_n(0)} > 1$$

for n large enough. Since $g_n(0)$ is then smaller than $g_n(k)$ for $n^{q/2} \le k \le n-m$, we may conclude that a.s.,

$$\begin{aligned} \arg\min_{0\le k\le n-m} g_n(k) &= \arg\min_{0\le k\le n^{q/2}} g_n(k) \\ &= \arg\min_{0\le k\le n^{q/2}} (g_n(k) - g_n(0)) \end{aligned} \tag{36}$$

for n sufficiently large.

Additionally, for $k \le n^{q/2}$,

$$g_n(k) - g_n(0) = \frac{1}{(n-k)^2} \sum_{j=k}^{n-1} W_j - \frac{1}{n^2} \sum_{j=0}^{n-1} W_j - \frac{1}{n-k}(\overline{Y}_{n,k} - \alpha)^2 + \frac{1}{n}(\overline{Y}_n - \alpha)^2. \tag{37}$$

But

$$\begin{aligned}\overline{Y}_{n,k} &= \overline{Y}_n\left(\frac{n}{n-k}\right) - \frac{\sum_{j=0}^{k-1} Y_j}{n-k}\\ &= \overline{Y}_n + \overline{Y}_n\left(\frac{k}{n}\right)(1+o(1)) - \overline{Y}_k\left(\frac{k}{n}\right)(1+o(1))\\ &= \overline{Y}_n + \frac{k}{n}(\overline{Y}_n - \overline{Y}_k) + o\left(\frac{k}{n}\right),\end{aligned}$$

so (34) and the fact that $k \le n^{q/2}$ imply that

$$\begin{aligned}\frac{1}{n-k}(\overline{Y}_{n,k}-\alpha)^2 &= \frac{1}{n}\left((\overline{Y}_n-\alpha) + \frac{k}{n}(\overline{Y}_n - \overline{Y}_k) + o\left(\frac{k}{n}\right)\right)^2\left(1+\frac{k}{n}+o\left(\frac{k}{n}\right)\right)\\ &= \frac{1}{n}(\overline{Y}_n-\alpha)^2 + O\left(\frac{2k}{n^2}(\overline{Y}_n-\alpha)(\overline{Y}_n-\overline{Y}_k)\right)\\ &= \frac{1}{n}(\overline{Y}_n-\alpha)^2 + O(n^{-2-q/2}) \qquad (38)\end{aligned}$$

a.s. as $n \to \infty$ (where $O(f(n))$ denotes a sequence of rv's $(\beta_j : j \ge 0)$ such that $\beta_n/f(n)$ remains bounded a.s. as $n \to \infty$), in view of the observations that $\overline{Y}_n - \overline{Y}_k = O(1)$ and $\overline{Y}_n - \alpha = O(n^{-q})$ a.s. as $n \to \infty$.

We also observe that (again keeping in mind that $k \le n^{q/2}$)

$$\begin{aligned}\frac{1}{(n-k)^2}\sum_{j=k}^{n-1} W_j - \frac{1}{n^2}\sum_{j=0}^{n-1} W_j &= \frac{1}{n^2}\sum_{j=k}^{n-1} W_j\left(1+\frac{2k}{n}+O\left(\frac{k^2}{n^2}\right)\right) - \frac{1}{n^2}\sum_{j=0}^{n-1} W_j\\ &= \frac{2k}{n^2}\frac{\sum_{j=k}^{n-1} W_j}{n} - \frac{1}{n^2}\sum_{j=0}^{k-1} W_j + O\left(\frac{k^2}{n^3}\right)\\ &= \frac{2k}{n^2}(\sigma^2 + o(n^{-q})) - \frac{1}{n^2}\sum_{j=0}^{k-1} W_j + O(n^{-3+q}) \qquad (39)\\ &= \frac{2k\sigma^2}{n^2} + o(n^{-2-q/2}) - \frac{1}{n^2}\sum_{j=0}^{k-1} W_j + O(n^{-3+q}) \qquad (40)\end{aligned}$$

a.s. as $n \to \infty$.

Results (37), (38), and (39) together establish that

$$\begin{aligned}\arg\min_{0\le k\le n^{q/2}}(g_n(k)-g_n(0)) &= \arg\min_{0\le k\le n^{q/2}} n^2(g_n(k)-g_n(0))\\ &= \arg\min_{0\le k\le n^{q/2}}\left(2k\sigma^2 - \sum_{j=0}^{k-1} W_j + o(n^{-q/2}) + O(n^{-1+q}) + O(n^{-q/2})\right)\\ &= \arg\min_{0\le k\le n^{q/2}}\left(2k\sigma^2 - \sum_{j=0}^{k-1} W_j + o(1)\right)\\ &\overset{a.s.}{\to} \arg\min_{k\ge 0}\left(2k\sigma^2 - \sum_{j=0}^{k-1} W_j\right)\end{aligned}$$

as $n \to \infty$. In view of (36), this proves the theorem. □

Proof of Proposition 4.1. Put $Z_j \triangleq V_{j-1} - \chi_j$,

$$\tilde{S}_n \triangleq Z_1 + \cdots + Z_n,$$

and set

$$\begin{aligned} \tau(x) &\triangleq \inf\{j \geq 0 : \ W_j(x) = 0\} \\ &= \inf\{j \geq 0 : \ \tilde{S}_j \leq -x\}. \end{aligned}$$

It is well known that

$$\frac{\tau(x)}{x} \overset{a.s.}{\to} \lambda \tag{41}$$

as $x \to \infty$ (e.g., see Rahimov and Abdurakhmanov [2007]). For $n < \tau(x)$, $W_n(x) = x + \tilde{S}_n$. For $t < \lambda$, set $\epsilon = \lambda - t$ and note that

$$\begin{aligned} P(W_{\lfloor tx \rfloor}(x) \in B) &= P(\tilde{S}_{\lfloor tx \rfloor}(x) \in B - x, \ \tau(x) \geq (\lambda - \epsilon/2)x) \\ &\quad + P(W_{\lfloor tx \rfloor}(x) \in B, \ \tau(x) < (\lambda - \epsilon/2)x) \\ &= P(\tilde{S}_{\lfloor tx \rfloor}(x) \in B - x) \\ &\quad + P(W_{\lfloor tx \rfloor}(x) \in B, \ \tau(x) < (\lambda - \epsilon/2)x) \\ &\quad - P(\tilde{S}_{\lfloor tx \rfloor}(x) \in B - x, \ \tau(x) < (\lambda - \epsilon/2)x), \end{aligned}$$

so

$$\sup_B |P(W_{\lfloor tx \rfloor}(x) \in B) - P(\tilde{S}_{\lfloor tx \rfloor} \in B - x)| \leq P(\tau(x) < (\lambda - \epsilon/2)x) \to 0$$

as $x \to \infty$. But

$$P(\tilde{S}_{\lfloor tx \rfloor} > \lambda^{-1}(\epsilon/2)x - x) - P(W_\infty > \lambda^{-1}(\epsilon/2)x) \to 1$$

as $x \to \infty$, proving (4) for $t < \lambda$.

On the other hand, for $t > \lambda$, let

$$\phi(n, \ B) \triangleq P(W_n(0) \in B)$$

and put

$$\phi^*(n) \triangleq \sup_B |P(W_n(0) \in B) - P(W_\infty \in B)|.$$

Then $\phi^*(n) \to 0$ as $n \to \infty$ (e.g., see p. 326 of Meyn and Tweedie [2009]). Observe that

$$\begin{aligned} |P(W_{\lfloor tx \rfloor}(x) \in B) - P(W_\infty \in B)| &\leq |P(W_{\lfloor tx \rfloor}(x) \in B, \ \tau(x) \leq \lfloor tx \rfloor) - P(W_\infty \in B)| \\ &\quad + P(\tau(x) > \lfloor tx \rfloor) \\ &= |E\phi(\lfloor tx \rfloor - \tau(x), \ B)I(\tau(x) \leq \lfloor tx \rfloor) - P(W_\infty \in B)| \\ &\quad + P(\tau(x) > \lfloor tx \rfloor) \\ &\leq |E(\phi(\lfloor tx \rfloor - \tau(x), \ B) - P(W_\infty \in B))I(\tau(x) \leq \lfloor tx \rfloor)| \\ &\quad + 2P(\tau(x) > \lfloor tx \rfloor) \\ &\leq E\phi^*(\lfloor tx \rfloor - \tau(x))I(\tau(x) \leq \lfloor tx \rfloor) \\ &\quad + 2P(\tau(x) > \lfloor tx \rfloor). \end{aligned}$$

Since $P(\tau(x) > \lfloor tx \rfloor) \to 0$ and $\phi^*(\lfloor tx \rfloor - \tau(x)) \overset{a.s.}{\to} 0$ as $x \to \infty$, the bounded convergence theorem proves (4) for $t > \lambda$. □

Proof of Proposition 4.2. Let

$$\nu(n(x),\ j) \triangleq P(\hat{d}_\gamma(n(x),\ 0) \le j).$$

Given our assumptions, Theorem 2 applies to $W(0)$ and consequently $\hat{d}_\gamma(n,\ 0)$ is a.s. bounded as a function of n. Thus, $(\hat{d}_\gamma(n,\ 0) : n \ge 0)$ is a tight sequence of rv's so that $\nu(n,\ j) \to 1$ as $j \to \infty$ uniformly in n. For $\beta > t > \lambda$, set $\epsilon = t - \lambda$. Then

$$\begin{aligned} P(\hat{d}_\gamma(n(x),\ x) \le tx) &\ge P(\hat{d}_\gamma(n(x),\ x) \le tx,\ \tau(x) \le (t - \epsilon/2)x) \\ &\ge E\nu(\lfloor n(x) - n(x)^\gamma \rfloor - \tau(x),\ tx - \tau(x)) I(\tau(x) \le (t - \epsilon/2)x). \quad (42) \end{aligned}$$

The last inequality uses the fact that if

$$\operatorname*{arg\,min}_{\tau(x) \le k \le \lfloor n(x) - n(x)^\gamma \rfloor} g_{n(x)}(k,\ x) \le tx,$$

then

$$\operatorname*{arg\,min}_{0 \le k \le \lfloor n(x) - n(x)^\gamma \rfloor} g_{n(x)}(k,\ x) \le tx.$$

It also exploits the strong Markov property at the hitting time $\tau(x)$ of level 0 to express the probability in terms of $\nu(\lfloor n(x) - n(x)^\gamma \rfloor - \tau(x),\ tx - \tau(x))$. Returning to (42), we recognize that since $\lfloor n(x) - n(x)^\gamma \rfloor - \tau(x) \to \infty$ when $\tau(x) \le (t - \epsilon/2)x$,

$$\nu(\lfloor n(x) - n(x)^\gamma \rfloor - \tau(x),\ tx - \tau(x)) \to 1$$

as $x \to \infty$, given the preceding tightness comment. Hence, the bounded convergence theorem yields the conclusion that $P(\hat{d}_\gamma(n(x),\ x) \le tx) \to 1$ as $x \to \infty$, provided that $P(\tau(x) \le (t - \epsilon/2)x) \to 1$ as $x \to \infty$. But this is an immediate consequence of (41).

Next, observe that for $k = \lfloor rx \rfloor$ with $r < \lambda$, and on the event that $n(x) \ge \tau(x)$ (whose probability tends to 1 as $x \to \infty$),

$$\begin{aligned} g_{n(x)}(k,\ x) &\ge \frac{1}{(n(x) - k)^2} \sum_{j=k}^{\tau(x)-1} (W_j(x) - \overline{W}_{n(x),\,k}(x))^2 \\ &= \frac{1}{(n(x) - k)^2} \sum_{j=k}^{\tau(x)-1} (x + \tilde{S}_j - \overline{W}_{n(x),\,k}(x))^2, \end{aligned}$$

where

$$\overline{W}_{n(x),\,k}(x) = \frac{\tau(x)}{n(x) - k} \sum_{j=k}^{\tau(x)-1} \frac{x + \tilde{S}_j}{\tau(x)} + \frac{n(x) - \tau(x)}{n(x) - k} \sum_{j=\tau(x)}^{n(x)-1} \frac{W_j(x)}{n(x) - \tau(x)}.$$

Of course, $(W_{\tau(x)+j}(x) : j \ge 0)$ has the same distribution as $(W_j(0) : j \ge 0)$, so it follows that

$$\frac{1}{m} \sum_{j=\tau(x)}^{\tau(x)+m-1} W_j(x) \overset{a.s.}{\to} EW_\infty < \infty$$

as $m \to \infty$ (e.g., see pp. 427 and 428 of Meyn and Tweedie [2009]), uniformly in x. In addition, the strong law of large numbers for $(\tilde{S}_j : j \ge 0)$ implies that for each $t \ge 0$,

$$\sup_{0 \le u \le t} \left| \frac{\tilde{S}_{\lfloor ux \rfloor}}{x} + \frac{u}{\lambda} \right| \overset{a.s.}{\to} 0$$

as $x \to \infty$ (e.g., see p. 20 of Whitt [2002]). Hence,

$$\frac{\overline{W}_{n(x),\,k}(x)}{x} = \frac{x}{n(x)-k}\sum_{j=k}^{\tau(x)-1}\frac{1}{x}\left(1-\frac{j}{\lambda x}\right)+o(1)$$

a.s. as $x \to \infty$. But the sum on the right-hand side is a Riemann sum approximation to the integral of

$$h(v) \triangleq 1-\frac{v}{\lambda}$$

over $[r,\ \lambda]$, proving that

$$\frac{\overline{W}_{n(x),\,k}(x)}{x} \to \frac{1}{\beta-r}\int_r^{\lambda}(1-\lambda^{-1}v)dv$$

a.s. as $x \to \infty$ (and, in fact, the convergence is uniform in r over $[0,\ \lambda]$). A similar Riemann sum approximation establishes that

$$\begin{aligned}\frac{1}{x}\frac{1}{(n(x)-k)^2}\sum_{j=k}^{\tau(x)-1}(x+\tilde{S}_j-\overline{W}_{n(x),\,k}(x))^2 &= \frac{x^2}{(n(x)-k)^2}\sum_{j=k}^{\tau(x)-1}\frac{1}{x}\left(1+\frac{\tilde{S}_j}{x}-\frac{\overline{W}_{n(x),\,k}(x)}{x}\right)^2\\ &\overset{a.s.}{\to} \frac{1}{(\beta-r)^2}\int_r^{\lambda}\left(1-\lambda^{-1}u-\frac{1}{\beta-r}\int_r^{\lambda}(1-\lambda^{-1}v)dv\right)^2du\end{aligned}$$

as $x \to \infty$, uniformly in $r \in [0,\ \lambda]$.

As a consequence, for any $\epsilon > 0$, $g_{n(x)}(k,\ x)$ is a.s. of order x for $k \le (\lambda-\epsilon)x$, whereas $g_{n(x)}(\tau(x),\ x)$ is a.s. $o(1)$ as $x \to \infty$. Thus, $P(\hat{d}_\gamma(n(x),\ x) \ge tx) \to 1$ for each $t < \lambda$, proving that

$$\frac{\hat{d}_\gamma(n(x),\ x)}{x} \xrightarrow{p} \lambda$$

as $x \to \infty$. □

ACKNOWLEDGMENTS

We thank the referees and the associate editor for their many detailed suggestions, leading to significant improvements to this article. We particularly thank one of the referees for an especially close reading of the proofs.

REFERENCES

V. Anantharam. 1988. How large delays build up in a GI/G/1 queue. *Queueing Systems* 5, 345–368.

B. Avi-Itzhak, W. L. Maxwell, and L. W. Miller. 1965. Queueing with alternating priorities. *Operations Research* 13, 2, 306–318.

S. Axler. 1997. *Linear Algebra Done Right* (2nd ed.). Springer-Verlag, New York, NY.

R. Bhattacharya, M. Majumdar, and N. Hashimzade. 2010. Limit theorems for monotone Markov processes. *Sankhya: The Indian Journal of Statistics* 72, 1, 170–190.

M. Bramson. 2008. Stability of queueing networks. *Probability Surveys* 5, 169–345.

E. Cinlar. 1975. *Introduction to Stochastic Processes*. Dover Publications Inc., Mineola, New York.

J. G. Dai. 1996. A fluid limit model criterion for instability of multiclass queueing networks. *Annals of Applied Probability* 6, 3, 751–757.

A. DasGupta. 2011. *Probability for Statistics and Machine Learning: Fundamentals and Advanced Topics*. Springer, New York, NY.

W. W. Franklin and K. P. White Jr. 2008. Stationarity tests and MSER-5: Exploring the intuition behind mean-squared-error-reduction in detecting and correcting initialization bias. In *Proceedings of the 2008 Winter Simulation Conference.* 541–546.

W. W. Franklin and K. P. White Jr. 2010. Parametric expression for MSER with geometrically decaying bias. In *Proceedings of the 2010 Winter Simulation Conference*. 957–964.

P. W. Glynn. 1989. A GSMP formalism for discrete event systems. *Proceedings of the IEEE* 77, 1, 14–23.

Winfried K. Grassmann. 2011. Rethinking the initialization bias problem in steady-state discrete event simulation. In *Proceedings of the 2011 Winter Simulation Conference.* 593–599.

K. Hoad and S. Robinson. 2011. Implementing MSER-5 in commercial simulation software and its wider implications. In *Proceedings of the 2011 Winter Simulation Conference.* 495–503.

K. Hoad, S. Robinson, and R. Davies. 2010. Automating warm-up length estimation. *Journal of the Operational Research Society* 61, 1389–1403.

E. K. Lada, N. M. Steiger, and J. R. Wilson. 2006. Performance evaluation of recent procedures for steady-state simulation analysis. *IIE Transactions* 38, 9, 711–727.

A. M. Law. 2015. *Simulation Modeling and Analysis* (5th ed.). McGraw-Hill, New York, NY.

S. Meyn and R. L. Tweedie. 2009. *Markov Chains and Stochastic Stability* (2nd ed.). Cambridge University Press, Cambridge, UK.

A. C. Mokashi, J. J. Tejada, S. Yousefi, T. Xu, J. R. Wilson, A. Tafazzoli, and N.M. Steiger. 2010. Performance comparison of MSER-5 and N-Skart on the simulation start-up problem. In *Proceedings of the 2010 Winter Simulation Conference.* 971–982.

R. Pasupathy and B. Schmeiser. 2010. The initial transient in steady-state point estimation: Contexts, a bibliography, the MSE criterion, and the MSER statistic. In *Proceedings of the 2010 Winter Simulation Conference*. 184–197.

F. G. Rahimov and V. A. Abdurakhmanov. 2007. On limit behavior of linear first passage time of the Markov chain. *Proceedings of the Institute of Mathematics and Mechanics of the National Academy of Sciences of Azerbaijan* 27, 69–74.

S. Robinson. 2002. New simulation output analysis techniques: A statistical process control approach for estimating the warm-up period. In *Proceedings of the 2002 Winter Simulation Conference*. 439–446.

L. W. Schruben. 1982. Detecting initialization bias in simulation output. *Operations Research* 30, 3, 569–590.

L. W. Schruben, H. Singh, and L. Tierney. 1983. Optimal tests for initialization bias in simulation output. *Operations Research* 71, 6, 1167–1178.

A. Tafazzoli, N. M. Steiger, and J. R. Wilson. 2011. N-Skart: A nonsequential skewness- and autoregression-adjusted batch-means procedure for simulation analysis. *IEEE Transactions on Automatic Control* 56, 2, 254–264.

R. J. Wang and P. W. Glynn. 2014. Measuring the initial transient: Reflected Brownian motion. In *Proceedings of the 2014 Winter Simulation Conference*. 652–661.

R. J. Wang and P. W. Glynn. 2016. On the rate of convergence to equilibrium for reflected Brownian motion. In preparation.

K. P. White Jr. 1997. An effective truncation heuristic for bias reduction in simulation output. *Simulation* 69, 6, 323–334.

K. P. White Jr., M. J. Cobb, and S. C. Spratt. 2000. A comparison of five steady-state truncation heuristics for simulation. In *Proceedings of the 2000 Winter Simulation Conference.* 755–760.

W. Whitt. 2002. *Stochastic-Process Limits*. Springer-Verlag, New York, NY.

W. Whitt. 2006. Analysis for design. In *Handbooks in Operations Research and Management Science*, S. G. Henderson and B. L. Nelson (Eds.). North-Holland, Amsterdam, Netherlands.

J. R. Wilson and A. A. B. Pritsker. 1978. A survey of research on the simulation startup problem. *Simulation* 31, 2, 55–58.

E. M. M. Winands, I. J. B. F. Adan, and G. van Houtum. 2006. Mean value analysis for polling systems. *Queueing Systems* 54, 35–44.

Received May 2013; revised September 2015; accepted February 2016