Les cahiers du GERAD ISSN: 0711–2440

Comparing Alternative Methods for Derivative Estimation when IPA Does Not Apply Directly

F.J. Vázquez-Abad
P. L'Ecuyer

G–91–30

June 1991

Comparing Alternative Methods for Derivative Estimation when IPA Does Not Apply Directly.

Felisa J. Vázquez-Abad
INRS Télécommunications,
Université du Québec, 3 Place du Commerce, Ile-des-Soeurs,
Québec, H3E 1H6, Canada

Pierre L'Ecuyer
GERAD and Département d'IRO,
Université de Montréal, C.P. 6128, Succ. A,
Montréal, H3C 3J7, Canada

June 1991

Abstract

Infinitesimal Perturbation Analysis (IPA) is perhaps the most efficient derivative estimation method for many practical discrete-event stochastic systems, whenever it applies. But there are many situations where it does not apply directly. Alternative methods such as Likelihood Ratios (LR), Finite Perturbation Analysis (FPA), Smoothed Perturbation Analysis (SPA), Rare Perturbation Analysis (RPA), and a few others, have been proposed and could be used when IPA does not apply directly. In this paper, we discuss some links that exist between these methods and explain them by showing how each of them can be applied to a specific example, namely to estimate the derivative of the expected number of customers per regenerative cycle in a $GI/G/1$ queue, with respect to parameters of the interarrival and/or service-time distributions. We also give the results of numerical experiments to compare the performances of these methods.

Résumé

L'analyse de perturbation infinitesimale (IPA), lorsqu'elle s'applique, semble être la méthode d'estimation de dérivées la plus efficace pour la plupart des systèmes stochastiques à événements discrets. Mais dans plusieurs cas, cette méthode ne s'applique pas directement. D'autres méthodes, telles la méthode du rapport de vraisemblance (LR), l'analyse de perturbation finie (FPA), l'analyse de perturbation lissée (SPA), l'analyse de perturbation rare (RPA), etc., ont été proposées et peuvent être utilisées lorsque IPA ne s'applique pas directement. Dans cet article, nous discutons de certains liens qui existent entre ces méthodes et expliquons ces dernières en montrant comment les appliquer dans le cas d'un exemple simple, en l'occurence pour estimer la dérivée du nombre espéré de clients par cycle regénératif dans une file $GI/G/1$, par rapport á des paramètres de la loi inter-arrivées ou de la loi des durées de service. Nous donnons aussi les résultats de quelques expériences numériques comparant les performances de ces méthodes.

1. Introduction

During the past decade there has been an increasing number of new methods and algorithms for the estimation of the sensitivity of complex queueing systems with respect to some parameters of the underlying distributions. Among the many important recent references to this area, we can cite for instance Glasserman and Gong (1990), Glynn (1990), Heidelberger et al. (1988), Ho (1987), Ho and Strickland (1990), L'Ecuyer (1990), Reiman and Weiss (1989), Rubinstein (1989), Simon (1989), and Suri (1989).

Both theoretical and empirical results suggest that for most discrete-event stochastic systems of interest, Infinitesimal Perturbation Analysis (IPA) is usually the most efficient derivative estimation method when it applies. But in many cases, IPA does not apply directly. Various alternative methods have been proposed and could be used in these situations. Such methods include the use of a Likelihood Ratio (LR) or Score function (SF), Finite Perturbation Analysis (FPA), Smoothed Perturbation Analysis (SPA), Conditional Infinitesimal Perturbation Analysis (CIPA), Rare Perturbation Analysis (RPA), and a few others. In this paper, we discuss some links that exist between these methods and explain them by showing how each of them can be applied to a specific example, namely to estimate the derivative of the expected number of customers per regenerative cycle in a $GI/G/1$ queue, with respect to parameters of the interarrival and/or service-time distributions. Applying those methods is not straightforward, even for such a simple example. The development that we follow for deriving our estimators can be generalized or adapted to many similar problems. We also look at the performance of our estimators on that specific example through numerical illustrations. Of course, the best method for that example is not necessarily the best in general, but nevertheless, numerical experiments can give some insight into what goes on. The estimators that we examine are all based on a *single* simulation run. That excludes the family of finite-difference estimators with or without common random numbers (see, e.g., L'Ecuyer and Perron, 1990).

Section 2 introduces the model and explains why straightforward IPA will not work in that case. In Section 3, we recall briefly how the likelihood ratio (LR) method applies, yielding a simple unbiased derivative estimate. It is well known, however, that LR estimators are plagued with a large variance, especially when the regenerative cycles (busy periods) are long. We are therefore looking for better estimators. In Section 4, in the context of our $GI/G/1$ queueing example, we describe a finite perturbation analysis scheme called Finite-Difference Phantom RPA. This approach, based on the thinning of a point process, was introduced in Vázquez-Abad and Kushner (1990). It is inspired by ideas used in Suri and Cao (1986) to study the sensitivities with respect to the number of jobs circulating in a closed queueing network. In Section 5, following Brémaud and Vázquez-Abad (1991), we take the RPA approach to the limit and add some smoothing (conditioning) to obtain an infinitesimal phantom RPA estimator. In Section 6, we introduce two new estimators based on Smoothed Perturbation Analysis (SPA), that is estimators obtained by applying IPA after having replaced the objective function by a conditional expectation. Similar estimators have been studied in Glasserman and Gong (1990), L'Ecuyer and Perron (1990), and Wardi et al. (1990), for different problems. These SPA estimators can be used to estimate sensitivities

with respect to parameters of either the interarrival or service-time distributions. Further, we explain in Section 7 how any estimator of the derivative with respect to the arrival rate (whether or not the arrivals are Poisson) can be transformed into an estimator of the derivative with respect to the (average) service rate, and vice-versa. These kinds of indirect estimators are called *surrogate* estimators, following the terminology of Vázquez-Abad and Kushner (1990). In Ho and Cao (1985), the same idea was used to estimate sensitivities with respect to routing parameters via sensitivities with respect to service rates for a closed network. Section 8 reports the results of our numerical experiments and Section 9 gives concluding remarks. It turns out that our SPA estimators of Section 6 are the most effective for the cases that we have examined.

2. Number of customers per busy cycle in a $GI/G/1$ queue

We consider a $GI/G/1$ queue with inter-arrival time distribution F_λ and service-time distribution G_μ. For simplification, suppose that F_λ and G_μ have respective densities f_λ and g_μ. Here, λ and μ are continuous parameters with respect to which we might want to estimate the derivative of some "performance measure" expressed as a mathematical expectation. We will use θ as a generic name that could designate either λ or μ. The "performance measure" that we will concentrate on in this paper is the expected number of customers in a busy cycle.

The evolution of the $GI/G/1$ queue can be described conveniently as follows throught Lindley's equations (1–2) below. Suppose that the queue is started empty and let i denote the i-th customer in the system. For each $i \geq 1$, let:

$$\begin{aligned} A_i &= \text{interarrival time between customer } i \text{ and } i+1; \\ S_i &= \text{service requirement of customer } i; \\ W_i &= \text{waiting time of customer } i; \\ X_i &= \text{sojourn time of customer } i. \end{aligned}$$

Then, one has $W_1 = 0$, $X_1 = S_1$, and for each $i > 1$,

$$W_i = \max(0, X_{i-1} - A_{i-1}); \tag{1}$$

$$X_i = W_i + S_i. \tag{2}$$

That process evolves according to a probability measure that depends on the parameter θ. Assuming that the queue is stable, this is a regenerative "discrete-time" Markov chain, where the "time" is viewed as representing the customer number i (see Asmussen, 1987). The regenerative points can be defined as the indexes of the customers who find the system empty when they arrive. Let τ denote the number of customers in a given regenerative cycle, say the first one, and $\ell(\theta) = E_\theta[\tau]$ be its expectation at parameter value θ.

A standard way to estimate $\ell(\theta)$ is to run the system for say n regenerative cycles and count the total number of customers in those n cycles, divided by n. This gives the estimator

$$\hat{\ell}_n(\theta) = \frac{1}{n}\sum_{j=1}^{n} \tau_j = C_n/n, \tag{3}$$

where τ_j represents the number of customers in the j-th regenerative cycle and C_n is the total number of customers during the n cycles. But estimating $\ell'(\theta)$, the derivative of $\ell(\theta)$ with respect to θ, is less simple. An estimation of $\ell'(\theta)$ is necessary, for example, when one wishes to estimate the derivative of an average "cost" per customer over an infinite horizon, using LR and a regenerative approach (see, e.g., Glynn 1990).

One approach for building efficient derivative estimators is IPA. The basic idea of IPA is essentially to use the derivative of the sample estimator, for fixed underlying $U(0,1)$ uniform random numbers, as an estimator of the derivative of the expectation. But in our case here, when the sequence of $U(0,1)$ variates that are used to generate the interarrival and service times are fixed, (3) is piecewise constant as a function of θ. Therefore, whenever the derivative of (3) is defined, it is zero. This is clearly not a worthwhile estimator for $\ell'(\theta)$, which is usually not zero in the cases of interest. This means that straightforward IPA does not apply in this case: we cannot interchange the derivative and expectation.

3. Applying the LR method

A now well known alternative to IPA is the likelihood ratio (LR) approach, sometimes called the score function (SF) method. That method can be traced back to Aleksandrov et al. (1968). More recent references include Glynn (1990), Reiman and Weiss (1989), and Rubinstein (1989). L'Ecuyer (1990) has shown how LR, SF, and IPA can be presented into a unified framework under which IPA can be viewed as a special case of LR. L'Ecuyer and Perron (1990) showed how SPA also fits quite well into this framework and how LR can be viewed as a special case of IPA.

The standard LR derivative estimator is simply obtained as the product of the performance measure of interest (here, τ), by the so-called *score function*. For one regenerative cycle, this yields (see L'Ecuyer 1990 or Glynn 1990):

$$\psi^{LR} = \tau \sum_{i=1}^{\tau} \frac{\partial}{\partial\theta}(\ln f_\theta(A_i) + \ln g_\theta(S_i)), \tag{4}$$

where A_i and S_i are considered fixed when taking the derivative with respect to θ. When $\theta = \lambda$ [resp. $\theta = \mu$], the derivative of $\ln g_\mu(S_i)$ [resp. $\ln f_\lambda(A_i)$] vanishes. For n regenerative cycles, the derivative estimator is obtained by averaging out the n values of ψ^{LR} associated with these cycles.

4. Finite-difference RPA

Ordinary perturbation analysis is based on the idea of introducing tiny perturbations to the system, so that the nominal and perturbed paths differ almost surely by a very small amount. In contrast, Rare Perturbation Analysis (RPA) is based on introducing the perturbations only rarely, so that the nominal and perturbed paths differ only very rarely, but the amount of the difference may be large. For more on RPA, see Vázquez-Abad and Kushner (1990) and Brémaud and Vázquez-Abad (1991).

To simplify the presentation here, we assume that $\theta = \lambda$ and that the arrival process is Poisson with rate λ. We therefore have a $M/G/1$ queue. Suppose that λ is reduced to $\lambda - \Delta\lambda$, for some small constant $\Delta\lambda > 0$. Standard finite PA will take that reduction into account by sliding along the arrival times slightly into the future: each interarrival time A_i is multiplied by $\lambda/(\lambda + \Delta\lambda)$. RPA, in contrast, will thin down the process by removing any given arrival with probability $\Delta\lambda/\lambda$. The path of the original process is called the *nominal* path, while the path of the process with some arrivals removed is called the *phantom* path. The customers whose arrival is actually removed are called *phantom* customers.

We now explain how the finite-difference RPA derivative estimator can be computed in a *single run*. That derivative estimator will be the difference between the average number of customers per busy cycle in the nominal path and the average number of customers per busy cycle in the phantom path, divided by $\Delta\lambda$. So, while simulating the nominal path, we need to compute the total number of non-phantom customers as well as the number of busy cycles in the phantom path.

To each customer i, associate a Bernoulli $(\Delta\lambda/\lambda)$ random variable I_i which is independent of all the A_j's, S_j's, and other I_j's. When $I_i = 1$, customer i is a phantom customer in the phantom path. The arrival process of those customers which are not phantoms is a Poisson process with rate $\lambda - \Delta\lambda$. (Note that this development also applies to the situation where the arrival process is a more general point process.) Computing the total number of non-phantom customers is trivial: just sum the $(1 - I_i)$'s. It remains to see how to compute the number of busy cycles in the phantom path.

Observe that removing customers can split busy cycles, but can never merge them. Each time a busy cycle starts in the nominal path, then a busy cycle must also start in the phantom path (unless the first customer in that busy cycle has been phantomized, in which case the corresponding busy cycle in the phantom path will start with the arrival of the next non-phantom customer). Note that here, the non-phantom customers in the phantom path keep the same numbers and same attributes that they had in the nominal path. For example, if the third customer is the first phantom, then the third non-phantom customer in the phantom path will have service time S_4, not S_3. Let $\tilde{W}_i$ and $\tilde{X}_i$ denote the waiting time and system time, respectively, of customer i in the phantom path. When i is a phantom, these quantities can be viewed as "phantom" times but are nevertheless well defined. The simplest way of taking into account that a customer has been removed, in Lindley's equations, is to

replace its service time by zero. Using that trick, Lindley's equations for the phantom system become: $\tilde{W}_1 = W_1 = 0$, $\tilde{X}_1 = (1 - I_1)S_1$, and for $i > 1$,

$$\tilde{W}_i = \max(0, \tilde{X}_{i-1} - A_{i-1}); \tag{5}$$

$$\tilde{X}_i = \tilde{W}_i + (1 - I_i)S_i. \tag{6}$$

It follows immediately that $\tilde{W}_i \leq W_i$ and $\tilde{X}_i \leq X_i$ for each i. Note that from the sequence $\{(A_i, S_i, X_i, \tilde{X}_i),\ i = 1, 2, 3, \ldots\}$, one can compute finite-difference estimators for many different performance measures. Computing the number of busy cycles is equivalent to computing the number of customers who are first in their busy cycle, i.e. whose waiting time is zero. Therefore, the number of busy cycles in the phantom path is equal to the number of busy cycles in the nominal path, *plus* the number of non-phantom customers who did wait in the nominal path and are not waiting any more in the phantom path (i.e. such that $\tilde{W}_i = 0 < (1 - I_i)W_i$), *minus* the number of customers who did not wait in the nominal path but are now phantoms (i.e. such that $W_i = 0$ and $I_i = 1$). To compute the last two numbers during the nominal simulation, it suffices to maintain $(\tilde{W}_i, \tilde{X}_i)$ and a counter D. At the beginning of the simulation, initialize D to zero. Whenever $\tilde{W}_i = 0 < (1 - I_i)W_i$, add one to the counter and whenever $W_i = 1 - I_i = 0$, subtract one from the counter. Let n, C, and $\tilde{C}$ denote respectively the number of regenerative cycles in the nominal path, the total number of customers in the nominal path, and the total number of non-phantom customers in the phantom path. Then, the finite-difference RPA estimator becomes:

$$\psi_n^{FD-RPA} = \frac{1}{\Delta\lambda}\left(\frac{C}{n} - \frac{\tilde{C}}{n + D}\right). \tag{7}$$

This estimator satisfies

$$E_\lambda[\psi_n^{FD-RPA}] = \frac{1}{\Delta\lambda}\left(E_\lambda[\tau] - E_{\lambda-\Delta\lambda}[\tau]\right).$$

Therefore, it is biased for $\ell'(\lambda)$, due to the finite differences. However, it can be computed in a single run and could also be used in more complex situations where IPA would not apply directly. See Vázquez-Abad and Kushner (1990) for an extension to a queueing network problem.

5. An average infinitesimal RPA approach

The bias on the estimator (7) can be reduced by reducing $\Delta\lambda$, but at the cost of increasing the variance. The reason is that when $\Delta\lambda$ is very small, customer phantomizations become rare events that have a large impact on the estimator value. Further, that finite-difference estimator cannot be taken to the limit directly, that is take the limit as $\Delta\lambda \to 0$, because when $\Delta\lambda$ is small enough, there are no more phantom customers and (7) becomes zero. So, we have exactly the same problem as we had with IPA in Section 2.

Brémaud and Vázquez-Abad (1991) have developed a less straightforward way of taking RPA to the limit. Their approach yields an unbiased derivative estimator. Let us sketch this approach in the context of our example. Suppose that we simulate the system for *one* regenerative cycle and let τ be the number of customers in that cycle in the nominal path. The simulation starts at the arrival of the first customer, both in the nominal and phantom paths. After that first customer has arrived, since the arrival rate in the phantom process is smaller, some of the $\tau - 1$ arrivals that follow can be phantomized, with the appropriate probabilities. The number K of phantom customers is a binomial random variable with parameters $\tau - 1$ and $p = \Delta\lambda/\lambda$. To estimate the derivative, we condition on K: for each integer k in $\{0, \ldots, \tau - 1\}$, we multiply $P[K = k]$ by the derivative of the expected number of customers in the cycle, *conditional* on the event that $K = k$ and on the sequence of interarrival and service times in the nominal path. We then sum up over all values of k. For $k = 0$, the conditional derivative is clearly zero. When $\Delta\lambda$ becomes infinitesimal, it can be shown that the event $\{K \geq 2\}$ can be neglected, so that the problem comes down to estimating the derivative of the expected number of customers in the first cycle given the nominal path and given that $K = 1$. When $K = 1$, each customer after the first one has the same chance of being the phantom, that is $1/(\tau - 1)$. So, for $i = 2, \ldots, \tau$, we will look at what happens when customer i is the (only) phantom in its cycle, and take the average over all these values of i. Let $\tilde{\tau}^{(i)}$ denote the number of customers in the first cycle of the phantom path when i is the only phantom. This can be computed using the same kind of Lindley equations as in the previous section, with $I_i = 1$ and $I_j = 0$ for $i \neq j$. (Note that we need one set of Lindley equations for each i, $i = 2, \ldots, \tau$. This implies a non-negligible overhead.) The average (infinitesimal) RPA estimator (for one busy cycle) then becomes (for more details on its derivation, see Brémaud and Vázquez-Abad, 1991):

$$\psi^{A-RPA} = \frac{1}{\lambda} \sum_{i=2}^{\tau} (\tau - \tilde{\tau}^{(i)}). \tag{8}$$

Brémaud and Vázquez-Abad (1991) show that under reasonable conditions, that estimator is unbiased for $\ell'(\theta)$. They also show some links between this estimator and the LR method. Finally, a similar estimator has also been suggested by Gong (1988).

Of course, as for LR, one will use say n regenerative cycles and estimate the derivative by averaging out the n values of ψ^{A-RPA} associated with these cycles. Note that here, contrary to what we did for the finite-difference RPA in the previous section, the number of regenerative cycles that we consider is the same for both the nominal and phantom paths. Here, when a busy cycle splits up, we simply discard what happens in the phantom path until the start of the next busy cycle in the nominal path.

6. Two SPA Estimators

Let us return to the model formulation of Section 2, where θ can be either λ or μ. Observe that the (infinite-horizon) average number of customers per cycle period is the inverse of the

fraction of customers that are first in their busy cycles, i.e. whose waiting time is zero. That is

$$\ell(\theta) = E_\theta[\tau] = \frac{1}{P_\theta(W=0)} \tag{9}$$

where

$$P_\theta(W=0) \stackrel{\text{def}}{=} \lim_{k\to\infty} \frac{1}{k}\sum_{i=1}^{k} P_\theta(W_i=0)$$

represents the probability that a "random" customer in steady-state has zero waiting time. By differentiating (9), one obtains

$$\ell'(\theta) = \frac{\partial}{\partial\theta}E_\theta[\tau] = -\frac{1}{[P_\theta(W=0)]^2}\frac{\partial}{\partial\theta}P_\theta(W=0). \tag{10}$$

We can now estimate $\ell'(\theta)$ indirectly by estimating $P_\theta(W=0)$ and its derivative with respect to θ. To do so, we consider a simulation with a total number C of customers (when we simulate for a fixed number n of regenerative cycles, C is a random variable). Let $\mathcal{I}_i$ be the indicator function of the event $\{W_i = 0\}$. To estimate $P_\theta(W=0)$ we can take the sample average

$$\frac{1}{C}\sum_{i=1}^{C}\mathcal{I}_i \tag{11}$$

which converges to $P_\theta(W=0)$ *a.s.* and in expectation as $C \to \infty$ (or as $n \to \infty$), by the elementary renewal theorem (see, e.g., Wolff, 1989).

As we saw before, since $\mathcal{I}_i$ is generally discontinuous in θ for fixed underlying $U(0,1)$ random variates, straightforward IPA cannot be applied directly to (11). To smooth out the estimator, we can replace $\mathcal{I}_{i+1}$ by its conditional expectation given X_i. Following the notation in L'Ecuyer (1990) and L'Ecuyer and Perron (1990), for each $i \geq 1$, let ω_i represent the sequence of standard uniform $U(0,1)$ variates that have been used to generate $\{A_1,\ldots,A_{i-1},S_1,\ldots,S_i\}$. Then, X_i is a function of (θ,ω_i).

Now, let:

$$h_i(\theta,\omega_i) = P_\theta(W_{i+1}=0 \mid \omega_i) = p_\theta(A_i \geq X_i) = \bar{F}_\theta(X_i), \tag{12}$$

where $\bar{F}_\theta(X_i) \stackrel{\text{def}}{=} 1 - F_\theta(X_i)$. Then, we have $E_\theta(h_i(\theta,\omega_i)) = E_\theta\left(E_\theta(\mathcal{I}_{i+1} \mid \omega_i)\right) = P_\theta(W_{i+1} = 0)$, so that a second unbiased estimator for $P_\theta(W=0)$ is

$$\frac{1}{n}\sum_{i=1}^{n} h_i(\theta,\omega_i). \tag{13}$$

If $h_i(\theta,\omega_i)$ satisfies the assumptions of Theorem 1 in L'Ecuyer (1990), then an unbiased estimator of $\partial P_\theta(W_{i+1}=0)/\partial\theta$ is given by:

$$h_i'(\theta,\omega_i) = \frac{\partial}{\partial\theta}\bar{F}(X_i) = -\left(\frac{\partial}{\partial\theta}F_\theta\right)(X_i) - f_\theta(X_i)\frac{\partial}{\partial\theta}X_i, \tag{14}$$

where the derivatives are taken with respect to θ for ω_i *fixed*. In particular, when $\theta = \mu$ (and λ is fixed), the right-hand-side of (14) becomes $-f_\lambda(X_i)(\partial X_i/\partial\mu)$. From (10), (13), and (14), we obtain the following estimator for the derivative $\ell'(\theta)$:

$$\psi^{SPA} = -C\sum_{i=1}^{C} h_i'(\theta,\omega_i)\left(\sum_{i=1}^{C} h_i(\theta,\omega_i)\right)^{-2}. \tag{15}$$

Sufficient conditions for Theorem 1 in L'Ecuyer (1990) to apply in this case can be obtained in a similar way as for the examples examined in L'Ecuyer and Perron (1990).

As a second choice for smoothing, let ω_i represent the sequence of standard uniform $U(0,1)$ variates that have been used to generate $\{A_1,\ldots,A_i,S_1,\ldots,S_{i-1}\}$. Now, h_i is defined as

$$h_i(\theta,\omega_i) = P_\theta(W_{i+1} = 0 \mid \omega_i) = P_\theta(W_i + S_i \le A_i) = G_\theta(A_i - W_i). \tag{16}$$

Note that when $A_i - W_i \le 0$, this quantity is zero. Again, if this h_i satisfies the assumptions of Theorem 1 in L'Ecuyer (1990), an unbiased estimator of $\partial P_\theta(W_{i+1} = 0)/\partial\theta$ is given by:

$$h_i'(\theta,\omega_i) = \frac{\partial}{\partial\theta}G_\theta(A_i - W_i) = \left(\frac{\partial}{\partial\theta}G_\theta\right)(A_i - W_i) + g_\theta(A_i - W_i)\frac{\partial}{\partial\theta}(A_i - W_i) \tag{17}$$

for $A_i - W_i > 0$, and $h_i'(\theta,\omega_i) = 0$ otherwise. In particular, when $\theta = \lambda$ (and μ is fixed), the right-hand-side of (17) becomes $g_\mu(A_i - W_i)\partial(A_i - W_i)/\partial\lambda$. Using these new h_i and h_i' in (15) yields a second SPA estimator. Sufficient conditions for that second estimator to be unbiased can be obtained as in L'Ecuyer and Perron (1990).

7. Surrogate (or indirect) estimation

Suppose that λ and μ represent respectively the average arrival rate and the average service rate (whether or not the distributions are exponential). Then, if both λ and μ are multipled by the same constant c, the expected number of customers per busy cycle does not change, since this just corresponds to changing the time scale by the factor c. Therefore, in that case, $\ell(\theta)$ depends on θ only via $\rho = \lambda/\mu$. From that observation, we can express $\ell'(\lambda)$ as a function of $\ell'(\mu)$, and vice-versa:

$$\ell'(\lambda) = \frac{dE_\lambda[\tau]}{d\lambda} = \frac{dE_\mu[\tau]}{d\mu}\frac{d\mu}{d\rho}\frac{d\rho}{d\lambda} = -\frac{\mu}{\lambda}\ell'(\mu). \tag{18}$$

Using (18), any estimator of the derivative with respect to the average arrival rate λ can be transformed into an estimator of the derivative with respect to the average service rate μ, and vice-versa. Such indirect estimators were studied in Vázquez-Abad and Kushner (1990), where they were called *surrogate* estimators. For example, any of the two SPA estimators defined in the previous section can be used to estimate the derivative with respect to either λ or μ. Furthermore, each of these two estimators for λ [resp., for μ] can be transformed into an estimator for μ [resp., for λ] by (18). This gives four SPA estimators for the derivative with respect to λ and another four with respect to μ.

8. Numerical illustration with an $M/M/1$ queue

To illustrate the behavior of the above estimators, we take a simple $M/M/1$ queue with arrival rate λ and service rate μ. We have $F_\lambda(x) = 1-e^{-\lambda x}$, $f_\lambda(x) = \lambda e^{-\lambda x}$, $G_\mu(x) = 1-e^{-\mu x}$, and $g_\mu(x) = \mu e^{-\mu x}$. We want to estimate $\ell'(\lambda)$, the derivative with respect to λ. We simulate that queue for n regenerative cycles and compute each derivative estimator that we are interested in. Further, we do r replications of that. The exact value of the derivative $\ell'(\lambda)$ for the $M/M/1$ queue can be computed easily, which permits us to compare our estimations with the true values:

$$\begin{aligned} \ell(\theta) &= \frac{\mu}{\mu-\lambda}; \\ \ell'(\lambda) &\stackrel{\text{def}}{=} \frac{d}{d\lambda} E_\lambda[\tau] = \frac{\mu}{(\mu-\lambda)^2}; \end{aligned}$$

The standard LR estimator (4) with $\theta = \lambda$ becomes in this case

$$\psi_\lambda^{LR} = \tau \sum_{i=1}^{\tau} \frac{\partial}{\partial\lambda} \ln f_\lambda(A_i) = \left(\tau/\lambda - \sum_{i=1}^{\tau} A_i\right) \tau. \tag{19}$$

On the other hand, the standard LR estimator for the derivative with respect to μ is given by

$$\psi_\mu^{LR} = \tau \sum_{i=1}^{\tau} \frac{\partial}{\partial\mu} \ln g_\mu(S_i) = \left(\tau/\mu - \sum_{i=1}^{\tau} S_i\right) \tau. \tag{20}$$

An indirect (surrogate) estimator of the derivative with respect to λ is readily obtained by combining (18) and (20).

When we take the derivative with respect to $\theta = \lambda$, for our first SPA estimator, we obtain $h_i(\theta, \omega_i) = e^{-\lambda X_i}$ and

$$h_i'(\theta, \omega_i) = e^{-\lambda X_i} \left(-X_i - \lambda \frac{\partial X_i}{\partial \lambda}\right) = -e^{-\lambda X_i} \sum_{j \in \Phi_i} S_j,$$

where $\Phi_i \stackrel{\text{def}}{=} \{j \leq i \mid \text{customer } j \text{ is in the same busy cycle as customer } i\}$. The last equality above follows from standard IPA arguments, as in Suri (1989) or Suri and Zazanis (1988), for example. For $\theta = \lambda$, for the second SPA estimator, we have $h_i(\theta, \omega_i) = 1 - e^{-\mu \max(0, A_i - W_i)}$ and, since $W_i = \sum_{j \in \Phi_i \setminus \{i\}} (S_j - A_j)$ and $\partial A_j / \partial \lambda = -A_j/\lambda$,

$$h_i'(\theta, \omega_i) = \mu e^{-\mu(A_i - W_i)} \frac{\partial(A_i - W_i)}{\partial \lambda} = -\frac{\mu}{\lambda} e^{-\mu(A_i - W_i)} \sum_{j \in \Phi_i} A_j$$

when $A_i - W_i > 0$. When $A_i - W_i \leq 0$, one has $h_i'(\theta, \omega_i) = 0$. The corresponding SPA estimators (13) and (15) for $\ell(\theta)$ and $\ell'(\theta)$ can be computed easily in a single simulation run.

If we take the derivative with respect to μ, we obtain by analogous calculations

$$h_i'(\theta, \omega_i) = -\lambda e^{-\lambda X_i} \frac{\partial X_i}{\partial \mu} = \frac{\lambda}{\mu} e^{-\lambda X_i} \sum_{j \in \Phi_i} S_j$$

for the first SPA estimator and

$$h_i'(\theta, \omega_i) = e^{-\mu(A_i - W_i)} \sum_{j \in \Phi_i} A_j$$

for the second one. The corresponding derivative estimators given by (15) can be used, together with (18), to obtain surrogate estimators of the derivative with respect to λ. It turns out that for this particular $M/M/1$ example, these two surrogate SPA derivative estimators are exactly the same as the two direct SPA estimators developed in the previous paragraph. But this coincidence is not always true in general.

Table 1 gives the results of our numerical experiments. We took $r = 100$ replications and $n = 10000$ regenerative cycles per replication. In all cases, $\lambda = 1$, so that the traffic intensity is $\rho = 1/\mu$. We performed simulations with the following values of ρ: 1/4, 1/2, 2/3, and 3/4. For the finite differences, we took $\Delta\lambda = 0.02$. We have computed estimators using the following methods: direct LR based on equation (19) (D-LR), surrogate LR based on equation (20) (S-LR), finite-difference RPA (FD-RPA), average infinitesimal RPA (A-RPA), and the two SPA estimators of Section 6 (SPA1) and (SPA2). For each method and each value of ρ, we give the average (aver.) of the 100 derivative estimations that we have obtained, and the sample standard deviation (s.d.). Notice that for each column of the table, all estimators have been computed from the *same* simulations, that is with common random numbers.

ρ	1/4		1/2		2/3		3/4	
Exact derivative	0.4444		2.00		6.00		12.00	
	aver.	s.d.	aver.	s.d.	aver.	s.d.	aver.	s.d.
D-LR	0.444	0.035	1.987	0.168	6.051	0.582	12.163	1.446
S-LR	0.440	0.044	1.982	0.169	6.072	0.634	12.021	1.455
FD-RPA	0.437	0.078	1.976	0.249	5.843	0.531	11.416	0.893
A-RPA	0.445	0.014	1.990	0.085	6.015	0.366	12.071	0.796
SPA1	0.444	0.009	1.992	0.063	6.020	0.286	12.067	0.614
SPA2	0.446	0.012	1.994	0.074	6.011	0.324	12.073	0.683

Table 1: Simulation results for the $M/M/1$ example.

From the simulation results, one can see that our first and second SPA estimators, in that order, are those that perform the best. They are unbiased and have a lower standard deviation than the other ones. Then, comes the average RPA method. Finite-difference RPA

has not only more variance, but also significant bias for the value of $\Delta\lambda = 0.02$ that we have chosen. A smaller value of $\Delta\lambda$ will reduce the bias, but then the variance will be still higher. The LR methods have the highest variance.

9. Conclusion

We have shown through an example how efficient perturbation analysis estimators can be built in situations where straightforward IPA will not work. We have introduced two new (efficient) SPA estimators for estimating the derivative of the expected number of customers per regenerative cycle in a $GI/G/1$ queue. For the examples that we have examined, these new estimators outperformed those that had been proposed previously for the same problem. We took the $GI/G/1$ queue as an illustration, but our development can be generalized to different performance measures and more general systems.

References

Aleksandrov, V. M., Sysoyev, V. I., and Shemeneva, V. V. (1968), "Stochastic Optimization", *Engineering Cybernetics*, **5**, 11–16.

Asmussen, S. (1987), *Applied Probability and Queues*, Wiley.

Brémaud, P. and Vázquez-Abad, F. (1991), "On the Pathwise Computation of Derivatives with respect to the Rate of a Point Process: the Phantom RPA method", submitted to *QUESTA*.

Glasserman, P. and Gong, W.B. (1990), "Smoothed Perturbation Analysis for a Class of Discrete Event Systems", *IEEE Trans. on Automatic Control*, **AC-35**, 11, 1218-1230.

Glynn, P. W. (1990), "Likelihood Ratio Gradient Estimation for Stochastic Systems", *Communications of the ACM*, **33**, 10, 75–84.

Gong, W.-B. (1988), "Smoothed Perturbation Analysis Algorithm for a $GI/G/1$ Routing Problem", *Proceedings of the 1988 Winter Simulation Comference*, IEEE Press, 525–531.

Heidelberger, P., X.-R. Cao, M. A. Zazanis, and R. Suri. (1988), "Convergence Properties of Infinitesimal Perturbation Analysis Estimates", *Management Science*, **34**, 11, 1281–1302.

Ho, Y.-C. (1987), "Performance Evaluation and Perturbation Analysis of Discrete Event Dynamic Systems", *IEEE Transactions of Automatic Control*, **AC-32**, 7, 563–572.

Ho, Y. C. and Cao, X. (1985) "Performance Sensitivity to routing changes in Queueing Networks and Flexible Manufacturing Systems using Perturbation Analysis", *IEEE Jour. Robotics and Autom.*, **vol. RA-1**, No. 4, pp. 165-172.

Ho, Y. C. and Strickland, S. (1990), "A Taxonomy of Perturbation Analysis Techniques", manuscript, Harvard University.

L'Ecuyer, P. (1990), "A Unified Version of the IPA, SF, and LR Gradient Estimation Techniques", *Management Sciences*, **vol. 36**, No. 11, pp. 1364-1383.

L'Ecuyer, P. and Perron, G. (1990), "On the Convergence Rates of IPA and FDC Derivative Estimators for Finite-Horizon Stochastic Systems", submitted for publication.

Reiman, M. I. and Weiss, A. (1989) "Sensitivity Analysis for Simulations via Likelihood Ratios", *Op. Res.*, **vol. 37**, No. 5, pp. 830-844.

Rubinstein, R. Y. (1989), "Sensitivity Analysis and Performance Extrapolation for Computer Simulation Models", *Operations Research*, **37**, 1, 72–81.

Simon, B. (1989), "A New Estimator of Sensitivity Measures for Simulations Based on Light Traffic Theory", *ORSA Journal on Computing*, **1**, 3, 172–180.

Suri, R. and Cao, X. (1986), "The Phantom Customer and the Marked Customer Methods for Optimization of Closed Queueing Networks with Blocking and general Service Times", *ACM Performance Evaluation Review*, **vol. 12**, No. 3, pp. 243-256.

Suri, R. and Zazanis, M. (1988), "Perturbation Analysis gives strongly consistent Estimates for the $M/G/1$ Queue", *Management Sci.*, **vol 34**, No. 1, pp. 114-137.

Suri, R. (1989), "Perturbation Analysis: The State of the Art and Research Issues Explained via the $GI/G/1$ Queue", *Proceedings of the IEEE*, **77**, 114–137.

Vázquez-Abad, F. and Kushner, H. (1991), "A Surrogate Estimation Approach for Adaptive Routing in Communication Networks", submitted to *IEEE Trans. on Automatic Control.*

Wardi, Y., Gong, W.-B., Cassandras, C. G., and Kallmes, M. H. (1990), "A New Class of Perturbation Analysis Algorithms for Piecewise Continuous Sample Performance Functions", submitted for publication.

Wolff, R. (1989), *Stochastic Modeling and the Theory of Queues*, Prentice-Hall.