On the Validity of Long-Run Estimation Methods for Discrete-Event Systems

Peter J. Haas* Peter W. Glynn†

Performance evaluation of computer systems, networks, and applications often involves analysis of long-run system characteristics. Many characteristics of interest can be expressed as time-average limits of the form

$$r(f) = \lim_{t\to\infty} \frac{1}{t} \int_0^t f(X(u))\, du,$$

where f is a real-valued function and $\{ X(t) \colon t \geq 0 \}$ is the underlying stochastic process that records the state of the system as it evolves over continuous time. In this paper we assume that $\{ X(t) \colon t \geq 0 \}$ can be represented as a generalized semi-Markov process (GSMP) and consider simulation-based methods for obtaining point estimates and confidence intervals for time-average limits. We also consider time-average limits of the form

$$\tilde{r}(\tilde{f}) = \lim_{n\to\infty} \frac{1}{n} \sum_{j=0}^{n-1} \tilde{f}(S_j, C_j),$$

where $\{ (S_n, C_n) \colon n \geq 0 \}$ is the general state space Markov chain used to define the GSMP (see below).

When the output process $\{ f(X(t)) \colon t \geq 0 \}$ or $\{ \tilde{f}(S_n, C_n) \colon n \geq 0 \}$ obeys a central limit theorem (CLT), there are two main approaches to obtaining an asymptotic confidence interval for the time-average limit. The first approach is to derive the confidence interval using a limit theorem in which the variance constant that appears in the CLT is "cancelled out" [6] and hence need not be estimated. "Cancellation" procedures of this type include [5, 6, 7] the original "standardized time series" (STS) area and maximum methods, the original methods of batch means and spaced batch means (where the number of batches is independent of the simulation run length), and the STS-weighted-area method. The second approach is to consistently estimate the variance constant. Procedures of this type include the regenerative method [13], the method of "variable" batch means (where the number of batches increases as the run length increases), and spectral methods [1].

It is usually nontrivial to determine for a specified GSMP model whether time-average limits are well defined. It is even harder to determine whether the output process obeys a CLT and, if so, whether a specified estimation method is applicable. Most conditions in the literature involve unrealistic assumptions—such as stationarity of the output process—or are difficult to verify. In this paper we provide new conditions on the building blocks of a GSMP under which long-run estimation problems are well defined and a variety of cancellation and consistent estimation methods are provably valid.

Our first set of results provides building-block conditions under which time-average limits exist and the output process $\{ f(X(t)) : t \geq 0 \}$ or $\{ \tilde{f}(S_n, C_n) \colon n \geq 0 \}$ obeys a functional central limit theorem (FCLT). Roughly speaking, a continuous-time stochastic process with time-average limit r obeys an FCLT if the associated cumulative (i.e., time-integrated) process, centered about the deterministic function $g(t) = rt$ and suitably compressed in space and time, converges in distribution to a Brownian motion as the degree of compression increases; the definition of an FCLT for a discrete-time process is similar. When an FCLT holds, the output process obeys an ordinary CLT. Moreover, the validity of a broad class of cancellation methods—including all of those mentioned above—follows directly from results in [6]. Our moment conditions are significantly weaker than those in [9, 11] and, in fact, appear to be the weakest conditions possible.

Our remaining results provide building-block conditions under which various estimators of the variance constant in the CLT for the output process are (weakly) consistent, so that confidence intervals based on these variance estimators are asymptotically valid. We use a coupling approach to extend consistency results for variance estimators from a stationary to a non-stationary setting. By combining this approach with known results for stationary processes, we obtain sufficient conditions under which a class of "quadratic form" variance estimators are consistent. This class includes batch means and spectral estimators. Our results complement those of [4], which establish *strong* consistency for variance estimators under the harder-to-verify assumption that the output process obeys a strong invariance principle. For example, it appears difficult to establish strong consistency for the popular version of variable batch means in which the number of batches grows as the 2/3 power of the run length—our results can be used to establish weak consistency for this method.

Generalized Semi-Markov Processes

The GSMP [13] is the traditional model for a discrete-event stochastic system. A GSMP $\{ X(t)\ t \geq 0 \}$ is a continuous-time stochastic process that makes a state transition when one or more "events" associated with the occupied state occur. Events associated with a state compete to trigger the next state transition and each set of trigger events has its

*Almaden Research Center, IBM Research Division, peterh@almaden.ibm.com

†Dept. of Management Science and Engineering, Stanford University, glynn@leland.stanford.edu

own probability distribution for determining the new state. At each state transition, new events may be scheduled. For each of these new events, a clock indicating the time until the event is scheduled to occur is set according to a probability distribution that depends on the old state, the new state, and the set of events that triggers the state transition. If a scheduled event is not in the set of events that triggers a state transition but is associated with the new state, then its clock continues to run down (at a state-dependent speed); if such an event is not associated with the new state, it is canceled and the corresponding clock reading is discarded. A GSMP is formally defined in terms of an underlying Markov chain $\{ (S_n, C_n) : n \geq 0 \}$, where S_n is the state and C_n is the vector of clock readings just after the nth state transition.

Limit Theorems

Denote by S the state space of the GSMP and by E the (finite) set of events; see [13] for details. Also, for $s \in S$, denote by $E(s)$ $(\subseteq E)$ the set of events that can possibly occur in state s. Whenever the current state is s and event $e \in E(s)$ occurs, the GSMP makes a transition to state $s' \in S$ with probability $p(s'; s, e)$. For $s, s' \in S$ write $s \to s'$ if $p(s'; s, e) > 0$ for some $e \in E(s)$, and write $s \rightsquigarrow s'$ if $s = s^{(0)} \to s^{(1)} \to \cdots \to s^{(k)} = s'$ for some $k \geq 1$ and states $s^{(1)}, s^{(2)}, \ldots, s^{(k-1)} \in S$. The GSMP is *irreducible* if $s \rightsquigarrow s'$ for each $s, s' \in S$. Denote by $r(s, e)$ the speed at which the clock for event $e \in E(s)$ runs down to 0 in state s. Finally, denote by μ the initial distribution of the underlying chain: $\mu(A) = P\{ (S_0, C_0) \in A \}$. We assume that each event e is "simple" with clock-setting distribution $F(\cdot\,; e)$.

DEFINITION 1. *Assumption PD(q)* is said to hold for a specified GSMP and real number $q \geq 0$ if (i) the state space S is finite, (ii) the GSMP is irreducible, (iii) each clock speed $r(s, e)$ is positive, and (iv) there exists $0 < \bar{x} \leq \infty$ such that each clock-setting distribution function has finite qth moment and a density component that is positive and continuous on $(0, \bar{x})$. If the requirement of finite qth moments is replaced by the requirement that there exist a real number $v > 1$ such that $\int_0^\infty v^x \, dF(x; e) < \infty$ for each $e \in E$, then *Assumption PDE* is said to hold.

We first give conditions under which time-average limits are well defined. Let Σ be the state space of the underlying chain $\{ (S_n, C_n) : n \geq 0 \}$. For $(s, c) \in \Sigma$, denote by $t^*(s, c)$ the time, starting in state s with clock-reading vector c, until the next state transition of the GSMP. Also, for $u \geq 0$, denote by $\mathcal{H}_u$ the set real-valued functions $\tilde{h}$ defined on Σ that satisfy $|\tilde{h}(s, c)| \leq a + b\left(t^*(s, c)\right)^u$ for some $a, b \geq 0$. Write $x \vee y = \max(x, y)$.

THEOREM 1. *If Assumption PD(1) holds, then, for any function $f: S \mapsto \Re$, there exists a finite constant $r(f)$ such that $\lim_{t\to\infty}(1/t)\int_0^t f\left(X(u)\right) du = r(f)$ a.s. for any initial distribution μ. If Assumption PD$(u \vee 1)$ holds $(u \geq 0)$, then, for any $\tilde{f} \in \mathcal{H}_u$, there exists a finite constant $\tilde{r}(\tilde{f})$ such that $\lim_{n\to\infty}(1/n)\sum_{j=0}^{n-1} \tilde{f}(S_j, C_j) = \tilde{r}(\tilde{f})$ a.s. for any initial distribution μ.*

We now give FCLT's in continuous and discrete time. Denote by $C[0,1]$ the set of continuous real-valued functions defined on $[0,1]$, by $\Rightarrow$ weak convergence on $C[0,1]$, and by $W = \{ W(t) : 0 \leq t \leq 1 \}$ a standard Brownian motion; see [2, 10]. When quantities $r(f)$ and $\tilde{r}(\tilde{f})$ exist, define sequences of $C[0,1]$-valued random functions $\{ U_\nu(f) : \nu \geq 0 \}$ and $\tilde{U}_1(\tilde{f}), \tilde{U}_2(\tilde{f}), \ldots$ by setting

$$U_\nu(f)(t) = \frac{1}{\sqrt{\nu}} \int_0^{\nu t} \Big(f\big(X(u)\big) - r(f) \Big) \, du$$

and

$$\tilde{U}_n(\tilde{f})(t) = \frac{1}{\sqrt{n}} \int_0^{nt} \big(\tilde{f}(S_{\lfloor u \rfloor}, C_{\lfloor u \rfloor}) - \tilde{r}(\tilde{f}) \big) \, du$$

for $0 \leq t \leq 1$, $\nu \geq 0$, and $n \in \{1, 2, \ldots\}$. (Here $\lfloor x \rfloor$ is the greatest integer less than or equal to x.)

THEOREM 2. *If Assumption PD(2) holds, then, for any function $f: S \mapsto \Re$, there exists $\sigma(f) \in [0, \infty)$ such that $U_\nu(f) \Rightarrow \sigma(f) W$ as $\nu \to \infty$ for any initial distribution μ. If Assumption PD$(2(u \vee 1))$ holds $(u \geq 0)$, then, for any $\tilde{f} \in \mathcal{H}_u$, there exists $\tilde{\sigma}(\tilde{f}) \in [0, \infty)$ such that $\tilde{U}_n(\tilde{f}) \Rightarrow \tilde{\sigma}(\tilde{f}) W$ as $n \to \infty$ for any initial distribution μ.*

Theorems 1 and 2 are established by arguing, as in [9], that sample paths of the output process can be decomposed into identically distributed, one-dependent cycles under the conditions of the theorems. Then, using a drift condition in [9], a martingale-based bound for functionals of a Markov chain [12, Th. 14.2.3], an argument similar to the proof of Wald's moment identity, and a well known bound on moments of random iid sums [8, Theorem I.5.2], we show that the sum of the output process over a cycle has finite first and second moments under the conditions of Theorems 1 and 2, respectively. The desired results now follow, as in [9, 10], from well known theorems for wide-sense regenerative processes.

Consistent Estimation Methods

For ease of exposition, we assume henceforth that Assumption PDE holds. Let $\tilde{f}$ be a function that is *polynomially dominated* in the sense that $\tilde{f} \in \mathcal{H}_u$ for some $u \geq 0$, and set $\bar{r}(n; \tilde{f}) = (1/n)\sum_{j=0}^{n-1} \tilde{f}(S_j, C_j)$. It follows from the results of the previous section together with the continuous mapping theorem that $\lim_{n\to\infty} \bar{r}(n; \tilde{f}) = \tilde{r}(\tilde{f})$ a.s. for some finite constant $\tilde{r}(\tilde{f})$—so that the point estimator $\bar{r}(n; \tilde{f})$ is strongly consistent for $\tilde{r}(\tilde{f})$—and $n^{1/2}\left(\bar{r}(n; \tilde{f}) - \tilde{r}(\tilde{f})\right) \Rightarrow \tilde{\sigma}(\tilde{f}) N(0,1)$ as $n \to \infty$ for some constant $\tilde{\sigma}(\tilde{f}) \in [0, \infty)$. Suppose that $\tilde{\sigma}^2(\tilde{f}) > 0$ (the usual case) and we can find an estimator V_n that is *consistent* for the variance constant $\tilde{\sigma}^2(\tilde{f})$ in that $V_n \Rightarrow \tilde{\sigma}^2(\tilde{f})$ as $n \to \infty$. Then standard arguments show that $[\bar{r}(n; \tilde{f}) - z_p(V_n/n)^{1/2}, \bar{r}(n; \tilde{f}) + z_p(V_n/n)^{1/2}]$ is an asymptotic $100p\%$ confidence interval for $\tilde{r}(\tilde{f})$, where z_p is the $(1+p)/2$ quantile of the standard normal distribution. We now discuss methods for consistently estimating $\tilde{\sigma}^2(\tilde{f})$—our emphasis is on methods that do not require regenerative structure as in [13]. After presenting the main results in the discrete-time setting, we briefly describe the extension to continuous time.

Consider a variance estimator of the form

$$V_n = V_n(\tilde{f}) = \sum_{i=0}^{n} \sum_{j=0}^{n} \tilde{f}(S_i, C_i) \tilde{f}(S_j, C_j) q_{i,j}^{(n)},$$

where each $q_{i,j}^{(n)}$ is a finite constant and $q_{i,j}^{(n)} = q_{j,i}^{(n)}$ for all i, j. Such a *quadratic-form* estimator V_n is said to be *localized* if there exist a constant $a_1 \in (0, \infty)$ and sequences

$\{ a_2(n) : n \geq 0 \}$ and $\{ m(n) : n \geq 0 \}$ of nonnegative constants with $a_2(n) \to 0$ and $m(n)/n \to 0$ such that (i) $|q_{i,j}^{(n)}| \leq a_1/n$ if $|i-j| \leq m(n)$ and (ii) $|q_{i,j}^{(n)}| \leq a_2(n)/n$ if $|i-j| > m(n)$. A localized estimator has the property that, as more and more observations of the output process are obtained, the influence of any one observation on the value of the estimator becomes negligible.

For a process $\{ Z_n : n \geq 0 \}$ with variance constant $\tilde{\sigma}^2$, the discrete time batch means estimator of $\tilde{\sigma}^2$ based on b batches of length m is given by

$$V_n^{(B)} = \frac{m}{b-1} \sum_{j=1}^{b} \left(\bar{X}_n(j) - \bar{X}_n \right)^2 \tag{1}$$

for $n = bm$ (the case that we always consider), where $\bar{X}_n(j) = (1/m) \sum_{i=(j-1)m}^{jm-1} Z_i$ is the jth batch mean ($1 \leq j \leq b$) and $\bar{X}_n = (1/b) \sum_{j=1}^{b} \bar{X}_n(j)$. The class of *spectral estimators* comprises variance estimators of the form

$$V_n^{(S)} = \sum_{h=-(m-1)}^{m-1} \lambda(h/m) \hat{R}_h, \tag{2}$$

where $\hat{R}_h = (n - |h|)^{-1} \sum_{i=0}^{n-|h|-1} (Z_i - \bar{Z}_n)(Z_{i+|h|} - \bar{Z}_n)$ and $\bar{Z}_n = (1/n) \sum_{i=0}^{n-1} Z_i$. The function λ is the "lag window," and we restrict attention to the class Λ of lag windows such that (i) λ is continuous on $[-1,1]$, (ii) $\lambda(x) = \lambda(-x)$, (iii) $\lambda(0) = 1$, (iv) $\lambda(x) = 0$ for $x \notin [-1,1]$, (v) $\sup_{-1 \leq x \leq 1} |\lambda(x)| < \infty$, and (vi) $\lim_{x \to 0} (1 - \lambda(x))/|x|^q = \alpha$ for some $q, \alpha \in (0, \infty)$. This class includes modified-Bartlett, Hanning, and Parzen windows; see [1, p. 527]. As shown in [14], the foregoing batch means and spectral estimators can be represented as quadratic-form estimators, and it is easy to verify that these estimators are localized.

We also restrict attention to "aperiodic" GSMP's, which are defined as follows. A *d-cycle* ($d \geq 1$) exists for a GSMP with state space S if and only if S can be partitioned into mutually disjoint subsets $S_1, S_2, \ldots, S_d$ such that $s' \in S_{i+1}$ whenever $s \to s'$ and $s \in S_i$; here $\to$ is defined as before, and we take $S_{d+1} = S_1$. The *period* of a GSMP is the largest integer d for which a d-cycle exists. A GSMP with period 1 is called *aperiodic.* It can be shown that if Assumption PDE holds for an aperiodic GSMP, then the underlying chain $\{ (S_n, C_n) : n \geq 0 \}$ is Harris ergodic, and a coupling argument then establishes the following result.

THEOREM 3. *Let $\{ (S_n, C_n) : n \geq 0 \}$ be the underlying chain of an aperiodic* GSMP *and $\tilde{f}$ be a polynomially dominated real-valued function defined on Σ. Suppose that Assumption PDE holds, so that there exists an invariant distribution π for the chain and $\{ \tilde{f}(S_n, C_n) : n \geq 0 \}$ obeys a* CLT *with variance constant $\tilde{\sigma}^2(\tilde{f})$. If a localized quadratic-form estimator $V_n(\tilde{f})$ satisfies $V_n(\tilde{f}) \Rightarrow \tilde{\sigma}^2(\tilde{f})$ when the initial distribution is π, then $V_n(\tilde{f}) \Rightarrow \tilde{\sigma}^2(\tilde{f})$ for any initial distribution.*

To obtain sufficient conditions for consistency of, e.g., batch-means and spectral variance estimators under the hypotheses of Theorem 3, we use the Harris ergodicity of the underlying chain in combination with standard moment bounds for Markov chains [12] and general results on consistent estimation in the stationary regime [1, 3] to establish consistency when the initial distribution of the underlying chain is the invariant distribution. Finally, we apply Theorem 3 to obtain the following result.

THEOREM 4. *Let $\{ (S_n, C_n) : n \geq 0 \}$ be the underlying chain of an aperiodic* GSMP *and $\tilde{f}$ be a polynomially dominated real-valued function defined on Σ. Suppose that Assumption PDE holds, so that $\{ \tilde{f}(S_n, C_n) : n \geq 0 \}$ obeys a* CLT *with variance constant $\tilde{\sigma}^2(\tilde{f})$.*

(i) If the number of batches $b = b(n)$ and batch length $m = m(n)$ satisfy $b(n) \to \infty$ and $m(n) \to \infty$ as $n \to \infty$, then $V_n^{(B)} \Rightarrow \tilde{\sigma}^2(\tilde{f})$ as $n \to \infty$, where $V_n^{(B)}$ is as in (1) with $Z_n = \tilde{f}(S_n, C_n)$.

(ii) If the spectral window length $m = m(n)$ satisfies $m(n) \to \infty$ and $m^2(n)/n \to 0$, then $V_n^{(S)} \Rightarrow \tilde{\sigma}^2(\tilde{f})$ as $n \to \infty$, where $V_n^{(S)}$ is as in (2) with $\lambda \in \Lambda$ and $Z_n = \tilde{f}(S_n, C_n)$.

Standard Cramér–Wold arguments extend the foregoing development to $\Re^l$-valued functions $\tilde{f}$, and we can apply the delta method to extend the preceding methodology to handle functions of time-average limits of the form $\tilde{r}(\tilde{f})$. In particular, a time-average limit $r(f)$ in continuous time can be expressed as $r(f) = \tilde{r}(ft^*)/\tilde{r}(t^*)$, where $(ft^*)(s,c) = f(s)t^*(s,c)$, and our results extend to the continuous-time setting.

REFERENCES

[1] T. W. Anderson. *The Statistical Analysis of Time Series.* Wiley, New York, 1971.

[2] P. Billingsley. *Convergence of Probability Measures.* Wiley, New York, 1968.

[3] C. Chien, D. Goldsman, and B. Melamed. Large-sample results for batch means. *Management Sci.*, 43:1288–1295, 1997.

[4] H. Damerdji. Strong consistency of the variance estimator in steady-state simulation output analysis. *Math. Oper. Res.*, 19:494–512, 1994.

[5] B. L. Fox, D. Goldsman, and J. J. Swain. Spaced batch means. *Oper. Res. Lett.*, 10:255–263, 1991.

[6] P. W. Glynn and D. L. Iglehart. Simulation output analysis using standardized time series. *Math. Oper. Res.*, 15:1–16, 1990.

[7] D. Goldsman, M. S. Meketon, and L. W. Schruben. Properties of standardized time series weighted area variance estimators. *Management Sci.*, 36:602–612, 1990.

[8] A. Gut. *Stopped Random Walks: Limit Theorems and Applications.* Springer-Verlag, New York, 1988.

[9] P. J. Haas. On simulation output analysis for generalized semi-Markov processes. *Comm. Statist. Stochastic Models*, 15:53–80, 1999.

[10] P. J. Haas. *Stochastic Petri Nets: Modelling, Stability, Simulation.* Springer-Verlag, New York, 2002.

[11] P. J. Haas. Estimation methods for delays in non-regenerative discrete-event systems. *Comm. Statist. Stochastic Models*, 19, 2003. To appear.

[12] S. P. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability.* Springer-Verlag, London, 1993.

[13] G. S. Shedler. *Regenerative Stochastic Simulation.* Academic Press, New York, 1993.

[14] W. T. Song and B. W. Schmeiser. Variance of the sample mean: Properties and graphs of quadratic-form estimators. *Oper. Res.*, 41:501–517, 1993.