MANAGEMENT SCIENCE
Vol. 34, No. 11, November 1988
Printed in U.S.A.

CONVERGENCE PROPERTIES OF INFINITESIMAL PERTURBATION ANALYSIS ESTIMATES*

PHILIP HEIDELBERGER, XI-REN CAO, MICHAEL A. ZAZANIS AND RAJAN SURI
IBM Research Division, Thomas J. Watson Research Center, P.O. Box 704, Yorktown Heights, New York 10598
Digital Equipment Corporation, 200 Forest Street, Marlboro, Massachusetts 01752
Department of Industrial Engineering, Northwestern University, Evanston, Illinois 60201
Department of Industrial Engineering, University of Wisconsin, Madison, Wisconsin 53706

Infinitesimal Perturbation Analysis (IPA) is a method for computing a sample path derivative with respect to an input parameter in a discrete event simulation. The IPA algorithm is based on the fact that for certain parameters and any realization of a simulation, the change in parameter can be made small enough so that only the times of events get shifted, but their order does not change. This paper considers the convergence properties of the IPA sample path derivatives. In particular, the question of when an IPA estimate converges to the derivative of a steady state performance measure is studied. Necessary and sufficient conditions for this convergence are derived for a class of regenerative processes. Although these conditions are not guaranteed to be satisfied in general, they are satisfied for the mean stationary response time in the $M/G/1$ queue. A necessary condition for multiple IPA estimates to simultaneously converge to the derivatives of steady state throughputs in a queueing network is determined. The implications of this necessary condition are that, except in special cases, the original IPA algorithm cannot be used to consistently estimate steady state throughput derivatives in queueing networks with multiple types of customers, state-dependent routing or blocking. Numerical studies on IPA convergence properties are also presented.

(SIMULATION; SENSITIVITY ANALYSIS; PERTURBATION ANALYSIS; QUEUES; DISCRETE EVENT SYSTEMS)

1. Introduction

Infinitesimal Perturbation Analysis (IPA) is a technique for calculating a sample path derivative with respect to an input parameter in a discrete event simulation (see Ho, Eyler and Chien 1983, Cao and Ho 1983, Ho and Cao 1983, Ho, Cao and Cassandras 1983, Ho, Suri, Cao, Diehl, Dille and Zazanis 1984, Suri and Zazanis to appear, Cao 1987a, 1988, Suri 1983 and 1987, Zazanis and Suri 1985a, b, Cao and Ho 1986 and Zazanis 1986). A closely related algorithm for calculating sample path derivatives in certain queueing networks has also been described in Woodside (1984). For example, in a queueing system simulation we might be interested in estimating the mean response time and its derivative with respect to the mean service time. The primary assumption of IPA is that if the change in the input parameter is small enough, then the times at which events occur get shifted slightly, but their order does not change. The IPA algorithm shows how a very small change in a parameter generates event time "perturbations" and how the perturbation associated with one event affects the times of subsequent events. Since events do not change in order, the effect of these perturbations can be tracked efficiently during the simulation run, thereby obtaining the sample path derivative with only moderate overhead. A technique called First Order Perturbation Analysis, which is in general an approximation, has been proposed to estimate the effect of a finite change in the value of an input parameter (see, for example, Ho, Cao and Cassandras 1983). An alternative approach to estimating derivatives based on likelihood ratios and the variance reduction technique of importance sampling (see, for example, Hammersley

[1] Accepted by George S. Fishman, former Departmental Editor; received May 22, 1986. This paper has been with the authors 6 months for 3 revisions.

0025-1909/88/3411/1281$01.25

and Handscomb 1964 and Halton 1970) is described in Glynn (1986), Glynn and Sanders (1986), Reiman and Weiss (1986) and Rubinstein (1986). This paper will only consider the Infinitesimal Perturbation Analysis algorithm as it was originally proposed.

For example, consider a single server queue and suppose that the service time of the nth customer is $S_n = \theta X_n$ where X_n is a random variable with mean one and θ is the parameter of interest (θ is the mean service time). Here, for simplicity, we have assumed θ to be a scale parameter of the service time distribution, although more general parameters can be handled (see Suri and Zazanis 1988). If θ changes by $d\theta$ then S_n changes by $d\theta X_n$. The first customer's departure time changes by $d\theta X_1$. Similarly, the nth customer's departure time changes by $d\theta X_1 + \cdots + d\theta X_n$ provided customer n is served in the first busy period. If there are τ customers served in the first busy period, and if $d\theta$ is small enough, then there are τ customers served in the first busy period of the system when the parameter is $\theta + d\theta$. In this case, the departure time of customer $\tau + 1$ changes only by $d\theta X_{\tau+1}$, i.e., the departure times of the customers served in the second busy period are unaffected by the change in departure times of customers served in the first busy period. Fixing a positive $d\theta$ is never required since we only need keep track of terms of the form $\sum_{n=1}^{\tau} \sum_{k=1}^{n} X_k$.

The key questions associated with IPA concern the statistical properties of these sample path derivatives, in particular their convergence properties. This paper addresses the question of when an IPA sample path derivative converges to the derivative of a steady state quantity. Let $r(\theta)$ denote the steady state quantity and let $\hat{r}(\theta, t)$ be an estimate of $r(\theta)$ after a simulation of length t. Assume that $\lim_{t\to\infty} \hat{r}(\theta, t) = r(\theta)$ with probability one. Suppose that we are interested in estimating the derivative of $r(\theta)$, $r'(\theta) = dr(\theta)/d\theta$. Let the IPA sample path derivative be

$$\hat{r}'(\theta, t) = \frac{d}{d\theta}\hat{r}(\theta, t) = \lim_{d\theta\to 0} [\hat{r}(\theta + d\theta, t) - \hat{r}(\theta, t)]/d\theta.$$

The key question concerning IPA estimates (see also Cao and Ho 1987) is whether or not they are strongly consistent, i.e., does

$$\frac{d}{d\theta} r(\theta) = \frac{d}{d\theta} \lim_{t\to\infty} \hat{r}(\theta, t) \overset{?}{=} \lim_{t\to\infty} \frac{d}{d\theta} \hat{r}(\theta, t)$$

with probability one? This is the classical question of when the order of two limits can be interchanged.

The IPA algorithm is derived by observing that for any simulation realization, or sample path ω, and for any finite run length t, there exists a $\delta(\omega, t) > 0$ such that if $|d\theta| < \delta(\omega, t)$, then the order of events does not change and only the time shifts need to be considered in estimating the derivative. However, given a fixed $d\theta$, for some sample paths ω, $|d\theta| < \delta(\omega, t)$ in which case the order of events does not change, but for other sample paths ω, $|d\theta| \geq \delta(\omega, t)$ in which case the order of events does change. In particular, as $t \to \infty$, the probability should approach one that, for a fixed $d\theta$, the order of events changes.

This situation has been recognized in the IPA literature, but at the same time, IPA has been shown to give strongly consistent estimates for:

1. The derivative of the mean stationary response time in the $M/G/1$ queue (Suri and Zazanis 1988) and the first two derivatives of the mean stationary response time in certain $GI/G/1$ queues (Zazanis and Suri 1985a). Estimating the second derivative requires an extension to the IPA algorithm.

2. The derivative of the higher moments of the stationary response time in the $M/M/1$ queue (Zazanis 1986).

3. The derivative of the stationary throughput with respect to a mean service time in a closed Jackson network with a single type of customer (Cao 1987c). Cao (1988) and

Cao and Ho (1986) have also shown that IPA produces unbiased estimates of the derivatives of the transient throughput and transient mean response time in a closed Jackson network with a single type of customer. Experimental evidence in Cao and Ho (1983) suggests that IPA may produce strongly consistent estimates in open Jackson networks with a single type of customer.

For a fixed t, intuitively, if the probability that the order of events changes is $o(d\theta)$, then the interchange in order between expectation and limit should be justified and IPA will yield an unbiased estimate of the derivative. However, in the above mentioned cases the probability that events change in order is $O(d\theta)$, which is of the same order as the IPA approximation $\hat{r}(\theta + d\theta, t) \simeq \hat{r}(\theta, t) + d\theta\hat{r}'(\theta, t)$, and yet IPA produces consistent estimates! For example, the straightforward IPA estimate of the derivative of the number of customers served in a busy period is identically zero whereas the derivative of the expected number of customers served in a busy period is not zero. Similarly, in a finite state space continuous time Markov chain, such as a closed Jackson network, with generator matrix $\mathbf{Q} = (q_{ij})$, $dP_{ik}/dq_{ij} \neq 0$ where P_{ik} is the probability transition matrix of the embedded Markov chain. That is, if $\theta = q_{ij}$, then, with probability $O(d\theta)$, the next state in the chain with parameter $\theta + d\theta$ is different from the next state in the chain with parameter θ.

In this paper, by examining the IPA convergence question in detail, we will explain this apparent paradox and more clearly identify the domain of applicability of IPA.

It should be noted that the original IPA algorithm constitutes but a portion of the body of research on Perturbation Analysis. In particular, several extensions of the IPA algorithm have been developed, some heuristic and others with an analytical foundation. These extensions increase the domain of applicability of Perturbation Analysis, and specific references will be given at appropriate points in the paper.

In §2 necessary and sufficient conditions are given for IPA to produce strongly consistent estimates in a class of regenerative processes. It is then shown that the IPA estimates are strongly consistent in the $M/G/1$ queue not because the probability that events change in order during a regenerative cycle is $o(d\theta)$ (which it isn't), but rather because certain limiting expectations (as $d\theta \to 0$) that are not estimated by IPA cancel each other out. These expectations exactly measure the limiting effect of a change in event order. An intuitive explanation for this cancellation is provided. A heuristic argument is then given that shows why IPA works for closed Jackson networks. This heuristic argument clearly identifies an assumption of customer homogeneity that is required by IPA.

In §3, necessary conditions are derived that must be satisfied if the original IPA algorithm is to simultaneously produce strongly consistent estimates of throughput (event rate) derivatives in essentially arbitrary discrete event stochastic systems. If $\lambda_1(\theta)$ and $\lambda_2(\theta)$ are steady state throughputs and if IPA produces strongly consistent estimates for $\lambda_1'(\theta)$ and $\lambda_2'(\theta)$ for all θ in some interval, then there must exist a constant c such that

$$\lambda_1(\theta)/\lambda_2(\theta) = c \tag{1.1}$$

for all θ in the interval. Among the implications of this result are that, except in special cases, IPA cannot in general be used to consistently estimate throughput derivatives in queueing networks with multiple types of customers, state-dependent routing or blocking due to finite buffers. References will be given to several different extensions of IPA that can alleviate these problems in certain situations. These extensions either provide exact results (i.e., strongly consistent derivative estimates) in specific simple queueing systems or provide approximate, but potentially accurate, results (i.e., single run finite difference approximations to the derivatives) in more general systems.

The results of §§2 and 3 complement, extend, clarify and synthesize other results on IPA convergence. For example, Cao (1985) discusses how an IPA estimate will typically be biased if the sample output estimate is a discontinuous function of the input parameter

θ. Lemma 2.1 and Theorem 2.2 essentially extend this result to regenerative processes and clarify a result of Zazanis and Suri (1985b). The examples cited above and in §3 all exhibit such discontinuities. Similarly, Theorem 3.1 (equation (1.1) above) extends a result that was obtained by Cao (1987a).

In §4, empirical results on IPA convergence properties are presented that further demonstrate the requirement for customer homogeneity. One example of particular interest is a simple closed product form queueing network with three queues and two types of customers (see Baskett, Chandy, Muntz and Palacios 1975). In this example IPA produces exact, i.e., strongly consistent, estimates of both λ_1' and λ_2' with respect to the mean service time at queue 1 and an inexact, but fairly accurate, estimate of λ_1' with respect to the mean service time at queue 2. However, the IPA estimate of λ_2' with respect to the mean service time at queue 2 has the wrong sign. The IPA extension described in Ho and Li (1988) does produce accurate finite difference approximations to the throughput derivatives in this example.

In §5, an IPA-like algorithm is presented for estimating derivatives in Birth and Death processes. The analysis of this algorithm clearly shows the difficulty created by assuming that small changes in parameters affect only the times of events and not their order or relative number. For Birth and Death processes, empirical studies in Glasserman (1988) indicate that this difficulty can also be alleviated by applying IPA to a different representation, i.e., different event generation method, of the process, although this alternative representation does not necessarily lead to strongly consistent derivative estimates in general continuous time Markov chains.

Finally, §6 summarizes the results of this paper.

2. IPA Regenerative Estimates

Let $\mathbf{X}_\theta = \{X_s(\theta), s \geq 0\}$ be a regenerative process depending on a parameter θ with stationary, or steady state, random variable $X(\theta)$, i.e., $X_s(\theta)$ converges in distribution to $X(\theta)$. Assume that we are interested in estimating a steady state quantity, $r(\theta) = \mathrm{E}[f(X(\theta))]$ and its derivative $r'(\theta) = dr(\theta)/d\theta$. Then under broadly applicable conditions (see, e.g., Crane and Iglehart 1975)

$$r(\theta) = \frac{\mathrm{E}[Y_i(\theta)]}{\mathrm{E}[\tau_i(\theta)]} = \frac{y(\theta)}{t(\theta)} \tag{2.1}$$

where $\tau_i(\theta)$ is the length of the ith regenerative cycle (given parameter θ), $Y_i(\theta) = \int_{T_{i-1}(\theta)}^{T_i(\theta)} f(X_s(\theta))ds$ and $T_i(\theta)$ is the time at which the ith regeneration occurs ($T_0(\theta) \equiv 0$). For processes in discrete time the integral in the definition of $Y_i(\theta)$ is replaced by a sum. Note that even if $X_s(\theta)$ does not converge in distribution to a random variable $X(\theta)$ because of periodicities, $r(\theta)$ is still a meaningful quantity to estimate since $\lim_{t\to\infty} (1/t) \int_0^t f(X_s(\theta))ds = r(\theta)$ with probability one provided $\mathrm{E}[Y_i(\theta)]$ and $\mathrm{E}[\tau_i(\theta)]$ exist and are finite. Differentiating equation (2.1) we obtain

$$r'(\theta) = \frac{y'(\theta)t(\theta) - y(\theta)t'(\theta)}{t(\theta)^2} \tag{2.2}$$

provided all the derivatives exist and are finite. Assume that the IPA sample path derivatives

$$Y_i'(\theta) = \frac{d}{d\theta} Y_i(\theta) \qquad \text{and} \qquad \tau_i'(\theta) = \frac{d}{d\theta} \tau_i(\theta)$$

exist and are finite. Let $\bar{Y}_n(\theta)$, $\bar{\tau}_n(\theta)$, $\bar{Y}_n'(\theta)$ and $\bar{\tau}_n'(\theta)$ denote the sample averages of $\{Y_i(\theta)\}$, $\{\tau_i(\theta)\}$, $\{Y_i'(\theta)\}$ and $\{\tau_i'(\theta)\}$, respectively, after n regenerative cycles. For example, $\bar{Y}_n'(\theta) = \sum_{i=1}^n Y_i'(\theta)/n$. The IPA estimate $\hat{r}_n'(\theta)$ of $r'(\theta)$ is defined to be

$$\hat{r}'_n(\theta) = \frac{\bar{Y}'_n(\theta)\bar{\tau}_n(\theta) - \bar{Y}_n(\theta)\bar{\tau}'_n(\theta)}{\bar{\tau}_n(\theta)^2}. \tag{2.3}$$

Letting $n \to \infty$ we obtain the following lemma (see also Zazanis and Suri 1985b):

LEMMA 2.1. *In a regenerative process, if $Y'_i(\theta)$ and $\tau'_i(\theta)$ exist and if* $\mathrm{E}[Y_i(\theta)]$, $\mathrm{E}[\tau_i(\theta)]$, $\mathrm{E}[Y'_i(\theta)]$ *and* $\mathrm{E}[\tau'_i(\theta)]$ *all exist and are finite, then, with probability one,*

$$\lim_{n\to\infty} \hat{r}'_n(\theta) = \frac{\mathrm{E}[Y'_i(\theta)]\mathrm{E}[\tau_i(\theta)] - \mathrm{E}[Y_i(\theta)]\mathrm{E}[\tau'_i(\theta)]}{\mathrm{E}[\tau_i(\theta)]^2}$$

$$= \frac{\mathrm{E}[Y'_i(\theta)]t(\theta) - y(\theta)\mathrm{E}[\tau'_i(\theta)]}{t(\theta)^2}. \tag{2.4}$$

As an example consider the mean response time in the $GI/G/1$ queue. In this case $\tau_i(\theta)$ is the number of customers served in the ith busy period. For small enough $d\theta$, the number of customers served in the busy period does not change so that $\tau'_i(\theta) = 0$. Therefore $\hat{r}'_n(\theta) = \bar{Y}'_n(\theta)/\bar{\tau}_n(\theta)$ and $\lim_{n\to\infty} \hat{r}'_n(\theta) = \mathrm{E}[Y'_i(\theta)]/t(\theta)$. Zazanis and Suri (1985a) have shown that for the $GI/G/1$ queue $r'(\theta) = \mathrm{E}[Y'_i(\theta)]/t(\theta)$ so that the IPA estimate converges to the derivative of the ratio $r(\theta) = y(\theta)/t(\theta)$ even though $y'(\theta) \neq \mathrm{E}[Y'_i(\theta)]$ and $t'(\theta) \neq \mathrm{E}[\tau'_i(\theta)]$.

In Suri (1987) a class of stochastic processes, parameters and estimators is formally defined for which sample path derivatives exist. It is shown that, for a simulation of length n events in this class, there exists a $\delta(\omega, n) > 0$ such that if $|d\theta| < \delta(\omega, n)$, then the order of events in the simulation with parameter $\theta + d\theta$ is the same as in the simulation with parameter θ. Furthermore, if $T_n(\theta)$ is the time of the nth event and $Y(\theta, n) = \int_0^{T_n(\theta)} f(X_s(\theta))ds$, then for $|d\theta| < \delta(\omega, n)$,

$$Y(\theta + d\theta, n) = \int_0^{T_n(\theta+d\theta)} f(X_s(\theta + d\theta))ds = Y(\theta, n) + d\theta Y'(\theta, n) + o(d\theta).$$

We will assume throughout that the process $\{X_s(\theta),\ s \geq 0\}$ has piecewise constant sample paths.

Applying these results to regenerative processes yields the following lemma which gives an expression for $y'(\theta)$ in terms of the expected value of the IPA sample derivative $\mathrm{E}[Y'_i(\theta)]$. A similar representation, under a different set of technical assumptions, has been presented in Cao (1985) for the case of independent replications.

LEMMA 2.2. *Assume that there exists a random variable $\delta_i(\omega)$ that is positive with probability one such that if $|d\theta| < \delta_i(\omega)$, then the number and order of events in the ith regenerative cycle with parameter $\theta + d\theta$ are the same as in the ith cycle with parameter θ. Further assume that if $|d\theta| < \delta_i(\omega)$, then $Y_i(\theta + d\theta) = Y_i(\theta) + d\theta Y'_i(\theta) + o(d\theta)Z_i(\theta)$ where the term represented by $o(d\theta)$ does not depend on ω. Let $p(d\theta) = \mathrm{P}\{\delta_i(\omega) \leq |d\theta|\}$. If*

1. $\mathrm{E}[|Z_i(\theta)|] < \infty$,
2. $p_\theta = \lim_{d\theta\to 0} p(d\theta)/d\theta < \infty$,
3. *$y'(\theta)$ exists and is finite, and*
4. $\mathrm{E}[|Y'_i(\theta)|] < \infty$,

then

$$y'(\theta) = \mathrm{E}[Y'_i(\theta)] + p_\theta d_\theta(Y) \tag{2.5}$$

where $d_\theta(Y) = \lim_{d\theta\to 0} \mathrm{E}[Y_i(\theta + d\theta) - Y_i(\theta) \mid \delta_i(\omega) \leq |d\theta|]$.

The probability $p(d\theta) = \mathrm{P}\{\delta_i(\omega) \le |d\theta|\}$ is the probability that there is a change in the order of events. Note that $\lim_{d\theta \to 0} p(d\theta) = 0$, but that $\lim_{d\theta\to 0} p(d\theta)/d\theta$ may not exist in general. If this limit does exist, it is not in general equal to zero. In the Appendix, it is shown that $\lim_{d\theta\to 0} p(d\theta)/d\theta$ exists and is nonzero in the $M/G/1$ queue. An extension to the argument in the Introduction shows that this limit exists for n state transitions in a finite state space continuous time Markov chain (and thus the limit will exist for an entire regenerative cycle under suitable regularity conditions on the transition rate matrix).

PROOF.

$$\begin{aligned} \mathrm{E}[Y_i(\theta + d\theta)] = \mathrm{E}[Y_i(\theta + d\theta) \,|\, \delta_i(\omega) > |d\theta|](1 - p(d\theta)) \\ + \mathrm{E}[Y_i(\theta + d\theta) \,|\, \delta_i(\omega) \le |d\theta|]p(d\theta). \end{aligned}$$

Therefore

$$\begin{aligned} \mathrm{E}[Y_i(\theta + d\theta)] = \mathrm{E}[Y_i(\theta) + d\theta Y_i'(\theta) + o(d\theta)Z_i(\theta) \,|\, \delta_i(\omega) > |d\theta|](1 - p(d\theta)) \\ + \mathrm{E}[Y_i(\theta + d\theta) \,|\, \delta_i(\omega) \le |d\theta|]p(d\theta). \quad (2.6) \end{aligned}$$

Subtracting $\mathrm{E}[Y_i(\theta)] = \mathrm{E}[Y_i(\theta) \,|\, \delta_i(\omega) > |d\theta|](1 - p(d\theta)) + \mathrm{E}[Y_i(\theta) \,|\, \delta_i(\omega) \le |d\theta|]p(d\theta)$ from both sides of equation (2.6) yields

$$\begin{aligned} \mathrm{E}[Y_i(\theta + d\theta)] - \mathrm{E}[Y_i(\theta)] = \mathrm{E}[d\theta Y_i'(\theta) + o(d\theta)Z_i(\theta) \,|\, \delta_i(\omega) > |d\theta|](1 - p(d\theta)) \\ + \mathrm{E}[Y_i(\theta + d\theta) - Y_i(\theta) \,|\, \delta_i(\omega) \le |d\theta|]p(d\theta). \quad (2.7) \end{aligned}$$

Dividing equation (2.7) by $d\theta$ and letting $d\theta \to 0$ yields the result provided that

$$\lim_{d\theta\to 0} \mathrm{E}[Y_i'(\theta) \,|\, \delta_i(\omega) > |d\theta|] = \mathrm{E}[Y_i'(\theta)] \qquad \text{and}$$

$$\lim_{d\theta\to 0} \mathrm{E}[(o(d\theta)/d\theta)Z_i(\theta) \,|\, \delta_i(\omega) > |d\theta|] = 0.$$

Let $I_i(d\theta)$ be the indicator of the event $\{\delta_i(\omega) > |d\theta|\}$. Then $\mathrm{E}[Y_i'(\theta) \,|\, \delta_i(\omega) > |d\theta|] = \mathrm{E}[Y_i'(\theta)I_i(d\theta)]/\mathrm{E}[I_i(d\theta)]$. Furthermore $\lim_{d\theta\to 0} I_i(d\theta) = 1$ and $\lim_{d\theta\to 0} \mathrm{E}[I_i(d\theta)] = 1 - \lim_{d\theta\to 0} p(d\theta) = 1$ and $\mathrm{E}[|Y_i'(\theta)I_i(d\theta)|] \le \mathrm{E}[|Y_i'(\theta)|]$. Thus by the dominated convergence theorem $\lim_{d\theta\to 0} \mathrm{E}[Y_i'(\theta) \,|\, \delta_i(\omega) > |d\theta|] = \mathrm{E}[Y_i'(\theta)]$. A similar argument shows that

$$\begin{aligned} \lim_{d\theta\to 0} |\mathrm{E}[(o(d\theta)/d\theta)Z_i(\theta) \,|\, \delta_i(\omega) > |d\theta|]| \\ \le \lim_{d\theta\to 0} |o(d\theta)/d\theta| \mathrm{E}[|Z_i(\theta)|]/(1 - p(d\theta)) = 0. \end{aligned}$$

Combining Lemmas 2.1 and 2.2 yields the following theorem which gives necessary and sufficient conditions for IPA to produce strongly consistent estimates for $r'(\theta)$.

THEOREM 2.1. *If the conditions of Lemmas* 2.1 *and* 2.2 *are satisfied for both* $\mathrm{E}[Y_i(\theta)]$ *and* $\mathrm{E}[\tau_i(\theta)]$, *then* $\lim_{n\to\infty} \hat{r}_n'(\theta) = r'(\theta)$ *with probability one if and only if* $p_\theta = 0$ *or*

$$d_\theta(Y)\mathrm{E}[\tau_i(\theta)] - \mathrm{E}[Y_i(\theta)]d_\theta(\tau) = 0. \quad (2.8)$$

PROOF. Lemma 2.2 will be applied to both $y'(\theta)$ and $t'(\theta)$. From Lemma 2.2, the definition of $\delta_i(\omega)$ depends only on the underlying structure of the stochastic process and not upon the particular functionals $Y_i(\theta)$ and $\tau_i(\theta)$. Thus the p_θ in Lemma 2.2 is the same for both $y'(\theta)$ and $t'(\theta)$. From equations (2.2) and (2.5)

$$r'(\theta) = \frac{(\mathrm{E}[Y_i'(\theta)] + p_\theta d_\theta(Y))\mathrm{E}[\tau_i(\theta)] - \mathrm{E}[Y_i(\theta)](\mathrm{E}[\tau_i'(\theta)] + p_\theta d_\theta(\tau))}{\mathrm{E}[\tau_i(\theta)]^2}. \quad (2.9)$$

Therefore

$$r'(\theta) = \frac{\mathrm{E}[Y_i'(\theta)]\mathrm{E}[\tau_i(\theta)] - \mathrm{E}[\tau_i'(\theta)]\mathrm{E}[Y_i(\theta)]}{\mathrm{E}[\tau_i(\theta)]^2} + p_\theta \frac{d_\theta(Y)\mathrm{E}[\tau_i(\theta)] - \mathrm{E}[Y_i(\theta)]d_\theta(\tau)}{\mathrm{E}[\tau_i(\theta)]^2}. \quad (2.10)$$

By Lemma 2.1, the IPA estimate $\hat{r}_n'(\theta)$ converges to the first term on the right-hand side of equation (2.10). Therefore, the IPA estimate converges to $r'(\theta)$ if and only if the second term on the right-hand side of equation (2.10) is zero.

As we will see, typically $p_\theta \neq 0$, $d_\theta(Y) \neq 0$, and $d_\theta(\tau) \neq 0$ so that if IPA is strongly consistent for $r'(\theta)$, it is because the terms cancel in equation (2.8). Thus the key to determining when IPA is strongly consistent for $r'(\theta)$ is to consider what happens when the order of events changes. In the usual case that $p_\theta \neq 0$, the IPA estimate converges to $r'(\theta)$ if and only if

$$r(\theta) = \frac{\mathrm{E}[Y_i(\theta)]}{\mathrm{E}[\tau_i(\theta)]} = \frac{d_\theta(Y)}{d_\theta(\tau)} = \frac{\lim_{d\theta\to 0} \mathrm{E}[Y_i(\theta + d\theta) - Y_i(\theta) \mid \delta_i(\omega) \leq |d\theta|]}{\lim_{d\theta\to 0} \mathrm{E}[\tau_i(\theta + d\theta) - \tau_i(\theta) \mid \delta_i(\omega) \leq |d\theta|]}. \quad (2.11)$$

By Lemma 2.2, equation (2.8) is equivalent to the condition

$$\mathrm{E}[\tau_i(\theta)]\left(\mathrm{E}[Y_i'(\theta)] - \frac{d}{d\theta}\mathrm{E}[Y_i(\theta)]\right) = \mathrm{E}[Y_i(\theta)]\left(\mathrm{E}[\tau_i'(\theta)] - \frac{d}{d\theta}E[\tau_i(\theta)]\right) \quad (2.12)$$

given in Zazanis and Suri (1985b) which was obtained by equating terms in equations (2.2) and (2.4). We now reinterpret the Suri and Zazanis (1988) result for the mean response time in the $M/G/1$ queue in light of Theorem 2.1. In this case $\tau_i(\theta)$ is the number of customers served during the ith busy period and $Y_i(\theta)$ is the sum of the response times of all customers served during this busy period. In Appendix A we will show that $\mathrm{E}[\tau_i(\theta)] = d_\theta(\tau)$ and that $\mathrm{E}[Y_i(\theta)] = d_\theta(Y)$ so that $\mathrm{E}[Y_i(\theta)]/\mathrm{E}[\tau_i(\theta)] = d_\theta(Y)/d_\theta(\tau)$ which, by Theorem 2.1, proves that IPA is strongly consistent for $r'(\theta)$ in the $M/G/1$ queue (a similar result for a class of $GI/G/1$ queues, obtained under somewhat different technical conditions, has also been obtained in Zazanis and Suri 1985b). The basic idea of the proof here is that the way in which events change order in a single-server queue is for multiple busy periods to collapse into a single busy period. If two busy periods collapse into one, then $\tau_1(\theta + d\theta) = \tau_1(\theta) + \tau_2(\theta)$ so that $d_\theta(\tau) = \mathrm{E}[\tau_i(\theta)]$ provided the probability is small enough that more than two successive busy periods collapse together. Similarly, when two busy periods collapse into one, the sum of the response times becomes $Y_1(\theta + d\theta) = Y_1(\theta) + Y_2(\theta) + \psi(d\theta)$ where $\psi(d\theta)$ is the sum of the increases in individual response times. For the $M/G/1$ queue and for a class of $GI/G/1$ queues (see Zazanis and Suri 1985b), $\mathrm{E}[\psi(d\theta) \mid \delta_i(\omega) \leq |d\theta|] = O(d\theta)$ and therefore $d_\theta(Y) = \mathrm{E}[Y_i(\theta)]$ (again assuming that the effect of more than two busy periods collapsing into one is negligible). These arguments, which are formalized in the Appendix, explain intuitively why IPA works in certain single server queues. Using a somewhat different approach, this basic notion has been used in Zazanis and Suri (1985a) to show that IPA can be extended to estimate higher derivatives in the $GI/G/1$ queue. This extension basically requires deriving an estimate for p_θ. The approach in Zazanis and Suri (1985a) provides an alternative interpretation of the $GI/G/1$ result. Consider a customer looking back in time: the probability that the previous busy period merges with the customer's busy period is $O(d\theta)$; however, the increase in that customer's response time due to this merge is also $O(d\theta)$. The expected effect is thus $o(d\theta)$ and can therefore be neglected.

As described above, the key to identifying when IPA produces consistent estimates for steady state derivatives is to consider the effect when events change order. A general characterization of stochastic systems for which the conditions of Theorem 2.1 are satisfied is an open problem. In the single-server queue, a change in event order is such that the integral over the original cycle is lengthened by the addition of the integral over a typical cycle plus a negligible term, from which the conditions of Theorem 2.1 are satisfied via the relationships $d_\theta(Y) = \mathrm{E}[Y_i(\theta)]$ and $d_\theta(\tau) = \mathrm{E}[\tau_i(\theta)]$. A full characterization of systems for which Theorem 2.1 is satisfied in this specific manner is also an open problem.

We now argue heuristically why IPA works for Jackson networks. A formal proof may be found in Cao (1987c). Suppose θ is a mean service time at a queue and that θ is increased to $\theta + d\theta$. Events can change order in two ways. First, idle periods can be eliminated (or created). Although idle periods do not necessarily correspond to regenerative cycles, the argument given above shows that IPA is insensitive to such occurrences, at least in the $M/G/1$ queue. Second, customers can change order in the queue, i.e., if customer i arrives before customer $i + 1$ in the sample path with parameter θ, it could happen that customer i arrives after customer $i + 1$ in the sample path with parameter $\theta + d\theta$. Let $A_i(\theta)$ and $D_i(\theta)$ denote the arrival time and departure time, respectively, of customer i to the queue when the parameter is θ. Let $A_i'(\theta)$ and $D_i'(\theta)$ denote the IPA sample path derivatives of $A_i(\theta)$ and $D_i(\theta)$, respectively. Thus if $|d\theta| < \delta_1(\omega)$, then (ignoring the $o(d\theta)$ terms) $A_i(\theta + d\theta) = A_i(\theta) + d\theta A_i'(\theta)$ and $D_i(\theta + d\theta) = D_i(\theta) + d\theta D_i'(\theta)$. Let $R_i(\theta) = D_i(\theta) - A_i(\theta)$ denote the response time of customer i and let $\Delta R_i(\theta) = R_i(\theta + d\theta) - R_i(\theta)$. Then

$$\Delta R_i(\theta) + \Delta R_{i+1}(\theta) = d\theta(D_i'(\theta) - A_i'(\theta)) + d\theta(D_{i+1}'(\theta) - A_{i+1}'(\theta)). \qquad (2.13)$$

Consider now what happens if $|d\theta| \geq \delta_1(\omega)$ and that the first event for which a change in event order occurs is that now customer $i + 1$ arrives before customer i in the network with parameter $\theta + d\theta$. Assuming the queue is FCFS, then customer i's service time and routing indicators can be given to customer $i + 1$ and similarly customer $i + 1$'s service time and routing indicators can be given to customer i. Therefore customer $i + 1$ arrives to the queue at time $A_{i+1}(\theta + d\theta) = A_{i+1}(\theta) + d\theta A_{i+1}'(\theta)$ and leaves the queue at time $D_{i+1}(\theta + d\theta) = D_i(\theta) + d\theta D_i'(\theta)$. Similarly, $A_i(\theta + d\theta) = A_i(\theta) + d\theta A_i'(\theta)$ and $D_i(\theta + d\theta) = D_{i+1}(\theta) + d\theta D_{i+1}'(\theta)$. Therefore

$$\Delta R_i(\theta) + \Delta R_{i+1}(\theta) = d\theta(D_{i+1}'(\theta) - A_i'(\theta)) + d\theta(D_i'(\theta) - A_{i+1}'(\theta)) \qquad (2.14)$$

which equals the expression in equation (2.13) in which events do not change order. Thus the only effect of the change in event order is that customer i and $i + 1$ interchange departure times. Therefore the IPA algorithm is insensitive to this type of event change provided that all customers have identical service time demand distributions and identical statistical routing distributions. This heuristic argument will not work if customers are not homogeneous. In that case, the change in event order leads to a discontinuity and thus we cannot expect IPA to produce consistent estimates in queueing networks with multiple types of customers. Ho and Cao (1985) noted that such a discontinuity exists. The above argument identifies the cause of this discontinuity and shows why it is not a problem in networks with homogeneous customers. This will be verified theoretically in the next section and experimentally in §4.

3. IPA Throughput Estimates

In this section, we will derive necessary conditions that must hold if IPA is to simultaneously produce strongly consistent estimates for the derivatives of multiple steady state throughputs. Let $\mathbf{Q}(\theta)$ be a family of stochastic processes whose governing probability law depends on a real-valued input parameter θ. An example of $\mathbf{Q}(\theta)$ is a family of queueing networks in which case θ might be a mean service time at a queue, or set of queues, in the network. Let $E_1, \ldots, E_m$ denote m different types of events that can occur in $\mathbf{Q}(\theta)$, for example E_i could denote the event that a customer departs from node i in the network. Let $N_i(t, \theta)$ be the number of type i events that occur in the interval $[0, t]$ and let $T_n(\theta)$ denote the time at which the nth event occurs (the index n counts all events, not just those associated with $E_1, \ldots, E_m$). Let $\hat{\lambda}_i(\theta, n) = N_i(T_n(\theta), \theta)/T_n(\theta)$ be an estimate of the rate at which type i events occur.

DEFINITION 3.1. $\mathbf{Q}(\theta)$ is called λ-Consistent at θ_0 if $\lim_{n\to\infty} T_n(\theta_0) = \infty$ with probability one and there exist finite, positive constants $\lambda_1(\theta_0), \ldots, \lambda_m(\theta_0)$ such that $\lim_{n\to\infty} \hat{\lambda}_i(\theta_0, n) = \lambda_i(\theta_0)$ with probability one for $i = 1, \ldots, m$.

DEFINITION 3.2. $\mathbf{Q}(\theta)$ is called λ-Consistent on the interval (a, b), if $\mathbf{Q}(\theta)$ is λ-Consistent at θ_0 for all $\theta_0 \in (a, b)$.

In the queueing network example, $\lambda_i(\theta)$ is the steady state throughput of customers at queue i.

The definition of λ-Consistency requires that, as the length of the simulation increases, throughput estimates converge with probability one to so-called steady state throughputs. The existence of such limits is an underlying assumption of many simulations. The question of when a process is λ-Consistent is thus related to the question of when a process is ergodic. Assuming the expected number of events in a regenerative cycle is finite, then any regenerative process, including any irreducible, positive recurrent continuous time Markov chain, is λ-Consistent as are some processes with more general state spaces such as certain finite state space Generalized Semi-Markov processes (Whitt 1980 or Glynn 1983). In particular, any closed queueing network described by a continuous time Markov chain with an irreducible finite state space is λ-Consistent including any (irreducible) closed product form network with Coxian phase-type service distributions. Note that a process may be λ-Consistent for certain types of event rates even though the process itself is not ergodic. For example, consider the departure process $\{D(t), t \geq 0\}$ in a $GI/G/1$ queue with arrival rate λ, mean service time $1/\mu$ and traffic intensity $\rho = \lambda/\mu > 1$. In this case $\lim_{t\to\infty} D(t)/t = \mu$ with probability one even though the waiting time and queue length processes are neither regenerative nor ergodic.

We will assume that the derivatives $\lambda_i'(\theta) = d\lambda_i(\theta)/d\theta$ exist and are finite for all i and all $\theta \in (a, b)$. The goal of IPA is to estimate $\lambda_1'(\theta), \ldots, \lambda_m'(\theta)$. Although we assume that the interval (a, b) is open, the definition and all subsequent results can also be given for closed or half open and half closed intervals with either a or b infinite. If the interval is closed, the derivative is understood to be the appropriate one-sided limit at the end points. Let $\hat{\lambda}_i'(\theta, n)$ denote the IPA estimate of $\lambda_i'(\theta)$ after n events.

DEFINITION 3.3. IPA is called λ′-Consistent for $\mathbf{Q}(\theta)$ at θ_0 if $\lim_{n\to\infty} \hat{\lambda}_i'(\theta_0, n) = \lambda_i'(\theta_0)$ with probability one for $i = 1, \ldots, m$.

DEFINITION 3.4. IPA is called λ′-Consistent for $\mathbf{Q}(\theta)$ on the interval (a, b) if IPA is λ′-Consistent for $\mathbf{Q}(\theta)$ at θ_0 for all $\theta_0 \in (a, b)$.

The motivation for defining λ′-Consistency on an interval is to classify stochastic systems for which IPA produces strongly consistent estimates for all values (or a range of values) of an input parameter rather than at just certain parameter values. For example, if $\theta = \lambda$ is the arrival rate in a queueing system, then the stability conditions for the system often include a requirement that $\lambda < b$ for some constant b. In this case, λ′-Consistency on the interval $(0, b)$ corresponds to the requirement that IPA produce strongly consistent estimates at all values of the arrival rate for which the queueing system is stable.

Implicit in these definitions is the assumption that the appropriate sample path derivatives exist. The primary assumption of IPA is that for a small enough change $d\theta$ in θ, the event times get shifted but there is no change in the order of events. Thus, for any sample path realization ω, there exists a $\delta_n(\omega) > 0$ such that if $|d\theta| < \delta_n(\omega)$, then $T_n(\theta + d\theta) = T_n(\theta) + d\theta T_n'(\theta) + o(d\theta, n, \omega)$ where $T_n'(\theta)$ is the (finite) random variable accumulated by the IPA algorithm and $\lim_{d\theta\to 0} o(d\theta, n, \omega)/d\theta = 0$ with probability one. Note that the appropriate δ depends on both n and ω. In this case, the sample path derivative $dT_n(\theta)/d\theta = T'_n(\theta)$. Suri (1987) formally defines a general class of stochastic processes and input parameters θ for which this representation of $T_n(\theta + d\theta)$ is valid. Although there may be several different ways of generating events (and therefore event time perturbations, e.g., there may be more than one way to sample from a given service time distribution, see Glynn 1987 or Glasserman 1988, we assume that the simulation implements a particular, fixed event generation mechanism to which IPA is applied for the purpose of simultaneously estimating $\lambda'_1(\theta), \ldots, \lambda'_m(\theta)$ from a single simulation run.

The main result of this section is the following theorem which states necessary conditions for IPA to be λ-Consistent on an interval.

THEOREM 3.1. *Suppose* $\mathbf{Q}(\theta)$ *is* λ-*Consistent on* (a, b). *If IPA is* λ′-*Consistent for* $\mathbf{Q}(\theta)$ *on* (a, b) *then there exist positive constants* c_{ij} $(1 \leq i, j \leq m)$ *such that*

$$\frac{\lambda_i(\theta)}{\lambda_j(\theta)} = c_{ij} \qquad (a < \theta < b). \tag{3.1}$$

PROOF. Let $\theta \in (a, b)$. As discussed above, if $|d\theta| < \delta_n(\omega)$, then $T_n(\theta + d\theta) = T_n(\theta) + d\theta T'_n(\theta) + o(d\theta, n, \omega)$ so that for small enough $d\theta$

$$\frac{\hat{\lambda}_i(\theta + d\theta, n) - \hat{\lambda}_i(\theta, n)}{d\theta} = \frac{N_i(T_n(\theta), \theta)}{d\theta}\left(\frac{1}{T_n(\theta) + d\theta T'_n(\theta) + o(d\theta, n, \omega)} - \frac{1}{T_n(\theta)}\right). \tag{3.2}$$

By rearranging terms and taking the limit of equation (3.2) as $d\theta$ approaches zero we obtain the sample path derivative

$$\hat{\lambda}'_i(\theta, n) = \frac{-N_i(T_n(\theta), \theta)T'_n(\theta)}{T_n(\theta)^2} = \hat{\lambda}_i(\theta, n)\left(\frac{-T'_n(\theta)}{T_n(\theta)}\right). \tag{3.3}$$

Since, by assumption, $\lim_{n\to\infty} \hat{\lambda}_i(\theta, n) = \lambda_i(\theta) > 0$ with probability one, we can divide both sides of equation (3.3) by $\hat{\lambda}_i(\theta, n)$ provided n is large enough to obtain

$$\frac{\hat{\lambda}'_i(\theta, n)}{\hat{\lambda}_i(\theta, n)} = \frac{-T'_n(\theta)}{T_n(\theta)}. \tag{3.4}$$

Since we are applying a single IPA algorithm along a single sample path to simultaneously estimate $\lambda'_1(\theta), \ldots, \lambda'_m(\theta)$ (see comment above), the right-hand side of equation (3.4) is independent of i and

$$\frac{\hat{\lambda}'_i(\theta, n)}{\hat{\lambda}_i(\theta, n)} = \frac{\hat{\lambda}'_j(\theta, n)}{\hat{\lambda}_j(\theta, n)} \tag{3.5}$$

for all i and j. Since, by assumption, $\mathbf{Q}(\theta)$ is λ-Consistent and IPA is λ′-Consistent for $\mathbf{Q}(\theta)$, when n tends to infinity we obtain

$$\frac{\lambda'_i(\theta)}{\lambda_i(\theta)} = \frac{\lambda'_j(\theta)}{\lambda_j(\theta)}. \tag{3.6}$$

Note that equation (3.6) must hold for all $\theta \in (a, b)$ and that

$$\lambda'_i(\theta)/\lambda_i(\theta) = \frac{d}{d\theta} \ln(\lambda_i(\theta)).$$

Therefore

$$\frac{d}{d\theta} \ln(\lambda_i(\theta)) = \frac{d}{d\theta} \ln(\lambda_j(\theta))$$

which, since $\lambda_i(\theta)$ and $\lambda_j(\theta)$ are continuous functions of θ, implies that $\ln(\lambda_i(\theta)) = \ln(\lambda_j(\theta)) + k_{ij}$ or that $\lambda_i(\theta) = c_{ij}\lambda_j(\theta)$.

If $\mathbf{Q}(\theta)$ is a closed product form network with a single type of customer, θ is a mean service time at some queue and $\lambda_i(\theta)$ is the throughput at queue i, then the necessary conditions of Theorem 3.1 are satisfied. This is consistent with Cao (1987c) who showed that IPA produces strongly consistent estimates of $\lambda_i'(\theta)$ for such networks (with single-server, fixed rate FCFS queues).

As an immediate corollary, we have the point-wise version of Theorem 3.1:

COROLLARY 3.1. *If $\mathbf{Q}(\theta)$ is λ-Consistent at $\theta = \theta_0$ and if IPA is λ'-Consistent for $\mathbf{Q}(\theta)$ at $\theta = \theta_0$, then $\lambda_i'(\theta_0)/\lambda_i(\theta_0) = \lambda_j'(\theta_0)/\lambda_j(\theta_0)$.*

Corollary 3.1 is just a restatement of equation (3.6). Equation (3.5) and Corollary 3.1 are more general versions of results obtained by Cao (1987a); Cao also used his versions of these results to conclude that IPA may not lead to exact estimates of throughput derivatives in closed queueing networks with multiple customer types. We now examine the implications of Theorem 3.1 in closer detail via the following series of corollaries.

COROLLARY 3.2. *Let $\mathbf{Q}(\theta)$ be a λ-Consistent mixed open and closed network with two types of customers (type 1 open, type 2 closed). Let $\lambda_1(\theta)$ be the departure rate of type 1 customers from the network and let $\lambda_2(\theta)$ be the throughput of type 2 customers at some queue in the network. Suppose $\lambda_1(\theta) = \lambda_1$ where λ_1 is the arrival rate of type one customers to the network. If IPA is λ'-Consistent for $\mathbf{Q}(\theta)$ on (a, b), then there exists a constant c such that $\lambda_2(\theta) = c$ for all $\theta \in (a, b)$.*

As a specific example of the corollary let θ be the mean service time at some queue. This corollary means that if the arrival rate equals the departure rate of open customers, i.e., if what goes in comes out, then the only networks for which IPA can be λ'-Consistent are ones in which the closed chain throughput is independent of θ. It is unlikely that this condition would be satisfied if the closed chain customers visit the queue associated with θ.

COROLLARY 3.3. *Let $\mathbf{Q}(\theta)$ be a λ-Consistent multiple chain closed product form network with population vector $\mathbf{N} = (N_1, \ldots, N_m)$ and let $\lambda_i(\theta) > 0$ be the throughput of type i jobs at some fixed queue, say queue number k, in the network. If IPA is λ'-Consistent for $\mathbf{Q}(\theta)$ on (a, b), then there exist constants $c_{ij} > 0$ such that*

$$\frac{\mathbf{G}(\mathbf{N} - \mathbf{e}_i, \theta)}{\mathbf{G}(\mathbf{N} - \mathbf{e}_j, \theta)} = c_{ij} \qquad (a < \theta < b) \tag{3.7}$$

where $\mathbf{G}(\mathbf{N}, \theta)$ is the normalization constant of the network with population $\mathbf{N}$, and $\mathbf{e}_i$ is a vector of zeros except for a one in the ith component.

PROOF. For such a network $\lambda_i(\theta)$ is proportional to $\mathbf{G}(\mathbf{N} - \mathbf{e}_i, \theta)/\mathbf{G}(\mathbf{N}, \theta)$ (see, for example, Reiser and Lavenberg 1980).

This corollary means that the effect on the normalization constant of removing a type i customer from the network must be a constant multiple, for all θ, of the effect of removing a type j customer. This implies that IPA can only be λ'-Consistent in the case that customers are essentially identical. Note that Gong and Ho (1987) have described an IPA extension that does produce strongly consistent throughput derivative estimates in a simple network with two types of customers (specifically, the network of Figure 1 in §4 with $s_2 = 0$ and $N_1 = N_2 = 1$). In addition, Cao (1987a) has described an IPA extension that produces finite difference approximations to the throughput derivatives for the above mentioned network and Ho and Li (1988) have described a different IPA extension that produces potentially accurate finite difference approximations to the throughput derivatives in more general networks with multiple customer types.

COROLLARY 3.4. *Let $\mathbf{Q}(\theta)$ be a multiple chain closed product form network with population vector $\mathbf{N} = (N_1, N_2)$ and let $\lambda_i(\theta) > 0$ be the throughput of type i jobs at some fixed queue, say queue 0, in the network. Let y_{ik} be the relative visit ratios of type i customers to queue k, s_{ik} be the mean service requirement of type i customers at queue k and let $\mu_1(n) = f(n)\theta$ be the queue-dependent service rate at queue 1 when there are a total of n jobs at the queue. If $y_{11} > 0$, $s_{11} > 0$ and $y_{21}s_{21} = 0$, then IPA cannot be λ'-Consistent for $\mathbf{Q}(\theta)$ on any interval.*

PROOF. Because $y_{11}s_{11} > 0$ and $y_{21}s_{21} = 0$, $\mathbf{G}(\mathbf{N} - \mathbf{e}_1, \theta)$ is a polynomial (in $1/\theta$) of degree $N_1 - 1$. However, in this case $\mathbf{G}(\mathbf{N} - \mathbf{e}_2, \theta)$ is a polynomial of degree N_1. By Corollary 3.3, $\mathbf{G}(\mathbf{N} - \mathbf{e}_1, \theta)/\mathbf{G}(\mathbf{N} - \mathbf{e}_2, \theta)$ must be independent of θ which is impossible.

This corollary means that IPA cannot be λ'-Consistent if θ affects the service rate at a queue that is visited by only one type of customer.

COROLLARY 3.5. *Let $\mathbf{Q}(\theta)$ be a queueing network with state-dependent routing, i.e., when a customer departs from queue i at time t it goes to queue j with probability $p_{ij}(\mathbf{X}_t(\theta))$ where $\mathbf{X}_t(\theta)$ is the state of the queueing network at time t. Let $\lambda_{ij}(\theta)$ denote the steady state throughput over the path from queue i to queue j and let $\lambda_i(\theta) = \sum_k \lambda_{ik}(\theta)$. If $\mathbf{Q}(\theta)$ is λ-Consistent on (a, b) and if IPA is λ'-Consistent for $\mathbf{Q}(\theta)$ on (a, b) then there exist constants p_{ij} such that $\lambda_{ij}(\theta)/\lambda_i(\theta) = p_{ij}$ for all $\theta \in (a, b)$.*

PROOF. Assume $\lambda_{ij}(\theta) > 0$ and $\lambda_{ik}(\theta) > 0$. By Theorem 3.1 there exists a positive constant c_{ijk} such that $\lambda_{ij}(\theta)/\lambda_{ik}(\theta) = c_{ijk}$. Therefore $\lambda_{ij}(\theta) \sum_k (1/c_{ijk}) = \sum_k \lambda_{ik}(\theta) = \lambda_i(\theta)$ where the sum is over all queues k such that $\lambda_{ik}(\theta) > 0$.

The ratio $\lambda_{ij}(\theta)/\lambda_i(\theta)$ is the steady state fraction of jobs routed from queue i to queue j. Thus, this corollary means that in a network with state-dependent routing, IPA can only be λ'-Consistent if the steady state routing fractions are independent of θ.

This corollary also proves that IPA cannot in general be used to exactly estimate throughput derivatives in networks that exhibit blocking because of finite capacity queues. This has been observed empirically for response times in Cao and Ho (1983). Suppose queue 2 has a finite capacity buffer and that when customers finish service at queue 1 they move to queue 2 if the buffer is not full and move to a fictitious queue, queue 0, otherwise. While there are any customers at queue 0, no other customers are served at queue 1. This is a particular form of state-dependent routing and the necessary condition for IPA to produce consistent estimates is that $\lambda_{02}(\theta)/\lambda_1(\theta) = c$, i.e., the fraction of jobs blocked must be independent of θ. In addition, this corollary also provides an alternative interpretation to the fact that IPA cannot be used directly to accurately estimate derivatives with respect to routing probabilities (an indirect method, based on the properties of stationary product form networks, was proposed in Ho and Cao 1985).

COROLLARY 3.6. *Let $\mathbf{Q}(\theta)$ be a queueing network and let E_i denote the event that an arriving (departing) customer to a fixed queue finds (leaves) i customers already at the queue. If $\mathbf{Q}(\theta)$ is λ-Consistent on (a, b) and if IPA is λ'-Consistent for $\mathbf{Q}(\theta)$ on (a, b) then $\lambda_i(\theta)/\lambda_j(\theta) = c_{ij}$ for all $\theta \in (a, b)$.*

This corollary shows that derivatives of the so-called "on-arrival" and "on-departure" distributions cannot be estimated by IPA. In fact this is obvious even without Theorem 3.1 since IPA assumes that events do not change in order so that when a customer arrives to a queue he/she sees the same queue length for both θ and $\theta + d\theta$ for small enough $d\theta$. Therefore the IPA estimate of the derivative of the on-arrival distribution is identically equal to 0. Zazanis and Suri (1985a) and Gong and Ho (1987) discuss an extension to IPA to estimate the derivative of the probability of zero wait in the $GI/G/1$ queue.

There are anomalous situations in which IPA can produce strongly consistent estimates but for which equation (3.1) in Theorem 3.1 does not hold. For example, consider a queueing network with two types of customers and let Q_i denote the set of queues visited by type i customers. Suppose that $Q_1 \cap Q_2 = \varnothing$ and that θ affects only type 1 customers. If the network associated with type i customers is a closed product form network and θ is a mean service time, then the results of Cao (1987c) imply that IPA produces a consistent estimate of $\lambda_1'(\theta) \neq 0$ and IPA also estimates $\lambda_2'(\theta) = 0$ which is also correct, but in apparent violation of Theorem 3.1. The explanation lies in the fact that the assumption of λ'-Consistency does not hold in this case since the accumulated perturbation $|T_n'(\theta)| \to \infty$ if event n corresponds to an event in Q_1 whereas $T_n'(\theta) = 0$ for all Q_2 events. Thus, in equation (3.4), $\lim_{n\to\infty} \hat{\lambda}_i'(n, \theta)$ does not exist, however there exist subsequences $m_i(n)$ corresponding to Q_i events such that $\lim_{n\to\infty} \hat{\lambda}_i'(m_i(n), \theta)$ exists for $i = 1, 2$. Thus Theorem 3.1 does not apply. However, the following lemma gives conditions under which this type of behavior cannot occur.

LEMMA 3.1. *Suppose* $\lim_{n\to\infty} \hat{\lambda}_i(n, \theta) = \lambda_i(\theta)$ *with probability one and there exists an increasing subsequence* $\{m(n), n \geq 1\}$ *such that, with probability one*:

1. $\lim_{n\to\infty} m(n) = \infty$,
2. $\lim_{n\to\infty} \hat{\lambda}_1'(m(n), \theta) = \alpha_1(\theta)$,
3. $\lim_{n\to\infty} T_{m(n+1)}(\theta)/T_{m(n)}(\theta) = 1$,
4. $\lim_{n\to\infty} (T_n'(\theta) - T_{l(n)}'(\theta))/T_{l(n)}(\theta) = 0$ *where* $l(n) = \sup\{m(k): m(k) \leq n\}$,

then $\lim_{n\to\infty} \hat{\lambda}_i'(n, \theta)$ *exists with probability one.*

PROOF. We suppress the dependency on θ in the notation. From equation (3.4) it suffices to show that $\lim_{n\to\infty} T_n'/T_n$ exists. Write

$$\frac{T_n'}{T_n} = \frac{T_n' - T_{l(n)}'}{T_n} + \frac{T_{l(n)}'}{T_n} = \left(\frac{T_n' - T_{l(n)}'}{T_{l(n)}} + \frac{T_{l(n)}'}{T_{l(n)}}\right)\left(\frac{T_{l(n)}}{T_n}\right). \tag{3.8}$$

By assumption 4, $(T_n' - T_{l(n)}')/T_{l(n)} \to 0$. Since $\hat{\lambda}_i(n, \theta) \to \lambda_i(\theta)$, and by assumptions 1 and 2, $-T_{l(n)}'/T_{l(n)} = \hat{\lambda}_1'(l(n), \theta)/\hat{\lambda}_1(l(n), \theta) \to \alpha_1(\theta)/\lambda_1(\theta)$. Define $k(n) = \sup\{j: m(j) \leq n\}$. Then $l(n) = m(k(n))$ and since T_n is increasing in n, $T_{l(n)} = T_{m(k(n))} \leq T_n \leq T_{m(k(n)+1)}$ and $(T_{m(k(n))}/T_{m(k(n)+1)}) \leq (T_{l(n)}/T_n) \leq 1$. Therefore, by assumption 3, $T_{l(n)}/T_n \to 1$ and $\lim_{n\to\infty} - T_n'/T_n = \alpha_1(\theta)/\lambda_1(\theta)$ with probability one.

The conditions in Lemma 3.1, in effect, require that if two events do not occur far apart in time, then their accumulated perturbations cannot be far apart. These conditions will be satisfied if the queueing network contains a single-server queue such that, with probability one, there exists an infinite sequence of busy periods in which both type 1 and type 2 customers are served (and the length of busy periods and the accumulated perturbations within a busy period are finite with probability one). In this example $T_{m(n)}$ is the nth time that a type 1 customer initiates a busy period. Thus if the different types of customers compete for resources at a shared queue, the above type of anomalous behavior cannot occur and the necessary conditions of Theorem 3.1 apply. A related concept of indecomposable network has been proposed in Cao (1987b).

Note that the definition of λ'-Consistency requires that $\lim_{n\to\infty} \hat{\lambda}_i'(\theta, n) = \lambda_i'(\theta)$ with probability one for all $i = 1, \ldots, m$. Thus λ'-Consistency does not hold if, say, $m = 2$ and $\lim_{n\to\infty} \hat{\lambda}_1'(\theta, n) = \lambda_1'(\theta)$ but $\lim_{n\to\infty} \hat{\lambda}_2'(\theta, n) \neq \lambda_2'(\theta)$ (with probability one). In this case IPA produces a strongly consistent estimate of $\lambda_1'(\theta)$ but not $\lambda_2'(\theta)$. However, our general computational experience has been that unless equation (3.1) holds, then (with probability one) both $\lim_{n\to\infty} \hat{\lambda}_1'(\theta, n) \neq \lambda_1'(\theta)$ and $\lim_{n\to\infty} \hat{\lambda}_2'(\theta, n) \neq \lambda_2'(\theta)$ (see, e.g., §4 for simulations of networks with multiple types of customer and §VI of Cao 1985 for simulations of networks with blocking).

In Theorem 3.1, we assumed that there is a single IPA algorithm applied along a single sample path to estimate all of the throughput derivatives simultaneously. As mentioned above, there may be more than one representation of the simulation, e.g., more than one way to sample from the same service time distribution. Thus it may be possible to, say, run two different IPA algorithms using two such different representations and use one representation to consistently estimate $\lambda_1'(\theta)$ and the other representation to consistently estimate $\lambda_2'(\theta)$. Strictly speaking Theorem 3.1 does not apply in this situation and equation (3.1) is no longer a necessary condition for strong consistency (although it does agree with our computational experience as described above). Implementation of such multiple IPA algorithms would appear, in general, to require running multiple simulations thereby losing the single run advantage of IPA.

4. IPA Experimental Results

In this section we investigate the convergence properties of the IPA estimates experimentally. We performed two sets of experiments. These experiments were performed using the IBM Research Queueing (RESQ) simulation package (see, e.g., Sauer and MacNair 1979). The simulation run lengths were sufficiently long so that very precise estimates (with estimates of their standard deviations) were produced. Furthermore, since analytical results are available for all of the systems simulated, it is possible to experimentally determine those cases for which the IPA estimates converge to the steady state derivatives.

The first set of experiments involved variations of the $M/M/1$ queue. Each system was simulated for 1,000,000 customers and, since the point estimates are all simple ratios, the Regenerative method (see, e.g., Crane and Iglehart 1975) was used to estimate standard deviations of the point estimates. The systems simulated were:

1. The first and second moments of the stationary response time in the $M/M/1$ queue,
2. The mean of the stationary response time in the $M/M/1$ queue with feedback,
3. The mean of the stationary response time in the $M/M/2$ queue, and
4. The mean of the stationary response time in the $M/M/1$ queue with nonpreemptive priorities and two types of customers.

Table 1 lists the input parameters for each of the experiments. In Table 1, λ denotes the arrival rate, s denotes the mean service time and p denotes the feedback probability. In the nonpreemptive queue, type 1 customers have priority over type 2 customers and λ and s are indexed by i. Table 2 lists the results of the experiments including the performance measure's true value, its estimate and the standard deviation of the estimate. In Table 2, R denotes the stationary response time (R is indexed by i for the nonpreemptive queue).

IPA is known to produce consistent estimates in the $M/M/1$ queue and our experimental results confirm this. The experiments also suggest that IPA produces strongly consistent estimates in the $M/M/1$ queue with feedback and in the $M/M/2$ queue. However, as expected from the theory presented here, IPA does not produce strongly

TABLE 1
Input Parameters for IPA Experiments on Variations of the M/M/1 Queue

Experiment	System	Traffic Intensity	Parameters		
I.1	M/M/1 Queue	0.10	$\lambda = 0.1$	$s = 1.0$	
I.2	M/M/1 with Feedback	0.10	$\lambda = 0.1$	$s = 0.5$	$p = 0.5$
I.3	M/M/1 with Feedback	0.50	$\lambda = 0.5$	$s = 0.5$	$p = 0.5$
I.4	M/M/2 Queue	0.30	$\lambda = 0.6$	$s = 1.0$	
I.5	M/M/2 Queue	0.60	$\lambda = 1.2$	$s = 1.0$	
I.6	Nonpreemptive M/M/1	0.15	$\lambda_i = 0.1$	$s_1 = 1.0$	$s_2 = 0.5$
I.7	Nonpreemptive M/M/1	0.30	$\lambda_i = 0.2$	$s_1 = 1.0$	$s_2 = 0.5$

1,000,000 Customers/Experiment

TABLE 2

IPA Experimental Results on Variations of the M/M/1 Queue

Experiment	Parameter	True Value	Estimate	Std. Dev.
I.1	$E(R)$	1.1111	1.1112	0.0016
	$dE(R)/ds$	1.2346	1.2349	0.0022
	$dE(R^2)/ds$	5.4870	5.4959	0.0216
I.2	$E(R)$	1.1111	1.1113	0.0016
	$dE(R)/ds$	2.4691	2.4678	0.0045
I.3	$E(R)$	2.0000	1.9904	0.0060
	$dE(R)/ds$	8.0000	7.9060	0.0419
I.4	$E(R)$	1.0989	1.0995	0.0015
	$dE(R)/ds$	1.3163	1.3166	0.0030
I.5	$E(R)$	1.5625	1.5603	0.0049
	$dE(R)/ds$	3.3203	3.3073	0.0229
I.6	$E(R_1)$	1.1389	1.1399	0.0023
	$E(R_2)$	0.6634	0.6636	0.0017
	$dE(R_1)/ds_1$	1.2377	1.2543	0.0033
	$dE(R_2)/ds_1$	0.2241	0.2699	0.0023
I.7	$E(R_1)$	1.3125	1.3124	0.0032
	$E(R_2)$	0.9464	0.9477	0.0036
	$dE(R_1)/ds_1$	1.5781	1.6740	0.0066
	$dE(R_2)/ds_1$	0.4751	0.7649	0.0062

consistent estimates for the derivative of the mean stationary response time in the $M/M/1$ queue with priorities and multiple customer types. The errors in the IPA estimates increase with the traffic intensity.

The second set of experiments involved estimating throughput derivatives in extremely simple closed product form queueing networks with one or two types of customers. In order to compare the IPA simulation results with exact analytical results, a simple extension to the Mean Value Analysis algorithm (Reiser and Lavenberg 1980) for computing derivatives in closed multiple chain product form networks may be used (Strelen 1986). The general network is pictured in Figure 1. There are N_i customers of type i. Type 1 customers are routed from queue 1 to queue 2 and back to queue 1 again. Type 2

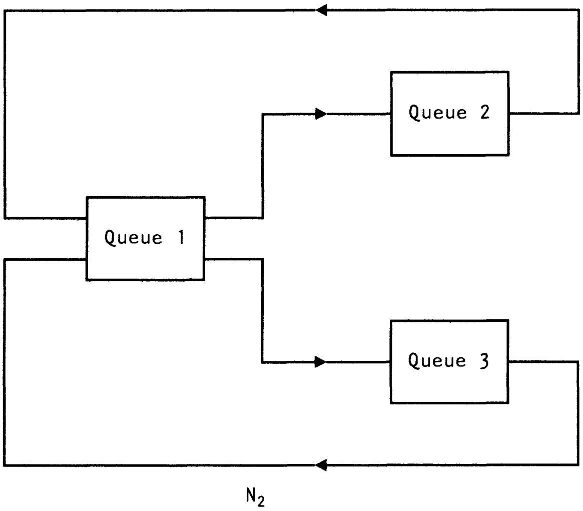

FIGURE 1. Closed Product Form Network with Two Types of Customers.

TABLE 3

Input Parameters for IPA Experiments on Closed Product Form Networks with Two Types of Customers

Experiment	N_1	N_2	s_1	s_2	s_3
II.1	2	0	1.0	1.2	
II.2	2	2	1.0	1.2	1.2
II.3	2	2	1.0	1.2	0.9

10 Replications/Experiment
50,000 departures from Queue 1/Replication

customers are routed from queue 1 to queue 3 and back to queue 1 again. Queue i is a single-server FCFS queue with exponential service times having mean s_i (Cao 1987a has considered a special case of this system with $s_2 = 0$ and $N_1 = N_2 = 1$). Each experiment was run for 10 replications with 50,000 departures from queue 1 per replication. The method of independent replications was used to estimate variances since estimating the variance of the derivative estimate using the Regenerative method is not entirely straightforward. Table 3 lists the input parameters for these experiments and Table 4 lists the results of these experiments.

The first of these experiments had only a single type of customer ($N_2 = 0$) and, in agreement with theory, the IPA estimates did converge to $d\lambda_1/ds_j$. The second experiment satisfied the Corollary 3.3 necessary condition that $\mathbf{G}(\mathbf{N} - \mathbf{e}_1, s_1) = \mathbf{G}(\mathbf{N} - \mathbf{e}_2, s_1)$ (with all other parameters fixed) and indeed the IPA estimates were consistent for $d\lambda_i/ds_1$. However, for $j \neq 1$, $\mathbf{G}(\mathbf{N} - \mathbf{e}_1, s_j) \neq c\mathbf{G}(\mathbf{N} - \mathbf{e}_2, s_j)$ and the IPA estimates were not consistent for $d\lambda_i/ds_j$ for $j \neq 1$. In fact the IPA estimate for $d\lambda_2/ds_2$ had the wrong sign. Increasing s_2 slows down the type 1 customers and speeds up the type 2 customers so that $d\lambda_2/ds_2 > 0$. However the IPA estimate for $d\lambda_2/ds_2$ is negative since IPA does not allow type 2 customers to pass the slowed down type 1 customers (events cannot change in order). In fact when a type 1 customer starts a busy period at queue 1, then all the departures from queue 1 during that busy period get delayed, including those of type 2 customers. The third experiment does not satisfy the Corollary 3.3 necessary condition for any of the parameters and none of the IPA estimates converged to $d\lambda_i/ds_i$. As mentioned above, Cao (1987a) and Gong and Ho (1987) have developed IPA extensions for

TABLE 4

IPA Experimental Results on Closed Product Form Networks with Two Types of Customers

Experiment	Parameter	True Value	Estimate	Std. Dev.
II.1	λ_1	0.6044	0.6039	0.0010
	$d\lambda_1/ds_1$	−0.2566	−0.2571	0.0006
	$d\lambda_1/ds_2$	−0.2898	−0.2889	0.0005
II.2	λ_1	0.4577	0.4566	0.0010
	λ_2	0.4577	0.4578	0.0010
	$d\lambda_1/ds_1$	−0.3529	−0.3522	0.0005
	$d\lambda_2/ds_1$	−0.3529	−0.3531	0.0006
	$d\lambda_1/ds_2$	−0.1425	−0.0438	0.0002
	$d\lambda_2/ds_2$	0.0552	−0.0439	0.0002
II.3	λ_1	0.4399	0.4393	0.0010
	λ_2	0.5022	0.5024	0.0010
	$d\lambda_1/ds_1$	−0.3406	−0.3662	0.0008
	$d\lambda_2/ds_1$	−0.4463	−0.4189	0.0006
	$d\lambda_1/ds_2$	−0.1300	−0.0308	0.0002
	$d\lambda_2/ds_2$	0.0683	−0.0352	0.0002

particular parameter settings of this network. The IPA extension described in Ho and Li (1988) provides accurate finite difference approximations to the throughput derivatives for this class of network as well.

It is also interesting to estimate the overhead required to implement IPA when used in a high level simulation package like RESQ. In RESQ, customers are routed from one node to another, joining queues, entering service or performing numerical computations depending on the type of node. To implement the closed queueing model described above without IPA required a RESQ model with 3 active queues, 4 classes and 1 passive queue (to measure response times). The model became significantly more complicated when IPA was added; 12 additional set nodes were required, 3 per class node. A set node is where computations are done and these were used to implement the IPA bookkeeping. In addition state-dependent routing had to be introduced since the algorithm is different depending upon whether or not an arrival sees an empty queue. For the same number of queue 1 departures, the CPU time to run the model with IPA was approximately 2.8 times the CPU time without IPA. Thus the total IPA overhead was 180% and since 6 derivatives were estimated, the overhead to implement IPA using RESQ for this model was 30% per derivative. Of course, if the IPA algorithm were built into the simulation package as part of the normal event handling procedures, the overhead would be much lower. For example, Ho, Suri, Cao, Diehl, Dille and Zazanis (1984) included the IPA algorithm as a basic part of the simulation package and reported an average of only 1% overhead per derivative for a variety of queueing networks.

5. IPA-Like Algorithm for Birth and Death Processes

In this section we analyze an IPA-like algorithm for a particular representation of a Birth and Death process simulation (an alternative representation is described in Glasserman 1988, see also discussion below). We will consider the effect of a change in a single parameter, λ_k, of the process. Let $\{\lambda_i > 0, i = 0, \dots, N-1\}$ and $\{\mu_i > 0, i = 1, \dots, N\}$ be the birth and death rates (define $\mu_0 = \lambda_N = 0$). Because the state space is finite and irreducible, the process is positive recurrent. Let π_i be the stationary probability of state i. Then $\pi_i = p_i/(\sum_{j=0}^{N} p_j)$ where $p_i = \prod_{j=1}^{i} \lambda_{j-1}/\mu_j$ for $i \geq 1$ and $p_0 = 1$. In this representation, the IPA-like algorithm only recognizes changes in the holding times of state k but does not take into account the possibility that the order of events might change. Let $\hat{\pi}'_n(i)$ be the IPA-like estimate of $\pi'_i = d\pi_i/d\lambda_k$ after n regenerative cycles. We will show that for $1 < k < N$, $\pi'_0 < 0$ whereas $\lim_{n\to\infty} \hat{\pi}'_n(0) > 0$. Let $p(i,j)$ denote the probability that the embedded Markov chain goes from state i to state j. The problem encountered by this IPA-like algorithm is that it does not consider the possibility of a change in the sequence of states in the embedded Markov chain whereas

$$\frac{d}{d\lambda_k} p(k, k+1) = \mu_k/(\lambda_k + \mu_k)^2 \neq 0.$$

Let E_{ij} denote the jth holding time in state i. The process is simulated as follows. When in state i for the jth time, two independent exponentials with mean one, E_{ij1} and E_{ij2}, are generated and we set $E_{ij}(\mu_i) = E_{ij1}/\mu_i$ and $E_{ij}(\lambda_i) = E_{ij2}/\lambda_i$. Then $E_{ij} = \min(E_{ij}(\mu_i), E_{ij}(\lambda_i))$ and if $E_{ij} = E_{ij}(\mu_i)$, then the next state is $i-1$, otherwise the next state is $i+1$. Consider now a small increase $d\lambda_k$ in λ_k. If $E_{kj} = E_{kj}(\lambda_k) < E_{kj}(\mu_k)$, then $E_{kj}(\lambda_k + d\lambda_k) < E_{kj}(\mu_k)$ and the holding time with parameter $\lambda_k + d\lambda_k$ is $E_{kj2}/(\lambda_k + d\lambda_k)$. If $E_{kj} = E_{kj}(\mu_k) < E_{kj}(\lambda_k)$, then for a small enough $d\lambda_k$, $E_{kj}(\mu_k) < E_{kj}(\lambda_k + d\lambda_k)$ and the holding time is unaffected by the change. Note, however, that if $d\lambda_k > 0$ is fixed prior to sampling, then for some sample paths $E_{kj}(\mu_k) < E_{kj}(\lambda_k)$ but $E_{kj}(\mu_k) > E_{kj}(\lambda_k + d\lambda_k)$ in which case a change in jump occurs. This possibility is not taken into account by the IPA-like algorithm. Letting E'_{ij} denote the sample path derivative of the holding times with respect to a change in λ_k we obtain

$$E'_{kj} = \begin{cases} 0 & \text{if jump down,} \\ -E_{kj2}/\lambda_k^2 = -E_{kj}/\lambda_k & \text{if jump up,} \end{cases} \tag{5.1}$$

and $E'_{ij} = 0$ for $i \neq k$. Therefore

$$\mathrm{E}[E'_{kj}] = (\lambda_k/(\lambda_k + \mu_k))(-1/\lambda_k)\mathrm{E}[E_{kj} \,|\, \text{jump up}] = -1/(\lambda_k + \mu_k)^2. \tag{5.2}$$

Now consider a regenerative cycle that begins and ends in state 0. Let τ_n denote the length of the nth cycle, $N_n(i)$ denote the number of visits to state i during the nth cycle and let $\alpha_n(i)$ denote the amount of time spent in state i during the nth cycle. Then $\mathrm{E}[\tau_n] = (\sum_{j=0}^{N} p_j)/\lambda_0$, $\mathrm{E}[N_n(k)] = (\lambda_k + \mu_k)p_k/\lambda_0$, and $\mathrm{E}[\alpha_n(k)] = p_k/\lambda_0$. Let τ'_n and $\alpha'_n(i)$ denote the IPA-like sample path derivatives of τ_n and $\alpha_n(i)$, respectively, and let $\bar{\tau}_n$, $\bar{\alpha}_n(i)$, $\bar{\tau}'_n$, and $\bar{\alpha}'_n(i)$ denote the sample averages after n cycles. The fraction of time spent in state i during n cycles is $\hat{\pi}_n(i) = \bar{\alpha}_n(i)/\bar{\tau}_n$ so the IPA-like derivative estimate of π'_0 is

$$\hat{\pi}'_n(0) = \frac{\bar{\alpha}'_n(0)\bar{\tau}_n - \bar{\tau}'_n\bar{\alpha}_n(0)}{\bar{\tau}_n^2} = -\frac{\bar{\tau}'_n\bar{\alpha}_n(0)}{\bar{\tau}_n^2} \tag{5.3}$$

since $\bar{\alpha}'_n(0) = 0$. Therefore

$$\lim_{n\to\infty} \hat{\pi}'_n(0) = -\frac{\mathrm{E}[\tau'_n]\mathrm{E}[\alpha_n(0)]}{\mathrm{E}[\tau_n]^2}. \tag{5.4}$$

Using the fact that $\mathrm{E}[\tau'_n] = \mathrm{E}[N_n(k)]\mathrm{E}[E'_{kj}]$, it can be shown that

$$\lim_{n\to\infty} \hat{\pi}'_n(0) = \frac{\pi_0\pi_k}{\lambda_k + \mu_k} > 0. \tag{5.5}$$

Direct differentiation of π_0 with respect to λ_k yields

$$\frac{d}{d\lambda_k}\pi_0 = -\frac{\pi_0 \sum_{j=k+1}^{N} \pi_j}{\lambda_k} < 0. \tag{5.6}$$

Thus for this representation of the Birth and Death process, the IPA-like algorithm predicts that the amount of time spent in state 0 increases when λ_k increases when in fact it decreases.

Glasserman (1988) considers an alternative representation of the Birth and Death simulation: a standard IPA algorithm incorporating load dependent servers. He presents empirical results that indicate that this representation does produce strongly consistent estimates for steady state derivatives. Interestingly, although Glasserman's representation also does not consider the possibility of a change in event order, he discusses how the so-called nominal and perturbed sample paths are closer to one another in this representation than in the representation described above. This is consistent with Glynn (1987) who considers different process-differentiable representations of distributions and shows that some are smoother, in a certain sense, than others (see also Ho and Cao 1985 for a specific queueing network example of this). Glasserman's analysis appears to extend to Jackson queueing networks, although not to arbitrary continuous time Markov chains; this is also consistent with the results presented here and elsewhere on IPA convergence.

6. Conclusions

In this paper we have examined the convergence properties of Infinitesimal Perturbation Analysis derivative estimates. In regenerative processes, steady state performance measures take the form $r(\theta) = \mathrm{E}[Y_i(\theta)]/\mathrm{E}[\tau_i(\theta)]$ where $\mathrm{E}[Y_i(\theta)]$ is the expected cumulative reward earned during a regenerative cycle and $\mathrm{E}[\tau_i(\theta)]$ is the expected length of a cycle. For a class of regenerative processes, the IPA estimate converges to $r'(\theta)$, the derivative of the steady state performance measure, if and only if either the probability of a change in event order during a regenerative cycle is $o(d\theta)$, or $d_\theta(Y) = d_\theta(\tau) = 0$, or

$$r(\theta) = \frac{\mathrm{E}[Y_i(\theta)]}{\mathrm{E}[\tau_i(\theta)]} = \frac{d_\theta(Y)}{d_\theta(\tau)} \tag{6.1}$$

where $d_\theta(Y)$ and $d_\theta(\tau)$ are the limiting expected change in the cumulative reward and cycle length, respectively, given that a change in event order has occurred during a cycle. The terms $d_\theta(Y)$ and $d_\theta(\tau)$ are not estimated by IPA. It was shown that for the mean waiting time in the $M/G/1$ queue the probability of a change in event order cycle is $O(d\theta)$, not $o(d\theta)$. Furthermore, even though IPA does not estimate $d_\theta(Y)$ and $d_\theta(\tau)$, IPA consistently estimates $r'(\theta)$ in the $M/G/1$ queue because the term in $r'(\theta)$ involving $d_\theta(Y)$ and $d_\theta(\tau)$ vanishes when equation (6.1) holds. The intuitive explanation for this cancellation was given. However, this cancellation is not guaranteed to occur in all regenerative processes.

In the event that this cancellation does not occur, the IPA algorithm would have to be extended to estimate the limiting probability of a change in event order, $d_\theta(Y)$ and $d_\theta(\tau)$ in order to produce strongly consistent estimates for steady state derivatives. This has been done for the single server queue (Zazanis and Suri 1985a). Using the notion of conditional expectation, Gong and Ho (1987) have described an IPA extension for some specific simple queueing systems: certain performance measures in the $GI/G/1$ queue, the $GI/G/1$ queue with a capacity constraint of one and the network of Figure 1 with $s_2 = 0$ and $N_1 = N_2 = 1$.

General necessary conditions were then derived for multiple IPA throughput derivative estimates to simultaneously converge to the derivatives of steady state throughputs. If $\lambda_1(\theta)$ and $\lambda_2(\theta)$ are steady state throughputs and if the IPA estimates simultaneously converge to $\lambda_1'(\theta)$ and $\lambda_2'(\theta)$ for all θ in an interval, then there must exist a constant c such that

$$\frac{\lambda_1(\theta)}{\lambda_2(\theta)} = c \tag{6.2}$$

for all values of θ in the interval. This is a strong requirement and it means that, except in special cases, the original IPA algorithm cannot be used to obtain strongly consistent estimates of throughput derivatives in queueing networks with multiple types of customers, state-dependent routing or blocking. Numerical results confirming the theory were then given. Ho and Li (1988) have described a different IPA extension that produces single run finite difference approximations to steady state derivatives in more general systems such as irreducible continuous time Markov chains with a finite state space. This method explicitly accounts for changes in event order and has been found to be quite accurate in simulations of some simple queueing systems with multiple customer types such as the network of Figure 1. Its accuracy and efficiency in more complex queueing systems remain as open issues.

Thus the original IPA algorithm cannot be applied with guaranteed confidence except in those cases for which either equation (6.1) or (6.2) holds. Even if equation (6.2) holds, IPA is not guaranteed to converge since equation (6.2) is only a necessary condition. In addition, equation (6.2) is a condition for consistency for steady state throughput sensitivities and does not ensure consistency for sensitivities of other steady state performance measures. For example, equation (6.2) holds for the throughput in the $M/M/1$ queue with nonpreemptive priorities when θ is a service time parameter, but it was shown experimentally that the corresponding IPA estimates do not converge to derivatives of the mean stationary response times. On the other hand, if IPA is known to produce strongly consistent estimates in a given situation, then its mean squared error is asymptotically smaller than the optimal (multiple run) finite difference approximation to the derivative (see Zazanis and Suri 1985b) and the gradient information supplied by IPA is potentially useful in optimization of stochastic systems (see Suri and Leung 1987).

As indicated in Zazanis and Suri (1985a), Gong and Ho (1987), Cao (1987a) and Ho and Li (1988), work is in progress to extend IPA to handle more general situations.[1]

[1] Thanks are due to the anonymous referees for their many suggestions. In particular, we express our gratitude to the referee who suggested the use of indicator functions to bridge a gap in the original proof of Lemma 2.2. Additional thanks are due to Sheldon H. Jacobson for spotting a numerical error in one of the tables.

Appendix A. IPA for the $M/G/1$ Queue Revisited

For simplicity, we assume that θ is a scale parameter of the service time distribution (see Suri and Zazanis (1988) for more general parameters), i.e., the nth service time in the ith busy period can be written as $S_{in} = \theta X_{in}$ where $\mathrm{E}[X_{in}] = 1$ and θ is the mean service time. Let $\tau_i(\theta)$ denote the number of customers served in the ith busy period, $B_i(\theta) = \sum_{n=1}^{\tau_i(\theta)} S_{in}$ be the length of the ith busy period, I_i be the length of the ith idle period and let $B_i(\theta) = \theta C_i(\theta)$ where $C_i(\theta) = \sum_{n=1}^{\tau_i(\theta)} X_{in}$. By Wald's equation (see, e.g., Ross 1970) $\mathrm{E}[B_i(\theta)] = \theta \mathrm{E}[C_i(\theta)] = \theta \mathrm{E}[\tau_i(\theta)]$. Since the arrival process is Poisson, $\mathrm{E}[\tau_i(\theta)] = 1/(1 - \lambda\theta)$. Furthermore, since $\tau_i'(\theta) = 0$, when Lemma 2.2 is applied to $t(\theta) = \mathrm{E}[\tau_i(\theta)]$ we obtain

$$t'(\theta) = \frac{d}{d\theta} 1/(1 - \lambda\theta) = \lambda/(1 - \lambda\theta)^2 = p_\theta d_\theta(\tau). \tag{A.1}$$

Note that $B_i(\theta)$ is increased by $d\theta C_i(\theta)$ so that $p(d\theta) = \mathrm{P}\{d\theta C_1(\theta) \geq I_1\} = \int [1 - \exp(-\lambda d\theta c)] dF_\theta(c)$ where $F_\theta(c)$ is the distribution function of $C_i(\theta)$. Therefore $p(d\theta) = 1 - \hat{F}_\theta(\lambda d\theta)$ where $\hat{F}_\theta(s)$ is the Laplace-Stieltjes transform of $F_\theta(c)$. Since $\rho = \lambda\theta < 1$, $E[C_i(\theta)] < \infty$ and therefore $\hat{F}_\theta(s) = 1 - sE[C_i(\theta)] + o(s)$ (see, e.g., pp. 435–436 of Feller 1971). Thus

$$p(d\theta) = 1 - \hat{F}_\theta(\lambda d\theta) = \lambda d\theta \mathrm{E}[C_i(\theta)] + o(d\theta) = \lambda d\theta/(1 - \lambda\theta) + o(d\theta) \tag{A.2}$$

and therefore $p_\theta = \lim_{d\theta\to 0} p(d\theta)/d\theta = \lambda/(1 - \lambda\theta)$. Substituting this expression for p_θ into equation (A.1) proves that $\mathrm{E}[\tau_i(\theta)] = d_\theta(\tau)$.

We now consider $d_\theta(Y)$. As shown in the Introduction, $Y_i'(\theta) = \sum_{k=1}^{\tau_i(\theta)} \sum_{j=1}^{k} X_{ij}$. Since the arrival process is renewal, $\{(Y_i(\theta), Y_i'(\theta), \tau_i(\theta), C_i(\theta), I_i), i \geq 1\}$ is a sequence of iid vectors. Let θ be increased to $\theta + d\theta$. Note that if $d\theta C_1(\theta) > I_1$, then the first two busy periods come together. If this happens the first customer in busy period two still arrives at time $B_1(\theta) + I_1$, but enters service at time $B_1(\theta) + d\theta C_1(\theta)$ (the time at which the last customer in busy period one departs) rather than immediately. Thus this customer's response time is increased by an additional $d\theta C_1(\theta) - I_1$ due to the merging of busy periods. Furthermore, each of the $\tau_2(\theta)$ response times in the second busy period is increased by this same factor of $d\theta C_1(\theta) - I_1$ due to the busy period merging. Similarly, if busy periods one, two and three merge, then each of the $\tau_3(\theta)$ response times in the third busy period is increased by an additional factor of $(d\theta C_1(\theta) - I_1) + (d\theta C_2(\theta) - I_2)$. This generalizes so that if $n_1(d\theta)$ is the number of busy periods that come together, then (see Zazanis and Suri 1985a)

$$Y_1(\theta + d\theta) = \sum_{i=1}^{n_1(d\theta)} (Y_i(\theta) + d\theta Y_i'(\theta)) + 1_{\{n_1(d\theta)>1\}} \sum_{i=2}^{n_1(d\theta)} \tau_i(\theta) W_i(d\theta) \tag{A.3}$$

where $W_i(d\theta) = \sum_{j=1}^{i-1} (d\theta C_j(\theta) - I_j)$. Note that $n_1(d\theta) = \inf\{n: \sum_{i=1}^{n} (d\theta C_i(\theta) - I_i) \leq 0\}$ is the number of customers served in a busy period of an $M/G/1$ queue with service times $\{d\theta C_i(\theta), i \geq 1\}$ and interarrival times $\{I_i, i \geq 1\}$. This modified $M/G/1$ queue has traffic intensity $\lambda d\theta \mathrm{E}[C_i(\theta)] = \lambda d\theta/(1 - \lambda\theta)$. Furthermore $W_i(d\theta)$ is the waiting time of the ith customer in this modified queue, $i = 2, \ldots, n_1(d\theta)$. Since $n_1(d\theta)$ is a stopping time for $\{(Y_i(\theta), Y_i'(\theta), \tau_i(\theta), C_i(\theta), I_i), i \geq 1\}$,

$$\mathrm{E}[Y_1(\theta + d\theta)] = \mathrm{E}[n_1(d\theta)](\mathrm{E}[Y_1(\theta)] + d\theta \mathrm{E}[Y_1'(\theta)]) + h(d\theta) \qquad \text{where} \tag{A.4}$$

$$h(d\theta) = \mathrm{E}[1_{\{n_1(d\theta)>1\}} \sum_{i=2}^{n_1(d\theta)} \tau_i(\theta) W_i(d\theta)]. \tag{A.5}$$

Assuming that $h(d\theta) = o(d\theta)$ (as will be shown later), then

$$y'(\theta) = \lim_{d\theta\to 0} \left(\mathrm{E}[Y_1(\theta)] \frac{\mathrm{E}[n_1(d\theta)] - 1}{d\theta} + \mathrm{E}[Y_1'(\theta)] \mathrm{E}[n_1(d\theta)] + \frac{o(d\theta)}{d\theta} \right). \tag{A.6}$$

Since $\mathrm{E}[n_1(d\theta)] = 1/(1 - \lambda d\theta \mathrm{E}[C_1(\theta)])$ where $\mathrm{E}[C_1(\theta)] = 1/(1 - \lambda\theta)$, we obtain

$$y'(\theta) = \mathrm{E}[Y'_1(\theta)] + \frac{\lambda}{1 - \lambda\theta}\mathrm{E}[Y_1(\theta)]. \tag{A.7}$$

By Lemma 2.2, $y'(\theta) = \mathrm{E}[Y'_1(\theta)] + p_\theta d_\theta(Y)$ and since $p_\theta = \lambda/(1 - \lambda\theta)$, $\mathrm{E}[Y_1(\theta)] = d_\theta(Y)$. It remains to be shown that $h(d\theta) = o(d\theta)$. Define $W_1(d\theta) = 0$. Then by equation (A.5)

$$h(d\theta) = \sum_{i=1}^{\infty} \mathrm{E}[1_{\{n_1(d\theta)\geq i\}}\tau_i(\theta)W_i(d\theta)]. \tag{A.8}$$

Since $n_1(d\theta)$ is a stopping time, $\tau_i(\theta)$ is independent of $1_{\{n_1(d\theta)\geq i\}} = 1 - 1_{\{n_1(d\theta)\leq i-1\}}$. Furthermore $\tau_i(\theta)$ is independent of $W_i(d\theta)$ since $W_i(d\theta)$ only depends on events in busy periods $1, \ldots, i - 1$. Therefore

$$h(d\theta) = \mathrm{E}[\tau_1(\theta)] \sum_{i=1}^{\infty} \mathrm{E}[1_{\{n_1(d\theta)\geq i\}}W_i(d\theta)] = \mathrm{E}[\tau_1(\theta)]\mathrm{E}[\sum_{i=1}^{n_1(d\theta)} W_i(d\theta)]. \tag{A.9}$$

But $\mathrm{E}[\sum_{i=1}^{n_1(d\theta)} W_i(d\theta)] = \mathrm{E}[W(d\theta)]\mathrm{E}[n_1(d\theta)]$ where $\mathrm{E}[W(d\theta)]$ is the expected stationary waiting time in the modified $M/G/1$ queue. By the Pollaczek-Khinchin formula (see, e.g., Kleinrock 1975)

$$\mathrm{E}[W(d\theta)] = \frac{\lambda(d\theta)^2(\mathrm{E}[C_i(\theta)]^2 + \mathrm{Var}\,[C_i(\theta)])}{2(1 - \lambda d\theta \mathrm{E}[C_i(\theta)])} = o(d\theta) \tag{A.10}$$

provided $\mathrm{Var}\,[C_i(\theta)] < \infty$ and therefore $h(d\theta) = o(d\theta)$.

A similar result for a class of $GI/G/1$ queues, derived under somewhat different conditions, has also been obtained in Zazanis and Suri (1985b). It appears that the argument presented here can be extended using Taylor series expansions to show that IPA is strongly consistent for

$$r'(\theta) = \frac{d}{d\theta}\mathrm{E}[f(R(\theta))]$$

where f is a differentiable function and $R(\theta)$ is the stationary response time in the $M/G/1$ queue.

References

BASKETT, F., K. M. CHANDY, R. R. MUNTZ AND F. G. PALACIOS, "Open, Closed, and Mixed Networks of Queues with Different Classes of Customers," *J. Assoc. Comput. Mach.*, 22 (1975), 248–260.

CAO, X. R., "A Sample Performance Function of Closed Jackson Queueing Networks," *Oper. Res.*, 36 (1988), 128–136.

———, "Convergence of Parameter Sensitivity Estimates in a Stochastic Experiment," *IEEE Trans. Automatic Control*, AC-30 (1985), 845–853.

———, "First-order Perturbation Analysis of a Simple Multi-Class Finite Source Queue," *Performance Evaluation*, 7 (1987a), 31–41.

———, "The Static Property of Perturbed Queueing Networks," *Proc. 26th IEEE Conf. Decision and Control*, Los Angeles, CA, 1987b, 257–262.

———, "Realization Probability in Closed Jackson Queueing Networks and Its Application," *Adv. in Appl. Probab.*, 19 (1987c), 708–738.

——— AND Y. C. HO, "Perturbation Analysis of Sojourn Time in Queueing Networks," *Proc. 22nd IEEE Conf. Decision and Control*, San Antonio, TX, 1983, 1025–1029.

——— AND ———, "Perturbation Analysis of Sojourn Times in Closed Jackson Queueing Networks." Technical Report, Division of Applied Sciences, Harvard University, Cambridge, MA, 1986.

——— AND ———, "Estimating the Sojourn Time Sensitivity in Queueing Networks using Perturbation Analysis," *J. Optim. Theory Appl.*, 53 (1987), 353–375.

CRANE, M. A. AND D. L. IGLEHART, "Simulating Stable Stochastic Systems. III. Regenerative Processes and Discrete-Event Simulations,"*Oper. Res.*, 23 (1975), 33–45.

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

GLASSERMAN, P. "Infinitesimal Perturbation Analysis of a Birth and Death Process," *Oper. Res. Lett.*, 7 (1988), 43–49.

GLYNN, P. W. "On the role of Generalized Semi-Markov Processes in Simulation Output Analysis," *Proc.* 1983 *Winter Simulation Conf.*, Volume 1, S. Roberts, J. Banks, B. Schmeiser (Eds.), IEEE Press, 1983.

———, "Stochastic Approximation for Monte Carlo Optimization," *Proc.* 1986 *Winter Simulation Conf.*, J. Wilson, J. Henriksen, S. Roberts (Eds.), IEEE Press, 1986, 356–365.

———, "Construction of Process-Differentiable Representations for Parametric Families of Distributions," Technical Report, Mathematics Research Center, University of Wisconsin-Madison, 610 Walnut Street, Madison, 1987.

——— AND J. L. SANDERS, "Monte Carlo Optimization of Stochastic Systems: Two New Approaches," *Proc.* 1986 *ASME Computers in Engineering Conf.*, 1986.

GONG, W. B. AND Y. C. HO, "Smoothed (Conditional) Perturbation Analysis of Discrete Event Dynamic Systems," *IEEE Trans. Automatic Control*, AC-32 (1987), 856–866.

HALTON, J. H., "A Retrospective and Prospective Survey of the Monte Carlo Method," *SIAM Rev.*, 12 (1970), 1–60.

HAMMERSLEY, J. M. AND D. C. HANDSCOMB, *Monte Carlo Methods*, Methuen, London, 1964.

HO, Y. C. AND X. R. CAO, "Perturbation Analysis and Optimization of Queueing Networks," *J. Optim. Theory Appl.*, 40 (1983), 559–582.

——— AND ———, "Performance Sensitivity to Routing Changes in Queueing Networks and Flexible Manufacturing Systems using Perturbation Analysis," *IEEE J. Robotics and Automation*, RA-1 (1985), 165–172.

———, ——— AND C. G. CASSANDRAS, "Infinitesimal and Finite Perturbation Analysis for Queueing Networks." *Automatica*, 4 (1983), 439–445.

———, M. A. EYLER AND T. T. CHIEN, "A New Approach to Determine Parameter Sensitivities of Transfer Lines," *Management Sci.*, 29 (1983), 700–714.

——— AND S. LI, "Extensions of Infinitesimal Perturbation Analysis," *IEEE Trans. Automatic Control,* AC-33 (1988), 427–438.

———, R. SURI, X. R. CAO, G. W. DIEHL, J. W. DILLE AND M. ZAZANIS, "Perturbation Analysis and Optimization of Large Multiclass (Non-Product-Form) Queueing Networks Using Perturbation Analysis," *Large Scale Systems*, 7 (1984), 165–180.

KLEINROCK, L., *Queueing Systems. Volume* 1. *Theory*, John Wiley and Sons, New York, 1975.

REIMAN, M. I. AND A. WEISS, "Sensitivity Analysis via Likelihood Ratios," *Proc.* 1986 *Winter Simulation Conf.*, J. Wilson, J. Henriksen, S. Roberts (Eds.), IEEE Press, 1986, 285–289.

REISER, M. AND S. S. LAVENBERG, "Mean-Value Analysis of Closed Multichain Queueing Networks," *J. Assoc. Comput. Mach.*, 27 (1980), 314–322.

ROSS, S. M. *Applied Probability Models with Optimization Applications*, Holden-Day, San Francisco, 1970.

RUBINSTEIN, R. Y., "Sensitivity Analysis and Performance Extrapolation for Computer Simulation Models," Technical Report, Division of Applied Sciences, Harvard University, Cambridge, MA, 1986.

SAUER, C. H. AND E. A. MACNAIR, "Queueing Network Software for Systems Modelling," *Software-Practice and Experience*, 9 (1979), 369–380.

STRELEN, J., "A Generalization of Mean Value Analysis to Higher Moments," *Performance Evaluation Rev.*, 14 (1986), *Proc. Performance* '86 *and ACM Sigmetrics* 1986, 129–140.

SURI, R., "Infinitesimal Perturbation Analysis of Discrete Event Dynamic Systems: A General Theory," *Proc.* 22*nd IEEE Conf. Decision and Control*, San Antonio, TX, 1983, 1030–1039.

———, "Infinitesimal Perturbation Analysis for General Discrete Event Systems," *J. Assoc. Comput. Mach.*, 34 (1987), 686–717.

——— AND Y. T. LEUNG, "Single Run Optimization of a SIMAN Model for Closed Loop Flexible Assembly Systems," *Proc.* 1987 *Winter Simulation Conf.*, A. Thesen, H. Grant and W. D. Kelton (Eds.), IEEE Press, 1987, 738–748.

——— AND M. A. ZAZANIS, "Perturbation Analysis Gives Strongly Consistent Sensitivity Estimates for the $M/G/1$ Queue," *Management Sci.*, 34 (January 1988), 39–64.

WHITT, W., "Continuity of Generalized Semi-Markov Processes," *Math. Oper. Res.*, 5 (1980), 494–501.

WOODSIDE, C. M., "Response Time Sensitivity Measurement for Computer Systems and General Closed Queueing Networks," *Performance Evaluation*, 4 (1984), 199–210.

ZAZANIS, M. A., "Unbiasedness of Perturbation Analysis Estimates for Higher Moments of the Response Time of an $M/M/1$ Queue," Technical Report, Division of Applied Sciences, Harvard University, Cambridge, MA, 1986.

——— AND R. SURI, "Estimating First and Second Derivatives of Response Time for $G/G/1$ Queues from a Single Sample Path," Technical Report, Division of Applied Sciences, Harvard University, Cambridge, MA, 1985a.

——— AND ———, "Comparison of Perturbation Analysis with Conventional Sensitivity Estimates for Regenerative Stochastic Systems," Technical Report, Division of Applied Sciences, Harvard University, Cambridge, MA (Revision date June, 1986), 1985b.