Second Derivative Sample Path Estimators for the $GI/G/m$ Queue

Michael C. Fu • Jian-Qiang Hu
College of Business and Management, University of Maryland, College Park, Maryland 20742
Boston University, Boston, Massachusetts 02215

Applying the technique of smoothed perturbation analysis (SPA) to the $GI/G/m$ queue with first-come, first-served (FCFS) queue discipline, we derive sample path estimators for the second derivative of mean steady-state system time with respect to a parameter of the service time distribution. Such estimators provide a possible means for speeding up the convergence of gradient-based stochastic optimization algorithms. The derivation of the estimators sheds some new light on the complications encountered in applying the technique of SPA. The most general cases require the simulation of additional sample subpaths; however, an approximation procedure is also introduced which eliminates the need for additional simulation. Simulation results indicate that the approximation procedure is reasonably accurate. When the service times are exponential or deterministic, the estimator simplifies and the approximation procedure becomes exact. For the $M/M/2$ queue, the estimator is proved to be strongly consistent.

(*Perturbation Analysis; Sample Path Analysis; Simulation; Queues*)

1. Introduction

Perturbation analysis is a technique for estimating derivatives from a *single* sample path (or simulation) of a stochastic discrete-event system. An extensive bibliography on perturbation analysis can be found in Ho and Cao (1991). With the exception of Zazanis and Suri (1989), however, all previous work on derivative estimation has focused on the problem of first derivative estimation. In their work, which extended earlier results in the seminal paper of Suri and Zazanis (1988), Zazanis and Suri (1989) derived second derivative estimators for mean system time of a $GI/G/1$ queue through the use of conditional expectation in a technique generalized, formalized, and given the name smoothed perturbation analysis (SPA) by Gong and Ho (1987). In this paper, we apply SPA to second derivative estimation for multiserver queueing systems. The motivation for the work is twofold:

(a) The topic of estimating second derivatives from a single sample path, aside from being an interesting topic in itself, is important because second derivatives provide a possible means for speeding up the convergence rate of gradient-based stochastic optimization algorithms. The empirical work by Fu and Ho (1988) has demonstrated markedly improved convergence rates for second derivative-enhanced algorithms over the ordinary gradient algorithms in relatively short simulation experiments. Furthermore, second derivatives have been shown to be useful for characterizing, via interpolation, the entire performance curve as a function of parameters of interest such as mean service time or the arrival rate (Reiman et al. 1987).

(b) In applying the technique of smoothed perturbation analysis, practical issues confront one in implementing the estimators in an efficient manner. The main complication is the question of whether or not it is possible to easily estimate the conditional expectation from the sample path, with the difficulty arising from two sources: the complexity of the system and the form of the performance measure. Multiserver queues and second derivatives, respectively, correspond to each of the two sources of difficulties. Thus, the topic of

0025-1909/93/3903/0359$01.25

investigating these difficulties and finding efficient, albeit possibly approximate, procedures to get around them is an important one.

Related work on SPA includes Glasserman and Gong (1990), who consider special transient performance measures for systems satisfying a certain structural condition. The performance measures are such that the complications alluded to in (b) are not present, and they are able to come up with very clean estimators, for which unbiasedness can be proved. Also considering transient performance measures, Fu and Hu (1992) extend the work by deriving SPA estimators which are applicable to more general systems. However, the generality comes at a cost: as we demonstrate for the specific queueing system considered in this paper, the general form of the estimator they derive may require the need for additional simulation of sample subpaths to estimate certain quantities in the estimators. In Fu and Hu (1991b), the work on second derivatives for the $GI/G/1$ queue is extended by showing that for deterministic arrivals, the estimator derived by Zazanis and Suri (1989) fails. An alternative estimator is derived by choosing a different set of conditioning quantities with which to apply SPA, thus demonstrating that the appropriate choice of conditioning quantities may be crucial in applying SPA.

In our work, we consider the $GI/G/m$ queue and derive an estimator for the second derivative of mean steady-state system time with respect to a parameter of the service time distribution. The estimator we derive has three sources of contribution, one more than the estimator for the single-server case. The additional contribution comes from the interaction between the multiple servers. Two of the contributions can be estimated directly and easily from the usual simulation used to estimate the performance measure itself, whereas the remaining contribution requires additional simulation in the form of generating additional sample subpaths, which we call a splitting procedure. Since experimental evidence shows that this procedure is a practical alternative only at light traffic intensities, we introduce an approximation procedure to replace the splitting procedure. The approximation procedure requires no additional simulation, and experimental results seem to indicate reasonable accuracy. We also consider some special cases. For exponential service times, the approximation procedure yields an estimator stochastically equivalent to the estimator using the splitting procedure, but without the need for additional simulation. We provide a strong consistency proof for the $M/M/2$ queue. For phase-type service time distributions, a modified version of the approximation procedure also results in a consistent estimator that does not require additional simulation. Since a general distribution can be approximated arbitrarily closely by a phase-type distribution, this procedure in principle also provides a practical alternative to the splitting procedure. Lastly, for deterministic service times, the estimator simplifies to a single positive term which can again be estimated exactly without the splitting procedure.

The remainder of the paper is organized as follows. In §2, we give a statement of the two versions of the algorithm corresponding to the splitting procedure and to the approximation procedure. In §3, we derive the estimators and discuss the splitting procedure, the proposed approximation procedure, and the use of phase-type service time distributions. Section 4 contains the special cases of exponential and deterministic service times, and §5 contains simulation results for a number of examples, where the efficiency of the splitting procedure is discussed and the accuracy of the approximation procedure is tested. Section 6 contains a summary, conclusions, and some possible extensions of the work.

2. The Estimation Algorithm

Mean system time of a customer (time in queue plus in service) in steady state, denoted by ET, is our performance measure of interest, and we are interested in estimating $d^2ET/d\theta^2$, where $\theta \in \Theta$ is a parameter of the service time distribution and Θ is an open set. The estimator we derive consists of three components: an IPA (infinitesimal perturbation analysis) contribution, a positive SPA contribution, and a negative SPA contribution. Each of the contributions may be zero in special cases.

Assume that the system starts empty and that time 0 is defined as the arrival time of the first customer to the system. Let C_i be the ith customer to arrive to the system, and T_i be the system time of C_i. Let A_i represent the interarrival time between the $(i-1)$st and ith

customers, and $X_i(\theta)$ represent the service time of the ith customer to arrive. The interarrival times are i.i.d. with c.d.f. F and p.d.f. f, and the service times are i.i.d. (and independent of the interarrival times) with c.d.f. G and p.d.f. g.

We now introduce the derivative of a service time random variable with respect to a parameter of its underlying distribution. In particular, using the inverse representation, $X = G^{-1}(U; \theta)$, where $U \sim U(0, 1)$, we have w.p.1 (Zazanis and Suri 1989)

$$\frac{dX_i}{d\theta} = \frac{dX}{d\theta}\bigg|_{X_i} = -\frac{G_\theta}{G_x}\bigg|_{(X_i,\theta)} \tag{1}$$

and

$$\frac{d^2X_i}{d\theta^2} = -\frac{G_{\theta\theta}G_x^2 + G_{xx}G_\theta^2 - 2G_{\theta x}G_\theta G_x}{G_x^3}\bigg|_{(X_i,\theta)}, \tag{2}$$

where the subscripts of G denote partial differentiation with respect to the subscripted argument and $G(x, \theta)$ is assumed twice continuously differentiable with respect to both arguments on $\mathcal{B} \times \Theta$, where $\mathcal{B}$ is the union of the support sets of $G(\cdot, \theta)$ over $\theta \in \Theta$.

We also define the following quantities, which are generated during the simulation:

D_i = departure epoch of C_i,
$S(i)$ = server of C_i,
i' = the index of the next customer to depart after C_i,
$\xi_j(t)$ = age of service time at server j at epoch t (0 if idle), $j = 1, \ldots, m$,
$\xi_0(t)$ = age of interarrival time at epoch t,
$\tau_j(t)$ = residual service time at server j at epoch t (∞ if idle), $j = 1, \ldots, m$,
$\tau_0(t)$ = residual interarrival time at epoch t,
$\xi_i = \xi_0(D_i)$ = age of interarrival time at epoch D_i,
$\tau_i = \min_{j \neq 0} \tau_j(D_i)$ = time until next scheduled departure following C_i's departure at epoch D_i (∞ if empty),
$\nu_i = \xi_{S(i')}(D_i)$ = age of service time at server where next scheduled departure will occur at epoch D_i (0 if empty),
$\eta_i = \min_{j \neq S(i')} \tau_j(D_i)$ = time until next scheduled event not from $S(i')$ following C_i's departure at epoch D_i.

The last four quantities are illustrated in Figures 1 and 2, where in Figure 2, we have $S(i) = S$ and $S(i') = S'$. Depicted in these and subsequent figures is the status of a server (or servers) as a function of time:

—when the server is busy: indicated by the horizontal line above the time axis, and

—when services commence and are completed at the server: indicated by vertical lines.

Note that Figures 1 and 2 also include other quantities needed for the derivation of the estimators given in the next section.

In addition, two additional quantities must be estimated. We define a server's *local* busy period as *the interval between two consecutive idle times of the (same) server*. Thus, a single system busy period (which we will sometimes refer to as a *global* busy period) may contain any number of local busy periods of a particular server. Analogously, one can define a server's local idle period. The figures thus show alternating local busy-idle periods of the servers. We define the following important random variable:

$n(l, \mathbf{R})$ = r.v. number of customers served by server l from $t = 0^+$ until the next time server l is idle, given initial state at $t = 0^-$ is $\mathbf{R}$, where

$\mathbf{R} = (Q, \{R_j\}_{j=0}^m)$,

Q = number in system,

Figure 1 Sample Path Conditions and Quantities for SPA1 Contribution

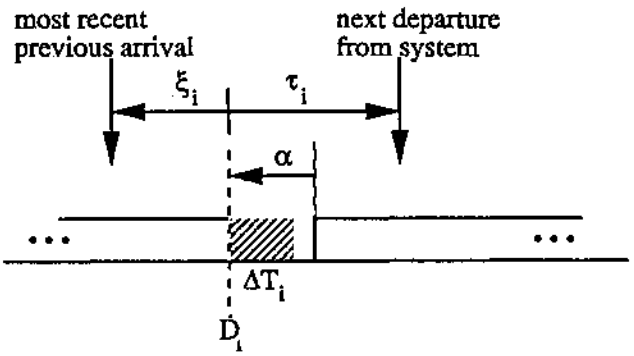
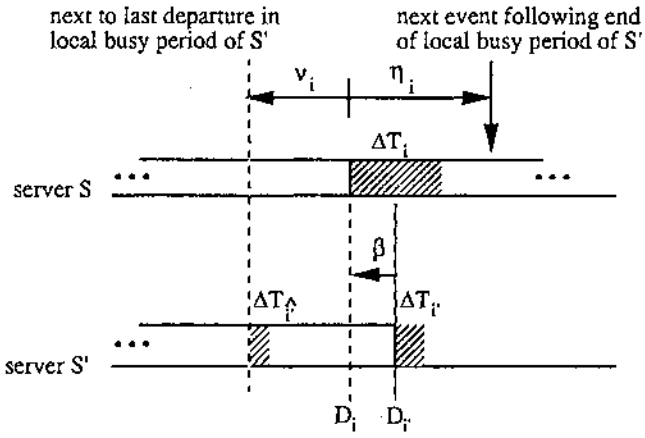


Figure 2 Sample Path Conditions and Quantities for SPA2 Contribution



R_0 = residual interarrival time,

R_j = residual service time at server j (∞ if idle),

$j = 1, \ldots, m$,

i.e., $n(l, \mathbf{R})$ is a random variable with the distribution of the number of customers served by server l until the end of l's present local busy period, given an initial state $\mathbf{R}$. The two additional quantities involve the following initial states:

$$\mathbf{R}_{(1),i} = (m - 1, \{R_j = \tau_j(D_i), j \neq 0, S(i);$$
$$R_{S(i)} = \infty; R_0 = 0\}),$$

$$\mathbf{R}_{(2),i} = (m + 1, \{R_j = \tau_j(D_i), j \neq S(i'), S(i);$$
$$R_{S(i)} = 0; R_{S(i')} = 0^+\}).$$

The initial state $\mathbf{R}_{(1),i}$ is such that at time 0, $S(i)$ is the only idle server, but since $R_0 = 0$, an arrival occurs at time 0, and this arrival goes to $S(i)$, beginning a new local busy period for $S(i)$. The initial state $\mathbf{R}_{(2),i}$ is such that at time 0, $S(i)$ will complete service of its present customer and receive the only customer in queue. The departure at $S(i')$ occurs at time 0^+, so $S(i')$ will go idle then, since there will no longer be any customer in the queue.

We now introduce the following notation to define the quantities of interest, which are conditional expectations:

$$n_{(1),i} = E[n(S(i), \mathbf{R}_{(1),i})], \tag{3}$$

$$n_{(2),i} = E[n(S(i), \mathbf{R}_{(2),i})], \tag{4}$$

where the subscript i serves as a reminder that the quantity is a conditional expectation dependent on various quantities associated with C_i. Intuitively, $n_{(1),i}$ is the expected number of customers served in a local busy period that begins with an arrival to the only idle server, with the residual service times at the other servers given by the residual service times at the departure time of C_i. Similarly, $n_{(2),i}$ is the expected number of customers served in the continuation of a local busy period of a server that has just finished serving a customer and will receive the only customer in queue and such that another server will just finish serving a customer, with the residual interarrival time and residual service times at the other servers given by the corresponding quantities at the departure time of C_i.

We now state the algorithm, Algorithm 1, which is given in the form of additional calculations that must be implemented in the subroutines already in place in the simulation program. For convenience, we have also included the estimator for the first derivative.

ALGORITHM 1.

INITIALIZATION:

DTSUM=IPA=0

DT(S)=D2T(S)=0 for S=1,...,m

DEPARTURE—At the end of service of customer i at server S:

DT(S)=DT(S)+$dX_i/d\theta$

D2T(S)=D2T(S)+$d^2X_i/d\theta^2$

IPA=IPA+D2T(S)

DTSUM=DTSUM+DT(S)

NB: Scheduling of next departure at server S, if any, must be done prior to following IF statements.

IF departure brings the number in system to $m - 1$ AND the next event is an arrival

THEN SPA1=SPA1+$n_{(1),i} * f(\xi_i)/(F(\xi_i + \tau_i) - F(\xi_i)) * (\text{DT(S)})^2$

IF departure brings the number in system to m
AND the next event is a departure of customer i' from server $S' \neq S$
THEN SPA2=SPA2+$n_{(2),i} * g(v_i)/(G(v_i + \eta_i) - G(v_i)) * ([DT(S) - DT(S') - (dX/d\theta)|_{v_i}]^+)^2$
IF no one waiting in queue THEN DT(S)=D2T(S)=0

OUTPUT STATISTICS—At the end of N customers served:
$(d^2ET/d\theta^2)_{est}$=(IPA+SPA1−SPA2)/N
$(dET/d\theta)_{est}$=DTSUM/N

The term $(dX/d\theta)|_{v_i}$ is the rightmost side of equation (1) evaluated at v_i. Also note that for the single-server case, the condition for the SPA2 term will never be satisfied, so the SPA2 contribution is zero. Since $\tau_i = \infty$, $F(\xi_i + \tau_i) = 1$ in the SPA1 term, and the estimator reduces to the estimator derived in Zazanis and Suri (1989) for the $GI/G/1$ queue.

We can see that the estimator contains three contributions, which intuitively estimate the following effects due to the introduction of a perturbation into the sample path:

(1) IPA contribution—effects when no changes in the order of events take place;

(2) SPA1 contribution—effects due to a coalescing of two consecutive local busy periods of the same server;

(3) SPA2 contribution—effects due to a change in order of two adjacent departures from *different* servers.

The first two effects are analogous to those used in deriving the estimator for the second derivative of the $GI/G/1$ queue, while the third is unique to the phenomenon of multiple servers. We will see that the two SPA contributions have opposite effects on the estimator, e.g., under a positive perturbation in the service times, the SPA1 contribution would be positive, while the SPA2 contribution would be negative.

One way to estimate $n_{(k),i}$, $k = 1, 2$, is the following obvious manner which we shall call a *splitting* procedure, because it involves the generation of additional sample subpaths split off from the main simulation. At a point where an estimate is needed, simply run a sample subpath simulation which begins with the state defining $n_{(k),i}$ and ends at the end of the local busy period of $S(i)$. This will give a single sample which provides an unbiased estimate. Of course, it is obvious that this can be done multiple times to get a more accurate estimate, so there is much leeway in how this estimation is actually implemented. In our simulation examples, we use just a single sample, and we shall see that even for this choice, the amount of extra simulation quickly becomes impractical for moderate traffic intensities. For $n_{(2),i}$, there is a way to get around the need for additional simulation, and the procedure is exact. For $n_{(1),i}$, we propose an approximation procedure that eliminates the need for additional simulation, the idea being to average over the residual service times that occur in the sample path.

We define the initial condition

$$\mathbf{R}_{(1)} = (m - 1, \{R_j > 0, j \neq 0, l; R_l = \infty; R_0 = 0\}),$$

which is simply the beginning of a local busy period of the server l which starts with m in the system (since an arrival occurs at time $t = 0$). We now define the quantity

$$n_{(1)} = E[n(l, \mathbf{R}_{(1)})]. \tag{5}$$

We note that this quantity is independent of l and of any of the other quantities associated with C_i. In words, it is the expected number of customers served in a local busy period of a server which starts with m in the system (since an arrival occurs at time $t = 0$). Its expectation can be easily estimated from the original simulation simply by taking samples every time an arrival brings the number in system to m. This procedure can be thought of as approximately averaging $n_{(1),i}$ over the possible vector of residual service times at the other servers.

On the other hand, $n_{(2),i}$ can be easily estimated from the original sample path. In addition, in actual implementation, there is a way to avoid explicitly estimating $n_{(2),i}$ as a separate quantity by replacing its estimation with a "propagation" procedure. The details and validity of this estimation scheme are discussed in the next section where the SPA2 contribution is derived. We now state Algorithm 2, which implements the two modifications discussed in the last two paragraphs.

ALGORITHM 2.

INITIALIZATION:

DTSUM=IPA=0

DT(S)=D2T(S)=SPA2(S)=0 for S=1,...,m

DEPARTURE—At the end of service of customer i at server S:

DT(S)=DT(S)+$dX_i/d\theta$

D2T(S)=D2T(S)+$d^2X_i/d\theta^2$

SPA2=SPA2+SPA2(S)

IPA=IPA+D2T(S)

DTSUM=DTSUM+DT(S)

NB: Scheduling of next departure at server S, if any, must be done prior to following IF statements.

IF departure brings the number in system to $m-1$ AND the next event is an arrival

THEN SPA1=SPA1+$f(\xi_i)/(F(\xi_i+\tau_i)-F(\xi_i))*(\text{DT(S)})^2$

IF departure brings the number in system to m

AND the next event is a departure of customer i' from server S'≠S

THEN SPA2(S)=SPA2(S)+$g(\nu_i)/(G(\nu_i+\eta_i)-G(\nu_i))*([\text{DT(S)}-\text{DT(S')}-(dX/d\theta)|_{\nu_i}]^+)^2$

IF no one waiting in queue THEN DT(S)=D2T(S)=SPA2(S)=0

OUTPUT STATISTICS—At the end of N customers served:

$(d^2ET/d\theta^2)_{est}$=(IPA+$n_{(1)}$*SPA1−SPA2)/N

$(dET/d\theta)_{est}$=DTSUM/N

The differences between Algorithm 1 and Algorithm 2 are the replacement of the estimate of $n_{(2),i}$ with the use of SPA2(·) for each server, and the replacement of the local (i.e., at each departure) estimate of $n_{(1),i}$ by a global (over the entire simulation) estimate of $n_{(1)}$, which can be easily estimated from the original simulation without the need for additional simulation. As we have discussed already, the latter is an approximation except for certain service time distributions.

3. Derivation of the Estimators

In this section, we derive the three contributions in the estimator for $d^2ET/d\theta^2$. In our actual derivation, our sample performance will be the customer average $\sum_{i=1}^{N} T_i/N$, which under mild conditions converges (as $N \to \infty$) w.p.1 to ET.

3.1. IPA Contribution

We assume that the system empties infinitely often and that the system time has a steady-state distribution (see, e.g., Whitt 1972 for sufficient conditions). We first concentrate on a single customer C_i. Recall that a local busy period was defined as the interval between two consecutive idle times of the same server. We say that two customers are in the same local busy period if they are served by the same server in the same local busy period, i.e., that server has not been idle at any time between the service of the two customers. To derive the IPA contribution, we introduce the following notation:

$$L(i) = \{j < i : C_j \text{ in the same local busy period as } C_i\},$$

$$\hat{i} = \max_{j \in L(i)} j,$$

i.e., $L(i)$ is the set of indices of all customers preceding C_i in the same local busy period, and $\hat{i}$ is the index of the customer preceding C_i in the local busy period.

The IPA contribution considers only sufficiently small perturbations and ignores possible changes in the order of events (arrivals and departures). The IPA contribution for the kth derivative can be written as (cf. e.g., Fu and Hu 1991a)

$$\frac{d^kT_i}{d\theta^k} = \sum_{j \in L(i)} \frac{d^kX_j}{d\theta^k} + \frac{d^kX_i}{d\theta^k}, \tag{6}$$

and in recursive form as

$$\frac{d^kT_i}{d\theta^k} = \begin{cases} d^kX_i/d\theta^k & \text{if } C_i \text{ initiates local busy period,} \\ d^kT_{\hat{i}}/d\theta^k + d^kX_i/d\theta^k & \text{otherwise.} \end{cases} \tag{7}$$

Intuitively, the IPA contribution to the perturbation in C_i's system time is the perturbation in customer C_i's service time, X_i, plus all the IPA contributions of the customers preceding C_i in the same local busy period. Taking $k = 2$, the sample mean over all customers simulated constitutes the IPA contribution of the estimator. The $k = 1$ case will be needed for both SPA contributions.

For reference, we now write out the complete estimator in mathematical form before deriving the two SPA contributions:

$$\left(\frac{d^2ET}{d\theta^2}\right)_{\text{est}} = \frac{1}{N}\left[\sum_{i=1}^{N}\frac{d^2T_i}{d\theta^2} + \sum_{i\in C_{(1)}}\frac{f(\xi_i)}{F(\xi_i+\tau_i)-F(\xi_i)} \times\left(\frac{dT_i}{d\theta}\right)^2 n_{(1),i} - \sum_{i\in C_{(2)}}\frac{g(\nu_i)}{G(\nu_i+\eta_i)-G(\nu_i)} \times\left(\left[\frac{dT_i}{d\theta}-\frac{dT_{\hat{i}}}{d\theta}-\frac{dX}{d\theta}\bigg|_{\nu_i}\right]^+\right)^2 n_{(2),i}\right], \quad (8)$$

where N is the number of customers simulated, and $C_{(1)}$ and $C_{(2)}$ are sets of customer indices indicating the two types of SPA contributions, defined by the following:

$C_{(1)}$ = {indices of customers that depart leaving $m-1$ in the system and such that the next event is an arrival}, (9)

$C_{(2)}$ = {indices of customers that depart leaving m in the system and such that the next event is a departure at another server}. (10)

The three sums averaged over N are the IPA, SPA1, and SPA2 contributions, respectively, as implemented by Algorithm 1. The squared terms in both SPA contributions are the first derivative IPA accumulations for the customers in $C_{(1)}$ and $C_{(2)}$. The two SPA contributions will be derived in the following subsections.

3.2. Origins of the SPA Contributions

Two events will be called *adjacent* if no other event occurs between them. When mean system time is the performance measure of interest (see Suri and Zazanis 1988 for a detailed discussion for the single-server case), the expected effect of a change in the order of two adjacent events along a sample path due to the introduction of the perturbation $\Delta\theta$ is typically of $O((\Delta\theta)^2)$, because the probability of the change is of $O(\Delta\theta)$ and the effect given the change is also of $O(\Delta\theta)$. Therefore, IPA suffices by itself for estimating first derivatives of mean system time, but for second derivatives, these effects must be included in the estimators. Nonadjacent event order changes can be ignored for the purpose of calculating second derivatives because they occur with probability of $o(\Delta\theta)$ and for identical servers have conditional effect of $O(\Delta\theta)$, thus having expected effects of $o((\Delta\theta)^2)$. Similar arguments are made rigorous in, e.g., Glasserman and Gong (1990), Fu and Hu (1992).

Thus, we consider only adjacent event order changes. We derive the SPA contributions under the assumption that $\Delta\theta > 0$, i.e., we calculate the right-hand limit as $\Delta\theta \to 0^+$. For the $GI/G/m$ queue, there are only four types of adjacent events: arrival/arrival, arrival/departure, departure/arrival, departure/departure. However, it is obvious that arrivals will never change order due to a perturbation and that departures from the same server will also never change order due to a perturbation. Furthermore, for a positive perturbation in the service times, an arrival will never overtake a departure. Thus, an event order change can occur only between a departure and an arrival and between departures of different servers. Unless the departure/arrival interchange occurs at the same server, there is no additional contribution from the interchange, since the arrival would be unaffected by the departure. Hence, by elimination, the only adjacent event situations which potentially cause additional contribution are the following:

(i) between two adjacent local busy periods of the same server, i.e., a departure ending a local busy period of a server overtakes an arrival beginning the next local busy period of the same server, referred to as coalescing of local busy periods;

(ii) between local busy periods of different servers, causing customers to change servers in the perturbed path, i.e., a departure from one server overtakes the departure from another server, causing the next customer in queue to be served by a different server.

These two cases correspond to the SPA1 and SPA2 contributions, respectively. Intuitively, SPA contributions

are derived by the following procedure under their respective adjacent event order changes:

(a) introduce a (virtual) perturbation $\Delta\theta$ into the sample path;

(b) trace its effect on the performance measure by deriving an expression for $E[\Delta_n T \mid z]$;

(c) take $\lim_{\Delta\theta\to 0} E[\Delta_n T|z]/(\Delta\theta)^n/n!$ to get an estimator of the contribution for $d^n ET/d\theta^n$,

where $\Delta_n T$ represents the nth term in the Taylor series representation of $T(\theta + \Delta\theta)$ at θ, z represents a set of conditioning quantities available from the sample path that must be selected, called the characterization in Gong and Ho (1987). The original (unperturbed) path at parameter value θ is referred to as the *nominal* path, whereas the path at parameter value $\theta + \Delta\theta$ is referred to as the *perturbed* path. Since we are considering a customer average, we will be considering

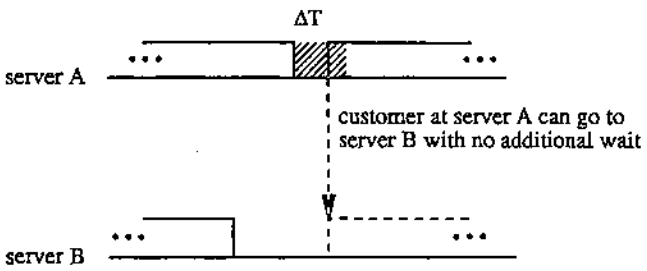
$$\frac{1}{N}\sum_{i=1}^{N} \lim_{\Delta\theta\to 0} \frac{E[\Delta_2 T_i|z_i]}{(\Delta\theta)^2/2}. \qquad (11)$$

In tackling these two cases, we hope that the accompanying figures aid in the understanding of the derivations.

3.3. SPA1 Contribution

A coalescing of local busy periods results in an additional contribution only if in addition to the departure and arrival being adjacent events, all other servers—of which there are $(m-1)$—are busy between these two events. Otherwise, even under the event order change of the departure overtaking the arrival, the arriving customer can go to another idle server, so that no coalescing of busy periods will occur and hence the arriving customer experiences no increase in its system time. Figure 3 demonstrates this case (recall that vertical lines within local busy periods indicate customer departures). Thus, the first SPA term, which we will refer to henceforth as the SPA1 term, is calculated by conditioning at the end of each local busy period that takes the number in system from m down to $(m-1)$, (i.e., just before the end of the local busy period, there were no customers in queue and all servers were busy), and under the condition that the next event is an arrival which begins the local busy period. Equivalently, we say that only departing customers under the above conditions generate a SPA1 term, which leads to the definition of $C_{(1)}$ given by equation (9).

Figure 3 Case (i) Not Resulting in Additional Contribution



Now consider a customer C_i, $i \in C_{(1)}$, that departs at D_i, followed subsequently by an arrival at $D_i + \alpha$, as shown in Figure 1. What we wish to estimate is the change in total system time due to a perturbation. If the total accumulated IPA perturbation ΔT_i, which is a function of $\Delta\theta$, is greater than the server's local idle period α, then coalescing occurs, and customers in the next local busy period of the same server experience an additional wait of $(\Delta T_i - \alpha)$. If there are n^* customers in the next local busy period, then the total increase in system time is $n^*(\Delta T_i - \alpha)^+$, so we need to estimate

$$\lim_{\Delta\theta\to 0} E[n^*(\Delta T_i - \alpha)^+ | z_i(\Delta\theta)],$$

where z_i is the set of conditioning quantities along the sample path. Conditioned on the $z_i(\Delta\theta)$ we will be choosing, the two terms are independent, so we have

$$\lim_{\Delta\theta\to 0} \frac{E[\Delta_2 T_i|z_i(\Delta\theta)]}{(\Delta\theta)^2/2} = \lim_{\Delta\theta\to 0} \frac{E[(\Delta T_i - \alpha)^+|z_i(\Delta\theta)]}{(\Delta\theta)^2/2} \times \lim_{\Delta\theta\to 0} E[n^*|z_i(\Delta\theta)]. \qquad (12)$$

The latter term is by definition the expected number of customers in a local busy period that begins with state described by m in the system, residual service times $\{\tau_j(D_i)\}_{j\neq 0,S(i)}$, a full service time at $S(i)$, and a zero residual interarrival time, which is equivalent to our definition of $n_{(1),i}$ given by equation (3).

Thus, we need to estimate two terms: $E[(\Delta T_i - \alpha)^+|z_i]$ and $n_{(1),i}$. We start with the latter term. Ideally, we would like to be able to estimate $n_{(1),i}$ from the nominal sample path at $D_i + \alpha$. Unfortunately, the nominal path gives the number of customers in a local busy period which begins with residual service times $\{\tau_j(D_i) - \alpha\}_{j\neq 0}$, with $\alpha = \tau_0(D_i) \neq 0$, which in general does not have expectation equal to $n_{(1),i}$, the latter being the

expectation of the limiting distribution as $\alpha \to 0$. In the single-server case, the number of customers in the succeeding busy period is unaffected by this condition (trivially, since there are no τ_i's at other servers to consider), but in the multiserver case, the number in the succeeding local busy period also depends on the residual service times at the other servers, which in general will change as $\alpha \to 0$. Thus in general $n_{(1),i}$ cannot be estimated exactly from the nominal sample path. There are two interesting exceptions, when the service times are exponential or deterministic, which we discuss in the next section. In order to calculate $n_{(1),i}$ for the general case, we propose two procedures:

(1) *Splitting*. At D_i, begin a separate simulation to estimate $n_{(1),i}$ by forcing an arrival at D_i, ending the inserted simulation when the local busy ends; this can be repeated to get multiple samples and hence a better estimate, but even for the case of a single sample for each $i \in C_{(1)}$, simulation results seem to indicate that this procedure is practical only in light traffic.

(2) *Approximation*. Approximate $n_{(1),i}$ for all i by the average number served in a local busy period beginning with m in the system, a quantity which can be estimated easily from the nominal path, and which we defined as $n_{(1)}$ in equation (5). It is obviously an approximation, since it ignores the residual service times, but because it is an average over different possible residual service times, it might be expected to do fairly well, and simulation results seem to support this view.

The crux of the approximation procedure is the averaging over all possible vectors of residual service times. The advantage is that no additional simulation is required. For exponential service times, which have the memoryless property, the estimate obtained through this procedure is unbiased. Otherwise, in general, it is not clear as to the nature of any possible bias. Under our assumptions of continuous service and interarrival times, the initial state for $n_{(1),i}$ is a continuous random variable, with an uncountable number of possible initial states, each state occurring on the sample path with probability zero. This is why the splitting procedure is needed to estimate $n_{(1),i}$; one has to set the initial state in running the additional simulation subpath. However, the memoryless property of the exponential distribution implies that the distribution of $n(l, \mathbf{R}_{(1),i})$ is the same for all i, and thus $n_{(1),i}$ can be estimated by the number served in any local busy period which begins with m in the system, *irrespective* of the residual service times at the other servers; since this condition occurs on the sample path with probability one, no additional simulation is needed.

Another way to incorporate this memoryless property is through the use of phase-type distributions, where we must keep track of the phases rather than ages. Since the initial state defining $n_{(1),i}$ has the interarrival residual time zero, we only need the service times to be phase-type; the interarrival time distribution can be general. Now, the number of different initial distributions which are possible is simply the (finite) number of different combinations of phases under the initial state defining $n_{(1),i}$. For example, for a k-Erlang service time distribution, the number of different $n_{(1),i}$ that have to be considered is simply k^{m-1}. Perhaps more important is the fact that these finite number of $n_{(1),i}$ can be estimated without the need for additional simulation, as long as we keep track of the phase that the server is in. For example, for a two-server, 2-Erlang service time distribution, we would just need to estimate $n_{(1),i}$ from two possible initial states: those which start a local busy period (bringing the number in system up to 2) with the other server in its first phase of service, and those which start a local busy period (bringing the number in system up to 2) with the other server in its second phase of service.

Now we return to the problem of deriving an estimator for $E[(\Delta T_i - \alpha)^+ | z_i]$. The set of conditioning quantities z_i should be chosen such that the conditional density of α can be calculated at D_i (see, e.g., Gong and Ho 1987 or Zazanis and Suri 1989). Since α is the residual interarrival time, we can condition on the age. Next, we need to satisfy the conditions defining $C_{(1)}$ under which the adjacent event order change (coalescing of local busy periods) occurs. Under this condition, the departure at D_i occurs with no customers in queue, and the next event should be an arrival. This requires that α be less than the minimum of the outstanding residual service times at D_i. Thus, the choice for z_i we use is the following: the accumulated IPA perturbation (for ΔT_i), the minimum of all the residual service times, and the age of the interarrival time. This choice appears to be the minimum set which allows us to easily compute a (conditional) density for α. Other choices are

possible (e.g., we could include all the residual service times). Using the conditional density for α, we can then calculate the conditional expectation

$$E[(\Delta T_i - \alpha)^+ | \Delta T_i, \tau_i, \xi_i, \alpha \le \tau_i],$$

where recall that ξ_i is the age of the interarrival time at D_i, and τ_i is the minimum of all outstanding residual service times at D_i.

The random variable α is simply $\tau_0(D_i)$, and we wish to find its conditional distribution given the age at D_i and under the additional condition that it is less than the outstanding residual service times at D_i. We note that without the addition of the latter condition, the density would be the same as that found in Zazanis and Suri (1989) for the single-server case, a hazard rate-like function. Let A denote a generic interarrival time with p.d.f. $f(\cdot)$ and c.d.f. $F(\cdot)$. For fixed t and τ, we have

$$\begin{aligned} P(\tau_0(t) \le y | \xi_0(t) = \xi, \tau_0(t) \le \tau) &= P(A - \xi \le y | A \ge \xi, A - \xi \le \tau) \\ &= \frac{P(\xi \le A \le \xi + \min(y, \tau))}{P(\xi \le A \le \xi + \tau)} \\ &= \frac{F(\xi + \min(y, \tau)) - F(\xi)}{F(\xi + \tau) - F(\xi)} \\ &= \begin{cases} 1 & \text{if } \tau \le y, \\ \dfrac{F(\xi + y) - F(\xi)}{F(\xi + \tau) - F(\xi)} & \text{if } \tau > y. \end{cases} \end{aligned} \tag{13}$$

Since the point of interest for us is D_i, and we are interested in the case where τ represents the minimum of the residual service times at time D_i, we take $t = D_i$ and $\tau = \tau_i$, and differentiating equation (13), we get the conditional density for α:

$$f_\alpha(y) = \begin{cases} 0 & \text{if } \tau_i \le y, \\ \dfrac{f(\xi_i + y)}{F(\xi_i + \tau_i) - F(\xi_i)} & \text{if } \tau_i > y, \end{cases} \tag{14}$$

where recall that ξ_i is the age of the interarrival time at D_i, and τ_i is the minimum of all outstanding residual service times at D_i. We note that since D_i is actually random and not fixed, a completely rigorous justification of this derivation would require showing that the appropriately supplemented queue-length process is strong Markov and that D_i is a stopping time for it.

In terms of the IPA accumulations, we have $\Delta T_i = (dT_i/d\theta)\Delta\theta$, where $(dT_i/d\theta)$ is given by equation (6), with $k = 1$. Integrating w.r.t. f_α we get the expectation

$$\begin{aligned} &E[(\Delta T_i - \alpha)^+ | \Delta T_i, \tau_i, \xi_i, \alpha \le \tau_i] \\ &= \int_0^\infty \left[\frac{dT_i}{d\theta}\Delta\theta - y\right]^+ \frac{f(\xi_i + y)}{F(\xi_i + \tau_i) - F(\xi_i)}\, dy \\ &= \frac{1}{2}\frac{f(\xi_i + \gamma)}{F(\xi_i + \tau_i) - F(\xi_i)}(\Delta T_i)^2 \quad \text{for some } \gamma \in [0, \Delta T_i], \end{aligned}$$

where we have applied the mean value theorem for integrals under the assumption that f is continuous. Dividing by $(\Delta\theta)^2/2$ and taking the limit, we get the SPA1 contribution for $i \in C_{(1)}$:

$$\frac{f(\xi_i)}{F(\xi_i + \tau_i) - F(\xi_i)}\left(\frac{dT_i}{d\theta}\right)^2 n_{(1),i}. \tag{15}$$

Taking the sample average over N customers gives the SPA1 contribution in equation (8).

We make one final remark with regard to the implementation in Algorithm 2. For the approximation procedure, we can estimate $n_{(1)}$ every time an arrival takes the number in system to m, instead of the more restrictive $C_{(1)}$ condition, which requires a departure preceding the beginning of the local busy period. Thus, to implement the approximation procedure, simply replace the second summation in the complete estimator given by equation (8) by the following:

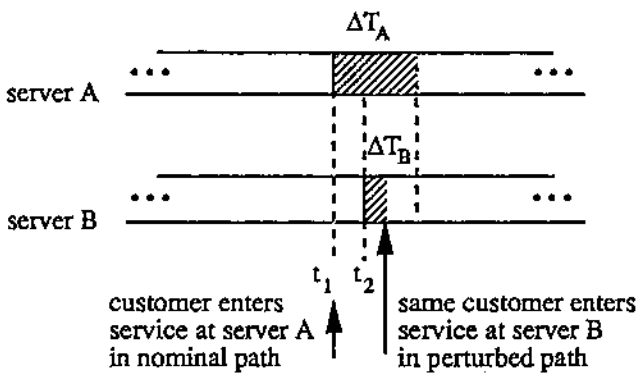
$$\frac{1}{N}\sum_{i \in C_{(1)}} \frac{f(\xi_i)}{F(\xi_i + \tau_i) - F(\xi_i)}\left(\frac{dT_i}{d\theta}\right)^2 * \frac{1}{\tilde{N}}\sum_{j \in \tilde{C}_{(1)}} \hat{n}_{(1),j}, \tag{16}$$

where $\tilde{C}_{(1)}$ is the set of indices of all the customers who begin a local busy period by bringing the number in system to m, $\hat{n}_{(1),j}$ is the number of customers served in the local busy period that $j \in \tilde{C}_{(1)}$ initiates, and $\tilde{N}$ is the cardinality of $\tilde{C}_{(1)} \cap \{C_1, C_2, \ldots, C_N\}$. For exponential and deterministic service times, we argue in the next section that this approximation is exact, and then go on to give a consistency proof for the $M/M/2$ queue. The accuracy of the approximation procedure in general and the computational burden of the splitting procedure are discussed in §5, where simulation examples are presented.

3.4. SPA2 Contribution

We turn to the SPA2 contribution now. Whereas the SPA1 contribution involved coalescing of local busy periods of the same server, the SPA2 contribution involves interaction between local busy periods of different servers. This type of event order change has not appeared in previous SPA applications. An example of the event order change is shown in Figure 4. If ΔT_A is sufficiently larger than ΔT_B ($\Delta T_A > \Delta T_B + t_2 - t_1$), then the departures at t_1 and t_2 would switch order in the perturbed path, since the service is FCFS. The customer who entered service at server A at t_1 in the nominal path (and at $t_1 + \Delta T_A$ in the perturbed path if no event order change took place) would switch service and enter service at server B at time $t_2 + \Delta T_B$ in the perturbed path. Similarly, the customer who entered service at server B at t_2 in the nominal path would switch to server A, entering service at time $t_1 + \Delta T_A$. As a result of the switching, one of the customers would experience a shortening of his wait over the nonswitching situation—by the amount $(t_1 + \Delta T_A - [t_2 + \Delta T_B])$—and the other would experience a lengthening of the same amount. This perturbation due to switching would be propagated to succeeding customers in the same local busy periods and thus the net effect would depend on the number in each of the two server's local busy periods. Since the servers are identical, and noting that if switching occurs, then the condition $\Delta\theta \to 0$ implies that $t_2 \to t_1$, the following symmetry argument tells us that the net expected effect would be zero: Both servers have customers at t_1 and t_2, so there is a positive and negative perturbation generated by the switching. Furthermore, as $t_2 \to t_1$, the remaining number in each of the two server's local busy periods would be equal in distribution; thus, the accumulated positive and negative perturbations would be equal in expectation, yielding no net contribution.

Figure 4 Event Order Change with Switching Between Servers



However, if one of the servers has a customer to switch, but the other does not, then only one perturbation is generated. Since customers go to a server as soon as one is available, there is only one case to consider, i.e., the situation in Figure 5—where a customer enters service at server A at the vertical dotted line—is impossible, because the customer could have gone to the idle server B earlier. Thus, the appropriate situation for the SPA2 term is a departure which takes the number in system from $m + 1$ down to m, and under the condition that the next event is a departure at another server (which will end that server's local busy period since it will take the number in system down to $m - 1$). Only departing customers under the above conditions generate a SPA2 term, which leads to the definition of $C_{(2)}$ given by equation (10).

Again, we consider a customer who fulfills conditions of interest, i.e., in Figure 2, we consider a customer C_i, $i \in C_{(2)}$, that departs at D_i from server $S(i) = S$, followed subsequently by a departure at another server, $S(i') = S'$, at $D_i + \beta$. Under no switching, the customer who entered service at server S at time D_i in the nominal sample path would enter service at server S at $D_i + \Delta T_i$ in the perturbed path. Under the switching condition, the customer enters service at server S' at $D_i + \beta + \Delta T_{i'}$, experiencing a net *shorter* wait of $(\Delta T_i - \Delta T_{i'} - \beta)$. It is crucial to note here that the switch will preserve all other local busy period relationships, i.e., all customers in C_i's local busy period served subsequently by server S in the nominal sample path will also switch to S' in the perturbed path. Thus, all succeeding customers in the local busy period of C_i (at server S in the nominal path) will experience this *negative* perturbation in system time of size $(\Delta T_i - \Delta T_{i'} - \beta)$. We again need to estimate two terms: $E[(\Delta T_i - \Delta T_{i'} - \beta)^+ | z_i]$ and the number of customers which are affected by this term. We start with the latter term, which we would like to be able to estimate this quantity from the nominal sample path at D_i. In this case, we are more fortunate than for the SPA1 case. Looking at Figure 2, we see that the perturbation is generated under the condition $\beta \to 0$, i.e., we are interested in the expected number of

customers in the continuation of the local busy period at server $S(i)$, with state described by $m - 1$ in the system, ages $\{\xi_j(D_i)\}_{j \neq S(i')}$, and a zero residual service time at $S(i')$, which is equivalent to our definition of $n_{(2),i}$ given by equation (4). Analogous to the analysis of the previous subsection, we have

$$\lim_{\Delta\theta \to 0} \frac{E[\Delta_2 T_i | z_i]}{(\Delta\theta)^2/2} = \lim_{\Delta\theta \to 0} \frac{E[(\Delta T_i - \Delta T_{i'} - \beta)^+ | z_i]}{(\Delta\theta)^2/2} n_{(2),i}. \quad (17)$$

The nominal path gives the number of customers in the continuation of the local busy period at $S(i)$ with state described by m in the system, residual times $\{\xi_j(D_i)\}_{j \neq S(i')}$, and $\beta = \tau_{S(i')}(D_i) \neq 0$ residual service time at $S(i')$, which we shall now argue has expectation given by $n_{(2),i}$. The only difference between the nominal path and the perturbed path is the difference in the residual service times at $S(i')$. Under the conditions of $C_{(1)}$, this service time triggers the next event, a departure which ends the local busy period at server $S(i')$. From Figure 2, it is clear that taking $\beta \to 0$ only shortens the end of the local busy period at $S(i') = S'$. Thus, the number of customers served in the continuation of $S(i)$'s local busy period is unaffected in this limit, so we can use the original nominal sample path to easily get a single unbiased sample of $n_{(2),i}$, which is implemented in Algorithm 2 through a "propagation" procedure in which all subsequent customers in the continuation of the local busy period of $S(i)$ receive the perturbation. Essentially, instead of multiplying by an estimate of $n_{(2),i}$ at the end of the simulation, Algorithm 2 updates a sum every time a customer completes service in the continuation of an appropriate local busy period. We reiterate that for the SPA2 contribution, this procedure is exact.

Figure 5 **Impossible Situation**

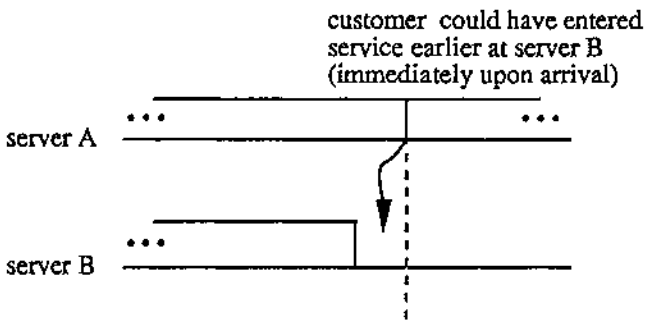
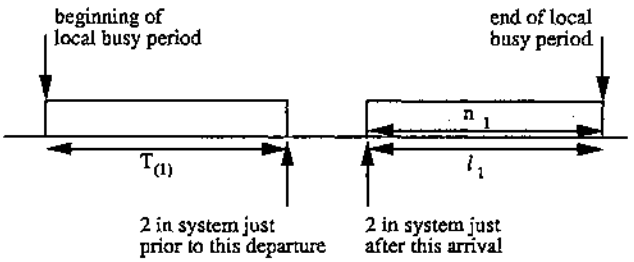


Figure 6 **In Reversed Sample Path, $T_{(1)} = l_1$**



Now, we derive $E[(\Delta T_i - \Delta T_{i'} - \beta)^+ | z_i]$. Again, we must condition on the age, this time of the appropriate service time, since β is a residual service time. To satisfy the conditions defining $C_{(2)}$, β must be less than the minimum of the outstanding residual times excluding the departure at $S(i')$ at time D_i. Thus, analogous to our previous analysis, we choose the following set of conditioning quantities z_i: the accumulated IPA perturbations at servers $S(i)$ and $S(i')$, the age of $C_{i'}$'s service time, and the minimum of all the residual events excluding the departure at $S(i')$, all evaluated at D_i. We use SPA to calculate the conditional expectation

$$E[(\Delta T_i - \Delta T_{i'} - \beta)^+ | \Delta T_i, \Delta T_{i'}, \eta_i, \nu_i, \beta \leq \eta_i].$$

Analogous to our derivation for f_α, it can then be shown that the conditional density for β is given by

$$f_\beta(y) = \begin{cases} 0 & \text{if } \eta_i \leq y, \\ \dfrac{g(\nu_i + y)}{G(\nu_i + \eta_i) - G(\nu_i)} & \text{if } \eta_i > y, \end{cases} \quad (18)$$

where recall that ν_i is the age of the service time at server $S(i')$ at time D_i, and η_i is the minimum of all outstanding residual times (excluding the next departure at $S(i')$) at time D_i. In terms of the IPA accumulations, we have

$$\Delta T_i = (dT_i/d\theta)\Delta\theta \quad \text{and}$$

$$\Delta T_{i'} = (dT_{i'}/d\theta + (dX/d\theta)|_{\nu+\beta})\Delta\theta.$$

Integrating w.r.t. f_β to get the expectation, we have

$$E[(\Delta T_i - \Delta T_{i'} - \beta)^+ | \Delta T_i, \Delta T_{i'}, \eta_i, \nu_i, \beta \le \eta_i]$$
$$= \int_0^\infty \left[\frac{dT_i}{d\theta}\Delta\theta - \frac{dT_{\hat{i}}}{d\theta}\Delta\theta - \frac{dX}{d\theta}\bigg|_{\nu_i+y}\Delta\theta - y \right]^+ \times \frac{g(\nu_i + y)}{G(\nu_i + \eta_i) - G(\nu_i)}\, dy. \quad (19)$$

We define y^* as the point at which the bracketed term in equation (19) becomes zero, given by the implicit equation

$$y^* = \left[\frac{dT_i}{d\theta} - \frac{dT_{\hat{i}}}{d\theta} - \frac{dX}{d\theta}\bigg|_{\nu_i+y^*} \right]^+ \Delta\theta. \quad (20)$$

We assume that y^* is uniquely defined and that $dX/d\theta$ is increasing in the value x taken by X. Then, applying the mean value theorem to $g(\cdot)$ and $dX/d\theta$, again assuming the continuity of each, we have

$$E[(\Delta T_i - \Delta T_{i'} - \beta)^+ | \Delta T_i, \Delta T_{i'}, \eta_i, \nu_i, \beta \le \eta_i]$$
$$= \int_0^{y^*} \left[\frac{dT_i}{d\theta}\Delta\theta - \frac{dT_{\hat{i}}}{d\theta}\Delta\theta - \frac{dX}{d\theta}\bigg|_{\nu_i+y}\Delta\theta - y \right]^+ \times \frac{g(\nu_i + y)}{G(\nu_i + \eta_i) - G(\nu_i)}\, dy$$
$$= \left(\left[\frac{dT_i}{d\theta} - \frac{dT_{\hat{i}}}{d\theta} - \frac{dX}{d\theta}\bigg|_{\nu_i+\zeta} \right]^+ \Delta\theta y^* - \frac{(y^*)^2}{2} \right) \times \frac{g(\nu_i + \zeta)}{G(\nu_i + \eta_i) - G(\nu_i)} \quad \text{for some} \quad \zeta \in [0, y^*]$$
$$= \frac{(\Delta\theta)^2}{2} \left(2\left[\frac{dT_i}{d\theta} - \frac{dT_{\hat{i}}}{d\theta} - \frac{dX}{d\theta}\bigg|_{\nu_i+\zeta} \right]^+ \frac{y^*}{\Delta\theta} - \left(\frac{y^*}{\Delta\theta}\right)^2 \right) \times \frac{g(\nu_i + \zeta)}{G(\nu_i + \eta_i) - G(\nu_i)}. \quad (21)$$

Dividing by $(\Delta\theta)^2/2$ and taking the limit—noting that $\zeta, y^* \to 0$ as $\Delta\theta \to 0$ and using equation (20), we get the SPA2 contribution for $i \in C_{(2)}$:

$$\frac{g(\nu_i)}{G(\nu_i + \eta_i) - G(\nu_i)} \times \left(\left[\frac{dT_i}{d\theta} - \frac{dT_{\hat{i}}}{d\theta} - \frac{dX}{d\theta}\bigg|_{\nu_i} \right]^+ \right)^2 n_{(2),i}. \quad (22)$$

Taking the sample average over N customers gives the SPA2 contribution in equation (8). Here we give the calculations for some simple, but useful, cases.

EXAMPLE 1. *θ a scale parameter.* Then $dX/d\theta = X/\theta$, so we have

$$\left(\left[\frac{dT_i}{d\theta} - \frac{dT_{\hat{i}}}{d\theta} - \frac{dX}{d\theta}\bigg|_{\nu_i} \right]^+ \right)^2 = \left(\frac{l_i - l_{\hat{i}}}{\theta} \right)^2, \quad (23)$$

where l_i and $l_{\hat{i}}$ are the lengths of the local busy periods of servers $S(i)$ and $S(i')$, respectively, at time $D_{i'}$, assuming that $l_i > l_{\hat{i}}$. If $l_i \le l_{\hat{i}}$, then the term is zero.

EXAMPLE 2. *θ a location parameter.* Then $dX/d\theta = 1$, so we have

$$\left(\left[\frac{dT_i}{d\theta} - \frac{dT_{\hat{i}}}{d\theta} - \frac{dX}{d\theta}\bigg|_{\nu_i} \right]^+ \right)^2 = (n_i - n_{\hat{i}})^2, \quad (24)$$

where n_i and $n_{\hat{i}}$ are the number of customers who have completed service in the local busy periods of servers $S(i)$ and $S(i')$, respectively, at time $D_{i'}$, assuming $n_i > n_{\hat{i}}$. Again, if $n_i \le n_{\hat{i}}$, then the term is zero.

4. Special Cases

Here, we consider the special cases of exponential and deterministic service times, for which the estimator simplifies. For exponential service times with mean θ, $dX/d\theta = X/\theta$ and $d^2X/d\theta^2 = 0$, so the IPA contribution is 0. Also, the memoryless property of the exponential distribution implies that the distributions for the set of perturbed sample path quantities that determine $n_{(1),i}$ are unchanged as $\alpha \to 0$, i.e., $n_{(1),i}$ would be unchanged by the substitution of $\{\tau_j(D_i) - \alpha\}_{j\neq 0}$, where $\alpha \neq 0$ from the nominal path; thus, we can estimate the contribution from the nominal path. Since the generated perturbation and the number in the succeeding local busy period are independent for exponential service times, we can lower the variance by estimating the two separately and taking the product, which was the method proposed in the approximation procedure for the general case. In summary, for exponential service times, we have two results: the splitting procedure becomes trivial, in the sense that no additional simulation is required for an exact estimation, and the approximation procedure is exact. As we already said, the latter is preferable, because it has lower variance, but both will give the same limiting value w.p.1.

Furthermore, for the $M/M/m$ queue, we can prove consistency of the estimator by direct comparison with the analytic expressions, using the time-reversibility of the system. This technique was utilized in unbiasedness proofs by Zazanis (1990) and Fu and Hu (1991a) for IPA estimators, but this is the first such proof involving SPA terms. Here, we do the $m = 2$ case. In principle, the calculations could be done for any m, but the calculations become extremely cumbersome as m increases. Taking $\theta = 1/\mu$ to be the mean of the service time distribution and $\lambda = 1/\alpha$ the rate of the Poisson arrival stream, we prove:

THEOREM. *For the $M/M/2$ queue, the SPA estimator for $d^2ET/d\theta^2$ is strongly consistent.*

PROOF. The estimator given by equation (8) without the IPA contribution (since it's zero) can be rewritten as

$$\frac{N_{(1)}}{N}\frac{1}{N_{(1)}}\sum_{i\in C_{(1)}}\left(\frac{d^2T_i}{d\theta^2}\right)_{\text{SPA1}} n_{(1),i} - \frac{N_{(2)}}{N}\frac{1}{N_{(2)}}\sum_{i\in C_{(2)}}\left(\frac{d^2T_i}{d\theta^2}\right)_{\text{SPA2}} n_{(2),i} \quad \text{where}$$

$$\left(\frac{d^2T_i}{d\theta^2}\right)_{\text{SPA1}} = \frac{f(\xi_i)}{F(\xi_i+\tau_i)-F(\xi_i)}\left(\frac{dT_i}{d\theta}\right)^2, \quad (25)$$

$$\left(\frac{d^2T_i}{d\theta^2}\right)_{\text{SPA2}} = \frac{g(\nu_i)}{G(\nu_i+\eta_i)-G(\nu_i)} \times \left(\left[\frac{dT_i}{d\theta} - \frac{dT_{\hat{i}}}{d\theta} - \frac{dX}{d\theta}\bigg|_{\nu_i}\right]^+\right)^2, \quad (26)$$

and where $N_{(k)} = |C_{(k)}|$, $k = 1, 2$, which by the strong law of large numbers converges w.p.1 (as $N \to \infty$) to

$$p_{(1)}E\left[\frac{d^2T}{d\theta^2}\right]_{\text{SPA1}} n_{(1)} - p_{(2)}E\left[\frac{d^2T}{d\theta^2}\right]_{\text{SPA2}} n_{(2)}, \quad (27)$$

where $p_{(k)}$ is the steady-state probability that $i \in C_{(k)}$ for an arriving customer C_i, and where we have assumed that $(d^2T_i/d\theta^2)_{\text{SPA}k}$ converge weakly to some steady-state random variables $(d^2T/d\theta^2)_{\text{SPA}k}$, $k = 1, 2$. Also, by the memoryless property of the exponential distribution, $n_{(k),i}$, $k = 1, 2$, are independent of i (and for $k = 1$, corresponds to the previous definition of $n_{(1)}$ given by equation (5)). The remainder of the proof consists of calculating the quantities and comparing the result to the analytical result. The details of the calculations can be found in Appendix I.

For deterministic service times, where θ is the service time, we have the following simplifications: the IPA contribution is zero; the SPA2 term is zero; and the SPA1 term can be estimated exactly from the nominal sample path. The IPA contribution is zero, because θ is a location parameter, so $d^2X/d\theta^2 = 0$. The reason that the SPA2 term is zero is that for deterministic service times, the adjacent event order change of switching of servers implies the occurrence of a *nonadjacent* event order change, which has expected effect $o((\Delta\theta)^2)$ and hence can be ignored for the purposes of calculating second derivatives. The complete argument is given in Appendix II.

Furthermore, the approximation is also exact for the $GI/D/m$ queue, because for deterministic service times, FCFS implies FIFO (first-in, first-out), so intuitively, the $GI/D/m$ is decomposable into m (nonindependent, since the arrival processes are correlated) $G/D/1$ queues, for which a Lindley-like equation exists and for which the nominal sample path can be used directly to estimate second derivatives. In terms of the sample path, for deterministic service times, taking $\alpha \to 0$ in Figure 1 will not change the shape of the sample path, in the sense that all subsequent ordering of departures and arrivals would be unchanged under such a limit, so that the number of customers in the succeeding local busy period would be the same as in the nominal path.

Note that since the SPA2 contribution is zero, the estimator is always positive, which implies convexity of mean system time with respect to the service time if the estimator is in fact consistent, which is not proven here. Actual proofs of convexity by sample path means can be found in Harel (1990) and also in Fu and Hu (1993). In the latter, the convexity is then used to prove the unbiasedness of the first derivative IPA estimator.

5. Simulation Examples

This section contains simulation results for a number of examples for the purpose of validating the estimator, investigating the efficiency of the splitting procedure, and empirically testing the accuracy of the approximation procedure. Where analytical results are available, they are used for comparison purposes. Otherwise, the

simulation results are compared with two types of symmetric finite differences calculated via common random numbers—a second-order estimate derived from estimates of T itself (denoted by $\hat{T}$), and a hybrid estimate derived from the IPA first derivative estimates (denoted by T'_{IPA}):

$$\left(\frac{d^2ET}{d\theta^2}\right)_{\text{SD}} = \frac{\hat{T}(\theta + \Delta\theta, \omega) - 2\hat{T}(\theta, \omega) + \hat{T}(\theta - \Delta\theta, \omega)}{(\Delta\theta)^2},$$

$$\left(\frac{d^2ET}{d\theta^2}\right)_{\text{SD/IPA}} = \frac{T'_{\text{IPA}}(\theta + \Delta\theta, \omega) - T'_{\text{IPA}}(\theta - \Delta\theta, \omega)}{2\Delta\theta}.$$

For the cases where differences are calculated, the table entries also include all the PA derivative estimates at the $\theta - \Delta\theta$ and $\theta + \Delta\theta$ values, indicated in the table headings for ρ by "−" and "+", respectively. In addition, 95% paired-t confidence intervals are calculated for the difference between the SPA estimate and the SD/IPA estimate to give an indication of any bias possibly introduced by the approximation procedure. The other entries are given in the form mean ± standard deviation, taken over 40 independent replications. Without loss of generality, the arrival rate was fixed at 1, i.e., $\lambda = 1$. For the service time distributions which were neither exponential nor deterministic, coefficients of variation are also provided. Except where noted, three different cases were simulated, corresponding to light traffic ($\rho = 0.2$), medium traffic ($\rho = 0.5$), and heavy traffic ($\rho = 0.8$).

EXAMPLE 1. $M/M/2$. An $M/M/2$ queue was simulated, with mean interarrival time $1/\lambda$ and mean service time θ. For Poisson arrivals, we have

$$f(z)/(F(z + \tau) - F(z)) = \lambda/(1 - e^{-\lambda\tau}).$$

The experimental results are given in Table 1, along with the analytical values, and show good agreement.

EXAMPLE 2. $M/D/2$. An $M/D/2$ queue was simulated, with mean interarrival time $1/\lambda$. The experimental results are given in Table 2, along with the analytical values, which are calculated by differentiating the formulas given in Syski (1962). Again, good agreement is indicated.

EXAMPLE 3. $U/D/2$. A $U/D/2$ queue was simulated, with interarrival times uniform on $[0, 2/\lambda]$, for which we have

$$\frac{f(z)}{F(z + \tau) - F(z)} = \begin{cases} 1/\tau & \text{if } z \in [0, 2/\lambda - \tau], \\ 1/(2/\lambda - z) & \text{if } z \in [2/\lambda - \tau, 2/\lambda]. \end{cases}$$

For the $U/D/2$, no analytic results exist with which to compare the numerical results, so finite differences were used, where $\Delta\theta = 0.025$, 0.05, and 0.005, for $\rho = 0.2$, 0.5, and 0.8, respectively. The experimental results are given in Table 3, along with both types of difference estimates. From the 95% confidence intervals for the difference, no bias is indicated, as should be expected.

EXAMPLE 4. $U/M/2$. A $U/M/2$ queue was simulated, with interarrival times uniform on $[0, 2/\lambda]$ and mean service time θ. The experimental results are given in Table 4, along with the analytical values, and again excellent agreement is demonstrated.

For the first four examples, the estimator should be exact, and experimental results seem to validate this. The next example shows that splitting, though theoretically a solution, is not a practical alternative except at low traffic intensities, which leads us to investigate the quality of the approximation procedure in the remaining set of examples. $\Delta\theta = 0.01$ was used for all of the finite difference calculations in the remaining examples, unlike in Example 3, where $\Delta\theta$ was larger for smaller ρ. In these examples, the effect of this relatively small choice on the second-order estimates is often evident in the large variances in the SD estimate. However, it should be noted that the choice of $\Delta\theta$ in using finite differences to estimate derivatives is always a difficult one a priori, this difficulty constituting another principal advantage of perturbation analysis, where no such choice has to be made. On the other hand, it is very interesting to note that the SD/IPA estimates show reasonable variance, using the same choice of $\Delta\theta$.

EXAMPLE 5. $M/U/2$. An $M/U/2$ queue was simulated, with mean interarrival time $1/\lambda$ and mean service time θ (and width 2θ, giving a coefficient of variation of $\frac{1}{3}$). In this case, the contributions for the SPA1 term cannot be estimated directly from the nominal sample path, so either the splitting procedure or the approximation procedure (i.e., ignoring the fact that the nominal path does not estimate the correct quantity exactly)

Table 1 **Estimates for the *M/M/2* Queue (1,000,000 Busy Periods)**

ρ	$(ET)_{est}$	ET	$(dET/d\theta)_{est}$	$dET/d\theta$	$(d^2ET/d\theta^2)_{est}$	$d^2ET/d\theta^2$
0.2	0.417 ± 0.0005	0.417	1.129 ± 0.002	1.128	0.689 ± 0.018	0.687
0.5	1.333 ± 0.002	1.333	2.223 ± 0.007	2.222	3.850 ± 0.055	3.852
0.8	4.439 ± 0.011	4.444	12.61 ± 0.07	12.65	61.9 ± 1.2	62.4

Table 2 **Estimates for the *M/D/2* Queue (100,000 Busy Periods)**

ρ	$(ET)_{est}$	ET	$(dET/d\theta)_{est}$	$dET/d\theta$	$(d^2ET/d\theta^2)_{est}$	$d^2ET/d\theta^2$
0.2	0.410 ± 0.0001	0.410	1.072 ± 0.001	1.072	0.368 ± 0.009	0.369
0.5	1.177 ± 0.001	1.177	1.630 ± 0.006	1.630	1.928 ± 0.045	1.937
0.8	3.040 ± 0.011	3.045	6.79 ± 0.07	6.85	30.4 ± 1.0	31.2

Table 3 **Estimates for the *U/D/2* Queue (100,000 Busy Periods)**

	ρ	PA	SD		
	0.2−	1.019 ± 0.0005	—		
	0.2	1.021 ± 0.0006	1.021 ± 0.0006		
	0.2+	1.024 ± 0.0006	—		
	0.5−	1.165 ± 0.002	—		
$dET/d\theta$	0.5	1.193 ± 0.003	1.194 ± 0.003		
	0.5+	1.226 ± 0.003	—		
	0.8−	2.847 ± 0.020	—		
	0.8	2.900 ± 0.020	2.905 ± 0.068		
	0.8+	2.952 ± 0.020	—	SD/IPA	95% difference
	0.2−	0.103 ± 0.005	—	—	—
	0.2	0.111 ± 0.004	0.111 ± 0.030	0.111 ± 0.005	0.00001 ± 0.0018
	0.2+	0.122 ± 0.004	—	—	—
	0.5−	0.513 ± 0.009	—	—	—
$d^2ET/d\theta^2$	0.5	0.606 ± 0.015	0.602 ± 0.091	0.610 ± 0.019	0.0049 ± 0.0070
	0.5+	0.710 ± 0.015	—	—	—
	0.8−	10.03 ± 0.23	—	—	—
	0.8	10.35 ± 0.24	8.08 ± 24.08	10.47 ± 0.56	0.1157 ± 0.1750
	0.8+	10.72 ± 0.29	—	—	—

Table 4 **Estimates for the *U/M/2* Queue (1,000,000 Busy Periods)**

ρ	$(ET)_{est}$	ET	$(dET/d\theta)_{est}$	$dET/d\theta$	$(d^2ET/d\theta^2)_{est}$	$d^2ET/d\theta^2$
0.2	0.406 ± 0.0004	0.406	1.048 ± 0.001	1.048	0.300 ± 0.010	0.302
0.5	1.165 ± 0.001	1.164	1.680 ± 0.004	1.679	2.425 ± 0.032	2.421
0.8	3.337 ± 0.008	3.334	8.59 ± 0.05	8.56	41.5 ± 1.5	41.5

must be used. In order to have a measure of efficiency, we counted the amount of extra simulation required for the splitting procedure. As we pointed out earlier, the number of samples taken in the splitting procedure is arbitrary, with more samples requiring more "outside" simulation. We ran the simplest version of taking just one split sample every time. Again, we planned to simulate three different cases corresponding to light, medium, and heavy traffic. However, only for the $\rho = 0.2$ case was the insertion procedure practical. In this case, the average amount of additional simulation required was about 46% of the original simulation length, where simulation length is measured in number of customers served. For the higher traffic intensity cases, the additional amount of simulation length required exceeded by orders of magnitude the original simulation length. To get an idea of just how many orders of magnitude, for the $\rho = 0.5$ we did a set of "short" runlength replications (100 busy periods), for which the amount of additional simulation averaged about 27 times the original runlength, ranging from less than three times the original length for one replication to an extreme case of well over 4,000 times the original length in another replication. The significant increase in the amount of additional simulation required is due to a combination of there being more occurrences of departures satisfying the $C_{(1)}$ conditions (so the *number* of subpaths generated is increased) and the fact that at higher values of ρ, the value of $n_{(1),i}$ increases substantially (so the simulation effort devoted to generating *each* individual subpath is greatly increased). Thus, it appears that the insertion procedure is a very limited option, applicable only in low traffic situations. The remainder of the simulation results presented in this section primarily concern the approximation procedure, with results for the splitting procedure estimate provided in some examples for the $\rho = 0.2$ cases. For the $M/U/2$, the experimental results for the approximation procedure estimate are given in Table 5, along with the finite difference calculations. For the $\rho = 0.2$ case, the "exact" estimate using additional simulations generated by the splitting procedure yields 0.483 ± 0.006, while the approximation yields 0.481 ± 0.006. Thus, it seems that the approximation matches the insertion answer very closely for this case, and the 95% confidence interval for the difference indicates no clear bias at $\rho = 0.2$. At the higher values of ρ, the confidence intervals fail to cover 0, so that there is a conclusive difference (at the asymptotically valid 95% confidence level), but of an order which is less than 5% of the

Table 5 **Estimates for the *M/U/2* Queue (1,000,000 Busy Periods)**

	ρ	PA	SD		
	0.2−	1.089 ± 0.0007	—		
	0.2	1.093 ± 0.0008	1.093 ± 0.0015		
	0.2+	1.098 ± 0.0007	—		
	0.5−	1.810 ± 0.003	—		
$dET/d\theta$	0.5	1.836 ± 0.003	1.837 ± 0.008		
	0.5+	1.862 ± 0.003	—		
	0.8−	8.340 ± 0.029	—		
	0.8	8.737 ± 0.032	8.755 ± 0.073		
	0.8+	9.166 ± 0.036	—	SD/IPA	95% difference
	0.2−	0.469 ± 0.007	—	—	—
	0.2	0.481 ± 0.006	0.976 ± 3.228	0.485 ± 0.008	0.0027 ± 0.0031
	0.2+	0.498 ± 0.007	—	—	—
	0.5−	2.463 ± 0.022	—	—	—
$d^2ET/d\theta^2$	0.5	2.544 ± 0.021	1.966 ± 0.133	2.582 ± 0.037	0.039 ± 0.013
	0.5+	2.620 ± 0.022	—	—	—
	0.8−	37.20 ± 0.70	—	—	—
	0.8	40.00 ± 0.79	40.34 ± 16.03	41.29 ± 0.56	1.30 ± 0.27
	0.8+	42.95 ± 0.67	—	—	—

estimate value. Thus, the approximation seems to do quite well. (It should be kept in mind that the SD/IPA estimator is also biased in general, since it estimates a difference.)

EXAMPLE 6. $M/T/2$. An $M/T/2$ (where "T" indicates a symmetric triangular density) queue was simulated, with mean interarrival time $1/\lambda$ and mean service time θ (and width 2θ, giving a coefficient of variation of $\frac{1}{6}$). For the $\rho = 0.2$ case, the splitting procedure estimate, which again required an average of 46% additional simulation, is 0.428 ± 0.004, compared with the approximate procedure estimate of 0.425 ± 0.004. The experimental results for the approximate procedure are given in Table 6, along with the finite difference calculations. The agreement is in accordance with the previous example, with no discernible difference at $\rho = 0.2$ and no more than 5% difference at the higher values of ρ.

EXAMPLE 7. $U/U/2$. A $U/U/2$ queue was simulated, with mean interarrival time $1/\lambda$ (and width $1/(2\lambda)$), and mean service time θ (and width 2θ, giving a coefficient of variation of $\frac{1}{3}$). For the $\rho = 0.2$ case, the splitting procedure estimate, which required an average of 26% additional simulation, is 0.162 ± 0.003, the same as the value of the approximate procedure estimate (to that precision). The experimental results for the approximate procedure are given in Table 7, along with the finite difference calculations. The agreement is similar to the trend shown in the previous two examples.

EXAMPLE 8. $T/T/2$. A $T/T/2$ queue was simulated, with mean interarrival time $1/\lambda$ (and width $1/(2\lambda)$), and mean service time θ (and width 2θ, giving a coefficient of variation of $\frac{1}{6}$). For the $\rho = 0.2$ case, the splitting procedure estimate, which required an average of 11% additional simulation, is 0.0225 ± 0.0007, compared with the approximate procedure estimate of 0.0224 ± 0.0007. The experimental results for the approximate procedure are given in Table 8, along with the finite difference calculations. The agreement is again similar to the previous examples, though percentage-wise the agreement is slightly worse than the previous three examples, due to the lower absolute magnitudes of the estimates themselves.

EXAMPLE 9. $M/We/2$. An $M/We/2$ queue was simulated, where "We" represents a Weibull distribution given by

$$F(x) = \begin{cases} 1 - \exp(-x^\theta/\beta) & \text{for } x > 0, \\ 0 & \text{otherwise,} \end{cases}$$

where we again take θ as our parameter of interest, but

Table 6 Estimates for the *M/T/2* Queue (1,000,000 Busy Periods)

	ρ	PA	SD		
	0.2−	1.078 ± 0.0006	—		
	0.2	1.082 ± 0.0006	1.083 ± 0.0012		
	0.2+	1.087 ± 0.0006	—		
	0.5−	1.711 ± 0.002	—		
$dET/d\theta$	0.5	1.733 ± 0.002	1.735 ± 0.005		
	0.5+	1.757 ± 0.002	—		
	0.8−	7.419 ± 0.024	—		
	0.8	7.765 ± 0.026	7.792 ± 0.067		
	0.8+	8.140 ± 0.026	—	SD/IPA	95% difference
	0.2−	0.412 ± 0.005	—	—	—
	0.2	0.425 ± 0.004	1.003 ± 2.618	0.427 ± 0.007	0.00001 ± 0.0018
	0.2+	0.437 ± 0.005	—	—	—
	0.5−	2.148 ± 0.020	—	—	—
$d^2ET/d\theta^2$	0.5	2.216 ± 0.021	2.079 ± 0.962	2.274 ± 0.031	0.058 ± 0.010
	0.5+	2.290 ± 0.021	—	—	—
	0.8−	32.26 ± 0.51	—	—	—
	0.8	34.85 ± 0.74	35.51 ± 12.49	36.05 ± 0.46	1.20 ± 0.26
	0.8+	37.50 ± 0.50	—	—	—

Table 7 **Estimates for the *U/U/2* Queue (1,000,000 Busy Periods)**

	ρ	PA	SD		
	0.2−	1.027 ± 0.0007	—		
	0.2	1.029 ± 0.0007	1.029 ± 0.0013		
	0.2+	1.030 ± 0.0007	—		
	0.5−	1.321 ± 0.002	—		
$dET/d\theta$	0.5	1.333 ± 0.002	1.332 ± 0.005		
	0.5+	1.345 ± 0.002	—		
	0.8−	4.553 ± 0.014	—		
	0.8	4.752 ± 0.016	4.756 ± 0.037		
	0.8+	4.967 ± 0.017	—	SD/IPA	95% difference
	0.2−	0.155 ± 0.003	—	—	—
	0.2	0.162 ± 0.003	0.134 ± 2.469	0.162 ± 0.006	−0.00026 ± 0.0018
	0.2+	0.166 ± 0.002	—	—	—
	0.5−	1.112 ± 0.009	—	—	—
$d^2ET/d\theta^2$	0.5	1.152 ± 0.014	1.092 ± 1.117	1.182 ± 0.019	0.030 ± 0.006
	0.5+	1.195 ± 0.011	—	—	—
	0.8−	18.12 ± 0.21	—	—	—
	0.8	19.63 ± 0.42	21.20 ± 5.75	20.71 ± 0.32	1.08 ± 0.14
	0.8+	21.01 ± 0.28	—	—	—

this time it is neither a scale nor a location parameter. Thus, unlike all the previous examples where the IPA term was zero, all three terms are nonzero here. Two different Weibull distributions were considered: $\theta = 0.5$ and $\theta = 2.0$, which give coefficients of variations of $\sqrt{5} \approx 2.236 > 1$ and $\sqrt{4/\pi - 1} \approx 0.5227 < 1$,

Table 8 **Estimates for the *T/T/2* Queue (1,000,000 Busy Periods)**

	ρ	PA	SD		
	0.2−	1.002 ± 0.0005	—		
	0.2	1.002 ± 0.0004	1.002 ± 0.0008		
	0.2+	1.002 ± 0.0005	—		
	0.5−	1.107 ± 0.007	—		
$dET/d\theta$	0.5	1.112 ± 0.0007	1.112 ± 0.0022		
	0.5+	1.118 ± 0.0007	—		
	0.8−	2.697 ± 0.005	—		
	0.8	2.796 ± 0.005	2.801 ± 0.013		
	0.8+	2.904 ± 0.006	—	SD/IPA	95% difference
	0.2−	0.0208 ± 0.0008	—	—	—
	0.2	0.0224 ± 0.0007	0.458 ± 1.536	0.0226 ± 0.0025	0.00010 ± 0.00082
	0.2+	0.0245 ± 0.0009	—	—	—
	0.5−	0.494 ± 0.003	—	—	—
$d^2ET/d\theta^2$	0.5	0.514 ± 0.003	0.537 ± 0.564	0.526 ± 0.008	0.012 ± 0.002
	0.5+	0.535 ± 0.004	—	—	—
	0.8−	8.82 ± 0.06	—	—	—
	0.8	9.47 ± 0.07	10.19 ± 2.46	10.36 ± 0.11	0.88 ± 0.04
	0.8+	10.24 ± 0.06	—	—	—

Table 9 **Estimates for the *M/We/2* Queue: $\theta = 0.5$ (1,000,000 Busy Periods)**

	ρ	PA	SD		
	0.2−	−0.473 ± 0.010	—		
	0.2	−0.406 ± 0.009	−0.407 ± 0.010		
	0.2+	−0.347 ± 0.009	—		
	0.5−	−17.02 ± 0.23	—		
$dET/d\theta$	0.5	−14.10 ± 0.19	−14.23 ± 0.19		
	0.5+	−11.83 ± 0.15	—		
	0.8−	−342.1 ± 4.3	—		
	0.8	−223.3 ± 3.0	−231.4 ± 2.5		
	0.8+	−155.0 ± 1.9	—	SD/IPA	95% difference
	0.2−	7.44 ± 1.08	—	—	—
	0.2	6.42 ± 1.27	6.30 ± 0.71	6.30 ± 0.14	−0.12 ± 0.40
	0.2+	5.48 ± 1.13	—	—	—
	0.5−	339.0 ± 14.5	—	—	—
$d^2ET/d\theta^2$	0.5	260.9 ± 14.7	264.1 ± 21.3	259.9 ± 5.0	−1.0 ± 4.7
	0.5+	204.6 ± 15.7	—	—	—
	0.8−	16639. ± 1161.	—	—	—
	0.8	9072. ± 881.	9045. ± 461.	9353. ± 163.	280. ± 281.
	0.8+	5494. ± 234.	—	—	—

respectively. As before, light, medium, and heavy traffic cases were simulated for both values of θ, this time obtained by varying the value of β appropriately (as opposed to the previous cases, where θ was varied). The experimental results are given in Tables 9 and 10, along with the finite difference calculations. For the $\theta = 2.0$

Table 10 **Estimates for the *M/We/2* Queue: $\theta = 2.0$ (1,000,000 Busy Periods)**

	ρ	PA	SD		
	0.2−	0.1681 ± 0.00006	—		
	0.2	0.1675 ± 0.00006	0.1674 ± 0.0006		
	0.2+	0.1668 ± 0.00006	—		
	0.5−	−0.1671 ± 0.0006	—		
$dET/d\theta$	0.5	−0.1638 ± 0.0006	−0.1640 ± 0.0015		
	0.5+	−0.1606 ± 0.0006	—		
	0.8−	−4.607 ± 0.020	—		
	0.8	−4.429 ± 0.019	−4.438 ± 0.043		
	0.8+	−4.261 ± 0.015	—	SD/IPA	95% difference
	0.2−	−0.0691 ± 0.0002	—	—	—
	0.2	−0.0688 ± 0.0001	−0.0650 ± 0.117	−0.0686 ± 0.0004	0.00021 ± 0.00014
	0.2+	−0.0685 ± 0.0001	—	—	—
	0.5−	0.332 ± 0.0015	—	—	—
$d^2ET/d\theta^2$	0.5	0.324 ± 0.0015	0.338 ± 0.329	0.326 ± 0.0036	0.0024 ± 0.0011
	0.5+	0.316 ± 0.0015	—	—	—
	0.8−	18.00 ± 0.17	—	—	—
	0.8	17.00 ± 0.17	16.54 ± 9.32	17.35 ± 0.21	0.35 ± 0.08
	0.8+	16.05 ± 0.15	—	—	—

Table 11 Estimates for *GI/G/5* Queues (100,000 Busy Periods)

		ρ	PA	SD	SD/IPA	95% difference
M/U/5		0.5−	1.1996 ± 0.0016	—		
	$dET/d\theta$	0.5	1.2032 ± 0.0016	1.2052 ± 0.0120		
		0.5+	1.2067 ± 0.0016	—		
		0.5−	0.3623 ± 0.0073	—	—	—
	$d^2ET/d\theta^2$	0.5	0.3686 ± 0.0082	−0.1545 ± 2.3982	0.3578 ± 0.0168	−0.0109 ± 0.0058
		0.5+	0.3734 ± 0.0084	—	—	—
U/M/5		0.5−	1.1191 ± 0.0019	—		
	$dET/d\theta$	0.5	1.1218 ± 0.0020	1.1237 ± 0.0259		
		0.5+	1.1246 ± 0.0020	—		
		0.5−	0.2917 ± 0.0080	—	—	—
	$d^2ET/d\theta^2$	0.5	0.2953 ± 0.0088	1.3388 ± 4.5725	0.2755 ± 0.0257	−0.0198 ± 0.0079
		0.5+	0.3030 ± 0.0110	—	—	—
U/U/5		0.5−	1.0581 ± 0.0008	—		
	$dET/d\theta$	0.5	1.0594 ± 0.0008	1.0582 ± 0.0103		
		0.5+	1.0607 ± 0.0008	—		
		0.5−	0.1371 ± 0.0023	—	—	—
	$d^2ET/d\theta^2$	0.5	0.1393 ± 0.0023	−0.4960 ± 1.8739	0.1334 ± 0.0081	−0.0060 ± 0.0025
		0.5+	0.1429 ± 0.0026	—	—	—

case, the same conclusions as in the previous examples can be drawn, while for the $\theta = 0.5$ case, no conclusion on the differences can be drawn due to the high variances of all the estimates. It appears that for higher values of the coefficient of variation the variance of the PA estimate degenerates faster than the finite difference estimates.

EXAMPLE 10. $G/G/5$. To see if there might be any dramatic decrease in the performance of the estimator for more than two servers, $U/M/5$, $M/U/5$, and $U/U/5$ queues were simulated at $\rho = 0.5$ using the approximation procedure. The experimental results are given in Table 11. The performance of the estimates are in line with the previous $m = 2$ examples (errors of about 3%, 7%, and 4%, respectively), which seems to indicate that the approximation does not degrade with higher m. Notice how poorly the second-order finite difference estimate does (in contrast with how well the first-order IPA-based finite difference does).

6. Conclusions and Extensions

We have derived a sample path estimator for the second derivative of mean system time in a $GI/G/m$ queue. The estimator consists of an IPA contribution and two SPA contributions which are of opposite signs. The derivation reveals many interesting issues that arise in applying SPA to estimating derivatives. A principal difficulty in applying SPA to complex systems is the ability to compute the conditional expectation easily from the nominal sample path; in terms of simulation, this means without the need for additional simulation. This problem was originally called the "propagation problem" in Gong and Ho (1987), who at that time conjectured that the nominal path could always be used to directly "propagate" the generated perturbation. For our second derivative estimator, however, this propagation could not be carried out for the SPA1 contribution, requiring in general the generation of additional sample subpaths. This splitting technique appears to be practical only for light traffic systems. In lieu of this procedure, an approximation procedure requiring no additional simulation was proposed (the phase-type distribution option is discussed in the next paragraph), with simulation results indicating that the approximation procedure is reasonably accurate. Further research on deriving some kind of bounds on the accuracy of the approximation procedure would be very useful.

As we have pointed out, the potential applications of second derivative estimation include use in speeding up the convergence of gradient-based stochastic optimization algorithms and in response surface fitting. One natural extension of the work related to these applications is to derive estimators with respect to other parameters, i.e., for interarrival time parameters; this should not be too difficult, given the framework set up here. Another avenue of further research worth investigating is the use of phase-type distributions to model general nonexponential distributions, instead of using the approximation procedure. The main reason the splitting technique is infeasible for general distributions is because there is a continuum of possible states **R** defining $n_{(1),i}$. For exponential service times, the states defining $n_{(1),i}$ are equivalent in distribution by the memoryless property, reducing the possibility to a single case, which occurs naturally on the simulated sample path. If phase-type distributions are considered, the continuum is again reduced to finitely many initial states, and by defining the states in the simulation property, i.e., by keeping track of the phases, the corresponding $n_{(1),i}$ can be estimated exactly from the original sample path without the need for additional simulation. Since this procedure would be exact and substantially more efficient than the general splitting case, it would be fruitful to investigate the performance of estimators incorporating these types of distributions.[1]

[1] This work was supported in part by the National Science Foundation under Grants Nos. ECS85-15449 and CDR-8803012, by the Office of Naval Research under Contracts Nos. N00014-90-K-1093 and N00014-89-J-1023, and by the Army under Contract No. DAAL-03-83-K-1071. The authors are grateful to the departmental editor, the associate editor, and the anonymous referees whose comments helped improve this paper.

Appendix I: Completion of Proof for the $M/M/2$ Queue

We consider a customer C_i who arrives to the system in steady state—by which we mean that upon arrival, C_i sees the steady-state distribution of the system (which we can assume since we have Poisson arrivals). Let $\lambda = 1/\alpha$ be the arrival rate, $\mu = 1/\theta$ be the service rate, $\rho = \lambda\theta/m$.

For $m = 2$, we have $ET = \theta/(1-\rho^2)$, where $\rho = \lambda\theta/2$. Differentiating twice w.r.t. θ, we get

$$\frac{d^2ET}{d\theta^2} = \frac{\lambda\rho(3+\rho^2)}{(1-\rho^2)^3}. \tag{28}$$

We proceed to show that equation (27), the w.p.1 limiting (as $N \to \infty$) value of our estimator, yields the same result.

The probability $p_{(1)}$ is defined by $C_{(1)}$: a departure bringing the number in system to $(m-1)$ and the subsequent event being an arrival. First, for this Markovian system, the latter is conditionally (on the number in system) independent of the former. Second, for an $M/M/m$ system, the departure process is also Poisson. Thus, a departure in the reversed process is a Poisson arrival which sees time averages. Third, the probability that the next event is an arrival, given there are $(m-1)$ in the system, is simply $\lambda/(\lambda+(m-1)\mu)$. Putting this all together, we have

$$p_{(1)} = p_{m-1}\frac{\lambda}{\lambda+(m-1)\mu}, \tag{29}$$

where p_n denotes the stationary probability for number in system.

Analogously, for the $p_{(2)}$, we have

$$p_{(2)} = p_m\frac{(m-1)\mu}{\lambda+m\mu}. \tag{30}$$

The SPA1 term given by equation (15) is comprised of two parts. For exponential interarrivals, the first becomes

$$\frac{f(\xi)}{F(\xi+\tau)-F(\xi)} = \frac{\lambda}{1-e^{-\lambda\tau}}.$$

Since the service times are exponential, the memoryless property means that this part is independent of the second part, which becomes $(T_{(1)}/\theta)^2$, where $T_{(1)}$ = length of the local busy period at the time of C_i's departure. This follows from equation (6), since $dX/d\theta = X/\theta$ for exponential service times (see Fu and Hu 1991a for details). Thus,

$$E\left[\frac{d^2T}{d\theta^2}\right]_{\text{SPA1}} = E\left[\frac{\lambda}{1-e^{-\lambda\tau}}\right]\frac{E[T_{(1)}^2]}{\theta^2}. \tag{31}$$

Similarly, for the SPA2 term, we have

$$E\left[\frac{d^2T}{d\theta^2}\right]_{\text{SPA2}} = E\left[\frac{\mu}{1-e^{-\mu\eta}}\right]\frac{E[T_{(2)}^2]}{2\theta^2}, \tag{32}$$

where $T_{(2)}$ = difference between the lengths of the local busy period at which C_i is served and the local busy period at which the next departure will take place, at the time of C_i's departure. The factor of $\frac{1}{2}$ arises from the fact that $C_{(1)}$ is in the longer local busy period w.p. $\frac{1}{2}$, in which case the term is $(T_{(2)}/\theta)^2$; otherwise, the term is 0. This follows from equation (23), since θ is a scale parameter.

We now specialize to $m = 2$, for which we have

$$p_n = 2\rho^n p_0, \quad p_0 = (1-\rho)/(1+\rho), \tag{33}$$

and calculate the other terms in equation (27). We first consider C_i, $i \in C_{(1)}$, in order to calculate $E[\lambda/(1-e^{-\lambda\tau})]$. Since there are just two servers, τ is simply the single residual service time active at D_i, under the condition that the residual service time is greater than the next interarrival time. For the memoryless exponential distribution, τ is simply the r.v. $(X \mid X > A)$, which has (conditional) p.d.f.

$$f_\tau(x) = \frac{d}{dx} P(\tau \le x) = \frac{d}{dx} P(X \le x | X > A)$$

$$= \frac{d}{dx} \frac{P(X \le x | X > A)}{P(X > A)} = \frac{(d/dx)P(A < X \le x)}{\lambda/(\lambda+\mu)}$$

$$= \frac{\lambda+\mu}{\lambda} \frac{d}{dx} \int_0^x P(X \le x, X > y) g(y) dy$$

$$= \frac{\lambda+\mu}{\lambda} \int_0^x \frac{\partial}{\partial x} [e^{-\mu y} - e^{-\mu x}] g(y) dy$$

$$= \frac{\lambda+\mu}{\lambda} \mu e^{-\mu x} \int_0^x g(y) dy = \frac{\lambda+\mu}{\lambda} \mu e^{-\mu x}(1 - e^{-\lambda x}).$$

Thus,

$$E\left[\frac{\lambda}{1-e^{-\lambda\tau}}\right] = \int_0^\infty \frac{\lambda}{1-e^{-\lambda x}} \frac{\lambda+\mu}{\lambda} \mu e^{-\mu x}(1 - e^{-\lambda x}) dx$$

$$= \lambda + \mu. \quad (34)$$

Next, we consider C_i, $i \in C_{(2)}$, who departs from server $S(i) = S$ (see Figure 2), in order to calculate $E[\mu/(1 - e^{-\mu\eta})]$. For the two-server case, η is the minimum of the residual interarrival time and the full (since a departure just occurred) service time at server S, under the condition that this minimum is greater than the residual service time at server S', where the next departure takes place. For the memoryless exponential distribution, η is simply the r.v. $(\min(X_A, A) | \min(X_A, A) > X_B)$, where X_A and X_B are i.i.d. exponential service time random variables. Let $Z = \min(X_A, A)$. Then, Z is an exponential random variable with rate $\lambda + \mu$, so following the same derivation as above, η has (conditional) p.d.f.

$$f_\eta(x) = \frac{d}{dx} P(\eta \le x) = \frac{d}{dx} P(Z \le x | Z > X)$$

$$= \frac{d}{dx} \frac{P(Z \le x | Z > X)}{P(Z > X)} = \frac{(d/dx)P(X < Z \le x)}{\mu/(\lambda+2\mu)}$$

$$= \frac{\lambda+2\mu}{\mu} \frac{d}{dx} \int_0^x P(Z \le x, Z > y) f(y) dy$$

$$= \frac{\lambda+2\mu}{\mu} \int_0^x \frac{\partial}{\partial x} [e^{-(\lambda+\mu)y} - e^{-(\lambda+\mu)x}] f(y) dy$$

$$= \frac{\lambda+2\mu}{\mu} (\lambda+\mu) e^{-(\lambda+\mu)x} \int_0^x f(y) dy$$

$$= \frac{\lambda+2\mu}{\mu} (\lambda+\mu) e^{-(\lambda+\mu)x}(1 - e^{-\mu x}).$$

Thus,

$$E\left[\frac{\mu}{1-e^{\mu\eta}}\right] = \int_0^\infty \frac{\mu}{1-e^{-\mu x}} \frac{\lambda+2\mu}{\mu} (\lambda+\mu) e^{-(\lambda+\mu)x}(1 - e^{-\mu x}) dx$$

$$= \lambda + 2\mu. \quad (35)$$

To calculate $E[T_{(1)}^2]$, $E[T_{(2)}^2]$, $n_{(1)}$, $n_{(2)}$, we define the following r.v.'s (see also Figures 6 and 7):

l_1 = time until end of local busy period, given that there are 2 in the system,

n_1 = number of customers served until end of local busy period, given that there are 2 in the system,

l_2 = time until end of local busy period, given that there is 1 in the system,

n_2 = number of customers served until end of local busy period, given that there is 1 in the system.

We have $n_{(1)} = E[n_1]$ and $n_{(2)} = E[n_2]$. Using the memorylessness property and reversibility of the system, we note that $T_{(1)}$ and l_1 are equal in distribution (as are $T_{(2)}$ and l_2), since $T_{(1)}$ in the reversed path is the time from 2 in the system to the end of the local busy period, i.e., l_1 in the forward path (and $T_{(2)}$ in the reversed path is the time from 1 in the system to the end of the local busy period, i.e., l_2 in the forward path). Furthermore, $n_{(1)}$ and $n_{(2)}$ are the corresponding number of customers served during the times of l_1 and l_2, respectively.

To derive the desired quantities of interest, we use the following relationships:

$$l_1 = Y_1 + \begin{cases} 0 & \text{w.p. } \mu/(\lambda+2\mu), \\ l_2 & \text{w.p. } \mu/(\lambda+2\mu), \\ \tilde{B} + l_1 & \text{w.p. } \lambda/(\lambda+2\mu), \end{cases}$$

$$l_2 = Y_2 + \begin{cases} 0 & \text{w.p. } \mu/(\lambda+\mu), \\ l_1 & \text{w.p. } \lambda/(\lambda+\mu), \end{cases} \quad \text{where}$$

$$Y_1 \sim \exp(\lambda+2\mu), \quad Y_2 \sim \exp(\lambda+\mu),$$

$\tilde{B}$ = time to go from 3 to 2 in the system.

The relationships follow easily by considering the possible events that can occur in the given situation. Since the system is Markovian, the time to occurrence of one of these events is an exponential random variable with rate equal to the sum of the rates of the possible events, while the probability that a particular event occurs is given by the ratio of the individual rate to the summed rate. For l_1, either a departure at server A, a departure at server B, or an arrival can occur. In the first case, the local busy period has ended, so there is no additional contribution. In the second case, we come to the situation which defines l_2, since there is now 1 in the system. In the third case, we now have three in the system. After adding the time for the first return to 2 from 3, we are back where we started, i.e., the situation that defines l_1. Similarly, for l_2, either a departure (at server A, since server B is idle) or an arrival can occur. The former ends the local busy period, while the latter brings the number in system to 2 and hence to the situation that defines l_1.

$\tilde{B}$ has the distribution of the busy period of an $M/M/1$ queue with arrival rate λ and service rate 2μ (mean service time $\theta/2$), so in particular its first two moments are given by (see, e.g., Kleinrock 1975)

$$E[\tilde{B}] = \frac{\theta/2}{1-\rho}, \quad E[\tilde{B}^2] = \frac{\theta^2/2}{(1-\rho)^3},$$

where $\rho = \lambda\theta/2$.

Furthermore, $E[l_i] = \theta n_{(i)}$, $i = 1, 2$, so our task reduces to calculating the first two moments of l_1 and l_2. Taking expectations of the l_i's, we have

$$E[l_1] = \frac{1}{\lambda + 2\mu} + E[l_2]\frac{\mu}{\lambda + 2\mu} + (E[\tilde{B}] + E[l_1])\frac{\lambda}{\lambda + 2\mu},$$

$$E[l_2] = \frac{1}{\lambda + \mu} + E[l_1]\frac{\lambda}{\lambda + \mu}.$$

Solving the set of simultaneous equations and dividing by θ, we get

$$n_{(1)} = \frac{1 + \rho/2}{1 - \rho^2}, \tag{36}$$

$$n_{(2)} = \frac{1}{1 - \rho^2}. \tag{37}$$

Taking expectations of the l_i^2's and noting the mutual independence of $\tilde{B}$, l_1, and Y_1, and the independence of l_1 and Y_2, we have

$$\begin{aligned}
E[l_1^2] &= E[Y_1^2]\frac{\mu}{\lambda + 2\mu} + E[Y_1 + l_2]^2\frac{\mu}{\lambda + 2\mu} \\
&\quad + E[Y_1 + \tilde{B} + l_1]^2\frac{\lambda}{\lambda + 2\mu} \\
&= E[Y_1^2] + (E[l_2^2] + 2E[l_2Y_1])\frac{\mu}{\lambda + 2\mu} \\
&\quad + (E[l_1^2] + E[\tilde{B}^2] + 2E[\tilde{B}l_1] + 2E[\tilde{B}Y_1] + 2E[l_1Y_1])\frac{\lambda}{\lambda + 2\mu} \\
&= \frac{2}{(\lambda + 2\mu)^2} + \left(E[l_2^2] + \frac{2\mu}{\lambda + 2\mu}\frac{\theta}{1 - \rho^2}\right)\frac{\mu}{\lambda + 2\mu} \\
&\quad + \left(E[l_1^2] + \frac{\theta^2}{2(1 - \rho)^3} + \frac{\theta}{1 - \rho}\frac{\theta(1 + \rho/2)}{1 - \rho^2} + \frac{\theta}{1 - \rho}\frac{1}{\lambda + 2\mu}\right. \\
&\quad \left. + \frac{2\theta(1 + \rho/2)}{1 - \rho^2}\frac{1}{\lambda + 2\mu}\right)\frac{\lambda}{\lambda + 2\mu} \\
&= \left[\mu E[l_2^2] + \lambda E[l_1^2] + \frac{2\theta}{(1 + \rho)(1 - \rho)^3}\right]\frac{1}{\lambda + 2\mu}, \quad \text{so} \\
E[l_1^2] &= \frac{1}{2}E[l_2^2] + \frac{\theta^2}{(1 + \rho)(1 - \rho)^3}, \quad \text{and}
\end{aligned}$$

$$\begin{aligned}
E[l_2^2] &= E[Y_2^2]\frac{\mu}{\lambda + \mu} + E[Y_2 + l_1]^2\frac{\lambda}{\lambda + \mu} \\
&= E[Y_2^2] + (E[l_1^2] + 2E[l_1Y_2])\frac{\lambda}{\lambda + \mu} \\
&= \frac{\lambda + \mu}{(\lambda + 2\mu)^2} + \left(E[l_1^2] + \frac{2\theta(1 + \rho/2)}{1 - \rho^2}\frac{1}{\lambda + \mu}\right)\frac{\lambda}{\lambda + \mu} \\
&= \frac{2\rho}{1 + 2\rho}E[l_1^2] + \frac{2\theta^2}{(1 + 2\rho)(1 - \rho^2)}.
\end{aligned}$$

Solving the set of simultaneous equations, we get

$$E[T_{(1)}^2] = E[l_1^2] = \frac{\theta^2(2 + \rho^2)}{(1 + \rho)^2(1 - \rho)^3}, \tag{38}$$

$$E[T_{(2)}^2] = E[l_2^2] = \frac{2\theta^2(1 - \rho + \rho^2)}{(1 + \rho)^2(1 - \rho)^3}. \tag{39}$$

Substituting equations (29)–(39) into equation (27), we get

$$\begin{aligned}
E\left[\frac{d^2T}{d\theta^2}\right] &= (\lambda + \mu)\frac{2 + \rho^2}{(1 + \rho)^2(1 - \rho)^3}\frac{1 + \rho/2}{1 - \rho^2}\frac{2\rho(1 - \rho)}{1 + \rho}\frac{\lambda}{\lambda + \mu} \\
&\quad - (\lambda + 2\mu)\frac{1 - \rho + \rho^2}{(1 + \rho)^2(1 - \rho)^3}\frac{\frac{1}{2}}{1 - \rho^2}\frac{2\rho^2(1 - \rho)}{1 + \rho}\frac{\mu}{\lambda + 2\mu} \\
&= \frac{\lambda\rho(3 + \rho^2)}{(1 - \rho^2)^3},
\end{aligned}$$

matching equation (28) and completing the proof.

Appendix II: Completion of Argument for the $G/D/m$ Queue

Here, we complete the argument as to why the SPA2 term is zero for the $G/D/m$ queue, the gist of the argument being that for deterministic service times, the phenomenon of switching between two servers due to a perturbation in the sample path implies the occurrence of a *nonadjacent* event order change elsewhere in the sample path. To make the argument clearer, we show in Figure 8 what must happen in a perturbed sample path that experiences a switching of server. We assume of course that the service completion events occurring at D_A and D_B are adjacent events. If these two adjacent events incur an order change in the perturbed path, then our argument will show that this will imply a nonadjacent event order change earlier in the sample path; in fact, it must be of the type shown in Figure 8—a service completion overtaking both another service completion and an arrival to the system, indicated by $\tilde{D}_A$ overtaking both $\tilde{D}_B$ and $\tilde{A}$ in the figure. As in the last section, for a general $GI/G/m$ queue, an event order change causing switching of customers between servers occurs when

$$\Delta T_A > \Delta T_B + (D_B - D_A), \tag{40}$$

which of course implies $\Delta T_A > \Delta T_B$. But for deterministic services ΔT is simply equal to $n\,\Delta\theta$, where n is the number of customers preceding the given customer in the local busy period. So, referring to Figure 8,

Figure 7 **In Reversed Sample Path, $T_{(2)} = l_2$**

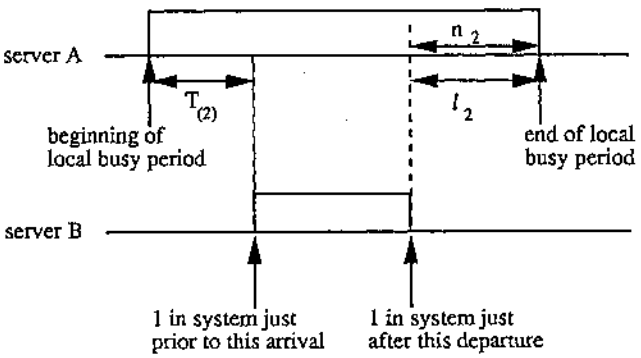
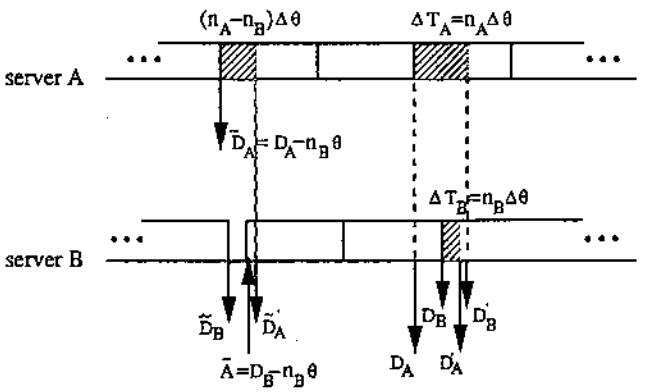


Figure 8 **Switching Implies Nonadjacent Event Order Change Elsewhere**



we have $\Delta T_A = n_A\Delta\theta$ and $\Delta T_B = n_B\Delta\theta$. Since $\Delta T_A > \Delta T_B$, we know that $n_A > n_B$, so equation (40) becomes

$$(n_A - n_B)\Delta\theta > D_B - D_A. \tag{41}$$

Since n_A and n_B are integers, we can write $n_A \geq n_B + 1$, and again using the fact that the service times are deterministic, this means that the local busy period at server A must have started at least one service time earlier than that at server B. This in turn implies that another service completion at server B must have occurred in the interval $[D_A - n_B\theta, D_B - n_B\theta]$. The point $(D_B - n_B\theta)$ is the beginning of the local busy period at server B, denoted by $\tilde{A}$ in our figure. At this point, since it is the beginning of the "local" busy period, there is no perturbation, while the perturbation at server A at point $(D_A - n_B\theta)$ is $(n_A - n_B)\Delta\theta$. We now claim that the events at these two points, $(D_B - n_B\theta)$ and $(D_A - n_B\theta)$, incur an event order change in the perturbed path. Since another service completion occurs between these two events, this would indicate a nonadjacent event order change, and hence our argument would be complete. An event order change occurs in the perturbed path if

$$D_A - n_B\theta + (n_A - n_B)\Delta\theta > D_B - n_B\theta,$$

which follows from equation (41) above. The example shown in Figure 7 is for $n_B = 2$ and $n_A \geq 3$, with the service completion at $\tilde{D}_B$ occurring in $[\tilde{D}_A', \tilde{A}]$, where we have denoted the start of B's "local" busy period by $\tilde{A}$. $\tilde{D}_A'$ shows the perturbation causing a nonadjacent event order change, overtaking both $\tilde{D}_B$ and $\tilde{A}$.

References

Fu, M. C. and Y. C. Ho, "Using Perturbation Analysis for Gradient Estimation, Averaging, and Updating in a Stochastic Approximation Algorithm," M. Abrams, P. Haigh and J. Comfort (Eds.), *Proc. Winter Simulation Conf.*, 1988, 509–517.

Fu, M. C. and J. Q. Hu, "Consistency of Infinitesimal Perturbation Analysis for the $GI/G/m$ Queue," *European J. Oper. Res.*, 54 (1991a), 121–139.

—— and ——, "On Choosing the Characterization for Smoothed Perturbation Analysis," *IEEE Trans. Automatic Control*, AC-36 (1991b), 1331–1336.

—— and ——, "Extensions and Generalizations of Smoothed Perturbation Analysis in a Generalized Semi-Markov Process Framework," *IEEE Trans. Automatic Control*, 37 (1992), 1483–1500.

—— and ——, "Sample Path Properties of the $G/D/m$ Queue," *European J. Oper. Res.*, 65, 2 (1993), 270–273.

Glasserman, P., "Structural Conditions for Perturbation Analysis of Queueing Systems," *J. Association Computing Machinery*, 38, 4 (1991), 1005–1025.

—— and W. B. Gong, "Smoothed Perturbation Analysis for a Class of Discrete Event Systems," *IEEE Trans. Automatic Control*, AC-35 (1990), 1218–1230.

Gong, W. B. and Y. C. Ho, "Smoothed Perturbation Analysis of Discrete-Event Dynamic Systems," *IEEE Trans. Automatic Control*, AC-32 (1987), 858–867.

Harel, A., "Convexity Results for Single Server Queues and for Multiserver Queues with Constant Service Times," *J. Appl. Prob.*, 27, 2 (1990), 465–468.

Ho, Y. C. and X. R. Cao, *Discrete Event Dynamic Systems and Perturbation Analysis*, Kluwer Academic, Boston, 1991.

Kleinrock, L., *Queueing Systems: Volume 1*, John Wiley & Sons, New York, 1975.

Reiman, M. I., B. Simon and J. S. Willie, "Siminterpolations: Estimating an Entire Queueing Function From a Single Sample Path," A. Thesen, H. Grant, and W. David Kelton (Eds.), *Proc. Winter Simulation Conf.*, 1987, 358–363.

Suri, R. and M. Zazanis, "Perturbation Analysis Gives Strongly Consistent Sensitivity Estimates for the $M/G/1$ Queue," *Management Sci.*, 34 (1988), 39–64.

Syski, R., *Introduction to Congestion Theory in Telephone Systems*, Oliver and Boyd, London, 1962.

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

Zazanis, M., "Infinitesimal Perturbation Analysis Estimates for Moments of the System Time of an $M/M/1$ Queue," *Oper. Res.*, 38 (1990), 364–369.

Zazanis, M. and R. Suri, "Perturbation Analysis for the $GI/G/1$ Queue," Technical Report (revised version), IE/MS Department, Northwestern University, Evanston, IL, 1989.

Accepted by James R. Wilson; received May 18, 1990. This paper has been with the authors 4 months for 4 revisions.