WEAK CONVERGENCE OF SAMPLE PATH DERIVATIVES FOR THE WAITING TIME IN A SINGLE SERVER QUEUE

MICHAEL A. ZAZANIS
Dept. of Industrial Engineering and Management Sciences
Northwestern University
Evanston, IL 60201

ABSTRACT

A simple theoretical framework is provided to address the problem of unbiasedness of infinitesimal perturbation analysis estimates in steady state and some simple sufficient conditions are presented. These are illustrated for the case of a GI/G/1 queue and improved conditions for unbiasedness of the estimators are given.

1. Introduction

Sensitivity analysis of queueing systems is a topic that has recently received attention from a number of authors (e.g. see [GS], [HC], [RW], [RZ]). The general idea is to provide estimates for the derivative of a performance criterion of a queueing system with respect to a parameter, from the information contained in a single sample path, without the use of finite differences.

In this paper we will examine more specifically the question of unbiasedness of derivative estimates given by the infinitesimal perturbation analysis (IPA) algorithm. (For background on IPA see e.g. [HC] and [SZ]). Even though this question has already been the subject of some papers (see [C], [SZ], and references therein) the issue of unbiasedness of the *steady state* estimates has by no means been exhausted. In §, a simple theoretical framework is given which enables us to provide additional sufficient conditions insuring that IPA estimates which are unbiased for sample paths of finite length will also be unbiased in steady state.

In the second part we verify these conditions in the case of a GI/G/1 queue. The simplicity of this system allows us to obtain an explicit expression for the derivative of the expected waiting time in steady state. In fact, in order to simplify the exposition in this summary, we will restrict ourselves to the estimation of derivatives with respect to a scale parameter of the service time distribution. The case of a general parameter, though more complicated, is not conceptually different. Based on these results, a simple estimator is obtained which is shown to be superior to classical finite difference estimators.

2. Unbiased IPA estimates in steady state

Let $\{F_i\}_{i=1,2,\ldots}$ be a filtration and let $W_i(\theta)$, $i=1, 2, \ldots$, $\theta \in [a, b]$, be a sequence of random functions on (Ω, F, P) adapted to $\{F_i\}_{i=1,2,\ldots}$. Let $f_i(\theta) = E[W_i(\theta)]$. Suppose also that

A1: for all θ $W_i(\theta)$ converges weakly to a r.v. $W(\theta)$,

A2: $\sup_i E \mid W_i(\theta) \mid^p \le M < \infty$ with $p > 1$.

Let $f(\theta) = E[W(\theta)]$. The above conditions then guarantee that $\lim_i f_i(\theta) = f(\theta)$. We will also assume that

A3: $f_i'(\theta)$, $i=1, 2, \ldots$, and $f'(\theta)$ exist for all $\theta \in [a, b]$.

The above process is an appropriate model for a large class of systems of practical interest. For instance, $W_i(\theta)$ might be the response time of the i^{th} customer arriving to a tandem network of queues whereas θ could be a parameter of the service time distribution of one of the servers in the network or a parameter of the arrival process. Assuming the system to be stable for $\theta \in [a, b]$, it is then well known that $W_i(\theta)$ converges weakly to a r.v. $W(\theta)$ representing the delay in steady state. We are interested in estimating the derivative of $E[W(\theta)]$ for sensitivity analysis or optimization purposes.

Returning to our model, let us further assume that

A4: $\dfrac{dW_i}{d\theta}$ exists w.p.1 and is F_i- measurable for all i.

(In the above example $\dfrac{dW_i}{d\theta}$ would be computed by the IPA algorithm and F_i-measurability would represent our ability to compute $\dfrac{dW_i}{d\theta}$ from the information available up to that point. For further details see [G] and [SZ]). We will also assume that

A5: $\dfrac{dW_i}{d\theta}$, $i=1, 2, \ldots$, converges weakly to some steady state r.v. $\dfrac{dW}{d\theta}$ for all $\theta \in [a, b]$.

We can now state the following

Theorem 1: Let $W_i(\theta)$, $i=1, 2, \ldots$ satisfy *A1-A5* and suppose that, addditionally, it satisfies

A6: $\dfrac{d}{d\theta} E[\dfrac{dW_i}{d\theta}] = \dfrac{d}{d\theta} E[W_i] \quad i=1, 2, \ldots$.

A7: $\dfrac{d}{d\theta} E[W_i] = f_i'(\theta)$ converges uniformly for all $\theta \in [a, b]$.

Then

$$\frac{d}{d\theta} E[W] = E[\frac{dW}{d\theta}] \; .$$

Proof: The proof follows immediately from a standard theorem on differentiation and uniform convergence (e.g. see [R], p.152).

Remarks: (i) In the above context, suppose that the IPA algorithm provides unbiased estimates for finite length sample paths. (This of course is assumption *A6)*. The above theorem suggests that additional conditions are required in order for IPA to also give unbiased estimates in *steady state*. A sufficient condition is *A7*.

(ii) Assumption *A7* in the above theorem can be replaced with the alternative Assumption

A7$'$: $f_i'(\theta)$, $i=1, 2, \ldots$, and $\lim_i f_i'(\theta)$ are continuous on $[a, b]$ and, for all i, $f_{i+1}' \geq f_i'$ on $[a, b]$.

Then from a theorem of Dini, ([R], p150), follows that $f_i'(\theta)$ converges uniformly for $\theta \in [a, b]$. Hence *A7$'$* implies *A7* and, for a number of systems, the former may be easier to verify than the latter, since a great deal is known about continuity and monotonicity properties of stochastic systems.

3. Sample path derivatives for the GI/G/1 queue

In this section we illustrate the above theorem for a GI/G/1 system. We will denote by C_i the i^{th} customer, by W_i the waiting time of C_i, by θX_i his service requirement, (depending on a scale parameter θ) and by A_i the interarrival time between C_i and C_{i+1}. Also, let $P(X_1 \leq x) = F(x)$ and $P(A_1 \leq x) = G(x)$. We will assume that $E[A_1] < \infty$, $E[X_1^3] < \infty$, and that the system is ergodic for all $\theta \in [a, b]$, which is equivalent to the requirement $bE[X_1] < E[A_1]$. Let us denote by W a random variable distributed according to the steady state distribution of the waiting time. Our goal is to obtain an estimator for $\frac{d}{d\theta} E[W]$ without the use of finite differences.

Let us assume that at time $t=0$ the first customer, C_1, arrives to an empty system. Also, let us designate by L_i the index of the customer who initiates the busy period in which C_i belongs. Evidently, $L_i \leq i$, the equality holding in the case where C_i initiates a busy period. For a First Come First Served system, the waiting time of C_i is given by the relationship:

$$W_i(\theta) = \max\Big(0,\ W_{i-1} + X_{i-1}\theta - A_{i-1} \Big), \quad i = 2, 3, \cdots \tag{3.1}$$

and

$$W_1(\theta) = 0 \ .$$

Under very mild assumptions it is easy to see, using for example the conditions in [Wh] that, for any given θ, $W_i(\theta)$ is differentiable with respect to θ w.p.1. (For instance, for a GI/G/1 system, the assumption that $G(\ )$, or $F(\ )$ is absolutely continuous is more than enough to guarantee this). Hence, differentiating $W_i(\theta)$ with respect to θ, we get

$$\frac{dW_i}{d\theta} = \begin{cases} 0 & \text{if } L_i = i \\ \dfrac{dW_{i-1}}{d\theta} + X_{i-1} & \text{if } L_i < i \end{cases} \tag{3.2}$$

Using (3.2) iteratively we obtain

$$\frac{dW_i}{d\theta} = \sum_{j=L_i}^{i-1} X_j \ , \tag{3.3}$$

with ill defined sums interpreted as being equal to zero.

In terms of the model of the process of the previous paragraph, we could define F_i to be σ–$\{A_1, \cdots, A_{i-1} ; X_1, \cdots X_{i-1}\}$, $i=1, 2,\ldots$, and F_1 to be the trivial σ–field. Then it is clear that both W_i and $\frac{dW_i}{d\theta}$ are F_i--measurable.

Consider now the discrete time renewal process defined by the indices of customers who initiate busy periods, i.e. $M_0 = 0$, and $M_k = \inf\{i : L_i > M_{k-1}\}$, for $k = 1, 2, \cdots$. The increment $N_k = M_k - M_{k-1}$ is of course equal to the number of customers served in the k^{th} busy period. (Our ergodicity assumption also guarantees that $E[N_k] < \infty$).

As it is well known, W_i, $i=1, 2,\ldots$, is a discrete time regenerative process with respect to the renewal process M_k and it is not hard to see from (3.3) that $\frac{dW_i}{d\theta}$ is also regenerative with respect to the same renewal process. It follows then from [CI] that, for all $\theta \in [a, b]$, the sequences W_i and $\frac{dW_i}{d\theta}$ converge weakly to the r.v.'s W and $\frac{dW}{d\theta}$, and furthermore that

$$E[\frac{dW}{d\theta}] = \frac{E[\sum_{i=1}^{N_1} \frac{dW_i}{d\theta}]}{E[N_1]} \ . \tag{3.4}$$

Combining (3.3) and (3.4), we get

$$E[\frac{dW}{d\theta}] = \frac{E[\sum_{i=1}^{N_1-1} \sum_{j=1}^{i} X_j]}{E[N_1]} \ . \tag{3.5}$$

Provided that one can establish that

$$E[\frac{dW}{d\theta}] = \frac{d}{d\theta} E[W] \ , \tag{3.6}$$

equation (3.5) suggests an estimator for $\frac{d}{d\theta}E[W]$ which as we show in §5 is superior to the classical estimators involving finite differences.

4. Sketch of the proof

In §3 we have argued that the sequence of waiting times W_i in a GI/G/1 queue satisfies conditions *A1*, *A4*, and *A5*. In this section we will briefly describe how to verify the remaining conditions, thus establishing (3.6). To check *A2*, choose $p = 2$ and notice that the Kiefer-Wolfovitz conditions (see [KW]) and our moment assumptions in the begining of §3 imply that $E[W^2] < \infty$. Since for all i, $E[W_i^2] < E[W^2]$, *A2* is satisfied.

For the rest of the assumptions, our task becomes much easier if, instead of the sequence of the waiting times W_i defined in (3.1), we consider the sequence

$$W_i^*(\theta) = \max\Big(0, X_1\theta - A_1, \cdots, (X_1+...+X_{i-1})\theta - A_1 - ... - A_{i-1} \Big), \quad i = 2, 3, ... \tag{4.1}$$

and

$$W_1^* = 0 .$$

It is easy to see that $W_i(\theta)$ and $W_i^*(\theta)$ have the same distribution for all i. Also, according to a standard result, $W^* = \lim_i W_i^*$ exists, is finite w.p.1, and is distributed according to the steady state distribution, provided that the system is ergodic. In particular,

$$E[W^*] = E[W] , \tag{4.2}$$

a result that we will use in the sequel. Next, let

$$L_i^*(\theta) = \min\{ j : 0 \le j \le i-1, (X_1+...+X_j)\theta - A_1 - ... - A_j = W_i^*(\theta) \} . \tag{4.3}$$

Thus $L_i^*(\theta)$ is the (smallest) value of the index that maximizes the expression in (4.1). In particular notice that, when $\theta_1 < \theta_2$, $L_i^*(\theta_1) \le L_i^*(\theta_2)$. Also, define

$$L^*(\theta) = \inf\{ j : (X_1+...+X_j)\theta - (A_1+...+A_j) = W^*(\theta) \} . \tag{4.4}$$

Under the conditions mentioned in the previous section, one can easily show that, for any given θ, $\dfrac{dW_i^*}{d\theta}$ exists w.p.1 and is equal to

$$\frac{dW_i^*}{d\theta} = \sum_{j=1}^{L_i^*} X_j . \tag{4.5}$$

The proof of (3.6) is based on the following lemmata whose proofs are straightforward and will be described briefly.

Lemma 1:

The r.v.'s $\dfrac{dW_i}{d\theta}$ and $\dfrac{dW_i^*}{d\theta}$ have the same distribution and in particular

$$E\,[\,\frac{dW_i}{d\,\theta}\,] \;=\; E\,[\,\frac{dW_i^*}{d\,\theta}\,]\,.$$

The proof is very simple and will be omitted.

Lemma 2: $E\,[\frac{dW_i^*}{d\,\theta}] \;=\; \frac{d}{d\,\theta}E\,[W_i^*]$ for all $\theta \in [a\,,b\,]$ and $i=1,\,2,\ldots$

Proof: Let $\delta > 0$ and consider

$$\frac{1}{\delta}\,[W_i^*(\theta{+}\delta) - W_i^*(\theta)] \;=\; \frac{1}{\delta}[\,\sum_{j=1}^{L_i^*(\theta+\delta)} (\theta{+}\delta)X_j - A_j \;-\; \sum_{j=1}^{L_i^*(\theta)} \theta X_j - A_j\;] \tag{4.6}$$

$$=\; \sum_{j=1}^{L_i^*(\theta+\delta)} X_j \;+\; \frac{1}{\delta}\,[\;\sum_{j=L_i^*(\theta)+1}^{L_i^*(\theta+\delta)} \theta X_j - A_j\;]\;.$$

But $\sum_{j=L_i^*(\theta)+1}^{L_i^*(\theta+\delta)} \theta X_j - A_j$ is either 0 (when $L_i^*(\theta) = L_i^*(\theta{+}\delta)$), or negative, (when $L_i^*(\theta) < L_i^*(\theta{+}\delta)$), because of definition (4.3). Hence,

$$0 \;\le\; \frac{1}{\delta}\,[\;W_i^*(\theta{+}\delta) - W_i^*(\theta)\;] \;\le\; \sum_{j=1}^{L_i^*(\theta+\delta)} X_j \;\le\; \sum_{j=1}^{L_i^*(b)} X_j \quad ,$$

and an appeal to the Dominated Convergence Theorem complets the proof.

We can now state

Lemma 3: $E\,[\frac{dW_{i+1}^*}{d\,\theta}] \ge E\,[\frac{dW_i^*}{d\,\theta}]$, $i=1,\,2,\ldots$, and $\frac{d}{d\,\theta}E\,[W_i^*]$ and $\lim_i \frac{d}{d\,\theta}E\,[W_i^*]$ are continuous on $[a\,,b\,]$.

Proof: Monotonicity with respect to i is trivial. To establish the continuity of $\frac{d}{d\,\theta}E\,[W_i^*] = E\,[\frac{dW_i^*}{d\,\theta}] \;=\; E\,[\sum_{j=1}^{L_i^*(\theta)} X_j\,]$, it is sufficient to consider $E\,[\;\sum_{j=1}^{L_i^*(\theta+\delta)} X_j \;-\; \sum_{j=1}^{L_i^*(\theta)} X_j\;]$, $\delta{>}0$, and to appeal to the Monotone Convergence Theorem. The case $\delta{<}0$ is similar.

Finally,

$$\lim_i \frac{d}{d\,\theta}E\,[W_i^*] \;=\; \lim_i E\,[\frac{dW_i^*}{d\,\theta}] \;=\; E\,[\;\lim_i \frac{dW_i^*}{d\,\theta}]\;,$$

the last equality following from the Monotone Convergence Theorem. Hence it is sufficient to show that $E\,[\sum_{j=1}^{L^*(\theta)} X_j\,]$ is continuous with respect to θ, and the proof for that is similar to the one establishing the continuity of $\frac{d}{d\,\theta}E\,[W_i^*]$.

From the above three lemmata then, and (4.2), the remaining conditions *A3*, *A6*, and $A7'$, follow, and an appeal to Theorem 1 establishes (3.6). We can thus state

Theorem 2: For the class of GI/G/1 systems discussed in this section, IPA gives unbiased estimates of the derivative of the expected waiting time in steady state.

5. Statistical Aspects

If a sample path consisting of M busy periods is available, the following ratio estimator for $\frac{d}{d\,\theta} E\,[W\,]$ is suggested by (3.5)

$$\hat{W}(M\,) = \frac{\sum_{k=1}^{M} \sum_{i=M_{k-1}+1}^{M_k-1} \sum_{j=M_{k-1}+1}^{i} X_j}{\sum_{k=1}^{M} N_k} \,. \tag{5.1}$$

From the Strong Law of Large Numbers follows that $\lim_{M} \hat{W}(M\,) = \frac{E\,[\,\sum_{i=1}^{N_1-1} \sum_{j=1}^{i} X_j\,]}{E\,[N_1]}$ with probability 1, and hence from (3.5) that $\hat{W}(M\,)$ is a strongly consistent estimator of $\frac{d}{d\,\theta} E\,[W\,]$. It is also straightforward to show that the Mean Square Error, $E\,[\,\hat{W}(M\,) - \frac{d}{d\,\theta} E\,[W\,]\,]^2$, of this estimator as a function of the number of cycles observed is asymptotically $O\,(\frac{1}{M})$. (The details of this derivation are complicated however by the presence of ratio estimator bias). On the other hand, classical estimates requiring finite differences can be shown to have Mean Square Errors that are at best $O\,(\frac{1}{M^{2/3}})$ (see [FG], [ZS]). This, together with the simplicity of the estimator given in (5.1) illustrates the effectiveness of the approach presented here, whenever it is applicable.

REFERENCES

[C] Cao, X.R. (1985). "On the Sample Performance Functions of Jackson Queueing Networks", Technical Report, Division of Applied Sciences, Harvard Univ., Cambridge, MA.

[CI] Crane, M.A. and D.L. Iglehart (1975). "Simulating Stable Stochastic Systems: III. Regenerative Processes and Discrete-Event Simulations", *Operations Research, Vol. 23,* 1, 33-45.

[FG] Fox, B.L. and P.W. Glynn (1987). "Replication Schemes for Limiting Expectations", Working Paper, Dept. of Industrial Engineering, University of Wisconsin-Madison.

[G] Glynn, P.W. (1987). "Construction of Process-Differentiable Representations for Parametric Families of Distributions", Technical Report, Mathematics Research Center, Univ. of Wisconsin-Madison.

[GS] Glynn, P.W. and J.L. Sanders, (1986). "Monte Carlo optimization of stochastic systems: Two new approaches", *Proceedings of the 1986 ASME Computers in Engineering Conference.*

[HC] Ho, Y.C., and X. Cao (1983). "Perturbation Analysis and Optimization of Queueing Networks", *J. Optim. Theory Applic. 40,* 4, 559-582.

[KW] Kiefer, J. and J. Wolfowitz (1956). "On the Characterizations of the General Queueing Process with Applications to Random Walks", *Ann. Math. Stat. 27,* 147-161.

[RW] Reiman, M.I. and A. Weiss (1986). "Sensitivity Analysis of Simulations via Likelihood Ratios", *Proceedings of the 1986 Winter Simulation Conference,* J. Wilson, J. Henriksen, S. Roberts (eds.), IEEE Press, 285-289.

[SZ] Suri, R. and Zazanis, M.A. (1986). "Perturbation Analysis gives strongly consistent sensitivity estimates for the M/G/1 queue", to appear in *Management Science*

[Wh] Whitt, W. (1974). "The Continuity of Queues", *Adv. Appl. Prob. 6,* 175-183.

[ZS] Zazanis, M.A. and R. Suri (1986). "Comparison of Perturbation Analysis with Conventional Sensitivity Estimates for Stochastic Systems", Working paper # 86-123, Dept. of Industrial Engineering, University of Wisconsin-Madison.