Discrete Event Dyn Syst (2012) 22:197–219
DOI 10.1007/s10626-011-0120-0

Using infinitesimal perturbation analysis of stochastic flow models to recover performance sensitivity estimates of discrete event systems

Chen Yao · Christos G. Cassandras

Received: 22 December 2010 / Accepted: 26 October 2011 / Published online: 7 December 2011

Abstract Stochastic Flow Models (SFMs) form a class of hybrid systems used as abstractions of complex Discrete Event Systems (DES) for the purpose of deriving performance sensitivity estimates through Infinitesimal Perturbation Analysis (IPA) techniques when these cannot be applied to the original DES. In this paper, we establish explicit connections between gradient estimators obtained through a SFM and those obtained in the underlying DES, thus providing analytical evidence for the effectiveness of these estimators which has so far been limited to empirical observations. We consider DES for which analytical expressions of IPA (or finite difference) estimators are available, specifically $G/G/1$ and $G/G/1/K$ queueing systems. In the case of the $G/G/1$ system, we show that, when evaluated on the same sample path of the underlying DES, the IPA gradient estimators of states, event times, and various performance metrics derived through SFMs are, under certain conditions, the same as those of the associated DES or their expected values are asymptotically the same under large traffic rates. For $G/G/1/K$ systems without and with feedback, we show that SFM-based derivative estimates capture basic properties of finite difference estimates evaluated on a sample path of the underlying DES.

Keywords Discrete event systems · Stochastic flow models · Infinitesimal perturbation analysis

C.G. Cassandras was supported in part by the National Science Foundation under Grant EFRI-0735794, by AFOSR under grants FA9550-07-1-0361 and FA9550-09-1-0095, by DOE under grant DE-FG52-06NA27490, and by ONR under grant N00014-09-1-1051.

C. Yao (✉) · C. G. Cassandras
Division of Systems Engineering and Center for Information and Systems Engineering, Boston University, Brookline, MA 02446, USA
e-mail: cyao@bu.edu

C. G. Cassandras
e-mail: cgc@bu.edu

1 Introduction

The study of Discrete Event Systems (DES) is based on well-developed modeling frameworks in which the system dynamics are driven by the occurrence of different events defined over some given event set (Cassandras and Lafortune 2008). When event occurrence rates get extremely high, however, analysis becomes prohibitively complex; even well-designed discrete event simulations have impractically slow execution times. In this case, one seeks alternative models through which the system dynamics are *abstracted* to an appropriate level that retains essential features enabling effective and accurate control and optimization. This is often the case in systems where random phenomena play different roles at different time scales and typically gives rise to stochastic hybrid system models (Cassandras and Lygeros 2006); in such systems some event-driven dynamics are retained to capture switches between different "modes" while the remaining dynamics are abstracted into differential equations describing the system state evolution within each such mode.

Fluid models are an example of this abstraction process applied to a large class of DES. Fluid models have been shown to be very useful in studying communication networks (Anick et al. 1982; Liu et al. 1999), manufacturing systems (Connor et al. 1994) and, more generally, settings where users compete over different sharable resources. While in most traditional fluid models the flow rates involved are treated as fixed parameters, a *Stochastic Flow Model* (SFM), as introduced in Cassandras et al. (2002), has the extra feature of treating the flow rates themselves as *stochastic processes*. With virtually no limitations imposed on the properties of such processes, a new approach for sensitivity analysis and optimization was recently proposed, based on Infinitesimal Perturbation Analysis (IPA). The essence of this approach is the on-line estimation of gradients (sensitivities) of certain performance measures with respect to various controllable parameters. These estimates may be incorporated in standard gradient-based algorithms to optimize parameter settings of the underlying DES. IPA was originally developed as a technique for evaluating gradients of sample performance functions in queueing systems and using them as unbiased gradient estimates of performance metrics expressed as expectations of these sample functions (Cassandras and Lafortune 2008). However, IPA estimates become biased (hence unreliable for control purposes) when dealing with aspects of queueing systems such as multiple user classes, blocking due to limited resource capacities, and various forms of feedback control. The emergence of SFMs has rekindled the interest in IPA because SFMs allow us to circumvent these limitations, yielding simple unbiased gradient estimates of useful metrics even in the presence of blocking and a variety of feedback control mechanisms, as in Cassandras (2006) and Wardi et al. (2009). In addition, recent work has also extended this approach to multiclass SFMs and to the study of non-cooperative stochastic resource contention games (Yao and Cassandras 2009a, b). It should be stressed that, although the IPA gradient estimators are derived on the SFM abstraction, they are evaluated using the data observed from the underlying DES sample path, and are ultimately used to drive the online optimization of the original DES.

The effectiveness of this approach that combines IPA and SFMs has been supported by successful implementation in various problems (Cassandras et al. 2002; Yao and Cassandras 2009a; Wardi et al. 2009; Yu and Cassandras 2004). However, there still lacks an explicit connection between the gradient estimators obtained through a SFM and those obtained in the underlying DES; this is because the

overall approach is designed to target complex systems, which is precisely where it is impossible to obtain gradient information directly. Thus, there has been no analytical evidence verifying the effectiveness thus far empirically observed.

In this paper, we aim to make such explicit connections between performance gradient estimates of SFMs and their underlying DES for some systems where analytical expressions are available. We specifically consider $G/G/1$ and $G/G/1/K$ queueing systems, where performance gradient estimates are available through either IPA or through Finite Perturbation Analysis (FPA) when IPA is not applicable. In the case of the $G/G/1$ system, we show that, when evaluated on the same sample path of the underlying DES, the IPA gradient estimators of states, event times, and various performance metrics derived through SFMs are, for at least certain classes of distributions that are analytically tractable, the same as those of the associated DES or their expected values are asymptotically the same under large traffic rates. Thus, the results in this paper complement previous research by demonstrating that a SFM not only provides a model abstraction for obtaining gradient estimates of systems where this cannot be accomplished directly, but it also recovers the same or approximate gradient estimates for systems where such information can be obtained directly.

The paper is organized as follows. In Section 2, IPA is applied to both a $G/G/1$ queueing system and its SFM counterpart, and relationships between the two are derived. We show that for certain classes of service time distributions the two IPA estimators are asymptotically identical or their expected values are asymptotically identical. In Section 3, we consider a $G/G/1/K$ system where IPA cannot be applied but finite difference estimates can be derived. We show that these finite differences are under certain conditions the same as IPA estimates derived for the SFM of such a system. Section 4 analyzes a $G/G/1/K$ system with feedback, and it is shown that state perturbations derived on the SFM recover properties of sample path state perturbations in the original DES.

2 IPA for a $G/G/1$ queueing system and its SFM

In this section, we study the SFM associated with a $G/G/1$ queueing system (Cassandras et al. 2002; Cassandras 2006). This SFM is shown in Fig. 1 and has state dynamics given by:

$$\frac{dx(t)}{dt^+} = \begin{cases} 0 & x(t) = 0 \text{ and } \alpha(t) \leq \beta(t) \\ \alpha(t) - \beta(t) & \text{otherwise} \end{cases} \tag{1}$$

where $\alpha(t)$, $\beta(t)$ represent the input and output rate respectively, both stochastic processes, and $x(t)$ is the queue content of the system. The random processes $\{\alpha(t)\}$, $\{\beta(t)\}$ are arbitrary (except for mild technical conditions, see Cassandras et al. 2010) typically taken to be piecewise continuous w.p. 1. We will compare the IPA derivative estimator derived through the SFM with the estimator obtained when IPA is applied directly on the original $G/G/1$ system for a common performance metric and establish relationships between the two, including some cases where they are shown to be asymptotically the same as traffic intensity increases. In addition, we show that the event time perturbations of the "common events" shared by the SFM and its DES counterpart have the same expected values under certain conditions.

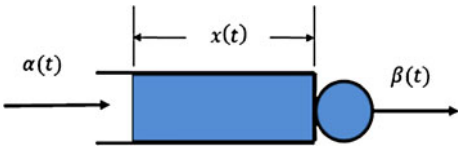

Fig. 1 SFM for $G/G/1$ queueing system

It is well known that IPA can be applied directly on the $G/G/1$ system to derive gradient estimates of performance metrics $J(\theta)$ as a function of some parameter θ. A widely-used performance metric is the mean system time over the first N customers viewed as a function of some parameter θ of the service time distribution. It has been shown (see Section 11.4 in Cassandras and Lafortune 2008) that the IPA estimator of the derivative of $J(\theta)$ with respect to θ is given by

$$\left[\frac{dJ}{d\theta}\right]_{IPA} = \sum_{b=1}^{B}\sum_{i=1}^{n_b}\sum_{j=1}^{i}\frac{dZ_j(\theta)}{d\theta} + \sum_{i=1}^{N-\sum_{b=1}^{B} n_b}\sum_{j=1}^{i}\frac{dZ_j(\theta)}{d\theta} \tag{2}$$

where $\{Z_j\}$ are service times, assumed to be i.i.d random variables. n_b is the number of customers served in the bth busy period, where busy periods are defined as time intervals during which the server of the queue remains busy. B is the number of busy periods for the first N customers, the last double sum accounting for a generally partial final busy period. The derivatives $\frac{dZ_j(\theta)}{d\theta}$ can be evaluated based on the observed value of Z_j and knowledge of the distribution of $\{Z_j\}$ (see Cassandras and Lafortune 2008); for certain classes of distributions, these derivatives take special forms independent of the specific distribution. It has also been shown that $\left[\frac{dJ}{d\theta}\right]_{IPA}$ above is an unbiased estimator of $\frac{dJ}{d\theta}$. Since expressions for all busy periods are the same, in the following we focus on an individual busy period in which the IPA gradient estimator is given by

$$\left[\frac{dJ(\theta)}{d\theta}\right]^{DES} = \sum_{i=1}^{n_b}\sum_{j=1}^{i}\frac{dZ_j(\theta)}{d\theta} \tag{3}$$

In the SFM of the $G/G/1$ system, discrete customers are replaced by continuous flows, hence there is no notion of "system time". Instead, in the SFM we use the total workload as a performance metric, which can be shown to be the same as the overall system time in the long run (Wardi and Melamed 2001). Briefly, if $J^w(T)$ is the total workload and N_T is the number of customers over $[0, T]$, then, by Little's Law, the average workload $\frac{J^w(T)}{T}$ and the average system time $\frac{J(T)}{N_T}$ satisfy $\frac{J^w}{T} = \lambda \cdot \frac{J}{N_T}$, where λ is the arrival rate. When T is large, the conservation law $\lambda \cdot T = N_T$ is satisfied and the previous equation reduces to $J^w = J$, which implies that the total workload can be used to capture the overall system time. Thus, we will use J to denote the workload function in what follows.

Consider a busy period in the $G/G/1$ system which starts at time τ_b and ends at time τ_e. In the corresponding SFM, we fix a busy period with the same starting and ending times so that the workload function is

$$J(\theta) = \int_{\tau_b}^{\tau_e} x(t)dt$$

The sample path derivative with respect to θ is

$$\left[\frac{dJ(\theta)}{d\theta}\right]^{SFM} = \frac{d\tau_e}{d\theta}x(\tau_e) - \frac{d\tau_b}{d\theta}x(\tau_b) + \int_{\tau_b}^{\tau_e} x'(t)dt \tag{4}$$

where $x'(t) \equiv \frac{dx(t)}{d\theta}$. Since $x(\tau_e) = x(\tau_b) = 0$, the above equation reduces to

$$\left[\frac{dJ(\theta)}{d\theta}\right]^{SFM} = \int_{\tau_b}^{\tau_e} x'(t)dt \tag{5}$$

Before proceeding, we provide a brief review of the IPA framework for general stochastic hybrid systems presented in Cassandras et al. (2010) based on which we can evaluate $x'(t)$ above. Let $\theta \in \Theta$ for a given compact, convex set $\Theta \subset \mathbb{R}^l$ be a controllable parameter vector and consider a sample path of such a system. Let $\{\tau_k(\theta)\}$, $k = 1, 2, \ldots$, denote the occurrence times of all events in the sample path. Over an interval $[\tau_k(\theta), \tau_{k+1}(\theta))$, the system is at some mode during which the time-driven state satisfies $\dot{x} = f_k(x, \theta, t)$. An event at τ_k is classified as (i) *Exogenous* if it causes a discrete state transition independent of θ and satisfies $\frac{d\tau_k}{d\theta} = 0$; (ii) *Endogenous*, if there exists a continuously differentiable function $g_k : \mathbb{R}^n \times \Theta \to \mathbb{R}$ such that $\tau_k = \min\{t > \tau_{k-1} : g_k(x(\theta, t), \theta) = 0\}$; and (iii) *Induced* if it is triggered by the occurrence of another event at time $\tau_m \leq \tau_k$. Since the systems considered in this paper do not include induced events, we will limit ourselves to the first two event types. We will use the notation $x'(t) \equiv \frac{\partial x(\theta,t)}{\partial \theta}$, $\tau_k' \equiv \frac{\partial \tau_k}{\partial \theta}$, $k = 0, \ldots, N$, for all state and event time sample derivatives. Then, as shown in Cassandras et al. (2010), $x'(t)$ satisfies:

$$x'(\tau_k^+) = x'(\tau_k^-) + \left[f_{k-1}(\tau_k^-) - f_k(\tau_k^+)\right]\tau_k' \tag{6}$$

$$x'(t) = e^{\int_{\tau_k}^t \frac{\partial f_k(u)}{\partial x}du}\left[\int_{\tau_k}^t \frac{\partial f_k(v)}{\partial \theta}e^{-\int_{\tau_k}^t \frac{\partial f_k(u)}{\partial x}du}dv + \xi_k\right] \tag{7}$$

where $t \in [\tau_k(\theta), \tau_{k+1}(\theta))$ and $\xi_k = x'(\tau_k^+)$ obtained from Eq. 6 unless $x(t)$ experiences a discontinuity and ξ_k must be specified by an explicit state reset condition. In addition, τ_k' in Eq. 6 is either $\tau_k' = 0$ for exogenous events or

$$\tau_k' = -\left[\frac{\partial g_k}{\partial x}f_k(\tau_k^-)\right]^{-1}\left(\frac{\partial g_k}{\partial \theta} + \frac{\partial g_k}{\partial x}x'(\tau_k^-)\right) \tag{8}$$

for endogenous events occurring when $g_k(x(\theta, \tau_k), \theta) = 0$ (with $\frac{\partial g_k}{\partial x}f_k(\tau_k^-) \neq 0$).

To apply the three fundamental IPA Eqs. 6–8 to our SFM, first note that the end of a busy period is an endogenous event satisfying $x(\tau_e) = 0$. In addition, over $[\tau_b, \tau_e)$ we have

$$f = \frac{dx(t)}{dt} = \alpha(t) - \beta(t, \theta), \quad t \in [\tau_b, \tau_e) \tag{9}$$

where Eq. 1 is used with an explicit dependence of the service rate on θ; otherwise, $f = \frac{dx(t)}{dt} = 0$. Then, with $g_k\left(x\left(\theta, \tau_e\right), \theta\right) = x$, Eq. 8 implies that

$$\tau_e' = -\frac{x'(\tau_e^-)}{\alpha(\tau_e) - \beta(\tau_e, \theta)}$$

and using Eq. 6 we get

$$x'(\tau_e^+) \;=\; x'(\tau_e^-) + [\alpha(\tau_e) - \beta(\tau_e, \theta) - 0]\,\tau_e' \tag{10}$$

Combining these two equations results in $x'(\tau_e^+) = 0$. Moreover, applying Eq. 7 over any idle period ending at τ_b we get $x'\left(\tau_b^-\right) = x'(\tau_e^+) = 0$. At τ_b, there are two cases to consider, depending on whether $\alpha(t) - \beta(t, \theta)$ is continuous at this event time. First, if $\alpha(t) - \beta(t, \theta)$ is continuous at τ_b, in view of Eq. 1 we have $0 \geq \alpha(\tau_b^-) - \beta(\tau_b^-, \theta) = \alpha(\tau_b^+) - \beta(\tau_b^+, \theta) \geq 0$, hence $\alpha(\tau_b^-) - \beta(\tau_b^-, \theta) = \alpha(\tau_b^+) - \beta(\tau_b^+, \theta) = 0$. On the other hand, if $\alpha(t) - \beta(t, \theta)$ is not continuous at τ_b, then in order to start the busy period there must exist a jump in $\alpha(t)$ at τ_b such that $\alpha(\tau_b^-) - \beta(\tau_b^-, \theta) \leq 0$ and $\alpha(\tau_b^+) - \beta(\tau_b^+, \theta) > 0$; this is obviously an exogenous event with $\tau_b' = 0$. Thus, using Eq. 6, we get

$$\begin{aligned} x'\left(\tau_b^+\right) \;&=\; x'\left(\tau_b^-\right) + \left[\alpha\left(\tau_b^-\right) - \beta\left(\tau_b^-, \theta\right) - \left(\alpha\left(\tau_b^+\right) - \beta\left(\tau_b^+, \theta\right)\right)\right]\tau_b' \\ &= 0 \end{aligned} \tag{11}$$

in either case.

It now remains to apply Eq. 7 over a busy period so as to evaluate $x'(t)$ in Eq. 5 for all $t \in [\tau_b, \tau_e)$. To do so, consider a busy period of the DES and let n_t be the index of the customer in this busy period that is served at time t, i.e.,

$$n_t = \max\left\{n : n \in \mathbb{N},\; \sum_{j=1}^{n-1} Z_j \leq t - \tau_b\right\} \tag{12}$$

where Z_j is the service time of the jth customer, and the server has obviously already processed $n_t - 1$ customers by time t since the start of this busy period. We can now apply Eq. 5 for all $t \in \left[\tau_b + \sum_{j=1}^{n_t-1} Z_j, \tau_b + \sum_{j=1}^{n_t} Z_j\right)$ observing that in this interval (Eq. 9) holds, therefore $\frac{\partial f_k(u)}{\partial x} = 0$ in Eq. 7. For ease of notation, let $Z(i) \equiv \tau_b + \sum_{j=1}^{i} Z_j$ and we get

$$x'(t) = x'\left(\tau_b^+\right) + \sum_{i=1}^{n_t-1} \int_{Z(i-1)}^{Z(i)} \frac{\partial f}{\partial \theta}(s)\, ds + \int_{Z(n_t)}^{t} \frac{\partial f}{\partial \theta}(s)\, ds \tag{13}$$

where $x'(\tau_b^+) = 0$ from Eq. 11, and $\frac{\partial f}{\partial \theta}(s) = -\frac{\partial \beta(s)}{\partial \theta}$ from Eq. 9. In the SFM, the instantaneous service rate $\beta(s)$ is defined as

$$\beta(s) = \frac{1}{Z_{n_s}} \tag{14}$$

so that $\frac{\partial f}{\partial \theta}(s) = \frac{1}{Z_{n_s}^2} \cdot \frac{dZ_{n_s}}{d\theta}$ and Eq. 13 becomes

$$\begin{aligned} x'(t) &= \sum_{i=1}^{n_t-1} \int_{Z(i-1)}^{Z(i)} \frac{1}{Z_i^2} \cdot \frac{dZ_i}{d\theta} ds + \int_{Z(n_t)}^{t} \frac{1}{Z_{n_t}^2} \cdot \frac{dZ_{n_t}}{d\theta} ds \\ &= \sum_{i=1}^{n_t-1} \frac{1}{Z_i^2} \cdot \frac{dZ_i}{d\theta} \cdot Z_i + \int_{Z(n_t)}^{t} \frac{1}{Z_{n_t}^2} \cdot \frac{dZ_{n_t}}{d\theta} ds \\ &= \sum_{i=1}^{n_t-1} \frac{1}{Z_i} \cdot \frac{dZ_i}{d\theta} + \frac{1}{Z_{n_t}^2} \cdot \frac{dZ_{n_t}}{d\theta} \cdot (t - Z(n_t - 1)) \end{aligned} \tag{15}$$

and Eq. 5 yields

$$\begin{aligned} \left[\frac{dJ(\theta)}{d\theta}\right]^{SFM} &= \sum_{i=1}^{n_b} \int_{Z(i-1)}^{Z(i)} x'(t) dt \\ &= \sum_{i=1}^{n_b} \int_{Z(i-1)}^{Z(i)} \sum_{k=1}^{i-1} \frac{1}{Z_k} \cdot \frac{dZ_k}{d\theta} dt \\ &\quad + \sum_{i=1}^{n_b} \int_{Z(i-1)}^{Z(i)} \frac{1}{Z_i^2} \cdot \frac{dZ_i}{d\theta} \cdot (t - Z(n_t - 1))\, dt \\ &= \sum_{i=1}^{n_b} \left\{ \sum_{k=1}^{i-1} \frac{Z_i}{Z_k} \cdot \frac{dZ_k}{d\theta} + \frac{1}{2} \cdot Z_i^2 \cdot \frac{1}{Z_i^2} \cdot \frac{dZ_i}{d\theta} \right\} \\ &= \frac{1}{2} \sum_{i=1}^{n_b} \frac{dZ_i}{d\theta} + \sum_{i=1}^{n_b} \sum_{k=1}^{i-1} \frac{Z_i}{Z_k} \cdot \frac{dZ_k}{d\theta} \end{aligned} \tag{16}$$

which is the IPA gradient estimator of $\frac{dJ(\theta)}{d\theta}$ obtained through the SFM of the $G/G/1$ queue with output rate $\beta(t, \theta)$ defined as in Eq. 14, and evaluated on the same busy period as the underlying $G/G/1$ queue used to derive Eq. 3.

In what follows, we compare the IPA derivative estimators in Eqs. 16 and 3 for two classes of service time distributions.

Case 1 The system is a $G/D/1$ queue. In this case, $Z_j = Z = \theta$ for all j, so that $\frac{dZ_j(\theta)}{d\theta} = 1$ and Eq. 3 reduces to

$$\left[\frac{dJ(\theta)}{d\theta}\right]^{DES} = \frac{n_b \cdot (n_b + 1)}{2} \tag{17}$$

On the other hand, Eq. 16 becomes

$$\left[\frac{dJ(\theta)}{d\theta}\right]^{SFM} = \frac{n_b}{2} + \frac{n_b\,(n_b - 1)}{2} = \frac{n_b^2}{2} \tag{18}$$

Comparing Eq. 18 with Eq. 17 we have the following asymptotic property:

$$\lim_{n_b\to\infty}\frac{\left[\frac{dJ(\theta)}{d\theta}\right]^{SFM}}{\left[\frac{dJ(\theta)}{d\theta}\right]^{DES}}=\lim_{n_b\to\infty}\frac{\frac{n_b^2}{2}}{\frac{n_b\cdot(n_b+1)}{2}}=1 \tag{19}$$

Thus, in high-traffic settings (implying long busy periods, hence n_b is large), the IPA gradient estimators obtained from the SFM provide highly accurate approximations of the estimators derived when IPA is applied directly to $G/D/1$ systems.

Case 2 The parameter θ is a *scale parameter* of the service time distribution, i.e.,

$$\frac{dZ_k}{d\theta}=\frac{Z_k}{\theta} \tag{20}$$

This applies to a large class of service time distributions, including the entire Erlang family and the uniform distribution. In this case, Eq. 3 reduces to

$$\left[\frac{dJ(\theta)}{d\theta}\right]^{DES}=\sum_{i=1}^{n_b}\sum_{j=1}^{i}\frac{Z_j(\theta)}{\theta}$$

Recalling the fact that $\{Z_j\}$ are i.i.d, we have $E\left[Z_j(\theta)\right]=m$ (constant). Taking expectations (conditioned on the value of n_b) we get

$$\begin{aligned}E\left[\frac{dJ(\theta)}{d\theta}\right]^{DES}&=\frac{1}{\theta}\sum_{i=n_{b-1}+1}^{n_b}\sum_{j=n_{b-1}+1}^{i}E\left[Z_j(\theta)\right]\\&=\frac{n_b(n_b+1)}{2\theta}m\end{aligned} \tag{21}$$

Similarly, taking expectations (conditioned on the value of n_b) on both sides of Eq. 16 we get

$$\begin{aligned}\left[\frac{dJ(\theta)}{d\theta}\right]^{SFM}&=E\left[\frac{1}{2}\sum_{i=1}^{n_b}\frac{Z_i}{\theta}+\sum_{i=1}^{n_b}\sum_{k=1}^{i-1}\frac{Z_i}{Z_k}\cdot\frac{Z_k}{\theta}\right]\\&=\frac{n_b}{2\theta}m+\frac{1}{\theta}\sum_{i=1}^{n_b}\sum_{k=1}^{i-1}E[Z_i]\\&=\frac{n_b}{2\theta}m+\frac{m}{\theta}\frac{n_b(n_b+1)}{2}-\frac{m}{\theta}n_b\\&=\frac{n_b^2}{2\theta}m\end{aligned} \tag{22}$$

Comparing Eq. 22 with Eq. 21, we have

$$\lim_{n_b\to\infty}\frac{E\left[\frac{dJ(\theta)}{d\theta}\right]^{SFM}}{E\left[\frac{dJ(\theta)}{d\theta}\right]^{DES}}=\lim_{n_b\to\infty}\frac{\frac{n_b^2}{2\theta}m}{\frac{n_b(n_b+1)}{2\theta}m}=1 \tag{23}$$

Thus, in high-traffic settings where n_b is large, the expected value of the IPA gradient estimator obtained from the SFM provides a highly accurate approximation of the expected value of the estimator derived when IPA is applied directly to the actual $G/G/1$ system.

In fact, the result in this case can be extended to a more general class of systems, where the parameter θ satisfies the following condition for service times (denoted by Z):

$$E\left[\frac{dZ}{d\theta}\right] = E[Z] \cdot E\left[\frac{1}{Z} \cdot \frac{dZ}{d\theta}\right] \tag{24}$$

with θ being a scale parameter as a special case. If Eq. 24 holds, we have

$$\begin{aligned} E\left[\frac{dJ(\theta)}{d\theta}\right]^{SFM} &= E\left[\frac{1}{2}\sum_{i=1}^{n_b}\frac{dZ_i}{d\theta} + \sum_{i=1}^{n_b}\sum_{k=1}^{i-1}\frac{Z_i}{Z_k}\cdot\frac{dZ_k}{d\theta}\right] \\ &= \frac{n_b}{2}E\left[\frac{dZ_i}{d\theta}\right] + \sum_{i=1}^{n_b}\sum_{k=1}^{i-1}E[Z_i]\cdot E\left[\frac{1}{Z_k}\cdot\frac{dZ_k}{d\theta}\right] \\ &= \frac{n_b}{2}E\left[\frac{dZ}{d\theta}\right] + \sum_{i=1}^{n_b}\sum_{k=1}^{i-1}E\left[\frac{dZ}{d\theta}\right] \\ &= \frac{n_b^2}{2}E\left[\frac{dZ}{d\theta}\right] \end{aligned} \tag{25}$$

and Eq. 23 can also be similarly established.

2.1 Event time derivatives

Another interesting feature of the SFM in Fig. 1 is that, under certain conditions, it gives the same event time derivatives as the actual $G/G/1$ system for the events it shares with it, i.e., starts and ends of busy periods. For a busy period of the $G/G/1$ system, let τ_b and τ_e denote the occurrence times of these two events. As shown in Cassandras and Lafortune (2008), the corresponding event time derivatives are given by

$$\left[\frac{d\tau_b}{d\theta}\right]^{DES} = 0\,, \quad \left[\frac{d\tau_e}{d\theta}\right]^{DES} = \sum_{i=1}^{n_b}\frac{dZ_i}{d\theta} \tag{26}$$

In the associated SFM, the event at τ_b is not necessarily exogenous, as already discussed in the previous section. However, our analysis here is based on the sample path of the actual DES where the start of a busy period is independent of the parameter θ which influences only service times. Therefore, the event at τ_b is exogenous in the context of this discussion and we have

$$\left[\frac{d\tau_b}{d\theta}\right]^{SFM} = 0 \tag{27}$$

which is the same as $\left[\frac{d\tau_b}{d\theta}\right]^{DES}$ in Eq. 26. As for the end of the busy period at τ_e, it is an endogenous event with a switching function $x(\tau_e) = 0$. Taking derivatives with respect to θ gives

$$\frac{dx}{dt}(\tau_e)\left[\frac{d\tau_e}{d\theta}\right]^{SFM} + \frac{dx}{d\theta}(\tau_e) = 0$$

Using Eqs. 9, 14, and 15, the above equation becomes

$$\left(\alpha(\tau_e) - \frac{1}{Z_{n_{\tau_e}}}\right)\cdot\left[\frac{d\tau_e}{d\theta}\right]^{SFM} + \sum_{i=1}^{n_{\tau_e}-1}\frac{1}{Z_i}\cdot\frac{dZ_i}{d\theta}$$

$$+\frac{1}{Z_{n_{\tau_e}}^2}\cdot\frac{dZ_{n_{\tau_e}}}{d\theta}\cdot\left(\tau_e - Z(n_{\tau_e}-1)\right) = 0$$

and since $\tau_e - Z(n_{\tau_e}-1) = Z_{n_{\tau_e}}$ this reduces to

$$\left(\alpha(\tau_e) - \frac{1}{Z_{n_{\tau_e}}}\right)\cdot\left[\frac{d\tau_e}{d\theta}\right]^{SFM}$$

$$+\sum_{i=1}^{n_{\tau_e}-1}\frac{1}{Z_i}\cdot\frac{dZ_i}{d\theta} + \frac{1}{Z_{n_{\tau_e}}}\cdot\frac{dZ_{n_{\tau_e}}}{d\theta} = 0$$

from which we obtain

$$\left[\frac{d\tau_e}{d\theta}\right]^{SFM} = \frac{\sum_{i=1}^{n_{\tau_e}-1}\frac{Z_{n_{\tau_e}}}{Z_i}\cdot\frac{dZ_i}{d\theta} + \frac{dZ_{n_{\tau_e}}}{d\theta}}{1-\alpha(\tau_e)\cdot Z_{n_{\tau_e}}} \tag{28}$$

We now make the following assumption regarding the arrival rate process $\{\alpha(t)\}$ in the SFM:

Assumption 1 At the end of busy periods the arrival rate is zero, i.e., $\alpha(\tau_e) = 0$.

This assumption is motivated by the fact that in the DES there is always a finite time interval between the last arrival event in a busy period and the end of the busy period itself, since there is at least a full service time interval between these two events. Therefore, the instantaneous arrival rate in the SFM must mirror this fact, hence $\alpha(\tau_e) = 0$. Alternatively, we may simply view this assumption as limiting the class of arrival rate processes used in the SFM. Under Assumption 1, Eq. 28 reduces to

$$\left[\frac{d\tau_e}{d\theta}\right]^{SFM} = \sum_{i=1}^{n_{\tau_e}-1}\frac{Z_{n_{\tau_e}}}{Z_i}\cdot\frac{dZ_i}{d\theta} + \frac{dZ_{n_{\tau_e}}}{d\theta} \tag{29}$$

Considering once again the same two cases as in the last section, we have the following.

Case 1 The system is a $G/D/1$ queue. Then, $Z_j = Z = \theta$ for all j, so that $\frac{dZ_j(\theta)}{d\theta} = 1$ and Eqs. 26 and 29 further reduce to

$$\left[\frac{d\tau_e}{d\theta}\right]^{SFM} = n_b = \left[\frac{d\tau_e}{d\theta}\right]^{DES} \tag{30}$$

i.e., event time derivatives obtained through the DES and SFM are the same.

Case 2 The parameter θ is a *scale parameter* of the service time distribution. In this case, taking expectations (conditioned on the value of n_b) in Eqs. 26 and 29 gives

$$E\left[\frac{d\tau_e}{d\theta}\right]^{DES} = E\left[\sum_{i=1}^{n_b}\frac{dZ_i}{d\theta}\right] = n_b \cdot E\left[\frac{dZ}{d\theta}\right]$$

$$\begin{aligned}\left[\frac{d\tau_e}{d\theta}\right]^{SFM} &= E\left[\sum_{i=1}^{n_{\tau_e}-1}\frac{Z_{n_{\tau_e}}}{Z_i}\cdot\frac{dZ_i}{d\theta}+\frac{dZ_{n_{\tau_e}}}{d\theta}\right] \\ &= E\left[\sum_{i=1}^{n_t-1} Z_{n_{\tau_e}}\cdot\frac{1}{\theta}+\frac{Z_{n_{\tau_e}}}{\theta}\right] \\ &= n_b\cdot E\left[\frac{Z}{\theta}\right] = n_b\cdot E\left[\frac{dZ}{d\theta}\right]\end{aligned} \tag{31}$$

which demonstrates that event time perturbations obtained through the SFM and DES have the same expected values in this case. The significance of these properties lies in the fact that it allows us to use performance sensitivities estimated through SFMs (rather than DES) for metrics that depend entirely on event times, such as throughput and resource utilization.

3 SFM for $G/G/1/K$ queueing system

In this section, we study the SFM associated with the $G/G/1/K$ queueing system and compare IPA derivative estimators obtained through this SFM (Sun et al. 2004a, b; Yao and Cassandras 2009a) to the finite difference estimators derived from the actual $G/G/1/K$ system. In particular, we treat the queue capacity K as the parameter of interest. In the underlying DES, K is integer-valued; however, we could easily replace it by a real-valued parameter θ and set $K = \lfloor\theta\rfloor$, the closest integer less than θ. Obviously, derivatives of performance metrics $J(K)$ with respect to K do not exist, but we can evaluate finite differences of the form $\Delta J(K) = J(K) - J(K-1)$. In the associated SFM, however, we can obtain derivative estimates with respect to the real-valued parameter θ.

A typical busy period in a sample path of the $G/G/1/K$ system is shown in Fig. 2. If this is the jth busy period, the first time an arrival event occurs while the queue is at capacity is denoted by $\tau_{j,1}$ and the time when the busy period ends is denoted by at $\tau_{j,e}$. One can observe in the figure that, when the buffer capacity K increases from 2 to 3, there is a workload increase represented by the shaded area, which indicates a discontinuity in the workload function with respect to the parameter K. Such discontinuities arise even if the parameter of interest is a real-valued one such as a parameter of the service time distribution and, as mentioned in the introduction, they result in biased IPA gradient estimators. However, these discontinuities are eliminated in the SFM abstraction, which enables the use of IPA (Fig. 3).

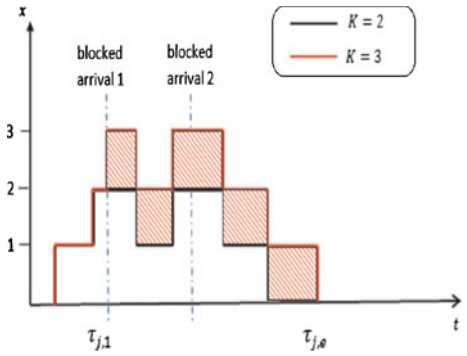

Fig. 2 Sample path of a busy period of $G/G/1/K$ system

The dynamics of the SFM for the $G/G/1/K$ system are

$$\frac{dx(t)}{dt^+} = \begin{cases} 0 & \text{if } x = 0 \text{ and } \alpha(t) \le \beta(t) \\ 0 & \text{if } x = \theta \text{ and } \alpha(t) \ge \beta(t) \\ \alpha(t) - \beta(t) & \text{otherwise} \end{cases} \tag{32}$$

where θ is the buffer capacity that corresponds to K in the $G/G/1/K$ counterpart. The loss rate resulting from buffer overflows is given by

$$l(t) = \begin{cases} \alpha(t) - \beta(t) & \text{if } x = \theta \text{ and } \alpha(t) \ge \beta(t) \\ 0 & \text{otherwise} \end{cases} \tag{33}$$

Similar to prior work on SFMs (e.g., Cassandras 2006; Wardi et al. 2009), the queue content can be either empty, full, or neither. Accordingly, the sample path can be decomposed into three types of intervals: an interval over which $x(t) = 0$ corresponds to an *empty period* (EP); an interval over which $x(t) = \theta$ corresponds to a *full period* (FP); a *nonboundary* period (NBP) is a supremal interval during which $0 < x(t) < \theta$. A *boundary period* (BP) is either an empty or a full period. The performance metrics we consider for this SFM are the average workload:

$$Q_T(x, \theta) = \frac{1}{T} \int_0^T x\,(t)\, dt \tag{34}$$

and the average loss rate:

$$L_T(x, \theta) = \frac{1}{T} \int_0^T l\,(t)\, dt \tag{35}$$

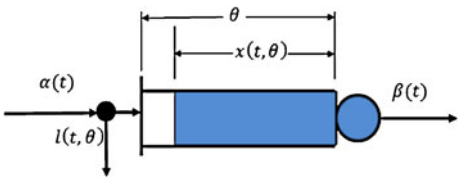

Fig. 3 SFM for $G/G/1/K$ systems

The IPA derivative estimators of these performance metrics have been derived in Cassandras and Lafortune (2008), Sun et al. (2004a) and are given by

$$\frac{dQ_T(x,\theta)}{d\theta} = \sum_{j=1}^{j=N_B} \left(v_{j,e} - v_{j,1}\right), \quad \frac{dL_T(x,\theta)}{d\theta} = N_B \tag{36}$$

where N_B is the number of "qualifying" busy periods, defined as busy periods that include at least one FP; $v_{j,e}$ is the time of the event that ends the jth qualifying busy period, and $v_{j,1}$ is the time of the event when the queue content first reaches θ in that busy period. A typical qualifying busy period is shown in Fig. 4. As in the previous section, we evaluate Eq. 36 using the values of N_B, $v_{j,1}$, and $v_{j,e}$ directly observed on the sample path of the underlying $G/G/1/K$ system with busy periods shown in Fig. 2. Observe that, comparing Figs. 2 and 4, $v_{j,1}$ and $v_{j,e}$ in Eq. 36 are given by

$$v_{j,1} = \tau_{j,1}, \quad v_{j,e} = \tau_{j,e} \tag{37}$$

It was also shown in Sun et al. (2004a) that the state derivatives $\frac{dx}{d\theta}(t)$ in a busy period of the SFM shown in Fig. 4 are given by

$$\frac{dx}{d\theta}(t) = \begin{cases} 0 & t \in [v_{j,0}, v_{j,1}) \\ 1 & t \in [v_{j,1}, v_{j,e}) \end{cases} \tag{38}$$

$$\frac{dx}{d\theta}(v_{j,e}^+) = 0$$

In view of Eq. 37, when these expressions are evaluated on the busy period of Fig. 2, Eq. 38 becomes

$$\frac{dx}{d\theta}(t) = \begin{cases} 0 & t \in [\tau_b, \tau_{j,1}) \\ 1 & t \in [\tau_{j,1}, \tau_{j,e}) \end{cases} \tag{39}$$

$$\frac{dx}{d\theta}(\tau_{j,e}^+) = 0$$

In what follows, we will show that the derivative estimators in Eq. 36 indeed reflect the sensitivities of the original DES to changes in K, when evaluated based on the sample path of the underlying $G/G/1/K$ system under certain conditions such that the change in K does not cause busy periods to coalesce. Thus, although in general (36) cannot recover these sensitivities, we are able to show that there is a connection between them. Let $x_{K-1}(t)$ and $x_K(t)$ denote the state of this system under queue

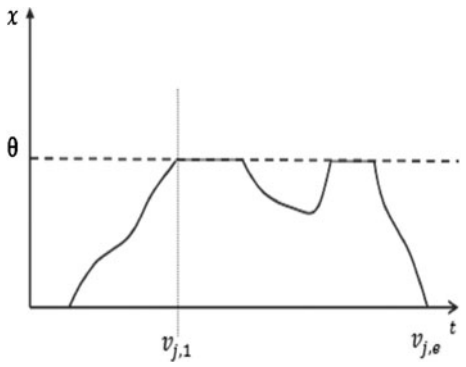

Fig. 4 Sample path of a typical qualifying busy period of the SFM of a $G/G/1/K$ system

capacities $K-1$ and K respectively, which we shall refer to as the *nominal* and and the *perturbed* system respectively. Set $\Delta x(t) \equiv x_K(t) - x_{K-1}(t)$ and observe that $\Delta x(t) = 0$ from the start of a busy period until the time when an arrival blocked in the nominal sample path is accepted in the perturbed sample path, i.e.,

$$\Delta x\left(\tau_{j,1}^{-}\right) = 0, \quad \Delta x\left(\tau_{j,1}^{+}\right) = 1 \tag{40}$$

In the next lemma, we show that $\Delta x(t) = 1$ over the time interval $[\tau_{j,1}, \tau_{j,e})$.

Lemma 1 *In the jth busy period of the $G/G/1/K$ system viewed in isolation from all other busy periods:*

$$\Delta x(t) = \begin{cases} 0 & t \in [\tau_{b,j}, \tau_{j,1}) \\ 1 & t \in [\tau_{j,1}, \tau_{j,e}) \end{cases} \tag{41}$$

$$\Delta x\left(\tau_{j,e}^{+}\right) = 0$$

where $\tau_{b,j}$ is the start of the busy period. In addition, all arrivals that are blocked in the nominal system are also blocked in the perturbed system except the first one at $\tau_{j,1}$.

Proof Clearly, $\Delta x(t) = 0$ for $t \in [\tau_{b,j}, \tau_{j,1})$. Next, let $\{\tau_1, \tau_2, ..., \tau_m\}$ be all event times during $(\tau_{j,1}, \tau_{j,e})$ in increasing order, and $\tau_0 = \tau_{j,1}$. For the interval $[\tau_0, \tau_1)$, it is obvious that $\Delta x(t) = 1$, $t \in [\tau_0, \tau_1)$. Assume that for some integer $k > 0$, $\Delta x(t) = 1$ for all $t \in [\tau_0, \tau_k)$; we will show that $\Delta x(t) = 1$ for all $t \in [\tau_0, \tau_{k+1})$. In the interval (τ_k, τ_{k+1}) there is no event occurring, therefore, $\Delta x(t) = 1$ based on the induction hypothesis. At $t = \tau_{k+1}$, there are two cases depending on whether the event at this time also occurs in the nominal sample path.

Case 1 If the event also occurs in the nominal system, then $x(t)$ changes by the same amount at this event for both nominal and perturbed systems, hence $\Delta x(t)$ remains unchanged, i.e., $\Delta x(\tau_{k+1}) = \Delta x(\tau_{k+1}^{-})$.

Case 2 If the event is an arrival that has been blocked in the nominal system, i.e., $x(\tau_{k+1}^{-}) = K-1$ in the nominal sample path, then by the induction hypothesis, $\Delta x(\tau_{k+1}^{-}) = 1$, hence $x(\tau_{k+1}^{-}) = K$, which implies that the arrival will also get blocked in the perturbed system. Therefore, $\Delta x(\tau_{k+1}) = \Delta x(\tau_{k+1}^{-}) = 1$. This also establishes the sample path property stated in the lemma, i.e., that all arrivals blocked in the nominal system will also get blocked in the perturbed system except the first one at $\tau_{j,1}$.

By the induction argument above, we have

$$\Delta x(t) = 1 \text{ for all } t \in [\tau_0, \tau_k) \tag{42}$$

At $\tau_{j,e}$ the queue becomes empty in both nominal and perturbed system, hence Δx resets to 0, i.e.,

$$\Delta x\left(\tau_{j,e}^{+}\right) = 0 \tag{43}$$

and the results of the lemma are established. □

In view of Lemma 1, we also have the following lemma, whose proof follows immediately by comparing Eq. 41 with the state derivatives (Eq. 39) for the SFM.

Lemma 2 *Suppose that for $G/G/1/K$ sample paths under queue capacities $K-1$ and K respectively, $\Delta x(t) \equiv x_K(t) - x_{K-1}(t) = 0$ for some $t \in (\tau_{j,e}, \tau_{b,j+1})$ for all j. Then, state derivatives $\frac{dx}{d\theta}(t)$ obtained through the SFM have the same values as the perturbations $\Delta x(t)$ obtained from the underlying $G/G/1/K$ system when evaluated on a given $G/G/1/K$ sample path:*

$$\frac{dx}{d\theta}(t) = \Delta x(t), \quad t \in [0, T] \tag{44}$$

Based on the above lemmas, we now show that $\frac{dQ_T(x,\theta)}{d\theta}$ and $\frac{dL_T(x,\theta)}{d\theta}$ given in Eq. 36 are the same as the finite differences $\Delta Q(K) \equiv Q(K) - Q(K-1)$ and $\Delta L(K) \equiv L(K) - L(K-1)$ with $K = \lfloor \theta \rfloor$. For the original $G/G/1/K$ systems, when evaluated on the same $G/G/1/K$ sample path, provided the condition in Lemma 2 holds.

Theorem 1 *Suppose that for $G/G/1/K$ sample paths under queue capacities $K-1$ and K respectively, $\Delta x(t) \equiv x_K(t) - x_{K-1}(t) = 0$ for some $t \in (\tau_{j,e}, \tau_{b,j+1})$ for all j. Then, the SFM-based IPA derivative estimators $\frac{dQ_T(x,\theta)}{d\theta}$ and $\frac{dL_T(x,\theta)}{d\theta}$ given in Eq. 36 and the finite differences $\Delta Q(K), \Delta L(K)$ with $K = \lfloor \theta \rfloor$ obtained for a $G/G/1/K$ sample path satisfy:*

$$\begin{aligned} \frac{dQ_T}{d\theta} &= \Delta Q(K) \equiv Q(K) - Q(K-1) \\ \frac{dL_T}{d\theta} &= \Delta L(K) \equiv L(K) - L(K-1) \end{aligned} \tag{45}$$

Proof In any qualifying busy period of the $G/G/1/K$ system, let $Q_j(K)$ and $L_j(K)$ denote the workload and loss respectively over the jth busy period. Using Lemma 1, we have

$$\Delta Q_j(K) = \Delta x \cdot \left(\tau_{j,e} - \tau_{j,1}\right) = \left(\tau_{j,e} - \tau_{j,1}\right)$$

Moreover, by Lemma 2, there is only one arrival that was previously blocked and will get accepted, therefore

$$\Delta L_j(K) = 1$$

Summing over all qualifying busy periods in $[0, T]$, we get

$$\begin{aligned} \Delta Q(K) &= \sum_j \Delta Q_j(K) = \sum_j \left(\tau_{j,e} - \tau_{j,1}\right) \\ \Delta L(K) &= \sum_j \Delta L_j(K) = N_B \end{aligned} \tag{46}$$

Recalling Eqs. 37 and 36, we have

$$\begin{aligned} \frac{dQ_T}{d\theta}(K) &= \sum_{j=1}^{j=N_B} \left(v_{j,e} - v_{j,1}\right) = \Delta Q(K) \\ \frac{dL_T}{d\theta}(K) &= N_B = \Delta L(K) \end{aligned}$$

which completes the proof. □

We emphasize again that this property of the SFM estimators provides a connection with a set of sample paths of the $G/G/1/K$ system where the change in K does not cause busy periods to coalesce. In addition, the following lemma shows that the SFM also provides event time derivatives that are the same as the associated event time perturbations in the underlying $G/G/1/K$ system.

Lemma 3 *Consider the jth busy period of the $G/G/1/K$ system starting at $\tau_{b,j}$ and ending at $\tau_{j,e}$ viewed in isolation from all other busy periods and the associated SFM over the same busy period. Then, under Assumption* 1*,*

$$\frac{dx}{dt}\left(\tau_{b,j}\right) = \Delta\tau_{b,j} = 0$$
$$\frac{dx}{dt}\left(\tau_{j,e}\right) = \Delta\tau_{j,e}$$

Proof First, based on the event classification reviewed in Section 2, events that start busy periods are exogenous, and independent of the perturbed parameter θ in the SFM or K in the DES, therefore $\frac{dx}{dt}\left(\tau_{b,j}\right) = \Delta\tau_{b,j} = 0$.

As for the event that ends the busy period, i.e., the event at $\tau_{j,e}$ in Fig. 2, in the $G/G/1/K$ system:

$$\Delta\tau_{j,e} = Z$$

where Z is the service time of an additional customer admitted in the perturbed but not the nominal systems. In the associated SFM, the event ending the busy period is endogenous and its occurrence time satisfies the switching function

$$x\left(\tau_{j,e}\right) = 0$$

Taking derivatives on both sides with respect to θ, we get

$$\frac{dx}{dt}\left(\tau_{j,e}\right)\cdot\frac{d\tau_{j,e}}{d\theta} + \frac{dx}{d\theta}\left(\tau_{j,e}\right) = 0$$

Recall that $\frac{dx}{dt}\left(\tau_{j,e}\right) = \alpha\left(\tau_{j,e}\right) - \beta\left(\tau_{j,e}\right)$. Then, using Eq. 39 and Assumption 1, we have

$$-\beta\left(\tau_{j,e}\right)\cdot\frac{d\tau_{j,e}}{d\theta} + 1 = 0$$

and we get

$$\frac{d\tau_{j,e}}{d\theta} = \frac{1}{\beta\left(\tau_{j,e}\right)} = Z = \Delta\tau_{j,e}$$

and the results of the lemma are established. □

4 SFM for $G/G/1/K$ system with feedback

SFMs have been extended to study queueing systems with feedback mechanisms that typically arise in many applications which involve admission or flow control (Yu and Cassandras 2002, 2006, 2004). In this section, we focus on a $G/G/1/K$ system with negative state feedback and its associated SFM studied in Yu and Cassandras (2004) and show that the IPA state derivative estimates derived through the SFM recover some basic properties of sample path finite differences of the original queueing

system. The SFM for a $G/G/1/K$ system with feedback is shown in Fig. 5, where $p(x)$ is a non-decreasing feedback function and the system dynamics are given by

$$\frac{dx(t)}{dt^+} = \begin{cases} 0 & \text{if } x = 0 \text{ and } \alpha(t) - \beta(t) \leq p(0) \\ 0 & \text{if } x = \theta \text{ and } \alpha(t) - \beta(t) \geq p(\theta) \\ \alpha(t) - \beta(t) - p(x(t)) & \text{otherwise} \end{cases} \tag{47}$$

The feedback mechanism is implemented by controlling θ and the function $p(x(t))$, selected to be monotonically nondecreasing in $x(t)$. The net effect is to suppress the incoming flow through rejection of arriving customers (see Yu and Cassandras 2004 for details.) IPA is applied on the SFM of Fig. 5 and derivative estimators are derived for various performance metrics, such as the average workload $Q_T(x, \theta)$ in Eq. 34, as follows:

$$\frac{dQ_T}{d\theta} = \int_0^T \frac{dx}{d\theta}(t)\, dt \tag{48}$$

For the special case of linear feedback, i.e. $p(x) = cx$, the state perturbations $\frac{dx}{d\theta}(t)$ in Eq. 48 were shown in Yu and Cassandras (2004) to be

$$\frac{dx}{d\theta}(t) = \begin{cases} \mathbf{1}[x(\eta_n) = \theta] \cdot e^{-c(t-\eta_n)} & t \in [\eta_n, \xi_n) \\ \mathbf{1}[x(t) = \theta] & t \in [\xi_n, \eta_{n+1}) \end{cases} \tag{49}$$

where $[\eta_n, \xi_n)$ is the nth *nonboundary* period in a sample path of the SFM, i.e., an interval when the queue is neither empty nor full. As in all previous sections, the IPA estimator in Eq. 48 is evaluated on the sample path of the original queueing system, hence η_n, ξ_n are all identified in the underlying DES sample path. In particular, η_n starts an interval which is either (i) the start of a busy period, in which case it ends at some ξ_n with $x(\xi_n)$ reaching the queue capacity or $x(\xi_n) = 0$ ending a busy period without ever reaching the queue capacity, or (ii) the start of an interval with $x(\eta_n)$ at queue capacity followed by the end of a busy period or x reaching the queue capacity again at ξ_n. Thus, $\frac{dx}{d\theta}(t) = e^{-c(t-\eta_n)}$ in Eq. 49 if η_n starts an interval where the queue is at capacity and is otherwise zero.

Considering the underlying $G/G/1/K$ system with negative feedback, the controller operates so that the value of $p(x_K(t))$ determines a fraction of arriving customers that are deliberately rejected even if $x_K(t) < K$. Figure 6 shows examples of the busy period of two such systems with different feedback functions, and in both examples, both nominal and perturbed sample paths are illustrated where the buffer capacity K is increased by 1. The difference between the two examples lies in the "intensities" of the feedback functions selected so that, for all x, we have

$$p_1(x) < p_2(x) \tag{50}$$

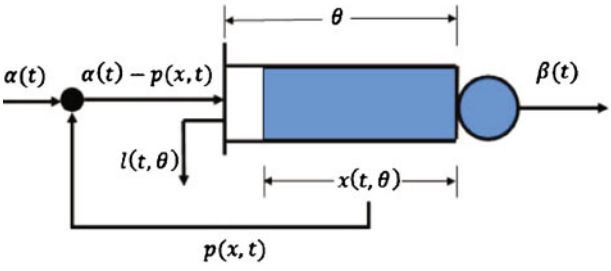

Fig. 5 SFM for $G/G/1/K$ system with feedback

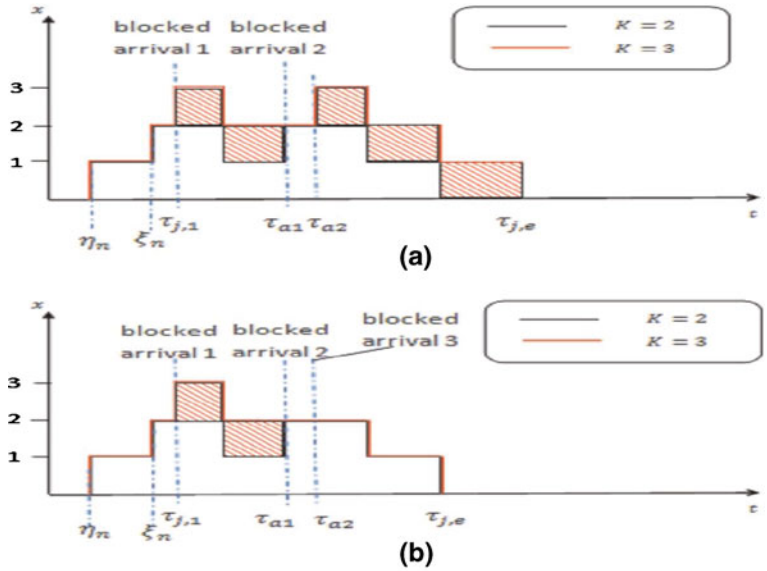

Fig. 6 Examples of a busy period of the $G/G/1/K$ system with feedback function: **a** $p_1(x)$ **b** $p_2(x) > p_1(x)$

In particular, if both systems use linear feedback, i.e., $p_1(x) = c_1 x$, $p_2(x) = c_2 x$, where $c_1, c_2 \in \left(0, \frac{1}{K}\right)$, so that $c_i x \in (0, 1)$, $i = 1, 2$, and the feedback in the *ith* system is manifested by rejecting $c_i x$ fraction of arrivals through a simple randomization admission control scheme. Moreover, Eq. 50 is equivalent to

$$c_1 < c_2 \tag{51}$$

First, we compare Fig. 6 with Fig. 2 where there is no feedback. Observe that, in the perturbed sample paths of these systems, some arrivals that are accepted in the $G/G/1/K$ system of Fig. 2 are rejected in both systems of Fig. 6. For instance, the arrival at τ_{a1} is rejected in the perturbed ($K = 3$) sample path of Fig. 6, while it gets accepted in the perturbed sample path of Fig. 2. This additional rejection is the effect of feedback control, because as the capacity K increases, $x_K(t)$ also increases, which leads to larger $p(x_K(t))$, hence more arrivals are deliberately rejected. This observation implies that the queue in a sample path of the perturbed system with feedback, i.e., $x_K(t)$, is smaller than that of the system without feedback, hence $\Delta x(t) \equiv x_K(t) - x_{K-1}(t)$ is also smaller. Then, together with Lemma 1, we conclude that $\Delta x(t) \leq 1$, for all t, in the systems of Fig. 5.

In addition, by comparing Fig. 6b with Fig. 6a we observe that there are arrivals that are accepted in the perturbed sample path of Fig. 6a but are rejected in the perturbed sample path of Fig. 6b, e.g., the arrival at τ_{a2} in Fig. 6a, b. This is due to the difference in the feedback "intensities" c_1, c_2 , since $p_1(x_K(t)) < p_2(x_K(t))$ in Eq. 50, hence, more arrivals are rejected in the system of Fig. 6b, which further results in smaller $\Delta x(t) \equiv x_K(t) - x_{K-1}(t)$. In the case of linear feedback, this implies that $\Delta x(t)$ is a decreasing function of the feedback intensity parameter c.

Finally, another simple observation is that, the longer the system is in a non-full period, the smaller the perturbation $\Delta x(t)$ is. The simple intuitive reasoning is that, the effect caused by the rejections in the previous full period decays as that full period ends.

These observations lead to the following proposition, which describes three properties of the state perturbations $\Delta x(t)$ in a $G/G/1/K$ system with linear feedback, under the same condition as in Theorem 1.

Proposition 1 *In the $G/G/1/K$ queue with negative feedback, suppose that for sample paths under queue capacities $K-1$ and K respectively, the state perturbation is such that $\Delta x(t) = x_K(t) - x_{K-1}(t) = 0$ for some $t \in (\tau_{j,e}, \tau_{b,j+1})$ for all j. Then, $\Delta x(t)$ in the busy period has the following properties:*

Property 1 $\Delta x(t) \in \{0, 1\}$.

Property 2 *In the case of linear feedback, $\Delta x(t)$ is a non-increasing function of the feedback gain parameter c.*

Property 3 *$\Delta x(t)$ is a non-increasing function of $(t - \eta_n)$, where η_n is the time when the previous full period ends.*

Proof First, we adopt similar notation as in the proof of Lemma 1, so that $\tau_{j,1}$ denotes the first time an arrival event occurs in the jth busy period while the queue is at capacity and $\tau_{j,e}$ is the time when the busy period ends; in addition, $\{\tau_1, \tau_2, ..., \tau_m\}$ are all event times during $(\tau_{j,1}, \tau_{j,e})$ in increasing order, and $\tau_0 = \tau_{j,1}$. Under the condition in the Proposition, clearly, $\Delta x(t) = 0$ for $t \in [\tau_{b,j}, \tau_{j,1})$, and $\Delta x(t) = 1$, $t \in [\tau_0, \tau_1)$. Assume that for some integer $k > 0$, $\Delta x(t) \in \{0, 1\}$ for all $t \in [\tau_0, \tau_k)$; we will show that $\Delta x(t) \in \{0, 1\}$ for all $t \in [\tau_0, \tau_{k+1})$. In the interval (τ_k, τ_{k+1}) there is no event occurring, therefore, $\Delta x(t) = 1$ based on the induction hypothesis. At $t = \tau_{k+1}$, there are 4 cases depending on the type of event occurring at this time, whether it also occurs in the nominal sample path, and the value of $\Delta x(\tau_k^-)$.

Case 1 If the event also occurs in the nominal sample path, then $x(t)$ changes by the same amount at this event for both nominal and perturbed systems, hence $\Delta x(t)$ remains unchanged, i.e., $\Delta x(\tau_k) = \Delta x(\tau_k^-)$.

Case 2 If the event is an arrival that has been blocked in the nominal sample path, i.e., $x(\tau_k^-) = K - 1$ in the nominal sample path, and $\Delta x(\tau_k^-) = 1$, then $x(\tau_k^-) = K - 1 + 1 = K$ in the perturbed sample path, which implies that the arrival will also get blocked in the perturbed system. Therefore, $\Delta x(\tau_k^+) = \Delta x(\tau_k^-) = 1$.

Case 3 If the event is an arrival that has been blocked in the nominal sample path, i.e., $x(\tau_k^-) = K - 1$ in the nominal sample path, and $\Delta x(\tau_k^-) = 0$, then $x(\tau_k^-) = K - 1 + 0 = K - 1$ in the perturbed sample path, which implies that the arrival will get accepted in the perturbed system. Therefore, $\Delta x(\tau_k^+) = \Delta x(\tau_k^-) + 1 = 1$.

Case 4 If $\Delta x(\tau_k^-) = 1$, and the event is an arrival that is rejected in the perturbed sample path due to the feedback control applied, but it has been accepted in the nominal system, then $\Delta x(\tau_k^+) = \Delta x(\tau_k^-) - 1 = 0$. This case arises because of the increase in the feedback intensity, i.e., $c\Delta x(\tau_k^-) = c$, using the fact that $\Delta x(\tau_k^-)$ is 1 in this case.

By the induction argument above, we have

$$\Delta x\,(t) \in \{0, 1\} \text{ for all } t \in [\tau_0, \tau_k) \tag{52}$$

At $\tau_{j,e}$ the queue becomes empty in both nominal and perturbed system, hence Δx resets to 0, i.e.,

$$\Delta x(\tau_{j,e}^{+}) = 0 \tag{53}$$

and *Property* 1 follows.

To prove *Property* 2, we note that in all 4 cases discussed above where $\Delta x\,(t)$ changes, only *Case* 4 depends on the feedback parameter c. In addition, it follows from the analysis therein that, the larger the value of c is, the larger the value of $c\Delta x(\tau_k^{-})$ is, which implies a more frequent occurrence of *Case* 4, resulting in the decrease of $\Delta x\,(t)$. Therefore, $\Delta x\,(t)$ is a non-increasing function of the feedback gain parameter c, and *Property* 2 is established.

Property 3 follows from the fact that, after the full period ends, i.e., $x\,(t) < K - 1$, then only *Case* 1 and *Case* 4 can occur, where $\Delta x\,(t)$ either remains the same or decreases. □

Looking at the state derivative $\frac{dx}{d\theta}\,(t)$ derived through the SFM as given in Eq. 49, it is easy to verify that it satisfies *Property* 2 and *Property* 3 above, i.e., $\frac{dx}{d\theta}\,(t)$ decreases as $(t - \eta_n)$ increases, or as c increases. In addition, note that *Property 1* implies that $0 \leq \Delta x\,(t) \leq 1$, which is also true for $\frac{dx}{d\theta}\,(t)$ in Eq. 49.

Finally, following an analysis similar to that of Theorem 1, we can show that the sample path finite difference in the workload is

$$\Delta Q\,(K) = Q\,(K) - Q\,(K-1) = \int_0^T \Delta x\,(t)\,dt$$

Although we can no longer establish the equality of $\Delta Q\,(K)$ with $\frac{dQ_T}{d\theta}$ obtained through Eqs. 48–49, as we did in Theorem 1 for the $G/G/1/K$ system, the fact that $\frac{dx}{d\theta}\,(t)$ shares the properties of $\Delta x\,(t)$ under the same condition suggests that the SFM-based performance sensitivity estimate provides good approximations of the sensitivity estimate $\Delta Q\,(K)$; this is consistent with the simulation-based results presented in Yu and Cassandras (2004).

5 Conclusions

Motivated by ample empirical evidence to date that performance sensitivity estimates obtained by IPA for SFMs used as abstractions of underlying DES provide accurate approximations of the performance sensitivity estimates of the actual DES, we have established in this paper explicit connections between such estimates for cases where analytical expressions for IPA (or finite difference) estimates are available. In particular, we have considered $G/G/1$ queueing systems, where we

have shown exact and asymptotic results for the equivalence of SFM and DES-based estimates. For $G/G/1/K$ systems without and with feedback, our results are much weaker, showing only that SFM-based derivative estimates capture some basic properties of finite difference estimates evaluated on a sample path of the underlying DES. Whereas in the study of DES IPA applies to a limited system class, in SFMs IPA has been shown in the authors' prior work to boil down to three fundamental equations of virtually arbitrary applicability. These provide the cornerstones for a very general unbiased estimation theory, playing a role similar to the general-purpose equations one uses, for example, in optimal control (state equations, costate equations, optimality conditions, etc). Although the expressions required to describe the performance sensitivity estimators generated by these fundamental equations often appear complicated, their actual implementation is in fact very simple.

References

Anick D, Mitra D, Sondhi MM (1982) Stochastic theory of a data-handling system with multiple sources. Bell Syst Tech J 61:1871–1894

Cassandras CG (2006) Stochastic flow systems: modeling and sensitivity analysis. In: Cassandras CG, Lygeros J (eds) Stochastic hybrid systems, pp 139–167. Taylor and Francis

Cassandras CG, Lafortune S (2008) Introduction to discrete event systems, 2nd edn. Springer

Cassandras CG, Lygeros J (eds) (2006) Stochastic hybrid systems. Taylor and Francis

Cassandras CG, Wardi Y, Melamed B, Sun G, Panayiotou CG (2002) Perturbation analysis for on-line control and optimization of stochastic fluid models. IEEE Trans Automat Contr 47(8): 1234–1248

Cassandras CG, Wardi Y, Panayiotou CG, Yao C (2010) Perturbation analysis and optimization of stochastic hybrid systems. Eur J Control 16(6):642–664

Connor D, Feigin G, Yao DD (1994) Scheduling semiconductor lines using a fluid network model. IEEE Trans Robot Autom 10(2):88–98

Liu B, Guo Y, Kurose J, Towsley D, Gong WB (1999) Fluid simulation of large scale networks: issues and tradeoffs. In: Proceedings of the intl. conf. on parallel and distributed processing techniques and applications, pp 2136–2142

Sun G, Cassandras CG, Panayiotou CG (2004a) Perturbation analysis and optimization of stochastic flow networks. IEEE Trans Automat Contr 49(12):2113–2128

Sun G, Cassandras CG, Panayiotou CG (2004b) Perturbation analysis of multiclass stochastic fluid models. J of Discrete Event Dynamic Systems 14(3):267–307

Wardi Y, Adams R, Melamed B (2009) A unified approach to infinitesimal perturbation analysis in stochastic flow models: the single-stage case. IEEE Trans Automat Contr 55(1):89–103

Wardi Y, Melamed B (2001) Variational bounds and sensitivity analysis of traffic processes in continuous flow models. J of Discrete Event Dynamic Systems 11(3):249–282

Yao C, Cassandras CG (2009a) Perturbation analysis and optimization of multiclass multiobjective stochastic flow models. In: Proceedings of 48th IEEE conference of decision and control, pp 914–919

Yao C, Cassandras CG (2009b) Perturbation analysis and resource contention games in multiclass stochastic fluid models. In: Proceedings of 3rd IFAC conference on analysis and design of hybrid systems, pp 256–261

Yu H, Cassandras CG (2002) Perturbation analysis and optimization of a flow controlled manufacturing system. In: Proceedings of 2002 international workshop on discrete event systems, pp 258–263

Yu H, Cassandras CG (2004) Perturbation analysis of feedback-controlled stochastic flow systems. IEEE Trans Automat Contr 49(8):1317–1332

Yu H, Cassandras CG (2006) Perturbation analysis and feedback control of communication networks using stochastic hybrid models. Nonlinear Analysis 65(6):1251–1280

Chen Yao is a Senior Research Engineer at Global Automation Research at Nalco Company. He received Bachelor degree in Automatic Control from Zhejiang University in 2006, Master degree in Systems Engineering from Boston University in 2009, and Ph.D. in Systems Engineering from Boston University in 2011. He has worked as a Research Assistant in the Center of Information and Systems Engineering (CISE) at Boston University between 2006 and 2010. His research interests lie in the areas of discrete event and hybrid systems, stochastic optimization, and cooperative control, with applications to manufacturing systems, communication systems, and robotics. He is the recipient of several awards, including the General Chair's Recognition Award of Interactive Sessions in the 48th IEEE Conference of Decision and Control (CDC), and Dean's Fellowship at Boston University in 2006.

Christos G. Cassandras is Head of the Division of Systems Engineering and Professor of Electrical and Computer Engineering at Boston University. He is also co-founder of Boston University's Center for Information and Systems Engineering (CISE). He received degrees from Yale University, Stanford University, and Harvard University. In 1982–84 he was with ITP Boston, Inc. where he worked on the design of automated manufacturing systems. In 1984–1996 he was a faculty member at the Department of Electrical and Computer Engineering, University of Massachusetts/Amherst. He specializes in the areas of discrete event and hybrid systems, cooperative control, stochastic optimization, and computer simulation, with applications to computer and sensor networks, manufacturing systems, and transportation systems. He has published over 300 refereed papers in these areas, and five books. He has guest-edited several technical journal issues and serves on several journal

Editorial Boards. He has collaborated with The MathWorks, Inc. in the development of the discrete event and hybrid system simulator SimEvents. Dr. Cassandras was Editor-in-Chief of the *IEEE Transactions on Automatic Control* (1998–2009) and has also served as Editor for Technical Notes and Correspondence and Associate Editor. He is the 2012 President of the IEEE Control Systems Society (CSS) and has served as Vice President for Publications and on the Board of Governors of the CSS. He has chaired the CSS Technical Committee on Control Theory, and served as Chair of several conferences. He has been a plenary speaker at many international conferences, including the *American Control Conference* in 2001 and the *IEEE Conference on Decision and Control* in 2002, and an IEEE Distinguished Lecturer. He is the recipient of several awards, including the 2011 IEEE Control Systems Technology Award, the Distinguished Member Award of the IEEE Control Systems Society (2006), the 1999 Harold Chestnut Prize (IFAC Best Control Engineering Textbook) for *Discrete Event Systems: Modeling and Performance Analysis*, and a 1991 Lilly Fellowship. He is a member of Phi Beta Kappa and Tau Beta Pi. He is also a Fellow of the IEEE and a Fellow of the IFAC.