

An Infinitesimal Perturbation Analysis Algorithm for a Multiclass G/G/1 Queue*

(submitted to the *OR Letters*)

Yu-Chi Ho and Jian-Qiang Hu
Division of Applied Science
Harvard University
Cambridge, MA 02138

(November 1988)

Abstract It has been shown that infinitesimal perturbation analysis (IPA) provides very efficient algorithms for estimating derivatives in a broad class of discrete event dynamic systems (DEDS). But in its simple form, it fails in most multiclass DEDS. In this paper, we propose a revised IPA for a multiclass G/G/1 queue.

Keywords: simulation*Markov processes*networks of queues*probability distributions

* The work in this paper was partly supported by the NSF under grants Nos. ECS85-15449 and CDR-8803012, under ONR Contracts Nos. N00014-86-K-0075 and N00014-84-K-0465 and under Army Contract No. DAAL-03-83-K-0171.

1. Introduction

With development of recent technologies, there has been increased interest in the study of Discrete Event Dynamic Systems (DEDS). DEDS can often be modelled as queueing networks. But, in most cases, seeking close form solutions is impossible due to their complexity. As an alternative approach to describe DEDS, simulation is widely used. It can model real systems more closely, but is often time-consuming and expensive.

One of the main issues in studying DEDS is performance sensitivity analysis, which is often used in modeling, analyzing, optimizing and controlling systems. Without any close form formula, the conventional way is to run two simulations under normal and perturbed parameter values for the system and form the finite difference ratio to give sensitivity estimates. We shall call this the brute force (BF) method. For a large system with many parameters, in order to get the n -dimensional gradient vector, $n+1$ simulations have to be run; so a large amount of computation is involved in the BF method. The BF method also suffers from the twin numerical evils of noise (if perturbations are too small) and nonlinearity (if perturbations are too large). Furthermore, BF method is invasive; and hence for a real system, sometimes it may be impossible to change the parameters to repeat experiments.

Perturbation Analysis (PA) and other single run gradient estimation techniques, on the other hand, takes the viewpoint that a single sample path or experiment on a DEDS contains inherently much more information about the system than conventional simulations utilize in their output analysis. Judicious and timely processing of single run data can yield much more information, including gradients [e.g. see Ho, 1987]. The basic idea of PA is to "reconstruct" a perturbed sample path from the nominal one with slight changes in parameters. To achieve this goal, several algorithms have been proposed under different assumptions (see a review of these algorithms in Ho [1987]). Among these algorithms, IPA is the most efficient one in obtaining sensitivity estimates. In addition to saving a lot of computation, another remarkable advantage of IPA is that it estimates performance gradients directly, not through finite differences; thus leading to superior variance properties. One simplifying assumption used in constructing the new perturbed path by IPA is that of deterministic similarity, which assumes that the event order of the nominal and the perturbed path is the same. Under this assumption, perturbation propagations can be computed easily through IPA. It has been proved that in some cases the event order changes can be ignored, so that IPA rule gives strongly consistent estimates of performance gradients [e.g. see Suri & Zazanis, 1987]. But there are cases

where IPA cannot routinely give strongly consistent estimates due to discontinuous sample performance functions resulting from parameter changes. The sufficient conditions that IPA gives strongly consistent estimates are studied by Suri [1987], Cao [1985] and Glasserman [1988a, 1988b]. The typical counterexample is multiclass networks. The reason that IPA does not work for multiclass networks can be briefly explained in what follows.

In general, only infinitesimal perturbations in the perturbed sample can be calculated by IPA rules. But IPA fails to account for finite changes. In multiclass networks, finite changes, which are caused by event order changes, play a very important role in performance sensitivity analysis. So applying IPA directly to multiclass networks gives us biased sensitivity estimates. Other algorithms, such as EPA and SPA, have been developed to overcome this difficulty with some additional cost of computation. Our purpose here is to develop a revised IPA retaining all its computational advantage but applicable in the multi-class setting. As we can see, the key point to make IPA applicable to multiclass networks is how we deal with those finite changes in the perturbed sample path. In this paper we present a revised IPA algorithm to extend IPA to multiclass G/G/1 queue. The remainder of this paper is organized as follows. In section 2, we briefly discuss IPA in a single class G/G/1 queue and show the difficulty when customers are from more than one classes. In section 3, the basic idea of revised IPA is described. An algorithm is also proposed for estimating the derivative of system time in a multiclass G/G/1 queue. In section 4, We prove that our revised IPA algorithm gives strongly consistent estimates when arrival processes are Poisson, and provide some simulation results for the G/G/1 case as well. Finally, section 5 gives summary of our new idea.

2. Infinitesimal Perturbation Analysis

The goal of PA is to answer the following "what if" question: what would happen if some parameters in DEDS were changed. There are two main aspects involved in all PA algorithms: perturbation generation and perturbation propagation. Given a sample path of a DEDS, perturbations in the timing of events corresponding to the changes of parameters in the system are first fictitiously introduced to the system, then propagated along the sample path. IPA has a most efficient way to achieve this computationally by assuming that event orders in both nominal and perturbed path are the same.

Consider a single class G/G/1 queue. We want to compute the first derivative of system time (waiting time plus service time) with respect to interarrival time parameter θ . To simplify notation, suppose that interarrival times are independently identically distributed (i.i.d) random variables with probability distribution function $F(x,\theta)$. The interarrival time in simulation then can be usually generated by $F^{-1}(u,\theta)$, where u is uniform random variable between zero and one. Given a sample path of the queue with parameter θ , which we call the nominal path, we are interested in what would happen if θ had been $\theta+\Delta\theta$, particularly when $\Delta\theta$ becomes infinitesimal small. First, it is easy to see that x_i , the interarrival time between customer i , C_i , and customer $i+1$, C_{i+1} , in the sample path would be perturbed by the amount of $\Delta x_i=(\partial x_i/\partial\theta)\Delta\theta + O(\Delta\theta^2)$ [Suri 1987]. For the rest of this paper, we always assume that $\partial x_i/\partial\theta$ exists. Now we look at the start of a busy period initiated by, say, C_{k+1} . Under the assumption that event order remains the same when perturbations are introduced into the nominal path, two busy period will not coalesce or one busy period will not split into two (the deterministic similarity assumption). So, in the perturbed path, C_{k+1} would also find system idle and the n th customer C_{k+n} in this period would still remain in the same position. Then total time delayed for C_{k+n} to enter the system is

$$\Delta T_{n+k} = \sum_{i=1}^{n+k} \frac{\partial x_i}{\partial \theta} \Delta \theta \quad (1)$$

Since service times for all customers are unchanged and the arrival time of the first customer C_{k+1} in the busy period is delayed by ΔT_{k+1} , the departure time of every customer in this busy period is changed by ΔT_{k+1} . We can think the busy period is shifted by a amount of ΔT_{k+1} . Thus, the total change in system time of C_{k+n} is

$$\Delta T_{k+1} - \Delta T_{n+k} = - \sum_{i=k+2}^{n+k} \frac{\partial x_i}{\partial \theta} \Delta \theta \quad (2)$$

That is to say, perturbations of interarrival times will be continuously accumulated during a busy period, and even propagated to the next busy period. That this is not at variance with our assumption of no coalescing and splitting of busy periods can be explained as follows. If we take a close look at the nominal sample path, we will find that event order changes only cause the Δ order change in system time, and the probability that event order changes is also the order of Δ . Therefore the total effect in system time caused by event order changes is of higher order and can be ignored when first derivatives are concerned. We show this process in

figure.1. Therefore, the IPA estimate of the first derivative of mean system time summed over all customers and busy periods is given by

$$\frac{\partial T}{\partial \theta} = -\frac{1}{N} \sum_{m=1}^M \sum_{n=k_m+1}^{k_{m+1}} \sum_{i=k_m+2}^n \frac{\partial x_i}{\partial \theta} \quad (3)$$

where N is total number of customers served and M is the number of busy periods observed. The same argument can also be used to obtain estimate for the first derivative of mean system time with respect to service time parameter. If φ is a parameter in service time, then IPA estimate for $\partial T/\partial \varphi$ is given by

$$\frac{\partial T}{\partial \varphi} = -\frac{1}{N} \sum_{m=1}^M \sum_{n=k_m+1}^{k_{m+1}} \sum_{i=k_m+1}^n \frac{\partial s_i}{\partial \varphi} \quad (4)$$

where s_i is the service time for customer C_i . In this case, perturbations of service times are accumulated during a busy period, but disappear when next busy period begins under the deterministic similarity assumption. It has been shown both theoretically and experimentally that IPA gives very good estimates for the $G/G/1$ queue. The main results can be found in Zazanis [1986].

But there is a difficulty in applying IPA directly to multiclass networks. Still consider our $G/G/1$ example. Suppose we now have two arrival streams which corresponding to two classes customers, and different class customers have different service time at the server. If we change, say, the mean interarrival time of class 1 customers, by Δ , then the arrival order of two classes customers will be changed sooner or later in the perturbed path. Suppose that this change happens in the busy period starting with customer $C_{k+1,1}$. (see figure.2 where the second subscript indicate the class of the customer) Let $C_{k+n,i}$ denote the n th customer in this busy period and $s_{k+n,i}$ be its service time. Let us assume that $C_{k+n,i}$ is class 1 and $C_{k+n+1,i}$ is class 2 and they are going to change the arrival order in the perturbed path. Then changes in system time for $C_{k+n,1}$ and $C_{k+n+1,2}$ are $s_{k+n+1,2} + \Delta_{k+n,1}$ and $-s_{k+n,1} + \Delta_{k+n+1,2}$, respectively. While $\Delta_{k+n,1}$ and $\Delta_{k+n+1,1}$ are both the order of Δ , but the total change changes in the system time are now finite. The finite terms in these changes cannot be cancelled out when being added together, because $s_{k+n,1} \neq s_{k+n+1,2}$. Therefore, the total average effect in system time due to arrival order change is now of the order of Δ and not of Δ^2 and thus not ignorable. Naively applying IPA will give us biased estimate.

3. Revised IPA algorithm

To overcome the above difficulty, we shall propose a revised version of IPA algorithm. There are two basic ideas contained in the method. One is viewing a multiclass customer queue as a single class customer queue. Another is converting finite changes in system time to infinitesimally small changes. Under this set-up, new perturbation generation law is obtained. We then apply it in conjunction with the original IPA propagation rule to the original queue to solve the problem .

To simplify discussions, suppose the multiclass G/G/1 queue we consider here only has two classes customers. Let the arrival rate or the reciprocal of the mean interarrival time of class i customer be λ_i , and service time of class i customer has mean \bar{s}_i and second moment σ_i^2 ($i=1, 2$). Let $A_i(x)$ be the p.d.f. (probability distribution function) of interarrival time and $G_i(s)$ be the p.d.f. of service time. Denote T the mean system time of a customer (averaged over both classes of customers). We wish to estimate $\partial T/\partial \lambda_1$.

First, instead of thinking there are two arrival streams coming to the queue, we superpose them together as one arrival process. The superposed process can be described by an ordered pair of random variable (Z, V) , where Z is "one" or "two" depending on the arrival which produced the customer. The random variable V is a non-negative real number: the time elapsed since the last customer of the component arrival stream which did not produce the customer in the superposed process. We illustrate the notation in figure.3. W_n is defined as the random variable representing the time between the n th and $(n+1)$ th customers in the superposed arrival process. We let $W_0=0$. The following results about the properties of the superposed process can be found in Cherry[1972]. Here we assume that all conditions needed for these results are satisfied , since they are general enough for our purpose.

Property 1. Stochastic process $\{Z_n, V_n, W_n; n \geq 0\}$ is a Markov renewal process.

Property 2. The underlying Markov chain has a unique stationary distribution given by:

$$\text{Prob}[Z_n=1, V_n \in [v, v+dv)] = \frac{\lambda_1 \lambda_2}{\lambda_1 + \lambda_2} \bar{A}_2(v) dv ,$$

$$\text{and Prob}[Z_n=2, V_n \in [v, v+dv)] = \frac{\lambda_1 \lambda_2}{\lambda_1 + \lambda_2} \bar{A}_1(v) dv ,$$

$$\text{where } \bar{A}_i(x) = 1 - A_i(x). \quad (5)$$

Property 3. In steady state, we have

$$\text{Prob}[Z=i] = \frac{\lambda_i}{\lambda_1 + \lambda_2}, \quad i = 1, 2. \quad (6)$$

Property 4. In steady state, the marginal distribution for W_n is

$$\begin{aligned} W_n(x) &= 1 - \text{Prob}[W_n \geq x] \\ &= 1 - \frac{\lambda_1 \lambda_2}{\lambda_1 + \lambda_2} \left(\int_0^\infty \bar{A}_1(x) \bar{A}_2(x+v) dv + \int_0^\infty \bar{A}_2(x) \bar{A}_1(x+v) dv \right). \end{aligned} \quad (7)$$

Since it is independent of n , we denote it by $W(x)$.

Property 3 and 4 can be easily derived from property 2. From property 3, we know that a customer in the superposed arrival process belongs to class 1 with probability d and class 2 with probability $1-d$, where $d = \lambda_1/(\lambda_1 + \lambda_2)$. In this sense, we can think the original multiclass queue is equivalent to a single class $G/G/1$ queue, in which all the customers are from one arrival process (superposed process) and the service time of a customer is given by the following rule: it is equal to s_1 with probability d and s_2 with probability $1-d$, where s_i is the random variable with p.d.f $G_i(s)$. In the "equivalent" queue, we notice that λ_1 is not only a parameter in arrival process, but also a parameter in service time through parameter d . Let T be the mean system time in the equivalent queue, it is obvious that $T = \bar{T}$. Then from the chain rule we have

$$\frac{dT}{d\lambda_1} = \frac{\partial T}{\partial \lambda_1} + \frac{\partial T}{\partial d} \frac{\partial d}{\partial \lambda_1}. \quad (8)$$

We can evaluate $\partial d/\partial \lambda_1$ from the definition of d . We also know that $\partial T/\partial \lambda_1$ can be obtained using the single class IPA algorithm. Because all customers are now from one arrival process, their arrival order will remain unchanged. In (3), replacing $\partial x_i/\partial \theta$ by $\partial W_i/\partial \lambda_1$ we then get the IPA estimate for $\partial T/\partial \lambda_1$. The only problem left is to estimate $\partial T/\partial d$.

If u and v are two independent uniform random variables over $[0,1]$, then by the way we establish the equivalent queue, we can generate the service time as:

$$s = \begin{cases} s_1 = G_1^{-1}(v) & \text{if } 0 \leq u \leq d \\ s_2 = G_2^{-1}(v) & \text{if } d < u \leq 1 \end{cases} \quad (9)$$

Now we change d , say, from d to $d + \Delta d$ ($\Delta d > 0$). By (9), we see that service time for some

customers may change from s_2 to s_1 . Actually this happens when $d < u \leq d + \Delta d$. We then say that these customers are to be switched. Obviously, switched customers will carry finite changes in service time. Hence finite changes in system time will be generated by these switched customers as well. Since IPA can only deal with infinitesimally small changes, it cannot be used to estimate $\partial T / \partial d$ directly. One way to get over this problem is to convert finite changes to infinitesimal small changes. The basic idea was first introduced by Ho and Cao [1985] to estimate the derivative of throughput with respect to routing probability in a Jackson network to smooth the sample path. The method we are going to use was originally used by Ho and Vakili [1987] to solve the routing problem in a parallel queue network.

Notice the following fact: under the condition that $0 \leq u \leq d$ (resp. $d < u \leq 1$), u/d (resp. $(1-u)/(1-d)$) is also a uniform random variable over $[0,1]$. So, instead of using (9) to determine service time we can use

$$s = \begin{cases} G_1^{-1}(1 - \frac{u}{d}) & 0 \leq u \leq d \\ G_2^{-1}(\frac{u-d}{1-d}) & d < u \leq 1 \end{cases} \quad (10)$$

It is a simple observation that if we generate service time via equation (10), the service time of every customer changes when d is changed. That is to say we are distributing the finite changes in service time of switched customers to infinitesimal changes in service time of all customers.

Now what we are concerned is whether the switched customers can be ignored or not after we carried out (10). Suppose that lower and upper bounds of service time s_i are l_i and u_i respectively ($l_i \geq 0$ and u_i may be $+\infty$), so $G_i(l_i) = 0$ and $G_i(u_i) = 1$. Assume that $G_i(s)$ has density function $g_i(s)$, and $g_i(s) > 0$ for $l_i \leq s \leq u_i$. From equation (10), we can see that when a customer is switched, i.e. $d < u \leq d + \Delta d$, the change in service time is $G_1^{-1}(0) - G_2^{-1}(0) + O(\Delta d) =$

$l_1 - l_2 + O(\Delta d)$. In order to make switched customers ignorable, we need

(A1) $l_1 = l_2 = 0$, or more general, $l_1 = l_2$.

Under the assumption A1, we have the order of Δd change in service time for switched customers. On the other hand, the probability that customers in the nominal path are to be switched is also the order of Δd . Therefore, total changes caused by switched customers is of order $(\Delta d)^2$ and can be ignored in calculating the first derivative.

After we rule out the possibility that some customers may be switched in the nominal path, then we can get the perturbation generation rule for service time by differentiating the both sides of equation (10):

$$\frac{\partial s}{\partial d} = \begin{cases} \frac{1 - G_1(s)}{d g_1(s)} & 0 \leq u \leq d \\ -\frac{1 - G_2(s)}{(1-d) g_2(s)} & d < u \leq 1 \end{cases} \quad (11)$$

In other word,

$$\frac{\partial s}{\partial d} = \begin{cases} \frac{1 - G_1(s)}{d g_1(s)} & \text{if customer is class 1} \\ -\frac{1 - G_2(s)}{(1-d) g_2(s)} & \text{if customer is class 2} \end{cases} \quad (12)$$

It means that in (12) we can assume the arrival order of all customers both in nominal and perturbed path are the same. Thus the perturbation propagation rule is still valid in constructing the perturbed path via the original one.

It now becomes clear that the reason we set up an equivalent queue is because we want to get better generation rule so that event order changes can be ignored under the new generation rule. There is a distinct difference between the old and the new generation rules. The former one only perturbs arrival time for class 1 customers. But the later one perturbs both arrival time and service time for all customers. After getting the new perturbation rule, we apply it the original multiclass G/G/1 queue, and the way to propagate these perturbations is still the same. So, substituting (12) to equation (4), we have IPA estimate for $\partial T / \partial d$.

Finally we should point out that the assumption A1 is not a critical condition for our method. It can be relaxed to general service time distributions by doing some modification on the generation rule (11) or (12). Since we want to address on the basic idea of our methodology, we would like not to make our algorithm too complicated by considering the general case.

4. Analysis and simulation results

We proposed a revised IPA algorithm to estimate the derivatives with respect to the arrival rates in a multiclass G/G/1 queue. In this section, the algorithm is evaluated from both

theoretical and experimental aspects. From now on, we call the algorithm we developed Revised IPA. First we have

Theorem Suppose assumption A1 is satisfied. If all arrival processes are Poisson, then the Revised IPA gives strongly consistent estimate of $dT/d\lambda_1$.

Outline of proof: Since all arrival processes are Poisson, we simply have a multiclass M/G/1 queue. Then we know that the equivalent queue is a single class M/G/1 queue. It has been shown in Suri and Zazanis [1988] that IPA gives strongly consistent derivative estimate of mean system time with respect to both arrival rate and mean service time in a M/G/1 queue. Therefore the IPA estimates for $\partial T/\partial\lambda_1$ and $\partial T/\partial d$ are strongly consistent. Substituting the results to (8), the theorem can be finally verified.

So far we cannot offer any theoretical proof to show that the Revised IPA gives strongly consistent estimate for general multiclass G/G/1 queue. This is not only because that we do not have any close form formula for mean system time, but because it still remains unproven that IPA provides strongly consistent estimate for single class G/G/1, though extensive experiment results suggest that it is true. In what follows, some numerical results are provided to test our algorithm. In all these experiments, we still focus on estimating $dT/d\lambda_1$. Since there do not exist close formulas to compute $dT/d\lambda_1$ for general multiclass G/G/1 queue, the brute force (BF) estimates are employed to compare with our Revised IPA estimates. We use symmetric difference estimates in the brute force method. The formula is given by

$$\left(\frac{\partial T}{\partial \lambda_1}\right)_{\text{BF}} = \frac{T(\lambda_1 + \Delta\lambda/2) - T(\lambda_1 - \Delta\lambda/2)}{\Delta\lambda} \quad (13)$$

The length of every sample path is total 1,000,000 customers, and we run 100 replications for every experiment to get the confidence interval at a 95% level by using normal distribution approximation. The results are present for three traffic intensities ρ ($\rho = \lambda_1\bar{s}_1 + \lambda_2\bar{s}_2$), namely for $\rho = 0.2, 0.5,$ and 0.8 . It is interesting to notice that IPA estimates have tighter confidence intervals than BF estimates. This is the variance property mentioned earlier which is retained in our Revised IPA.

Experiment.1 $A_i(x)$ is uniform distribution over $[0, 2/\lambda_i]$; $G_i(s)$ is exponential distribution with mean \bar{s}_i . ($i=1, 2$) We choose $\Delta\lambda=0.1$ for $\lambda_1=2$, $\Delta\lambda=0.05$ for $\lambda_1=0.6$ and 0.2 . The results are given in table.1

Experiment.2 $A_1(x)$ is exponential distribution with λ_1 ; $A_2(x)$ is uniform distribution over $[0, 2/\lambda_2]$; $G_1(s)$ is uniform distribution over $[l_1, 2\bar{s}_1 - l_1]$; $G_2(s) = \alpha(s - l_2)^3$, $s \in [l_2, 1+l_2]$, where α is a constant. We choose $\Delta\lambda=0.1$ for $\lambda_1=1$, $\Delta\lambda=0.05$ for $\lambda_1=0.5$ and 0.3 . The results are given in table.2.

5. Conclusion

IPA provides very efficient algorithms in estimating derivatives of performances in DEDS. The basic assumption used in IPA is that the event orders are the same in both nominal and perturbed sample path. This is equivalent to assume that the event order changes due to parameter perturbations only produce infinitesimal small changes, not finite changes in performance measure. The probability that the event order changes happen is usually the same order of parameter change. Therefore, given the assumption, the total effect caused by parameter change is then the higher order and can be ignored. But this assumption is usually violated in most multiclass networks. The goal of Revised IPA is to smooth the discontinuity (finite changes) of the sample path such that IPA can be applied.

General speaking, there are usually more than one representations to simulate a queue network. IPA algorithms are certainly related to the way we represent the systems, because for different representations different rules will be derived to generate the perturbations in the nominal sample path. The condition required in IPA may or may not be satisfied under these generation rules. In Glasserman [1988c], it has been shown that IPA works for one representation of a birth and death process, but fails for the other representation. The study in this paper gives another example to illustrate this fact. By finding an equivalent single class G/G/1 queue representation for the original multiclass G/G/1 queue and applying IPA to the new representation, we made it to yield consistent estimates of the first derivative of mean system time including both class 1 and 2 customers in a G/G/1 queue. As yet, we cannot compute the first derivative of mean system time for individual class of customer. This problem is still under investigation.

REFERENCES

- Cao, X. R. (1985), "Convergence of Parameter Sensitivity Estimates in a Stochastic Experiment," *IEEE Trans. Automatic Control*, Vol. AC-30, pp. 845-853.
- Cherry, W.P. (1972), *The Superposition of Two Independent Morkov Renewal Processes*, Ph.D. thesis, the University of Michigan.
- Glasserman, P. (1988a), "Structural Conditions for Perturbation Analysis Derivative Estimation, I: Finite Time Performance," submitted to publication.
- Glasserman, P. (1988b), "Structural Conditions for Perturbation Analysis Derivative Estimation, II: Networks of Queues," submitted to publication.
- Glasserman, P. (1988c), "Infinitesimal Perturbation Analysis of a Birth and Death Process," *Operation Research Letters*, 7, pp.43-49, 1988.
- Ho, Y. C. (1987), "Performance Evaluation and Perturbation Analysis of Discrete Systems: Perspective and Open Problems," *IEEE Trans. Automatic Control*, Vol. AC-32, pp. 563-572.
- Ho, Y.C. and Cao, X.R. (1985), "Perturbation Sensitivity to Routing Changes in Queueing Networks and Flexible Manufacturing Systems Using Perturbation Analysis," *IEEE J. of Robotics and Automation*, Vol. RA-1, No. 4, pp. 165-172.
- Suri, R. (1987), "Infinitesimal Perturbation Analysis for General Discrete Event Systems," *J. of ACM*, pp. 686-717.
- Suri, R. and Zazanis, M.A. (1987), "Infinitesimal Perturbation Analysis and the Regenerative Structure of GI/G/1 Queue," *Proc. of the 26th Conf. on Dec. and Contr.*, Los Angeles, CA.
- Suri, R. and Zazanis, M. A. (1988), "Perturbation Analysis Gives Strongly Consistent Estimates for the M/G/1 Queue," *Management Science*, Vol. 34, pp. 39-64.
- Ho, Y.C. and Vakili, P. (1987), "Infinitesimal perturbation analysis algorithm for a routing problem," *Allerton Control Conference*.
- Zazanis, M. A. (1986), *Statistical Properties of Perturbation Analysis Estimates for Discrete Event Systems*, Ph.D. thesis, Harvard University.

| λ_1 | λ_2 | \bar{s}_1 | \bar{s}_2 | ρ | BF est. | Revised IPA est. |
|-------------|-------------|-------------|-------------|--------|---------------------|---------------------|
| 2 | 2.5 | 0.3 | 0.08 | 0.8 | 1.3030 ± 0.1189 | 1.2525 ± 0.0258 |
| 0.6 | 4 | 0.5 | 0.05 | 0.5 | 0.6563 ± 0.1175 | 0.6423 ± 0.0112 |
| 0.2 | 0.5 | 0.6 | 0.16 | 0.2 | 0.8124 ± 0.0249 | 0.8840 ± 0.0053 |

Table.1

| λ_1 | λ_2 | \bar{s}_1 | \bar{s}_2 | l_1, l_2 | ρ | BF est. | Revised IPA est. |
|-------------|-------------|-------------|-------------|------------|--------|---------------------|---------------------|
| 1 | 0.5 | 0.35 | 0.9 | 0.15 | 0.8 | 1.5394 ± 0.1281 | 1.7274 ± 0.0723 |
| 0.5 | 0.2 | 0.8 | 0.5 | 0.1 | 0.5 | 1.6043 ± 0.0642 | 1.6615 ± 0.0239 |
| 0.3 | 0.1 | 0.6 | 0.2 | 0 | 0.2 | 0.6212 ± 0.0151 | 0.6359 ± 0.0062 |

Table.2

Proof of theorem: Our proof will be based on the properties of IPA estimate derived by Suri and Zazanis [1988] for a single class M/G/1 queue.

If all arrival processes are Poisson, we simply have a multiclass M/G/1 queue. Then we know that the superposed process is also Poisson. So, the equivalent queue is now a single class M/G/1 queue with arrival rate $\lambda = \lambda_1 + \lambda_2$. From Kleinrock [1975], we have formula for T and \bar{T} :

$$\bar{T} = \bar{T} = \bar{s} + \frac{\lambda \sigma^2}{2(1 - \lambda \bar{s})} \quad (14)$$

where $\bar{s} = d\bar{s}_1 + (1-d)\bar{s}_2$ is the mean service time in the equivalent M/G/1 queue and $\sigma^2 = d\sigma_1^2 + (1-d)\sigma_2^2$ the second moment of service time.

In Suri and Zazanis [1988], it has been proved that for a single class M/G/1 queue IPA gives strongly consistent estimate of the first derivative of mean system time of a customer with respect to arrival rate, i.e.

$$\left(\frac{\partial \bar{T}}{\partial \lambda} \right)_{\text{IPA}} = \left(\frac{\partial \bar{T}}{\partial \lambda} \right)_{\text{EXACT}} = \frac{\sigma^2}{2(1 - \lambda \bar{s})^2} \quad (15)$$

$$\text{Because } \frac{\partial W_i}{\partial \lambda_1} = \frac{\partial W_i}{\partial \lambda} \frac{\partial \lambda}{\partial \lambda_1} = \frac{\partial W_i}{\partial \lambda}, \text{ we have } \left(\frac{\partial \bar{T}}{\partial \lambda_1} \right)_{\text{IPA}} = \left(\frac{\partial \bar{T}}{\partial \lambda} \right)_{\text{IPA}} = \frac{\sigma^2}{2(1 - \lambda \bar{s})^2} \quad (16)$$

It also has been shown in Suri and Zazanis [1988] that by using IPA in a single class M/G/1 queue we can express the sensitivity of \bar{T} with respect to any parameter, say, d as

$$\left(\frac{\partial \bar{T}}{\partial d} \right)_{\text{IPA}} = E \left(\frac{\partial s}{\partial d} \right) + \frac{\lambda E \left(s \frac{\partial s}{\partial d} \right)}{1 - \lambda \bar{s}} + \frac{\lambda^2 \sigma^2 E \left(\frac{\partial s}{\partial d} \right)}{2(1 - \lambda \bar{s})^2}. \quad (17)$$

From equation (4), and using assumption A1 we can get

$$E \left(\frac{\partial s}{\partial d} \right) = \bar{s}_1 - \bar{s}_2 \quad \text{and} \quad E \left(s \frac{\partial s}{\partial d} \right) = \frac{1}{2} (\sigma_1^2 - \sigma_2^2) \quad (18)$$

Substitute (18) into (17):

$$\begin{aligned}
\left(\frac{\partial T}{\partial d}\right)_{\text{IPA}} &= (\bar{s}_1 - \bar{s}_2) + \frac{\lambda(\sigma_1^2 - \sigma_2^2)}{2(1-\lambda\bar{s})} + \frac{\lambda^2 \sigma^2 (\bar{s}_1 - \bar{s}_2)}{2(1-\lambda\bar{s})^2} \\
&= (\bar{s}_1 - \bar{s}_2) + \frac{\lambda(\sigma_1^2 - \sigma_2^2) + \lambda^2(\bar{s}_1\sigma_2^2 - \bar{s}_2\sigma_1^2)}{2(1-\lambda\bar{s})^2}
\end{aligned} \tag{19}$$

Therefore, from the chain rule and (1) we obtain the IPA estimate for $dT/d\lambda_1$,

$$\begin{aligned}
\left(\frac{dT}{d\lambda_1}\right)_{\text{IPA}} &= \left(\frac{\partial T}{\partial \lambda_1}\right)_{\text{IPA}} + \left(\frac{\partial T}{\partial d}\right)_{\text{IPA}} \frac{\partial d}{\partial \lambda_1} \\
&= \frac{\sigma^2}{2(1-\lambda\bar{s})^2} + \left((\bar{s}_1 - \bar{s}_2) + \frac{\lambda(\sigma_1^2 - \sigma_2^2) + \lambda^2(\bar{s}_1\sigma_2^2 - \bar{s}_2\sigma_1^2)}{2(1-\lambda\bar{s})^2} \right) \frac{\lambda_2}{\lambda^2} \\
&= \frac{\lambda_2(\bar{s}_1 - \bar{s}_2)}{\lambda^2} + \frac{(1-d)(\sigma_1^2 - \sigma_2^2) + \lambda_2(\bar{s}_1\sigma_2^2 - \bar{s}_2\sigma_1^2) + d\sigma_1^2 + (1-d)\sigma_2^2}{2(1-\lambda\bar{s})^2} \\
&= \frac{\lambda_2(\bar{s}_1 - \bar{s}_2)}{\lambda^2} + \frac{\sigma_1^2 + \lambda_2(\bar{s}_1\sigma_2^2 - \bar{s}_2\sigma_1^2)}{2(1-\lambda\bar{s})^2}
\end{aligned} \tag{20}$$

But by direct substitution of the definition of \bar{s} and σ^2 into (14) we get the theoretical value for $dT/d\lambda_1$:

$$\frac{dT}{d\lambda_1} = \frac{\lambda_2(\bar{s}_1 - \bar{s}_2)}{\lambda^2} + \frac{\sigma_1^2 + \lambda_2(\bar{s}_1\sigma_2^2 - \bar{s}_2\sigma_1^2)}{2(1-\lambda\bar{s})^2} \tag{21}$$

So, we prove that Revised IPA algorithm gives the strongly consistent estimate for $dT/d\lambda_1$.