MANAGEMENT SCIENCE
Vol. 34, No. 1, January 1988
Printed in U.S.A.

PERTURBATION ANALYSIS GIVES STRONGLY CONSISTENT SENSITIVITY ESTIMATES FOR THE $M/G/1$ QUEUE*

RAJAN SURI AND MICHAEL A. ZAZANIS
Department of Industrial Engineering, University of Wisconsin, Madison, Wisconsin 53706
Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, Illinois 60201

The technique of perturbation analysis has recently been introduced as an efficient way to compute parameter sensitivities for discrete event systems. Thus far, the statistical properties of perturbation analysis have been validated mainly through experiments. This paper considers, for an $M/G/1$ queueing system, the sensitivity of mean system time of a customer to a parameter of the arrival or service distribution. It shows analytically that (i) the steady state value of the perturbation analysis estimate of this sensitivity is unbiased, and (ii) a perturbation analysis algorithm implemented on a *single* sample path of the system gives asymptotically unbiased and strongly consistent estimates of this sensitivity. (No previous knowledge of perturbation analysis is assumed, so the paper also serves to introduce this technique to the unfamiliar reader.) Numerical extensions to $GI/G/1$ queues, and applications to optimization problems, are also illustrated.
(DISCRETE EVENT SYSTEMS; SIMULATION; QUEUEING SYSTEMS; SAMPLE PATH ANALYSIS; STOCHASTIC SYSTEMS)

1. Introduction

The perturbation analysis technique has recently been developed as an effective method for sensitivity analysis of complex discrete event systems. It enables the sensitivity of a performance measure to be calculated while observing a *single* sample path of the system. It therefore offers computational savings for computer simulations, and it also has the ability to be applied directly on actual systems. Thus far, much of the research on perturbation analysis has focused on experimental results demonstrating its accuracy for various complex systems (further details and references follow below). It is clear that for a practically useful new approach such as this, it is also important to study its theoretical properties. As a step in this direction we study here perturbation analysis applied to an $M/G/1$ queue. While this is a simple and analytically well understood system from the viewpoint of queueing theory, it nevertheless is nontrivial, and provides a good test case for perturbation analysis techniques. Only by understanding the behavior of perturbation analysis for simpler systems can we hope to study its properties for more complex cases.

The current paper requires no previous knowledge of perturbation analysis. Indeed, it is written so as to give the reader an introduction to this technique as well as present the new results. An overview of our result is as follows. Let $y(\omega, \theta)$ be some performance measure of a discrete event system, where θ is a decision parameter and ω denotes the outcome of various random events (formalized later). Examples would be y = throughput of a system, θ = mean service time of a server in the system, ω = the values of actual interarrival and service times that occur during a sample observation on the system. Let $\bar{y}(\theta) = E[y(\omega, \theta)]$ where $E(\cdot)$ denotes expectation w.r.t. ω. We assume that analytic expressions are not available for $\bar{y}(\theta)$, and the system designer must use Monte Carlo experiments to study the system. In system design we are often interested in the sensitivity of $\bar{y}(\theta)$ to θ, i.e. the value of $d\bar{y}/d\theta$, so that we may optimize the performance with

* Accepted by George Fishman; received September 1984. This paper has been with the authors 15 months for 4 revisions.

0025-1909/88/3401/0039$01.25

respect to θ. Conventionally, this might be estimated by a quantity such as $[y(\omega', \theta + \Delta\theta) - y(\omega, \theta)]/\Delta\theta$, where ω and ω' may or may not be the same (depending on experimental design). This involves two Monte Carlo experiments. The perturbation analysis approach shows that an alternative statistic, $g(\omega, \theta)$ defined by

$$g(\omega, \theta) = \lim_{\Delta\theta \to 0} \frac{y(\omega, \theta + \Delta\theta) - y(\omega, \theta)}{\Delta\theta}$$

can be obtained from a *single* Monte Carlo experiment, the same "nominal" experiment used to obtain $y(\omega, \theta)$. The question then arises, is $g(\omega, \theta)$ an unbiased estimate of $d\bar{y}/d\theta$, i.e. does $E[g(\omega, \theta)] = d\bar{y}/d\theta$? Note that $d\bar{y}/d\theta$ involves expectation first, then differentiation, while $E[g(\omega, \theta)]$ involves the converse. While apparently a simple question in conditions for changing the order of operations, this turns out to be quite hard to answer for general discrete event systems. Here we consider an $M/G/1$ queue, with λ the parameter of the (Poisson) arrival rate, and θ a parameter of the (general) service time distribution. The performance measure is taken to be $\bar{T}(\lambda, \theta)$ = steady state value of mean time spent in the system by a customer. We show analytically that a perturbation analysis algorithm, implemented on a single sample path of such a queue, gives a strongly consistent and asymptotically unbiased estimate of $d\bar{T}/d\theta$. The same is proved also for $d\bar{T}/d\lambda$.

This result is of interest because first, it shows that the reversal of operations above is valid for this system. Second, researchers working on perturbation analysis have encountered frequent criticism that their algorithms are derived by using assumptions that are not valid for sufficiently long sample paths. (This criticism will become clearer below.) This paper reassures us that, at least for the system under study here, perturbation analysis is indeed valid. We hope that similar results will become available for more complex systems in due course.

Lastly, this paper also illustrates applications to $GI/G/1$ queues, in particular to parameter optimization for these systems. It is seen that perturbation analysis provides an interesting optimization ability, as well as computational savings, for such cases.

2. Background on Perturbation Analysis

The perturbation analysis approach has its origins in a paper by Ho et al. (1979) for buffer storage optimization in a production line. Since then, considerable progress has been made in formalizing as well as extending the basic concepts. It is useful to classify the perturbation analysis research into two categories: *infinitesimal* and *finite* perturbation analysis. While the original paper above, and many other applications (Ho et al. 1983, Suri and Cao 1982, 1983), have produced useful experimental results using finite perturbation analysis, theoretical understanding of these algorithms is still lacking. In contrast, the theory of infinitesimal perturbation analysis is much better developed. Ho et al. (1983) and Ho and Cassandras (1983) derive algorithms for throughput in production lines (tandem queues), while Ho and Cao (1983) and Cassandras and Ho (1984) derive corresponding algorithms for throughput in queueing networks with general service times and finite buffers. Suri and Dille (1985) apply these algorithms to flexible manufacturing systems (FMS). Suri (1987) uses infinitesimal perturbation analysis to derive an algorithm for a *general* performance measure and *general* discrete event system. All these papers show that the sensitivity $g(\omega, \theta)$ *as defined above*, can be computed exactly from one sample path (i.e. one observation on the system). However, the question of equality of $E[g(\omega, \theta)]$ and $d\bar{y}/d\theta$ is not addressed there, except via experimental results.

Two earlier papers contain results related to our work here. Cao (1985a) has given conditions under which perturbation analysis gives unbiased estimates of parameter

sensitivity. Qualitatively, his results are illuminating, but some of the technical conditions are hard to interpret for practical systems. Ho and Cao (1983) showed that perturbation analysis is exact (in the expected value sense) for a certain restricted class of queueing networks, specifically, closed queueing networks with M servers that have identical service rates, exponentially distributed service times, and unlimited buffer space, and the customers have equal probability of going to any server next. This could be considered the first proof that perturbation analysis works for a "classic" discrete event system. Although reassuring in its results, a possible criticism might be that some potential error effects in the perturbation analysis estimates cancel out due to the symmetry of the system. The current paper considers another "classic" discrete event system, simple yet nontrivial. Certainly no criticism on the basis of symmetry is possible here, and we show conclusively that the steady state values of the perturbation analysis estimates of $d\bar{T}/d\theta$ and $d\bar{T}/d\lambda$ are both unbiased.

Since the original writing of this manuscript (1984), Heidelberger et al. (1987) have also given some results on the consistency of infinitesimal perturbation analysis for regenerative systems. However, his conditions and assumptions resemble those of Cao (1985a) discussed above and are equally hard to interpret. Also, Heidelberger et al. (1987) only consider infinitesimal perturbation analysis. The limitations reported in that paper can be circumvented by means of more sophisticated techniques. For example Ho and Li (1987) report such a sophisticated perturbation analysis algorithm that gives consistent gradient estimates for a network for which Heidelberger et al. (1987) show that infinitesimal perturbation analysis estimates are biased. Thus, determining the class of systems for which consistent perturbation analysis algorithms can be derived, remains an open question. The current paper represents one of the first steps in verifying consistency for a given class of systems.

3. Definitions

3.1. *System Model*

Consider the $M/G/1$ queueing system (Kleinrock 1975) with the following notation. The arrival process is Poisson with rate λ. We use the dummy variable x to represent a particular value of the service time. The (cumulative) service time distribution is $F(x, \theta)$. Note that the service time distribution depends on a parameter θ. The idea here is that θ is a decision parameter that could be chosen by a designer/operator of the system. In general θ could be a vector. Also there are some restrictions on θ. These points will be elaborated later. The first two moments of this distribution are $\bar{x}$ and $\overline{x^2}$ (which therefore, also depend on θ). We are interested in the mean system time of a customer, in steady state, given by the Pollaczek-Khinchin, or P-K, formula (Kleinrock 1975)

$$\bar{T}(\lambda, \theta) = \bar{x} + \frac{\lambda \overline{x^2}}{2(1 - \lambda \bar{x})}. \tag{3.1}$$

The sensitivity of this mean system time to θ is then obtained by straightforward differentiation to be

$$\frac{d\bar{T}}{d\theta} = \frac{d\bar{x}}{d\theta} + \frac{\lambda}{2(1 - \lambda\bar{x})} \frac{d\overline{x^2}}{d\theta} + \frac{\lambda^2 \overline{x^2}}{2(1 - \lambda\bar{x})^2} \frac{d\bar{x}}{d\theta}. \tag{3.2}$$

Similarly, the sensitivity with respect to λ is

$$\frac{d\bar{T}}{d\lambda} = \frac{\overline{x^2}}{2(1 - \lambda\bar{x})^2}. \tag{3.3}$$

These are the two sensitivities that will be estimated in an alternative way using perturbation analysis.

We assume that $\bar{x}$, $\overline{x^2}$, $d\bar{x}/d\theta$, $d\overline{x^2}/d\theta$ all exist and are finite at the given value of θ. We will also need to define the inverse mapping (dependent on θ)

$$F_\theta^{-1}(u) = \inf\{x: F(x, \theta) \geq u\} \qquad \text{for} \qquad 0 \leq u \leq 1. \tag{3.4}$$

The use of this mapping will become apparent below.

Throughout the paper, we use the notation C_k to denote the kth customer.

3.2. *Nominal and Perturbed Sample Paths*

Let $x(\omega, \theta)$ be the value of the service time for a particular customer. For clarity of later analysis, it is necessary to define precisely what we mean by $x(\omega, \theta + \Delta\theta)$, i.e. the "same realization" but for a random variable with different parameters. This section is based on some concepts in Suri (1983), and should be intuitively obvious to anyone familiar with discrete event simulation. However, since perturbation analysis applies equally to actual systems as to computer simulations, Suri (1983) justifies these concepts for experiments on actual systems as well.

Since we are concerned here only with the $M/G/1$ queue, we will specialize our definitions for this case. Formally, we define a realization ω to be a pair of infinite sequences

$$\omega = [u_{11}, u_{12}, \ldots; u_{21}, u_{22}, \cdots] \tag{3.5}$$

where each u_{ij} is an independent realization of a random variable (r.v.) uniformly distributed in [0, 1]. For a given value of θ, a map from a value of ω to a sample path of the $M/G/1$ queue is obtained by the following correspondence. For customer C_j, $(-1/\lambda)\log(u_{1j})$ is the value for the inter-arrival time, and $F_\theta^{-1}(u_{2j})$ the value for the service time. It is well known that this procedure gives rise to the correct distributions, e.g. see Fishman (1978). We call the sample path so generated the *nominal* sample path. A *perturbed* sample path, obtained when the parameter is $\theta + \Delta\theta$, would then simply be generated by letting the service time of C_j be $F_{\theta+\Delta\theta}^{-1}(u_{2j})$ instead. Similarly, another perturbed sample path, for the case where the arrival parameter is $\lambda + \Delta\lambda$, is generated by letting the interarrival time of C_j be $[-1/(\lambda + \Delta\lambda)]\log(u_{1j})$. The point to be noted is that this procedure separates the realization (which is now independent of the parameter values) from the (parameter dependent) sample paths of the queueing system (see Suri 1983).

3.3. *Performance Measures*

Formally, a performance measure is a real-valued map on the detailed sample path (Suri 1987). Various performance measures can be defined for the system being studied. In the present case we deal only with the average time spent in the system by an arriving customer, in steady state. Use of Little's Law (Kleinrock 1975) allows this to be extended to expected queue lengths as well. Other performance measures are discussed in Suri (1987).

We consider first, a finite time sample path, consisting of M busy periods. (A busy period is the period between successive times when the system is empty.) Suppose N customers pass through the system during this time. Their interarrival times and service times are given via the first N values of each of the two infinite sequences in (3.5), using the mappings as above. For these values of arrival and service times, we simulate the system and observe s_i = system time of C_i (time from arrival to departure). We define our performance measure to be

$$T(\omega, \lambda, \theta, M) = \frac{1}{N}\sum_{i=1}^{N} s_i. \tag{3.6}$$

Note that N depends implicitly upon the evolution of the queue, and on M, i.e. we should write $N(\omega, \lambda, \theta, M)$ to be rigorous.

We assume that the system is ergodic (implied by $\lambda\bar{x} < 1$), so that w.p. 1

$$\lim_{M\to\infty} T(\omega, \lambda, \theta, M) = \bar{T}(\lambda, \theta), \tag{3.7}$$

where the RHS is the steady state value, which no longer depends on ω. In practice, we can only observe finite sample paths, so we use (3.6) with large M to estimate the RHS of (3.7). As discussed later, using other well-known techniques we can also estimate the variance of our estimator for a given observation.

4. Stochastic Derivatives and Service Time Perturbations

As we saw in the previous section, according to the perturbation analysis model, a change $\Delta\theta$ in a parameter θ of the service time distribution results in perturbation in the service time of C_j equal to

$$x_j(\omega, \theta + \Delta\theta) - x_j(\omega, \theta) = F^{-1}_{\theta+\Delta\theta}(u_{2j}) - F^{-1}_{\theta}(u_{2j}). \tag{4.1}$$

A necessary assumption for infinitesimal perturbation analysis is that small changes in θ introduce small changes in the r.v.'s of interest to the system (Ho and Cassandras 1983, Ho and Cao 1983, Suri 1987). Here we will be thinking of $\Delta\theta$ as vanishingly small and hence, we will examine the limit

$$\lim_{\Delta\theta\to 0} \frac{1}{\Delta\theta}[x_j(\omega, \theta + \Delta\theta) - x_j(\omega, \theta)] = \frac{dx_j}{d\theta}. \tag{4.2}$$

To this extent we will make the following

Assumption A1. $F^{-1}_{\theta}(u_{2j})$ is a differentiable function of θ for almost all u_{2j}.

Under A1, according to the perturbation analysis model (as expressed in (4.1)) the derivative w.r.t. θ of the service time of the jth customer is a well defined r.v. given by

$$\frac{dx_j}{d\theta} = \lim_{\Delta\theta\to 0} \frac{1}{\Delta\theta}[F^{-1}_{\theta+\Delta\theta}(u_{2j}) - F^{-1}_{\theta}(u_{2j})] \qquad \text{a.e.} \tag{4.3}$$

(For those values of u_{2j} where $F^{-1}_{\theta}(u_{2j})$ is not differentiable we can arbitrarily ascribe to $dx_j/d\theta$ the value zero. We need not worry about them since by A1 they constitute a set of measure zero.)

In (4.3), $dx_j/d\theta$ is defined as a function of the random variable u_{2j} which we will not assume to be an observable quantity even though this may be the case when we simulate the system (Suri 1983). We would like to express $dx_j/d\theta$ as a function of x_j which *is* an observable quantity. So we will introduce the following

Assumption A2. The derivative $dx_j/d\theta$ as defined in (4.3) depends only on the value of the service time x_j, i.e. it can be put in the form

$$\frac{dx_j}{d\theta} = \phi(x_j) \tag{4.4}$$

where $\phi(\ )$ is a deterministic function that may depend on θ.

The introduction of this assumption does not constitute a serious restriction to the range of applicability of the method since a wide class of parameters and distributions satisfy it. In particular, as we show in Theorem A at the end of this section, A2 is satisfied whenever θ is a location or a scale parameter of any distribution.

As another example, consider the class of distributions $F(x, \theta)$ which are strictly increasing in some interval $[a(\theta), b(\theta)]$, where $a(\theta) \geq 0$, with $F(a(\theta), \theta) = 0$ and $F(b(\theta), \theta) = 1$ (or in some interval $[a(\theta), \infty)$ with $F(a(\theta), \theta) = 0$) and for which the limit in (4.3) exists. This class of functions satisfies A2 because then, as it can be easily seen, the equation $u_{2j} = F(x_j, \theta)$ specifies a unique value of u_{2j} from a given x_j. As a result, (4.3) can be written as

$$\frac{dx_j}{d\theta} = \lim_{\Delta\theta \to 0} \frac{1}{\Delta\theta} [F^{-1}_{\theta+\Delta\theta}(F(x_j, \theta)) - x_j] \tag{4.5}$$

which is a relation of the form (4.4).

A case of particular interest is the one where $F(x, \theta)$ is a differentiable function of both x and θ for almost all x. According to the perturbation analysis model, for any given $\Delta\theta$ the nominal and the perturbed values of service times must obey

$$F(x_j(\omega, \theta + \Delta\theta), \theta + \Delta\theta) = F(x_j(\omega, \theta), \theta) \tag{4.6}$$

since they both equal u_{2j}. Using the formula for differentiating implicitly defined functions, we have

$$\frac{dx_j}{d\theta} = -\left.\frac{\partial F/\partial\theta}{\partial F/\partial x}\right|_{(x_j,\theta)} \tag{4.7}$$

provided that $\partial F/\partial x(x_j, \theta) \neq 0$. (Notice that (4.7) is the value of the limit in (4.3) when $\partial F/\partial x$ and $\partial F/\partial\theta$ exist and are continuous. Also notice that in this case A2 is satisfied since (4.7) defines $dx_j/d\theta$ as a function of x_j). Although (4.7) is very useful when dealing with "smooth" distribution functions, sometimes only (4.3) is applicable as is shown in the following

EXAMPLE 4.1. Consider the deterministic distribution

$$F(x, \theta) = \begin{cases} 0 & \text{if} \quad x < \theta, \\ 1 & \text{if} \quad x \geq \theta. \end{cases}$$

(In all the examples we will use x to denote x_j and u to denote u_{2j}.) Obviously neither (4.7) nor (4.5) is applicable. ((4.5) fails because the equation $u = F(x, \theta)$ does not specify a unique u for a given x.) However it is easy to see that the inverse mapping is given by $x = F^{-1}_{\theta}(u) = \theta$ for all u, which is obviously differentiable w.r.t. θ. Hence by (4.3), $dx_j/d\theta = 1$.

Let us finally introduce one additional assumption concerning some statistics of the r.v. $dx_j/d\theta$ defined above.

Assumption A3. The random variables $dx_j/d\theta$ and $dx_j^2/d\theta$ are integrable and satisfy the relations:

(i) $E(dx_j/d\theta) = dE(x_j)/d\theta$,

(ii) $E(dx_j^2/d\theta) = dE(x_j^2)/d\theta$,

(iii) $E(dx_j/d\theta)^2 < \infty$.

The validity of the interchange of expectation and differentiation which is assumed in A3 can be verified for many distributions used in practice, e.g. exponential, uniform, deterministic, etc. It was also pointed out to us by a referee that this interchange can be ensured whenever $\partial F/\partial\theta$ exists by the following condition: (e.g. see Bickel and Doksum 1977)

$$\int_0^\infty |\partial F/\partial\theta(x, \theta)|dx \qquad \text{and} \qquad \int_0^\infty x|\partial F/\partial\theta(x, \theta)|dx$$

must be continuous functions of θ for $\theta \in \Theta$ where Θ is the parameter space. This test however fails whenever the density $\partial F/\partial\theta$ does not exist as is the case for the deterministic distribution. In these cases the more direct approach described in Examples 4.1 through 4.4 can be very effective. For a further discussion of the mathematical questions arising from our definition of $dx/d\theta$ in (4.3) the reader is referred to Glynn (1987).

The above three assumptions, A1, A2, and A3, are necessary for infinitesimal perturbation analysis of the $M/G/1$ queue.

DEFINITION (also see Suri (1987)). A parameter θ is called *admissible* if $F(x, \theta)$ satisfies A1, A2, and A3.

A number of instances of admissible parameters for many common distributions are given below. We end this section by giving some examples illustrating Assumptions A1 through A3.

EXAMPLE 4.2: Exponential Distribution. Let θ be the mean of the distribution. Then $u = F(x, \theta) = 1 - e^{-x/\theta}$ $(x \geq 0)$. Hence $F_\theta^{-1}(u) = -\theta \log (1 - u)$, and

$$\frac{d}{d\theta} F_\theta^{-1}(u) = -\log (1 - u) = -\log e^{-x/\theta} = \frac{x}{\theta},$$

so A1 and A2 are satisfied at any given $\theta > 0$. (We can also see that (4.7) applies.) Now let us illustrate A3. By definition, $dx_j/d\theta = dF_\theta^{-1}(u)/d\theta$ and so $dx(\omega, \theta)/d\theta = x/\theta$. Now $E[dx(\omega, \theta)/d\theta] = E[x/\theta] = 1 = d\bar{x}/d\theta$, which illustrates the first part of A3. Next, $E[dx^2/d\theta] = E[2xdx/d\theta] = 2E[x(x/\theta)] = 2\overline{x^2}/\theta = 4\theta$, using the fact that $\overline{x^2} = 2\theta^2$ for the exponential distribution. Also from this fact we have $d\overline{x^2}/d\theta = 4\theta$, and thus the second part of the assumption is verified. Finally, $E(dx/d\theta)^2 = E[x^2/\theta^2] = \overline{x^2}/\theta^2 = 2 < \infty$. Thus any $\theta > 0$ is admissible.

EXAMPLE 4.3: Uniform Distribution. Let x be uniformly distributed in $[\theta - \delta, \theta + \delta]$. Here we have two parameters, the mean θ and the spread δ, with $F(x, \theta, \delta) = \frac{1}{2} + (x - \theta)/2\delta$ for $\theta - \delta \leq x \leq \theta + \delta$ (F is identically zero below this range, unity above it). So $F_{\theta,\delta}^{-1}(u) = \theta + (2u - 1)\delta$ and A1 is satisfied for both parameters. A2 is also satisfied since $F(x, \theta, \delta)$ is continuous in x. Moreover, (4.7) applies in $(\theta - \delta, \theta + \delta)$, as $F(x, \theta, \delta)$ is clearly well behaved in this range. Finally, one can verify easily A3 proving that both θ and δ are admissible.

EXAMPLE 4.4: A Discrete Distribution. Let this have probability mass p at θ_1 and $1 - p$ at θ_2. Now there are *three* parameters, $\theta = (\theta_1, \theta_2, p)$, and

$$F(x, \theta) = \begin{cases} 0 & \text{if} \quad x < \theta_1, \\ p & \text{if} \quad \theta_1 \leq x < \theta_2, \\ 1 & \text{if} \quad \theta_2 \leq x. \end{cases}$$

Accordingly, $F_\theta^{-1}(u)$ is θ_1 if $u \leq p$ and θ_2 if $u > p$. Here, as in the case of the deterministic service time, we can see that θ_1 and θ_2 both satisfy A1, A2 and A3 and hence are admissible. Let us verify, for instance, A3 for θ_1: $dx/d\theta_1$ is 1 if $x = \theta_1$ and 0 if $x = \theta_2$ and thus $E[dx/d\theta_1] = p = dE[x]/d\theta_1$. Also,

$$E\left[\frac{dx^2}{d\theta_1}\right] = 2E\left[x \frac{dx}{d\theta_1}\right] = 2p\theta_1 = \frac{d}{d\theta_1} [p\theta_1^2 + (1 - p)\theta_2^2] = \frac{d}{d\theta_1} E[x^2].$$

Now consider the parameter p. We see that F_θ^{-1} is discontinuous only at the point $u = p$. Hence it is differentiable almost everywhere with derivative $dx/dp = 0$ and A1 is satisfied. However it is very easy to check that A3 is violated (notice that $d\bar{x}/dp \neq 0 = E[dx/dp]$), hence p is not admissible. This example thereby illustrates a common distribution which has both admissible and non-admissible parameters.

We now close this section with a theorem showing that a large class of parameters interesting in applications are admissible.

THEOREM A. *Location and scale parameters of distributions which have finite second moments are admissible.*

PROOF. We start with the case where θ is a location parameter. Then $F(x, \theta) = F_1(x - \theta)$ and using (3.4),

$$\begin{aligned} F_\theta^{-1}(u) &= \inf \{x: F_1(x - \theta) \geq u\} = \theta + \inf \{x - \theta: F_1(x - \theta) \geq u\} \\ &= \theta + \inf \{y: F_1(y) \geq u\} = \theta + l_1(u) \end{aligned} \tag{4.8}$$

where $l_1(u)$ does not depend on θ. From the definition in (4.3) we have

$$\lim_{\Delta\theta \to 0} \frac{1}{\Delta\theta} [F^{-1}_{\theta+\Delta\theta}(u) - F^{-1}_{\theta}(u)] = 1,$$

$$\frac{dx}{d\theta} = \phi(x) = 1. \tag{4.9}$$

Thus, A1 and A2 are satisfied. One can easily see that A3 is satisfied as well. Notice that since θ is location parameter, $E[x] = \theta + c_1$ and $E[x^2] = \theta^2 + 2\theta c_1 + c_2$ where c_1 and c_2 are constants not depending on θ. Thus $dE[x]/d\theta = 1 = E[dx/d\theta]$. Also,

$$E\left[\frac{dx^2}{d\theta}\right] = E\left[2x\frac{dx}{d\theta}\right] = 2E[x] = 2(\theta + c_1) = \frac{d}{d\theta} E[x^2].$$

Finally, since $dx/d\theta$ is a constant, the third part of A3 is trivially satisfied. This concludes the proof for the case of a location parameter.

The case where θ is a scale parameter of $F(x, \theta)$ is completely analogous. In this case, $F(x, \theta) = F_2(x/\theta)$ and working in the same way as above we have that

$$F^{-1}_{\theta}(u) = \theta \inf \{y: F_2(y) \geq u\} = \theta l_2(u) \tag{4.10}$$

where $l_2(u)$ does not depend on θ. From (4.3) and (4.10),

$$\frac{dx}{d\theta} = \lim_{\Delta\theta \to 0} \frac{1}{\Delta\theta} [(\theta + \Delta\theta) l_2(u) - \theta l_2(u)] = l_2(u) = \frac{F^{-1}_{\theta}(u)}{\theta} = \frac{x}{\theta}.$$

Thus we can again see that A1 and A2 are satisfied and that

$$\frac{dx}{d\theta} = \phi(x) = \frac{x}{\theta}. \tag{4.11}$$

Checking the validity of A3 can easily be done in the same way as above.

As a consequence of the above theorem we can see that the parameters of the commonly used distributions we state below are admissible: the mean (and rate) of the exponential distribution, the mean and spread of the uniform distribution, the scale parameter of the gamma distribution, the constant that characterizes the deterministic distribution, the scale parameter of the Weibull distribution. Finally let us give some more examples of admissible (and nonadmissible) parameters of common distributions. Let $p_i \geq 0$, $i = 1, 2, \ldots, n$ and $\sum_{i=1}^{n} p_i = 1$. Then, for the general discrete distribution $F(x; \theta_1, \ldots, \theta_n; p_1, \ldots, p_n) = \sum_{\{i: x \leq \theta_i\}} p_i$ the parameters $\theta_1, \ldots, \theta_n$ are admissible whereas $p_1, \ldots, p_n$ are *not*. For the hyperexponential distribution $F(x; \theta_1, \ldots, \theta_n; p_1, \ldots, p_n) = 1 - \sum_{k=1}^{n} p_i e^{-x/\theta_i}$, $\theta_1, \ldots, \theta_n$ are admissible parameters, while $p_1, \ldots, p_n$ are *not*. Both parameters of the Beta, the Weibull and the lognormal distribution are admissible.

5. Perturbation Analysis Algorithm for Service Time in $G/G/1$ Queue

Next we develop a perturbation analysis algorithm to compute the sensitivity of mean time in system w.r.t the service time parameter θ. Since application of perturbation analysis does not require the Markovian ("M") assumption for arrivals, in this section we may as well develop the algorithm for the $GI/G/1$ queue. This algorithm is a special case of the general technique in Suri (1987), or the network algorithm in Ho and Cao (1983), but for the $GI/G/1$ queue it is easily derived from first principles, which we do here for the benefit of the unfamiliar reader.

5.1. *Derivation of Algorithm*

The basic idea of (infinitesimal) perturbation analysis is to consider what would have happened to a given (nominal) sample path, if θ had been $\theta + \Delta\theta$ instead. We are interested in particular in the case where $\Delta\theta$ becomes infinitesimally small. The service time of C_j would then be changed by the amount $\Delta x_j = dx_j \Delta\theta/d\theta$ $(=\phi(x_j)\Delta\theta$ whenever A2 applies). Now let us see how such changes would affect the system time of all customers. The standard assumption that makes infinitesimal perturbation analysis easy to implement is that the order of events in the nominal and perturbed paths remains the same. (As we will see later, for the $GI/G/1$ system we need only worry about the possibility that in the perturbed path two busy periods may coalesce because of accumulated perturbations, or conversely, that one busy period may split into two.) So we will make the following assumption.

Assumption of Infinitesimal Perturbations (AIP). The set of indices (j) of customers (C_j) that initiate busy periods are the same for the nominal and perturbed paths.

In other words we assume that if C_j initiates a busy period in the nominal path then, and only then, does C_j initiate a busy period in the perturbed path. This assumption will be discussed below, but it is typical of infinitesimal perturbation analysis. (Notice that the above assumption (AIP) expresses a way of thinking in order to derive an algorithm and does not constitute a premise of the analysis in §6.)

Now we look at the start of a busy period initiated by (say) C_{k+1}, and by (AIP) it follows that in the perturbed path C_{k+1} would also find the system idle. Suppose n customers $C_{k+1}, \ldots, C_{k+n}$ are served during this busy period. Then, in the perturbed path, C_{k+1} would spend an additional time Δx_{k+1} (with value as above) in the system. C_{k+2} would spend an additional time $\Delta x_{k+1} + \Delta x_{k+2}$ in the system—an additional Δx_{k+1} waiting for C_1 and an additional Δx_{k+2} for its own service time. In general then,

$$\textit{additional time for} \quad C_{k+i} = \sum_{j=1}^{i} \Delta x_{k+j} \qquad (1 \le i \le n). \tag{5.1}$$

Now suppose there are M busy periods observed, during a total time t, to get the estimate T as in (3.6). Assuming still that (AIP) holds, the effect of $\Delta\theta$ on T is

$$\Delta T = \frac{1}{N} \sum_{m=1}^{M} \sum_{i=1}^{n_m} \sum_{j=1}^{i} \Delta x_{k_m+j} = \frac{1}{N} \sum_{m=1}^{M} \sum_{i=1}^{n_m} \sum_{j=1}^{i} \frac{dx_{k_m+j}}{d\theta} \Delta\theta \tag{5.2}$$

where n_m is the number of customers served during the mth busy period, $k_m + 1$ is the index of the customer initiating this busy period $(k_1 \equiv 0)$, and $N = \sum_{m=1}^{M} n_m$. Dividing by $\Delta\theta$ we get our estimate

$$\left.\frac{d\bar{T}}{d\theta}\right]_{\text{est}} = \frac{\Delta T}{\Delta\theta} = \frac{1}{N} \sum_{m=1}^{M} \sum_{i=1}^{n_m} \sum_{j=1}^{i} \frac{dx_{k_m+j}}{d\theta}. \tag{5.3}$$

Note first, that to calculate (5.3), values of $dx_{k_m+j}/d\theta$ are determined from the nominal observations of x_{k_m+j}, by evaluation of (4.3) (or (4.7) if applicable). In other words, this estimate can be obtained without observing the perturbed path. Second, the notation "est" above reminds us that this is an estimate, with as yet no proven relation to the true value $d\bar{T}/d\theta$.

It is illustrative to write this calculation down in algorithmic form, to see the basics of perturbation analysis implemented during observation of a single sample path: this is shown in Algorithm 1. The value EST in Algorithm 1 is the perturbation analysis estimate of $d\bar{T}/d\theta$.

Even this simple algorithm illustrates the *three basic elements* of perturbation analysis, namely

Algorithm 1.

Initialize:

0. DXSUM = 0
   DTSUM = 0

At end of service of customer j:

1. DXDTHETA = $\phi(x_j)$ [x_j is observed]
2. DXSUM = DXSUM + DXDTHETA
3. DTSUM = DTSUM + DXSUM
4. If server is now idle, then DXSUM = 0

At end of M busy periods (with N customers served):

5. EST = DTSUM/N
6. Stop

(i) *Perturbation Generation.* In step 1, the effect of $\Delta\theta$ on the outcome of a single random variable is calculated.

(ii) *Perturbation Propagation.* In step 2, we "propagate" the effects of a change in time of a current event to the times of future events. Also, step 4 takes into account the fact that perturbations do not propagate across idle periods.

(iii) *Effect on Performance.* Step 3 reflects how the perturbation in time of a specified event contributes to perturbations in the overall performance measure.

In the example here, these three elements are quite simple, but for complex systems their calculation can be more involved, although still relatively efficient (Suri 1987). In concept, all the references cited on infinitesimal perturbation analysis just use these three steps above.

Thus, if we perform a computer simulation of a $GI/G/1$ queue, in the usual way, to get an estimate of $\bar{T}$ (e.g. Meketon and Heidelberger 1982), then, with the above calculations included, we will get estimates of both $\bar{T}$ and $d\bar{T}/d\theta$ from a single simulation. Also note that there is no requirement that the nominal sample path be generated from a simulation—*we could be observing an actual queue*—in which case our ability to estimate $d\bar{T}/d\theta$ without repeating the experiment is a rather interesting capability of perturbation analysis! Finally notice how simple Algorithm 1 becomes when θ is either a location or a scale parameter. In this case, as we can see from (4.9) and (4.11), $\phi(x_j)$ is equal to 1 or to x_j/θ respectively and step 1 does not involve any computation at all.

Extension of this method for the case of a *vector* of parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$ is immediate. Let $\phi_i(x_j) = dx_j/d\theta_i$. In Algorithm 1 we use p-dimensional arrays for the variables, and the extended algorithm just requires some "loops" to be added: see Algorithm 2. The value EST[i] in Algorithm 2 is the perturbation analysis estimate of $\partial\bar{T}/\partial\theta_i$.

So perturbation analysis gives us the capability to compute the gradient of a performance measure with respect to p parameters from a single sample path. To do that by conventional approaches would require $p + 1$ sample paths.

5.2. *Application to M/G/1 Queue*

Returning to our original scalar case, from (5.3) we see that for the mth busy period, perturbation analysis calculates h_m defined by

$$h_m = \sum_{i=1}^{n_m} \sum_{j=1}^{i} \frac{dx_{k_m+j}}{d\theta}. \tag{5.4}$$

Algorithm 2.

Initialize:

0. For i:=1 to p do DXSUM[i] = 0
   For i:=1 to p do DTSUM[i] = 0

At end of service of customer j:

For i:=1 to p do
begin

1. DXDTHETA[i] = $\phi_i(x_j)$ [x_j is observed]
2. DXSUM[i] = DXSUM[i] + DXDTHETA[i]
3. DTSUM[i] = DTSUM[i] + DXSUM[i]
4. If server is now idle, then DXSUM[i] = 0

end

At end of M busy periods (with N customers served):

5. For i:=1 to p do EST[i] = DTSUM[i]/N
6. Stop

Using h_m, we can write (5.3) as

$$\left.\frac{d\bar{T}}{d\theta}\right]_{\text{est}} = \left[\frac{1}{M}\sum_{m=1}^{M} h_m\right] \Big/ \left[\frac{1}{M}\sum_{m=1}^{M} n_m\right]. \tag{5.5}$$

The first bracket is the average value of h_m for a busy period, and the second is the average number of customers served in a busy period. Thus by the strong law of large numbers

$$\lim_{M\to\infty} \left.\frac{d\bar{T}}{d\theta}\right]_{\text{est}} = \frac{E[h_m]}{E[n_m]} \qquad \text{w.p. 1.} \tag{5.6}$$

(The existence of $E[h_m]$ is established in the Appendix.) Our aim here is to derive an analytic expression for the RHS of this equation, and compare it with the known value of $d\bar{T}/d\theta$ for an $M/G/1$ queue, namely equation (3.2).

5.3. *Remark on the Analysis*

The reader may wonder at this point whether we are simply undertaking an exercise in algebra, to derive a known result in a roundabout way. This is not so! Recall that for our derivation of the perturbation analysis algorithm above we required the assumption (AIP), namely, that if C_k initiates a busy period in the nominal path, then C_k also initiates a busy period in the perturbed path (and vice versa). A similar assumption is required for all infinitesimal perturbation analysis algorithms (Ho and Cassandras 1983, Ho and Cao 1983, Suri 1987). A justification of such an assumption is that it can be replaced by the alternative (Suri 1987),

Assumption (*AIP'*): Two events do not occur at the same instant in time, and the number of events in a sample path of finite length is finite with probability one.

Then, within this observation period, we will have a finite value $t_{\min}$ which is the minimum time separating two events. So, from Assumption A1 there exists a $\Delta\theta$ small enough so that the cumulative effect of all perturbations will not exceed $t_{\min}$ during the observation period, and the order of events will be the same in the nominal and perturbed paths. Assumption AIP will then be implied by A1. Now, because of the definition of $dy/d\theta$ in §1, we see that we are only interested in infinitesimal perturbations anyway, so the analysis becomes exact for this finite observation period. (Formal arguments, for a general discrete event system, are in Suri 1987.)

A potential criticism of this perturbation analysis approach is that such an estimate may be biased because of the following argument. For any $\Delta\theta$ "small enough" for the above argument, there exists a *new* time t "large enough" so that the probability that the order of events will change becomes arbitrarily close to unity. Thus the estimate $d\bar{T}/d\theta]_{\text{est}}$ may not be accurate "in the long run", i.e. its limiting value may not equal the true value of $d\bar{T}/d\theta$. This argument is more formally stated by Cao (1985a), who also sheds some light on the issues involved, and gives examples of cases where perturbation analysis both is, and is not, biased (see also Heidelberger et al. 1987).

For these reasons, it is important for us to establish whether the perturbation analysis estimate is unbiased for those systems that are analytically tractable. In the following section we will establish the unbiasedness of these estimates for one of the simplest, analytically tractable discrete event systems, the $M/G/1$ queue.

6. Behavior of Estimates for Service Parameter Sensitivities

Now let us consider the perturbation analysis algorithm applied to the service time parameter (i.e. θ) of an $M/G/1$ queue. In order to do that, we focus our attention on the mth busy period. For ease of notation, we will drop the subscript m throughout this section, and count customers from 1 on. So, during this busy period there are n customers served, C_i is the ith customer, and x_i the service time of C_i. Also, as in (5.4), the contribution that this busy period makes to the sensitivity estimate is

$$h = \sum_{i=1}^{n} \sum_{j=1}^{i} \frac{dx_j}{d\theta} = \sum_{i=1}^{n} \sum_{j=1}^{i} \phi(x_j), \tag{6.1}$$

the second equality following from A2.

Let us consider (5.6) again. From Kleinrock (1975) we have $E[n] = 1/(1 - \lambda\bar{x})$, and thus

$$\lim_{M\to\infty} \frac{d\bar{T}}{d\theta}\Bigg]_{\text{est}} = (1 - \lambda\bar{x})E[h] \qquad \text{w.p. 1.} \tag{6.2}$$

So the problem reduces to computing, for a busy period, the value of

$$E[\sum_{i=1}^{n} \sum_{j=1}^{i} \phi(x_j)]. \tag{6.3}$$

Unfortunately, the difficulty of the problem now becomes apparent. First, each $\phi(x_j)$ depends on each x_j (as shown in §4), and second, n depends on the *entire sequence of* x_j's in a complicated fashion. So the evaluation of the above expectation is not straightforward.

We proceed by using three devices to simplify our task. The first, well known in queueing theory, is to decompose the busy period into sub-busy periods. The advantage of doing this is that the sub-busy periods are statistically independent, and identically distributed. The second device, unique to our approach, is to derive a recursion in *two* r.v.'s associated with the busy period. The final device is to note that since we are interested only in expected values and not distributions, we can considerably simplify our task by taking expectations at an appropriate point in the analysis.

We follow the approach in Kleinrock (1975, p. 209) for busy period analysis. Let k customers arrive during the service time of C_1. Since we are interested only in mean system time, we can use LCFS service discipline, and as explained in Kleinrock (1975) each of C_2 through C_{k+1} initiates a sub-busy period *statistically identical* to the "parent" busy period initiated by C_1. Furthermore, these sub-busy periods are *statistically independent.*

Let us define the quantity

$$g = \sum_{i=1}^{n} \phi(x_i). \tag{6.4}$$

Notice that g and h are the values of DXSUM and DTSUM obtained from Algorithm 1 executed for only one busy period.

Now we will number the sub-busy periods in the order that they occur, and we will number the customers *in the order that they are served* using the LCFS discipline. Let m_r be the number of customers in the rth sub-busy period and define $m^{(r)} = 1 + m_1 + \cdots + m_r$ with $m^{(0)} = 1$. Then $C_{m^{(r-1)}+1}$ through $C_{m^{(r)}}$ are the customers that belong to the rth sub-busy period. Now consider the quantities

$$g^{(r)} = \sum_{i=1}^{m_r} \phi(x_{m^{(r-1)}+i}) \qquad \text{with} \qquad g^{(0)} = \phi(x_1) \qquad \text{and} \tag{6.5}$$

$$h^{(r)} = \sum_{i=1}^{m_r} \sum_{j=1}^{i} \phi(x_{m^{(r-1)}+j}). \tag{6.6}$$

Notice that $g^{(r)}$ and $h^{(r)}$ are the values of g and h that would be obtained if Algorithm 1 was applied only to customers of the rth sub-busy period. For this reason we will refer to $g^{(r)}$ and $h^{(r)}$ as the *stand-alone values* for the rth sub-busy period.

Next we derive some relations between the above quantities: From (6.4) and (6.6) follows that

$$g = \sum_{i=1}^{n} \phi(x_i) = \phi(x_1) + \sum_{r=1}^{k} \sum_{i=1}^{m_r} \phi(x_{m^{(r-1)}+i}) = \sum_{r=0}^{k} g^{(r)}. \tag{6.7}$$

This can also be seen from the fact that the value of DXSUM at the end of the busy period is the sum of $\phi(x_1)$ plus the stand-alone values of DXSUM for all the sub-busy periods.

Now let us derive a corresponding expression for h. The contribution of the rth sub-busy period to the final value of the register DTSUM in Algorithm 1 consists of two parts: the first part is the stand-alone value $h^{(r)}$, while the second is due to the fact that when the contribution of $C_{m^{(r-1)}+1}$ (i.e. of the customer who initiates the rth sub-busy period) is taken into account, DXSUM does not equal zero (as it would for the stand-alone value). In fact it has the value

$$\sum_{i=1}^{m^{(r-1)}} \phi(x_i) = \sum_{s=0}^{r-1} g^{(s)}. \tag{6.8}$$

The quantity in (6.8) will be added to DTSUM as many times as there are customers in the rth sub-busy period. Hence the contribution of the rth sub-busy period to h is

$$h^{(r)} + m_r[\sum_{s=0}^{r-1} g^{(s)}]. \tag{6.9}$$

Summing (6.9) over all sub-busy periods (and taking into account the contribution of C_1 who initiates the busy period) we get

$$h = \phi(x_1) + \sum_{r=1}^{k} [h^{(r)} + m_r \sum_{s=0}^{r-1} g^{(s)}]. \tag{6.10}$$

(This relation can also be verified by simple rearrangement of the RHS of the above equation.)

Now we wish to derive the expected values of g and h from (6.7) and (6.10). (In the Appendix we show that these expected values exist and are finite.) Let us first consider

g. On the RHS of (6.7) we have a sum of the r.v. $\phi(x_1)$ with a random number k of r.v.'s $g^{(r)}$. Now, the power of sub-busy periods lies in exploiting the independence of these r.v.'s. We note first that k is the number of arrivals during the service of C_1, and so depends only on x_1 and is independent of the $g^{(r)}$ values for $r \geq 1$. Also, the $g^{(r)}$ are distributed identically to g. So taking expectations in (6.5) we get

$$E(g) = E(\phi(x_1)) + E(k)E(g). \tag{6.11}$$

Noting that $E(k)$ equals $\lambda\bar{x}$ and solving for $E(g)$ gives

$$E(g) = \frac{E(\phi(x_1))}{1 - \lambda\bar{x}}. \tag{6.12}$$

Now we turn to h in (6.10). Clearly, the $h^{(r)}$ terms are independent and distributed as h. The interesting point to note is that each term such as $m_r(\phi(x_1) + g^{(1)} + \cdots + g^{(r-1)})$ involves the product of the number of customers arriving in a given sub-busy period, with quantities accumulated through the *preceding* sub-busy periods, but not including the current sub-busy period. Thus again, these r.v.'s are independent. Let us now take expected values in (6.10) conditioned on x_1 and k.

$$E(h|x_1, k) = \phi(x_1) + \sum_{r=1}^{k} E[h^{(r)}|x_1, k] + \sum_{r=1}^{k} E[m_r \sum_{s=0}^{r-1} g^{(s)}|x_1, k]$$

Taking into account the fact that quantities referring to different sub-busy periods are independent from each other and from x_1 and k and identically distributed to the parent busy period and also that m_r is independent of $g^{(s)}$ for $s < r$, we get

$$\begin{aligned} E(h|x_1, k) &= \phi(x_1) + kE(h) + kE(m_r)\phi(x_1) \\ &\quad + (E[m_2] + 2E[m_3] + \cdots + (k-1)E[m_k])E(g) \\ &= \phi(x_1) + k[E(h) + E(m_r)\phi(x_1)] + E(m_r)E(g)(k^2 - k)/2. \end{aligned} \tag{6.13}$$

(Here $\phi(x_1)$ is known when x_1 is given.) Now let us take expectations with respect to k in the above equation conditioned on x_1. Then $E(k|x_1)$ is the average number of Poisson arrivals in an interval of length x_1 and so it is equal to λx_1. Similarly, $E(k^2|x_1)$ is equal to $\lambda x_1 + (\lambda x_1)^2$. Taking also into account that m_r is identically distributed to n (the number of customers in the parent busy period) and thus $E(m_r) = E(n) = 1/(1 - \lambda\bar{x})$, and using the expression for $E(g)$ already derived, we get

$$E(h|x_1) = \phi(x_1) + \lambda x_1[E(h) + \phi(x_1)/(1 - \lambda\bar{x})] + (\lambda x_1)^2 E(\phi(x_1))/2(1 - \lambda\bar{x})^2. \tag{6.14}$$

For the next step, we take expectation w.r.t x_1 to get

$$E(h) = E(\phi(x)) + \lambda\bar{x}E(h) + \lambda E(x\phi(x))/(1 - \lambda\bar{x}) + \lambda^2\overline{x^2}E(\phi(x))/2(1 - \lambda\bar{x})^2 \tag{6.15}$$

(the subscript on x is no longer necessary). Hence solving (6.15) for $E(h)$ gives

$$E(h) = E(\phi(x))/(1 - \lambda\bar{x}) + \lambda E(x\phi(x))/(1 - \lambda\bar{x})^2 + \lambda^2\overline{x^2}E(\phi(x))/2(1 - \lambda\bar{x})^3. \tag{6.16}$$

Substituting values from (4.4), using A3, and multiplying by $1 - \lambda\bar{x}$, we get for (6.2) the value

$$\frac{d\bar{x}}{d\theta} + \frac{\lambda}{2(1 - \lambda\bar{x})}\frac{d\overline{x^2}}{d\theta} + \frac{\lambda^2\overline{x^2}}{2(1 - \lambda\bar{x})^2}\frac{d\bar{x}}{d\theta}. \tag{6.17}$$

Voila! This is exactly the value obtained by differentiating the P-K formula earlier (3.2).

We consolidate and summarize what we have proved in the following.

THEOREM 1. *The perturbation analysis algorithm (Algorithm 1) gives strongly consistent estimates for the sensitivity of the steady state mean system time of a customer, with respect to an admissible service time parameter, for an M/G/1 queue.*

PROOF. Our busy period analysis proved that $E[h_m]/E[n_m] = d\bar{T}/d\theta$. Since the busy periods are independent and identically distributed, and since $E|h_m|$ exists (see Appendix), the strong law of large numbers then implies that our estimate

$$[\sum_{m=1}^{M}\sum_{i=1}^{n_m}\sum_{j=1}^{i}\phi(x_{k_m+j})]/[\sum_{m=1}^{M} n_m] \to \frac{d\bar{T}}{d\theta} \qquad \text{with probability 1} \tag{6.18}$$

as $M \to \infty$.

THEOREM 2. *The perturbation analysis estimate obtained from Algorithm 1 for the sensitivity of the steady state mean system time of a customer with respect to an admissible parameter is asymptotically unbiased.*

The proof of this statement is given in the Appendix.

COROLLARY 1. *For a customer entering an M/G/1 queue in steady state, let ΔW be the perturbation in this customer's system time calculated by perturbation analysis due to a parameter change $\Delta\theta$, considered infinitesimal. Then $\Delta W/\Delta\theta$ is an unbiased estimator of $d\bar{T}/d\theta$.*

Since the arrivals are Markovian, they also find the system in steady state (Kleinrock 1975), and the corollary follows by straightforward application of the strong law of large numbers, so the details are omitted here. This result says that if we start in steady state, then the average of $dW/d\theta$ observed over any N customers will be an unbiased estimator of $d\bar{T}/d\theta$ (as opposed to asymptotically unbiased when we started our algorithm at the beginning of a busy period).

7. Perturbation Analysis for Parameter of Arrival Distribution

7.1. *Algorithm for GI/G/1 Queue*

In the same vein as §5, we can estimate the derivative of mean system time w.r.t. an admissible parameter of the (general) arrival distribution for a $GI/G/1$ queue. For this section, let λ denote this more general parameter, and let $G(a, \lambda)$ be the (cumulative) probability distribution for interarrival times (a is the dummy variable for the interarrival time).

In order to use infinitesimal perturbation analysis, we will need assumption (AIP). Let a_j denote interarrival time between C_{j-1} and C_j, and let Δa_j be the perturbation in this time due to $\Delta\lambda$ (C_j arrives Δa_j later relative to C_{j-1}). Consider the busy period consisting of C_{k_m+1} through $C_{k_m+n_m}$. (We are using the same notation as in §5.) Then the change in system time for C_{k_m+i} is

$$\Delta T_{k_m+i} = -\sum_{j=2}^{i}\Delta a_{k_m+j} = -\sum_{j=2}^{i}\frac{da_{k_m+j}}{d\lambda}\Delta\lambda \qquad \text{if} \quad 2 \le i \le n, \qquad \text{and } 0 \quad \text{if} \quad i = 1. \tag{7.1}$$

The minus sign in (7.1) is due to the fact that a customer who arrives at the system a little later will have to wait less. Notice in particular that there is no change in the system time of the customer who initiates a busy period, and that perturbations are not propagated from one busy period to the next. Let us now define

$$q_m = -\sum_{i=2}^{n_m}\sum_{j=2}^{i}\frac{da_{k_m+j}}{d\lambda}. \tag{7.2}$$

Then the perturbation analysis estimate based on M regenerative periods is, for this case,

$$\left.\frac{d\bar{T}}{d\lambda}\right]_{\text{est}} = -\sum_{m=1}^{M}\sum_{i=1}^{n_m}\sum_{j=2}^{i}\frac{da_{k_m+j}}{d\lambda}\Big/\sum_{m=1}^{M} n_m = \sum_{m=1}^{M} q_m/\sum_{m=1}^{M} n_m, \tag{7.3}$$

with notation as in (5.3), and with the further understanding that $q_m = 0$ when $n_m = 1$. An algorithm similar to Algorithm 1 is easily written to calculate (7.3) for an observed sample path. Also note that we can extend Algorithm 2 to simultaneously calculate gradients with respect to several arrival and service parameters.

7.2. *Behavior of Estimates for Arrival Parameter Sensitivity*

We now study the above perturbation analysis estimate (7.3) in the case of an $M/G/1$ queue. Since the interarrival times are exponentially distributed, with $G(a, \lambda) = 1 - \exp(-\lambda a)$, the arrival time parameter is admissible (see Example 4.2—note however that λ here is the inverse of the mean, and this gives rise to a negative sign in (7.4) below as compared to $dx/d\theta$ in Example 4.2).

With $G(a, \lambda)$ as above, we have from the analog of (4.7) that

$$\frac{da_{k_m+j}}{d\lambda} = -\left.\frac{\partial G/\partial\lambda}{\partial G/\partial a}\right|_{a_{k_m+j},\lambda} = -\frac{a_{k_m+j}}{\lambda}. \tag{7.4}$$

Hence, (7.1) becomes

$$\Delta T_{k_m+i} = \frac{\Delta\lambda}{\lambda}\sum_{j=2}^{i} a_{k_m+j} \quad \text{if} \quad i \geq 0, \quad \text{and } 0 \quad \text{if} \quad i = 1. \tag{7.5}$$

Now let

$$z_{k_m+i} = \sum_{j=2}^{i} a_{k_m+j} \quad \text{if} \quad i \geq 2 \quad \text{and } 0 \quad \text{if} \quad i = 1. \tag{7.6}$$

Notice that z_{k_m+i} is the time that has elapsed from the beginning of the mth busy period till the arrival of the ith customer of that busy period. From (7.2), (7.4) and (7.6) we have

$$q_m = \frac{1}{\lambda}\sum_{i=1}^{n_m} z_{k_m+i}. \tag{7.7}$$

Thus from (7.3), using the strong law of large numbers, we have

$$\lim_{M\to\infty}\left.\frac{d\bar{T}}{d\lambda}\right]_{\text{est}} = \lim_{M\to\infty}\left[\frac{1}{M}\sum_{m=1}^{M} q_m\Big/\frac{1}{M}\sum_{m=1}^{M} n_m\right] = \frac{E[q_1]}{E[n_1]} \quad \text{w.p. 1.} \tag{7.8}$$

Notice that $E[q_1]$ exists since q_1 is nonnegative and in fact it can be shown that $E[q_1]$ is finite: if we denote by y_1 the length of the first busy period, then $z_i \leq y_1$ for $i = 1, 2, \cdots n_1$. Hence $q_1 < n_1 y_1/\lambda$, which implies that

$$E[q_1] < \frac{1}{\lambda}E[n_1 y_1] \leq \frac{1}{\lambda}E[n_1^2]^{1/2}E[y_1^2]^{1/2}.$$

But $E[n_1^2]$ and $E[y_1^2]$ are finite since we have assumed the queue to be stable and the second moment of the service time finite (e.g. see Kleinrock 1975, pp. 214–218). Thus

$$E[q_1] = \frac{1}{\lambda}E\Big[\sum_{i=1}^{n_1} z_i\Big] < \infty. \tag{7.9}$$

The process z_k, $k = 1, 2, \cdots$ is clearly a discrete time regenerative process with regeneration points the integers $k_m + 1$, $m = 1, 2, \cdots$ which correspond to customers who initiate busy periods. From a standard result in regenerative processes (Crane and

Iglehart, 1975) follows that as k goes to infinity, z_k converges in distribution to a random variable z with expected value

$$E[z] = \frac{E[\sum_{i=1}^{n_1} z_i]}{E[n_1]}. \tag{7.10}$$

Hence, from (7.8), (7.9) and (7.10) follows that

$$\lim_{M\to\infty} \frac{d\bar{T}}{d\lambda}\Big]_{\text{est}} = \frac{1}{\lambda} E[z] \qquad \text{w.p. 1.} \tag{7.11}$$

To verify that our estimate is strongly consistent, it is sufficient to show that $E[z]/\lambda$ is equal to the true value of the derivative. (The reason for transforming (7.8) into (7.11) is that $E[z]$ is much easier to compute than $E[q_1]$.)

To compute $E[z]$ we will make use of the fact that Poisson arrivals take a "random look" into the system (Wolff 1982). We can regard the sample path of an $M/G/1$ queue as an alternating renewal process consisting of busy and idle periods. Let us suppose that this process is in steady-state and let us make a random observation. Define the random variable $\hat{z}$ as follows. $\hat{z} = 0$ if the observation point falls on an idle period whereas $\hat{z} = \tau$ if the observation point falls on a busy period that has started τ time units ago. Since customers arriving according to a Poisson process see the system in the same way as an observer who arrives at a "random instant" we have

$$E[z] = E[\hat{z}]. \tag{7.12}$$

(In fact, as it was pointed out by a referee, (7.12) does not follow directly from Wolff (1982). A proof of (7.12) is included in the Appendix.) But $E[\hat{z}]$ can be computed very easily from a standard result in the theory of continuous time regenerative processes (Crane and Iglehart 1975). In fact,

$$E[\hat{z}] = \frac{E[\int_0^{y_1} \tau d\tau]}{E[y_1] + E[I_1]} \tag{7.13}$$

where y_1 is again the length of the first busy period and I_1 is the length of the first idle period. Hence,

$$E[\hat{z}] = \frac{E[y_1^2]}{2(E[y_1] + E[I_1])}. \tag{7.14}$$

Taking into account that $E[y_1] = \bar{x}/(1 - \lambda\bar{x})$, $E[y_1^2] = \overline{x^2}/(1 - \lambda\bar{x})^3$ and $E[I_1] = 1/\lambda$ (e.g. see Kleinrock 1975), we get

$$E[\hat{z}] = \frac{\overline{\lambda x^2}}{2(1 - \lambda\bar{x})^2}. \tag{7.15}$$

So from (7.11), (7.12) and (7.15) follows that

$$\lim_{M\to\infty} \frac{d\bar{T}}{d\lambda}\Big]_{\text{est}} = \overline{x^2}/2(1 - \lambda\bar{x})^2 \tag{7.16}$$

which is the value in (3.3). Thus we have proved

THEOREM 3. *The perturbation analysis estimate of the derivative of the mean response time of an $M/G/1$ queue with respect to the arrival rate is strongly consistent.*

One can also very easily establish the following

THEOREM 4. *Perturbation analysis gives asymptotically unbiased estimates of the derivative of the mean response time of an $M/G/1$ queue with respect to the arrival rate.*

The proof is given in the Appendix.

8. Experiments with Some $M/G/1$ and $GI/G/1$ Queues

This section contains results of experiments to estimate the gradient $d\bar{T}/d\theta$ for some single server queues, from one sample path. It concludes with an interesting optimization experiment, also conducted on a single sample path.

First we validate our analytical results with simulation experiments on $M/G/1$ queues. Then we use perturbation analysis on some $GI/G/1$ queues. In the $GI/G/1$ cases, analytic expressions are not available for $\bar{T}$ and $d\bar{T}/d\theta$. Nevertheless, the experimental results are encouraging. More usefully, we show how perturbation analysis enables surprisingly efficient optimization of $GI/G/1$ systems.

Given that we are using experimental observations here, it is appropriate to remark on the use of *confidence intervals* with perturbation analysis. Since the estimates obtained by perturbation analysis are just some functions of observations on a sample path, the usual techniques can be applied to get confidence intervals for these gradient estimates, as would be used for any other conventional estimates (e.g. mean system time) obtained from a sample path. In particular, we can use independent replications, or batch means, or regenerative methods (Fishman 1978). The use of *regenerative methods* is a particularly appropriate choice with perturbation analysis, since they too operate by observing a *single* sample path, and so we have chosen to work with regenerative techniques here.

In using the regenerative approach, we also used the bias reduction technique suggested by Meketon and Heidelberger (1982). To implement this, one first chooses a stopping criterion, which is a number N. At the end of busy period k, let $N(k)$ be the total number of customers served since the start of the experiment. Their approach terminates the experiment when $N(k) \geq N$. All simulations below used $N = 100{,}000$ and all confidence intervals shown are at the 95% level.

8.1. *Numerical Validation of Perturbation Analysis Estimates*

EXAMPLE 8.1: An $M/M/1$ Queue. We simulated an $M/M/1$ system with $\lambda = 0.01$ (mean interarrival time = 100). The service time parameter is $\theta = \bar{x}$. Three different cases were tried namely, light traffic ($\theta = 20$, $\rho = \lambda\bar{x} = 0.2$), medium traffic ($\theta = 50$, $\rho = 0.5$), and heavy traffic ($\theta = 80$, $\rho = 0.8$). The experimental results are in Table 1, where they are also compared with the known analytic values.

EXAMPLE 8.2: An $M/U/1$ Queue. Here the service time distribution, and its parameters, are as in Example 4.3. Again, the system was simulated for three different traffic intensities. For all cases λ was fixed at 0.01, and we chose for light traffic: $\theta = 20$, $\delta = 16$; for medium traffic: $\theta = 50$, $\delta = 40$ and for heavy traffic: $\theta = 80$, $\delta = 64$. In this example, we simultaneously estimated three gradients, w.r.t. the parameters θ, δ, and λ, from one simulation at each traffic intensity. Experimental results are compared with analytically derived values in Table 2.

EXAMPLE 8.3: A $GI/G/1$ Queue. In this case the service time *density* $f(x, \theta)$ is triangular with

$$f(x, \theta) = \begin{cases} x/\theta^2 & \text{if} \quad 0 \leq x < \theta, \\ 2/\theta - x/\theta^2 & \text{if} \quad \theta \leq x < 2\theta, \\ 0 & \text{elsewhere,} \end{cases} \tag{8.1}$$

and the interarrival times are uniform in [0, 200]. Analytic expressions are not available for this queue. We will estimate $d\bar{T}/d\theta$ at the value $\theta = 70.61$, using both perturbation analysis and by a finite difference method with additional experiments at $\theta = 73.61$.

TABLE 1

Gradient Estimates for M/M/1 Queue

Traffic Intensity ρ	Method Used	Quantity Estimated		
		$\bar{T}$	$d\bar{T}/d\theta$	$d\bar{T}/d\lambda$
0.2	Experiment	25.00 ± 0.23	1.562 ± 0.021	623 ± 22
	Theory	25.00	1.563	625
0.5	Experiment	99.98 ± 1.78	3.95 ± 0.12	9,816 ± 487
	Theory	100.00	4.00	10,000
0.8	Experiment	403 ± 22	25.6 ± 3.0	$165 \times 10^3 \pm 21 \times 10^3$
	Theory	400	25.0	160×10^3

Note. See §8 and Example 8.1 for details of the experiments.

TABLE 2

Gradient Estimates for M/U/1 Queue

Traffic Intensity ρ	Method Used	Quantity Estimated			
		$\bar{T}$	$d\bar{T}/d\theta$	$d\bar{T}/d\delta$	$d\bar{T}/d\lambda$
0.2	Experiment	23.05 ± 0.10	1.288 ± 0.007	0.066 ± 0.005	377 ± 11
	Theory	23.03	1.288	0.067	379
0.5	Experiment	80.3 ± 0.8	2.58 ± 0.05	0.262 ± 0.015	5893 ± 215
	Theory	80.8	2.61	0.267	6067
0.8	Experiment	273 ± 13	15.3 ± 1.9	1.09 ± 0.12	$102 \pm 15 \times 10^3$
	Theory	274	14.7	1.07	97×10^3

Notes.
1. P/A stands for perturbation analysis.
2. See §8 and Example 8.3 for details of the experiments.

TABLE 3

Gradient Estimates for G/G/1 Queue

Parameter Value θ	Method Used	Quantity Estimated	
		$\bar{T}$	$d\bar{T}/d\theta$
70.61	Experiment including P/A	111.70 ± 0.78	3.54 ± 0.07
73.61	Experiment including P/A	123.35 ± 0.96	4.29 ± 0.10
	Average of P/A Estimates		3.91 ± 0.06
	Finite Difference Estimate		3.88 ± 0.21

Notes.
1. P/A stands for perturbation analysis.
2. See §8 and Example 8.3 for details of the experiments.

(The reason for choosing the points 70.61 and 73.61 is to compare the result with another experiment that follows.) Results are in Table 3. Note that we used perturbation analysis to estimate the derivative at both points, and the two values compare well

with the difference estimate. The difference estimate is actually estimating the slope of the secant connecting the two points, while perturbation analysis is estimating the two tangent slopes. In fact, it can be shown that, to *second* order, the secant slope is the average of the two tangent slopes (Cao 1983). The average of the two perturbation analysis estimates is also shown, which now compares very well with the difference estimate, giving us additional confidence in the method. Also notice that the perturbation analysis estimates give tighter confidence intervals. We have shown elsewhere that this is always the case for certain queueing systems (Zazanis and Suri 1986a).

8.2. *Single-Run Optimization*

Now we come to a very interesting use of perturbation analysis, suggested independently by Ho (1982) and Meketon (1983), for optimization of discrete event systems. Our method is a further modification of that of Meketon (1983). In view of our preceding results, an obvious optimization approach would be to conduct an experiment (on a simulation or real system), get an estimate of the sensitivity(ies) using perturbation analysis, then use this to improve the parameter value(s), and repeat. Since each experiment is stochastic, we would use a stochastic approximation algorithm to update the parameter values. Such an approach was used successfully by Ho and Cao (1983) for several queueing network examples.

However, it seems that we can do considerably better than this first approach. The new idea, suggested by Ho (1982) and Meketon (1983), is that, since the perturbation analysis estimate is available as the experiment is being observed, why not use this estimate to improve the parameter value *while the system is evolving* and thus optimize the system during a single experiment! We should note that this keeps introducing transient phenomena into the system, and therefore, none of our analytic results are applicable, and as yet no other analysis is available on the convergence of this scheme. (Standard stochastic approximation results do not apply either.) On the other hand, from a practical point of view, our experimental results, and also those of Meketon (1983) are exciting and worth mentioning, even if this technique is still in a "heuristic" stage. The excitement is basically due to the fact that we have observed convergence at very fast rates.

The general optimization problem that we solve, for a $GI/G/1$ queue, can be stated as follows: find θ^* to minimize some cost function $J(\bar{T}, \theta)$. Note that $\bar{T}$ is itself a function of θ, to be estimated experimentally. We assume J is differentiable in its arguments. Then

$$\frac{dJ}{d\theta} = \frac{\partial J}{\partial \bar{T}} \frac{d\bar{T}}{d\theta} + \frac{\partial J}{\partial \theta}. \tag{8.2}$$

We will use perturbation analysis to estimate $d\bar{T}/d\theta$. Hence we can get $dJ/d\theta$ from (8.2), and use this to update the value of θ.

Our modification of Meketon's method for updating θ, is to use an algorithm analogous to Kesten's (1958) accelerated version of the Robbins-Monro procedure. The idea is to estimate the gradient $dJ/d\theta$ for a fixed number of customers, say L, using perturbation analysis as explained. Then the value of θ is adjusted by a step size proportional to the gradient. We then re-start the gradient estimation for the next L customers. Every time there is a reversal in the sign of the $dJ/d\theta$ estimate, the step size proportion is decreased. The algorithm stops after the step size proportion becomes smaller than a given value. The procedure is summarized in Algorithm 3. (The usual stochastic approximation approach would update K, in Algorithm 3, at each iteration. This can lead to very slow convergence if θ is started far from the optimal value.) Our procedure is now tested against some queueing systems.

Algorithm 3.

0. INPUTS
   A = starting step-size proportion
   AMIN = stopping step-size proportion
   THETA = starting value of parameter
   L = number of customers per iteration
1. INITIALIZE
   $K = 1$
2. GRADIENT
   Obtain EST = perturbation analysis estimate of $d\bar{T}/d\theta$
   over next L customers
3. UPDATE THETA
   NEWGRAD = $(\partial J/\partial \bar{T}) * \text{EST} + \partial J/\partial \theta$
   THETA = THETA + $(A/K) *$ NEWGRAD
4. UPDATE K
   if $K = 1$ then OLDGRAD = NEWGRAD
   if sign(OLDGRAD) $\neq$ sign(NEWGRAD) then $K = K + 1$
   OLDGRAD = NEWGRAD
5. STOPPING CRITERION
   if $(A/K) \geq$ AMIN go to step 2
   stop

EXAMPLE 8.4: Optimization of $M/M/1$ Queue. We start with this analytically tractable system, with notation as in Example 8.1. Our objective is to find θ^* to minimize $c_0\bar{T} + c_1/\theta$, where c_0 is the cost per unit of time spent by a customer in the system, while c_1 is the cost per unit speed of a server. The analytic solution is $\theta^* = \sqrt{c_1}/(\sqrt{c_0} + \lambda\sqrt{c_1})$. With $c_0 = 1$, $\lambda = 0.01$, and $c_1 = 22500$, the theoretical optimum is $\theta^* = 60$ with corresponding objective value $J(\theta^*) = 525$. We implemented our algorithm for two cases, $L = 5$ and $L = 10$. The results are in Table 4. The value θ_{pa} is the final value of θ attained by our algorithm. The interesting point is that we obtained near-optimal behaviour in under 1000 customers in both cases. If we measure optimality in terms of cost (J) rather than parameter value, it is seen the algorithm arrives very close indeed. The reason that 1000 customers can be considered as "fast" convergence is that it typically takes 100,000 customers to get reasonable confidence intervals just for the mean system time in this queue (e.g. Table 1, and also Meketon and Heidelberger 1982). Also, since the first version of this manuscript was written, Suri and Leung

TABLE 4

Single-Run Optimization of M/M/1 Queue

	Initial Values			Final Values					
Update Freq L	θ	$J(\theta)$	A	AMIN	θ_{pa}	Error in θ_{pa}	$J(\theta_{pa})$	Error in $J(\theta_{pa})$	Run Length
5	20	1150	1	0.02	59.55	0.75%	525.05	0.01%	975
10	20	1150	1	0.05	59.47	0.88%	525.07	0.01%	610

Notes.

1. Run Length is the total number of customers served.
2. See §8 and Example 8.4 for details of the experiments.

TABLE 5
Single-Run Optimization of G/G/1 Queue (Experiments to Test for Local Minimum)

Parameter Value θ	Method Used	Quantity Estimated	
		J	$dJ/d\theta$
70.61	Experiment including P/A	430.36 ± 0.78	−0.977 ± 0.067
73.61	Experiment including P/A	429.02 ± 0.96	0.134 ± 0.098
76.61	Experiment including P/A	431.31 ± 1.31	1.405 ± 0.140

Notes.
1. P/A stands for perturbation analysis.
2. See §8 and Example 8.5 for details of the experiments.

(1986) have performed an extensive experimental study which clearly shows the speed of this algorithm compared with other approaches.

EXAMPLE 8.5: Optimization of $GI/G/1$ Queue. Now we consider the same cost function as in Example 8.4, but the system is the one from Example 8.3. We choose $L = 5, A = 1$, AMIN = 0.02. Starting at $\theta = 20$, our algorithm converged at $\theta_{pa} = 73.61$, after serving 1145 customers. In order to check whether this is indeed a (local) minimum or not, we used additional simulations to estimate the cost at θ_{pa} and two neighbouring points. The results are shown in Table 5. Now we illustrate another interesting point. Although the mean value of J, estimated at θ_{pa}, is lower than its neighbors in Table 5, the confidence limits are wide enough that we cannot derive a conclusion about θ_{pa} being superior to its neighbors. (More experiments would be required.) However, we also used perturbation analysis to estimate the gradient of J at each of the three points, while conducting the same simulations. If we believe the perturbation analysis estimates, then the experiment is conclusive at the 95% confidence level, since the gradient at 70.61 is conclusively negative, while that at 76.61, positive. So this assures us that these two points bracket the minimum, and θ_{pa} is a reasonable stopping point for our algorithm. Here we see a side benefit of perturbation analysis, as it gave us additional information, and in fact tighter confidence intervals, by which to judge the outcome of some experiments.

9. Extensions

The main aim of the current paper was to study the behaviour of perturbation analysis applied to the $M/G/1$ queue. Under additional restrictions, we have recently been able to show similar results for $GI/G/1$ systems as well (Zazanis and Suri 1986b). We were also able to show that perturbation analysis gives unbiased estimates for higher moments of the response time as well, in the case of an $M/M/1$ queue (Zazanis 1986). As illustration of some uses of perturbation analysis, we gave here experiments on $GI/G/1$ queues, as well as a fast optimization algorithm. In a similar way, we can apply these methods to the case where a *vector* of parameters needs to be optimized. In this case the efficiency gained through perturbation analysis can be substantial.

Some other interesting issues arise out of this work. It turns out that we can also estimate second derivatives for a $GI/G/1$ queue, using perturbation analysis on a single sample path (Zazanis and Suri 1986b). Also, as mentioned, we can show that whenever perturbation analysis estimates for gradients are unbiased, the corresponding confidence interval are always tighter than those obtained by doing a second experiment (Zazanis and Suri 1986a). Finally, we have considered here a classic, but simple, system

and it would be useful to see similar results for other practical systems. Steps in this direction are Cao (1985b) and Zazanis and Suri (1986b).[1]

[1] This work was partly supported by the U.S. Office of Naval Research Contracts N00014-75-C-0648 and N00014-79-C-0776, and by NSF Grant ENG82-13680 while the authors were at Harvard University. We would like to thank Arnold Buss of Cornell University and especially an anonymous referee for comments that led to improved proofs and stronger results.

Appendix

Here we will establish the integrability of the random variables h and g and we will supply the proof of Theorem 2.

LEMMA 1. *The random variables h and g, defined in* (6.1) *and* (6.4), *are integrable, i.e.* $E|h|$ *and* $E|g|$ *exist and are finite.*

PROOF. First we will establish the integrability of h. We start with the inequality

$$|h| = |\sum_{i=1}^{n}\sum_{j=1}^{i}\phi(x_j)| \le \sum_{i=1}^{n}\sum_{j=1}^{i}|\phi(x_j)| \le \sum_{i=1}^{n}\sum_{j=1}^{n}|\phi(x_j)| \quad \text{or}$$

$$|h| \le n\sum_{j=1}^{n}|\phi(x_j)|. \tag{App. 1}$$

Let us set $|\phi(x_j)| = \zeta_j$ for convenience. Now we will show that the expectation of the r.v. on the RHS of (App. 1) is finite. Indeed,

$$E(n\sum_{j=1}^{n}\zeta_j) = E(n\sum_{j=1}^{n}[\zeta_j - E(\zeta_j)] + n^2E(\zeta_1))$$

$$= E(n^2)E(\zeta_1) + E(n\sum_{j=1}^{n}[\zeta_j - E(\zeta_j)]). \tag{App. 2}$$

Using the Cauchy-Schwartz inequality we get

$$E(n\sum_{j=1}^{n}[\zeta_j - E(\zeta_j)]) \le (En^2)^{1/2}[E(\sum_{j=1}^{n}[\zeta_j - E(\zeta_j)])^2]^{1/2}. \tag{App. 3}$$

Now notice that ζ_i, $i = 1, 2, \cdots$ are i.i.d. random variables with $E[\zeta_1] < \infty$ and $E[\zeta_1^2] < \infty$ (the finiteness of the first two moments of $\zeta_1 = |\phi(x_1)|$ is postulated in A3). Also n, the number of customers in the busy period, is a stopping time with $E[n] < \infty$ (the finiteness of this expectation is guaranteed by the stability of the system). Hence we can apply one of Wald's identities (Shiryayev 1984) to the RHS of (App. 3) and conclude that

$$E(\sum_{j=1}^{n}\zeta_j - E(\zeta_j))^2 = \text{Var}\,(\zeta_1)E(n) < \infty. \tag{App. 4}$$

Hence, from (App. 2), (App. 3) and (App. 4) it follows that

$$E(n\sum_{j=1}^{n}|\phi(x_j)|) \le E(n^2)E(|\phi(x_j)|) + [E(n^2)\,\text{Var}\,(|\phi(x_j)|)E(n)]^{1/2} < \infty. \tag{App. 5}$$

(The fact that the system is stable and that $\overline{x^2} < \infty$ implies the finiteness of $E(n)$ and $E(n^2)$, e.g. see Kleinrock 1975.) From (App. 1) and (App. 5) we have finally that

$$E(|h|) \le E(n\sum_{j=1}^{n}|\phi(x_j)|) < \infty. \tag{App. 6}$$

The integrability of g is an immediate consequence of the inequalities

$$|g| < \sum_{j=1}^{n}|\phi(x_j)| < n\sum_{j=1}^{n}|\phi(x_j)|$$

and of the integrability of $n\sum_{j=1}^{n}|\phi(x_j)|$ which was established above.

Now we give the proof of Theorem 2 which establishes the asymptotic unbiasedness of the perturbation analysis estimates.

PROOF OF THEOREM 2. We need to show that

$$\lim_{M\to\infty} E\left\{\frac{\sum_{m=1}^{M}h_m}{\sum_{m=1}^{M}n_m}\right\} = \frac{d\bar{T}}{d\theta} \quad \text{w.p. 1.} \tag{App. 7}$$

We note that, since n_m is the number of customers in the mth busy period, $n_m \geq 1$. Hence,

$$\frac{\sum_{m=1}^{M} h_m}{\sum_{m=1}^{M} n_m} \leq \frac{1}{M} \sum_{m=1}^{M} |h_m|. \tag{App. 8}$$

Since $E|h_m|$ exists, as we showed in the lemma above, by the strong law of large numbers

$$\frac{1}{M} \sum_{m=1}^{M} |h_m| \to E|h_1| \qquad \text{w.p. 1.} \tag{App. 9}$$

So we see that the sequence $\sum_{m=1}^{M} h_m / \sum_{m=1}^{M} n_m$ is dominated by the sequence of random variables

$$R_M(\omega) = \frac{1}{M} \sum_{m=1}^{M} |h_m|.$$

Now $E\{R_M(\omega)\} = E|h_1|$, and combining this with (App. 9) we see that

$$\lim_{M\to\infty} E\{R_M(\omega)\} = E\{\lim_{M\to\infty} R_M(\omega)\} \qquad \text{w.p. 1.} \tag{App. 10}$$

Hence we can apply a generalization of the Lebesgue Convergence Theorem (Royden 1968) and conclude that

$$\lim_{M\to\infty} E\left\{\frac{\sum_{m=1}^{M} h_m}{\sum_{m=1}^{M} n_m}\right\} = E\left\{\lim_{M\to\infty} \frac{\sum_{m=1}^{M} h_m}{\sum_{m=1}^{M} n_m}\right\}. \tag{App. 11}$$

But (6.18) establishes the strong consistency of the perturbation analysis estimates, hence we have

$$\lim_{M\to\infty} \frac{\sum_{m=1}^{M} h_m}{\sum_{m=1}^{M} n_m} = \frac{d\bar{T}}{d\theta} \qquad \text{w.p. 1.} \tag{App. 12}$$

From (App. 11) and (App. 12) the asymptotic unbiasedness of the estimates is established.

Next we give a sketch of the

PROOF OF THEOREM 4. We need to show that

$$\lim_{M\to\infty} E\left\{\frac{\sum_{m=1}^{M} q_m}{\sum_{m=1}^{M} n_m}\right\} = E\left\{\lim_{M\to\infty} \frac{\sum_{m=1}^{M} q_m}{\sum_{m=1}^{M} n_m}\right\} = \frac{d\bar{T}}{d\lambda}. \tag{App. 13}$$

The proof is exactly the same as that of Theorem 2. The argument here is further simplified by the fact that $q_m \geq 0$ and $E[q_m] < \infty$ as was shown in §7.2. Hence a counterpart of Lemma 1 is not needed here.

Finally a proof for equation (7.12) is presented. This proof was given by an anonymous referee.

PROOF OF (7.12). Let

$$S_M = \sum_{m=1}^{M} (Y_m + I_m), \qquad M \geq 1, \qquad (S_0 \equiv 0), \tag{App. 14}$$

denote the ending of the Mth busy cycle of the $M/G/1$ queue, and let

$$\nu(t) = \max\{M: S_M \leq t\}, \qquad t \geq 0, \tag{App. 15}$$

denote the number of busy cycles that are completed up to time t. Thus,

$$\hat{z}(t) = \begin{cases} t - S_{\nu(t)}, & t \leq S_{\nu(t)} + Y_{\nu(t)+1}, \\ 0, & \text{otherwise,} \end{cases} \tag{App. 16}$$

denotes the age of the busy period in progress at time t. Given a fixed $\xi \geq 0$, we can use the standard renewal argument to show that

$$\lim_{t\to\infty} P(\hat{z}(t) > \xi) = E[\max(0, Y_1 - \xi)]/(E[Y_1] + E[I_1]) \qquad \text{for all} \qquad \xi \geq 0. \tag{App. 17}$$

If we let $\hat{z}$ denote a random variable whose complementary cumulative distribution function is given by the right-hand side of (App. 17), then we have $\hat{z}(t) \Rightarrow \hat{z}$ as $t \to \infty$, where of course $\Rightarrow$ designates weak convergence, a continuous-time version of the discrete-time result $z_k \Rightarrow z$ as $k \to \infty$ that was established prior to (7.10). Then, the ergodicity of the $M/G/1$ queue when $\lambda\bar{x} < 1$ ensures that

$$\lim_{t\to\infty} \frac{1}{t} \int_0^t 1_{\{\hat{z}(s)>\xi\}}\, ds = P(\hat{z} > \xi) \qquad \text{w.p. 1} \qquad \text{and} \tag{App. 18}$$

$$\lim_{t\to\infty} \frac{1}{A(t)} \int_0^t 1_{\{\hat{z}(s)>\xi\}}\, dA(s) = P(z > \xi) \qquad \text{w.p. 1} \tag{App. 19}$$

where $A(t)$ is the number of Poisson arrivals to the system in $[0, t]$ (see also equation (A.8) of Crane and Iglehart 1975). Now Theorem 1 of Wolff (1982) ensures that the left-hand sides of (App. 18) and (App. 19) are equal with probability 1 and since ξ was selected arbitrarily, we see that $P(\hat{z} > \xi) = P(z > \xi)$ for all $\xi \geq 0$. It follows immediately that

$$E[z] = \int_0^\infty P(z > \xi)d\xi = \int_0^\infty P(\hat{z} > \xi)d\xi = E[\hat{z}]$$

$$= \frac{E[Y_1^2]}{2(E[Y_1] + E[I_1])}. \qquad \text{(App. 20)}$$

References

BICKEL, P. J. AND K. A. DOKSUM, *Mathematical Statistics: Basic Ideas and Selected Topics*, Holden-Day, San Francisco, 1977.

CAO, X. R., Division of Applied Sciences, Harvard University, unpublished memorandum, 1983.

———, "Convergence of Parameter Sensitivity Estimates in a Stochastic Experiment," *IEEE Trans. Automat. Control*, 30, 9 (1985a), 845–853.

———, "On the Sample Functions of Queueing Networks with Applications to Perturbation Analysis," submitted to *Oper. Res.*, (1985b).

CASSANDRAS, C. AND Y. C. HO, "An Event Domain Formalism for Sample Path Perturbation Analysis of Discrete Event Dynamic Systems," *IEEE Trans. Automat. Control*, (1985), 1217–1221.

CRANE, M. A. AND D. L. IGLEHART, "Simulating Stable Stochastic Systems. III. Regenerative Processes and Discrete-Event Simulations," *Oper. Res.*, 23 (1975), 33–35.

GLYNN, P. W., "Construction of Process-Differentiable Representations for Parametric Families of Distributions," Technical Report, Mathematics Research Center, University of Wisconsin-Madison, 1987.

FISHMAN, G. S., *Principles of Discrete Event Simulation*, John Wiley, New York, 1978.

HEIDELBERGER, P., X. R. CAO, M. A. ZAZANIS AND R. SURI, "Convergence Properties of Infinitesimal Perturbation Analysis," submitted to *Management Sci.*, 1987.

HO, Y. C., Personal communication, 1982.

———, M. A. EYLER AND T. T. CHIEN, "A Gradient Technique for General Buffer Storage Design in a Serial Production Line," *Internat. J. Production Res.*, 17, 6 (1979), 557–580.

———, ——— AND ———, "A New Approach to Determine Parameter Sensitivities of Transfer Lines," *Management Sci.*, 29, 6 (1983), 700–714.

——— AND X. CAO, "Perturbation Analysis and Optimization of Queueing Networks," *J. Optim. Theory Appl.*, 40, 4 (1983), 559–582.

———, ——— AND C. CASSANDRAS, "Infinitesimal and Finite Perturbation Analysis for Queueing Networks," *Automatica*, 19, 4 (1983), 439–445.

——— AND C. CASSANDRAS, "A New Approach to the Analysis of Discrete Event Dynamic Systems," *Automatica*, 19, 2 (1983), 149–167.

——— AND SHU LI, "Extended Perturbation Analysis for Multiclass Networks," submitted to *IEEE Trans. Automat. Control*, 1987.

KESTEN, H., "Accelerated Stochastic Approximation," *Ann. Math. Statist.*, 29 (1958), 41–49.

KLEINROCK, L., *Queueing Systems. I*, John Wiley, New York, 1975.

MEKETON, M. S., "A Tutorial on Optimization in Simulations," presented at Winter Simulation Conference (WSC-83), 1983.

——— AND P. HEIDELBERGER, "A Renewal Theoretic Approach to Bias Reduction in Regenerative Simulations," *Management Sci.*, 28, 2 (1982), 173–181.

ROYDEN, H. L., *Real Analysis*, Macmillan, New York, 1968, p. 89.

SHIRYAYEV, A. N., *Probability*, Springer Verlag, Berlin and New York, 1984.

SURI, R., "Implementation of Sensitivity Calculations on a Monte Carlo Experiment," *J. Optim. Theory Appl.*, 40, 4 (1983), 625–630.

———, "Infinitesimal Perturbation Analysis of General Discrete Event Dynamic Systems," *J. Assoc. Comput. Mach.*, 34, 3 (1987), 686–717.

——— AND X. CAO, "The Phantom Customer and Marked Customer Methods for Optimization of Closed Queueing Networks with Blocking and General Service Times," *ACM Performance Evaluation Rev.*, 12 (1983), 243–256.

——— AND J. W. DILLE, "A Technique for On-Line Sensitivity Analysis of Flexible Manufacturing Systems," *Ann. Oper. Res.*, (1985), 379–384.

——— AND Y. T. LEUNG, "Single Run Optimization of Discrete Event Systems: Experimental Investigation for the $M/M/1$ Queue," Working Paper, Dept. of Industrial Engineering, University of Wisconsin-Madison, 1986.

WOLFF, R. W., "Poisson Arrivals See Time Averages," *Oper. Res.*, 30, 2 (1982), 223–231.

ZAZANIS, M. A., "Unbiasedness of Infinitesimal Perturbation Analysis Estimates for Higher Moments of the Response Time of an $M/M/1$ Queue," submitted to *Oper. Res.*, (1986).

——— AND R. SURI, "Comparison of Perturbation Analysis with Conventional Sensitivity Estimates for Stochastic Systems," Working paper #86-123, Dept. of Industrial Engineering, University of Wisconsin-Madison, 1986a.

——— AND ———, "Estimating First and Second Derivatives of Response Time for $GI/G/1$ Queues from a Single Sample Path," Working paper #86-121, Dept. of Industrial Engineering, University of Wisconsin-Madison, 1986b.