INDIRECT ESTIMATION VIA $L = \lambda W$

PETER W. GLYNN
Stanford University, Stanford, California

WARD WHITT
AT&T Bell Laboratories, Murray Hill, New Jersey

(Received September 1985; revisions received December 1986, October 1987; accepted November 1987)

For a large class of queueing systems, Little's law ($L = \lambda W$) helps provide a variety of statistical estimators for the long-run time-average queue length L and the long-run customer-average waiting time W. We apply central limit theorem versions of Little's law to investigate the asymptotic efficiency of these estimators. We show that an indirect estimator for L using the natural estimator for W plus the known arrival rate λ is more efficient than a direct estimator for L, provided that the interarrival and waiting times are negatively correlated, thus extending a variance-reduction principle for the GI/G/s model due to A. M. Law and J. S. Carson. We also introduce a general framework for indirect estimation which can be applied to other problems besides $L = \lambda W$. We show that the issue of indirect-versus-direct estimation is related to estimation using nonlinear control variables. We also show, under mild regularity conditions, that any nonlinear control-variable scheme is equivalent to a linear control-variable scheme from the point of view of asymptotic efficiency. Finally, we show that asymptotic bias is typically asymptotically negligible compared to asymptotic efficiency.

The formula $L = \lambda W$ (Little's law) expresses a fundamental principle in queueing theory: Under very general conditions, the time-average queue length L is equal to the product of the arrival rate λ and the customer-average waiting time W (see Little 1961 and Stidham 1974). Little's law is very useful because the assumptions are minimal; it applies to general systems such as queueing networks and subnetworks as well as to individual queues (see Section 11.3 of Heyman and Sobel 1982).

As shown by Law (1974, 1975), Carson (1978) and Carson and Law (1980), Little's law also helps construct new statistical estimators for the basic queueing parameters. These estimators are useful for both direct system measurements and computer simulation experiments. It is naturally of interest to compare these estimators and determine which have more desirable properties. For the GI/G/s queue, Carson and Law applied regenerative process theory to compare several of the basic estimators in terms of their asymptotic efficiency (the normalization constant in the central limit theorem). The asymptotic efficiency is important because it determines the size of confidence intervals when there are large samples.

Our goal in this paper is to compare the asymptotic efficiency of basic estimators for L and W in more general queueing systems. It turns out that the fundamental relation $L = \lambda W$, when appropriately generalized, not only helps identify estimators, but also provides the basis for comparing their asymptotic efficiency. In particular, central limit theorem (CLT) versions of $L = \lambda W$ can be applied for this purpose (see Glynn and Whitt 1986, 1987, 1988).

The problem is easily stated in rough general terms: For a queueing system satisfying Little's law, suppose that we have natural estimators $\hat{L}$, $\hat{\lambda}$ and $\hat{W}$ for L, λ and W, respectively. Several questions arise:

Q1. When are the estimators $\hat{\lambda}\hat{W}$ and $\hat{L}/\hat{\lambda}$ more efficient for estimating L and W, respectively, than the natural estimators $\hat{L}$ and $\hat{W}$?

Q2. If λ is known, as is often the case (e.g., with simulation), when are the estimators $\lambda\hat{W}$ and $\hat{L}/\lambda$ more efficient for estimating L and W than the natural estimators $\hat{L}$ and $\hat{W}$?

Q3. Can the differences in asymptotic efficiency be significant? Can the variance reduction using the more efficient estimators be substantial?

Q4. In each of these situations (using $\hat{\lambda}$ or λ), are there other estimators that are even more efficient than either of these two? Is it possible to determine the most efficient estimators in some sense?

Q5. When there is substantial potential variance reduction with a new estimator, is it easy to realize? Are the new estimators easy to construct and widely applicable?

Subject classification: Queues: applications to statistical estimation. Queues: asymptotic efficiency via $L = \lambda W$. Simulation: control variables.

0030-364X/89/3701-0082 $01.25

Q6. Do the conclusions here generalize? Are there additional ramifications? Are there even more general principles?

Q7. The focus above is on asymptotic efficiency. What about asymptotic bias (the difference between the expected value of the estimator and the actual parameter value)? Which is more important: asymptotic efficiency or asymptotic bias?

The purpose of this paper is to properly formulate and answer these questions. We obtain simple answers that are valid in considerable generality. Omitting the qualifications now, the answers are:

A1. Little's law does not change the asymptotic efficiency when λ needs to be estimated.

A2. When λ is known and used, the asymptotic efficiencies of the estimators are indeed typically different. A simple criterion is available to determine which estimator is more asymptotically efficient in terms of the covariance matrix in the central limit theorem version of $L = \lambda W$. Of course, the covariance matrix elements are hard to calculate even for elementary models, but in considerable generality, we show that $\lambda\hat{W}$ and $\hat{W}$ are more asymptotically efficient for estimating L and W than $\hat{L}$ and $\hat{L}/\lambda$ (see Section 7).

A3. The differences in asymptotic efficiency can be dramatic. We give examples in Section 7 in which the asymptotic variance of one estimator (natural or indirect) is zero, while the asymptotic variance of the other estimator is positive, so that the relative advantage can be infinite (either way). Typical variance reduction amounts from using $\lambda\hat{W}$ instead of $\hat{L}$ can be seen from numerical examples for GI/G/s queues in Law; Carson; and Carson and Law. The typical improvement is significant but not overwhelming; for a typical GI/G/s model one might achieve on the order of 10% variance reduction. (The ratio of the confidence interval lengths is the square root of the ratio of the variances.) The improvement usually increases as the arrival process gets more highly variable, so that for bursty arrival processes the variance reduction can well be 50–90%.

A4. Any discussion of "most efficient" requires caution and qualifications; much depends on the information and the context. Given appropriate qualifications, when λ is estimated, all the estimators are equally efficient asymptotically. *When λ is known, it is typically possible to do better than either the indirect or the natural estimator.* The most asymptotically efficient estimators of L and W in the general $L = \lambda W$ framework (Section 2) are $\lambda\hat{W} + \hat{a}(\hat{\lambda}^{-1} - \lambda^{-1})$ and $\hat{W} + (\hat{a}/\lambda)(\hat{\lambda}^{-1} - \lambda^{-1})$, where $\hat{a}$ is a consistent estimator for a constant times a ratio of covariance matrix elements. (It is no accident that these new estimators look like linear control-variable estimators; see **A6** below.) This optimality is restricted to the general $L = \lambda W$ framework in Section 2; *if other quantities such as mean service times are known in models with additional structure, then it is possible to do even better.*

A5. The alternative estimators are easy to construct, so that the gains in asymptotic efficiency are easy to realize. However, the indirect estimators $\lambda\hat{W}$ and $\hat{L}/\lambda$ clearly require no extra work. The improved estimators $\lambda\hat{W} + \hat{a}(\hat{\lambda}^{-1} - \lambda^{-1})$ and $\hat{W} + (\hat{a}/\lambda)(\hat{\lambda}^{-1} - \lambda^{-1})$ require somewhat more work because $\hat{a}$ needs to be constructed, but this is not difficult. Moreover, the same construction applies to a wide variety of queueing systems and other models. The improved estimators are worthwhile because they usually provide significantly more variance reduction than the direct and indirect estimators. However, with small samples the more elementary direct and indirect estimators might be preferred because they avoid the estimator $\hat{a}$. Experience indicates that there can be significant degradation of confidence interval coverage due to $\hat{a}$ with small samples (e.g., Lavenberg, Moeller and Sauer 1979).

A6. The conclusions generalize. We show that the issue of indirect-versus-direct estimation using $L = \lambda W$ can be regarded as a special case of estimation using nonlinear control variables; see Kleijnen (1974) and Nelson (1987). We also show that a nonlinear control-variable scheme is asymptotically equivalent to a linear control-variable scheme from the point of view of asymptotic efficiency. This is a rather direct consequence of Taylor's theorem, but it is very important. (Our analysis supplements p. 53 of Cheng and Feast (1980) on this point.) Hence, *from the point of view of asymptotic efficiency, the problem of exploiting $L = \lambda W$ for estimation efficiency can be viewed as a special case of estimation with linear controls.* (As noted in **A5** above, though, indirect estimation via $L = \lambda W$ is convenient because the weight $\hat{a}$ is not needed.) More generally, from the point of view of asymptotic efficiency, indirect estimation is covered by the theory of linear control variables; see Theorem 9 in Section 8. See Bratley, Fox and Schrage (1987); Iglehart and Lewis (1979); Chapter III of Kleijnen; Lavenberg, Moeller and Sauer; Lavenberg and Welch (1981); Lavenberg, Moeller and Welch (1982); Nozari, Arnold and Pegden (1984); Rubinstein and Marcus (1985); and Wilson and Pritsker (1984a,b) for background on linear control variables and references to

the relevant statistics literature, e.g., Hansen, Hurwitz and Madow (1953) and Cochran (1977). To a large extent, our analysis can be viewed as providing additional motivation for using linear control estimators. More generally, we present a convenient framework for evaluating the asymptotic efficiency of many estimators.

A7. We investigate asymptotic bias as well as asymptotic efficiency (Section 9). In a large-sample context, asymptotic efficiency is usually more important than asymptotic bias: Typically the size of confidence intervals (asymptotic efficiency) is on the order $n^{-1/2}$, whereas the bias is on the order n^{-1}, where n is the sample size. (This conclusion seems to be the accepted view; e.g., p. 278 of Fishman 1973; we present additional supporting arguments.) We identify two kinds of bias: initial and nonlinearity, both of which tend to be of the order n^{-1}. The linear control estimators have the advantage of having no nonlinearity bias, but bias is also introduced when the linear weight $\hat{a}$ in **A4** is estimated. The bias associated with estimating $\hat{a}$ is also typically of the order n^{-1}.

The key to answering these questions is formulating them carefully. Thus, we begin in Section 1 by introducing a general framework for considering indirect estimation. Then in Section 2 we introduce the framework for $L = \lambda W$. It is much less general than the framework of Section 1, but much more general than a specific model such as the standard GI/G/s queue. Some conclusions hold in the general estimation framework of Section 1, whereas others depend on the $L = \lambda W$ framework. We try to highlight the differences. In Section 3, we state more precisely the fundamental principles emerging from our analysis. The remaining sections develop the theory in more detail. We outline the rest of the paper at the end of Section 3. For much of the theory, we draw on Glynn and Whitt (1986, 1988); when we do, we often omit proofs.

We close this introduction by emphasizing that the focus of this paper is entirely on asymptotic analysis (the limiting behavior as the sample size n increases). We are concerned primarily with asymptotic efficiency, but we also consider asymptotic bias. We believe that asymptotic analysis is appropriate for most simulations, because simulations usually permit large samples. Moreover, experience indicates that the asymptotic analysis does indeed capture the dominant effects in a large sample context. However, in a small sample context many other statistical issues arise; we do not address these small-sample issues here.

1. General Framework for Indirect Estimation

We believe that it is useful to define three kinds of estimators for each parameter: *natural*, *direct* and *indirect*. Of course, there are many different specific estimators, but this classification captures the essential properties. To focus on the main ideas, we first define these estimators in a more general framework. Let $\{(X_n, Y_n, Z_n): n \geq 1\}$ be a sequence of random vectors, with $X_n \in R^k$, $Y_n \in R^l$ and $Z_n \in R^1$, that satisfy a Weak Law of Large Numbers (WLLN), i.e.,

$$n^{-1}\left(\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} Y_i, \sum_{i=1}^{n} Z_i\right) \Rightarrow (x, y, z) \quad \text{in } R^{k+l+1} \tag{1}$$

as $n \to \infty$, where (x, y, z) is a nonrandom vector and $\Rightarrow$ denotes weak convergence or convergence in distribution. (Recall that weak convergence to a nonrandom limit is equivalent to convergence in probability and that joint convergence in probability is equivalent to convergence in probability of the marginals separately; pp. 25–27 of Billingsley 1968.) Furthermore, suppose x, y and z are related by

$$z = f(x, y) \tag{2}$$

where $f: R^{k+l} \to R^1$ is a suitably smooth function, i.e., having continuous partial derivatives in all coordinates in a neighborhood of (x, y). In this framework, we regard x, y and z as the basic parameters and some initial segment $\{(X_i, Y_i, Z_i): 1 \leq i \leq n\}$ of the sequence $\{(X_i, Y_i, Z_i) : i \geq 1\}$ as the model for the observed data. The *natural estimators* for x, y and z are then, respectively,

$$\mathbf{x}_n^N = n^{-1} \sum_{i=1}^{n} X_i,$$

$$\mathbf{y}_n^N = n^{-1} \sum_{i=1}^{n} Y_i,$$

$$\mathbf{x}_n^N = n^{-1} \sum_{i=1}^{n} Z_i. \tag{3}$$

The *direct estimator* for z exploits (2) and the other natural estimators via

$$\mathbf{z}_n^D = f(\mathbf{x}_n^N, \mathbf{y}_n^N). \tag{4}$$

The *indirect estimator* for z also exploits the fact that one of the parameters x or y, say x, is known, so that

$$\mathbf{z}_n^I = f(x, \mathbf{y}_n^N). \tag{5}$$

For simplicity, we have assumed that Z_n and z are real-valued, but the ideas and results extend easily to vectors in R^d (e.g., see Rubinstein and Marcus). Also

note that the sequence $\{Z_n\}$ is needed in this general framework only for the natural estimator $\mathbf{z}_n^N$ in (3); we can obtain the direct estimator (4) and the indirect estimator (5) directly from $\{(X_i, Y_i) : 1 \leq i \leq n\}$ and (2). Moreover, other direct and indirect estimators for z are obtained by substituting different estimators for the natural estimators $\mathbf{x}_n^N$ and $\mathbf{y}_n^N$ in (3); i.e., the ideas easily generalize further.

By (1), the natural estimators in the general framework are all consistent, in the sense that $\mathbf{z}_n^N \Rightarrow z$ as $n \to \infty$. Since f is locally continuous, a minor modification of the continuous mapping theorem (Theorem 5.1 of Billingsley) implies that the other estimators for z in (4) and (5) are consistent too; e.g., $\mathbf{z}_n^D \Rightarrow f(x, y) = z$. To investigate the asymptotic efficiency of the estimators, we will introduce extra conditions so that the estimators satisfy CLTs. In particular, we will have

$$n^{1/2}(\mathbf{z}_n^N - z) \Rightarrow N(0, \sigma_N^2),$$

$$n^{1/2}(\mathbf{z}_n^D - z) \Rightarrow N(0, \sigma_D^2)$$

and

$$n^{1/2}(\mathbf{z}_n^I - z) \Rightarrow N(0, \sigma_I^2) \tag{6}$$

where $N(\mu, \sigma^2)$ is a random variable with a normal distribution having mean μ and variance σ^2. The variances, σ_N^2, σ_D^2 and σ_I^2 in (6) are the *asymptotic efficiency parameters* of the estimators $\mathbf{z}_n^N$, $\mathbf{z}_n^D$ and $\mathbf{z}_n^I$, respectively. We say that one estimator is more asymptotically efficient than another if its asymptotic efficiency parameter is less than the other.

In order to compare the direct and indirect estimators, we also assume that the basic sequence $\{(X_n, Y_n) : n \geq 1\}$ satisfies a joint CLT; i.e.,

$$n^{1/2}(\mathbf{x}_n^N - x, \mathbf{y}_n^N - y)$$

$$= n^{-1/2}\left(\sum_{i=1}^n X_i - nx, \sum_{i=1}^n Y_i - ny\right) \Rightarrow N(0, C) \tag{7}$$

where C is a covariance matrix, 0 is a vector of 0's and $N(0, C)$ is a random vector with a multivariable normal distribution with parameters $(0, C)$. We analyze asymptotic efficiency in this general framework in Section 8.

It is also important to consider the *bias* of each estimator. For example, for the direct estimator $\mathbf{z}_n^D$ in (4) the bias is $E(\mathbf{z}_n^D - z)$. Under reasonable regularity conditions, $n(E\mathbf{z}_n^D - z) \to \beta_D$ as $n \to \infty$, and similarly for the other estimators of z. We then call β_D the *asymptotic bias parameter* of the estimator $\mathbf{z}_n^D$. What is more important is that $E\mathbf{z}_n^D$ approaches z an order of magnitude faster than $\mathbf{z}_n^D$ does in (6), so that $E(\mathbf{z}_n^D - z)$ is asymptotically negligible as $n \to \infty$ compared to the size of the confidence interval based on $\mathbf{z}_n^D$. We develop the supporting asymptotic theory and describe the asymptotic bias parameter in more detail in Section 9.

2. A Framework for $L = \lambda W$

We now introduce a framework for $L = \lambda W$. We start with a sequence of ordered pairs of real-valued random variables $\{(A_n, D_n) : n \geq 1\}$, satisfying

$$0 \leq A_n \leq A_{n+1} \to \infty \quad \text{as } n \to \infty \quad \text{and } A_n \leq D_n < \infty$$

w.p. 1 (with probability one). We usually interpret A_n and D_n as the arrival and departure epochs of the nth arriving customer. (However, arrival and departure should be interpreted with respect to the system under consideration. For example, if the system refers to a queue, excluding the servers, then D_n is the epoch when the nth customer leaves the queue, which usually occurs when the customer begins service.) We do not require that $D_n \leq D_{n+1}$, i.e., customers depart in the same order that they arrive, although this does in fact hold for many queues. We view the nth customer as being in the system during the interval $[A_n, D_n]$, so that $Q(t)$, the number of customers in the system at time t, is given by

$$Q(t) = \sum_{n=1}^{N(t)} I(A_n \leq t \leq D_n) \tag{8}$$

where I is the set indicator function, defined for any event B and sample point ω by $I(B)(\omega) = 1$ if $\omega \in B$ and $I(B)(\omega) = 0$ otherwise, and

$$N(t) = \max\{n \geq 0 : A_n \leq t\}, \; t \geq 0, \tag{9}$$

with $A_0 = 0$ without there being a 0th customer. Thus, $N(t)$ is the arrival counting process. The waiting time for the nth customer is of course $W_n = D_n - A_n$. (For further discussion, see Section 2 of Glynn and Whitt (1986). A more general framework encompassing the extension of $L = \lambda W$ to $H = \lambda G$ is introduced and CLTs are proved for it in Glynn and Whitt (1989). The $H = \lambda G$ relations are also covered by the general estimation framework of Section 1.)

The standard statement of Little's law relates the w.p. 1 limit of the time average $t^{-1}\int_0^t Q(s)\,ds$ as $t \to \infty$ to the w.p. 1 limits of the customer averages $n^{-1}\sum_{k=1}^n W_k$ and $n^{-1}A_n$ as $n \to \infty$. To study

asymptotic efficiency, we will assume instead a joint CLT.

Basic CLT Assumption

The sequence $\{(A_n, \sum_{k=1}^n W_k): n \geq 1\}$ obeys a joint CLT: There exist constants λ and w with $0 < \lambda$, $w < \infty$ and a covariance matrix $C = \{C_{ij}: 1 \leq i, j \leq 2\}$ such that

$$n^{-1/2}\left(A_n - n\lambda^{-1}, \sum_{k=1}^{n} W_k - nw\right) \Rightarrow N(0, C). \tag{10}$$

It is of course important that the CLT (10) and stronger FCLTs (functional central limit theorems) actually hold in many circumstances (see Iglehart 1971; Whitt 1972; Glynn and Whitt 1987; and Propositions 2 and 3 in Section 7 here.

Henceforth, we follow the convention that ordinary lower case letters usually represent nonrandom elements, while capitals and boldface letters such as $\mathbf{z}_n^N$ in (3) usually represent random variables. Thus, we let w replace W as the long-run customer-average waiting time and we let q replace L as the long-run time-average queue length. This explains why w appears as a translation constant in (10).

To interpret what follows it is useful to have a concrete example (which we use throughout the paper).

Example 1a

Consider the standard M/M/1 queue with service rate 1 and arrival rate ρ with $\rho < 1$. Let W_n be the waiting time of the nth customer before beginning service. Then (10) holds with $C_{11} = 1/\rho^2$, $C_{12} = -1(/1 - \rho)^2$ and $C_{22} = \rho[2 + 5\rho - 4\rho^2 + \rho^3]/(1 - \rho)^4$. First, C_{11} is obvious because it is just the variance of an exponential with mean $1/\rho$. The term C_{22} comes from (2.1) of Law (1975), while the term C_{12} comes from (2.2) of Law plus Theorem 2 here. See Daley and Jacobs (1969), Iglehart (1971) and Abate and Whitt (1988) for different approaches.

By the continuous mapping theorem (Theorem 5.1 of Billingsley), (10) immediately implies the associated Weak Law of Large Numbers (WLLN), i.e.,

$$n^{-1}A_n \Rightarrow \lambda^{-1} \quad \text{and} \quad n^{-1}\sum_{k=1}^{n} W_k \Rightarrow w. \tag{11}$$

Less obvious is the associated WLLN for the arrival counting process $N(t)$ and the queue length process $Q(t)$ in (8) and (9).

Theorem 1. *If the WLLNs* (11) *holds, which is implied by* (10), *then*

$$t^{-1}N(t) \Rightarrow \lambda \quad \text{and} \quad t^{-1}\int_0^t Q(s)\,ds \Rightarrow q$$

as $t \to \infty$, *where* $q = \lambda w$.

This is Theorem 3 of Glynn and Whitt (1988). Thus, (10) yields WLLNs for $Q(t)$, W_n and A_n and requires that the limits be related by $q = \lambda w$, which, of course, is $L = \lambda W$ in our notation. In other words, (10) is a hypothesis guaranteeing the existence of the relevant limits and the validity of Little's formula (see Franken, König, Arndt and Schmidt 1981; Stidham, and Section 2 of Glynn and Whitt 1986 for other such hypotheses). Neither (10) nor the standard hypotheses imply each other (see Glynn and Whitt 1988). More important for our purposes, however, is the fact that (10) also has important implications for statistical estimation.

Note that the $L = \lambda W$ framework is represented as a special case of the general framework in Section 1 by letting $z = L = q$, $x = \lambda$, $y = W = w$ and $f(x, y) = xy$ (or $z = W = w$, $x = \lambda^{-1}$, $y = L = q$ and $f(x, y) = y/x$). Of course, the $L = \lambda W$ framework does not fit into Section 1 exactly as given because the data for q consist of the continuous time queue length process instead of some sequence $\{Z_n: n \geq 1\}$, but the ideas easily extend to cover this modification. As we will show, the key point is that the arrival rate λ is often known in advance, so that there is an opportunity to exploit it in the statistical estimation of q and w. For example, in an open network of queues, λ is typically known in a simulation experiment. In a system measurement context, indirect estimation may also be relevant. For example, suppose that we are estimating q for a newly installed telephone switching system. The historical calling record might be used to obtain a highly accurate estimate for λ. Moreover, the arrival process might be unchanged by the addition of the new switching system, even though the service mechanism and, thus, the waiting times and queue-length process, would typically be very different. On the other hand, for a closed network of queues, λ is typically unknown even in a simulation experiment. In either case, but especially (as it turns out) when λ is known, it is clearly important to know whether the direct or the indirect estimator is more efficient. It is also important to know if there are even more efficient estimators.

We close this section by remarking that (10) is

weaker than the conditions in Glynn and Whitt (1986, 1987, 1988) for the continuous time processes $(N(t), \int_0^t Q(s)\,ds)$ to satisfy a joint CLT, i.e., for

$$t^{-1/2}\left(N(t) - \lambda t, \int_0^t Q(s)\,ds - qt\right) \Rightarrow N(0, C^*) \qquad (12)$$

for some covariance matrix C^*. If the joint CLT (10) is strengthened to a joint FCLT (functional central limit theorem), then $(N(t), \int_0^t Q(s)\,ds)$ obeys a joint FCLT, and thus also (12), by Theorem 4 of Glynn and Whitt (1986). If, instead, the sequence $\{(A_k - A_{k-1}, W_k): k \geq 1\}$ is stationary in addition to (10), then (12) holds by Theorem 1 of Glynn and Whitt (1988). (In either case, we can then express C^* in terms of C.) We will introduce the stronger FCLT hypothesis in Section 5, but first, in Section 4 we see what can be done only with (10). Many of the results in this paper can be obtained without considering the continuous time processes, and so do not depend on the extra FCLT assumption.

3. Main Conclusions

Now that we have formulated a general indirect estimation framework (Section 1) and the $L = \lambda W$ framework (Section 2), we can state our main conclusions more precisely. We use the new notation: q for L and w for W. Six general principles emerge from our analysis:

P1. In the $L = \lambda W$ framework, the asymptotic efficiency of the natural and direct estimators for q coincide (and similarly for w). This principle very much depends on the special structure of the queueing model; it is valid in the $L = \lambda W$ framework of Section 2, but not in the general indirect estimation framework of Section 1. This nice property occurs for queueing systems because the relation $L = \lambda W$, when fully developed, embodies much more than a relation (2) among the parameters q, λ and w. Properly interpreted, the relation $L = \lambda W$ also entails a relation among the associated stochastic processes; see Theorem 1 of Glynn and Whitt (1986). As a consequence, *in the queueing context* we are able to show that

$$\mathbf{z}_n^N = \mathbf{z}_n^D + o_P(n^{-1/2}), \qquad (13)$$

which means that modulo a term that converges in probability to zero after dividing by $n^{-1/2}$, $\mathbf{z}_n^N$ is equal to $\mathbf{z}_n^D$; i.e., (13) means that

$$n^{1/2}(\mathbf{z}_n^N - \mathbf{z}_n^D) \Rightarrow 0 \quad \text{as } n \to \infty. \qquad (14)$$

As a consequence of (14), $\sigma_N^2 = \sigma_D^2$, but (13) and (14) are much stronger than the separate convergence in (6) with $\sigma_N^2 = \sigma_D^2$. Formulas 13 and 14 say that in a large-sample context $n^{1/2}(\mathbf{z}_n^N - z)$ and $n^{1/2}(\mathbf{z}_n^D - z)$ are essentially the same random variable. They will have approximately the same value for any sufficiently long segment of any realization of the underlying stochastic process. (It is trivial that $\mathbf{z}_n^N$ and $\mathbf{z}_n^D$ are essentially the same random variable for large n, both being nearly z.)

In queueing, this principle means that we can restrict attention to the direct and indirect estimators; i.e., the fundamental issue as far as asymptotic efficiency is concerned is whether to use λ instead of an estimator for λ when λ is known. As far as asymptotic efficiency is concerned, there is no advantage or disadvantage to using a segment of the continuous time queue length process instead of a segment of the discrete time waiting time sequence, given the same basic data.

P2. In the special case of $L = \lambda W$, we can exploit the special structure of f in (2) to write $w = \lambda^{-1}q$ and relate the asymptotic efficiency parameters of w and q. Let $\sigma^2(q)$ and $\sigma^2(w)$ denote the asymptotic efficiency parameters of estimators of q and w, respectively, with subscripts indicating the method. In particular, it is immediate that

$$\sigma_I^2(q) = \lambda^2\sigma_N^2(w) \quad \text{and} \quad \sigma_I^2(w) = \lambda^{-2}\sigma_N^2(q),$$

but $\sigma_D^2(w) = \sigma_N^2(w)$ and $\sigma_D^2(q) = \sigma_N^2(q)$ by P1 above, so that

$$\sigma_I^2(q) - \sigma_D^2(q) = \lambda^2(\sigma_D^2(w) - \sigma_I^2(w)).$$

As a consequence, the indirect estimator for q is more efficient than the direct estimator for q if and only if the direct estimator for w is more efficient than the indirect estimator for w. Moreover, we will show that the estimators for q and w are actually related in the stronger sense of (13) and (14); see (20). As a consequence, as far as asymptotic efficiency is concerned, it suffices to consider only one of q and w.

P3. In considerable generality, the indirect estimator for q (the direct estimator for w) is asymptotically more efficient than the direct estimator for q (the indirect estimator of w); see Section 7. In particular, this is true provided the waiting times and interarrival times are negatively correlated, which, in turn, is true of the conditional expected waiting time given that an interarrival time is a nonincreasing function of the interarrival time, for all interarrival times and waiting times, which we would expect to be satisfied in most queueing systems. This conclusion is valid for the

standard GI/G/s queue with the first-come first-served discipline, as was shown by Carson and Law. Moreover, we show that the independence and identical-distribution assumptions of the GI/G/s model can be relaxed (Theorem 8).

P4. The time scale has no stochastic effect on the asymptotic efficiency of the estimators. In other words, under mild regularity conditions, it does not matter whether we collect data for a specified number n of customers, for a random number of customers, for a specified time interval $[0, t]$ or for a random time interval. There is a deterministic effect due to the expected average number of customers arriving; e.g., in $[0, t]$ it is λt (asymptotically). As a consequence, working with the time interval $[0, t]$ instead of n simply causes each asymptotic efficiency parameter to be multiplied by λ^{-1} (see Section 6). It thus has no effect on a comparison of estimators in terms of their asymptotic efficiency. (This effect is the same as caused by regenerative cycles in regenerative simulation.)

P5. As indicated in answer **A6** to question **Q6**, the relationship between indirect and direct estimation is fruitfully viewed from the perspective of control variables. This is true in the general framework of Section 1. In particular, direct estimation can be identified with estimation using nonlinear control variables, whereas indirect estimation corresponds to not using the control variable. Nonlinear control-variable schemes have been proposed for statistical estimation in simulation experiments by Kleijnen and Nelson. However, we show that under mild regularity conditions a nonlinear control-variable scheme is equivalent to an associated linear control-variable scheme as far as the asymptotic efficiency is concerned. Finally, by optimizing over the linear control variable schemes in the usual way, we obtain an estimator that is at least as good and usually strictly better than either the direct or indirect estimator, with the criterion of asymptotic efficiency. In fact, in a strong sense, the optimal linear control estimators are best possible in the $L = \lambda W$ framework of Section 2 with the criterion of asymptotic efficiency; see Theorem 10. Similarly, the natural estimators are best possible when λ is unknown. This optimality goes beyond optimality within the linear control-variable framework because our $L = \lambda W$ framework in Section 2 is more general. However, for a particular model with additional structure (e.g., when service times with known mean are specified too), it is typically possible to do even better by using multiple control variables; see Section 10 and Lavenberg, Moeller and Sauer.

P6. As indicated in answer **A7**, asymptotic bias is usually asymptotically negligible compared to asymptotic efficiency. For example, consider the direct estimator $\mathbf{z}_n^D$ in (4): the bias is $(E\mathbf{z}_n^D - z)$. Asymptotic analysis, based on reasonable regularity conditions, shows that

$$E\mathbf{z}_n^D - z = n^{-1}\beta_D + o(n^{-1})$$

or, equivalently, that $n(E\mathbf{z}_n^D - z) \to \beta_D$ as $n \to \infty$. The main conclusion upon comparison with (6) is that the bias is indeed asymptotically negligible compared to the efficiency as $n \to \infty$. We also describe the asymptotic bias parameter β_D in more detail. In particular, we can identify separate contributions to the bias, i.e.,

$$\beta_D = \beta_x + \beta_y + \beta_{f''}$$

where β_x and β_y are the contributions due to the initial bias of $\mathbf{x}_n^N$ and $\mathbf{y}_n^N$, respectively (which also involves f in 2), and $\beta_{f''}$ is the contribution due to the nonlinearity of f (which also involves $(\mathbf{x}_n^N, \mathbf{y}_n^N)$). The linear control estimators have the advantage that $\beta_{f''}$ vanishes.

Here is how the rest of this paper is organized. Sections 4 and 5 discuss estimation in the customer (discrete) time scale, and Section 6 discusses estimation in the intrinsic (continuous) time scale (establishing **P4**). Section 4 explores what can be done given the basic CLT assumption (10), while Sections 5 and 6 require the addition of a stronger FCLT assumption to treat the continuous time processes $Q(t)$ and $N(t)$ in (8) and (9).

Section 7 is devoted to the question of when indirect estimation is asymptotically more efficient than direct estimation (**P3**). Section 8 returns to the general framework in Section 1, establishes the connection to control variable estimation and develops the new, more efficient estimator (**P5**). Section 9 investigates asymptotic bias, supporting the conclusions in **A7** and **P6**.

Finally, Section 10 illustrates the value of the general framework in Section 1 and the associated general results in Sections 8 and 9 by studying the asymptotic efficiency and the asymptotic bias of estimators of the time-average limit of the workload in a GI/G/s queue at time t, using the extension of $L = \lambda W$ to $H = \lambda G$; e.g., see pp. 408–412 of Heyman and Sobel. This example shows how we can exploit a known service time distribution as well as a known arrival rate.

4. Estimation in the Customer Time Scale

Consider the $L = \lambda W$ framework in Section 2 with the basic CLT assumption (10). Suppose that we

observe the system over the time interval required for the first n customers to arrive and depart, i.e., over the interval $[0, D_n^{\uparrow}]$ where $D_n^{\uparrow} = \max\{D_k : 1 \leq k \leq n\}$, with the purpose of estimating q. Such an observation may be obtained either by a direct system measurement or by a computer simulation experiment. Over the time interval $[0, D_n^{\uparrow}]$, the variables $W_1, W_2, \ldots, W_n$, and A_n are observable, so that the estimators $n^{-1}A_n$ and $\sum_{k=1}^{n} W_k$ for λ^{-1} and w can be constructed.

Theorem 1 suggests the following *direct estimator* for $q = \lambda w$:

$$\mathbf{q}_n^D = (nA_n^{-1})n^{-1}\sum_{k=1}^{n} W_k = A_n^{-1}\sum_{k=1}^{n} W_k. \tag{15}$$

Note that (15) is a special case of (4), with the modification that $\{Z_n\}$ is replaced by the continuous time process $\{Q(t): t \geq 0\}$. (As remarked in the introduction, the sequence $\{Z_n\}$ or its analog $\{Q(t)t \geq 0\}$ is not needed for the direct and indirect estimators; here the basic sequence $\{(X_n, Y_n): n \geq 1\}$ for the general framework in Section 1 is defined by $X_n = A_n - A_{n-1}$ and $Y_n = W_n$, $n \geq 1$, using the data we have chosen to observe.) Clearly, the basic assumption (10) implies that $\mathbf{q}_n^D$ is a consistent estimator for q, in the sense that $\mathbf{q}_n^D \Rightarrow q$. The rate of convergence of q_n^D to q is described by the following CLT. Let $=_d$ denote equality in distribution.

Theorem 2. *Under* (10), $n^{1/2}(\mathbf{q}_n^D - q) \Rightarrow N(0, \sigma_D^2)$ *where* $\sigma_D^2 = \lambda^2(q^2C_{11} - 2qC_{12} + C_{22})$.

Proof. Note that

$$n^{1/2}[\mathbf{q}_n^D - q]$$

$$= n^{1/2}A_n^{-1}\left[\sum_{k=1}^{n} W_k - qA_n\right]$$

$$= (n/A_n)n^{-1/2}\left(A_n - n\lambda^{-1}, \sum_{k=1}^{n} W_k - nw\right)\cdot\begin{pmatrix}-q\\1\end{pmatrix}.$$

By (10) and Theorem 5.1 of Billingsley,

$$n^{-1/2}\left(A_n - n\lambda^{-1}, \sum_{k=1}^{n} W_k - nw\right)\cdot\begin{pmatrix}-q\\1\end{pmatrix}$$

$$\Rightarrow N(0, C)\cdot\begin{pmatrix}-q\\1\end{pmatrix} =_d \lambda^{-1}\sigma_D N(0, 1).$$

Recalling that $A_n/n \Rightarrow \lambda^{-1}$, we apply Theorems 4.1 and 4.4 of Billingsley to complete the proof.

If λ is known, then we may use instead the following indirect estimator of q:

$$\mathbf{q}_n^I = \lambda n^{-1}\sum_{k=1}^{n} W_k. \tag{16}$$

A CLT for $\mathbf{q}_n^I$ follows immediately from (10): under (10) $n^{1/2}(\mathbf{q}_n^I - q) \Rightarrow N(0, \sigma_I^2)$, where $\sigma_I^2 = \lambda^2 C_{22}$. We say that the indirect estimator $\mathbf{q}_n^I$ is more asymptotically efficient for estimating q than the direct estimator $\mathbf{q}_n^D$ if $\sigma_I^2 < \sigma_D^2$. In Section 7 we investigate when this inequality holds.

Example 1b

For the M/M/1 queue, since we let W_n be the waiting time before beginning service, the associated queue length process counts the number of customers waiting, excluding any in service, and has steady state mean $q = \rho^2(1 - \rho)$. The asymptotic efficiency parameters for the direct and indirect estimators of q are

$$\sigma_D^2 = 2\rho^3(1 + 4\rho - 4\rho^2 + \rho^3)/(1 - \rho)^4$$

and

$$\sigma_I^2 = \rho^2 C_{22} = \rho^3(2 + 5\rho - 4\rho^2 + \rho^3)/(1 - \rho)^4.$$

The variance reduction using the indirect estimator is

$$\sigma_D^2 - \sigma_I^2 = \rho^2(q^2C_{11} - 2qC_{12}) = \rho^4(3 - \rho)/(1 - \rho)^3$$

and the relative savings is

$$\frac{\sigma_D^2 - \sigma_I^2}{\sigma_D^2} = \frac{\rho(1 - \rho)(3 - \rho)}{2 + 8\rho - 8\rho^2 + 2\rho^3},$$

which converges to 0 as ρ approaches both 0 and 1. For $\rho = 0.5$, 0.7, 0.8 and 0.9, the relative savings is 0.15, 0.11, 0.08, and 0.05, respectively.

5. Using the Stronger FCLT Assumption

Given the data over $[0, D_n^{\uparrow}]$ in Section 4, we also can construct the natural estimator for q, namely

$$\mathbf{q}_n^N = (D_n^{\uparrow})^{-1}\int_0^{D_n^{\uparrow}} Q(s)\, ds. \tag{17}$$

However, to treat (17) we need a stronger condition than (10). In particular, we introduce the FCLT condition in Glynn and Whitt (1986). (An alternative is to assume stationarity and invoke Theorem 1 of Glynn and Whitt 1988.) For this purpose, let $D[0, 1]$ be the function space of all right-continuous real-valued functions on [0, 1] with left limits everywhere,

as in Chapter 3 of Billingsley. Then let

$$\mathbf{A}_n(t) = n^{-1/2}[A_{[nt]} - \lambda^{-1}nt]$$

and

$$\mathbf{W}_n(t) = n^{-1/2}\left[\sum_{k=1}^{[nt]} W_k - wnt\right], \quad 0 \leq t \leq 1 \tag{18}$$

by random elements of $D[0, 1]$, where $[x]$ is the greatest integer less than or equal to x. Also let $\mathbf{B}(b, C)(t)$, $0 \leq t \leq 1$, represent two-dimensional Brownian motion on $D[0, 1]^2 \equiv D[0, 1] \times D[0, 1]$ with drift vector $b = (b_1, b_2)$ and covariance matrix C. ($B(b, C)$ can be represented as $B(b, I)M$ where I is the identity matrix, M is a 2×2 rotation matrix such that $C = M^T DM$, D is a diagonal matrix and $B(b, I)$ is standard two-dimensional Brownian motion with independent marginal processes (p. 365 of Karlin and Taylor (1975) and pp. 70, 84 of Feller (1971); for each t, $B(b, C)(t)$ is distributed as $N(tb, tC)$.)

Stronger FCLT Assumption

The sequence $\{(\mathbf{A}_n, \mathbf{W}_n): n \geq 1\}$ in (18) satisfies a joint FCLT: there exist constants λ and w with $0 < \lambda$, $w < \infty$ and a covariance matrix $C = \{C_{ij}: 1 \leq i, j \leq 2\}$ such that

$$(\mathbf{A}_n, \mathbf{W}_n) \Rightarrow \mathbf{B}(0, C) \quad \text{in } D[0, 1]^2. \tag{19}$$

Note that (19) implies (10), by applying the continuous mapping theorem with the projection map π, defined by $\pi(x, y) = (x(1), y(1))$ for $x, y \in D[0, 1]$, so that (19) is indeed a stronger condition. However, for practical purposes there is little difference between (19) and (10). It is only in pathological situations that (10) holds without (19); e.g., see Glynn and Whitt (1988).

In order to treat $\mathbf{q}_n^N$ in (17), we use the stronger FCLT condition (19). As a consequence, we also get an FCLT conclusion, but we do not state it.

Theorem 3. *Under* (19), $n^{-1}D_n^{\uparrow} \Rightarrow \lambda^{-1}$, $n^{1/2}(\mathbf{q}_n^D - \mathbf{q}_n^N) \Rightarrow 0$, *and* $n^{1/2}(\mathbf{q}_n^N - q) \Rightarrow N(0, \sigma_D^2)$ *for* σ_D^2 *in Theorem* 2.

Proof. This follows easily from Theorem 4 of Glynn and Whitt (1986). The functional version of the first limit is contained directly there. The second limit, which corresponds to (13) and (14), follows from the third limit and the limit in Theorem 2 holding jointly with the same limit random variable, i.e.,

$$n^{1/2}(\mathbf{q}_n^D - q, \mathbf{q}_n^N - q) \Rightarrow (X, X) \quad \text{in } R^2$$

where $X =_d (0, \sigma_D^2)$. To establish this joint limit, use (19) and Theorem 4 of Glynn and Whitt (1986) to get

$$\left\{n^{-1}A_n, n^{-1}D_n^{\uparrow}, n^{-1/2}\left(A_n - n\lambda^{-1}, D_n^{\uparrow} - n\lambda^{-1}, \sum_{k=1}^{n} W_k - nw, \int_0^{D_n^{\uparrow}} Q(s)\,ds - nw\right)\right\},$$

$$\Rightarrow (\lambda^{-1}, \lambda^{-1}, A, A, W, W) \quad \text{in } R^6.$$

Then apply the argument in the proof of Theorem 2 twice to get the desired joint convergence in R^2 with limit $X = \lambda(-qA + W) =_d N(0, \sigma_D^2)$.

So far, we have discussed only estimators for q. Estimators for w can be treated in the same way; in fact, we can apply the previous results for the estimators $\mathbf{q}_n^N$, $\mathbf{q}_n^D$ and $\mathbf{q}_n^I$ of q. Simply observe that for the data over $[0, D_n^{\uparrow}]$, the corresponding three estimators for w are, by definition according to (3)–(5),

$$\mathbf{w}_n^N = n^{-1}\sum_{k=1}^{n} W_k = \lambda^{-1}\mathbf{q}_n^I,$$

$$\mathbf{w}_n^D = (n^{-1}A_n)\mathbf{q}_n^N \quad \text{and} \quad \mathbf{w}_n^I = \lambda^{-1}\mathbf{q}_n^N. \tag{20}$$

Only $\mathbf{w}_n^D$ in (20) requires some additional discussion: Since $\mathbf{w}_n^N = n^{-1}A_n\mathbf{q}_n^D$ and

$$\begin{aligned}\mathbf{w}_n^D &= (n^{-1}A_n)\mathbf{q}_n^N \\ &= (n^{-1}A_n - \lambda^{-1})(\mathbf{q}_n^N - \mathbf{q}_n^D) \\ &\quad + \lambda^{-1}(\mathbf{q}_n^N - \mathbf{q}_n^D) + n^{-1}A_n\mathbf{q}_n^D,\end{aligned}$$

$\mathbf{w}_n^D = \mathbf{w}_n^N + o_p(n^{-1/2})$ under (19) by (11) and Theorem 3, which in turn implies that $\mathbf{w}_n^D = \lambda^{-1}\mathbf{q}_n^I + o_p(n^{-1/2})$ under (19); i.e., we have established **P1** and **P2** in Section 3 for the estimators of w.

6. Estimation in the Intrinsic Time Scale

Now suppose that we observe the queue over the time interval $[0, t]$, again for the purpose of estimating q. It turns out that, from the point of view of asymptotic efficiency, this change in the basic data corresponds to a deterministic time transformation. As $n \to \infty$, the number of arrivals in $[0, D_n^{\uparrow}]$ is approximately n; i.e., $n^{-1}N(D_n^{\uparrow}) \Rightarrow 1$. On the other hand, as $t \to \infty$, the number of arrivals in $[0, t]$ is approximately λt; i.e., $t^{-1}N(t) \Rightarrow \lambda$. By changing to the intrinsic time scale, the limit theorems in Sections 4 and 5 are modified as if we replaced the number of customers, n, by λn in the customer time scale without changing the normalization to $(\lambda n)^{1/2}$; i.e., the change corresponds to $n^{1/2}(\mathbf{q}_{\lambda n}^D - q) \Rightarrow \lambda^{-1/2}N(0, \sigma_D^2) =_d N(0, \lambda^{-1}\sigma_D^2)$. See

Lemma 1 in Section 5 of Glynn and Whitt (1986) for additional theoretical justification.

Let $O(t)$ be the number of customers to depart the system by time t, which we can define by

$$O(t) = \sum_{k=1}^{\infty} I(D_k \leq t), \quad t \geq 0,$$

where $I(B)$ is the indicator function as in (8). Over the interval $[0, t]$, $W_1, W_2, \ldots, W_{O(t)}$, and $A_{O(t)}$ are observable, so that paralleling (15), we can define the direct estimator

$$\mathbf{q}^D(t) = \mathbf{q}^D_{O(t)} = A^{-1}_{O(t)} \sum_{k=1}^{O(t)} W_k. \tag{21}$$

From Theorem 1, we have $t^{-1}N(t) \Rightarrow \lambda$ and we should expect to have $t^{-1}O(t) \Rightarrow \lambda$ as well. However, this is less obvious; it holds under the conditions of Theorem 1 by Theorem 3 of Glynn and Whitt (1988).

It is to be expected from Theorem 2 that

$$O(t)^{1/2}(\mathbf{q}^D_{O(t)} - q) \Rightarrow \sigma_D D(0, 1). \tag{22}$$

In fact, (22) can be justified under (19), by appealing to Theorem 4 of Glynn and Whitt (1986). By combining (22) with the converging-together theorem (Theorem 4.1 of Billingsley) or by appealing directly to Theorem 4 of Glynn and Whitt (1986), we obtain the following CLT.

Theorem 4. *Under* (19), $t^{1/2}(\mathbf{q}^D(t) - q) \Rightarrow N(0, \lambda^{-1}\sigma_D^2)$ *for* σ_D^2 *in Theorem* 2.

A variety of other estimators can be constructed from data observed over the interval $[0, t]$, all of which are asymptotically equivalent to $\mathbf{q}^D(t)$. Let

$$\mathbf{q}^D_1(t) = \mathbf{q}^N(t) = t^{-1} \int_0^t Q(s)\, ds,$$

$$\mathbf{q}^D_2(t) = (A_{N(t)})^{-1} \sum_{k=1}^{O(t)} W_k$$

and

$$\mathbf{q}^D_3(t) = t^{-1} \sum_{k=1}^{O(t)} W_k.$$

The next CLT also follows from Theorem 4 in Glynn and Whitt (1986). The fact that the natural estimator $\mathbf{q}^N(t) = \mathbf{q}^D_1(t)$ has the same asymptotic efficiency parameter $\lambda^{-1}\sigma_D^2$ as $\mathbf{q}^D(t)$ establishes principle **P1** in the intrinsic time scale.

Theorem 5. *Under* (19), *for each* i, $\mathbf{q}^D_i(t) = \mathbf{q}^D(t) + o_p(t^{-1/2})$ *and* $t^{1/2}(\mathbf{q}^D_i(t) - q) \Rightarrow N(0, \lambda^{-1}\sigma_D^2)$.

As in Section 2, an indirect estimator of q is also available on the intrinsic time scale. In particular, assuming that λ is known, let

$$\mathbf{q}^I(t) = \lambda O(t)^{-1} \sum_{k=1}^{O(t)} W_k. \tag{24}$$

The following CLT has a proof similar to that of Theorem 4.

Theorem 6. *Under* (19), $t^{1/2}(\mathbf{q}^I(t) - q) \Rightarrow N(0, \lambda^{-1}\sigma_I^2)$ *for* $\sigma_I^2 = \lambda^2 C_{22}$ *as in Section* 4.

A corresponding simple transformation of the asymptotic efficiency parameters occurs if we measure time in regenerative cycles, as is often done in regenerative simulation; see p. 210 of Glynn and Whitt (1986).

Example 1c

For the M/M/1 queue, the asymptotic efficiency of the direct estimator of q in the discrete time scale is $\sigma_D^2 = 2\rho^3(1 + 4\rho - 4\rho^2 + \rho^3)/(1 - \rho)^4$. In the intrinsic time scale it is $\rho^{-1}\sigma_D^2$. If, instead, we want to measure time in busy cycles, as in regenerative simulation, then to obtain the asymptotic efficiency of the direct estimator of q we simply divide by the expected number of customers served in a busy cycle and get $(1 - \rho)\sigma_D^2$. This is $VAD(\hat{d}_2)$ in (2.2) of Law (1975).

7. Comparing the Asymptotic Efficiency

We compare the asymptotic efficiency of direct and indirect estimators by comparing the quantities σ_D^2 and σ_I^2. Since

$$\sigma_D^2 = \sigma_I^2 - 2q\lambda^2 C_{12} + q^2\lambda^2 C_{11},$$

a sufficient condition for $\sigma_I^2 \leq \sigma_D^2$ is $C_{12} \leq 0$. In fact, it is common to have $C_{12} \leq 0$, as we will show. Then $\sigma_D^2 - \sigma_I^2$ is bounded below by $q^2\lambda^2 C_{11}$, and the reduction in the variance relative to the square of the estimate q is

$$(\sigma_D^2 - \sigma_I^2)/q^2 \geq \lambda^2 C V_{11} = c_A^2$$

where c_A^2 is the asymptotic variability parameter of the arrival process, which plays a prominent role in heavy-traffic limit theorems and approximations; see Whitt (1982). The asymptotic variability parameter c_A^2 thus seems to be a good indicator of the relative variance reduction. This is illustrated by the Q data in Tables I and II of Carson and Law. We would expect significant variance reduction from the indirect estimation of q in models with very bursty arrival

processes as measured by c_A^2, e.g., $c_A^2 = 18$ for a multiplexer in Sriram and Whitt (1986).

To analyze the sign of C_{12}, it is convenient to impose the following uniform integrability (UI) assumption; see p. 32 of Billingsley or Section 4.5 of Chung (1974). As usual, the purpose of UI is to get convergence of moments from convergence in distribution.

Assumption UI. The sequences $\{n^{-1}(A_n - n\lambda^{-1})^2: n \geq 1\}$ and $\{n^{-1}(\sum_{k=1}^n W_k - nw)^2: n \geq 1\}$ are uniformly integrable.

Assumption **UI** immediately implies uniform integrability of several related sequences.

Proposition 1. *Under Assumption* **UI**, *the sequences*

$$\{n^{-1/2}(A_n - n\lambda^{-1}): n \geq 1\},$$

$$\left\{n^{-1/2}\left(\sum_{k=1}^n W_k - nw\right): n \geq 1\right\}$$

and

$$\left\{n^{-1}\left[(A_n - n\lambda^{-1})\left(\sum_{k=1}^n W_k - nw\right)\right]: n \geq 1\right\}$$

are uniformly integrable.

Proof. Observe that

$$n^{-1/2}\,|A_n - n^{-1}|$$

$$\leq 1 + n^{-1}(A_n - n\lambda^{-1})^2,$$

$$n^{-1/2}\left|\sum_{k=1}^n W_k - nw\right|$$

$$\leq 1 + n^{-1}\left(\sum_{k=1}^n W_k - nw\right)^2,$$

$$n^{-1}\left|(A_n - n\lambda^{-1})\left(\sum_{k=1}^n W_k - nw\right)\right|$$

$$\leq 2^{-1}\left[n^{-1}(A_n - n\lambda^{-1})^2 + n^{-1}\left(\sum_{k=1}^n W_k - nw\right)^2\right].$$

Thus, the three sequences in question are dominated by uniformly integrable sequences, which implies uniform integrability of the dominated sequences.

The continuous mapping theorem applied to (10) shows that

$$n^{-1/2}(A_n - n\lambda^{-1}) \Rightarrow Y_1,$$

$$n^{-1/2}\left(\sum_{k=1}^n W_k - nw\right) \Rightarrow Y_2$$

$$n^{-1}\left[(A_n - n\lambda^{-1})\left(\sum_{k=1}^n W_k - nw\right)\right] \Rightarrow Y_1Y_2$$

where (Y_1, Y_2) is distributed as $N(0, C)$. The uniform integrability provided by Proposition 1 allows us to pass the expectation through (25), yielding

$$n^{-1/2}E(A_n - n\lambda^{-1}) \to EY_1 = 0$$

$$n^{-1/2}E\left(\sum_{k=1}^n W_k - nw\right) \to EY_2 = 0$$

$$n^{-1}E\left[(A_n - n\lambda^{-1})\left(\sum_{k=1}^n W_k - nw\right)\right] \to EY_1Y_2 = C_{12}$$

as $n \to \infty$. Let $U_j = A_j - A_{j-1}$ be an interarrival time. We thus have

$$C_{12} = \lim_{n\to\infty} n^{-1}\mathrm{cov}\left[A_n, \sum_{k=1}^n W_k\right]$$

$$= \lim_{n\to\infty} n^{-1}\sum_{j=1}^n\sum_{k=1}^n \mathrm{cov}(U_j, W_k). \tag{26}$$

under Assumptions 10 and **UI**. To obtain nonpositivity of C_{12}, it, therefore, suffices to show that $\mathrm{cov}(U_j, W_k) \leq 0$ for all j and k. Since

$$\mathrm{cov}(U_j, W_k) = \mathrm{cov}(U_j, h_{jk}(U_j))$$

where

$$h_{jk}(u) = E(W_k \mid U_j = u) \tag{27}$$

by Theorem 9.1.3 of Chung (1974), we obtain the following result. For the proof, recall that a family of random variables $\{X_n\}$ is *associated* if $\mathrm{cov}(f(\{X_n\}), g(\{X_n\})) \geq 0$ for all pairs of nondecreasing real-valued functions f and g such that the expectations exist; see p. 29 of Barlow and Proschan (1975).

Theorem 7. *If, in addition to Assumptions* 10 *and* **UI**, $h_{jk}(u)$ *in* (27) *is a nonincreasing function of* u *for all* j *and* k, *then* $\sigma_I^2 \leq \sigma_D^2$.

Proof. We need only observe that if f and g are nondecreasing functions, then $\mathrm{cov}[f(X), g(X)] \geq 0$ provided that the covariance is well defined. In other words, a single random variable X is associated. In fact, $(f(X), g(X))$ then has the maximum possible

covariance among bivariate distributions with marginals distributed as $f(X)$ and $g(X)$ by virtue of the rearrangement theorem of Hardy, Littlewood and Polya (1952); see Whitt (1976): Theorem 2.1 and Lemmas 2.2 and 2.3 also establish the desired inequality.

From a practical view, Theorem 7 says that indirect estimation of q is more efficient than direct estimation provided that the waiting times tend to be nonincreasing functions of the interarrival times ($h_{jk}(u)$ in (27) involves an expectation and (26) involves an average). The conditions of Theorem 7 hold in many queues, including the standard GI/G/s queue with the FCFS (first-come first-served) discipline when W_k is the waiting time of the kth customer until beginning service (or the sojourn time, including service time), as was shown by Carson and Law, Lemma 3.

Our next result shows that the GI/G/s result in Carson and Law holds with the usual distributional assumptions (independence and identical distributions) greatly relaxed.

Theorem 8. *Consider the standard multiserver queueing model with unlimited waiting room and the FCFS discipline. If, in addition to Assumptions* 10 *and* **UI**, *either a) the family* $\{U_k, -V_k: k \geq 1\}$ *is associated, where* V_k *is the service time of the kth customer, or b) the service times are independent of the interarrival times and the interarrival times* $\{U_k\}$ *alone are associated, then* $Cov(U_j, W_k) \leq 0$ *for all* j *and* k, *so that* $\sigma_I^2 \leq \sigma_D^2$.

Proof. Kiefer and Wolfowitz (1955) proved that, for each sample path of interarrival time and service times, W_k is a nonincreasing function of the first $k-1$ interarrival times and a nondecreasing function of the first $k-1$ service times. Hence,

$$\begin{aligned}&\mathrm{Cov}(U_j, W_k)\\ &\quad = \mathrm{Cov}(U_j, f_k(U_1, \ldots, U_{k-1}, -V_1, \ldots, -V_{k-1}))\end{aligned}$$

where f_k: $R^{2k-2} \to R$ is nonincreasing. Since $\{(U_k, -V_k)\}$ is associated by assumption in (a), the conclusion follows. Under (b), we can first condition on the service times. By the assumed independence between $\{U_k\}$ and $\{V_k\}$, the conditional interarrival times given the service times are associated. Hence, we have

$$\mathrm{Cov}(U_j, W_k) = \mathrm{Cov}(U_j, g_k(U_1, \ldots, U_{k-1}))$$

where there is an implicit conditioning on the service times. However, g_k is nonincreasing for each realization of the service times, so that $\mathrm{Cov}(U_j, W_k)$ conditional on the service times is nonpositive. Finally, we obtain the desired conclusion by unconditioning.

Note that in Theorem 8 neither the service nor the interarrival times need be either independent or identically distributed. Of course, independence for random variables is a sufficient condition for them to be associated.

We now give examples in which each of the basic estimators for q is strongly preferred.

Example 2

Consider the standard GI/D/∞ service system, focusing on the number of customers in service. Obviously, $n^{-1}\sum_{k=1}^{n} W_k = w$ for each n so that $C_{22} = C_{12} = 0$. In general, however, A_n has variance so that $C_{11} > 0$. In this case, $\sigma_I^2 = 0$ whereas $\sigma_D^2 = \lambda^2 q^2 C_{11} > 0$.

Example 3

We now show that the reverse of Example 2 can prevail: it can be much better to use the direct estimator $\mathbf{q}_n^D$ than the indirect estimator $\mathbf{q}_n^I$. Suppose that $Q(t)$ is the number of jobs in service at a single-server facility, so that $Q(t) \in \{0, 1\}$. If there are an unlimited number of jobs to be processed at this facility, then the server is always busy so that $Q(t) = 1$ for all t w.p.1. Furthermore, the arrival time A_{k+1} of the $(k+1)$st customer is the departure time of the kth customer, so that $W_k = U_{k+1}$. Thus,

$$\mathbf{q}_n^D = A_n^{-1} \sum_{k=1}^{n} W_k = (A_{n+1} - A_1)/A_n. \tag{28}$$

If the interarrival times $\{U_k: k \geq 1\}$ are independent and identically distributed, with finite variance, it is easily shown using (28) that $n^{1/2}(\mathbf{q}_n^D - 1) \Rightarrow 0$ as $n \to \infty$, implying that $\sigma_D^2 = 0$. On the other hand, $\mathbf{q}_n^I = \lambda n^{-1}\sum_{k=1}^{n} W_k$ clearly satisfies $n^{1/2}(\mathbf{q}_n^I - 1) \Rightarrow N(0, \sigma_I^2)$, where $\sigma_I^2 = \lambda^2 \mathrm{var}(U_1)$. Note that in this example W_k and U_{k+1} are positively correlated; in fact, their correlation is 1, the maximum possible. Also, observe that this example arises approximately with a GI/G/1 queue in heavy traffic if the system we focus on consists of the server.

Before concluding this section, we discuss conditions guaranteeing Assumption **UI**. Frequently, the conditions necessary to obtain **UI** also imply (19) (and thus 10).

Proposition 2. *Let* $\{X_n: n \geq 0\}$ *be a real-valued nondelayed regenerative sequence with an associated*

sequence of regeneration times $\{T_n: n \geq 1\}$. *If*

$$E\left(\sum_{k=0}^{T_1-1} (|X_k| + 1)\right)^2 < \infty,$$

then

$$\left\{n^{-1}\left(\sum_{k=1}^{n-1} (X_k - EX_k)\right)^2 : n \geq 1\right\}$$

is uniformly integrable and an FCLT holds for the partial sums.

For the proof of uniform integrability in Proposition 2, see Chung (1966), p. 102; for the FCLT, see Freedman (1967). Both proofs are given for discrete time Markov chains but generalize without difficulty to the regenerative case.

Proposition 3. *Let* $\{X_n: n \geq 0\}$ *be a strictly stationary* ϕ*-mixing sequence via* ϕ*-mixing coefficients satisfying* $\sum_{n=0}^{\infty} \phi_n^{1/2} < \infty$. *If* $EX_0^2 < \infty$, *then*

$$\left\{n^{-1}\left(\sum_{k=0}^{n-1} X_k - nEX_0\right)^2 : n \geq 1\right\}$$

is uniformly integrable and an FCLT holds for the partial sums.

The proof and more details on the statement of Proposition 3 can be found on pgs. 172–177 of Billingsley. Similar results hold for strongly mixing sequences (Hall and Heyde 1980, p. 132) and associated sequences (Newman and Wright 1981).

8. Nonlinear Control Variables

In this section, we show that the relationship between direct and indirect estimation is fruitfully viewed from a control-variable perspective. This perspective allows us to construct an estimator that is asymptotically more efficient than either the direct or the indirect estimator just discussed. To accomplish this goal, we return to the general framework of Section 1. Since we are not going to consider natural estimators here, we do not need the sequence $\{Z_n\}$ in (1), but to treat asymptotic efficiency we need the CLT (7). Obviously the basic CLT assumption (10) is just (7) in the $L = \lambda W$ framework.

Direct and indirect estimators also arise naturally in nonlinear control-variable schemes, but in a different way. Our starting point is a function of one variable, say g. Given (7), suppose that we are interested in estimating $z = g(y)$ where $g: R^l \to R$. We then introduce a convenient function f and a vector x in R^k such that $f: R^{k+l} \to R$ and $z = g(y) = f(x, y)$ for that special x and all y. The new vector x is the *control parameter*, and the natural estimator for x, $\mathbf{x}_n^N$, is the *control*. The *uncontrolled estimator* for z is $g(\mathbf{y}_n^N)$, which coincides with the indirect estimator $f(x, \mathbf{y}_n^N)$. The *controlled estimator* is $f(\mathbf{x}_n^N, \mathbf{y}_n^N)$, which coincides with the direct estimator.

Below are examples of functions f that have been introduced for this purpose:

$$\begin{aligned}
&\text{(i)} \quad f(u, y) = g(y) + (u - x)a^t \\
&\text{(ii)} \quad f(u, y) = (x/u)g(y) \\
&\text{(iii)} \quad f(u, y) = (u/x)g(y) \\
&\text{(iv)} \quad f(u, y) = g(y)^{(x/u)} \\
&\text{(v)} \quad f(u, y) = g(y)^{(u/x)}
\end{aligned} \tag{29}$$

where u, x and α are row vectors in R^k and α^t is the transpose of α (associated column vector) in i, where $k = 1$ and $x \neq 0$ in ii–v. The key property is that $f(u, y) = g(y)$ when $u = x$. Control scheme i is the standard linear control-variable scheme which has been studied extensively in the literature. The ratio and product controls, ii and iii, can be found in Kleijnen. The power law controls, iv and v, were proposed recently by Nelson.

As we have already seen, direct and indirect estimation arises naturally in the context of Little's Law. As noted in (20), the natural estimator for w based on data over the interval $[0, D_n^{\uparrow}]$ is $\mathbf{w}_n^N = n^{-1} \sum_{i=1}^{n} W_i$ while the indirect estimator is

$$\mathbf{w}_n^I = \lambda^{-1}\mathbf{q}_n^N = \lambda^{-1}(D_n^{\uparrow})^{-1} \int_0^{D_n^{\uparrow}} Q(s)\, ds.$$

However, from Section 5, $\mathbf{w}_n^I = \lambda^{-1}\mathbf{q}_n^D + o_p(n^{-1/2})$ where

$$\lambda^{-1}\mathbf{q}_n^D = \lambda^{-1}A_n^{-1} \sum_{i=1}^{n} W_i = \lambda^{-1}(nA_n^{-1})n^{-1} \sum_{i=1}^{n} W_i$$

is precisely a ratio control estimator for w of the form ii in (29). (Set $g(y) = y = w$ and $x = \lambda^{-1}$.) Thus, $\mathbf{w}_n^I$ and $\mathbf{w}_n^N$ are related to one another via a ratio control scheme and, by our terminology, $\mathbf{w}_n^I$ and $\mathbf{w}_n^N$ are indirect and natural estimators for w, respectively, the latter being asymptotically equivalent to a direct estimator.

Now return to the general framework (1)–(7), and let

$$\nabla_x f(x, y) = \left(\frac{\partial f}{\partial x_1}(x, y), \ldots, \frac{\partial f}{\partial x_k}(x, y)\right)^t$$

and

$$\nabla_y f(x, y) = \left(\frac{\partial f}{\partial y_1}(x, y), \ldots, \frac{\partial f}{\partial y_l}(x, y)\right)^t,$$

so that $\nabla_x f(x, y)$ and $\nabla_y f(x, y)$ are column vectors. Unspecified vectors such as x are taken to be row vectors. Using Taylor's theorem to expand

$$f\left(n^{-1}\sum_{i=1}^{n} X_i, n^{-1}\sum_{i=1}^{n} Y_i\right)$$

around the point (x, y), we find that

$$\mathbf{z}_n^D = f(x, y) + \left(n^{-1}\sum_{i=1}^{n} X_i - x\right)\nabla_x f(\xi_n, \eta_n) + \left(n^{-1}\sum_{i=1}^{n} Y_i - y\right)\nabla_y f(\xi_n, \eta_n) \quad (30)$$

for n sufficiently large, where (ξ_n, η_n) lies on the line segment joining (x, y) with

$$n^{-1}\left(\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} Y_i\right),$$

and so is a random vector. By (1), it follows that $(\xi_n, \eta_n) \Rightarrow (x, y)$ as $n \to \infty$, so that the continuity of the gradient of f in the neighborhood of (x, y) implies that $\nabla_x f(\xi_n, \eta_n) \Rightarrow \nabla_x f(x, y)$ and $\nabla_y f(\xi_n, \eta_n) \Rightarrow \nabla_y f(x, y)$ as $n \to \infty$. Relations (7) and (30) together then show that

$$\mathbf{z}_n^D = z + \left(n^{-1}\sum_{i=1}^{n} X_i - x\right)\nabla_x f(x, y) + \left(n^{-1}\sum_{i=1}^{n} Y_i - y\right)\nabla_y f(x, y) + o_p(n^{-1/2}) \quad (31)$$

so that $n^{1/2}(\mathbf{z}_n^D - z) \Rightarrow N(0, \sigma_D^2)$, where

$$\sigma_D^2 = \nabla_x f(x, y)^t C_{11} \nabla_x f(x, y) + \nabla_y f(x, y)^t C_{21} \nabla_x f(x, y) + \nabla_x f(x, y)^t C_{12} \nabla_y f(x, y) + \nabla_y f(x, y)^t C_{22} \nabla_y f(x, y) \quad (32)$$

with C_{11}, C_{12}, C_{21} and C_{22} being the $k \times k$, $k \times l$, $l \times k$ and $l \times l$ submatrices of the covariance matrix C in (7).

Similarly, we can analyze the indirect estimator $\mathbf{z}_n^I$, thereby finding that

$$\mathbf{z}_n^I = z + \left(n^{-1}\sum_{i=1}^{n} Y_i - y\right)\nabla_y f(x, y) + o_p(n^{-1/2}), \quad (33)$$

so that $n^{1/2}(\mathbf{z}_n^I - z) \Rightarrow N(0, \sigma_I^2)$ where

$$\sigma_I^2 = \nabla_y f(x, y)^t C_{22} \nabla_y f(x, y). \quad (34)$$

We summarize our discussion thus far with the following theorem. (Part c follows immediately from (31) and (33).)

Theorem 9. *If* (7) *holds, then*

(a) $n^{1/2}(\mathbf{z}_n^D - z) \Rightarrow N(0, \sigma_D^2)$, *where* σ_D^2 *is given by* (32);
(b) $n^{1/2}(\mathbf{z}_n^I - z) \Rightarrow N(0, \sigma_I^2)$, *where* σ_I^2 *is given by* (34);
(c) $\mathbf{z}_n^D = \mathbf{z}_n^I + (n^{-1}\sum_{i=1}^{n} X_i - x)\nabla_x f(x, y) + o_p(n^{-1/2})$.

Specialized to Little's Law, i.e., the framework in Section 2, Theorem 9a and b are covered by Section 4. Perhaps the most interesting aspect of Theorem 9 is c, which states that modulo the $o_p(n^{-1/2})$ term (which is asymptotically negligible), we have a representation of $\mathbf{z}_n^D$ as $\mathbf{z}_n^I$ plus a linear control-variable term (see i in (29)). *Thus, direct and indirect estimators are related to one another, essentially (asymptotically), through linear control variable schemes.* In particular, recalling the fact that nonlinear control-variable schemes are but a special case of our framework, we conclude that every nonlinear control-variable method behaves asymptotically like a linear control-variable scheme. In other words, *no improvement in asymptotic efficiency can be achieved by generalizing the notion of control variables from the linear form to a nonlinear setting.* (Such a conclusion is also reached by Cheng and Feast, p. 53.) Of course, this does not preclude the possibility of better performance by nonlinear methods in a small sample context.

Theorem 9c is also the basis for improving the performance of $\mathbf{z}_n^D$. We can consider a general linear control estimator, replacing $\nabla_x f(x, y)$ by some α^t to obtain

$$\mathbf{z}_n^C(\alpha) = \mathbf{z}_n^I + \left(n^{-1}\sum_{i=1}^{n} X_i - x\right)\alpha^t. \quad (35)$$

Theorem 9 and the converging-together theorem (Theorem 4.1 of Billingsley) prove that

$$n^{-1/2}(\mathbf{z}_n^C(\alpha) - z) \Rightarrow N(0, \sigma_C^2(\alpha))$$

where

$$\sigma_C^2(\alpha) = \alpha C_{11}\alpha^t + 2\nabla_y f(x, y)^t C_{21}\alpha^t + \nabla_y f(x, y)^t C_{22}\nabla_y f(x, y). \quad (36)$$

Let α^* minimize the unconstrained quadratic program $\sigma_C^2(\alpha)$ in (36). It is well known (e.g., p. 31 of

Anderson 1958 and the Appendix of Rubinstein and Marcus) that if C_{11} is invertible, then

$$\alpha^* = -\nabla_y f(x, y)^t C_{21} C_{11}^{-1}, \tag{37}$$

so that

$$\sigma_C^2(\alpha^*)$$

$$= \nabla_y f(x, y)^t (C_{22} - C_{21} C_{11}^{-1} C_{12}) \nabla_y f(x, y). \tag{38}$$

If C_{11} is positive definite (and so nonsingular), then we have $\sigma_C^2(\alpha^*) < \sigma_D^2$ in (32) unless $\alpha^* = \nabla_x f(x, y)^t$, in which case $\sigma_C^2(\alpha^*) = \sigma_D^2$. We also have $\sigma_C^2(\alpha^*) < \sigma_I^2$ in (34) unless $C_{12} \nabla_y f(x, y) = 0$, in which case $\sigma_C^2(\alpha^*) = \sigma_I^2$. (See the linear control variable papers for further discussion.)

Specializing this discussion to the estimation of q, we have $k = l = 1$, so that C_{11}, C_{12} and C_{22} are scalars. If $C_{11} \neq 0$, then the controlled estimator for q in the framework of Section 2 that involves the best linear control is defined as

$$\mathbf{q}_n^C(\alpha^*) = \lambda n^{-1} \sum_{k=1}^{n} W_k + \alpha^*(n^{-1} A_n - \lambda^{-1}) \tag{39}$$

for $\alpha^* = -\lambda C_{12}/C_{11}$, which has asymptotic efficiency

$$\sigma_C^2(\alpha^*) = \lambda^2 (C^{22} - C_{12}^2/C_{11}).$$

Of course, usually the covariance matrix C is unknown in advance, so that α^* in (37) or (39) is unknown. However, α^* can be estimated from data. It suffices to use any consistent estimator for α^* in (37) or (39). The converging-together theorem (Theorem 4.1 of Billingsley) establishes the following result.

Corollary. Under (7), if $\mathbf{a}_n \Rightarrow \alpha^*$ in (37), then $n^{1/2}(\mathbf{z}_n^C(\mathbf{a}_n) - z) \Rightarrow N(0, \sigma_C^2(\alpha^*))$ for $\sigma_C^2(\alpha^*)$ in (38), so that the estimation of α^* entails no loss of asymptotic efficiency.

Of course, for a finite sample, there is loss of efficiency; see Lavenberg and Welch. From an asymptotic viewpoint, we conclude that the queueing estimators

$$\mathbf{q}_n^C(\mathbf{a}_n) = \lambda n^{-1} \sum_{k=1}^{n} W_k + \mathbf{a}_n(n^{-1} A_n - \lambda^{-1})$$

and

$$\mathbf{w}_n^C(\mathbf{a}_n) = \mathbf{q}_n^C(\mathbf{a}_n)/\lambda, \tag{40}$$

where $\mathbf{a}_n \Rightarrow \alpha^* = -\lambda C_{12}/C_{11}$ are more efficient than any of the direct or indirect estimators for q and w: From the construction, $\sigma_C^2 \leq \sigma_D^2$ and $\sigma_C^2 \leq \sigma_I^2$. Moreover, it is also easy to quantify the difference. In particular, since $\sigma_I^2 = \lambda^2 C_{22}$ and $\sigma_D^2 = \lambda^2(q^2 C_{11} - 2qC_{12} + C_{22})$,

$$\sigma_I^2 - \sigma_C^2 = \lambda^2 C_{12}^2/C_{11}$$

and

$$\sigma_D^2 - \sigma_C^2 = \lambda^2(qC_{11} - C_{12})^2/C_{11}. \tag{41}$$

Example 1d

For the M/M/1 queue, the optimal weight and variance of the linear control estimator in (39) are

$$\alpha^* = -\frac{\lambda C_{12}}{C_{11}} = \frac{\rho^3}{(1-\rho)^2}$$

and

$$\sigma_C^2(\alpha^*) = \lambda^2\left(C_{22} - \frac{C_{12}^2}{C_{11}}\right)$$

$$= \frac{\rho^3[2 + 4\rho - 4\rho^2 - \rho^3]}{(1-\rho)^4}.$$

The variance reduction from using the optimal linear control estimator of q instead of the indirect estimator is

$$\sigma_I^2 - \sigma_C^2 = \frac{\lambda^2 C_{12}^2}{C_{11}} = \frac{\rho^4}{(1-\rho)^4}$$

and the relative saving is $(\sigma_I^2 - \sigma_C^2)/\sigma_I^2 = \rho/(2 + 5\rho - 4\rho^2 + \rho^3)$. Note that the relative savings approaches 25% as $\rho \to 1$ for the optimal linear control estimator, as opposed to 0% for the indirect estimator. From the form of $\sigma_I^2 - \sigma_C^2$, we would expect that the value of using the optimal linear control estimator instead of the indirect estimator might tend to be a decreasing function of the arrival process variability, as measured by the asymptotic variability parameter $c_A^2 = \lambda^2 C_{11}$.

Of course, many estimators like (39) and (40) have been investigated previously as candidate linear control estimators for queues; e.g., Lavenberg, Moeller and Welch. Our contribution is to relate direct and indirect estimators to linear control estimators.

It should be intuitively clear that we have found the best possible estimators of the queueing parameters q and w in the framework of Section 2 when λ is known, using the criterion of asymptotic efficiency. However, such a strong statement is hard to make precise and, in fact, not nearly true without additional qualifications. We give one concrete expression of this idea. For this purpose, suppose that we observe data over the interval $[0, D_n^{\uparrow}]$.

Theorem 10. *Under* (19), *no estimator* $f(\mathbf{X}_n)$ *is asymptotically more efficient for estimating q and w than the*

two estimators in (40) *when* λ *is known, and the natural estimators* $\mathbf{q}_n^N$ *and* $\mathbf{w}_n^N$ *in* (17) *and* (20) *when* λ *is unknown, provided that* $f: R^{14} \to R^1$ *has continuous partial derivatives in all coordinates in the neighborhood of* $\mathbf{x}$, $\mathbf{X}_n$ *is*

$$\mathbf{X}_n = \Bigg(n^{-1}A_n, \quad n^{-1}D_n, \quad n^{-1}D_n^{\uparrow}, \quad n^{-1}\sum_{i=1}^{n} W_i,$$

$$n^{-1}\sum_{i=1}^{O(D_n^{\uparrow})} W_i, \quad n^{-1}\int_0^{A_n} Q(s)\,ds, \quad n^{-1}\int_0^{D_n} Q(s)\,ds,$$

$$n^{-1}\int_0^{D_n^{\uparrow}} Q(s)\,ds, \quad n^{-1}N(A_n), \quad n^{-1}N(D_n),$$

$$n^{-1}N(D_n^{\uparrow}), \quad n^{-1}O(A_n), \quad n^{-1}O(D_n), \quad n^{-1}O(D_n^{\uparrow}) \Bigg)$$

and

$$\mathbf{x} = (\lambda^{-1}, \lambda^{-1}, \lambda^{-1}, w, w, w, w, w, 1, 1, 1, 1, 1, 1).$$

Proof. Under (19), we can establish a joint (FCLT) in $D[0, 1]^{14}$ for $\mathbf{X}_n$ with appropriate normalization. This FCLT follows from the joint convergence in Theorem 4 of Glynn and Whitt (1986) plus a Taylor series expansion as in (30). For example, since $n^{-1}D_n = n^{-1}A_n + o_p(n^{-1/2})$, nothing is gained by incorporating $n^{-1}D_n$ as well as $n^{-1}A_n$ into the estimator.

The practical implication of Theorem 10 is that it is not possible to improve the asymptotic efficiency with estimators of q based on other quantities from Section 2 such as the departure epochs instead of or in addition to the arrival epochs; i.e., the estimator

$$\tilde{\mathbf{q}}_n(\mathbf{d}_n) = \lambda n^{-1}\sum_{i=1}^{n} W_i + \mathbf{d}_n(n^{-1}D_n - \lambda^{-1})$$

where $\mathbf{d}_n \Rightarrow \delta^*$ with δ^* the optimal linear control has the same asymptotic efficiency as $\mathbf{q}_n^C(\alpha^*)$ in (39). Detailed descriptions of the asymptotic efficiency of many related estimators also follow from Theorem 4 of Glynn and Whitt (1986).

The optimality in Theorem 10 obviously requires some qualifications, such as are contained in the conditions there. Much depends on the information. *The conclusion is not nearly true without qualifications.* For example, in the setting of Section 2 suppose that we know in advance that $W_1 = w$, which still allows the CLT (10) with a nondegenerate limit. Then W_1 is clearly the best estimator for w using data from $[0, D_n^{\uparrow}]$; the asymptotic efficiency parameter is zero.

The optimality in Theorem 10 also breaks down when there is additional model structure. For example, consider the M/M/1 queue with the FCFS discipline in the setting of Section 2 with known arrival rate λ and service rate μ. Let the traffic intensity be $\rho = \lambda/\mu < 1$. Let q represent the expected equilibrium number of customers waiting, not counting the customer in service, if any. Clearly, the exact value $q = \lambda^2/(\mu^2 - \lambda\mu)$ is the best estimator for q; we don't need any data.

For more general models, when both the service rate and the arrival rate are known, it is natural to use linear controls such as (39) and (40) involving both parameters λ^{-1} and μ^{-1}. Even if μ is unknown and λ is known, paralleling the information above, it is intuitively obvious that for the M/M/1 model it is often better to use an estimator for μ, say $\hat{\mu}$ to obtain $\hat{q} = \lambda^2(\hat{\mu}^2 - \lambda\hat{\mu})$ as an estimator for q than an estimator based on $\mathbf{w}_n^N$. (See Schruben and Kulkarni 1982 for some complications, which do not affect our CLT analysis.)

Example 1e

In particular, suppose that we observe the M/M/1 queue over the interval $[0, D_n^{\uparrow}]$ as in Sections 2 and 4, where $D_n^{\uparrow} = D_n$ is the epoch the nth customer completes service. From the observations of $\{(A_i, D_i): 1 \leq i \leq n\}$, we can determine the service times V_n as well as the waiting times $W_n = D_n - A_n$ from the basic data, because we have a single-server queue with the FCFS queue discipline. (We do not need extra data!) We can then form the natural estimator for μ^{-1}, namely, $\hat{\mu}^{-1} = n^{-1}\sum_{k=1}^{n} V_k$. (Doing better with $\hat{\mu}^{-1}$ does not contradict Theorem 10 because $\hat{\mu}^{-1}$ does not appear in $\mathbf{X}_n$.) We can use the general framework in Section 1 and the results in this section to make a detailed analysis of the asymptotic efficiency. Here $x = \lambda^{-1}$, $y = \mu^{-1}$ and $f(x, y) = y^2/(x^2 - xy)$ and the proposed alternative estimator with λ known, $\hat{q}$ above, is the indirect estimator $f(x, \mathbf{y}_n^N)$. Theorem 9b then describes the asymptotic efficiency. The partial derivatives are

$$\frac{\partial f(x, y)}{\partial y} = \frac{2x^2y - xy^2}{(x^2 - xy)^2} = \frac{\lambda^2(2\mu - \lambda)}{(\mu - \lambda)^2} = \frac{\mu\rho^2(2 - \rho)}{(1 - \rho)^2},$$

so that from (34) the asymptotic efficiency parameter associated with this estimator $\hat{q}$ is

$$\sigma^2(\hat{q}) = \frac{\lambda^4(2\mu - \lambda)^2}{\mu^2(\lambda - \mu)^4} = \frac{\rho^4(2 - \rho)^2}{(1 - \rho)^4}. \tag{42}$$

Note that the asymptotic efficiency parameter in (42) is less than the optimal linear control asymptotic

efficiency parameter in Example 1d for all ρ. Of course, in general we can do even better than $\hat{q}$ by using a linear control estimator of the form $\hat{q} + a(\hat{\lambda}^{-1} - \lambda^{-1})$ where $\hat{\lambda}^{-1} = n^{-1}A_n$.

To summarize, this section presents a strong case for using the optimal linear control estimators when some of the parameters are known. In considerable generality the estimator (40) is most efficient in the $L = \lambda W$ framework of Section 2 when λ is known, but it does not take into account extra information that may be available when the model has additional structure, as illustrated by the M/M/1 example.

The optimal linear control estimator has the drawback that it requires constructing the estimator $\mathbf{a}_n$ for α^*, but this is actually not difficult. Using the estimator $\mathbf{a}_n$ for α^* also introduces bias, but the analysis in Section 10 reveals that the bias contribution due to estimating α^* is typically asymptotically negligible compared to the size of the confidence intervals, being of order $O(n^{-1})$ just like the other bias terms, compared to $O(n^{-1/2})$ for the confidence intervals. However, since $\mathbf{a}_n$ is an estimator for a function of covariance matrix elements, for small n it is likely to have a large variance; $\mathbf{a}_n$ seems to be a term that requires a relatively large n to be in a large sample context. However, for simulation experiments with ample data the optimal linear control estimators seem desirable.

We conclude this section by giving a specific estimator $\mathbf{a}_n$ for α^* in (37) to use in the linear control estimator $\mathbf{z}_n^C(\mathbf{a}_n)$ in the corollary to Theorem 9 for the special case of regenerative structure. (See Iglehart and Lewis, and Lavenberg, Moeller and Sauer, and the other control-variable papers for related results.) In addition to the general framework (1)–(7), suppose that the basic sequence $\{(X_n, Y_n): n \geq 1\}$ in R^{k+1} is regenerative (possibly delayed) with regeneration times $\{T_n: n \geq 1\}$. Let

$$\hat{X}_i = \sum_{j=T_i}^{T_{i+1}-1} X_j, \quad \hat{Y}_i = \sum_{j=T_i}^{T_{i+1}-1} Y_j, \quad \Delta_i = T_{i+1} - T_i$$

and $L(n) = \max\{i \geq 0: T_i \leq n\}$ with $T_0 = 0$. Let the desired estimators be

$$C_{11}(n) = n^{-1} \sum_{i=1}^{L(n)} (\hat{X}_i - \mathbf{x}_n^N \Delta_i)^t (\hat{X}_i - \mathbf{x}_n^N \Delta_i),$$

$$C_{12}(n) = n^{-1} \sum_{i=1}^{L(n)} (\hat{X}_i - \mathbf{x}_n^N \Delta_i)^t (\hat{Y}_i - \mathbf{y}_n^N \Delta_i),$$

$$C_{21}(n) = n^{-1} \sum_{i=1}^{L(n)} (\hat{Y}_i - \mathbf{y}_n^N \Delta_i)^t (\hat{X}_i - \mathbf{x}_n^N \Delta_i),$$

$$C_{22}(n) = n^{-1} \sum_{i=1}^{L(n)} (\hat{Y}_i - \mathbf{y}_n^N \Delta_i)^t (\hat{Y}_i - \mathbf{y}_n^N \Delta_i),$$

$$\mathbf{a}_n = -\nabla_y f(\mathbf{x}_n^N, \mathbf{y}_n^N)^t C_{21}(n) C_{11}^{-1}(n) \tag{43}$$

assuming that $C_{11}(n)$ has an inverse and using $\mathbf{x}_n^N$ and $\mathbf{y}_n^N$ in (3). The following proposition is easily proved using a standard regenerative argument.

Proposition 4. (a) *If*

$$E(\Delta_i^2) < \infty,$$

$$E\left[\left(\sum_{i=T_1}^{T_2-1} |X_i|\right)^t \left(\sum_{i=T_1}^{T_2-1} |X_i|\right)\right] < \infty$$

and

$$E\left[\left(\sum_{i=T_1}^{T_2-1} |Y_i|\right)^t \left(\sum_{i=T_1}^{T_2-1} |X_i|\right)\right] < \infty,$$

where $|x| = (|x_1|, \ldots, |x_k|)$ *for* $x \in R^k$, *then* $C_{ij}(n) \to C_{ij}$ *w.p.*1 *as* $n \to \infty$ *for each* i, j.

(b) *If, in addition,* C_{11} *has an inverse, then* $C_{11}(n)$ *has an inverse for all sufficiently large* n *and* $\mathbf{a}_n \to \alpha^*$ *in* (37) *w.p.*1 *as* $n \to \infty$.

9. Asymptotic Bias

So far, we have only considered asymptotic efficiency; now we consider asymptotic bias. Our goal is to show that bias is typically asymptotically negligible compared to efficiency as the sample size increases, thus justifying paying more attention to asymptotic efficiency.

We conduct our analysis in the general framework of Section 1, but we simplify the notation by deleting the variable y. The parameter to be estimated can thus be represented as $z = f(x)$. As before, all vectors are taken to be row vectors; $x \in R^k$ is $1 \times k$. We consider only the direct estimator $\mathbf{z}_n^D = f(\mathbf{x}_n^N)$. (Of course, the results generalize.) The bias is $E\mathbf{z}_n^D - z$. This section is devoted, first, to showing that in considerable generality $E\mathbf{z}_n^D = z + n^{-1}\beta_D + o(n^{-1})$ and, second, to describing the asymptotic-bias parameter β_D that appears in this expansion. By Theorem 9a, the length of the confidence intervals are of the order $n^{-1/2}$, so this result will support our goal.

We make four new assumptions in addition to the

ones in Section 1:

(i) $n(E\mathbf{x}_n^N - x) \to \gamma_x$ as $n \to \infty$.
(ii) $\{n^2[(\mathbf{x}_n^N - x)(\mathbf{x}_n^N - x)^t]^2 : n \geq 1\}$ is uniformly integrable.
(iii) The function f has continuous second partial derivatives in all coordinates in a neighborhood of x.
(iv) $\{f(\mathbf{x}_n^N) : n \geq 1\}$ is a uniformly bounded sequence of random variables. (44)

Condition i is critical: It says that the initial bias for the natural estimator $\mathbf{x}_n^N$ is of the order n^{-1}. It is important that condition i is actually satisfied in most cases of practical interest. Typically, $X_n \Rightarrow X$ and $EX_n \to EX$ as $n \to \infty$ for the process $\{X_n\}$ in (1). Moreover, there often is *geometric ergodicity* for the expected values, i.e., $\| EX_n - x \| = O(\rho^n)$ for $x = EX$ and $|\rho| < 1$, so that

$$\sum_{n=1}^{\infty} |(EX_n - x)_j| < \infty \quad 1 \leq j \leq k. \tag{45}$$

The convergence (45) is common for ergodic Markov chains; see Lemma 7.2, p. 224 of Doob (1953) and the discussion about the fundamental matrix on pp. 75, 101 of Kemeny and Snell (1960). The fundamental matrix for ergodic Markov chains is essentially a representation of the sum in (45). Note that

$$\begin{aligned} n(E\mathbf{x}_n^N - x) &= \sum_{i=1}^{n} (EX_i - x) \\ &= \sum_{i=1}^{\infty} (EX_i - x) - \sum_{i=n+1}^{\infty} (EX_i - x) \\ &= \gamma_x + o(1). \end{aligned}$$

Thus, if (45) holds for the processes $\{X_n : n \geq 1\}$ in (1), then

$$n(E\mathbf{x}_n^N - x) \to \gamma_x = \sum_{n=1}^{\infty} (EX_n - x) \quad \text{as } n \to \infty. \tag{46}$$

Of course, (45) only provides an easy sufficient condition for condition i in (44); Condition i holds more generally. The remaining conditions in (44) are technical regularity conditions. Condition ii is the standard regularity condition to get convergence of moments from weak convergence; the weak convergence follows from (7) and the continuous mapping theorem. Condition iii typically holds, so it is not a serious restriction. Condition iv is a restriction as stated, but it can be relaxed. (Finding better statements and proofs appears to be a worthwhile direction for research.) Condition iv is a convenient sufficient condition to facilitate a relatively easy proof of our main result about asymptotic bias. Note that condition iv holds if the function f is continuous on R^k and the basic random variables X_n in (1) all have values in some compact subset K of R^k. Obviously, the averages $\mathbf{x}_n^N$ are then contained in K too for all n; and a continuous function on a compact subset is bounded. Conditions ii and iv address technical problems at infinity.

To state our result, let $H(x)$ be the $k \times k$ Hessian matrix of the second partial derivatives of f and let

$$A \circ B = \sum_{i=1}^{k} \sum_{i=j}^{k} A_{ij} B_{ij} \tag{47}$$

for two $(k \times k)$ matrices A and B.

Theorem 11. *Under conditions* i–iv *in* (44) *and* (7), $E\mathbf{z}_n^D = z + n^{-1}\beta_D + o(n^{-1})$, *where* $\beta_D = \beta_x + \beta_H$, *with*

$$\beta_x = \gamma_x \nabla f(x),$$

$$\beta_H = 2^{-1} H(x) \circ C,$$

C is the covariance matrix in (7) *and*

$$\nabla f(x) = \left(\frac{\partial f}{\partial x_1}(x), \ldots, \frac{\partial f}{\partial x_k}(x) \right)^t$$

as in Section 8.

We refer to β_x as the bias contribution due to the initial bias of $\mathbf{x}_n^N$. We refer to β_H as the bias contribution due to the nonlinearity of f. Obviously $\beta_H = 0$ if H is a zero matrix, i.e., if f is linear. We prove Theorem 11 in Appendix A.

We close this section by briefly discussing the bias associated with using $\mathbf{a}_n$ to estimate the optimal linear weight α^* in (37), as suggested by the corollary to Theorem 9. Note that

$$\begin{aligned} nE(\mathbf{x}_n^N - x)\mathbf{a}_n^t = nE(\mathbf{x}_n^N - x)\alpha^{*t} \\ + nE[(\mathbf{x}_n^N - x)(\mathbf{a}_n - \alpha^*)^t]. \end{aligned} \tag{48}$$

The first term on the right in (48) is asymptotically $\gamma_x(\alpha^*)^t + o(1)$ under Assumption i in (44). Assume that $n^{1/2}(\mathbf{a}_n - \alpha^*, \mathbf{x}_n^N - x)$ converges weakly to a nondegenerate limit as $n \to \infty$, which we would expect to have a joint normal distribution. Then $n(\mathbf{x}_n^N - x)(\mathbf{a}_n - \alpha^*)^t$ converges weakly too, by the continuous mapping theorem, so that

$$nE[(\mathbf{x}_n^N - x)(\mathbf{a}_n - \alpha^*)^t] = \beta_{\alpha^*} + o(1) \tag{49}$$

under an additional uniform integrability condition similar to ii in (44). The asymptotic bias associated

with estimating α^* by $\mathbf{a}_n$ is thus of the order $O(n^{-1})$, with β_{α^*} being the associated asymptotic bias parameter.

10. Another Application of the General Framework

The general results in Sections 1, 8 and 9 obviously go much beyond the $L = \lambda W$ framework in Section 2. For example, we can generalize all the other theorems in Carson and Law involving other queueing variables. We can compare the direct and indirect estimators for more general models than the GI/G/s queue, and we obtain the best possible estimators for each parameter in the sense of Theorem 10. We illustrate these ideas by briefly discussing one application, namely, to estimating the time-average of the workload (the uncompleted work in service time at time t) in the GI/G/s system. We assume that $\rho < 1$ and $P(U_1 > V_1) > 0$, where U_1 is an interarrival time and V_1 is a service time, so that the empty state is a regeneration epoch with finite expected regeneration interval; see Whitt (1972) and Carson and Law. In fact, we only need to assume that $\rho < 1$ in order to get regenerative structure, but, in general, the regeneration epochs are more complicated; see Charlot, Ghidouche and Hamami (1978) and Sigman (1988a, b).

To treat the workload, we use the extension of $L = \lambda W$ to $H = \lambda G$; see pp. 408–412 of Heyman and Sobel and references there. A CLT version of $H = \lambda G$ is contained in Glynn and Whitt (1989), which allows us to go beyond the specific GI/G/s model. Moreover, the CLT version of $H = \lambda G$ provides a generalization of the special queueing framework in Section 2. However, here we only discuss the application of Section 1.

The basic data sequence here is $\{(A_n, W_n, V_n): n \geq 1\}$ where A_n and W_n are as in Section 2, with W_n interpreted as the waiting time until beginning service, and V_n is the service time of the nth customer, which we assume satisfies $EV_n = \nu_1$ and $E(V_n^2) = \nu_2 < \infty$. (Unlike the single server queueing example in Section 8, here we need to observe $\{V_n\}$ in addition to $\{(A_n, D_n)\}$.) In terms of this basic sequence, we can define the workload in the system at time t, say $Z(t)$. Using the standard $H = \lambda G$ argument, we can show that if $\rho = \lambda\nu_1/s < 1$, then $t^{-1}(Z(t) \to z$ w.p.1 in addition to the usual $L = \lambda W$ limits, and

$$z = f(\lambda, w, \nu_1, \nu_2) = \lambda\nu_1 w + (\lambda\nu_2)/2. \tag{50}$$

A key assumption supporting (50), which is satisfied by the GI/G/s model, is that V_n is independent of W_n. If this assumption is dropped, then $\nu_1 w$ in (50) must be replaced by the limit of $n^{-1}\sum_{k=1}^{n} V_k W_k$; e.g., see Heyman and Sobel.

Under our regenerative assumptions and appropriate moment hypotheses, there is a joint CLT paralleling (10), i.e., (7) holds, so that we can apply Sections 8 and 9 to obtain the desired results. In particular, given data in $[0, D_n^{\uparrow}]$, the direct estimator for z is $\hat{\lambda}_n\hat{\nu}_{1n}\mathbf{w}_n^N + \hat{\lambda}_n\hat{\nu}_{2n}/2$ and the indirect estimator based on known (λ, ν_1, ν_2) is $\lambda\nu_1\mathbf{w}_n^N + \lambda\nu_2/2$, where

$$\mathbf{w}_n^N = n^{-1}\sum_{k=1}^{n} W_k,$$

$$\hat{\lambda}_n = n^{-1}A_n, \quad \hat{\nu}_{1n} = n^{-1}\sum_{k=1}^{n} V_k$$

and

$$\hat{\nu}_{2n} = n^{-1}\sum_{k=1}^{n} V_k^2.$$

(In the setting of Section 1, $\mathbf{y}_n^N = (\hat{\lambda}_n, \hat{\nu}_{1n}, \hat{\nu}_{2n})$.) Moreover, the most asymptotically efficient estimator for z in the sense of Theorem 10 given data in $[0, D_n^{\uparrow}]$ when λ, ν_1 and ν_2 are known is the linear control estimator.

$$\mathbf{z}_n^C(\mathbf{a}_n) = \lambda\nu_1\mathbf{w}_n^N + (\lambda\nu_2)/2 + \left[(n^{-1}A_n - \lambda^{-1}), \left(n^{-1}\sum_{k=1}^{n} V_k - \nu_1\right), \left(n^{-1}\sum_{k=1}^{n} V_k^2 - \nu_2\nu\right)\right]\mathbf{a}_n \tag{51}$$

where $\mathbf{a}_n$ is a consistent estimator for α^* in (37). A specific consistent estimator $\mathbf{a}_n$ is displayed in (43).

Appendix A: Proof of Theorem 11

Let $B_\varepsilon(x) = \{y \in R^k: \|x - y\| \leq \varepsilon\}$, where $\|\cdot\|$ is the Euclidean norm and ε is sufficiently small so that f possesses continuous second partial derivatives in $B_\varepsilon(x)$; invoke condition iii. We apply Taylor's theorem to $f(\mathbf{x}_n^N)$: When $\mathbf{x}_n^N \in B_\varepsilon(x)$, we can write

$$f(\mathbf{x}_n^N) = f(x) + (\mathbf{x}_n^N - x)\nabla f(x) + 2^{-1}(\mathbf{x}_n^N - x)H_n(\mathbf{x}_n^N - x)^t$$

where

$$H_n \equiv (H_n(i,j): 1 \leq i,j \leq k) = \left(\frac{\partial^2 f}{\partial x_i \partial x_j}(\zeta_n): 1 \leq i,j \leq k\right),$$

where ζ_n lies on the line segment joining $\mathbf{x}_n^N$ and x, and so is random, and $\nabla f(y)$ is the column vector of partial derivatives at y. Moreover, since $\mathbf{x}_n^N \Rightarrow x$ by (1) and H is continuous on $B_\varepsilon(x)$, $H_n(x) \Rightarrow H(x)$ by the continuous mapping theorem.

First we divide the expectation $E\mathbf{z}_n^D$ into two parts:

$$\begin{aligned} E\mathbf{z}_n^D &= Ef(\mathbf{x}_n^N) \\ &= E(f(\mathbf{x}_n^N); \| \mathbf{x}_n^N - x \| \leq \varepsilon) \\ &\quad + E(f(\mathbf{x}_n^N); \| \mathbf{x}_n^N - x \| > \varepsilon). \end{aligned} \tag{A-1}$$

Let $\|f\|$ be the bound on $\{f(\mathbf{x}_n^N): n \geq 1\}$ from condition iv. The second term in (A-1) is asymptotically negligible because

$$\begin{aligned} &nf(\mathbf{x}_n^N) I(\| \mathbf{x}_n^N - x \| > \varepsilon) \\ &\leq n \| f \| \, I([(\mathbf{x}_n^N - x)(\mathbf{x}_n^N - x)^t]^2 > \varepsilon^4) \\ &\leq n \| f \| ((\mathbf{x}_n^N - x)(\mathbf{x}_n^N - x)^t)^2 / \varepsilon^4 \end{aligned}$$

where

$$n^2((\mathbf{x}_n^N - x)(\mathbf{x}_n^N - x)^t)^2 \Rightarrow (N(0, C)N(0, C)^t)^2 \tag{A-2}$$

by (7) and the continuous mapping theorem. The expectations converge in (A-2) by the **UI** condition ii. Dividing by n, we obtain $nE((\mathbf{x}_n^N - x)(\mathbf{x}_n^N - x)^t)^2 \to 0$ as $n \to \infty$.

Turning to the first term in (A-1), we apply the Taylor series expansion to get

$$\begin{aligned} &E(f(\mathbf{x}_n^N); \| \mathbf{x}_n^N - x \| \leq \varepsilon) \\ &= E\{f(x) + (\mathbf{x}_n^N - x)\nabla f(x) \\ &\qquad + 2^{-1}(\mathbf{x}_n^N - x) H_n(\mathbf{x}_n^N - x)^t ; \| \mathbf{x}_n^N - x \| \leq \varepsilon\}. \end{aligned}$$

First,

$$\begin{aligned} &n | \, E(f(x) + (\mathbf{x}_n^N - x)\nabla f(x); \| \mathbf{x}_n^N - x \| \leq \varepsilon) \\ &\qquad - f(x) - E(\mathbf{x}_n^N - x)\nabla f(x) | \\ &\leq n | \, E(f(x) + (\mathbf{x}_n^N - x)\nabla f(x); \| \mathbf{x}_n^N - x \| > \varepsilon) | \\ &\leq n | f(x) | P(\| \mathbf{x}_n^N - x \| > \varepsilon) \\ &\quad + nE(\| \mathbf{x}_n^N - x \| \cdot \| \nabla f(x) \| I(\| \mathbf{x}_n^N - x \| > \varepsilon)) \\ &\leq n | f(x) | P(\| \mathbf{x}_n^N - x \| > \varepsilon) \\ &\quad + n \| \nabla f(x) \| E(\| \mathbf{x}_n^N - x \| I(\| x_n^N - x \| > \varepsilon)) \\ &\leq n | f(x) | P(\| \mathbf{x}_n^N - x \| > \varepsilon) \\ &\quad + n \| \nabla f(x) \| E(\| \mathbf{x}_n^N - x \|^4)/\varepsilon^3, \end{aligned}$$

which converges to zero by the argument used above to treat the second term in (A-1). Of course, $nE(\mathbf{x}_n^N - x)\nabla f(x) \to \gamma_x \nabla f(x)$ by condition i.

It remains to show that

$$\begin{aligned} &nE(2^{-1}(\mathbf{x}_n^N - x) H_n(\mathbf{x}_n^N - x)^t ; \| \mathbf{x}_n^N - x \| \leq \varepsilon) \\ &\quad \to 2^{-1}(H(x) \circ C \end{aligned}$$

where $H(x) \circ C$ is defined in (47). Since $\mathbf{x}_n^N \Rightarrow x$ by (1), $H_n \Rightarrow H(x)$ by condition iii and the continuous mapping theorem. By more of the same reasoning,

$$\begin{aligned} &n(\mathbf{x}_n^N - x) H_n(\mathbf{x}_n^N - x)^t I(\| \mathbf{x}_n^N - x \| \leq \varepsilon) \\ &\quad \Rightarrow \Gamma \equiv N(0, C) H(x) N(0, C)^t \end{aligned} \tag{A-3}$$

where $E(\Gamma) = H \circ C$. Hence, it suffices to show that the left side of (A-3) is uniformly integrable. To this end, note that $H(y)$ is bounded on the compact set $B_\varepsilon(x)$, by M. The following is based on two elementary inequalities: $|x| \leq 1 + x^2$ for a scalar x and $aAa^t \leq Mkaa^t$ where a is $1 \times k$, A is $k \times k$ and $M \geq \max\{|A_{ij}| : 1 \leq i, j \leq k\}$. In particular,

$$\begin{aligned} &n(\mathbf{x}_n^N - x) H_n(\mathbf{x}_n^N - x)^t I(\| \mathbf{x}_n^N - x \| \leq \varepsilon) \\ &\leq nM \sum_{i=1}^{k} \sum_{j=1}^{k} | \, \mathbf{x}_{ni}^N - x_i \, | \, | \, \mathbf{x}_{nj}^N - x_j \, | \\ &\leq nM 2^{-1} \sum_{i=1}^{k} \sum_{j=1}^{k} [(\mathbf{x}_{ni}^N - x_i)^2 + (\mathbf{x}_{nj}^N - x_j)^2] \\ &\leq nkM \sum_{i=1}^{k} (\mathbf{x}_{ni}^N - x_i)^2 = nkM(\mathbf{x}_n^N - x)(\mathbf{x}_n^N - x)^t \\ &\leq kM(1 + n^2[(\mathbf{x}_n^N - x)(\mathbf{x}_n^N - x)^t]^2) \end{aligned}$$

which is uniformly integrable by condition ii.

Acknowledgment

Peter Glynn's research was conducted at the University of Wisconsin–Madison with support from the National Science Foundation under grant ECS-8404809 and the U.S. Army under contract DAAG29-80-C-0041. The authors are grateful to the referees for their many helpful suggestions.

References

ABATE, J., AND W. WHITT. 1988. The Correlation Functions of RBM and M/M/1. *Stoc. Models* **4**, 315–359.

ANDERSON, T. W. 1958. *An Introduction to Multivariate Statistical Analysis.* John Wiley & Sons, New York.

BARLOW, R. E., AND F. PROSCHAN. 1975. *Statistical Theory of Reliability and Life Testing.* Holt, Rinehart & Winston, New York.

BILLINGSLEY, P. 1968. *Convergence of Probability Measures*, John Wiley & Sons, New York.

BRATLEY, P., B. L. FOX AND L. E. SCHRAGE. 1987. *A Guide to Simulation*, Ed. 2. Springer-Verlag, New York.

CARSON, J. S. 1978. Variance Reduction Techniques for Simulated Queueing Processes. Report 78-8, Department of Industrial Engineering, University of Wisconsin, Madison.

CARSON, J. S., AND A. M. LAW. 1980. Conservation Equations and Variance Reduction in Queueing Simulations. *Opns. Res.* **28,** 535–546.

CHARLOT, F., M. A. GHIDOUCHE AND M. HAMAMI. 1978. Irreducibility and Recurrence in the Sense of Harris for the Waiting Times in the GI/G/q Queue (in French). *Z. Wahrscheinlichkeit.* **43,** 187–203.

CHENG, R. C. H., AND G. M. FEAST. 1980. Control Variables with Known Mean and Variance. *J. Opnl. Res. Soc.* **31,** 51–56.

CHUNG, K. L. 1966. *Markov Chains and Stationary Transition Probabilities.* Springer-Verlag, New York.

CHUNG, K. L. 1974. *A Course in Probability Theory*, Ed. 2. Academic Press, New York.

COCHRAN, W. G. 1977. *Sampling Techniques.* John Wiley & Sons, New York.

DALEY, D. J., AND D. R. JACOBS, JR. 1969. The Total Waiting Time in a Busy Period of a Stable Single-Server Queue, II. *J. Appl. Prob.* **6,** 565–572.

DOOB, J. L. 1953. *Stochastic Processes.* John Wiley & Sons, New York.

FELLER, W. 1971. *An Introduction to Probability Theory and Its Applications*, Vol. II, Ed. 2. John Wiley & Sons, New York.

FISHMAN, G. S. 1973. *Concepts and Methods in Discrete Event Digital Simulation.* John Wiley & Sons, New York.

FRANKEN, P., D. KÖNIG, U. ARNDT AND V. SCHMIDT 1981. *Queues and Point Processes.* Akademie-Verlag, Berlin.

FREEDMAN, D. 1967. Some Invariance Principles for Functionals of a Markov Chain. *Ann. Math. Statist.* **38,** 1–7.

GLYNN, P. W., AND W. WHITT. 1986. A Central-Limit-Theorem Version of $L = \lambda W$. *Queueing Syst.* **1,** 191–215.

GLYNN, P. W., AND W. WHITT. 1987. Sufficient Conditions for Functional-Limit-Theorem Versions of $L = \lambda W$. *Queueing Syst.* **1,** 279–287.

GLYNN, P. W., AND W. WHITT. 1988. Ordinary CLT and WLLN Versions of $L = \lambda W$. *Math. Oper. Res.* **13,** 674–692.

GLYNN, P. W., AND W. WHITT. 1989. Extensions of the Queueing Relations $L = \lambda W$ and $H = \lambda G$. *Opns. Res.* (to appear).

HALL, P., AND C. C. HEYDE. 1980. *Martingale Limit Theory and its Applications.* Academic Press, New York.

HANSEN, M. H., W. N. HURWITZ AND W. G. MADOW. 1953. *Sample Survey Methods and Theory*, Vols. 1 and 2. John Wiley & Sons, New York.

HARDY, G. H., J. E. LITTLEWOOD AND G. POLYA. 1952. *Inequalities*, Ed. 2. University Press, Cambridge.

HEYMAN, D. P., AND M. J. SOBEL. 1982. *Stochastic Models in Operations Research*, Vol. 1. McGraw-Hill, New York.

IGLEHART, D. L. 1971. Functional Limit Theorems for the Queue GI/G/1 in Light Traffic. *Adv. Appl. Prob.* **3,** 269–281.

IGLEHART, D. L., AND P. A. W. LEWIS. 1979. Regenerative Simulation and Internal Controls. *J. Assoc. Comput. Mach.* **26,** 271–282.

KARLIN, S., AND H. M. TAYLOR. 1975. *A First Course in Stochastic Processes.* Ed. 2. Academic Press, New York.

KEMENY, J. G., AND J. L. SNELL. 1960. *Finite Markov Chains.* Van Nostrand, Princeton, N.J.

KIEFER, J., AND J. WOLFOWITZ. 1955. On the Theory of Queues with Many Servers. *Trans. Am. Math Soc.* **78,** 1–18.

KLEIJNEN, J. P. C. 1974. *Statistical Techniques in Simulation*, Part 1. Marcel Dekker, New York.

LAVENBERG, S. S., T. L. MOELLER AND C. H. SAUER. 1979. Concomitant Control Variables Applied to the Regenerative Simulation and Queueing Systems. *Opns. Res.* **27,** 134–160.

LAVENBERG, S. S., T. L. MOELLER AND P. D. WELCH. 1982. Statistical Results on Control Variables with Application to Queueing Network Simulation. *Opns. Res.* **30,** 182–202.

LAVENBERG, S. S., AND P. D. WELCH. 1981. A Perspective on the Use of Control Variables to Increase the Efficiency of Monte Carlo Simulations. *Mgmt. Sci.* **27,** 322–335.

LAW, A. M. 1974. Efficient Estimates for Simulated Queueing Systems. Operations Research Center Report 74-7, University of California, Berkeley.

LAW, A. M. 1975. Efficient Estimators for Simulated Queueing Systems. *Mgmt. Sci.* **22,** 30–41.

LITTLE, J. D. C. 1961. A Proof for the Queueing Formula: $L = \lambda W$. *Opns. Res.* **9,** 383–387.

NELSON, B. L. 1987. On Control Variate Estimators. *Comput. Opns. Res.* **14,** 219–225.

NEWMAN, C. M., AND A. L. WRIGHT. 1981. An Invariance Principle for Certain Dependent Sequences. *Ann. Prob.* **9,** 671–675.

NOZARI, A., S. F. ARNOLD AND C. D. PEGDEN. 1984. Control Variates for Multipopulation Simulation Experiments. *IIE Trans.* **16,** 159–169.

RUBINSTEIN, R. V., AND R. MARCUS. 1985. Efficiency of Multivariate Control Variates and Monte Carlo Simulation. *Opns. Res.* **33,** 661–677.

SCHRUBEN, L., AND R. KULKARNI. 1982. Some Consequences of Estimating Parameters for the M/M/1 Queue. *Opns. Res. Lett.* **1,** 75–78.

SIGMAN, K. 1988a. Regeneration in Tandem Queues with Multiserver Stations *J. Appl. Prob.* **25,** 391–403.

SIGMAN, K. 1988b. Queues as Harris Recurrent Markov Chains. *Queueing Syst.* **3,** 179–198.

SRIRAM, K., AND W. WHITT. 1986. Characterizing Superposition Arrival Processes in Packet Multiplexers for Voice and Data. *IEEE J. Select Areas Commun.* **SAC-4,** 833–846.

STIDHAM, S., JR. 1974. A Last Word on $L = \lambda W$. *Opns. Res.* **22,** 417–421.

WHITT, W. 1972. Embedded Renewal Processes in the GI/G/s Queue. *J. Appl. Prob.* **9,** 650–658.

WHITT, W. 1976. Bivariate Distributions with Given Marginals. *Ann. Statist.* **4,** 1280–1289.

WHITT, W. 1982. Approximating A Point Process by a Renewal Process: Two Basic Methods. *Opns. Res.* **30,** 125–147.

WILSON, J. R., AND A. A. B. PRITSKER. 1984a. Variance Reduction in Queueing Simulation Using Generalized Concomitant Variables. *J. Statist. Comput. Simul.* **19,** 129–153.

WILSON, J. R., AND A. A. B. PRITSKER. 1984b. Experimental Evaluation of Variance Reduction Techniques for Queueing Simulation Using Generalized Concomitant Variables. *Mgmt. Sci.* **30,** 1459–1472.