Perturbation Analysis for Stochastic Fluid Queueing Systems

Yong Liu and Weibo Gong
Department of Electrical and Computer Engineering
University of Massachusetts, Amherst
yonliu,gong@ecs.umass.edu

Abstract

Recent study for congestion control in high speed networks indicates that the derivative information for the congestion at the common buffer for multiple sources could be useful in achieving efficient and fair allocation of the bandwidth (Kelly 1997, 1998). In this paper we present an algorithm that estimates such derivatives for multiple on-off sources. The algorithm has its root in the infinitesimal perturbation analysis (IPA) for the classical queueing systems. Although the traditional IPA algorithm does not give unbiased derivative estimates for multi-class arrivals, we are able to prove the unbiasedness in the case of multi-class on-off sources. The results in this paper may motivate a new look at the end-to-end congestion control issue.

1 Introduction

Congestion control is one of the crucial components in high speed network operation. Despite the enormous success of the well-known TCP/IP congestion control protocol in the past decade, the Internet has evolved to the state that pure end to end congestion control may not be enough. Recently many studies on pricing for congestion control have shown that derivatives of congestion at the common buffer with respect to individual sources may provide useful information for efficient and fair use of the bandwidth (Kelly 1997,1998). This is quite natural viewed from the well-known Arrow-Hurwize resource allocation theory (Arrow 1968). In fact Gallager has used this theory to derive a distributed optimal routing algorithm for single class network (Gallager 1975). In essence this theory says that if a limited resource has to be shared by n users and that the system utility is the sum of each user's utility, then the best way to allocate the resource is to equalize the derivatives of each user's individual utility . In the case that the individual utility is not known to the system, the system can pose a common price for the users to "buy" the resource. Simple adjustment of such a price would eventually lead to an optimal allocation of the resource. Since the essence is to equalize the users' utility derivatives, it is clear that these

derivatives are important and one can use the difference between the derivatives to drive the price adjustment. In the case of deterministic setting this scheme has been done in, for example, in economic systems. However in our current setting, the users send in stochastic on-off flows to be processed by a common server and we are facing a sensitivity analysis problem for a fluid queue. Basically we have a deterministic server with service rate c and n on-off sources flow into the common buffer waiting to be processed. Thus the question is to find a simple algorithm to estimate the derivatives of the backlog v at the common buffer with respect to the parameters of each source. Our main result is that the IPA estimates for the steady state derivatives are unbiased. This paves the way for further discussion of using such derivative information. We make some remarks about our result.

- IPA for classical queueing systems was introduced in the early 80's. See Ho and Cao (1991) and Glasserman (1991) for details of IPA theory. Suri and Fu (1995) is the first paper discussing IPA for fluid queues. The setting and motivation of Suri and Fu (1995) are quite different from ours. A simpler version of this work was first presented in Liu and Gong (1999).

- Although the derivation of the main result is based on the analysis of an infinite buffer system, we will show that the principles carry over to the finite buffer case, with minor modifications to the algorithm (see section 5). This is quite different from the case of traditional IPA when applied to finite buffer queues, where the IPA algorithm usually would not give unbiased result.

- We emphasize that IPA does not give unbiased derivative estimates for traditional discrete event multi-class queues. Fluid models are much more friendly to the elegant IPA algorithm. Since fluid queues are increasingly being used in high speed network analysis, we hope our results would motivate more study along this line.

- The derivative estimates obtained in this paper are sample derivatives. The use of sample derivatives in stochastic approximation type of algorithms for optimization has been well studied and it is known that such algorithms do not converge very fast. However a study in Chong and Ramadge (1993) shows that in queueing systems one can update rather quickly in such schemes and expect quick convergence. For example in the setting of the classical G/G/1 queue Chong and Ramadge (1993) prove that the optimization algorithm could update once every departure. This gives the hope that sample path derivative based

optimization algorithm could be realistic in online use for high speed networks. It is also important to note that in quite generic situations it takes relatively few iterations to get into a suboptimal allocation, which could be good enough (Ho, Sreeniva, 1992).

The rest of this paper is organized as follows. In Section 2 we briefly describe the pricing theory introduced by Kelly et al. (Kelly 1997, 1998), which motivated our study of derivative estimates. Section 3 presents our main results. By looking into the sample path performance function, we prove that IPA gives unbiased derivative estimates for fluid queues feeded by multiple ON-OFF sources. Section 4 presents two numerical examples, one is a single source fluid queue and the other is a multiple source fluid queue. IPA estimates are compared with theoretical results. In Section 5, issues of unbiasedness of IPA for finite buffer systems and variance of IPA for multiple sources are discussed. We conclude this paper in Section 6.

2 Pricing for Rate Control of Elastic Traffic

In this section we summarize the pricing theory developed by Kelly et al. (Kelly 1997, 1998). Figure 1 depicts a small network with two users sharing one bottle-neck link, where x_1 and x_2

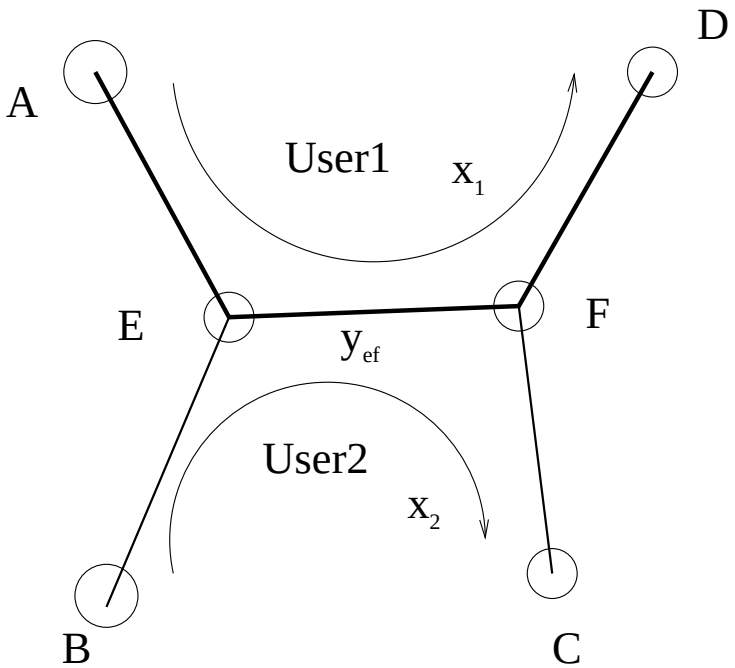


Figure 1: Network of Users with Elastic Traffic

are transmit rates of two users respectively; y_{ef} is available bandwidth at the bottle-neck link. More generally, let's consider a data communication network of links J. A group of users R transmit data through the network. Let $A = [A_{jr}]$ be the route matrix for all users: $A_{jr} = 1$ if user r traverses through link j; $A_{jr} = 0$, otherwise. Given a transmit rate x_r, the user r's utility is $U_r(x_r)$, which is an increasing, strictly concave and continuously differentiable function. We call traffic with this kind of utility function elastic traffic. The total rate on each link is constrained. We may have a hard constraint such as

$$AX \leq C,$$

where $X = [x_1, x_2, \cdots, x_{|R|}]^T$ is the transmit rate vector for all the users, $C = [c_1, c_2, \cdots, c_{|J|}]^T$ is the capacity vector for all the links of the network. We can also approximate the hard constraint by a group of cost functions for all links of the network

$$\begin{cases} C_j(y_j) = 0 & \text{if } y_j << c_j \\ C_j(y_j) \to +\infty & \text{if } y_j \to c_j \end{cases},$$

where y_j is the aggregate rate at link j. Let $\ell(s)$ be the set of links traversed by user s. The problem for the overall system is to maximize the objective function:

$$L(X) = \sum_{r \in R} U_r(X_r) - \sum_{j \in J} C_j(\sum_{s: j \in \ell(s)} x_s).$$

Given $\{U_r(X_r)\}$ and $\{C_j(y_j)\}$, the problem can be solved in a centralized way. However the utility function of each end user maybe unknown for other users in the network. A distributed algorithm is presented in Kelly's papers, which use the intelligence of end users to cooperatively drive the system to its optimum. Briefly, each link looks at its aggregate rate and sends back the derivative information $dC_j(y_j)/dy_j$ to all its users. This information is called shadow price. It can be understood as a price for an unit traffic at link j within the network pricing framework. In a cooperative environment, it can be treated as a marking mechanism for a router to communicate with end users. For each user, upon receiving feedbacks from all the links on its route, it will adjust its transmission rate according to current shadow prices and the derivative of its own utility function.

One example of utility function is $U_r(x_r) = w_r \log x_r$. End user's scheme of adjusting its rate is

$$\frac{d}{dt} x_r(t) = k\{w_r - x_r(t) \sum_{j \in \ell(r)} \mu_j(t)\},$$

where $\mu_j(t)$ is the feedback from link j:

$$\mu_j(t) = \frac{d}{dy} C_j(y)|_{y=\sum_{s:j\in\ell(s)} x_s(t)},$$

k is a positive constant which determines the magnitude of user's rate adaption. Within the pricing framework, w_r can be understood as the total amount of money that user r would like to pay for its network transmission. This equation suggests additive increase in transmit rate x_r if there is no congestion (or $\mu_j(t) = 0, \forall j \in L(r)$), and multiplicative decrease if there is congestion (or $x_r(t)\sum_{j\in r}\mu_j(t)$ dominates over w_r). This is related to the way TCP congestion control scheme works.

Kelly et al. have shown the stability of the algorithm by constructing a Lyapunov function

$$L(X) = \sum_{r\in R} w_r \log x_r - \sum_{j\in J} C_j(\sum_{s:j\in\ell(s)} x_s).$$

For more general utility function $U_r(x_r)$, the user adapting scheme is

$$\begin{cases} w_r(t) = x_r(t)\frac{d}{dx_r}U_r(x_r(t)) \\ \frac{d}{dt}x_r(t) = k\{w_r(t) - x_r(t)\sum_{j\in L(r)}\mu_j(t)\} \end{cases},$$

where $w_r(t)$ is called user adaption. The distributed algorithm will drive the system to the optimum of the objective function

$$L(X) = \sum_{r\in R} U_r(x_r) - \sum_{j\in J} C_j(\sum_{s:j\in\ell(s)} x_s).$$

More generally, if the link cost function is of the form $C_j(x_1, \cdots, x_{|R|})$, link j sends partial derivative $\partial C_j(X)/\partial x_r$ to user r. Then the system can still converge to the optimum of

$$L(X) = \sum_{r\in R} U_r(x_r) - \sum_{j\in J} C_j(X).$$

3 IPA Estimator for Infinite Buffer Queue

In Section 2, we see that derivatives of congestion at the common buffer with respect to the individual sources is very important in resource management of congested network. But the model used in Section 2 is a deterministic model. It doesn't capture the statistical characteristics of network traffic. In this paper, we use parameterized stochastic fluid models, such as Markov ON-OFF model, to model users' data transfer behavior. Each user's utility is a function of the

model parameters. In order to drive the network to its optimum state, derivatives information of congestion is needed. We use the average backlog at a link to measure the congestion. Infinitesimal Perturbation Analysis (IPA) enables us to get unbiased estimates for partial derivatives of the link backlog with respect to the parameters of individual users.

3.1 System Model

Consider a fluid queueing system (see Figure 2) with one server and feeded by M ON-OFF flows. Assume both ON and OFF periods of source i follow distributions with scale parameter θ_i^1 and θ_i^2 respectively. The average length of the ith flow 's ON periods is $1/\mu_i$. And the average length of the ith flow 's OFF periods is $1/\lambda_i$. When flow i is on, workload is feeded into the queue at rate h_i. We assume that the queueing system has an infinite buffer. When the buffer is non-empty, the server processes workload at rate c. Assume flow's peak rate $h_i > c$, for $i = 1, \cdots, M$. In order for the system to be stable, we enforce

$$\sum_{i=1}^{M} \frac{h_i \lambda_i}{\lambda_i + \mu_i} < c.$$

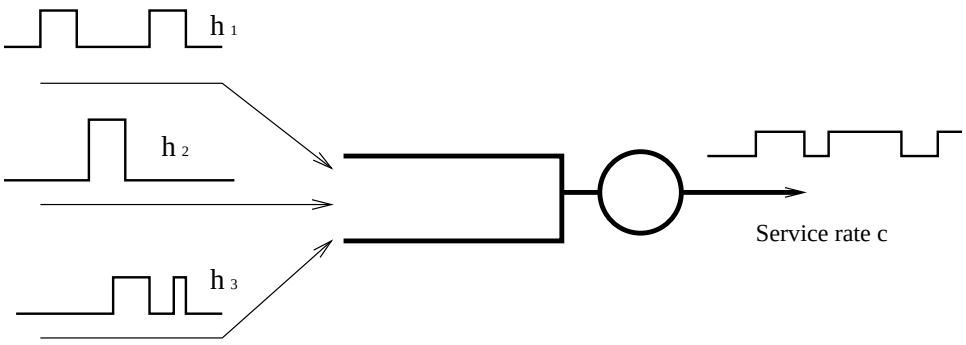


Figure 2: Fluid Queue Feeded by 3 ON-OFF Processes

For fluid queueing systems, Little's Law still holds (Konstantopoulos, Zazanis 1997). The expected waiting time is proportional to the average queue length. We are interested in the average queue length of the system over a finite time period T, the performance index is $J_T(\{\theta_i^1, \theta_i^2\}) = E\{L_T(\{\theta_i^1, \theta_i^2\}, \xi)\}$, where ξ is a sequence of i.i.d uniform random variables which determine the lengths of ON and OFF periods in a sample path. We try to estimate the partial derivatives $\{\partial J_T(\{\theta_i^1, \theta_i^2\})/\partial\theta_i^1, \partial J_T(\{\theta_i^1, \theta_i^2\})/\partial\theta_i^2\}$, for $i = 1, \cdots, M$. Our study focus on IPA for $\{\partial J_T(\{\theta_i^1, \theta_i^2\})/\partial\theta_i^1\}$. It can be easily generalized to obtain IPA for $\{\partial J_T(\{\theta_i^1, \theta_i^2\})/\partial\theta_i^2\}$.

Without lose of generality, we study the queue length derivative with respect to the first source. Let $\theta = 1/\theta_1^1, h = h_1$. The length of the source 1's ith active period $A_i = \theta \times U_i$, where $\{U_i, i = 1, \cdots\}$ are independent samples of random variable U which determines the distribution of source 1 's active periods. Let $q(t)$ be the queue length at time t, $f_1(t)$ be the total flow from source 1 until time t, $f_2(t)$ be the total flow from all other sources until time t, $c(t)$ be the total flow served by the server until t. Then

$$\begin{aligned} q(t) &= f_1(t) + f_2(t) - c(t), \\ \frac{dc(t)}{dt} &= c \times 1(q(t) > 0), \\ \frac{df_1(t)}{dt} &= h \times 1(\text{source 1 is on at } t). \end{aligned}$$

Sample path performance is

$$L_T(\theta, \xi) = \frac{1}{T} \int_0^T q(t, \theta, \xi) dt.$$

3.2 Unbiasedness of IPA Estimator

Given the above model, we are ready to show the unbiasedness of IPA estimator for multiple source fluid queue. Let

$$\begin{aligned} \delta f_1(t, \theta, \xi) &= f_1(t, \theta + \delta\theta, \xi) - f_1(t, \theta, \xi), \\ \delta c(t, \theta, \xi) &= c(t, \theta + \delta\theta, \xi) - c(t, \theta, \xi), \\ \delta q(t, \theta, \xi) &= q(t, \theta + \delta\theta, \xi) - q(t, \theta, \xi), \end{aligned}$$

then

$$L_T(\theta + \delta\theta, \xi) - L_T(\theta, \xi) = \frac{1}{T} \int_0^T \delta f_1(t, \theta, \xi) - \delta c(t, \theta, \xi) dt.$$

Lemma 1 $\sup_{t\in[0,T]} |\delta f_1(t, \theta, \xi) - \delta c(t, \theta, \xi)| \leq \sup_{t\in[0,T]} |\delta f_1(t, \theta, \xi)|$.

Proof: Suppose $\delta\theta > 0$, then $\delta A_i = \delta\theta \times U_i > 0$. It is easy to see that for $t \in [0, T]$, $f_1(t)$ is a non-decreasing function of $\{A_i\}$, so $\delta f_1(t) \geq 0$. Now we want to prove $\delta c(t) \geq 0$ and $\max_{t\in[0,T]} \delta c(t) \leq \max_{t\in[0,T]} \delta f_1(t)$.

If for some $t_1 \in [0, T]$, $\delta c(t_1) < 0$, we prove by contradiction. Since $\delta c(t)$ is continuous and $\delta c(0) = 0$, $\exists T_0 < t_1$, s.t. $\delta c(T_0) = 0$, and $\delta c(t) < 0$ for $t \in (T_0, t_1]$. $\delta c(t)$ is piecewise linear, thus

$\exists T_1 \in (T_0, t_1)$, s.t. for $t \in (T_0, T_1]$,

$$\begin{aligned}
& \frac{d\delta c(t)}{dt} = \frac{dc(t,\theta+\delta\theta)}{dt} - \frac{dc(t,\theta)}{dt} < 0, \\
\Rightarrow \quad & \frac{dc(t,\theta+\delta\theta)}{dt} = 0 \text{ and } \frac{dc(t,\theta)}{dt} = c, \\
\Rightarrow \quad & q(t,\theta+\delta\theta) = 0 \text{ and } q(t,\theta) > 0, \\
\Rightarrow \quad & \delta q(t) = \delta f_1(t) - \delta c(t) < 0, \\
\Rightarrow \quad & \delta f_1(t) < \delta c(t) < 0.
\end{aligned}$$

This contradicts $\delta f_1(t) \geq 0$, therefore if $\delta\theta > 0$, then $\delta c(t) \geq 0$ for all $t \in [0, T]$.

$\delta c(t)$ is a piecewise linear continuous function of t,

$$\frac{d\delta c(t)}{dt} = \begin{cases} c & \text{if } q(t,\theta+\delta\theta) > 0 \text{ and } q(t,\theta) = 0, \\ -c & \text{if } q(t,\theta+\delta\theta) = 0 \text{ and } q(t,\theta) > 0, \\ 0 & \text{otherwise.} \end{cases}$$

Suppose it reaches its maximum at point t^*. If $d^- c(t,\theta)/dt|_{t=t^*} = c$, since $c(t,\theta)$ is piecewise linear, we can find a $T_0 < t^*$, s.t. $dc(t,\theta)/dt|_{t=t^*} = c$ for $t \in (T_0, t^*)$ and $d^- c(t,\theta)/dt|_{t=T_0} = 0$. Because $dc(t,\theta+\delta\theta)/dt \leq c$, for $t \in (T_0, t^*)$,

$$\begin{aligned}
& \frac{d\delta c(t)}{dt} \leq 0, \quad \Rightarrow \quad \delta c(t^*) \leq \delta c(T_0), \quad \Rightarrow \quad \frac{d^- c(t,\theta)}{dt}|_{t=T_0} = 0, \\
\Rightarrow \quad & q(T_0,\theta) = 0 \text{ and } c(T_0,\theta) = f_1(T_0,\theta) + f_2(T_0).
\end{aligned}$$

So we have

$$\begin{aligned}
\delta c(t^*) \leq \delta c(T_0) \leq & f_1(T_0,\theta+\delta\theta) + f_2(T_0) - c(T_0,\theta) \\
& = f_1(T_0,\theta+\delta\theta) - f_1(T_0,\theta) = \delta f_1(T_0) \leq \sup_{t\in[0,T]} \delta f_1(t).
\end{aligned}$$

Thus for $t \in [0, T]$, if $\delta\theta > 0$, $\delta f_1(t) \geq 0, \delta c(t) \geq 0$,

$$\begin{aligned}
\sup_{t\in[0,T]} |\delta f_1(t,\theta,\xi) - \delta c(t,\theta,\xi)| \leq & \max\{ \sup_{t\in[0,T]} \delta c(t,\theta,\xi), \sup_{t\in[0,T]} \delta f_1(t,\theta,\xi)\} \\
& = \sup_{t\in[0,T]} |\delta f_1(t,\theta,\xi)|.
\end{aligned}$$

For $\delta\theta < 0$, we have similar argument for the conclusion $\delta c(t) \leq 0$ and $\min_{t\in[0,T]} \delta c(t) \geq \min_{t\in[0,T]} \delta f_1(t)$. So we have

$$\sup_{t\in[0,T]} |\delta f_1(t,\theta,\xi) - \delta c(t,\theta,\xi)| \leq \sup_{t\in[0,T]} |\delta f_1(t,\theta,\xi)|.$$

Lemma 2 *If $\theta_{\min} \leq \theta$ and $\theta_{\min} \leq \theta + \delta\theta$, then*

$$|\frac{L_T(\theta+\delta\theta,\xi) - L_T(\theta,\xi)}{\delta\theta}| \leq \frac{T \times h}{\theta_{\min}}. \tag{1}$$

Proof:

$$\begin{aligned}
|L_T(\theta+\delta\theta,\xi) - L_T(\theta,\xi)| &= \frac{1}{T}|\int_0^T \delta f_1(t,\theta,\xi) - \delta c(t,\theta,\xi)dt| \\
&\leq \frac{1}{T}\int_0^T |\delta f_1(t,\theta,\xi) - \delta c(t,\theta,\xi)|dt \\
&\leq \sup_{t\in[0,T]} |\delta f_1(t,\theta,\xi) - \delta c(t,\theta,\xi)| \\
&\leq \sup_{t\in[0,T]} |f_1(t,\theta+\delta\theta,\xi) - f_1(t,\theta,\xi)|.
\end{aligned}$$

If $\theta_{\min} \leq \theta \leq \theta + \delta\theta$, $f_1(t,\theta,\xi)$ is the total flow injected by source 1 until time t,

$$f_1(t,\theta,\xi) = h \times \{ \sum_{i=1}^{N(t,\theta,\xi)} A_i(\theta) + R(t,\theta,\xi)\},$$

where $N(t,\theta,\xi)$ is the number of complete ON periods within $[0,t]$, $R(t,\theta,\xi)$ is the truncated length of the possible ON period crossing over t.

$$\begin{aligned}
f_1(t,\theta+\delta\theta,\xi) - f_1(t,\theta,\xi) =& h \times \{ \sum_{i=1}^{N(t,\theta+\delta\theta,\xi)} A_i(\theta+\delta\theta) \\
&+ R(t,\theta+\delta\theta,\xi) - \sum_{i=1}^{N(t,\theta,\xi)} A_i(\theta) - R(t,\theta,\xi)\}.
\end{aligned}$$

It is clear that $N(t,\theta,\xi)$ is a non-increasing function of θ. If $N(t,\theta,\xi) > N(t,\theta+\delta\theta,\xi)$, then

$$\sum_{i=1}^{N(t,\theta+\delta\theta,\xi)} A_i(\theta+\delta\theta) + R(t,\theta+\delta\theta,\xi) \leq \sum_{i=1}^{N(t,\theta,\xi)} A_i(\theta+\delta\theta).$$

If $N(t,\theta,\xi) = N(t,\theta+\delta\theta,\xi) = k$, then the $(k+1)th$ ON period of θ must begin earlier then the $(k+1)th$ ON period of $\theta+\delta\theta$, which means $R(t,\theta+\delta\theta,\xi) < R(t,\theta,\xi)$. In both cases,

$$\begin{aligned}
f_1(t,\theta+\delta\theta,\xi) - f_1(t,\theta,\xi) &\leq h \sum_{i=1}^{N(t,\theta,\xi)} (A_i(\theta+\delta\theta) - A_i(\theta)) = h \sum_{i=1}^{N(t,\theta,\xi)} (\delta\theta \times U_i) \\
&\leq h\delta\theta \sum_{i=1}^{N(t,\theta_{\min},\xi)} U_i = \frac{h\delta\theta}{\theta_{\min}} \sum_{i=1}^{N(t,\theta_{\min},\xi)} A_i(\theta_{\min}) \leq \frac{h \times \delta\theta \times T}{\theta_{\min}}.
\end{aligned} \tag{2}$$

Then we have

$$|\frac{L_T(\theta+\delta\theta,\xi)-L_T(\theta,\xi)}{\delta\theta}| \leq \frac{h\times T}{\theta_{\min}}. \tag{3}$$

Similar argument holds for $\theta_{\min} \leq \theta+\delta\theta < \theta$. So, as long as $\theta_{\min} \leq \theta$ and $\theta_{\min} \leq \theta+\delta\theta$, inequality (3) holds.

Theorem 1 *For fixed ξ, $L_T(\theta,\xi)$ is uniformly continuous with respect to θ over $[\theta_{\min},\theta_{\max}]$.*

Proof: It is an immediate consequence of Lemma 2.

The trajectory of $q(t)$ over $[0,T]$ can be divided into N busy periods and one possible residual period crossing over T. In Figure 3, we have a trajectory consisting of three busy periods and one residual period. The trajectory is determined by the timing of all the jump events of Markov ON-OFF sources. We call the trajectory of θ the nominal trajectory, the trajectory of $\theta+\delta\theta$ the perturbed trajectory.

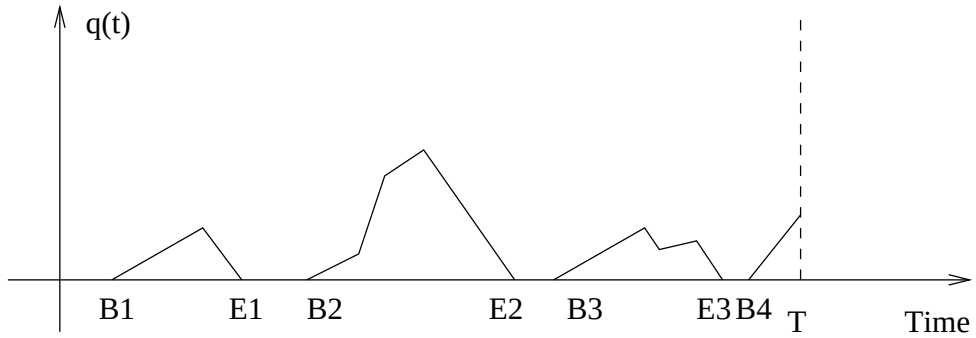


Figure 3: Sample Path of $q(t)$ over $[0,t]$

Definition 1 *The nominal and the perturbed trajectory is said to be similar as long as all the jump events within the ith busy period (or the residual period) of the nominal trajectory still belong to the ith busy period (or the residual period) of the perturbed trajectory and that the order of all these jump events remain unchanged after perturbation.*

In Figure 4, we show an example of similar nominal and perturbed sample paths.

Lemma 3 *Denote by $Trac(\theta,\xi)$, $Trac(\theta+\delta\theta,\xi)$ the nominal and the perturbed trajectory respectively, $\forall\theta\in[\theta_{\min},\theta_{\max}]$, for a.s. ξ, $\exists\ \varepsilon(\theta,\xi)$, s.t., $Trac(\theta,\xi)$ is similar to $Trac(\theta+\delta\theta,\xi)$, if $|\delta\theta|<\varepsilon(\theta,\xi)$.*

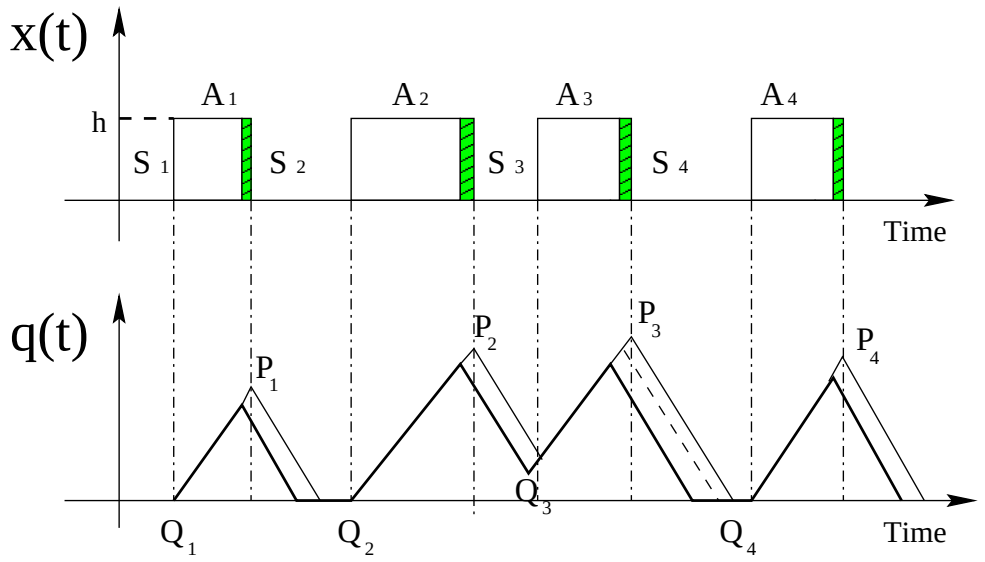


Figure 4: Similar Nominal and Perturbed Sample Path

Proof: 1) For *a.s.* ξ, $\exists\ \varepsilon_1(\theta,\xi)$, *s.t.*, the order of all jump events remains unchanged after a perturbation of $|\delta\theta| < \varepsilon_1(\theta,\xi)$.

Let $J_{\min}$ be the minimum of all intervals between jumps. Because intervals between two consecutive jumps of one source are continuous random variable, $P\{\xi : J_{\min}(\xi) = 0\} = 0$. The upper bound of perturbation on one particular jump event is

$$|\sum_{i=1}^{N(t,\theta_{\min},\xi)} \delta\theta \times U_i| = \frac{|\delta\theta|}{\theta_{\min}} \sum_{i=1}^{N(t,\theta_{\min},\xi)} A_i(\theta_{\min}) \leq \frac{|\delta\theta| \times T}{\theta_{\min}}.$$

So we can select $\varepsilon_1(\theta,\xi) = J_{\min} \times \theta_{\min}/T$ to ensure the order of jump events unchanged.

2) For *a.s.* ξ, $\exists\ \varepsilon_2(\theta,\xi)$, *s.t.*, two busy periods will not merge into one busy period after a perturbation of $|\delta\theta| < \varepsilon_2(\theta,\xi)$.

Let $B_i(\theta), E_i(\theta)$ be the start and end time of the ith busy period of the nominal trajectory, and $B_i^1(\theta), E_i^1(\theta)$ be the time of source 1 's first and last jump event within the ith busy period. Figure 5 is an example of a busy period starting with source 2 and ending with source 1. Let $B_i^1(\theta+\delta\theta), E_i^1(\theta+\delta\theta)$ be the time of those two jumps in the perturbed trajectory. As in 1), $|\delta B_i^1|, |\delta E_i^1| \leq T|\delta\theta|/\theta_{\min}$. For the perturbed trajectory, let $B_i(\theta+\delta\theta)$ be the time of the earliest jump of all jumps within the ith busy period of the nominal trajectory; let $E_i(\theta+\delta\theta)$ be the virtual finish time for all ON periods within the ith nominal busy period. If we have $B_{i+1}(\theta+\delta\theta) > E_i(\theta+\delta\theta)$ for all two adjacent busy periods, no two busy periods

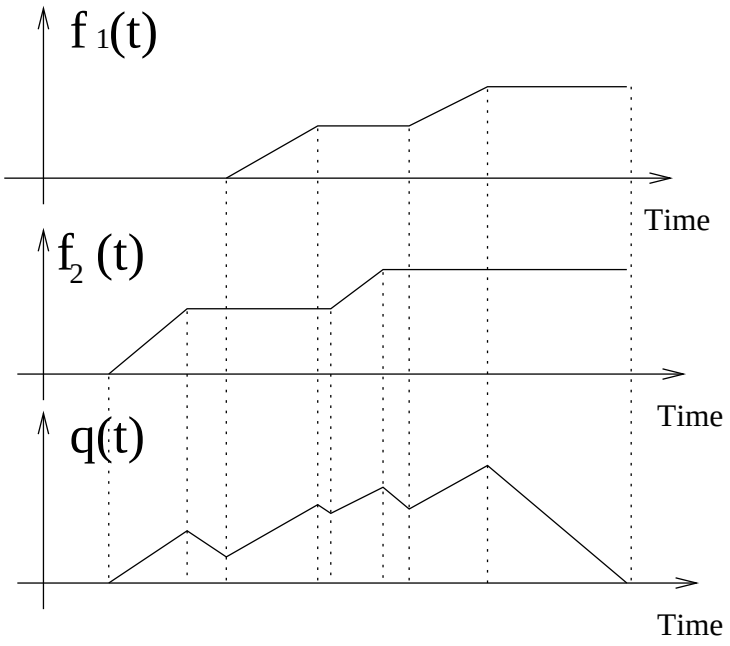


Figure 5: Server Busy Period starting with source 2

will merge together. Let $S_{\min}$ be the minimum length of server's silence periods. We have $P\{\xi : S_{\min}(\xi) = 0\} = 0$. It suffices to require $|\delta E_i| < S_{\min}/2$ and $|\delta B_{i+1}| < S_{\min}/2$. It is clear that $|\delta B_i| \leq |\delta B_i^1|$, and

$$\begin{aligned} |\delta E_i| &< |\delta E_i^1| + |\frac{f_1(E_i^1 + \delta E_i^1, \theta + \delta\theta) - f_1(E_i^1 + \delta E_i^1, \theta)}{c}| \\ &\leq |\delta E_i^1| + \frac{\sup_{t\in[0,T]} |\delta f_1(t)|}{c} \leq \frac{|\delta\theta| \times (h + c) \times T}{\theta_{\min} \times c}. \end{aligned}$$

So if we set $\varepsilon_2(\theta, \xi) = cS_{\min}\theta_{\min}/(2(h + c)T)$, for $|\delta\theta| < \varepsilon_2$, $|\delta E_i| < S_{\min}/2 < S_{i+1}/2$ and $|\delta B_{i+1}| < S_{\min}/2 < S_{i+1}/2$, then the ith and $(i+1)th$ busy period will not merge into one busy period.

3) For $a.s.$ ξ, $\exists$ $\varepsilon_3(\theta, \xi)$, $s.t.$, one busy period will not split into several busy periods after a perturbation of $|\delta\theta| < \varepsilon_3(\theta, \xi)$.

Suppose $\{VT_k(\theta, \xi)\}$, $k = 1, \cdots, K$ are the times of all the jump up events in the nominal trajectory within [0,T], $V_k(\theta, \xi) = q(VT_k(\theta, \xi), \theta, \xi)$. For $a.s.$ ξ, K is finite. If for all $\{k : V_k(\theta, \xi) > 0\}$, we have $V_k(\theta + \delta\theta, \xi) > 0$, then no busy period will split. If the jump up event is

not from source 1, then $VT_k(\theta+\delta\theta,\xi) = VT_k(\theta,\xi)$, so

$$|V_k(\theta+\delta\theta,\xi) - V_k(\theta,\xi)| \le \sup_{t\in[0,T]} |\delta q(t,\theta,\xi)| \le \frac{|\delta\theta| \times T \times h}{\theta_{\min}}.$$

If the jump up event is from source 1, then $\delta VT_k(\theta,\xi) < \frac{|\delta\theta|\times T}{\theta_{\min}}$, so

$$\begin{aligned}|V_k(\theta+\delta\theta,\xi) - V_k(\theta,\xi)| &< |(h + \sum_{i=2}^{M} h_i) \times \delta VT_k(\theta,\xi)| + \sup_{t\in[0,T]} |\delta q(t,\theta,\xi)| \\ &< \frac{|\delta\theta| \times T \times (2h + \sum_{i=2}^{M} h_i)}{\theta_{\min}}.\end{aligned}$$

Let $V_{\min}(\theta,\xi) = \min_{\{k:1\le k\le K, V_k>0\}} V_k(\theta,\xi)$, then $V_{\min}(\theta,\xi) > 0$. So if we set

$$\varepsilon_3(\theta,\xi) = \frac{V_{\min}(\theta,\xi) \times \theta_{\min}}{T \times (2h + \sum_{i=2}^{M} h_i)},$$

no busy period will split.

From argument 1), 2), 3), we have $\forall\theta$, for $a.s.\xi$, $Trac(\theta,\xi)$ is similar to $Trac(\theta+\delta\theta,\xi)$, provided that $\delta\theta < \varepsilon(\theta,\xi) = \min\{\varepsilon_1(\theta,\xi), \varepsilon_2(\theta,\xi), \varepsilon_3(\theta,\xi)\}$

Theorem 2 *$\partial L_T(\theta,\xi)/\partial\theta$ exists at any θ with probability 1.*

Proof: Suppose the sample path of the fluid queue consists of N complete busy periods and one residual period. Define

$$S_T^i(\theta,\xi) = \int_{B_i}^{E_i} q(t,\theta,\xi)dt \quad i = 1,\cdots,N,$$

which is the integration of queue length over the ith busy period. Similarly, for the residual period define

$$S_T^{N+1}(\theta,\xi) = \int_{B_{N+1}}^{T} q(t,\theta,\xi)dt.$$

Let

$$\Omega_1 = \{i: \text{ the } i\text{th busy period is started by source 1}\},$$
$$\Omega_2 = \{i: \text{ the } i\text{th busy period is started by a source other than source 1}\}.$$

Then

$$L_T(\theta,\xi) = \frac{1}{T}(\sum_{i\in\Omega_1} S_T^i(\theta,\xi) + \sum_{i\in\Omega_2} S_T^i(\theta,\xi) + S_T^{N+1}(\theta,\xi)).$$

From Lemma 3, for *a.s.* ξ, when $|\delta\theta| < \varepsilon(\theta,\xi)$, the perturbed sample path is similar to the nominal sample path. Then we have

$$\delta L_T(\theta,\xi) = \frac{1}{T}(\sum_{i\in\Omega_1} \delta S_T^i(\theta,\xi) + \sum_{i\in\Omega_2} \delta S_T^i(\theta,\xi) + \delta S_T^{N+1}(\theta,\xi)). \tag{4}$$

We will establish the existence of $\partial L_T(\theta,\xi)/\partial\theta$ by proving the existence of $\partial S_T^i(\theta,\xi)/\partial\theta$ for $i \in \Omega_1$, $i \in \Omega_2$ and $i = N+1$.

For the $t \in [B_i, E_i]$, let

$$\begin{aligned}
f_i^1(t,\theta,\xi) &\triangleq f_1(t,\theta,\xi) - f_1(B_i,\theta,\xi) \quad \text{and} \quad W_1^i(\theta,\xi) \triangleq \int_{B_i}^{E_i} f_i^1(t,\theta,\xi)dt,\\
f_i^2(t,\theta,\xi) &\triangleq f_2(t,\xi) - f_2(B_i,\xi) \qquad \text{and} \quad W_2^i(\theta,\xi) \triangleq \int_{B_i}^{E_i} f_i^2(t,\theta,\xi)dt,\\
c_i(t,\theta,\xi) &\triangleq c(t,\theta,\xi) - c(B_i,\theta,\xi) \qquad \text{and} \quad W_c^i(\theta,\xi) \triangleq \int_{B_i}^{E_i} c_i(t,\theta,\xi)dt.
\end{aligned}$$

Then

$$S_T^i(\theta,\xi) = W_1^i(\theta,\xi) + W_2^i(\theta,\xi) - W_c^i(\theta,\xi).$$

Let $\{A_{ij}, U_{ij}, j = 1,\cdots,J(i)\}$ be source 1's ON and OFF periods within the ith busy period. Then

$$W_1^i(\theta,\xi) = h \times (\frac{1}{2}(\sum_{j=1}^{J(i)} A_{ij})^2 + \sum_{j=1}^{J(i)}(\sum_{k=1}^{j} A_{ik}) \times U_{ij}).$$

For $k = 1,\cdots,J(i)-1$, $\delta U_{ik} = 0$, $\delta U_{iJ(i)} = \delta E_i - \delta B_i^1 - \sum_{k=1}^{J(i)} \delta A_{ik}$,

$$\begin{aligned}
\delta W_1^i(\theta,\xi) &= h(\sum_{j=1}^{J(i)} A_{ij} \sum_{j=1}^{J(i)} \delta A_{ij} + \sum_{j=1}^{J(i)}(\sum_{k=1}^{j} \delta A_{ik})U_{ij} + \sum_{j=1}^{J(i)} A_{ij}\delta U_{iJ(i)}) + o(|\delta\theta|)\\
&= h(\sum_{j=1}^{J(i)}(\sum_{k=1}^{j} \delta A_{ik})U_{ij} + \sum_{j=1}^{J(i)} A_{ij}(\delta E_i - \delta B_i^1)) + o(|\delta\theta|).
\end{aligned}$$

Case I: If $i \in \Omega_2$, $\delta B_i(\theta,\xi) = 0$, $\delta W_2^i(\theta,\xi) = f_i^2(E_i,\theta,\xi) \times \delta E_i$ and

$$\begin{aligned}
\delta W_c^i(\theta,\xi) &= c(E_i(\theta,\xi) - B_i(\theta,\xi))\delta E_i + o(\delta\theta) = c_i(E_i,\theta,\xi) \times \delta E_i + o(\delta\theta),\\
\delta S_T^i(\theta,\xi) &= \delta W_1^i(\theta,\xi) + \delta W_2^i(\theta,\xi) - \delta W_c^i(\theta,\xi)\\
&= h \times (\sum_{j=1}^{J(i)}(\sum_{k=1}^{j}\delta A_{ik}) \times U_{ij} - \sum_{j=1}^{J(i)} A_{ij} \times \delta B_i^1) + (f_i^1(E_i,\theta,\xi)\\
&\quad + f_i^2(E_i,\theta,\xi) - c_i(E_i,\theta,\xi)) \times \delta E_i + o(|\delta\theta|)\\
&= h \times \sum_{j=1}^{J(i)}(\sum_{k=1}^{j}\delta A_{ik}) \times U_{ij} - f_i^1(E_i,\theta,\xi) \times \delta B_i^1 + o(|\delta\theta|).
\end{aligned}$$

Let $\delta\theta \to 0$, $\partial S_T^i(\theta,\xi)/\partial\theta$ exists and

$$\frac{\partial S_T^i(\theta,\xi)}{\partial\theta} = \frac{1}{\theta}(h \times \sum_{j=1}^{J(i)}(\sum_{k=1}^{j} A_{ik}) \times U_{ij} - f_i^1(E_i,\theta,\xi) \times \frac{f_1(E_{i-1},\theta,\xi)}{h}). \tag{5}$$

Case II: If $i \in \Omega_1$, $\delta B_i(\theta,\xi) = \delta B_i^1(\theta,\xi)$, $\delta W_1^i(\theta,\xi)$ and $\delta W_2^i(\theta,\xi)$ are the same as in the previous case, and

$$\begin{aligned}
\delta W_c^i(\theta,\xi) &= \delta\frac{c}{2}(E_i(\theta,\xi) - B_i(\theta,\xi))^2 = c_i(E_i,\theta,\xi) \times (\delta E_i - \delta B_i^1) + o(\delta\theta),\\
\delta S_T^i(\theta,\xi) &= h \times \sum_{j=1}^{J(i)}(\sum_{k=1}^{j}\delta A_{ik}) \times U_{ij} + f_i^2(E_i,\theta,\xi) \times \delta B_i^1 + o(|\delta\theta|).
\end{aligned}$$

Let $\delta\theta \to 0$, we have

$$\frac{\partial S_T^i(\theta,\xi)}{\partial\theta} = \frac{1}{\theta}(h \times \sum_{j=1}^{J(i)}(\sum_{k=1}^{j} A_{ik}) \times U_{ij} + f_i^2(E_i,\theta,\xi) \times \frac{f_1(E_{i-1},\theta,\xi)}{h}). \tag{6}$$

Case III: For $i = N+1$, let $J(N+1)$ be the number of complete active periods of source 1, $E_{N+1} = T$. If the residual period begins with a jump event of a source other than source 1, then

$$\delta B_{N+1} = \delta E_{N+1} = 0 \quad \textit{and} \quad \delta W_2^{N+1}(\theta,\xi) = \delta W_c^{N+1}(\theta,\xi) = 0.$$

Similar to a busy period, we have

$$\delta W_1^{N+1}(\theta,\xi) = h \sum_{j=1}^{J(N+1)} (\sum_{k=1}^{j} \delta A_{(N+1)k}) U_{(N+1)j} - f_{N+1}^1(T,\theta,\xi)\delta B_{N+1}^1 + o(|\delta\theta|).$$

Then

$$\delta S_T^{N+1}(\theta,\xi) = h \sum_{j=1}^{J(N+1)} (\sum_{k=1}^{j} \delta A_{(N+1)k}) U_{(N+1)j} - f_{N+1}^1(T,\theta,\xi)\delta B_{N+1}^1 + o(|\delta\theta|).$$

Let $\delta\theta \to 0$, we have

$$\frac{\partial S_T^{N+1}(\theta,\xi)}{\partial\theta} = \frac{1}{\theta}(h \sum_{j=1}^{J(N+1)} (\sum_{k=1}^{j} A_{(N+1)k}) U_{(N+1)j} - f_{N+1}^1(T,\theta,\xi)\frac{f_1(E_N,\theta,\xi)}{h}). \tag{7}$$

For the residual period which begins with a jump event from source 1, $\delta W_1^{N+1}(\theta,\xi)$ is the same as in the previous case, and

$$\begin{aligned}
\delta B_{N+1} &= \delta B_{N+1}^1, \\
\delta W_2^{N+1}(\theta,\xi) &= 0, \\
\delta W_c^{N+1}(\theta,\xi) &= \delta\frac{c}{2}(T - B_{N+1}(\theta,\xi))^2 = -c_{N+1}(E_{N+1},\theta,\xi) \times \delta B_{N+1}^1 + o(|\delta\theta|), \\
\delta S_T^{N+1}(\theta,\xi) &= h \times \sum_{j=1}^{J(N+1)} (\sum_{k=1}^{j} \delta A_{(N+1)k}) \times U_{(N+1)j} + (c_{N+1}(E_{N+1},\theta,\xi) \\
&\quad - f_{N+1}^1(T,\theta,\xi)) \times \delta B_{N+1}^1 + o(|\delta\theta|).
\end{aligned}$$

Let $\delta\theta \to 0$, we have

$$\begin{aligned}
\frac{\partial S_T^{N+1}(\theta,\xi)}{\partial\theta} =& \frac{1}{\theta}(h \sum_{j=1}^{J(N+1)} (\sum_{k=1}^{j} A_{(N+1)k}) \times U_{(N+1)j} + (c_{N+1}(E_{N+1},\theta,\xi) \\
&- f_{N+1}^1(T,\theta,\xi)) \times \frac{f_1(E_N,\theta,\xi)}{h}).
\end{aligned} \tag{8}$$

In all three cases, $\partial S_T^i(\theta,\xi)/\partial\theta$ exists for *a.s.* ξ. We have

$$\frac{\partial L_T(\theta,\xi)}{\partial\theta} = \frac{1}{T}(\sum_{i\in\Omega_1} \frac{\partial S_T^i(\theta,\xi)}{\partial\theta} + \sum_{i\in\Omega_2} \frac{\partial S_T^i(\theta,\xi)}{\partial\theta} + \frac{\partial S_T^{N+1}(\theta,\xi)}{\partial\theta})$$

exists at any θ with probability 1.

Theorem 3 *IPA is an unbiased derivative estimate, namely we have*

$$E\{\frac{\partial}{\partial\theta}L_T(\theta,\xi)\} = \frac{d}{d\theta}E\{L_T(\theta,\xi)\}.$$

Proof: From Theorem 1, for $\theta \in [\theta_{\min}, \theta_{\max}]$

$$|\frac{L_T(\theta+\delta\theta,\xi) - L_T(\theta,\xi)}{\delta\theta}| \leq \frac{h \times T}{\theta_{\min}}.$$

From Theorem 2 at any θ,

$$\lim_{\delta\theta\to 0} \frac{L_T(\theta+\delta\theta,\xi) - L_T(\theta,\xi)}{\delta\theta} = \frac{\partial L_T(\theta,\xi)}{\partial\theta} \text{ exists with probability 1.}$$

By Lebesgue 's dominated convergence theorem, and choose $g(\xi) \triangleq h \times T/\theta_{\min}$ as the dominating function, we have

$$\begin{aligned}\frac{d}{d\theta}E\{L_T(\theta,\xi)\} &= \lim_{\delta\theta\to 0}\int_\Omega \frac{L_T(\theta+\delta\theta,\xi) - L_T(\theta,\xi)}{\delta\theta}dP \\ &= \int_\Omega \lim_{\delta\theta\to 0} \frac{L_T(\theta+\delta\theta,\xi) - L_T(\theta,\xi)}{\delta\theta}dP \\ &= E\{\frac{\partial L_T(\theta,\xi)}{\partial\theta}\}.\end{aligned}$$

3.3 IPA Algorithm

We have proved the unbiasedness of IPA estimator for multiple source fluid queue. For the clarity of presentation, we summarize the IPA algorithm here.

Suppose we are developing the IPA estimator for a fluid queue feeded by multiple Markov ON-OFF sources as described in Section 3.1. We can calculate IPA estimates from the queue's sample path between $[0, T]$:

- Divide the queue sample path into N busy periods and one possible residual period crossing over T. Let $\{A_{ij}, U_{ij}, j = 1, \cdots, J(i)\}$ be source 1's ON and OFF periods within the ith busy period. Suppose the length of the residual period is R, and it contains $J(N+1)$ complete active periods of source 1: $\{A_{(N+1)j}, j = 1, \cdots, J(N+1)\}$. Also introduce notations
  - v_i^1: the volume of fluid generated by source 1 during the ith busy period (for the residual period, $i = N+1$)
  - v_i^2: the volume of fluid generated by all other sources during the ith busy period

– L_i: the total volume of fluid generated by source 1 till the end of the ith busy period

- If the ith busy period begins with a jump up event of a source other than source 1, calculate

$$S_i = \frac{1}{\theta}(h \times \sum_{j=1}^{J(i)}(\sum_{k=1}^{j} A_{ik}) \times U_{ij} - \frac{v_i^1 \times L_{i-1}}{h}).$$

- If the ith busy period begins with a jump up event of source 1, calculate

$$S_i = \frac{1}{\theta}(h \times \sum_{j=1}^{J(i)}(\sum_{k=1}^{j} A_{ik}) \times U_{ij} + \frac{v_i^2 \times L_{i-1}}{h}).$$

- If the residual period begins with a jump up event of a source other than source 1, calculate

$$S_{N+1} = \frac{1}{\theta}(h \times \sum_{j=1}^{J(N+1)}(\sum_{k=1}^{j} A_{(N+1)k})U_{(N+1)j} - \frac{v_{N+1}^1 \times L_N}{h}).$$

- If the residual period begins with a jump up event of source 1, calculate

$$S_{N+1} = \frac{1}{\theta}(h \times \sum_{j=1}^{J(N+1)}(\sum_{k=1}^{j} A_{(N+1)k})U_{(N+1)j} + (c \times R - \frac{v_{N+1}^1 \times L_N}{h}).$$

- After calculating $\{S_i, i = 1 \cdots N+1\}$, the IPA estimate is simply $\sum_{i=1}^{N+1} S_i/T$.

4 Numerical Examples

Using the previous IPA algorithm, we do simulations for both single source fluid queue and multiple source fluid queue.

4.1 Single Source Fluid Queue

For the first set of simulations, the fluid queue is feeded by one Markov ON-OFF source. Let c be the service rate of the server; h be the peak rate of the ON-OFF source; $1/\mu$ and $1/\lambda$ be the average length of source ON and OFF periods respectively. The following formula characterizes the stationary queue length (see Anick and Mitra, 1982)

$$E[q(0)] = \frac{\lambda h(h-c)}{(\mu+\lambda)(\mu c - \lambda h + \lambda c)}.$$

θ	0.5000	0.6500	0.8000	0.9500	1.1000	1.2500	1.4000
Theoretical Value	0.3331	0.4281	0.5261	0.6302	0.7431	0.8685	1.0104
IPA Esitmate	0.3298	0.4334	0.5268	0.6272	0.7579	0.8731	1.0080
Sample Variance	0.0089	0.0115	0.0220	0.0304	0.0408	0.0658	0.1009
95% Conf. Intv.	0.0117	0.0133	0.0184	0.0216	0.0250	0.0318	0.0394

Table 1: IPA Estimates for Single Source Fluid Queue

T	200	400	800	1600	3200	6400
Theoretical Value	0.33306	0.33306	0.33306	0.33306	0.33306	0.33306
IPA Esitmate	0.33749	0.33690	0.33327	0.32656	0.33381	0.3345
Sample Variance	0.0174	0.01017	0.00842	0.00989	0.00701	0.00676

Table 2: Impact of Simulation Length on Single Source IPA Estimator

If we choose $\theta = 1/\mu$, the formula for queue length derivative with respect to θ is

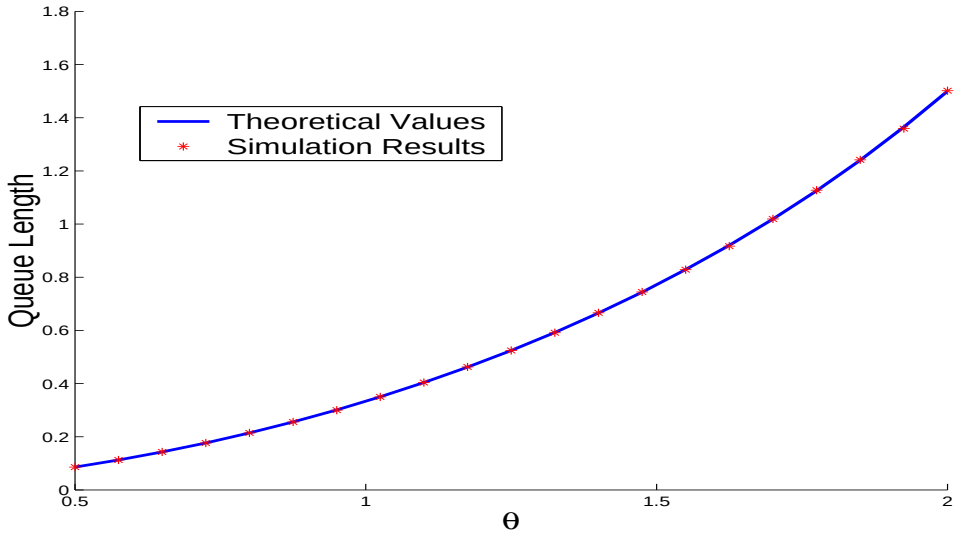
$$\frac{dE[q(0)]}{d\theta} = \frac{\lambda h(h-c)\mu^2(2\mu c - \lambda h + 2\lambda c)}{(\mu+\lambda)^2(\mu c - \lambda h + \lambda c)^2}.$$

We set $h = 1.5$, $c = 1.0$, $\lambda = 0.5$ and vary $\theta = 1/\mu$ from 0.5 to 2 with step size 0.015. At each θ, we run 1000 simulations, each of which simulates the fluid queue for 10000 seconds. We then take the average for both the mean queue length over $[0, T]$ and IPA estimates of queue length derivative with repect to θ. The total running time is 315 seconds on a 2GHz linux box. From Figure 6, we can see the simulation results match very well with the theoretical stationary formula. Table 1 lists statistics for IPA estimator for 7 different values of θ, including theoretical Value, sample variance and length of 95% confidence interval.

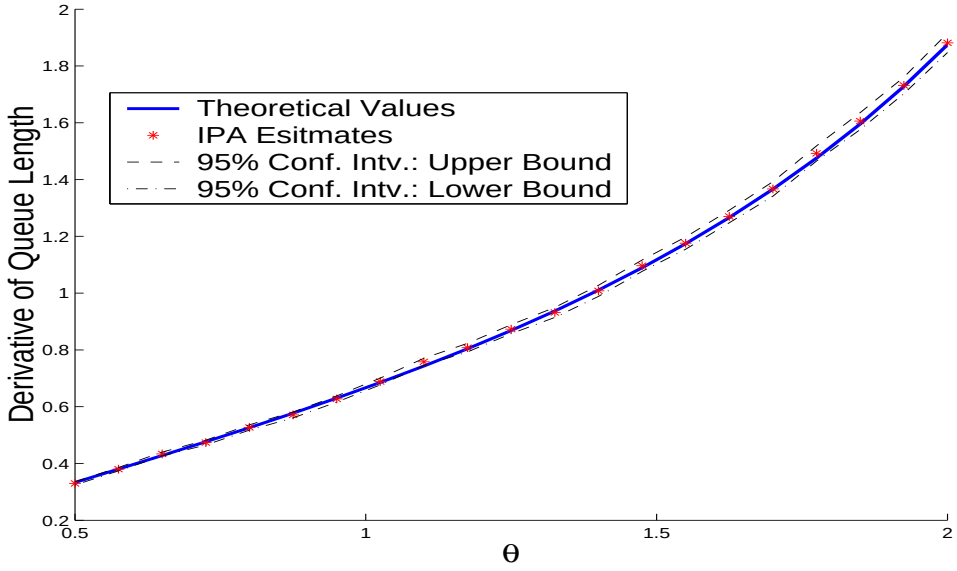
In order to assess the impact of simulation length T on IPA estimates, we do another set of experiments with fixed $\theta = 0.5$, but with T varying from 200 seconds to 6400 seconds. The results are presented in Table 2. For single source IPA estimator, the longer the simulation length, the better the estimate.

4.2 Multiple Source Fluid Queue

Now let's look at IPA for fluid queue with M Markov ON-OFF sources. Denote by c the service rate of the server. For source i, let h_i be the its peak rate; $1/\mu_i$ and $1/\lambda_i$ be the average length of its ON and OFF periods respectively. The stationary queue length formula can be established



(a) Stationary Queue Length



(b) Derivative of Queue Length

Figure 6: Simulation Result for single Source Fluid Queue

θ	0.4000	0.4500	0.5000	0.5500	0.6000	0.6500	0.7000
Theoretical Value	0.7138	0.8040	0.8993	1.0005	1.1086	1.2248	1.3503
IPA Esitmate	0.7351	0.8017	0.8510	1.0166	1.0507	1.3121	1.3172
Sample Variance	33.11	39.01	45.49	52.97	62.47	69.46	82.18
95% Conf. Intv.	0.1842	0.1999	0.2159	0.2329	0.2530	0.2667	0.2901

Table 3: IPA Estimates for Two Sources Fluid Queue

T	200	400	800	1600	3200	6400
Theoretical Value	0.71384	0.71384	0.71384	0.71384	0.71384	0.71384
IPA Esitmate	0.678	0.71230	0.74279	0.68437	0.65596	0.62039
Sample Variance	2.5758	5.217	10.284	20.672	41.204	83.901

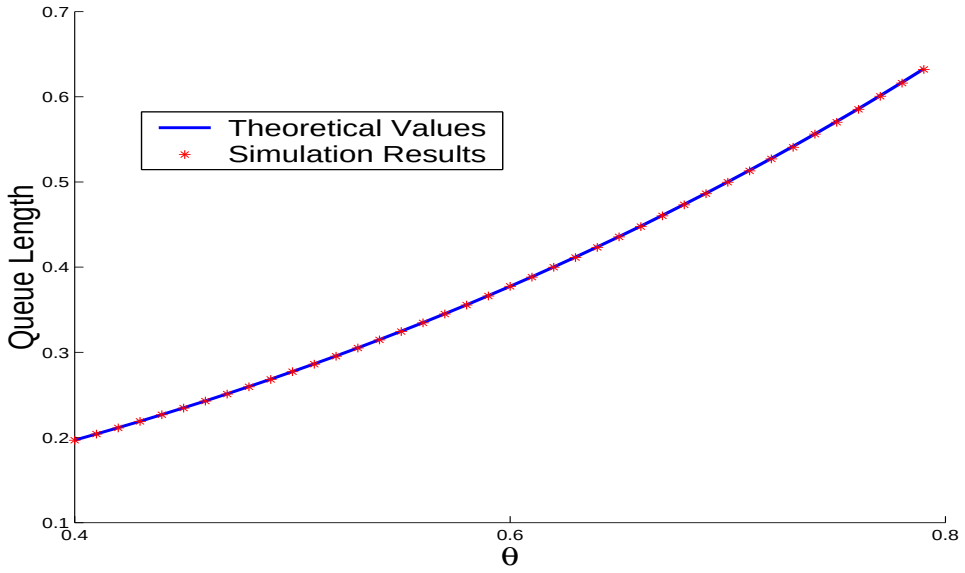
Table 4: Impact of Simulation Length on Two Sources IPA Estimator

through Poisson Driven Stochastic Differential Equation, (see Brockett and Gong, 1999)

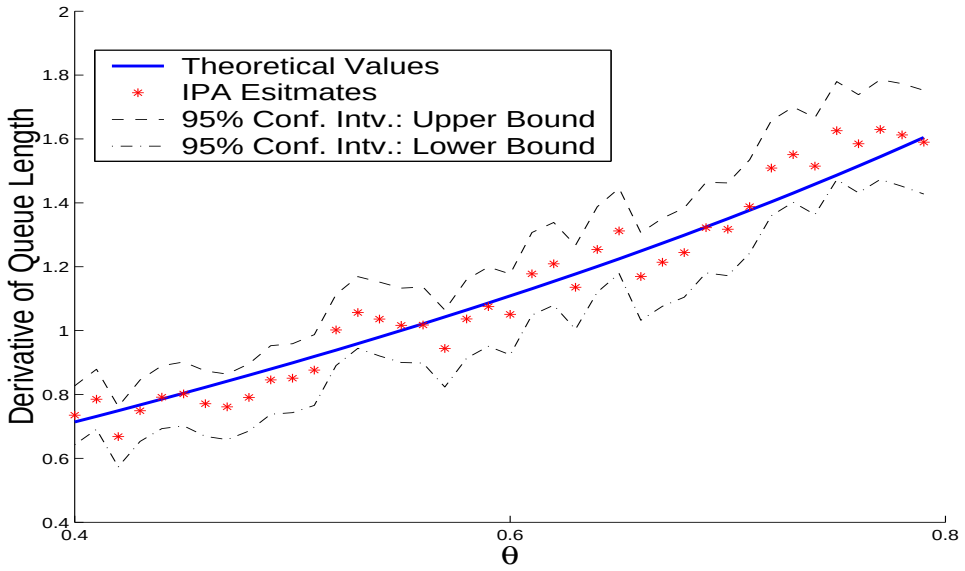
$$E[q(0)] = \frac{1}{c - \sum_{i=1}^{M} r_i} \sum_{i=1}^{M} r_i \tau_i (h_i - c + \sum_{k=1, k\neq i}^{M} r_k),$$

where $\tau_i = 1/(\lambda_i + \mu_i), r_i = h_i \lambda_i/(\lambda_i + \mu_i)$. We simulate a fluid queue feeded by two Markov ON-OFF sources. IPA is used to esitmate the mean queue length derivative with respect to the average length of source 1's ON periods, i.e., $\theta = 1/\mu_1$. We change $\theta = 1/\mu_1$ from 0.4 to 0.8 with step size 0.01. For each θ, we run 15000 simulations with different random sequences. Each simulation simulate the multiple source fluid queue for 2500 seconds. Averages are taken for both the mean queue length and its derivative with respect to θ. Other simulation parameters are: $c = 1.0, h_1 = 1.5, \lambda_1 = 0.5, h_2 = 1.5, \lambda_2 = 0.5, \mu_2 = 3$. The total running time is 1584 seconds on a 2GHz linux box. The results are shown in Figure 7. IPA estimate for multiple sources fluid queue is still unbiased, but the variance is much bigger than the single source case. Table 3 lists statistics for IPA estimator at 7 different θ.

The impact of simulation length T on IPA estimator is investigated by another set of experiments with fixed $\theta = 0.5$, but with T varying from 200 seconds to 6400 seconds. The results are presented in Table 2. We can see the variance of multiple source IPA increase linearly with the simulation length. We will discuss this variance problem in Section 5.



(a) Stationary Queue Length



(b) Derivative of Queue Length

Figure 7: Simulation Result for Two Sources Fluid Queue

5 Discussions

We have proved in Theorem 3 that IPA gives an unbiased estimate for the derivative of buffer content in the fluid queueing system with infinite buffer and multiple ON-OFF sources. In this section, we address some issues of both the unbiasedness of IPA for finite buffer system and the variance issue of multiple sources seen in the numerical example two above.

5.1 IPA for Finite Buffer Fluid Queue

For a fluid queueing system with finite buffer, using similar techniques as in Section 3, we are able to prove uniform continuity of the sample path function $L_T(\theta, \xi)$ by showing that inequality (1) still holds. We can still ensure the similarity of the perturbed trajectory to its nominal trajectory if the parameter perturbation $\delta\theta$ is small enough. Based on these, IPA estimate can be derived similarly and is unbiased.

Without showing the derivation, we give the IPA estimate formula for finite buffer fluid queue feeded by a single source. The trajectory of $q(t)$ can still be divided into busy periods and possible one residual period crossing over T. Denote by B_i, E_i the start and end time of busy period i. If $q(t)$ doesn't hit the buffer upper bound Q during the ith busy period, $\partial S_T^i(\theta, \xi)/\partial\theta$ can be derived in the same way as in infinite buffer case. If $q(t)$ hits the buffer upper bound m_i times during the ith busy period, let H_i^j denote the jth hitting time, D_i^j denote the jth departure time of $q(t)$ from the upper bound for $1 \leq j \leq m_i$, and $D_i^0 \triangleq B_i$, $D_i^{m_i+1} \triangleq E_i$. We can divide the ith busy period into $m_i + 1$ sub-periods: $\eta_{ij} = [D_i^{j-1}, D_i^j)$, for $1 \leq j \leq m_i + 1$. Figure 8 shows the trajectory of $q(t)$ within a busy period which is split into 3 sub-periods.

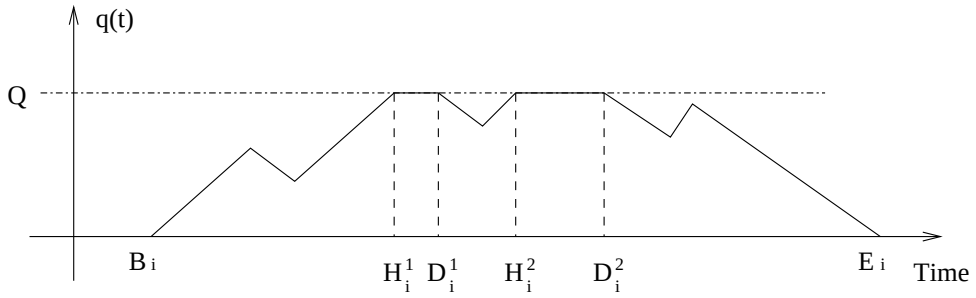


Figure 8: Trajectory of $q(t)$ with Finite Buffer

For the ith busy period we have

$$\frac{\partial S_T^i}{\partial \theta} = \sum_{j=1}^{m_i+1} \frac{\partial}{\partial \theta} \int_{\eta_{ij}} q(t)dt. \tag{9}$$

Suppose that there are $n(i,j)$ active periods of the source within η_{ij}: $A_{ij}^1, \cdots, A_{ij}^{n(i,j)}$. Let U_{ij}^k be the source silent period right after A_{ij}^k, then

$$\frac{\partial}{\partial \theta} \int_{\eta_{ij}} q(t)dt = \frac{h}{\theta} \sum_{k=1}^{n(i,j)-1} \sum_{l=1}^{k} A_{ij}^l \times U_{ij}^k + \frac{1}{\theta} \sum_{k=1}^{n(i,j)} A_{ij}^k \times Q \text{ for } 1 \le j \le m_i \tag{10}$$

and

$$\frac{\partial}{\partial \theta} \int_{\eta_{i(m_i+1)}} q(t)dt = \frac{h}{\theta} \sum_{k=1}^{n(i,m_i+1)} \sum_{l=1}^{k} A_{i(m_i+1)}^l \times U_{i(m_i+1)}^k. \tag{11}$$

For the residual busy period $[B_{N+1}, E_{N+1}]$, if $q(t)$ doesn't hit Q, then IPA is the same as in the infinite buffer case. If $q(t)$ does hit Q and $q(T) < Q$, divide the residual busy period into $m_{N+1}+1$ sub-periods in the same way as we did for a busy period. IPA estimates for all the sub-periods other than the last one are the same as in (10). For the last sub-period $\eta_l \triangleq \eta_{(N+1)(m_{N+1}+1)}$, let $n_l = n(N+1, m_{N+1}+1)$,

$$\frac{\partial}{\partial \theta} \int_{\eta_l} q(t)dt = \frac{h}{\theta} \sum_{k=1}^{n_l} \sum_{l=1}^{k} A_{(N+1)(m_{N+1}+1)}^l U_{(N+1)(m_{N+1}+1)}^k - \frac{f_1(D_{N+1}^{m_{N+1}})q(T)}{h\theta}. \tag{12}$$

If $q(t)$ does hit Q and $q(T) = Q$, which means $D_{N+1}^{m_{N+1}} = T$, then divide the residual busy period into m_{N+1} sub-periods. For the first $m_{N+1}-1$ sub-periods, IPA estimate is the same as in (10). For the last sub-period $\eta_l \triangleq \eta_{(N+1)m_{N+1}}$, let $n_l = n(N+1, m_{N+1})$,

$$\frac{\partial}{\partial \theta} \int_{\eta_l} q(t)dt = \frac{h}{\theta} \sum_{k=1}^{n_l-1} \sum_{l=1}^{k} A_{(N+1)m_{N+1}}^l U_{(N+1)m_{N+1}}^k - \frac{f_1(D_{N+1}^{m_{N+1}-1})Q}{h\theta}. \tag{13}$$

Using this algorithm, we repeat the single source example in Section 4, with the buffer length Q set to be 1. Theoretical value of $E[q(0)]$ can be obtained through (2.18) – (2.23) in Elwalid and Mitra (1991). From Figure 9, we can see IPA estimate for finite buffer case is still unbiased.

5.2 IPA Variance for Multiple Source Fluid Queue

Theorem 3 shows that IPA gives an unbiased estimate for derivative $\frac{d}{d\theta}E\{L_T(\theta,\xi)\}$. We also see in Table 4 that multiple source IPA variance increases with the simulation length. In order

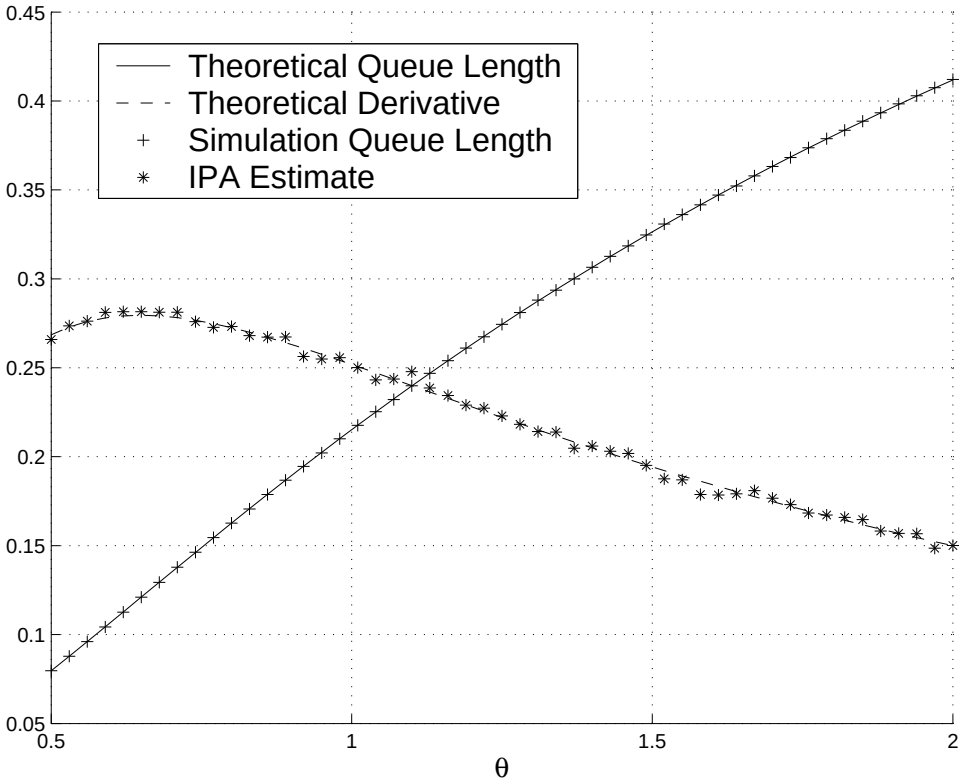


Figure 9: Simulation Result for Finite Buffer Fluid Queue

to obtain good estimate in practice, variance of the IPA estimate is also very important. In this Section, we discuss the variance issue of IPA estimates for multiple source fluid queue.

Based on equation (4), the IPA estimator reads

$$\frac{\partial}{\partial\theta}L_T(\theta,\xi) = \frac{1}{T}(\sum_{i=1}^{N}\frac{\partial S_T^i(\theta,\xi)}{\partial\theta} + \frac{\partial S_T^{N+1}(\theta,\xi)}{\partial\theta}).$$

For the convenience of derivation, let's omit the residual term $\partial S_T^{N+1}(\theta,\xi)/\partial\theta$, which corresponds to the case that queue is empty at time T. We consider two cases: 1) single source; 2) multiple sources.

In case 1), every busy period is started by source 1, by using equation (6), we can see

$$\frac{\partial}{\partial\theta}L_T(\theta,\xi) = \frac{h}{T\theta}\sum_{i=1}^{N}\sum_{j=1}^{J(i)}\sum_{k=1}^{j}A_{ik}\times U_{ij}, \tag{14}$$

where $J(i)$ is the number of the source's ON periods which belong to the server's ith busy period, A_{ij}, U_{ij} are the length of the jth ON and OFF period within ith busy period. Because

server's busy periods are independent, in long run, $Var(\frac{\partial}{\partial\theta}L_T(\theta,\xi))$ is inversely proportional to simulation length T. This suggests when $T \to \infty$, we can get consistent IPA estimate for single source fluid queue.

In case 2), let $\Omega_1 = \{i :$ the ith busy period is started by source 1 $\}$ and $\Omega_2 = \{i :$the ith busy period is started by a source other than 1 $\}$. Combining equation (5) and (6), we will have:

$$\begin{aligned}\frac{\partial}{\partial\theta}L_T(\theta,\xi) = \frac{h}{T\theta}\sum_{i=1}^{N}\sum_{j=1}^{J(i)}\sum_{k=1}^{j}A_{ik}U_{ij} + \frac{1}{T\theta h}\sum_{i\in\Omega_1}f_i^2(E_i,\theta,\xi)f_1(E_{i-1},\theta,\xi) \\ - \frac{1}{T\theta h}\sum_{i\in\Omega_2}f_i^1(E_i,\theta,\xi)f_1(E_{i-1},\theta,\xi),\end{aligned} \tag{15}$$

where $f_i^1(E_i,\theta,\xi)$ is the amount of workload generated by source 1 in the server's ith busy period; $f_i^2(E_i,\theta,\xi)$ is the workload generated by all other sources. $f_1(E_{i-1},\theta,\xi)$ is the total amount of workload generated by source 1 till the end of server's $i-1$th busy period. We have proved that $|\frac{\partial}{\partial\theta}L_T(\theta,\xi)|$ is bounded by $hT/\theta_{\min}$, which is a very loose bound when T is big. Let's look at the variance of the IPA estimate. The first term in equation (15) is exactly the same as the only term in equation (14). Its variance goes down to zero as T goes to ∞. When T is big, variance of the second and third term tend to grow proportionally to T and there is no explicit relationship between these two terms which can ensure their contribution to the variance of $\frac{\partial}{\partial\theta}L_T(\theta,\xi)$ can cancel each other. It has been shown in our numerical example for multiple sources that the experimental variance of IPA estimates increases with the simulation length. This explains why the IPA estimates for multiple sources are not as good as those for the single source case.

6 Conclusions and Future Work

We have seen that the derivative information is important to achieve high network resource utilization. However it is not easy to obtain such information in a stochastic network environment. Gibbens and Kelly (1999) have shown an example of sample path shadow prices, where they use a Poisson model for packet arrivals. They show that a simple marking scheme will send accurate shadow prices information to end users. Unfortunately, this is not true for arrivals with more general statistical properties. In this paper, we use parameterized stochastic model for end users' traffic. Based on this, we develop an infinitesimal perturbation analysis type of derivative estimation algorithm to obtain the shadow price at each common buffer. The unbiasedness of

the estimator has been established. Such derivative estimator could be useful in pricing schemes for congestion management in high speed communication networks.

We would like to provide some general comments in the conclusion of this paper. In the derivation of the derivative estimates, we saw that perturbation analysis techniques give better estimates for the single class queue than for the multi-class queue. Since in multi-class queue a small change of a parameter would lead to accumulated changes in the sample path performance index, we would expect that the derivative estimates to incur larger variance. In this paper, we actually show that the estimates are not consistent. In other words, the PA estimates for multi-class queues have increasing variance with the observation length. Although methods exploring regenerative structure of the sample paths could be applied to obtain consistent estimates, the variances are still much bigger than in the single class case. This observation could be significant for network queue management issues in general. Mathematically speaking, any "correct" queue management scheme should make use of the sensitivity information. We believe that to obtain such information based on sample path observations may not be trivial. In general, when multiple flows go through a queue, the sample path queue behavior does not contain enough information for the mean queue length sensitivities for on-line management. This issue needs to be investigated in more depth. We hope the study of this paper will draw more attentions in this direction.

References

[1] Anick, D., Mitra, D. and Sondhi, M. M. 1982. "Stochastic Theory of a Data-Handling System with Multiple Sources" *The Bell System Technical Journal*, HP. 1871-1894, Vol. 61, No. 8.

[2] Arrow, K. J. 1968. "Applications of Control Theory to Economic Growth", Lectures in Applied Mathematics, Vol. 12. (American Mathematical Society: Providence,R.I.).

[3] Boxma, O. J. and Dumas, V. 1998 "The Busy Period in the Fluid Queue " *Proceedings of Sigmetrics/Performance'98*, pp.100-110.

[4] Brockett, R.W., Gong, W.B. and Guo, Y. "Stochastic analysis for fluid queueing systems", *Proceedings of IEEE CDC'99* (1999) pp. 3077-3082.

[5] Chong, E. K. P. and Ramadge, P. J. 1993 "Optimization of Queues Using an Infinitesimal Perturbation Analysis Based Stochastic Algorithm with General Update Times", *SIAM J. Control and Optimization*, Vol. 31 , No. 3, pp. 698-732.

[6] Elwalid, A. and Mitra, D. 1991 "Analysis and design of rate-based congestion control of high speed networks, I: stochastic fluid models, access regulation", *Queueing Systems*, 9 (1991) 29-64

[7] Gallager, R. G. 1975. "A Minimum Delay Routing Algorithm Using Distributed Computation", *IEEE Trans Commun.* COM-23:73-85.

[8] Gibbens, R. J. and Kelly, F. P. 1999. "Resource pricing and the evolution of congestion control" *Automatica* 35(1999)

[9] Glasserman, P. 1991. *Perturbation Analysis for Gradient Estimation*, Kluwer.

[10] Ho, Y. C. and Cao, X. R. 1991. *Perturbation Analysis of Discrete Event Dynamic Systems* Kluwer.

[11] Ho, Y. C., Sreeniva, R. and Vakili, P. 1992. "Ordinal Optimization of Discrete Event Dynamic Systems", *Journal on DEDS*, Vol. 2 #2

[12] Kelly, F., Maulloo, A. and Tan, D. 1998. "Rate control in communication networks: shadow prices, proportional fairness and stability", *Journal of the Operational Research Society.*

[13] Kelly, F. 1997. "Charging and Rate Control for Elastic Traffic" *European Transaction on Telecommunications* pp. 33-37, volume 8.

[14] Konstantopoulos, T. and Zazanis, M. 1997. "Conservation Laws and Reflection Mappings with an Application to Multiclass Mean Value Analysis for Stochastic Fluid Queues", *Stoch. Proc. Appl.* 65, No. 1, 139-146.

[15] Kurose, J. F. and Simha, R. 1989 "Resource Sharing in Distributed Systems", *IEEE Transactions on Computers*, Vol. 38, No. 5, pp. 705-707.

[16] Liu, Y., and Gong, W.B. 1999 "Perturbation Analysis for Stochastic Fluid Queueing System", *Proceedings of 38th IEEE/CDC*, (December 1999).

[17] Suri, R., Fu, B. 1995. "On Using Continuous Lines to Model Discrete Production Lines", *Journal of Discrete Event Dynamic Systems,* Vol. 4, 129-169.