CONSERVATION EQUATIONS AND VARIANCE REDUCTION IN QUEUEING SIMULATIONS

John S. Carson and Averill M. Law
Department of Industrial Engineering
University of Wisconsin-Madison

ABSTRACT

We consider the efficient estimation of mean delay in queue, d, mean wait in system, w, time average number in queue, Q, and time average number in system, L, for simulated queueing systems. We prove for the $GI/G/s$ queue that it is more efficient to estimate w, Q, and L from an estimate of d than it is to estimate them directly. This generalizes previous results for the $M/G/1$ queue and also confirms empirical studies on other $GI/G/s$ queues.

I. INTRODUCTION

We are concerned with the efficient estimation of the parameters d, Q, w, and L when the data are collected from a computer simulation of a queueing system, where:

d = long-run average delay in queue per customer,
Q = long-run time average number in queue,
w = long-run average wait in system per customer,
L = long-run time average number in system.

(When we say "delay," we mean wait in queue only, not including the wait in service. The "system" consists of one or more queues (waiting lines) plus a service mechanism.)

More specifically, we consider how conservation equations and other structural relations between parameters can be used to achieve a variance reduction. The parameters d, Q, w, and L are related by the equations:

$$Q = \lambda d, \tag{1}$$

$$L = \lambda w, \tag{2}$$

$$w = d + E(S), \tag{3}$$

where λ is the arrival rate of customers to the system, and $E(S)$ is the expected total time in service of a typical customer. The relations (1)-(3) hold for a great variety of queueing systems without any distributional assumptions and in particular when the system has the regenerative property (which we assume throughout). A proof of (1) and (2) may be found in Stidham [5]; equation (3) is obvious.

We consider a brief example. Let $\hat{Q}$ be the standard regenerative estimator of Q in a regenerative queueing process. Relation (1) suggests an alternative estimator, $\lambda\hat{d}$, $\hat{d}$ being the standard regenerative estimator of d. We have shown that for the $GI/G/s$ queue the estimator $\lambda\hat{d}$ is more efficient than $\hat{Q}$ as an estimator of Q, i.e., $\text{Var}(\lambda\hat{d}) < \text{Var}(\hat{Q})$ at least for large enough sample sizes. This result, and the others discussed later, generalize previous results for the $M/G/1$ queue (see Law [4]). Moreover, the methods used to prove our results are quite general in scope and potentially can be applied to other queueing processes.

For the remaining sections, we discuss the regenerative method of estimation as applied to the $GI/G/s$ queue (see Crane and Iglehart [2]). Our main results are listed in Section 5. All proofs may be found in Carson [1]. The last section briefly discusses further work along the same lines.

II. THE $GI/G/s$ QUEUE

In the standard $GI/G/s$ queue, we have a single waiting line and s parallel servers, interarrival times distributed as a random variable (r.v.)A, and service times distributed as a r.v. S, with all of these r.v's being mutually independent. Let $\lambda = 1/E(A)$ be the arrival rate, and assume $0 < E(A) < \infty$ and $0 < E(S) < \infty$. It is known that the queue is stable if and only if $\rho = \lambda E(S)/s < 1$. If, in addition, $P(A > S) > 0$, then the queue is regenerative (see Whitt [6]), and the regenerative method of estimation can be applied (see Crane and Iglehart [2]). To say that the queue is regenerative means that it becomes completely empty of customers infinitely often with probability one (w.p.1). If we assume, for simplicity, that the first customer finds all servers free, then he begins the first busy cycle. Suppose customer number $N_c + 1$ arrives at time B_c and is the next customer to find all servers free. Then a second busy cycle begins at time B_c. If $\rho < 1$ and $P(A > S) > 0$ (as we shall henceforth assume), then $E(N_c) < \infty$ and $E(B_c) < \infty$ (see [6]) and there will be an infinite sequence of independent and identically distributed (i.i.d.) busy cycles.

We shall need the following notation:

N_c = the number of customers served in a busy cycle;

B_c = the length of a busy cycle;
D_c = the total delay of all customers served in a busy cycle;
W_c = the total waiting time of all customers served in a busy cycle.

We shall assume that N_c, B_c, D_c, and W_c have finite second moments. Note that, under our assumptions, the steady-state parameters d, Q, w, and L exist and are finite and constant w.p.1.

III. THE REGENERATIVE METHOD OF ESTIMATION

Suppose we simulate a $GI/G/s$ queue for m busy cycles and we are interested in the estimation of Q and d. The relevant data to collect are

$$D_{c_1}, D_{c_2}, \ldots, D_{c_m},$$

and

$$N_{c_1}, N_{c_2}, \ldots, N_{c_m},$$

$$B_{c_1}, B_{c_2}, \ldots, B_{c_m},$$

where the subscript indicates the busy cycle. Let $\overline{D}_c$, $\overline{N}_c$, and $\overline{B}_c$ be the sample means of the three sequences. It follows from the regenerative nature of the queue that the random vectors $(D_{c_i}, N_{c_i}, B_{c_i})$, $1 \le i \le m$, are i.i.d. and that

$$d = E(D_c)/E(N_c) \tag{3a}$$

and

$$Q = E(D_c)/E(B_c). \tag{3b}$$

The direct regenerative estimators of d and Q are, respectively,

$$\hat{d} = \overline{D}_c/\overline{N}_c \tag{4}$$

and

$$\hat{Q} = \overline{D}_c/\overline{B}_c. \tag{5}$$

Using the strong law of large numbers, the central limit theorem, and (3a,b), it is easily seen that $\hat{d}$ and $\hat{Q}$ are strongly consistent for d and Q, respectively, and, in addition, are asymptotically normally distributed. (See Carson [1] or [2] for a proof.)

In a similar manner, we obtain the direct estimators of w and L:

$$\hat{w} = \overline{W}_c/\overline{N}_c, \tag{6}$$

$$\hat{L} = \overline{W}_c/\overline{B}_c. \tag{7}$$

IV. INDIRECT ESTIMATION

In a simulation λ and $E(S)$ would be known. From the relations (1)-(3), we see that an estimate of any one of the four parameters of interest could be used to obtain an estimate of any other parameter. For comparison purposes, we concentrate on estimation of d.

Let $\hat{d}_1 = \hat{d}$ be the direct estimator of d, and consider the following indirect estimators:

$$\hat{d}_2 = \hat{Q}/\lambda, \tag{8}$$

$$\hat{d}_3 = \hat{w} - E(S), \tag{9}$$

$$\hat{d}_4 = \hat{L}/\lambda - E(S). \tag{10}$$

Using (1)-(3) and the asymptotic normality of the estimators (4)-(7), it is easily seen that each of the estimators (8)-(10) is strongly consistent and asymptotically normally distributed.

The variance of the asymptotic normal distribution provides a basis for the comparison of the estimators $\hat{d}_i$, $i = 1,2,3,4$. This variance can be expressed in the following way. Suppose that θ is some steady-state parameter of a regenerative process (such as d, Q, w, or L) and that

$$\theta = \alpha E(X)/E(Y) + \beta, \tag{11}$$

where X and Y are r.v.'s defined over a busy cycle (such as N_c and D_c), and α and β are constants. A regenerative estimator of θ is

$$\hat{\theta} = \alpha\overline{X}/\overline{Y} + \beta, \tag{12}$$

where $\overline{X}$ and $\overline{Y}$ are sample means taken over m cycles (such as $\overline{N}_c$ and $\overline{D}_c$). The estimator $\hat{\theta}$ satisfies:

$$\hat{\theta} \to \theta \quad \text{w.p.1}$$

and

$$\sqrt{m}(\hat{\theta} - \theta)/\sqrt{v} \xrightarrow{\mathscr{D}} N(0,1) \quad \text{as} \quad m \to \infty, \tag{13}$$

where

$$\begin{aligned} v &= \operatorname{Var}(\alpha X - (\theta - \beta)Y)/E^2(Y) \\ &= (\theta - \beta)^2\operatorname{Var}(X/EX - Y/EY). \end{aligned} \tag{14}$$

(Here, $\xrightarrow{\mathscr{D}}$ denotes convergence in distribution and $N(0,1)$ is a mean zero, variance one normal r.v. That the two expressions in (14) are equal follows trivially from (11). The expression for v given in [4, eq. (1.27)] should have θ^2 replaced by $(\theta - \beta)^2$.)

The quantity v given by (14) will be denoted by VAD($\hat{\theta}$) and called the *variance of the asymptotic distribution*. Our basis of comparison for two distinct ratio estimators $\hat{\theta}_1$ and $\hat{\theta}_2$ of θ (both of the form (12)) will be their VADs, since it follows from (13) and (14) that Var($\hat{\theta}$) is approximately VAD($\hat{\theta}$)/m (for large m). If VAD($\hat{\theta}_1$) $\le$ VAD($\hat{\theta}_2$) and equality holds only in a degenerate case (i.e., when either the interarrival or service times are constant), then we shall say that $\hat{\theta}_1$ is *more efficient* than $\hat{\theta}_2$ as an estimate of θ. In the next section, we discuss the efficiency of the estimators $\hat{d}_i$, $i = 1,2,3,4$.

V. MAIN RESULTS: *GI/G/s* QUEUE

We now come to our main results. The first theorem shows that $\lambda\hat{d}_1$ is a more efficient estimator of Q than its direct estimator $\hat{Q} = \lambda\hat{d}_2$.

Theorem 1. VAD($\hat{d}_1$) $\le$ VAD($\hat{d}_2$), with equality holding if and only if Var(A) = 0, i.e., the interarrival times are constant (w.p.1).

The next theorem combined with the previous one shows that the indirect estimator $\lambda\hat{d}_1 + \lambda E(S)$ of L is more efficient than its direct estimator $\hat{L}$.

Theorem 2. VAD($\hat{d}_2$) $\le$ VAD($\hat{d}_4$), with equality holding if and only if Var(A) = 0 and Var (S) = 0.

The third theorem shows that the indirect estimator $\hat{d}_1 + E(S)$ of w is more efficient than its direct estimator $\hat{w}$.

Theorem 3. VAD($\hat{d}_1$) $\le$ VAD($\hat{d}_3$), with equality holding if and only if Var(S) = 0.

For completeness, we also state:

Theorem 4. VAD($\hat{d}_3$) $\le$ VAD($\hat{d}_4$), with equality holding if and only if Var(A) = 0.

VI. CONCLUSIONS AND ADDITIONAL WORK

Theorems 1 through 4 tell us that for the estimation of d it is more efficient to use the direct estimator $\hat{d}$ given by (4) than any of the indirect estimators (8)-(10). On the other hand, for the estimation of Q, w, or L, it is more efficient to use the appropriate linear function of $\hat{d}$ suggested by (1)-(3), namely,

$$\tilde{Q} = \lambda\hat{d},$$

$$\tilde{w} = \hat{d} + E(S),$$

$$\tilde{L} = \lambda\hat{d} + \lambda E(S).$$

Empirical evidence given in [3] and [4] indicates that variance reductions from 0% to at least 76% can be obtained by using the appropriate indirect estimator based on $\hat{d}$. Thus, at least in the *GI/G/s* queue, it is only necessary to estimate d.

In additional work, we have investigated using linear combinations of estimators. For example, let

$$\hat{Q}(\alpha) = \alpha_1\hat{Q} + \alpha_2\lambda\hat{d},$$

where $\alpha = (\alpha_1,\alpha_2)$ and $\alpha_1 + \alpha_2 = 1$. Note that $\hat{Q}((1,0)) = \hat{Q}$ and $\hat{Q}((0,1)) = \lambda\hat{d}$, and thus VAD($\hat{Q}(\alpha)$) $\le$ VAD($\lambda\hat{d}$) $\le$ VAD($\hat{Q}$) for some choice of α. Thus, by proper choice of $\alpha = (\alpha_1,\alpha_2)$, a greater variance reduction can be achieved than by using the single alternative, $\lambda\hat{d}$. It is important to note that this method will work for any regenerative queueing process, provided alternative estimates are available. For the details, see [1].

VII. ACKNOWLEDGEMENT

The authors would like to thank the Office of Naval Research for their support of this research under contract N00014-76-C-0403(NR 047-145).

VIII. REFERENCES

1. Carson, J.S., "Efficient Estimators for Simulated Stochastic Processes," Technical Report No. 77-26, Department of Industrial Engineering, University of Wisconsin, Madison (1977).

2. Crane, M.A. and Iglehart, D.L., "Simulating Stable Stochastic Systems, I: General Multi-Server Queues," J. Assoc. Comput. Mach. 21, 103-113 (1974).

3. Law, A.M., "Efficient Estimators for Simulated Queueing Systems," ORC 74-7, Operations Res. Ctr., Univ. of Calif., Berkeley (1974).

4. Law, A.M., "Efficient Estimators for Simulated Queueing Systems," Management Sci. 22, 30-41 (1975).

5. Stidham, S., "A Last Word on $L = \lambda w$," Operations Res. 22, 417-421 (1974).

6. Whitt, W., "Embedded Renewal Processes in the *GI/G/s* Queue," J. Appl. Prob. 9, 650-658 (1972).