VARIANCE REDUCTION IN SIMULATIONS OF LOSS MODELS

by

Rayadurgam Srikant[1] and Ward Whitt[2]

October 20, 1995
Revision: October 14, 1997

Operations Research 47 (1999) 509–523

[1]Coordinated Science Laboratory, University of Illinois, 1308 W. Main Street, Urbana, IL 61801; rsrikant@uiuc.edu
[2]AT&T Labs, Room A117, 180 Park Avenue, Building 103, Florham Park, NJ 07932-0971; wow@research.att.com

Abstract

We propose a new estimator of steady-state blocking probabilities for simulations of stochastic loss models that can be much more efficient than the natural estimator (ratio of losses to arrivals). The proposed estimator is a convex combination of the natural estimator and an indirect estimator based on the average number of customers in service, obtained from Little's law ($L = \lambda W$). It exploits the known offered load (product of the arrival rate and the mean service time). The variance reduction is dramatic when the blocking probability is high and the service times are highly variable. The advantage of the combination estimator in this regime is partly due to the indirect estimator, which itself is much more efficient than the natural estimator in this regime, and partly due to strong correlation (most often negative) between the natural and indirect estimators. In general, when the variances of two component estimators are very different, the variance reduction from the optimal convex combination is about $1 - \rho^2$, where ρ is the correlation between the component estimators. For loss models, the variances of the natural and indirect estimators are very different under both light and heavy loads. The combination estimator is effective for estimating multiple blocking probabilities in loss networks with multiple traffic classes, some of which are in normal loading while others are in light and heavy loading, because the combination estimator does at least as well as either component estimator, and provides improvement as well.

Subject classifications: Simulation, efficiency: variance reduction for estimates of blocking probabilities; Queues, simulation: efficient simulation estimators for loss models; Communications: efficient simulation of loss networks

Area of Review: Simulation

This paper proposes a method for reducing variance in the estimation of blocking probabilities in simulations of stochastic loss models. A *stochastic loss model* has one or more arrival processes, modeled as stochastic processes, and has the property that not all of these arrivals are admitted. We are interested in a long-run-average or steady-state *blocking probability*, i.e., the long-run proportion of arrivals from one arrival process that are not admitted. The mathematical model is quite general; we assume that admitted arrivals each eventually spend some random time in service, possibly after waiting, and then depart. Otherwise, we only assume appropriate long-run averages exist; see (1)–(5) below. In particular, there are no Markov or independence assumptions; very general dependence is allowed among interarrival times and service times.

The allowed model generality means that the model can be a complex loss network or resource-sharing model, perhaps with alternative routing, such as a model of a communication network; see Ross (1995). Simulations of large complex loss networks can be very time consuming, often requiring hours or more. Thus, effective variance reduction methods can be very useful.

We propose an easily implemented estimator for blocking probabilities that can be remarkably efficient compared to the natural estimator (ratio of losses to arrivals). By "efficient" we mean low variance for given run length or, equivalently, short run length for given variance. The new estimator is a convex combination of the natural estimator and an indirect estimator based on the average number of customers in service, obtained from Little's law ($L = \lambda W$).

It turns out that the improvement over the natural estimator provided by the proposed method is especially dramatic when the holding times are highly variable and the blocking probability is relatively high. This is a practically important case for communication networks because, first, multiple services (e.g., voice and computer lines) lead to highly variable holding times and, second, interest in system response to failures leads to considering scenarios with relatively high blocking probabilities. Of course, the response to short-lived failures requires transient analysis, but since serious link failures in telecommunications networks, such as are caused by backhoe accidents, persist for a substantial time compared to call holding times, there is serious interest in the steady-state behavior in the presence of failures. Since continued reliable service is desired, effort is made to provide satisfactory service even in the presence of failures. Hence, simulation experiments are frequently conducted to estimate steady-state blocking probabilities under relatively heavy loads.

The proposed procedure is also effective for complex loss networks with multiple traffic classes, some of which are in normal loading while others are in light and heavy loading. The new combination estimator tends to be close to the appropriate component estimator depending on the loading, and provides improvement as well. The combination estimator would be useful even if it only selected the better component estimator, because their efficiency differs dramatically in light and heavy loading.

There is a substantial literature on variance reduction, as can be seen from Chapters 2 and 8 of Bratley, Fox and Schrage (1987). Fleming, Schaeffer and Simon (1995) also treat a class of loss models and achieve spectacular variance reduction in many cases by combining control variates and importance sampling.

1. Alternative Estimators

We consider a general system to which arrivals come according to some stochastic process $\{A(t) : t \geq 0\}$, i.e., $A(t)$ records the number of arrivals in the interval $[0, t]$. Some of these arrivals are admitted to the system, after which they stay for a random time and then depart, while other arrivals are blocked and lost. Let $\{L(t) : t \geq 0\}$ be the stochastic process representing losses, i.e., $L(t)$ is the number of losses in the interval $[0, t]$. Admitted customers may initially wait before beginning service, but they eventually enter service and then depart. Let $\{S_n : n \geq 1\}$ be the successive service times of the admitted calls. Let $N(t)$ and $W(t)$ represent the number of customers in service and waiting, respectively, at time t.

We make no detailed stochastic modeling assumptions, such as independence or Markov assumptions. We only assume that

$$t^{-1}A(t) \to \lambda \quad \text{as } t \to \infty \ , \tag{1}$$

$$L(t)/A(t) \to B \quad \text{as } t \to \infty \ , \tag{2}$$

$$(S_1 + \ldots + S_n)/n \to \mu^{-1} \quad \text{as } n \to \infty \ , \tag{3}$$

$$\hat{n}(t) \equiv t^{-1} \int_0^t N(u)du \to n \quad \text{as } t \to \infty \tag{4}$$

and

$$\frac{W(t)}{t} \to 0 \quad \text{as } t \to \infty \ , \tag{5}$$

all with probability 1 (w.p.1), where λ, B, μ^{-1} and n are positive finite real numbers. Equations (1) and (2) together imply that $L(t)/t \to \lambda B$ as $t \to \infty$ w.p.1 as well. The limits λ, B, μ^{-1} and n in (1), (2), (3) and (4) are the arrival rate, the (long-run-average or steady-state) blocking probability, the mean service time and the long-run-average or steady-state number of customers in service, respectively. Condition (5) implies that the long-run rate of customers entering service equals the long-run rate of admitted customers, $\lambda(1-B)$. Condition (5) is clearly satisfied by the classical $G/G/s/k$ model with s servers and k (finite) extra waiting spaces, but it is also satisfied for other models. For example, the number of available servers could be random. There need not even be separate identifiable servers for each customer. Alternatively, service might be completed in several stages at separate facilities.

The point is that the framework provided by equations (1)–(5) is very general, so that the proposed estimation procedure is widely applicable. Of course, wide applicability does not imply that the proposed estimation procedure is necessarily effective in reducing variance. However, it is our experience that the method is indeed effective for many parameter settings in many models.

In this setting our goal is to estimate the blocking probability B by simulation. The *natural estimator* is

$$\hat{B}_N(t) = L(t)/A(t) \ . \tag{6}$$

By (2), B is the limit of $\hat{B}_N(t)$ as $t \to \infty$, so that the natural estimator is consistent. However, since the natural estimator is the ratio of two random quantities, it is a *ratio estimator*. Ratio estimators have some complications; e.g., in general they are biased: If the processes $\{L(t) : t \geq 0\}$ and $\{A(t) : t \geq 0\}$ have stationary increments, then $EL(t)/EA(t) = B$ for each t, but in general $E\hat{B}_N(t) \neq B$.

An estimator closely related to the natural estimator, which we call the *simple estimator*, is

$$\hat{B}_S(t) = \frac{L(t)}{\lambda t} \ , \tag{7}$$

where λ is the *arrival rate* in (1) . Assuming that the process $\{A(t) : t \geq 0\}$ has stationary increments, $\lambda = EA(1)$. Assuming that the process $\{L(t) : t \geq 0\}$ has stationary increments, the simple estimator $\hat{B}_S(t)$ is unbiased for each t: $E\hat{B}_S(t) = EL(1)/\lambda = B$. Thus, the simple estimator might seem preferable to the natural estimator, but in Srikant and Whitt (1996) (hereafter referred to as SW) we showed, through examples and theory (Section 7 of that paper), that the simple and natural estimators tend to be nearly identical for large samples (in actual value as well as in distribution).

In this paper we propose an alternative estimator that in some circumstances has significantly lower variance and is nearly as easy to construct. Our starting point is the *indirect estimator*

$$\hat{B}_I(t) = 1 - \frac{\hat{n}(t)}{\alpha} \ , \tag{8}$$

where $\alpha \equiv \lambda/\mu$ is the *offered load* and $\hat{n}(t)$ is as in (4). The indirect estimator $\hat{B}_I(t)$ requires that we know the parameters λ and μ^{-1}, which is usually the case in simulations. (There are exceptions.

For example, we would not know λ if the arrival process of interest itself comes from overflows from another system with unknown blocking probability. This presumes that we are interested in the proportion of these overflows that are subsequently blocked. We would not know μ^{-1} if the service time included some unknown random waiting time.)

The indirect estimator also requires that we record the statistic $\hat{n}(t)$, but that is usually not difficult to do. The indirect estimator is obtained from Little's law ($L = \lambda W$); if λ, B, μ^{-1} and n are the limits in (1)–(4), then the relation $L = \lambda W$ applied to the service facility (but not the waiting room if there is any) yields $\lambda(1 - B)\mu^{-1} = n$ or, equivalently, $B = 1 - (n/\alpha)$, from which we obtain (8); see Whitt (1991, 1992).

Indirect estimation of queueing quantities by Little's law was studied by Law (1975), Carson and Law (1980) and Glynn and Whitt (1989), but they did not focus on loss models. SW studied the performance of the estimators $\hat{B}_I(t)$ and $\hat{B}_N(t)$, and showed that $\hat{B}_I(t)$ tends to be much more (less) efficient than $\hat{B}_N(t)$ in heavy (light) loading. The advantage of $\hat{B}_I(t)$ over $\hat{B}_N(t)$ in heavy loading is much more dramatic than the previous results for indirect estimators for delay models; e.g., the variance reduction might be by a factor of 1000 or more (e.g., see the case $\gamma = +6.0$ in Table 1 of SW).

Our proposed estimator is the *combination estimator*

$$\hat{B}_C(t) = p\hat{B}_N(t) + (1 - p)\hat{B}_I(t) \ , \tag{9}$$

where p is appropriately chosen to reduce variance (see Section 2). The idea behind the combination estimator in (9) is the observation that $\hat{B}_I(t)$ is decreasing in $\hat{n}(t)$, while $\hat{B}_N(t)$ should tend to be increasing in $\hat{n}(t)$, so that $\hat{B}_I(t)$ and $\hat{B}_N(t)$ should be negatively correlated. We prove a supporting covariance inequality for a class of $GI/GI/s/0$ models (having s servers, no extra waiting room and independent sequences of i.i.d. interarrival times and holding times) in Section 7, but the ordering is intuitively reasonable in general.

The general idea that variance can be reduced by combining different estimators as in (9) is well known, e.g., see p. 63 of Bratley, Fox and Schrage (1987). However, it was not apparent that the combination estimator in (9) can provide truly significant improvement for loss models, as is demonstrated by our examples in Section 4. In the best case in our examples of $GI/GI/s/0$ models with $s = 100$, the variance ratio is $Var\hat{B}_N(t)/Var\hat{B}_C(t) \approx 1800$ (see Table 1). Only part of this benefit would be achieved by the indirect estimator alone; in this case $Var\hat{B}_N(t)/Var\hat{B}_I(t) \approx 200$. The variance ratio of 1800 means that the run length for the combined estimator $\hat{B}_C(t)$ could be about 1800 times shorter than the run length for the natural estimator $\hat{B}_N(t)$ in order to produce the same statistical precision. That variance reduction would reduce a 30 minute run to less than 1 second.

We show that the variance reduction provided by the indirect and combination estimators is even greater when we add a finite waiting room. If a waiting room of size 100 is added to the $GI/GI/s/0$ model with $s = 100$, then the variance reduction in the best case jumps from 10^3 to 10^6 or more; see Section 4.4. The advantage of the waiting room should be evident, because then the mean occupancy $\hat{n}(t)$ is even less variable. (Recall that $\hat{n}(t)$ is the average number of customers in service, not the average number of customers in the system.)

However, it turns out that the benefit of the combination estimator is not uniform in the model parameters. The combination estimator tends to provide dramatic improvement under heavy loads, significant improvement under normal loads, and moderate improvement under light loads. We show that the performance of the combination estimator can be explained by the variance ratio $r^2 \equiv Var\hat{B}_I(t)/Var\hat{B}_N(t)$ and the correlation $\rho \equiv Corr(\hat{B}_I(t), \hat{B}_N(t))$. In Section 2 we show that, in general, the variance reduction of a combination estimator is about $1 - \rho^2$ when the variance ratio r^2 is either very large or very small. As shown by SW, the variance ratio r^2 tends to be very large under light loads and very small under heavy loads. Loss model examples show that the correlation ρ tends to be quite strongly negative under all loadings, but especially under heavy loads (e.g., see Table 1).

As shown for indirect estimators such as $\hat{B}_I(t)$ by Glynn and Whitt (1989), a key ingredient in the proposed estimator $\hat{B}_C(t)$ is exploiting the known parameters λ and μ^{-1}. However, there are other ways to take advantage of this knowledge, in particular, through linear control estimators. Thus, we also consider linear control estimators, using estimators of the arrival rate and mean service time as control variables. (Glynn and Whitt (1989) show that from the perspective of asymptotic efficiency it suffices to consider linear control estimators in the class of suitably smooth nonlinear control estimators.) For this purpose, let

$$\hat{\lambda}(t) = t^{-1}A(t) \tag{10}$$

and

$$\hat{\mu}^{-1}(t) = (1/D(t)) \sum_{i=1}^{D(t)} S_i \ , \tag{11}$$

where as before S_i is the service time of the i^{th} customer to complete service and $D(t)$ is the number of departures (of admitted customers after receiving service) in $[0, t]$. Linear control estimators can be considered with respect to each of the estimators $\hat{B}_N(t), \hat{B}_I(t)$ and $\hat{B}_C(t)$. One is

$$\hat{B}_{LN}(t) = \hat{B}_N(t) + a_1(\hat{\lambda}(t) - \lambda) + a_2(\hat{\mu}^{-1}(t) - \mu^{-1}) \ , \tag{12}$$

where a_1 and a_2 are chosen appropriately. The corresponding linear control estimator constructed from $\hat{B}_I(t)$ is denoted $\hat{B}_{LI}(t)$. The *grand combination estimator* is

$$\begin{aligned} \hat{B}_{GC}(t) &= \hat{B}_C(t) + b_1(\hat{\lambda}(t) - \lambda) + b_2(\hat{\mu}^{-1}(t) - \mu^{-1}) \\ &= p\hat{B}_N(t) + (1-p)\hat{B}_I(t) + b_1(\hat{\lambda}(t) - \lambda) + b_2(\hat{\mu}^{-1}(t) - \mu^{-1}) \ , \end{aligned} \tag{13}$$

where the three parameters p, b_1 and b_2 are chosen appropriately.

The grand combination estimator $\hat{B}_{GC}(t)$ in (13) (with the best parameters) clearly should be most efficient overall, and that is our experience. However, *we find that the combination estimator $\hat{B}_C(t)$ in (9) consistently performs nearly as well as the grand combination estimator $\hat{B}_{GC}(t)$* in (13), so that it should suffice to use the more elementary combination estimator.

Our examples show that linear control estimators can also significantly reduce variance. The variance reduction for estimates of blocking probabilities tends to be greater than the variance reduction for standard single-server queues using similar control variates; see Lavenberg, Moeller and Welch (1982). However, the combination estimator $\hat{B}_C(t)$ consistently does at least as well as, and in some cases does significantly better than, the linear control estimators $\hat{B}_{LN}(t)$ and $\hat{B}_{LI}(t)$.

It is well known that the blocking probabilities in the $M/GI/s/0$ model (with Poisson arrival process) are insensitive to the general holding-time distribution beyond its mean; e.g., see p. 271 of Wolff (1989). However, in SW we showed that the variances of the estimators $\hat{B}_N(t)$ and $\hat{B}_I(t)$ do *not* have this insensitivity property. Indeed, for the $M/GI/s/0$ model these variances tend to be proportional to $1 + c_s^2$, where c_s^2 is the squared coefficient of variation (SCV, variance divided by the square of the mean) of the holding-time distribution. In contrast, the variance of the new combination estimator $\hat{B}_C(t)$ tends to be nearly insensitive to the holding-time distribution beyond its mean; see Sections 4.1, 4.2 and 5. This partly explains the effectiveness of the combination estimator.

In our previous paper we developed predictions for the variance of the estimators $\hat{B}_N(t)$ and $\hat{B}_I(t)$ in the $G/G/s/0$ model, to be used before any data have been collected. We have yet to develop such predictions for the new estimators proposed here. We only know that the variance should be less than the minimum of the variances of $\hat{B}_N(t)$ and $\hat{B}_I(t)$ for the $G/G/s/0$ model. Hence, the previous predictions can yield upper bounds for $G/G/s/0$ models.

Our previous paper focused on the computational effort required to achieve a given statistical precision with the basic estimators. We remark that the story for loss models ($G/G/s/0$) is quite different from the story for delay models ($G/G/s/\infty$); see Whitt (1989). In particular, for loss models there is no precipitous rise in required computational effort as the traffic intensity approaches 1. Indeed, for loss systems the case in which the traffic intensity is 1 is called normal loading. Figure 1 of SW shows that the computational effort to achieve a given statistical precision (using a criterion of absolute error) increases with the offered load for the natural estimator. However, Figure 2 of SW shows that the computational effort decreases with the offered load for the indirect estimator. Given that we use the better of the two basic estimators, normal loading (the middle) requires the most computational effort. It is good, then, that the combination estimator provides significant variance reduction there.

The methods here would be broadly applicable to estimate blocking probabilities from real-time measurements of actual loss systems, provided that we could also estimate the offered load during the measurement process. Hence, it is also natural to consider the *modified indirect estimator*

$$\hat{B}_M(t) = 1 - \frac{\hat{n}(t)}{\hat{\alpha}(t)} \ , \tag{14}$$

where

$$\hat{\alpha}(t) = \hat{\lambda}(t)\hat{\mu}^{-1}(t) \ , \tag{15}$$

and the associated *modified combination estimator* $\hat{B}_{MC}(t)$, defined as in (4) with $\hat{B}_M(t)$ in place of $\hat{B}_I(t)$. Unfortunately, however, we found that these modified estimators do not provide significant improvement. The variance ratio $Var\hat{B}_N(t)/Var\hat{B}_{MC}(t)$ in our examples was consistently about 1. Hence we do not display results for these estimators.

It is of course possible that we could obtain good estimates of λ and μ^{-1} from previous measurements. In a network application we might monitor the system and have available estimates of λ^{-1} and μ^{-1}. There might then be a failure event, which would make it desirable to estimate blocking probabilities. Assuming that the parameters λ and μ^{-1} are not altered by the failure event, we can use the previous estimates of λ^{-1} and μ^{-1} in the combination estimator to estimate the blocking probability from measurements after the failure event.

We now investigate the general combination variance reduction approach more carefully.

2. Variance Reduction for Combination Estimators

Part of the benefit of the combination estimator $\hat{B}_C(t)$ in heavy loads comes from the indirect estimator $\hat{B}_I(t)$, which SW have shown to be significantly more efficient than the natural estimator $\hat{B}_N(t)$ in heavy loads. To understand the two different contributions to efficiency in heavy loads, it is useful to represent the variance ratio as the product of two separate variance ratios, i.e.,

$$\frac{Var\hat{B}_C(t)}{Var\hat{B}_N(t)} = \frac{Var\hat{B}_C(t)}{Var\hat{B}_I(t)} \frac{Var\hat{B}_I(t)}{Var\hat{B}_N(t)} . \tag{16}$$

It is interesting to see how the variance ratio $Var\hat{B}_C(t)/Var\hat{B}_I(t)$ is affected by the fact that the variance ratio $Var\hat{B}_I(t)/Var\hat{B}_N(t)$ is quite small. In this section we show that the variance ratio $Var\hat{B}_C(t)/Var\hat{B}_I(t)$ depends on two key factors: the *variance ratio* $Var\hat{B}_I(t)/Var\hat{B}_N(t)$ and the *correlation* $Corr(\hat{B}_I(t), \hat{B}_N(t))$.

To express the problem generically, let p be an arbitrary constant, let X and Y be arbitrary random variables with a common mean and let

$$Z = pX + (1-p)Y . \tag{17}$$

Let $\sigma_X^2 = VarX$, $\sigma_Y^2 = VarY$, $r = \sigma_Y/\sigma_X$ and $\rho = Cov(X,Y)/\sigma_X\sigma_Y$. Clearly, the variance ratio is r^2 and the correlation is ρ. By direct calculation,

$$VarZ \equiv V(p) = \sigma_Y^2 \left(\frac{p^2}{r^2} + (1-p)^2 + 2p(1-p)\frac{\rho}{r} \right) . \tag{18}$$

Differentiating, we find that $V''(p) > 0$ for all p, so that the minimum is found by setting $V'(p) = 0$. The minimum variance of the combination variable Z is attained at

$$p^* = \frac{r(r-\rho)}{1 + r^2 - 2\rho r} \tag{19}$$

and is

$$V(p^*) = \frac{\sigma_Y^2(1-\rho^2)}{1 + r^2 - 2r\rho} . \tag{20}$$

Note that in general we can have $p^* < 0$ and $p^* > 1$ in (19), but if $\rho \le 0$, then necessarily $0 < p^* < 1$.

Assume that $\sigma_Y^2 \le \sigma_X^2$, so that $r \le 1$. Then we want to compare $VarZ$ to σ_Y^2, since it is more efficient (has lower variance) than $VarX$. For this purpose, let the *combination variance reduction factor* as a function of p be

$$R(p) = \frac{V(p)}{\sigma_Y^2} = \frac{p^2}{r^2} + (1-p)^2 + 2p(1-p)\frac{\rho}{r} \tag{21}$$

and let the *optimal combination variance reduction factor* be

$$R(p^*) = \frac{V(p^*)}{\sigma_Y^2} = \frac{1-\rho^2}{1+r^2-2r\rho} \ . \tag{22}$$

We can use (22) to bound below the possible variance reduction:

$$R(p^*) \geq \frac{1-\rho^2}{(1+r)^2} \geq \frac{1-\rho^2}{4} \tag{23}$$

for $r \leq 1$. If $r = 1$, then $R(p^*) = (1+\rho)/2$, which is only significant when ρ is suitably close to its lower limit -1. If ρ is indeed close to -1, then the lower bound can be approximated by

$$\frac{1-\rho^2}{4} \approx \frac{1+\rho}{2} \ ,$$

which agrees with what is achieved when $r = 1$.

We are especially interested in the case of small r. From (22), we see that

$$\lim_{r \to 0} R(p^*) = 1 - \rho^2 \ , \tag{24}$$

which is independent of the sign of ρ. Note that the limit of $R(p^*)$ as $r \to 0$ differs from the lower bound over all r in (20), which is attained at $r = 1$, only by a factor of 4. In the case of small r, the variance reduction in (16) is approximately the product of $1 - \rho^2$ and r^2. The combination estimator helps under heavy loads because ρ is then often quite close to -1.

We can also to see how p^* behaves as $r \to 0$. From (19), we see that

$$\frac{p^*(r)}{r} \to -\rho \quad \text{as } r \to 0 \ ,$$

so that we have $p^* \approx -\rho r$ for small r. More generally, if we let $p/r \to c$ as $r \to 0$, then

$$R(p) \to c^2 + 1 + 2c\rho \ , \tag{25}$$

by (21). We can use (25) to see how errors in p^* affect the variance reduction. An ϵ asymptotic relative error in p^* corresponds to $p/r \to c$ as $r \to 0$ with $c = -\rho(1+\epsilon)$. Then

$$R(p^*(1+\epsilon)) = 1 - \rho^2 + \epsilon^2\rho^2 \ , \tag{26}$$

so that an ϵ asymptotic relative error in p^* yields an absolute loss of variance reduction (increase in R) of $\epsilon^2\rho^2$, which is less than ϵ^2. Hence, for small r, an ϵ relative error in p^* will have negligible impact if ϵ^2 is suitably small compared to $1 - \rho^2$.

If we do not know ρ, but we know r, then we could let $p = r$. From (21),

$$\begin{aligned} R(r) &= 1 + (1-p)^2 + 2(1-p)\rho \\ &= 1 + (1-p)^2 - 2(1-p) + 2(1-p)(1+\rho) \\ &= p^2 + 2(1-p)(1+\rho) \approx r^2 + 1 - \rho^2 \ , \end{aligned} \tag{27}$$

which is not too different from $1 - \rho^2$ when r is sufficiently small. Indeed, if $r^2 \approx 1 - \rho^2$, then the variance reduction in (16) is approximately r^4, i.e., each step then contributes equally and the overall reduction is the one-step reduction squared.

3. Estimation Procedures

There are a variety of ways to implement the estimation procedures presented so far. What is appropriate depends on the specific model. We now describe what we have done for the models considered here (in Section 4). In Section 3.1 we discuss the required simulation run lengths and the initial conditions. In Section 3.2 we discuss how we estimate variances and covariances. In Section 3.3 we discuss how we estimate the optimal combination parameter p^* in (19). Finally, in Section 3.4 we discuss linear control estimators.

3.1. Run Length and Initial Conditions

We try to avoid most serious statistical problems by having relatively long runs. For the $M/M/s/0$ model with $s = 100$ and service rate 1, we let the measurement interval be 10^4. When the arrival rate is $\lambda = 100$, this means that the expected number of arrivals during the run is 10^6. Since the steady-state blocking probability is then about 0.07, the expected number of losses is 7×10^4.

In the $M/M/s/0$ model and more general $GI/M/s/0$ model (with renewal arrival process), losses are regeneration points, so that segments between successive losses are i.i.d. Since B^{-1} is one plus the expected number of arrivals between successive losses, B^{-1} could be estimated in this framework by the sample mean of 7×10^4 i.i.d. random variables. We do not actually use this estimation procedure and we do not restrict attention to $GI/M/s/0$ models, but this analysis shows that the sample size is indeed quite large. We do not discuss the issue of required path length for loss models at length here, because we already did so in SW.

We start each run with an empty system. Since that initial condition introduces bias, we have a warmup period, i.e., we wait a fixed time before collecting any data. (The full run begins after the warmup period.) As shown in Section 11 of SW, the warmup period for loss models often need not be extraordinarily long to make the initial bias negligible. For the $M/M/s/0$ model, we let the warmup period be 50, which corresponds to 50 mean service times. This is more than adequate for the $M/M/s/0$ model (then about 5 is adequate), but is appropriate for the more variable hyperexponential service times that we also consider in some of our examples.

We note that finite-capacity models tend to require shorter warmup periods than infinite-capacity models, because the maximum number of customers that can be in the finite-capacity system is constrained. In an infinite-capacity system a longer time is required to reach levels that are captured by the tail of the steady-state content distribution. As indicated in Section 11 of SW, the selection of a warmup period can be aided by considering the behavior of associated infinite-server models. In the $M/G/\infty$ model with a holding-time cdf G having mean 1, the time-dependent number of busy servers starting empty has a Poisson distribution with mean

$$EN(t) = n(1 - \int_t^\infty G^c(u)du) \ , \tag{28}$$

where n is the steady-state mean; see (21) on p. 740 of Eick, Massey and Whitt (1993). (In (73) of SW, $E\hat{n}(t)$ should be replaced by $EN(t)$ or (73) should be

$$E\hat{n}(t) - n = -\frac{n}{t}\int_0^t H_e^c(u)du \ ,$$

where H_e is the stationary-excess service-time cdf there.) Since the Poisson distribution is fully characterized by its mean, it is reasonable to measure the time to approach steady state in terms of the time for the mean to approach within a proportion ϵ of its steady state mean. From (28),

$$\frac{n - EN(t)}{n} = \epsilon \tag{29}$$

if and only if

$$\int_t^\infty G^c(u)du = \epsilon \ . \tag{30}$$

Equation (30) leads us to choose a warmup period of 5 in the $M/M/s/0$ model. (Then the integral reduces to e^{-t}.)

It is significant that equation (28) remains valid in the much more general $G/GI/s/0$ model, see Remark 2.3 of Massey and Whitt (1993), so that it is reasonable to use (30) for such more general models. However, the Poisson distribution property is lost when the arrival process is not required to be Poisson. Thus the full distribution may not be close to the steady-state distribution when the means are close. Nevertheless, (30) seems like a useful practical criterion.

3.2. Estimating Variances and Covariances

To estimate variances and covariances, we use simple batch means; i.e., we divide the total run (after the warmup period) into k nonoverlapping batches of equal length and construct batch means. We typically use $k = 20$. Since the runs are relatively long, there tends to be negligible correlation between different batches. Since there are about 10^4 regeneration points within each run in the $GI/M/s/0$ model, it is evident that the batches should be very nearly independent in those cases.

It would also be possible to use other procedures, such as overlapping batch means or weighted batch means; see Meketon and Schmeiser (1984) and Bischak, Kelton and Pollock (1993). Our variance reduction technique does not require that we use simple batch means.

Given that we do use simple batch means, we estimate the covariance $Cov(X, Y)$ for arbitrary random variables X and Y by

$$\hat{C}(X,Y) = \frac{1}{k-1} \sum_{i=1}^{k} (X_i - \bar{X})(Y_i - \bar{Y}) \ , \tag{31}$$

where (X_i, Y_i) are the batch means from the i^{th} batch and $(\bar{X}, \bar{Y})$ are the averages of the batch means. The variance estimate $\hat{V}(X)$ is $\hat{C}(X, X)$. For instance, given a measurement interval $[0, T]$ (after warmup), X_i and Y_i might be the estimators $\hat{B}_N(t)$ in (6) and $\hat{B}_I(t)$ in (8) constructed over the subinterval $[(i-1)T/k, iT/k]$.

If we want the variance and covariance estimates themselves to have lower variance, in addition to making longer runs, we need to let the number of batches grow as we increase the run length; see Glynn and Whitt (1991). As described in Glynn and Whitt (1991), the standard deviations of the variance estimators $\hat{V}(\hat{B}_N(t))$ and $\hat{V}(\hat{B}_I(t))$ are about $\sqrt{2/(k-1)}$ times their means. To derive this relation, we assume that the usual asymptotic normality for estimators as the run length grows is valid. If the run is sufficiently long, then for the estimators $\hat{B}_N(t)$ and $\hat{B}_I(t)$ the batch means will be approximately k i.i.d. normal random variables, each with mean B and variance $k\sigma^2/T$, where σ^2 is the asymptotic variance (σ_N^2 or σ_I^2) and T is the total run length. Then the sample variance is approximately distributed as $k\sigma^2\chi_{k-1}^2/T(k-1)$, where χ_{k-1}^2 is a chi-square random variable with $k-1$ degrees of freedom. The random variable χ_{k-1}^2 has mean $k-1$ and variance $2(k-1)$, so that the sample variance has approximate mean $k\sigma^2/T$ and approximate variance $2k^2\sigma^2/T^2(k-1)$. Hence, the standard deviation of the sample variance is indeed approximately $\sqrt{2/(k-1)}$ times its mean. For $k = 20$, the ratio of the standard deviation to the mean is about $1/3$. This analysis shows how much statistical precision we can expect from the variance estimators. Obviously we can reduce the standard deviation of the variance estimator if we increase the number of batches. However, the analysis only remains correct if the batches remain approximately independent.

3.3. Estimating p^*

In the general setting of (17), we estimate p^* by using formula (19) with the estimates for r and ρ, i.e.,

$$\hat{p} = \frac{\hat{r}(\hat{r} - \hat{\rho})}{1 + \hat{r}^2 - 2\hat{\rho}\hat{r}} \ , \tag{32}$$

where

$$\hat{r}^2 = \frac{\hat{V}(Y)}{\hat{V}(X)} \quad \text{and} \quad \hat{\rho} = \frac{\hat{C}(X,Y)}{\sqrt{\hat{V}(X)\hat{V}(Y)}} \ . \tag{33}$$

By the same argument used to establish (19), $\hat{p}$ is the value of p minimizing the sample variance

$$\hat{V}(p) \equiv \frac{1}{k-1}\sum_{i=1}^{k}[pX_i + (1-p)Y_i - (p\bar{X} + (1-p)\bar{Y})]^2 \ . \tag{34}$$

Hence, $\hat{p}$ can also be found by computing $\hat{V}(p)$ and searching for the minimum p.

To avoid bias in the step, we should estimate p^* using a separate run, but in fact we do the estimation of p^* using the same run that we estimate B. This procedure clearly induces some underestimation of the variances. In general, it is important to be aware of this possibility, but in our context we found the effect to be minor. To reach this conclusion, we tested the procedure by performing multiple independent replications. We found that the estimates of p from any of several different runs produced similar variance reduction. Moreover, the fluctuation in variance estimates typically was greater between runs than within one run over the various optimal p values. We will illustrate this phenomenon later.

To further support estimating p^* within the same run that we estimate B, σ_X^2 and σ_Y^2, we show that the procedure tends to be asymptotically correct as the sample size, say t, increases, provided the number of batches increases with t. Now X and Y in (17) should be replaced by stochastic processes $X(t)$ and $Y(t)$ (e.g., they might be sample means). In great generality, $VarX(t) \to 0$ and $VarY(t) \to 0$ as $t \to \infty$, but $tVarX(t) \to \sigma_X^2$, $tVarY(t) \to \sigma_Y^2$ and $tCov(X(t), Y(t)) \to \rho\sigma_X\sigma_Y$ as $t \to \infty$, so that $VarY(t)/VarX(t) \to r$ and $Corr(X(t), Y(t)) \to \rho$ as $t \to \infty$. Under these limits, $\hat{p}(t) \to p^*$ and $R(\hat{p}(t)) \to R(p^*)$ as $t \to \infty$.

In a specific application we have a fixed small r. By the analysis in (25)–(27), we need to ensure that the error in $\hat{\rho}\hat{r}$ is then suitably small compared to r. If r is extraordinarily small, this step could be difficult, but then σ_Y^2 itself should be small.

3.4. Linear Control Variates

The standard theory of linear control variates implies that the optimal value of a_1 in the linear control estimator (12) is

$$a_1^* = -Cov(\hat{B}_N(t), \hat{\lambda}(t))/Var(\hat{\lambda}(t)) \tag{35}$$

and similarly for the others, e.g., see p. 96 of Glynn and Whitt (1989) and references cited there. The variance reduction (ratio of new to old variance) provided by using the optimal linear control is $1 - \gamma^2$, where γ is the correlation between the original estimator and the control. We obtain our linear control estimators by estimating a_1^* in (35) by estimating the quantities in the numerator and denominator. In the $GI/GI/s/0$ model the interarrival times and service times are independent, so that it suffices to treat the two controls separately.

For the grand combination estimator $\hat{B}_{GC}(t)$ in (14), the variance evidently is *not* in general a convex function of the parameters (p, b_1, b_2). Hence, we found the optimal values of b_1 and b_2 for each of a set of p-values and then optimized over p, again all within one run. This was easily done, requiring negligible computation time, for p values from 0 to 1 increasing by 0.01.

4. Simulation Experiments

We will illustrate how the variance reduction procedures perform by considering several examples.

4.1. The $GI/GI/s/0$ Model

We first consider the standard s-server loss model having no extra waiting space and i.i.d. service times that are independent of i.i.d. interarrival times. We first let $s = 100$ and $\mu = 1$. We consider three values of λ: $\lambda = 140$ (heavy loading), $\lambda = 100$ (normal loading) and $\lambda = 80$ (light loading). We do simulation experiments for these three cases using exponential (M) and hyperexponential (H_2, mixture of two exponentials) distributions for the interarrival times and service times. The exponential distribution has squared coefficient of variation (SCV, variance divided by the square of the mean) 1, while the H_2 distribution we consider has SCV 10. We let c_a^2 and c_s^2 denote the SCV of the interarrival times and service times, respectively.

Our H_2 distribution has "balanced means," i.e., it has density

$$f(x) = p\lambda_1 e^{-\lambda_1 x} + (1-p)\lambda_2 e^{-\lambda_2 x}, \ x \geq 0 \ , \tag{36}$$

with $p\lambda_1^{-1} = (1-p)\lambda_2^{-1}$. The other two parameters are determined by the mean m and the SCV c^2. In particular,

$$p = [1 + \sqrt{(c^2-1)/(c^2+1)}]/2 \tag{37}$$

and

$$p\lambda_1^{-1} = (1-p)\lambda_2^{-1} = m/2 \; . \tag{38}$$

The H_2 distribution is a natural highly variable distribution to consider for service times because it represents the mixture of two exponential distributions with different means. Such mixtures naturally arise when the customers being considered actually represent the combination of two or more different classes with different characteristics. Hyperexponential distributions also are natural to consider for arrival processes too, because they are equivalent to on/off arrival processes, i.e., a Markov modulated Poisson process with a two-state environment: There is an exponential holding time in each environment state; in one environment state there are no arrivals, while in the other environment state arrivals occur according to a Poisson process.

For these particular models it is not difficult to calculate the blocking probability analytically. First, for the $M/GI/s/0$ model, the blocking probability can be calculated easily from Erlang's formula. Second, for the $H_2/M/s/0$ model and $H_2/H_2/s/0$ model, the blocking probability can be calculated exactly by using continuous-time Markov chains. For $s = 100$, the number of states needed for the $H_2/H_2/s/0$ model is of order 10^4, which is manageable. However, it is clear that the variance reduction behavior will be similar for other distributions for which it is not possible to compute the blocking probability analytically. We use the analytic results for Poisson arrivals to help validate our results.

In this example, we let each simulation run length be 200,000 time units, which corresponds to an expected number of arrivals equal to 200,000 λ, (2×10^7 when $\lambda = 100$). We use 400 batches and delete an initial period of length 50 to allow the system to approach steady state.

Simulation results are displayed in Table 1. In each case we display the natural estimate $\hat{B}_N(t)$ and its estimated standard deviation $SD\ \hat{B}_N(t)$. We also display the estimated variance ratios $Var\hat{B}_N(t)/Var\hat{B}(t)$ for several alternative estimators $\hat{B}(t)$. In our simulation experiments we actually considered combination and linear control estimators based on $\hat{B}_S(t)$ as well as $\hat{B}_N(t)$, but as in our previous paper we found that $\hat{B}_S(t)$ and $\hat{B}_N(t)$ tend to be interchangeable, so we only report results for $\hat{B}_N(t)$.

As in SW, we find that the performance of the estimators in $GI/GI/s/0$ model depends on the loading. Roughly speaking, the loading can be regarded as light, normal or heavy when $\alpha < s - 2\sqrt{\alpha}$,

$s - 2\sqrt{\alpha} \leq \alpha \leq s + 2\sqrt{\alpha}$, or $\alpha > s + 2\sqrt{\alpha}$. A starting point is the result from our previous paper that $\hat{B}_I(t)$ is much more efficient than $\hat{B}_N(t)$ in heavy loading, much less efficient in light loading, and about equally efficient in normal loading.

Here are the conclusions we draw from Table 1: First, the efficiency of the grand combination estimator $\hat{B}_{GC}(t)$ and the combination estimator $\hat{B}_C(t)$ in (9) are essentially the same. Thus, we conclude that the combination estimator already includes the benefits from using controls λ and μ^{-1}. In every case, the combination estimator is at least as efficient as all the other estimators. For each other estimator, there is some case in which the combination estimator is substantially better.

As indicated earlier, the variance reduction is dramatic in heavy loading. This is due in part to the advantage of the indirect estimator, but the combination feature also contributes significantly. The variance reduction provided by the combination feature is also substantial in normal loading. In normal loading the combination improves the indirect estimator more than the indirect estimator improves the natural estimator (but much of the gain would be captured by the indirect estimator plus a linear control). The variance reduction tends to increase as the service time gets more variable. The effect of arrival process variability is less clear.

The linear control estimators $\hat{B}_{LN}(t)$ and $\hat{B}_{LI}(t)$ consistently offer improvement over the basic estimators $\hat{B}_N(t)$ and $B_I(t)$, respectively. In heavy loading, $\hat{B}_{LI}(t)$ is nearly as good as $\hat{B}_C(t)$, while in light loading $\hat{B}_{LN}(t)$ is nearly as good as $\hat{B}_C(t)$. In normal loading $\hat{B}_C(t)$ seems to be slightly better than $\hat{B}_{LN}(t)$ and $\hat{B}_{LI}(t)$, with $\hat{B}_{LI}(t)$ being slightly better than $\hat{B}_{LN}(t)$. A key point is that everything is not captured by the linear controls: The differences between the natural and indirect estimators are not removed by simply using linear controls.

In Table 2 we give variance ratios for the $M/M/s/0$ model with $\mu = 1$ as a function of λ and s. The intent here is to show the impact of system size as well as loading. With one exception (normal loading $\lambda = 100$ to 1000), large s means larger variance ratios, but the loading is clearly a more important factor. If we hold the blocking probability fixed, then size becomes a clearer factor; then larger size consistently yields larger variance ratios.

In order to validate our results, we performed independent replications. Table 3 displays the sample means and sample standard deviations of key quantities for four cases in Table 1 based on 20 independent replications or runs each of length $t = 10^4$ using 20 batches. (Thus the total simulation time and the length of each batch is the same.) In each case, the sample mean is the average of the 20 numbers obtained from the 20 runs, while the sample standard deviation is the estimated standard deviation of the quantity from a single run (not the estimated standard deviation of the sample mean, which would be smaller). Thus, the standard deviation estimates show the variability of the estimates from each run.

First, in these cases the exact blocking probabilities can be computed from Erlang's blocking formula. The exact blocking probabilities are $B = 0.07570$ for $\lambda = 100$ and $B = 0.30124$ for $\lambda = 140$. From Table 3 we see that there is no discernible bias in the estimators $\hat{B}_N(t)$ and $\hat{B}_C(t)$. The standard deviations of the estimators $\hat{B}_N(t)$ and $\hat{B}_I(t)$ are also consistent with the predictions in SW, which justifies our choice of run length. Note that the standard deviation of the blocking probability is about 1% of the estimated value, whereas the standard deviations of the standard deviation estimates are larger (relatively); e.g., for the natural estimator they are about 15%. Similarly, the standard deviations of the estimates $\hat{p}$, $\hat{r}$, $\hat{\rho}$ and $\hat{R}$ are also larger.

The main conclusions about variance reduction can be validated by comparing the sample means of the estimated standard deviations (of $\hat{B}_N(t)$ and $\hat{B}_C(t)$) to the sample standard deviations of the estimated means. Table 3 shows that these are close. The sample means of the estimated standard deviation of $\hat{B}_C(t)$ are consistently slightly less than the sample standard deviation of the estimated mean of $\hat{B}_C(t)$, revealing the underestimation of variance that occurs due to estimating p^* in the same run. Table 3 shows that the average predicted variance reductions in the four cases were 13, 410, 33 and 1634, respectively. After squaring the ratios of the displayed standard deviations, we see that the corresponding ratios of the sample variances of the means are 10, 367, 23 and 1067, respectively. Thus, the predicted variance reduction from the output of one run is slightly optimistic, but clearly genuine.

In order to gain further insight into the effect of estimating the optimal weight p^* from the same run in which we estimate $\hat{B}_N(t)$ and $\hat{B}_I(t)$, we plot in Figure 1 the variance $V(p)$ as a function of p for 5 different replications of the $M/H_2/s/0$ heavy-loading ($\lambda = 140$) example from Tables 1 and 3. The example shows that the estimate $\hat{p}$ from any one run would yield similar predicted variance reduction in any other run. Figure 1 is consistent with the slight underestimation of variance observed in Table 3.

A major conclusion of our previous paper was that, unlike the blocking probabilities themselves, the statistical precision of the basic estimators $\hat{B}_N(t)$ and $\hat{B}_I(t)$ in the $M/GI/s/0$ model strongly depends on the holding-time distribution beyond its mean. However, we observed a near insensitivity to the holding-time distribution (beyond the mean) in the standard deviations of the estimators $\hat{B}_C(t)$ and $\hat{B}_{GC}(t)$ in the $M/GI/s/0$ model. In Section 5 we show that the insensitivity is asymptotically correct as $\lambda \to \infty$. From statistical analysis of the simulation results, we are able to conclude, with very high probability, that in general full insensitivity does not hold for the standard deviations of the estimators $\hat{B}_C(t)$ and $\hat{B}_{GC}(t)$, but it is a close approximation.

To illustrate, in Table 4 we display the sample means of four estimators based on 10 runs of length 10,000 each for the $M/GI/s/0$ model with two holding-time distributions. The first holding time distribution is Erlang (E_{10}) with $c_s^2 = 0.1$, while the other is H_2 with $c_s^2 = 10.0$. As before, we consider heavy loading, normal loading and light loading; i.e., we consider $s = 100$, $\mu = 1$ and three values of λ: $\lambda = 140$, $\lambda = 100$ and $\lambda = 80$. The estimated standard deviations are quite close for $\hat{B}_C(t)$ and $\hat{B}_{GC}(t)$, but not for the other two estimators.

4.2. Loss Networks

To show that the estimation procedures also apply to more elaborate loss networks, as in Ross (1995), we also considered three-link triangle networks. Direct traffic is offered to each link, but if these requests are blocked, then they can be routed on the other links if there is space. We assume that each request uses one circuit, with alternate routed traffic requiring one circuit on both of the other two links. Alternate routed calls hold the circuits on both links for the duration of the call. Both circuits become free when the call is complete.

We also allow trunk reservation on each link. A trunk reservation parameter tr_i on link i means that alternate routed traffic is only accepted on that link if there are at least tr_i free circuits on that link. There must be sufficient free capacity on both links in order for a candidate alternate routed call to be admitted.

We consider examples with independent Poisson call arrival processes and exponential call holding times. For this continuous-time Markov chain model, we used uniformization to construct an associated discrete-time Markov chain with the same steady-state probabilities; e.g., see Keilson (1979). To simulate the full process, we would have to include i.i.d. exponential times between transitions (real or fictitious), but since we only wanted to estimate steady-state quantities, we directly simulated the discrete-time Markov chain. (This step itself serves to reduce variance; see Fox and Glynn (1986).)

In the specific examples we now discuss, the three links all have capacity 100 and trunk reservation parameter tr, and the holding times all have mean 1. The model is thus specified by the three arrival rates λ_i and the common trunk reservation parameter tr.

We apply the estimation procedures to estimate the blocking probabilities of each class and the overall (total) blocking probability. The results for six cases are displayed in Table 5. These results were obtained from single runs with 10^7 arrivals after a warmup period of 10^5 arrivals. Since the simulation is in discrete time, the integral in (4) is replaced by a sum.

In the first two cases the arrival rate is 140 on each link, with the common trunk reservation parameter being 5 in the first case and 0 in the second. If the trunk reservation parameter is high enough, then the example becomes like three separate links in heavy loading. However, the first example with $tr = 5$ differs noticeably from the $M/M/s/0$ heavy-loading cases in Tables 1–3. The combination estimator yields significant variance reduction when $tr = 5$, but not as great as for only one link.

However, there is a dramatic change when $tr = 0$. Evidently, the alternate routed calls make the occupancy levels for the individual classes much more variable, so that the indirect estimator becomes less efficient. The combination estimator does no worse than the natural estimator, but it only provides significant improvement for the overall blocking probability. This case also shows that the correlation ρ can be positive. (Positivity was confirmed by independent replications.)

The third case in Table 5 is a balanced network with normal loading. In this case, the trunk reservation parameter $tr = 5$ is sufficiently small that the model is very different from three separate links. Nevertheless, the combination estimator reduces variance by factors of about 4 and 13 for the individual classes and the total network. In this case the variance reduction is primarily due to the combination procedure ($R < r^2$).

The remaining three cases in Table 5 are unbalanced networks. For these cases, the advantage of the combination estimator fluctuates widely. Moreover, it is difficult to predict in advance whether the indirect or natural estimator is better. These examples show that the combination estimator can be good even if it just automatically selects the better of these two basic estimators. Of course, it does this and somewhat better still.

4.3. Finite Waiting Rooms

Our final example involves the addition of a finite waiting room. The addition of a finite waiting room clearly has negligible effect in light loading, but it can have a dramatic impact under heavy loading. To illustrate, we first consider the $M/M/s/k$ model with $s = 100$, $k = 100$ and $\lambda = 140$. This is the same heavy-loading example considered in Tables 1–3, except that we have added a waiting room of size 100. The waiting room slightly reduces the blocking probability from 0.3012 to 0.2857, but it has an enormous impact on the variance reduction. Since the number of busy servers remains at 100 much more frequently, the variance of the indirect estimator drops dramatically. In several independent runs of length 10^4, the estimated standard deviation of the indirect estimator was 3×10^{-6} while the estimated standard deviation of the estimated natural estimator was 1.2×10^{-3}. This is a variance reduction of 1.6×10^5. In this example, there was not much for the combination estimator to add. It yielded essentially the same estimated mean and standard deviation, and $\hat{\rho} = -0.054$. The corresponding example with hyperexponential service times having $c_s^2 = 10$ yielded a variance reduction for the indirect and combination estimators of 1.6×10^6. In this case the combination estimator itself provided slight further improvement; the estimated correlation was $\hat{\rho} = -0.192$.

With a finite waiting room, the indirect estimator can be much better than the natural estimator even with only a single server. To illustrate, we consider an $M/M/1/k$ model with $\mu = 1.0$ and $k = 100$. Based on runs of length 10^7, the variance reductions for the indirect and combination estimators were both 7.8×10^4 when $\lambda = 2.0$ and were 17.6 and 25.8, respectively, when $\lambda = 1.1$.

5. Heavy Loading Asymptotics

We know that the indirect estimator becomes much more efficient than the natural estimator in heavy loading. The examples have shown that the combination estimator can contribute even more variance reduction in heavy loading. Since r is small in heavy loading, (24) in Section 2 implies that the additional variance reduction provided by the combination estimator is then approximately $(1 - \rho^2)^{-1}$, where ρ is the correlation between the basic estimators $\hat{B}_N(t)$ and $\hat{B}_I(t)$. A natural question, then, is: What is the correlation ρ?

In this section we identify the limit of ρ as $\lambda \to \infty$. We first show that, in the general $G/G/s/0$ model, the correlation between the indirect estimator and another estimator approaches -1 as $\lambda \to \infty$. This other estimator is the *time congestion estimator*

$$\hat{B}_T(t) = t^{-1} \int_0^t 1_{\{N(u)=s\}} du \ . \tag{39}$$

The time-congestion estimator was considered in SW, where it was found to behave similarly to the natural and simple estimators. Since the time-congestion estimator is similar to the natural estimator, the simple analysis in this case supports our intuition in the actual case of interest (with $\hat{B}_N(t)$).

When the arrival rate becomes very large, the system is nearly full all the time. There tends to be only one free server for a short time after each service completion. Therefore,

$$\hat{B}_I(t) = 1 - \frac{\hat{n}(t)}{\alpha} \approx 1 - \frac{(s-1)}{\alpha} - \frac{B_T(t)}{\alpha} \ . \tag{40}$$

Hence we have established our first result.

Theorem 1. *In the $G/G/s/0$ model,*

$$\lim_{\lambda \to \infty} Corr(\hat{B}_I(t),\ \hat{B}_T(t)) = -1 \ .$$

We now study ρ (for the natural estimator) in the $G/G/s/0$ model with $\mu = 1$. We assume that the arrival process is an ergodic stationary point process independent of the service times, which form a stationary sequence. We let $\lambda \to \infty$ by scaling the interarrival times. We assume that the service times satisfy a functional central limit theorem (FCLT), e.g., see Billingsley (1968) and Whitt (1980). Let $\Rightarrow$ denote convergence in distribution and let $\lfloor x \rfloor$ be the greatest integer less than or equal to x. Then the assumed FCLT is

$$n^{-1/2} \left(\sum_{i=1}^{\lfloor nt \rfloor} S_i - nt \right) \Rightarrow \sqrt{c_s^2} W(t) \quad \text{as } n \to \infty \ , \tag{41}$$

where $\{W(t) : t \geq 0\}$ is a standard (drift 0, variance 1) Brownian motion or Wiener process. This condition is satisfied in the GI case provided that the service-time cdf has finite variance, in which case c_s^2 is the SCV.

As $\lambda \to \infty$, the system alternates between s servers busy and $(s-1)$ servers busy. After each service completion, there is a brief idle period until the next arrival. As $\lambda \to \infty$, this idle period tends to have the stationary-excess distribution of the interarrival-time distribution. To obtain a meaningful statement, we should consider the system as $\lambda \to \infty$ with time rescaled so that the arrival rate is 1. Then, if F_a is the interarrival-time cdf with mean 1, then the idle time cdf approaches

$$F_{ae}(t) = \int_0^t [1 - F_a(u)] du \ , \quad t \geq 0 \ , \tag{42}$$

which has mean $m_2/2$, second moment $m_3/3$ and, thus, SCV

$$c_{ae}^2 = \frac{4m_3}{3m_2^2} - 1 \ , \tag{43}$$

where m_k is the k^{th} moment of F_a, with $m_1 = 1$. This occurs as $\lambda \to \infty$, because there are then many arrivals between each service completion. This makes the epoch of a service completion fall at an arbitrary time in the stationary point process.

We also assume that successive idle times become i.i.d. as $\lambda \to \infty$, which will occur if the arrival process is only weakly dependent. In the following result, we assume the technical regularity condition of uniform integrability; see p. 32 of Billingsley (1968). The proof and some other asymptotic results of interest appear in an appendix (available on line).

Theorem 2. *In the G/G/s/0 model, assuming uniform integrability of $\hat{B}_N^2(t)$ and $\hat{B}_I^2(t)$,*

$$\lim_{t\to\infty} \lim_{\lambda\to\infty} Corr(\hat{B}_N(t), \hat{B}_I(t)) = -\sqrt{\frac{c_s^2}{c_s^2 + c_{ae}^2}} \ . \tag{44}$$

It is interesting to see how the limit in (44) behaves in special cases. For an M arrival process $c_{ae}^2 = c_a^2 = 1$; the minimum value of c_{ae}^2 is 1/3 for a D arrival process. For the $M/G/s/0$ examples in Table 1, $c_s^2 = 1$ and $c_s^2 = 10$, so that $\rho \to -1/\sqrt{2} \approx -0.707$ and $-\sqrt{10/11} \approx -0.953$, respectively. These limiting formulas agree remarkably well with the estimates $\hat{\rho} = -0.710$ and $\hat{\rho} = -0.937$ in the heavy-loading cases of Table 1. Formula (44) seems to provide useful rough approximations even outside the heavy-loading regime, as shown by the normal and light loading cases in Table 1.

It is significant that Theorem 2 is consistent with the approximate insensitivity we observed in Section 4.1 for the combination estimator in the $M/G/s/0$ model. Combining (17) of SW with (44) above, we obtain

$$\lim_{\lambda\to\infty} \frac{Var\hat{B}_C(t; G/G/s/0)}{Var\hat{B}_C(t; M/M/s/0)} \approx \frac{(1-\rho_G^2)}{(1-\rho_M^2)} \frac{(c_a^2 + c_s^2)}{2} \approx \left(\frac{c_{ae}^2}{c_{ae}^2 + c_s^2}\right)(c_a^2 + c_s^2) \ .$$

where μ and λ are fixed. In the case of M arrivals, $c_{ae}^2 = c_a^2 = 1$, so that the ratio becomes 1, showing asymptotic insensitivity in the $M/G/s/0$ model.

6. The Importance of Knowing λ and μ

The estimators $\hat{B}_I(t)$, $\hat{B}_C(t)$ and $\hat{B}_{GC}(t)$ all take advantage of our knowledge of λ and μ. To apply these estimators to real-time system measurements instead of simulations, we would like to achieve similar variance reduction using estimates of λ and μ, (i.e., via the modified indirect estimator $\hat{B}_M(t)$ in (14) and the associated modified combination estimator $\hat{B}_{MC}(t)$). Unfortunately, however, the good performance of the indirect and combination estimators evidently depends on knowing λ and μ. This is essentially the same conclusion reached in Glynn and Whitt (1989) about indirect estimation via $L = \lambda W$.

It is important to note that some attempts to achieve effective variance reduction when we do not know λ and μ are mere illusions. In order to estimate the final variances of our estimators, we consistently work with batch means. Thus the modified indirect estimator $\hat{B}_M(t)$ in (14) is obtained by taking estimates of $\hat{n}_i(t)$ and $\hat{\alpha}_i(t)$ within each batch and then forming the average of the ratios $n^{-1}\sum_{i=1}^{n}(\hat{n}_i(t)/\hat{\alpha}_i(t))$. Instead, we could determine the overall average $\hat{\alpha}(t)$ for the entire run and use that in each batch with the batch means of $\hat{n}_i(t)$, i.e., $n^{-1}\sum_{i=1}^{n}\hat{n}_i(t)/\hat{\alpha}(t)$. This alternative approach yields spectacular improvement in the direct sample estimates of the estimator variance in heavy loading, but the observed gain is not genuine. The actual estimates produced by this new version of the modified estimator $\hat{B}_M(t)$ turn out to be very similar to the estimates from the previous modified estimator. The putative decrease in sample variance occurs because we have ignored the strong positive correlation between batches caused by using the common factor $\hat{\alpha}(t)$ in each batch. The lack of variance reduction is confirmed when we estimate the variance by performing independent replications.

To illustrate, we give an example. Consider the $M/M/s/0$ model with $\lambda = 140$, $s = 100$ and $\mu = 1$, as in Tables 1–3. Since we are in heavy loading, we know that $\hat{B}_I(t)$ will have lower variance than $\hat{B}_N(t)$, and we would like to achieve this gain with $\hat{B}_M(t)$. In a run of length 10,000, we obtain estimates $\hat{B}_N(t) = 0.3020$, $SD(\hat{B}_N(t)) = 0.000993$, while $SD(\hat{B}_I(t)) = 0.000073$. The two modified estimators yielded estimates 0.301992 and 0.302004, and sample standard deviations 0.000994 and 0.000073. So at first glance, it looks as if we have succeeded with the modified estimator using the $\hat{\alpha}(t)$ for the entire run. However, multiple independent replications show that the real standard deviation for both modified estimators is actually about $SD(\hat{B}_N(t))$ – just as is the case for $\hat{B}_M(t)$.

The situation is different for the natural estimator $\hat{B}_N(t)$ in (6). If we know λ, then we are able to use the *simple estimator* $\hat{B}_S(t)$ in (7) instead of the natural estimator. However, we have found that the role of known λ and μ is very different in these cases. On the one hand, in our previous paper we found that the estimators $\hat{B}_S(t)$ and $\hat{B}_N(t)$ are almost identical (both in actual value and in variance), so that they can be used interchangeably with negligible difference. On the other hand, $\hat{B}_I(t)$ and $\hat{B}_M(t)$ turn out to be very different, so that $\hat{B}_M(t)$ fails to capture the advantage of $\hat{B}_I(t)$ in heavy loading. Similarly, $\hat{B}_{MC}(t)$ fails to capture the advantages of $\hat{B}_C(t)$.

There is a basis for understanding why these estimators perform as they do in the theory of indirect estimation in Sections 1 and 8 of Glynn and Whitt (1989). There, generic estimators that do not use known parameters are called *direct estimators*, while the corresponding ones that do are called *indirect estimators*. The relation between the efficiencies of these estimators is characterized in Theorem 9 of Glynn and Whitt (1989). In an appendix we apply this theorem to explain the consequences of estimating λ and μ in these two settings.

7. Correlation Inequalities

In Section 5 we identified the limiting correlation between $\hat{B}_N(t)$ and $\hat{B}_I(t)$ as the load increases. In this section we establish qualitative results for all loadings. We provide theoretical evidence showing that the estimators $\hat{B}_N(t), \hat{n}(t), -\hat{B}_I(t)$ and $\hat{\lambda}(t)$ are indeed all positively correlated in a large class of loss models (for any loading), which is consistent with intuition. (Unfortunately we are unable to treat $\hat{\mu}^{-1}(t)$.) In order to avoid having to treat ratios of random variables, we consider the estimator $\hat{B}_S(t) = L(t)/\lambda t$ in (7) instead of $\hat{B}_N(t)$. As indicated earlier, $\hat{B}_S(t)$ and $\hat{B}_N(t)$ are very similar.

The specific class of models we consider here we denote by $DFR/IFR/s/0$; it is the special case of the general $GI/GI/s/0$ model in which the interarrival-time distribution is DFR (has decreasing failure rate) and the service-time distribution is IFR (has increasing failure rate). If $F(t)$ is the cumulative distribution function with density $f(t)$, then the failure rate is

$$r(t) = f(t)/(1 - F(t)) \ . \tag{45}$$

The DFR (IFR) property means that $r(t)$ is a decreasing (increasing) function; see Barlow and Proschan (1975). The DFR class includes the hyperexponential (H_k, mixture of k exponentials) distribution, while the IFR class includes the Erlang (E_k, convolution of k identical exponentials) distribution. Both include the exponential distribution, so that the $M/M/s/0$ (Erlang) model is covered. However, the examples with H_2 service times in Section 3 are not included.

Here is our main correlation inequality result.

Theorem 3. *In the $DFR/IFR/s/0$ model, the estimators $\hat{B}_S(t), \hat{n}(t)$, $\hat{\lambda}(t)$ and $-\hat{B}_I(t)$ are all positively correlated.*

We prove Theorem 3 by representing the DFR/IFR/s/0 model as a limit of discrete-time models, and by establishing a related result for discrete-time models. Theorem 1 of Whitt (1980) can serve as the connecting continuity theorem. Related continuity results appear in Kalashnikov and Rachev (1990). The proof of Theorem 3 appears in the appendix.

8. Summary

In this paper we have proposed a new estimator for loss models, a combination of the natural and indirect estimators in (6) and (8). In this combination the simple estimator in (7) can be substituted for the natural estimator, yielding very similar performance. The combination is a convex combination as in (9) in which the optimal weight p^* depends on the variances and covariance of the two component estimators, as described in (19). We have estimated p^* using batch means from one run, as indicated in (32). We showed that using the same run causes minor underestimation of variances (see Table 3). This underestimation could be avoided, if deemed important, by estimating p^* in a separate pilot run.

In our previous paper we showed that the indirect estimator is much more (less) efficient than the natural and simple estimators in heavy (light) loading. Here we observed that this same property holds, with even more difference in heavy loading, when there is a finite waiting room.

In Section 2 we analyzed the benefit of a combination estimator in general, showing that the variance reduction factor is about $(1-\rho^2)^{-1}$ when the two variances are very unequal. Examples in Section 4 and theoretical results in Section 5 and the appendix show that ρ tends to be quite strongly negative, especially under heavy loading, so that the combination estimator provides significant variance reduction over the indirect estimator. In Section 5 we proved for the $G/G/s/0$ model that the correlation approaches $-\sqrt{c_s^2/(c_s^2+c_{ae}^2)}$ as the arrival rate increases, where c_s^2 and c_{ae}^2 are given in (41) and (43).

Even in normal loading, the combination estimator can yield variance reduction because the two component estimators tend to be negatively correlated. In Section 7 we established correlation inequalities for a large class of models to provide theoretical support for this conclusion. These analytical results do not nearly apply to all models for which the estimation procedure can be applied, but they serve as useful theoretical reference points. The examples in Section 4 show that the correlation is usually negative. (The balanced heavily loaded network without trunk reservation in Section 4.2 is a counterexample to a more general result.)

Finally, in Section 6 we showed that the variance reduction achieved by the indirect and combination estimators depends upon knowing the parameters λ and μ. Thus the variance reduction technique tends not to be directly applicable to system measurements in which λ and μ need to be estimated. Overall, the paper continues the longstanding tradition in the simulation literature of showing that, with some thought, simulations can be conducted more efficiently and effectively.

References

Barlow, R. E. and F. Proschan. 1975. *Statistical Theory of Reliability and Life Testing in Probability Models*, Holt, Rinehart and Winston, New York.

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

Bischak, D. P., W. D. Kelton and S. M. Pollock. 1993. Weighted batch means for confidence intervals in steady-state simulations. *Management Sci.* 39, 1002–1019.

Bratley, P., B. L. Fox and L. E. Schrage. 1987. *A Guide to Simulation*, second ed., Springer-Verlag, New York.

Carson, J. S. and A. M. Law. 1980. Conservation equations and variance reduction in queueing simulations. *Opns. Res.* 28, 535–546.

Eick, S. G., W. A. Massey and W. Whitt. 1993. the physics of the $M_t/G/\infty$ queue. *Opns. Res.* 41, 731–742.

Fleming, P. J., D. Schaeffer and B. Simon. 1995. Efficient Monte-Carlo simulation of a product-form model for a cellular system with dynamic resource sharing. *ACM Trans. Modeling Comp. Sim.* 5, 3–21.

Fox, B. L. and P. W. Glynn. 1986. Discrete-time conversion for simulating semi-Markov processes. *Opns. Res. Letters* 5, 191–196.

Glynn, P. W. and W. Whitt. 1989. Indirect estimation via $L = \lambda W$. *Opns. Res.* 37, 82–103.

Glynn, P. W. and W. Whitt. 1991. Estimating the asymptotic variance with batch means. *Opns. Res. Letters* 10, 431-435.

Kalashnikov, V. V. and S. T. Rachev. 1990. *Mathematical Methods for Construction of Queueing Models*, Wadsworth and Brooks/Cole, Pacific Grove, CA.

Keilson, J. 1979. *Markov Chain Models – Rarity and Exponentiality*, Springer-Verlag, New York.

Lavenberg, S. S., T. L. Moeller and P. D. Welch. 1982. Statistical results on control variables with application to queueing network simulations. *Opns. Res.* 30, 182–202.

LAW, A. M. 1975. Efficient estimators for simulated queueing systems. *Management Sci.* 22, 30–41.

MASSEY, W. A. AND W. WHITT. 1993. Networks of infinite-server queues with nonstationary Poisson input. *Queueing Systems* 13, 183–250.

MEKETON, M. S. AND B. SCHMEISER. 1984. Overlapping batch means: something for nothing? *Proc. 1984 Winter Simulation Conference*, 227–230.

ROSS, K. W. 1995. *Multiservice Loss Models for Broadband Telecommunication Networks*, Prentice-Hall, Englewood Cliffs, NJ.

SRIKANT, R. AND W. WHITT. 1996. Simulation run lengths to estimate blocking probabilities. *ACM Trans. Modeling Comp. Sim.* 6, 7–52.

WHITT, W. 1980. Continuity of generalized semi-Markov processes. *Math. Opns. Res.* 5, 494–501.

WHITT, W. 1989. Planning queueing simulations. *Management Sci.* 35, 1341–1366.

WHITT, W. 1991. A review of $L = \lambda W$ and extensions. *Queueing Systems* 9, 235–268.

WHITT, W. 1992. Correction note on $L = \lambda W$. *Queueing Systems* 12, 431–432.

WOLFF, R. W. 1989. *Stochastic Modeling and the Theory of Queues*, Prentice-Hall, Englewood Cliffs, NJ.

heavy loading: $\lambda = 140$				
the c_a^2	1	1	10	10
cases c_s^2	1	10	1	10
estimated $B_N(t)$	0.3010	0.3012	0.3468	0.3404
SD $B_N(t)$	0.00018	0.00052	0.00031	0.00053
variance ratios				
LN	147	41	24.1	30.7
I	114	209	7.7	17.8
LI	230	1606	23.3	100.8
C	253	1885	23.3	100.8
GC	253	1885	24.1	104.7
correlation ρ	-0.710	-0.937	-0.681	-0.847
normal loading: $\lambda = 100$				
the c_a^2	1	1	10	10
cases c_s^2	1	10	1	10
estimated $B_N(t)$	0.0744	0.0751	0.1609	0.1411
SD $B_N(t)$	0.00021	0.00043	0.00037	0.00056
variance ratios				
LN	11.5	15.1	6.3	14.7
I	2.5	2.6	1.3	1.9
LI	11.5	21.6	6.5	17.7
C	12.3	28.4	7.1	20.4
GC	12.3	28.4	7.2	21.4
correlation ρ	-0.727	-0.878	-0.682	-0.863
light loading: $\lambda = 80$				
the c_a^2	1	1	10	10
cases c_s^2	1	10	1	10
estimated $B_N(t)$	0.00394	0.00403	0.0587	0.0402
SD $B_N(t)$	0.00046	0.000091	0.00025	0.00037
variance ratios				
LN	1.39	2.2	2.16	3.9
I	0.021	0.015	0.020	0.22
LI	1.09	0.328	1.99	0.26
C	1.39	2.3	2.12	5.7
GC	1.39	2.3	2.18	6.0
correlation ρ	-0.408	-0.695	-0.482	-0.807

Table 1. Simulation estimates for the $GI/GI/s/0$ model with $s = 100$ and $\mu = 1$ using exponential ($SCV = 1$) and hyperexponential ($SCV = 10$) distributions, based on simulation runs for $t = 2 \times 10^5$ (which corresponds to an expected number of arrivals equal to $2\lambda \times 10^5$) using 400 batches.

heavy loading			
the case	$\lambda = 20$ $s = 10$	$\lambda = 140$ $s = 100$	$\lambda = 1200$ $s = 1000$
$\hat{B}_N(t)$	.5375	.3010	.1719
$SD\ \hat{B}_N(t)$	.000424	.00106	.00118
variance ratios			
LN	16.0	9	4
I	19.9	183	4
LI	45.3	412	1539
C	48.3	484	2049
GC	48.3	504	2049
normal loading			
the case	$\lambda = 10$ $s = 10$	$\lambda = 100$ $s = 100$	$\lambda = 1000$ $s = 1000$
$\hat{B}_N(T)$	.2144	.0751	.02338
$SD\ \hat{B}_N(t)$	.000636	.00186	.000984
variance ratios			
LN	5.3	19.9	10.2
I	2.5	3.3	3.0
LI	6.6	23.3	13.6
C	7.6	27.1	17.1
GC	7.6	32.9	18.1
light loading			
the case	$\lambda = 5$ $s = 10$	$\lambda = 80$ $s = 100$	$\lambda = 930$ $s = 1000$
$\hat{B}_N(t)$	.01817	.00394	.001276
$SD\ \hat{B}_N(t)$	.000226	.000221	.000251
variance ratios			
LN	1.13	1.9	1.7
I	0.016	0.035	0.027
LI	0.93	0.28	0.14
C	1.13	1.9	2.0
GC	1.13	2.1	2.0

Table 2. Variance ratios for the $M/M/s/0$ model with $\mu = 1$ as a function of λ and s. The simulation run length is $10^6/s$ with 20 batches in each case (corresponding to an expected number of arrivals equal to $(\lambda/s)10^6$).

	M service		H_2 service	
	$\lambda = 100$	$\lambda = 140$	$\lambda = 100$	$\lambda = 140$
$\hat{B}_N(t)$ mean	0.07588	0.30110	0.07563	0.30088
SD	0.00085	0.00115	0.00189	0.00196
$\hat{B}_C(t)$ mean	0.07572	0.30126	0.07568	0.30125
SD	0.00027	0.000061	0.00039	0.000060
SD $\hat{B}_N(t)$ mean	0.00086	0.00106	0.00199	0.00208
SD	0.00012	0.00017	0.00031	0.00031
SD $\hat{B}_C(t)$ mean	0.00023	0.000057	0.00036	0.000054
SD	0.00007	0.000008	0.000053	0.000006
$\hat{p}$ mean	0.363	0.0624	0.377	0.0645
SD	0.039	0.0090	0.025	0.0049
$\hat{r}$ mean	0.626	0.088	0.628	0.0743
SD	0.095	0.0128	0.063	0.0053
$\hat{\rho}$ mean	-0.697	-0.728	-0.876	-0.917
SD	0.108	0.120	0.053	0.041
$\hat{R}$ mean	0.227	0.404	0.299	0.362
SD	0.078	0.153	0.068	0.082
Var. Red mean	13.2	410.0	33.0	1634.0
SD	5.89	134.0	14.0	904.0
min	5.6	120.0	14.8	521.0
max	20.0	618.0	69.6	4896.0

Table 3. Sample means and standard deviations of estimates for the $M/GI/s/0$ model with $s = 100$ and $\mu = 1$ based on 20 independent replications of runs each with 10^6 arrivals and 20 batches.

loading	estimator	holding-time variability	
		$c_s^2 = 0.1$	$c_s^2 = 10.0$
heavy $\lambda = 140$	N	.000583	.002009
	I	.000054	.000140
	C	.000040	.000048
	GC	.000040	.000047
normal $\lambda = 100$	N	.000691	.001929
	I	.000459	.001225
	C	.000289	.000330
	GC	.000228	.000321
light $\lambda = 80$	N	.000190	.000364
	I	.001030	.003340
	C	.000165	.000230
	GC	.000161	.000223

Table 4. Average standard deviation estimates for four estimators in the $M/GI/s/0$ model for two different holding-time distributions with $s = 100$, $\mu = 1$ and three values of $\lambda : \lambda = 140$, $\lambda = 100$ and $\lambda = 80$, based on 10 independent replications, each of length 10,000 time units.

trk. res.	5	0	5	5	5	10
1	140.0	140.0	100	130.0	200.0	140.0
λ_i 2	140.0	140.0	100	90.0	40.0	80.0
3	140.0	140.0	100	110.0	40.0	120.0
1	0.3019	0.3893	0.0764	0.2308	0.2553	0.2959
$\hat{B}_N(t)$ 2	0.3022	0.3892	0.0758	0.0420	0.00145	0.00475
3	0.3029	0.3892	0.0767	0.1509	0.00146	0.1998
total	0.3023	0.3893	0.0763	0.1527	0.1827	0.1935
1	0.00050	0.00047	0.00042	0.00044	0.00049	0.00038
SD $\hat{B}_N(t)$ 2	0.00043	0.00036	0.00043	0.00030	0.000038	0.00016
3	0.00053	0.00041	0.00044	0.00043	0.000040	0.00062
total	0.00025	0.00028	0.00030	0.00028	0.00036	0.00029
1	0.185	0.861	0.947	0.526	0.237	0.284
r 2	0.195	1.540	0.895	1.606	20.38	5.464
3	0.161	0.924	0.926	0.700	26.26	0.266
total	0.130	0.315	0.543	0.490	0.288	0.727
1	−0.276	0.458	−0.530	−0.205	0.373	−0.505
ρ 2	−0.531	0.588	−0.493	−0.551	0.064	−0.466
3	−0.407	0.316	−0.554	−0.373	−0.221	−0.501
total	−0.644	0.536	−0.695	−0.442	−0.203	−0.035
1	0.813	0.829	0.248	0.641	0.979	0.545
R 2	0.577	0.419	0.282	0.130	0.0024	0.022
3	0.721	0.709	0.241	0.428	0.0013	0.560
total	0.493	0.936	0.252	0.481	0.799	0.632
1	0.075	0.364	0.482	0.258	−0.037	0.164
p^* 2	0.113	0.939	0.463	0.648	1.0007	0.901
3	0.079	0.442	0.475	0.373	0.9903	0.153
total	0.085	0.092	0.077	0.273	0.118	0.351
overall 1	36.0	1.6	4.5	5.6	18.2	22.7
variance 2	45.6	1.0	4.4	3.0	1.0	1.5
reduction 3	53.6	1.6	4.8	4.8	1.1	25.2
factor total	120.0	10.8	13.5	8.7	15.1	3.0

Table 5. Simulation results for six examples of three-link triangle networks with alternate routing. The capacity of each link is 100 and all mean service times are 1. All runs have 10^7 arrivals with 400 batches after a warmup of 10^5 arrivals.

Figure 1. The variance $V(p)$ is a function of p in five independent replications of the $M/H_2/s/0$ example with $\lambda = 140$, $\mu = 1$ and $s = 100$. Each replication was of length 10^4 (about 1.4×10^6 arrivals).

Variance

Reduction

Appendix

1. Summary

In this appendix we present additional supporting material. We start in Section 2 by discussing additional experimental results demonstrating the wide applicability of the variance reduction procedure. Next in Section 3 we present additional heavy-loading asymptotic results, including a proof of Theorem 2. In Section 4 we show how Theorem 9 of Glynn and Whitt (1989) can help explain the importance of knowing the parameters λ and μ (instead of using their estimates). In Section 5 we discuss additional insights that can be gained by considering the special case of deterministic service times. Finally, in Section 6 we prove the correlation inequalities in Theorem 3.

2. Experiments with Dependent Interarrival and Service Times

To show that the estimation procedure also applies to more complex models, we also considered the $G/G/s/0$ model in which the service times and the arrival process are allowed to be dependent. Specifically, we let the service time of each customer be exactly (λ/μ) times the last interarrival time. This gives the service time the correct mean μ^{-1}, but makes these two variables strongly dependent (correlation 1).

No extra work was required to implement the various estimators for these modified models. Moreover, essentially the same variance reduction behavior was observed in these modified models as was observed for the standard models.

Intuitively, having service times positively correlated with interarrival times should reduce congestion, but our simulation results showed that the blocking probabilities in the modified $M/M/s/0$ and $H_2/H_2/s/0$ systems did not differ too much from the blocking probabilities in the standard $M/M/s/0$ and $H_2/H_2/s/0$ systems. For example, for the modified $M/M/s/0$ model with $s = 2$, $\mu = 1$ and $\lambda = 2$, the estimated blocking probability was 0.380 compared to 0.400 in the standard $M/M/s/0$ model; for $s = 100$, $\mu = 1$ and $\lambda = 140$, the estimated blocking probability was 0.3008 compared to 0.3012. The decrease was statistically significant, but not large.

3. Heavy Loading Asymptotics

In this section we prove Theorem 2. To do so, we combine three new asymptotic results. Let $\{C_s(t) : t \geq 0\}$ be the counting process of successive busy periods. As $\lambda \to \infty$, $\{C_s(t) : t \geq 0\}$ will be the superposition of s stationary point processes generated by the service times at each server. By (41) and Section 7 of Whitt (1980) (recall $\mu = 1$),

$$n^{-1/2}(C_s(nt) - snt) \Rightarrow \sqrt{sc_s^2}W(t) \quad \text{as } n \to \infty \ . \tag{46}$$

We first consider the natural estimator.

Theorem 4. *In the G/G/s/0 model with $\mu = 1$,*

$$\lim_{\lambda\to\infty} \lambda(1-\hat{B}_N(t)) = t^{-1}C_s(t) \quad \text{w.p.1.} \tag{47}$$

Proof. As $\lambda \to \infty$, there is one admitted arrival in each busy cycle, and the busy cycles approach the busy periods. Thus,

$$\lambda(1-\hat{B}_N(t)) \approx \frac{\lambda C_s(t)}{A(t)} \to \frac{C_s(t)}{t} \quad \text{as } \lambda \to \infty \quad \text{w.p.1}$$

by the ergodic theorem for the arrival process. □

We now consider the indirect estimator.

Theorem 5. *In the G/G/s/0 model with $\mu = 1$,*

$$\lim_{\lambda\to\infty} \lambda^2(1-\hat{B}_I(t)) - \lambda s = t^{-1}Y(t) \equiv -t^{-1}\sum_{i=1}^{C_s(t)} X_i \quad w.p.1\ , \tag{48}$$

where $\{X_i\}$ is an i.i.d. sequence independent of $\{C_s(t) : t \geq 0\}$ with X_i distributed as F_{ae} in (42).

Proof. First note that

$$\lambda(1-\hat{B}_I(t)) = \hat{n}(t)\ .$$

Then note that for large λ

$$\hat{n}(t) \approx s - t^{-1}\sum_{i=1}^{D_s(t)} Z_i$$

where Z_i is the length of the i^{th} idle period and $D_s(t)$ is the counting process of the busy cycles. Hence, when we rescale by multiplying by λ,

$$\lambda^2(1-\hat{B}_I(t)) - \lambda s \approx -t^{-1}\sum_{i=1}^{D_s(t)} \lambda Z_i$$

As $\lambda \to \infty$, the variables λZ_i become distributed the same as the variables X_i, $\{\lambda Z_i\}$ becomes independent of $D_s(t)$ and $D_s(t) \to C_s(t)$. □

We now establish a joint CLT for $(C_s(t), Y(t))$ where $Y(t)$ is the limit in (48).

Theorem 6. *In the G/G/s/0 model with $\mu = 1$, if $\lambda \to \infty$ and then $t \to \infty$, then*

$$t^{-1/2}(C_s(t) - st\ ,\ Y(t) - st) \Rightarrow (\sqrt{sc_s^2}N_1(0,1)\ ,\ -\sqrt{sc_s^2}N_1(0,1) + \sqrt{sc_{ae}^2}N_2(0,1)) \tag{49}$$

for c_s^2 in (41) and c_{ae}^2 in (43), where $N_1(0,1)$ and $N_2(0,1)$ are independent standard (mean 0, variance 1) normal random variables.

Proof. Our starting point is the joint FCLT

$$n^{-1/2}(\sum_{i=1}^{\lfloor nt \rfloor} X_i - nt\ ,\ C_s(nt) - snt) \Rightarrow (\sqrt{c_{ae}^2}W_1(t), \sqrt{sc_s^2}W_2(t)) \ \text{ as } n \to \infty$$

in the function space $D[0,\infty)$, where $D[0,\infty)$ is the space of right-continuous real-valued functions with left limits everywhere in $(0,\infty)$, endowed with the Skorohod J_1 topology (see Whitt (1980)) and $W_1(t)$ and $W_2(t)$ are independent standard Brownian motions. This initial FCLT holds because of the independence which holds as $\lambda \to \infty$. Next we apply the continuous mapping theorem using a random time change, as in Section 5 of Whitt (1980) and the projection at $t = 1$ to obtain (49); i.e., we consider the process $n^{-1/2}(\sum_{i=1}^{\lfloor nt \rfloor} X_i - nt)$ evaluated at the random time $C_s(nt)/n$ and then add the process $n^{-1/2}(C_s(nt) - snt)$ to obtain

$$n^{-1/2}(\sum_{i=1}^{C_s(nt)} X_i - snt \ , \ C_s(nt) - snt)$$

$$\Rightarrow (\sqrt{c_{ae}^2} W_1(t) + \sqrt{sc_s^2} W_2(t) \ , \ \sqrt{sc_s^2} W_2(t)) \ \text{ as } n \to \infty \ .$$

Since $Y(t) = -\sum_{i=1}^{C_s(t)}$, we have the desired result. $\square$

The limit in (49) leads to the approximations

$$Cov(C_s(t),\ Y(t)) \approx -stc_s^2 \tag{50}$$

and

$$Corr(C_s(t),\ Y(t)) \approx -\sqrt{\frac{c_s^2}{c_s^2 + c_{ae}^2}} \tag{51}$$

assuming that t is suitably large. Moreover, Theorems 4 and 5 lead to the approximation

$$\rho \approx Corr(\hat{B}_N(t), \hat{B}_I(t)) \approx Corr(C_s(t),\ Y(t)) \ . \tag{52}$$

Combining Theorems 4–6 and (50)–(52), and employing uniform integrability arguments (which we omit), see p. 32 of Billingsley (1968), we obtain Theorem 2.

We can also use Theorems 2 and 4–6 to obtain insights about the performance of the linear control estimators in heavy loading. Note that the limit $C_s(t)/t$ in Theorem 4 is essentially the estimator $\hat{\mu}(t)$. Thus, as $\lambda \to \infty$, $Corr(\hat{B}_N(t), \hat{\mu}(t)) \to 1$ and $\hat{B}_{LN}$ using the control μ becomes a significant improvement over $\hat{B}_N$. On the other hand, $Corr(\hat{B}_N(t), \hat{\lambda}(t)) \to 0$, so that λ is not an effective control for large λ. Since $Corr(\hat{B}_N(t), \hat{\mu}(t)) \to 1$ as $\lambda \to \infty$, we see that $\hat{B}_{LI}$ using μ is asymptotically equivalent to $\hat{B}_C(t)$ as $\lambda \to \infty$, i.e., $Corr(\hat{B}_I(t), \hat{\mu}(t))$ has the same limit as in (44), which is consistent with the numerical results in Table 1.

We conclude this section by noting that the uniformization approach for treating Markovian networks in Section 4.3 makes ρ much more strong negatively. With the discrete time framework, the random variables X_i in $Y(t)$ in (48) become replaced with constants. This enables us to conclude the following.

Theorem 7. *In the $M/M/s/0$ model, if the discrete-time process is simulated using uniformization, then*

$$\lim_{t\to\infty}\ \lim_{\lambda\to\infty} Corr(\hat{B}_N(t),\ \hat{B}_I(t)) = -1\ .$$

To illustrate, we used the network simulation program with one link to simulate the $M/M/s/0$ model with $s = 100$, $\mu = 1$ and 2×10^7 arrivals with 2000 batches. The estimates of ρ were -0.690, -0.804, -0.874 and -0.967 for arrival rates 140, 500, 1000 and 5000, respectively. For the $M/M/s/0$ with $\mu = 1$, $s = 2$ and $\lambda = 100$, we obtained $\hat{\rho} \approx 0.9993$.

4. The Importance of Knowing λ and μ

We now show how to apply Theorem 9 of Glynn and Whitt (1989) to explain the importance of knowing λ and μ (instead of using estimates) in order to achieve the variance reduction with the indirect estimator $\hat{B}_I(t)$ and the combination estimator $\hat{B}_C(t)$.

For completeness, we restate Theorem 9 of Glynn and Whitt(1989) here, with minor modification in notation for our setting. The indirect estimation framework involves estimators $(\hat{x}_t, \hat{y}_t)$ of (x, y) such that

$$t^{1/2}(\hat{x}_t - x, \hat{y}_t - y) \Rightarrow N(0, C)\ ,$$

where x and y are vectors, and $N(0, C)$ is a random vector with a multivariate normal distribution having zero means and covariance matrix C. Let x and y be k and l dimensional, respectively. Let C_{11}, C_{12}, C_{21} and C_{22} be the $k \times k$, $k \times l$, $l \times k$ and $l \times l$ submatrices of C associated with x and y. Our goal is to estimate $f(x, y)$ for a smooth real-valued function f, i.e., with gradient

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

and

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

The direct estimator is $\hat{z}_t^D \equiv f(\hat{x}_t, \hat{y}_t)$ and the indirect estimator is $\hat{z}_t^I \equiv f(x, \hat{y}_t)$. Let $z \equiv f(x, y)$. Let $o_p(t^{-1/2})$ refer to a term which converges to 0 in probability after dividing by $t^{-1/2}$. Here is the result:

Theorem 8. (Glynn and Whitt (1989)) *Under the assumptions above,*

(a) $t^{1/2}(\hat{z}_t^D - z) \Rightarrow N(0, \hat{\sigma}_D^2)$ *as* $t \to \infty$,

(b) $t^{1/2}(\hat{z}_t^I - z) \Rightarrow N(0, \hat{\sigma}_I^2)$ *as* $t \to \infty$,

(c) $\hat{z}_t^D = \hat{z}_t^I + (\hat{x}_t - x) \bigtriangledown_x f(x, y) + o_p(t^{-1/2})$ *as* $t \to \infty$,

where

$$\hat{\sigma}_I^2 = \bigtriangledown_y f(x,y)^t C_{22} \bigtriangledown_y f(x,y)$$

and

$$\begin{aligned}\hat{\sigma}_D^2 &= \bigtriangledown_x f(x,y)^t C_{11} \bigtriangledown_x f(x,y) + \bigtriangledown_y f(x,y)^t C_{21} \bigtriangledown_x f(x,y) \\ &\quad + \bigtriangledown_x f(x,y)^t C_{12} \bigtriangledown_y f(x,y) + \bigtriangledown_y f(x,y)^t C_{22} \bigtriangledown_y f(x,y) \ .\end{aligned}$$

From Theorem 8 we see that the asymptotic efficiency (asymptotic variance) of the estimators $\hat{z}_t^D$ and $\hat{z}_t^I$ depends on the gradient $\bigtriangledown f(x,y)$ and the covariance matrix C. We now apply Theorem 8 to our estimators.

First, we relate the natural and simple estimators. The natural estimator $\hat{B}_N(t)$ is the ratio of the two estimators $L(t)/t$ and $\hat{\lambda}(t)$. Therefore, the connecting function is $f(x,y) = y/x$ where $x = \lambda$ and $y = \lambda B$. The associated gradient of f is

$$\bigtriangledown f(x,y) = (-y/x^2, 1/x) = (-B/\lambda, 1/\lambda) \ . \tag{53}$$

Theorem 8 (c) implies that

$$\hat{B}_N(t) = \hat{B}_S(t) - B\lambda^{-1}(\hat{\lambda}(t) - \lambda) + o_p(t^{-1/2}) \ . \tag{54}$$

By assumption, $t^{1/2}(\hat{\lambda}(t) - \lambda) \approx N(0, C_{11})$, which is equal in distribution to $\sqrt{C_{11}}N(0,1)$. When $A(t)$ is a Poisson process, $C_{11} = \lambda$. Using this as an approximation, we have

$$\hat{B}_N(t) - \hat{B}_S(t) \approx \frac{-B}{\sqrt{\lambda t}} N(0,1) \ , \tag{55}$$

which tends to be small compared to B when λt is large. For instance, in the examples in Table 1 we had $\lambda t \approx 10^6$, so that $1/\sqrt{\lambda t} = 10^{-3}$ there. Hence, we anticipate that $\hat{B}_N(t) \approx \hat{B}_S(t)$ when λ is large. We can also reach this conclusion by focusing on the variances, using Theorem 8(a) and (b), but this essentially repeats what was done in Section 7 of SW, so we stop.

Now we consider indirect estimation in the context of the indirect estimator $\hat{B}_I(t)$. Since $B = 1 - n/\alpha$ for $\alpha = \lambda/\mu$, the connecting function is $f(x_1, x_2, y) = 1 - (y/x_1 x_2)$ for $x_1 = \lambda, x_2 = \mu^{-1}$ and $y = n$. The gradient is

$$\bigtriangledown f(x_1, x_2, y) = \left(\frac{y}{x_1^2 x_2}, \frac{y}{x_1 x_2^2}, \frac{-1}{x_1 x_2}\right) = \left(\frac{n}{\lambda^2 \mu^{-1}}, \frac{n}{\lambda \mu^{-2}}, \frac{-1}{\lambda \mu^{-1}}\right) \ . \tag{56}$$

Theorem 8 (a) and (b) imply that

$$Var \hat{B}_I(t) = \frac{1}{\alpha^2} Var(\hat{n}(t)) \tag{57}$$

and

$$Var\hat{B}_M(t) \ - \ Var\hat{B}_I(t) = \frac{n^2}{\lambda^2\alpha^2} Var(\hat{\lambda}(t)) + \frac{n^2}{\alpha^2\mu^{-2}} Var(\hat{\mu}^{-1}(t))$$

$$+ \frac{2n^2}{\alpha^3} Cov(\hat{\lambda}(t), \hat{\mu}^{-1}(t)) - \frac{2n}{\lambda\alpha^2} Cov(\hat{\lambda}(t), \hat{n}(t)) - \frac{2n}{\mu^{-1}\alpha^2} Cov(\hat{\mu}^{-1}(t), \hat{n}(t)) \ . \tag{58}$$

Assuming that the arrival process and holding times are mutually independent, we expect the covariance term $Cov(\hat{\lambda}(t), \hat{\mu}^{-1}(t))$ in (58) to be negligible. Assuming that $\mu = 1$ and $\alpha \approx n$, we see that the prefactors of all terms not involving $\hat{\mu}^{-1}(t)$ are about the same size, namely, $1/\alpha^2$. However, the prefactors of terms involving $\hat{\mu}^{-1}(t)$ are larger, showing the potential for the service times to have a greater influence.

The case of primary interest, yielding the big advantage for $\hat{B}_I(t)$ and $\hat{B}_C(t)$, is heavy loading. In heavy loading we will have $\hat{n}(t) \approx s$ and $Var(\hat{n}(t))$ much reduced compared to $Var(\hat{\lambda}(t))$ and $Var(\hat{\mu}^{-1}(t))$. Similarly, covariance terms involving $\hat{n}(t)$ should be negligible. Thus, we can apply (58) to deduce that $Var\hat{B}_I(t)$ should be much smaller than $Var\hat{B}_M(t)$, as is borne out by experiments.

5. Deterministic Service Times

We can obtain additional insight into the variance reduction by considering the special case of the $G/D/s/0$ model, which has deterministic service times. Without loss of generality, let the service times all be of length 1. In general, we have the basic conservation law

$$N(t) = N(0) + A(t) - D(t) - L(t) \ , \tag{59}$$

where $D(t)$ records the number of departures in $[0, t]$, i.e., the number of admitted arrivals that have completed service. For simplicity, assume that $N(0) = 0$. Then, since we have deterministic service times,

$$D(t) = A(t-1) - L(t-1) \ . \tag{60}$$

Combining (59) and (60), we obtain

$$N(t) = [A(t) - A(t-1)] - [L(t) - L(t-1)] \tag{61}$$

and

$$\hat{n}(t) \equiv \frac{1}{t}\int_0^t N(u)du = \frac{1}{t}\int_{t-1}^t [A(u) - L(u)]du \ . \tag{62}$$

Since $A(t)$ and $L(t)$ increase as t increases, (62) implies that

$$\hat{n}(t) \approx \frac{A(t)}{t} - \frac{L(t)}{t} \ . \tag{63}$$

We first apply (63) to analyze the modified estimator. In our context, assuming (63),

$$\hat{B}_M(t) = 1 - \frac{t\hat{n}(t)}{A(t)} = \frac{L(t)}{A(t)} = \hat{B}_N(t) \ , \tag{64}$$

so that we can see why $\hat{B}_M(t)$ tends to have approximately the same variance as $\hat{B}_N(t)$ in general.

In the case of deterministic service times, equation (58) also simplifies because $\hat{\mu}^{-1}(t)$ is essentially constant. Equation (58) becomes

$$Var\hat{B}_M(t) = Var\hat{B}_I(t) + \frac{n^2}{\lambda^4}Var(\hat{\lambda}(t)) - \frac{2n}{\lambda^3}Cov(\hat{\lambda}(t) \ , \ \hat{n}(t)) = \frac{1}{\lambda^2}Var\left(\frac{n\hat{\lambda}(t)}{\lambda} - \hat{n}(t)\right) \ . \tag{65}$$

If in addition $\lambda = n$, then (63) and (65) imply that

$$Var\hat{B}_M(t) \approx \lambda^{-2}Var\frac{L(t)}{t} \approx Var\hat{B}_S(t) \ , \tag{66}$$

which further connects $\hat{B}_S(t), \hat{B}_N(t)$ and $\hat{B}_M(t)$.

From (63), we can also characterize the variance of the simple estimator in heavy loading. Assuming that $\hat{n}(t) \approx n$,

$$\hat{B}_S(t) = \frac{L(t)}{\lambda t} \approx \frac{A(t)}{\lambda t} - n$$

so that

$$Var\hat{B}_S(t) = \frac{Var\hat{\lambda}(t)}{\lambda^2} \ .$$

If $\{A(t) : t \geq 0\}$ is a Poisson process, then $Var\hat{B}_S(t) \approx t/\lambda$. In contrast,

$$Var\hat{B}_I(t) = \frac{Var \ \hat{n}(t)}{\lambda^2} \ .$$

Since $\hat{\lambda}(t)$ will be more variable than $\hat{n}(t)$ in heavy loading, we see the advantage of the indirect estimator.

6. Correlation Inequalities

In this section we prove Theorem 3 in Section 7. In particular, we establish the result by considering discrete-time loss models and invoking continuity theorems.

We now consider the discrete-time loss model. As before, there are s servers with no extra waiting room. We let arrivals occur before services in each period, so that blocking is determined before processing any service times. (In a limiting continuous-time model, usually two events never occur at the same time.)

Let a_k be the potential number of arrivals in period k. (This could be only 1 or 0, but we allow other possibilities.) Let c_{kj} be the potential service completion indicator variable for server j in period k; i.e., $c_{kj} = 1$ if the j^{th} customer has a potential service completion and $c_{kj} = 0$ otherwise; if server j is busy and if $c_{kj} = 1$, then there is a service completion. (Each server serves at most one customer in each period. We allow $c_{kj} = 1$ when the server is idle; then there is no actual service completion.)

Let x_{kj} be the j^{th} server occupancy indicator variable for period k; i.e., $x_{kj} = 1$ if the j^{th} server is busy after both arrivals and service completions in period k, and $x_{kj} = 0$ otherwise. Let y_{kj} be the j^{th} server pre-service occupancy indicator variable for period k; i.e., $y_{kj} = 1$ if the j^{th} server is busy after the arrivals but before the service completions in period k. (Our approach allows heterogeneous servers.)

We can describe the evolution of the system recursively through the equations

$$x_{kj} = 1 \text{ if } y_{kj} = 1 \text{ and } c_{kj} = 0 \ , \tag{67}$$

otherwise $x_{kj} = 0$; and

$$y_{kj} = 1 \text{ if } x_{k-1,j} = 1 \text{ or if } x_{k-1,j} = 0 \text{ and } j - \textstyle\sum_{i=1}^{j-1} x_{k-1,i} < a_k \ , \tag{68}$$

otherwise $y_{kj} = 0$.

Let n_k be the number of busy servers and let b_k be the number of blocked arrivals in period k. These are defined by

$$n_k = \sum_{j=1}^{s} x_{kj} \ , \quad k \geq 0 \ , \tag{69}$$

and

$$b_k = [a_k + n_{k-1} - s]^+, \quad k \geq 1 \ , \tag{70}$$

where $[x]^+ = \max\{x, 0\}$.

We start our treatment of the discrete-time model with the following elementary monotonicity result. See Berger and Whitt (1992) and references there for previous related work.

Theorem 9. *For $K \geq 1$, the variables x_{Kj}, y_{Kj}, n_K and b_K are nondecreasing functions of the vector $I_K \equiv (x_{0j}, a_k, -c_{kj}, 1 \leq j \leq s, 1 \leq k \leq K)$.*

In the most elementary setting, which includes the discrete-time analog of the $M/M/s/0$ model as a special case, we can regard the sequence $I_K \equiv \{x_{0j}, a_k, -c_{kj}, 1 \leq j \leq s, 1 \leq k \leq K\}$ as being composed of mutually independent random variables specified exogeneously. More generally, these variables will be dependent.

Recall that a collection of random variables $\{X_i : i \in \mathcal{I}\}$ is *associated* if the covariances satisfy

$$Cov(f(\{X_i : i \in \mathcal{I}\}),\ g(\{X_i : i \in \mathcal{I}\})) \geq 0 \tag{71}$$

for all nondecreasing real-valued functions f and g for which the covariance is well defined; e.g., see p. 29 of Barlow and Proschan (1975), p. 224 of Baccelli and Bremáud (1994) or p. 230 of Glasserman and Yao (1994). The following result is an immediate consequence of basic properties of associated random variables. It closely parallels Theorem 8 of Glynn and Whitt (1989).

Theorem 10. *If I_K is an associated set of random variables, then so is $\{I_K, x_{kj}, y_{kj}(1 \leq j \leq s), n_k, b_k,\ 1 \leq k \leq K\}$.*

The obvious sufficient condition for the condition of Theorem 10 is for the random variables in I_K to be mutually independent. This covers the discrete-time analog of the $M/M/s/0$ model.

We now define the discrete-time statistical estimators. Let

$$\hat{b}_K = K^{-1} \sum_{k=1}^{K} b_k \tag{72}$$

$$\hat{n}_K = K^{-1} \sum_{k=1}^{K} n_k \tag{73}$$

$$\hat{a}_K = K^{-1} \sum_{k=1}^{K} a_k \tag{74}$$

Corollary. *Under the condition of Theorem 10, $\hat{b}_K, \hat{n}_K$ and $\hat{a}_K$ are associated.*

The corollary implies that $\hat{\lambda}(t), \hat{B}_S(t)$ and $-\hat{B}_I(t)$ in the $M/M/s/0$ model are associated and thus are positively correlated. The $M/M/s/0$ result requires representing the $M/M/s/0$ models as a limit of a sequence of the discrete-time models.

We now want to treat models in which the variables in I_K are *not* mutually independent. It is natural to define the arrival variables a_k exogeneously. A natural sufficient condition for $\{a_1, \ldots, a_k\}$ to be associated is for the variables a_k to be *conditionally increasing in sequence,* i.e., for $E[f(a_k)|a_1, \ldots, a_{k-1}]$ to be nondecreasing in $(a_1, \ldots, a_{k-1})$ for all nondecreasing f and all k, $2 \leq k \leq K$; see Theorem 4.7 on p. 146 of Barlow and Proschan. Suppose that a_k are binary variables with the intervals between arrivals being i.i.d. Then the arrival process is a discrete-time renewal process. Let p_n be the probability that there are n periods between arrivals and let $p_n^c = 1 - (p_1 + \ldots + p_{n-1})$; i.e., p_n^c is the associated tail probability. It is easy to see that $\{a_k\}$ is conditionally increasing in sequence if and only if $\{p_n\}$ is DFR (has decreasing failure rate), i.e., if

$$\frac{p_n}{p_n^c} \geq \frac{p_{n+1}}{p_{n+1}^c} \tag{75}$$

because

$$P(a_n = 1 | a_{n-1} = a_{n-2} = a_{n-k+1} = 0, a_{n-k} = 1, a_j, j \leq n-k-1) = p_k / p_k^c\ . \tag{76}$$

We now want to consider dependent service completion variables c_{kj}. For this purpose, let w_{kj} be the server-j prolonged-service indicator variable defined by

$$w_{kj} = 1 \text{ if } y_{kj} = 1 \text{ and } c_{kj} = 0 \tag{77}$$

with $w_{kj} = 0$ otherwise. If $w_{kj} = 0$ and $w_{k+1,j} = \ldots = w_{k+m,j} = 1$, then there is a customer in service at server j in period $k + m$ who has been in service for m periods. Note that w_{kj} is increasing in $(y_{kj}, -c_{kj})$, so that Theorem 10 remains valid if we include the additional variables $w_{kj}, 1 \le j \le s, 1 \le k \le K$.

Now note that all the variables can be defined recursively. The order can be $x_{01}, x_{02}, \ldots, x_{0s}$, $n_0, a_1, y_{11}, y_{12}, \ldots y_{1s}, -c_{11}, -c_{12}, \ldots, -c_{1s}, w_{11}, w_{12}, \ldots, w_{1s}, x_{11}, x_{12}, \ldots, x_{1s}, n_1, b_1, a_2, y_{21}, y_{22}, \ldots, y_{2s}$, $-c_{21}, \ldots, -c_{2s}, w_{21}, w_{22}, \ldots, w_{2s}, x_{21}, \ldots, x_{2s}, n_2, b_2$, etc. Let $\mathcal{H}(\phi-)$ be the history to just before the symbol ϕ. For independent service processes, it is reasonable to assume that we have the following conditional distribution property

$$(c_{kj}|\mathcal{H}(c_{kj}-)) = (c_{kj}|c_{lj}, w_{lj}, 1 \le l \le k-1)) \quad \text{w.p.1} \; . \tag{78}$$

If the arrival process is exogeneously specified, then

$$(a_k|\mathcal{H}(a_k-)) = (a_k|a_1, \ldots, a_{k-1}) \quad \text{w.p.1} \tag{79}$$

which is what we considered above. The general theorem is as follows

Theorem 11. *Suppose that $(a_k|\mathcal{H}_k(a_k-))$ and $(-c_{kj}|\mathcal{H}_k(c_{kj}-))$ are conditionally increasing in sequence, then the conclusion of Theorem 10 holds.*

Proof. Since the other variables defined in (67), (68), (69), (70) and (77) are all monotone in preceding a_k and $-c_{kj}$ variables, the condition implies that the variables are conditionally increasing in sequence.

Corollary. *If (78) and (79) hold and if $(c_{kj}|c_{\ell j}, w_{\ell j},\ 1 \le j \le k-1)$ and $(a_k|a_1, \ldots, a_{k-1})$ are conditionally increasing, then the assumption of Theorem 11 holds, so that the conclusion of Theorem 10 holds.*

If the service times at each server are i.i.d. then

$$(c_{kj}|c_{\ell j}, w_{\ell j}, 1 \le \ell \le k-1) = (c_{kj}|w_{\ell j}, 1 \le \ell \le k-1) \; .$$

Moreover, paralleling the discussion of (76), $(-c_{kj}|w_{\ell j}, 1 \le \ell \le k-1)$ will be conditionally increasing if and only if the service times are IFR.

The DFR and IFR results above imply that the estimators are positively correlated in the $DFR/IFR/s/0$ model. Given DFR interarrival times, and IFR service times, we can represent the model as a limit of discrete-time models where the interarrival times are DFR and the service times are IFR. Hence, we have proved Theorem 3.

References

Baccelli, F. and P. Bremáud. 1994. *Elements of Queueing Theory*, Springer, New York.

Berger, A. W. and W. Whitt. 1992. Comparisons of multi-server queues with finite waiting rooms. *Stoch. Models* 8, 719–732.

Glasserman, P. and D. D. Yao. 1994. *Monotone Structure in Discrete-Event Systems*, Wiley, New York.

Whitt, W. 1980. Some useful functions for functional limit theorems. *Math. Opns. Res.* 5, 67–85.