Application of IPA to Fluid Petri Nets

A. Giua, C. Seatzu
Dept. of Electrical and Electronic Eng., University of Cagliari
Piazza d'Armi — 09134 Cagliari, Italy
Tel: +39-070-675-5759, Fax: +39-070-675-5782, {giua,seatzu}@diee.unica.it.

Y. Wardi
The School of Electrical and Computer Engineering,
Georgia Institute of Technology, Atlanta, USA
yorai.wardi@ece.gatech.edu.

Abstract

Infinitesimal Perturbation Analysis (IPA) recently has been extensively investigated in the setting of fluid queues, where it was shown to yield simple algorithms for computing the gradients of several performance functions. More lately, efforts have been made to extend its application domain from fluid queueing networks to other kinds of stochastic hybrid systems. In this vein, the present paper inaugurates a study of the application of IPA to a class of hybrid Petri nets. The main point of concern is the modeling element of the fluid transition with multiple input places, representing concurrency and synchronization in Petri nets, and not yet studied in the context of IPA. We first derive the IPA gradient of the throughput with respect to fluid flow parameters at the input places, and then consider an example of optimizing throughput in a fork-join system. Simulation experiments are presented in support of the theoretical results. We point out that the main purpose of the paper is to initiate a study of IPA in the setting of hybrid Petri nets, and not to consider application examples.

Published as:
A. Giua, C. Seatzu, Y.Wardi, "Application of IPA to Fluid Petri Nets," *ADHS'09: 3nd IFAC Conf. on the Analysis and Design of Hybrid Systems* (Zaragoza, Spain), Sep 2009.

This work has been partially supported by the European Communitys Seventh Framework Programme under project DISC (Grant Agreement n. INFSO-ICT-224498).

1 Introduction

Infinitesimal Perturbation Analysis (IPA) has been established as a general technique for sensitivity analysis of Discrete Event Dynamic Systems (DEDS), providing algorithms for the gradients (derivatives) of sample performance functions [9, 4]. Ever since its inception in the early nineteen-eighties [10], IPA has been applied mainly to queueing models, where performance metrics of delay and throughput are of primary interest. Its main appeal stems from the simplicity of the formulas it yields for the sample gradients of these performance functions. However, this initial appeal has been dampened by the discovery that in many realistic queueing models, the resulting gradients are statistically biased, thereby casting doubt on the viability of the IPA approach in sample-based performance optimization [7].

To overcome this problem, recently IPA has been investigated in the setting of fluid queues, labeled Stochastic Flow Models (SFM) [5, 13]. Not only does it yield unbiased sample derivatives for a far-larger class of systems and performance functions than in the traditional (discrete) queueing setting, but also the algorithms for the IPA gradients tend to be model-free and simple to compute. For example, one of the earliest results on IPA in the SFM setting concerns the average loss rate due to spillover in a queue with a finite buffer, as a function of the buffer size [5, 13]. The sample derivative of this performance function over a T seconds-long time interval was shown to be $-N_T/T$, where N_T is the number of busy periods in the time-interval $[0, T]$ that incur any loss. Figure 1 shows an example of the trajectory of the buffer contents (amount of fluid) where $N_T = 2$, since there are three busy periods but only two of them have loss (indicated in the figure when the buffer is full). This formula for the sample derivative is model-free in the sense that it can be computed directly from the sample path of the system, without any explicit knowledge of the flow processes or their underlying distributions. Furthermore, the inflow-rate process need not be piecewise constant, continuous, or Markovian.

The above result stimulated a further study of IPA in the setting of general SFM networks (see [3] and references therein). Performance metrics of interest include throughput, loss, and delay, and they are viewed as functions of variable parameters such as service rates, inflow rates, buffer sizes, and various flow control parameters. Several application areas have been considered, including telecommunications [1, 8], manufacturing [3], and traffic control [12]. In all of these studies the IPA gradients were computed by model-free and easily-computable formulas and algorithms. Moreover, these algorithms typically exhibited robustness with respect to model variations, and hence provided adequate sensitivity estimates when applied, under certain circumstances, to sample paths obtained from discrete queueing systems.

Current research in IPA aims at extending its scope from SFM to other types of stochastic hybrid systems, and in this context the present paper inaugurates an investigation of hybrid Petri nets. What is peculiar about such system-models is that they capture the concepts of concurrency and synchronization, which have not yet been investigated from the standpoint of

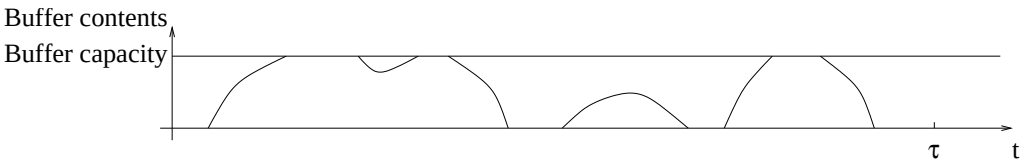

Figure 1: Buffer contents vs. time

IPA.

Hybrid Petri nets (HPN) [6] are a class of models that combine fluid and discrete event dynamics. Various formulations have been defined, and they all extend the "traditional" concept of Petri nets to include the flow of both fluid and tokens. Whereas the fluid-flow rates describe the continuous, time-driven dynamics, tokens' transitions represent the dynamics of discrete events. This paper considers an example of such a system, whose analysis will focus on the time-driven (continuous) dynamics. To this end we will view the fluid flow rates on a given arc in the network as a piecewise-continuous, random function of time (t), whose discontinuities correspond to the discrete events. In particular, we will perform sensitivity analysis for processes associated with fluid transitions, which has not yet been done in the setting of IPA. Our analysis focuses on the continuous dynamics and it does not explicitly use the discrete-event dynamics; for a comprehensive exposition of hybrid Petri nets please see [6] [2]).

2 Throughput Sensitivity in Fluid Transitions

This section concerns the throughput sensitivity of a fluid transition (continuous transition) in terms of various traffic flow parameters. A typical scenario is shown in Figure 2, where a transition, referred to as the *processing transition*, has n input places. Each input place has a fluid source that is represented via a fluid transition, called an *input transition*. Let $v_i(t)$ denote the instantaneous flow rate from the input transition i, and let $V_x(t)$ denote the maximum flow rate of the processing transition, where $t \in [0, T]$ for a given $T > 0$. Furthermore, let $m_i(t)$ denote the amount of fluid (workload) in place i, and let $v_x(t)$ denote the instantaneous flow rate from the processing transition. According to the axiomatic definition of dynamical systems [14] we can view $v_i(t)$ and $V_x(t)$ as the input to the system while $m_i(t)$, $i = 1, \ldots, n$, and $v_x(t)$ comprise its state variable. These random processes are defined on a common probability space $(\Omega, \mathcal{F}, P)$, and their realizations are assumed to be piecewise continuous; formal definitions will be stated later.

Define the set of empty places at time t via

$$I(t) \;=\; \{i = 1, \ldots, n \;:\; m_i(t) = 0\}. \tag{1}$$

Now the state variables are related to the input processes via the following two equations,

$$\dot{m}_i(t) \;=\; v_i(t) - v_x(t) \tag{2}$$

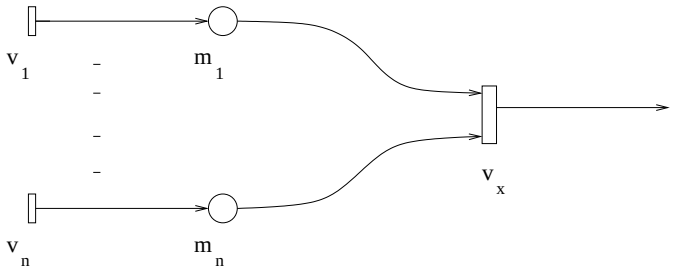

Figure 2: Fluid transition

with 'dot' denoting derivative with respect to time, and

$$v_x(t) \;=\; \begin{cases} V_x(t), & \text{if } I(t) = \emptyset \\ \min\{v_i(t), \quad i \in I(t)\}, & \text{if } I(t) \neq \emptyset. \end{cases} \tag{3}$$

A few remarks are due.

(i). Equation (3) implies that a transition "fires" (processes fluid) at its maximum rate as long as none of its input places is empty, while an empty input place would slow down the firing rate.

(ii). Because of the circular relation between the various functions in (1) – (3), these equations have to be viewed as jointly defining $I(t)$, $m_i(t)$, and $v_x(t)$. As mentioned earlier, additional, mild assumptions will be stated later to ensure the consistency of these definitions.

(iii). Equation (2) requires given initial conditions $m_i(0)$, $i = 1, \ldots, n$.

(iv). If $I(t)$ does not change its value in some open interval $(\bar{\tau}, \bar{t}) \subset [0, T]$, then $\forall t \in (\bar{\tau}, \bar{t})$, and $\forall i, j \in I(t)$, $v_i(t) = v_j(t)$. In this case,

$$v_x(t) \;=\; v_i(t), \qquad \forall t \in (\bar{\tau}, \bar{t}), \qquad \forall i \in I(t). \tag{4}$$

(v). Several papers on hybrid Petri nets impose structural assumptions on the input flow functions, like piecewise constancy (e.g., [2]). Our analysis can be carried out verbatim for the more general case of piecewise continuity.

Next, let θ be a real-valued parameter upon which the aforementioned processes depend. For example, θ can be the maximum flow rate of the processing transition, or a parameter of the probability law of any one of the input processes. To maintain generality, we assume henceforth that all of the input processes are functions of θ, and hence we denote their realizations by $v_i(\theta; t)$ and $V_x(\theta; t)$, respectively. Consequently, all of the other processes mentioned above are functions of θ as well, and we denote them by $m_i(\theta; t)$ and $v_x(\theta; t)$. This dependence on θ means the following: Fix a value of θ, and let the system evolve for $t \in [0, T]$ in the interval $[0, T]$. In other words, θ does not change while the system evolves in this way. Then the resulting processes, $\{v_i(\theta; t)\}$, $\{m_i(\theta; t)\}$, etc., satisfy Equations (1) – (3) for all $t \in [0, T]$. We assume that θ is confined to a given compact interval Θ, and we point out that generally θ can be a

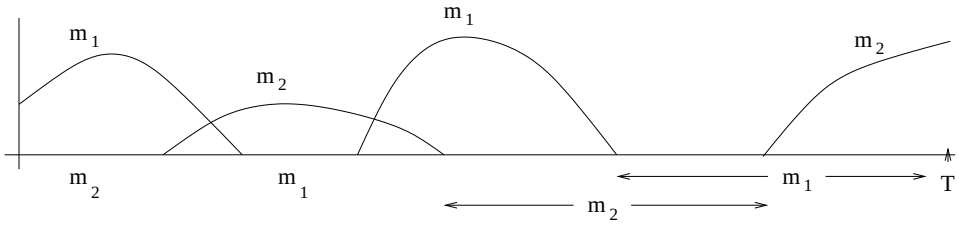

Figure 3: Fluid-level trajectories in the places

multivariable parameter as well, but we consider the scalar case in order to keep the discussion simple.

Our concern in this paper is with the average throughput over the interval $[0, T]$ as a function of θ. Denoting this function by $J(\theta)$, we define it as

$$J(\theta) \;=\; T^{-1} \int_0^T v_x(\theta; t) dt. \tag{5}$$

We next derive a formula for the IPA derivative $\frac{dJ}{d\theta}$. Fix $\theta \in \Theta$. Suppose that all of the derivative terms mentioned in the sequel exist; assumptions guaranteeing this will be made later, in the context of a specific example. We define an *empty period* of place i as a maximal positive-length subinterval of $[0, T]$ during which $m_i(t) = 0$. Suppose that there are k_i empty periods of place i in the interval $[0, T]$; we denote them by $[\tau_{i,j}, t_{i,j}]$, $j = 1, \ldots, k_i$, in increasing order. Note that we exclude from the definition of empty periods the case where $\tau_{i,j} = t_{i,j}$, i.e., empty periods consisting of a single point. A particular scenario is shown in Figure 3 for the case where $n = 2$.

We make the following assumptions for every fixed $\theta \in \Theta$.

Assumption 2.1 *With probability 1 (W.p.1) the functions $v_i(\theta; \cdot)$, $i = 1, \ldots, n$, and $V_x(\theta; \cdot)$, are piecewise Lipschitz continuous in the interval $t \in [0, T]$.*

Assumption 2.2 *All of the derivative terms that are mentioned below, exist w.p.1.*

Assumption 2.3 *For all $i = 1, \ldots, n$, $\frac{dm_i}{d\theta}(\theta; 0) = 0$.*

A few remarks are due.

Assumption 2.1 means that the input functions are piecewise continuous and Lipschitz continuous on the intervals where they are continuous. This assumption implies that all the terms defined by Equations (1) – (3) and (5) are well defined.

It is possible to make assumptions on the input processes from which the statement in Assumption 2.2 follows, but that would be abstract and cumbersome in the general case. On the other hand, such assumptions can be quite simple and intuitive when made in the context of specific examples, as will be seen in the next section.

Assumption 2.3 means that the initial conditions $m_i(\theta, 0)$ are independent of θ, certainly a mild assumption.

We have to make one more assumption about the starting times and end times of empty periods. We mentioned earlier that jumps (discontinuities) in the input processes $\{v_i(\theta;\cdot)\}$ and $\{V_x(\theta;\cdot)\}$ are due to the occurrence of certain discrete events. In addition to these, we will consider as discrete events the changeover in the state of a place from empty to non-empty and vice versa. Similarly to [5], we classify these events as follows. (i) *Exogenous events* are discontinuities in the input processes $\{v_i(\theta;\cdot)\}$ and $\{V_x(\theta;\cdot)\}$ whose timing is independent of θ. (ii). *Endogenous events* are the beginning of empty periods at the various places. We point out that these events are called macro-events in [2]. (iii) An *Induced event* is the end of an empty period at one of the places that is triggered by the start of an empty period at another place (an explanation follows). Now we will assume that all discontinuities in the input processes $\{v_i(\theta;\cdot)\}$ and $\{V_x(\theta;\cdot)\}$ are exogenous events. This assumption is justified in some situations, including the example discussed in the next section; cases where it does not hold require an extension of the analysis, as it was done for the application of IPA to SFMs [3]. Endogenous events, namely the beginning of empty periods, occur whenever one of the places is drained until it becomes empty; there is no reason to have a jump in any input process at that time, and this justifies the assumption that endogenous events and exogenous events do not co-occur. However, en endogenous event (the start of an empty period) may cause an abrupt decrease in $v_x(\theta,\cdot)$, and this can trigger the end of an empty period at another place, hence an induced event.

Recall that $t_{i,j}$ denotes the end of the jth empty period at place i.

Formally, we make the following assumption.

Assumption 2.4 *Fix $\theta \in \Theta$. W.p.1, the following holds: (i). All discontinuities in the input processes $\{v_i(\theta;\cdot)\}$ and $\{V_x(\theta;\cdot)\}$ are exogenous events. (ii) Only a single exogenous event or endogenous event can occur at a given time. (iii). If $t_{i,j}$ is not the time of an induced event, then either $\frac{dt_{i,j}}{d\theta} = 0$, or the function $v_i(\theta;\cdot) - v_x(\theta;\cdot)$ is continuous at $t = t_{i,j}$.*

Corollary 2.5 *If $t_{i,j}$ is not the time of an induced event, then either $\frac{dt_{i,j}}{d\theta} = 0$, or the function $v_i(\theta,\cdot) - v_x(\theta,\cdot)$ is continuous at $t = t_{i,j}$ and $v_i(t_{i,j}) - v_x(t_{i,j}) = 0$.*

Proof. Suppose that $t_{i,j}$ is not the time of an induced event. If the function $v_i(\theta,\cdot) - v_x(\theta,\cdot)$ has a jump at $t = t_{i,j}$ then the result is obvious from Assumption 2.4(iii). If that function is continuous at $t = t_{i,j}$ then, since $t_{i,j}$ is the end of an empty period, the function $v_i(\theta,\cdot) - v_x(\theta,\cdot)$ changes its sign there from nonpositive to non-negative, and hence $v_i(\theta, t_{i,j}) - v_x(\theta, t_{i,j}) = 0$. □

Next, define t_{max} to be the last time-point when an empty period at any place ends, namely, $t_{max} := \max\{t_{i,j} \ : \ i = 1,\ldots,n;\ j = 1,\ldots,k_i\}$. Then either $t_{max} = T$, or $t_{max} < T$. In the first case, shown in Figure 3, the final time T is contained in an empty period in at least one of the places, and hence $I(T) \neq \emptyset$ (in Figure 3, $I(T) = \{1\}$). On the other hand, in the second case,

none of the places is empty at time $t = T$, and therefore $I(T) = \emptyset$.

The following result yields the IPA derivative $\frac{dJ}{d\theta}(\theta)$.

Proposition 2.6 Fix $\theta \in \Theta$.

(1) If $t_{max} = T$, then for every $i \in I(t_{max})$,

$$\frac{dJ}{d\theta}(\theta) \;=\; T^{-1} \int_0^T \frac{\partial v_i}{\partial \theta}(\theta, t)dt. \tag{6}$$

(2) If $t_{max} < T$, then for every $i \in I(t_{max})$,

$$\frac{dJ}{d\theta}(\theta) \;=\; T^{-1}\Big(\int_0^{t_{max}} \frac{\partial v_i}{\partial \theta}(\theta, t)dt + \int_{t_{max}}^T \frac{\partial V_x}{d\theta}(\theta; t)dt\Big). \tag{7}$$

Proof. (1). Fix $i \in I(t_{max})$. Recall that, for every $t \in [0, T]$, the input flow rate to place i is $v_i(\theta; t)$, and the output flow rate from place i is $v_x(\theta; t)$; consequently,

$$\int_0^T \big(v_i(\theta; t) - v_x(\theta; t)\big)dt \;=\; m_i(\theta; T) - m_i(\theta; 0). \tag{8}$$

Since $t_{max} = T$ and $i \in I(t_{max})$, we have that $m_i(\theta; T) = 0$, and therefore, and by (8),

$$\int_0^T v_i(\theta; t)dt - \int_0^T v_x(\theta; t)dt \;=\; -m_i(\theta, 0). \tag{9}$$

By (5) and (9),

$$J(\theta) \;=\; T^{-1}\Big(\int_0^T v_i(\theta; t)dt + m_i(\theta; 0)\Big). \tag{10}$$

By Assumption 2.3, $\frac{\partial m_i}{\partial \theta}(\theta; 0) = 0$. By Assumption 2.4(i), every jump point of $v_i(\theta; \cdot)$ is the time of an exogenous event, and hence, if $\bar{t}$ is the time of such a discontinuity, then $\frac{d\bar{t}}{d\theta} = 0$. Consequently, and by (10) and Leibnitz rule, Equation (6) follows.

(2). Fix $i \in I(t_{max})$. Observe that Equation (8) is satisfied, but it is no longer true that $m_i(\theta; T) = 0$. However, place i is empty at time t_{max} and therefore $m_i(\theta; t_{max}) = 0$, and hence, similarly to Equation (9), we have that

$$\begin{aligned} &\textstyle\int_0^{t_{max}} \big(v_i(\theta; t) - v_x(\theta; t)\big)dt \\ &=\; m_i(\theta; t_{max}) - m_i(\theta; 0) = -m_i(\theta; 0). \end{aligned} \tag{11}$$

Moreover, for every $t \in [t_{max}, T]$, $v_x(\theta; t) = V_x(\theta; t)$, and therefore, and by (5) and (11),

$$\begin{aligned} J(\theta) \;&=\; T^{-1} \textstyle\int_0^T v_x(\theta; t)dt \\ &=\; T^{-1}\Big(\textstyle\int_0^{t_{max}} v_x(\theta; t)dt + \int_{t_{max}}^T v_x(\theta; t)dt\Big) \\ &=\; T^{-1}\Big(\textstyle\int_0^{t_{max}} v_i(\theta; t)dt + m_i(\theta; 0) + \int_{t_{max}}^T V_x(\theta; t)dt\Big). \end{aligned} \tag{12}$$

Taking derivatives in (12), and using Assumption 2.3 and Assumption 2.4(i) in conjunction with Leibnitz rule, we obtain that

$$\begin{aligned}\frac{dJ}{d\theta} &= T^{-1}\Big(\int_0^{t_{max}} \frac{\partial v_i}{\partial \theta}(\theta;t)dt + \int_{t_{max}}^{T} \frac{\partial V_x}{\partial \theta}(\theta;t)dt + \\ &\quad + \big(v_i(\theta;t_{max}^-) - V_x(\theta;t_{max}^+)\big)\frac{dt_{max}}{d\theta}\Big).\end{aligned} \tag{13}$$

But $V_x(\theta, t_{max}^+) = v_x(\theta; t_{max}^+)$. Moreover, t_{max} is defined as the last time any place is empty, and hence it cannot be the time of an induced event (since an induced event would be the start of another empty period, which would end at a later time). Therefore, and by Assumption 2.4(iii) and Corollary 2.5, either $\frac{dt_{max}}{d\theta} = 0$ or

$$v_i(\theta;t_{max}^-) - v_x(\theta;t_{max}^+) = v_i(\theta;t_{max}) - v_x(\theta;t_{max}) = 0;$$

in any event, $\big(v_i(\theta;t_{max}^-) - V_x(\theta;t_{max}^+)\big)\frac{dt_{max}}{d\theta} = 0$ and hence, and by (13), Equation (7) is satisfied. □

3 An application example

In this section we focus on a particular application example in the manufacturing domain. Raw material is supplied to the system at a constant rate, processed concurrently into parts by three machines, and then assembled into finished products. Whereas the assembly machine is reliable, the three processing machines can fail, and the durations of their failure modes and operational modes can have general distributions. Furthermore, we assume that machines' failure/repair schedules are time-dependent (as opposed to operation-dependent), and we denote the probability space underlying these schedules by $(\Omega, \mathcal{F}, P)$. We consider the performance metric of the system's throughput, namely the output rate from the assembly machine, and we attempt to maximize it as a function of the routing ratios of the raw material to the three processing machines.

The continuous Petri net model of the system is shown in Fig. 4. Fluid flow is generated by the transition t_0 at an assumed unitary rate, and then deposited in a buffer represented by the place p_0. The three processing machines, denoted by M_i, $i = 1, 2, 3$, are represented by the transitions t_i, $i = 1, 2, 3$, and the buffers that follow them are modeled by the places p_i, $i = 1, 2, 3$. The assembly machine is represented by the continuous (fluid) transition t_x, whose maximum rate is assumed to be unitary. Let θ_i, $i = 1, 2, 3$, represent the fraction of fluid that is routed at place p_0 to transition t_i when all of the machines are operational. Then $\theta_i \in [0, 1]$, and $\theta_1 + \theta_2 + \theta_3 = 1$. We call θ_i the *nominal routing fraction* to t_i. Let us denote by v_i the fluid flow rate through t_i. Since the fluid inflow rate from t_0 to p_0 is assumed to be 1, we have that $v_i = \theta_i$ whenever all three machines are operational. However, when some of the machines are down, $v_i \neq \theta_i$. We adopt the policy that whenever a machine is down but not all of the machines are down, the fluid that was supposed to flow to transition t_i is divided among the operational machines in

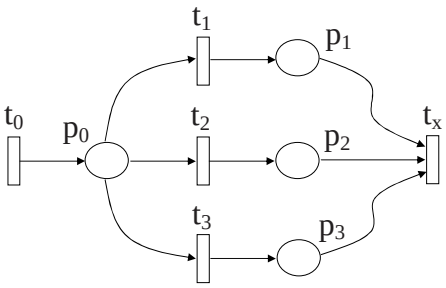

Figure 4: The Petri net system considered in Section 3

proportion to their respective nominal routing fractions. Thus, if only machine i is operational while machines j and k are down, then $v_i = 1$ while $v_j = v_k = 0$; and if machines i and j are operational while machine k is down, then $v_i = \theta_i + \dfrac{\theta_k \theta_i}{\theta_i + \theta_j}$ and $v_j = \theta_j + \dfrac{\theta_k \theta_j}{\theta_i + \theta_j}$.

Finally, if all of the machines are down then $v_i = 0$ for all $i = 1, 2, 3$. In all but that latter case, it is readily seen that $v_1 + v_2 + v_3 = 1$. Finally, let us denote by v_x the fluid flow rate through the assembly transition t_x. It is readily seen that v_x can be determined according to Equation (3), where we recall that the maximum flow rate is $V_x = 1$.

Our goal here is to determine the values of θ_1, θ_2 and θ_3 that maximize the average throughput of the system. Since θ_3 can be determined by θ_1 and θ_2 and the sample path, we can adopt the optimization parameter to be (θ_1, θ_2), which we denote by the two-dimensional vector $\theta \in R^2$. The sample performance function is

$$J(\theta) \ := \ T^{-1} \int_0^T v_x(\theta; t) dt, \tag{14}$$

and the optimization problem we consider is to maximize the expected-value function $E[J(\theta)]$, $E[\cdot]$ denoting expectation in $(\Omega, \mathcal{F}, P)$.

We now show how to solve this optimization problem using the framework described in the last section, and especially Proposition 2.6. In order to maximize $E[J(\theta)]$ we use a stochastic approximation algorithm of the Robbins-Monro type (see [11]) that computes an iteration-sequence $\theta(k) := (\theta_1(k), \theta_2(k)) \in R^2$, $k = 1, 2, \ldots$. The algorithm has the following form:

Algorithm 3.1 *Data: $\theta(1) \in R^2$ such that $\theta_i(1) \in [0, 1]$ for $i = 1, 2$, a small $\epsilon > 0$, and a positive step-size sequence $\{\lambda_k\}_{k=1}^{\infty}$ satisfying the convergence conditions for Robbins-Monro algorithms, namely $\sum_{k=1}^{\infty} \lambda_k = \infty$ and $\sum_{k=1}^{\infty} \lambda_k^2 < \infty$.*
Step 1: Set $k = 1$.
Step 2: Simulate the system for a T-second horizon, and compute the sample derivatives $\frac{\partial J}{\partial \theta_i}(\theta(k))$, $i = 1, 2$, by using Proposition 2.6.
Step 3: For $i = 1, 2$, if $\theta_i(k) + \lambda_k \frac{\partial J}{\partial \theta_i}(\theta(k)) \in (\epsilon, 1 - \epsilon)$, set

$$\theta_i(k+1) \ = \ \theta_i(k) + \lambda_k \frac{\partial J}{\partial \theta_i}(\theta(k)); \tag{15}$$

	$\partial v_1/\partial\theta_1$	$\partial v_1/\partial\theta_2$	$\partial v_2/\partial\theta_1$	$\partial v_2/\partial\theta_2$	$\partial v_3/\partial\theta_1$	$\partial v_3/\partial\theta_2$
M_1, M_2, M_3 ON	1	0	0	1	-1	-1
M_1, M_2 ON M_3 OFF	$\frac{\theta_2}{(\theta_1+\theta_2)^2}$	$-\frac{\theta_1}{(\theta_1+\theta_2)^2}$	$-\frac{\theta_2}{(\theta_1+\theta_2)^2}$	$\frac{\theta_1}{(\theta_1+\theta_2)^2}$	0	0
M_1, M_3 ON M_2 OFF	$\frac{1}{1-\theta_2}$	$\frac{\theta_1}{(1-\theta_2)^2}$	0	0	$-\frac{1}{1-\theta_2}$	$-\frac{\theta_1}{(1-\theta_2)^2}$
M_2, M_3 ON M_1 OFF	0	0	$\frac{\theta_2}{(1-\theta_1)^2}$	$\frac{1}{1-\theta_1}$	$-\frac{\theta_2}{(1-\theta_1)^2}$	$-\frac{1}{1-\theta_1}$
otherwise	0	0	0	0	0	0

Table 1: The values of $\partial v_i/\partial\theta_j$, $i = 1, 2, 3$, $j = 1, 2$, relative to the example in Section 3

and otherwise, set $\theta_i(k+1) = \theta_i(k)$.
Step 4: Set $k = k + 1$, *and go to Step 1.* ■

A few remarks are due. (i). Step 3 ensures that $\theta(k)$ remain feasible for all $k = 1, 2, \ldots$, namely $\theta(k) \in [0, 1] \times [0.1]$. (ii). The convergence of such algorithms (and many extensions thereof) to local minima have been proved in several references; see, e.g., [11]. (iii). The partial-derivative terms $\frac{\partial v_j}{\partial \theta_i}$, $j = 1, 2, 3$, $i = 1, 2$, that appear in Equations (6) and (7) depend on the enabling conditions (states) of the machines M_j, $j = 1, 2, 3$, and are summarized in Table 1.

Let us now present the results of some numerical simulations. In all considered cases the durations of the machines' states are exponentially distributed and mutually independent, with mean ON and OFF periods of machine i denoted by $\lambda_{i,f}$ and $\lambda_{i,r}$, respectively. We point out that this guarantees Assumptions 2.1, 2.3 and 2.4. Assumption 2.2 is also satisfied except, perhaps, if $\theta_i = \theta_j$ for some $i \neq j$; in which case the one-sided derivatives exist and can be used in the algorithm instead of the derivatives. The step size in Step 3 of the algorithm was chosen to be $\lambda_k = 0.5/k^{0.6}$. In all cases the simulation processed discrete parts, and the system is empty at time $t = 0$.

– **Case 1.** Let $\lambda_{1,f} = \lambda_{2,f} = \lambda_{3,f} = 0.9$ and $\lambda_{1,r} = \lambda_{2,r} = \lambda_{3,r} = 0.1$. Moreover, $\theta(1) = (0.2, 0.8)$, and $T = 30$ in equation (14). We ran the algorithm for $k_{\max} := 160$.

The values of $\theta_1(k)$ and $\theta_2(k)$ for $k = 0, 1, \ldots, k_{\max}$, are shown in Figure 5. As it can be observed both parameters converge to 1/3. Obviously, $\theta_3 = 1/3$ as well since $\theta_3 = 1 - \theta_1 - \theta_2$. This result is quite intuitive given the current values of $\lambda_{i,f}$ and $\lambda_{i,r}$. In fact, being such values equal for all the machines, their average operational times are the same, and consequently the flow will be equally shared among them.

The results of an evolution of markings m_1, m_2, m_3 and m_0 are shown in Figure 6 in the case of $\theta_1 = \theta_2 = \theta_3 = 1/3$. As it can be easily observed the marking of place p_0 is always empty apart from a very short time interval in which all the machines M_1, M_2 and M_3 are in the OFF state, as shown in Figure 7.

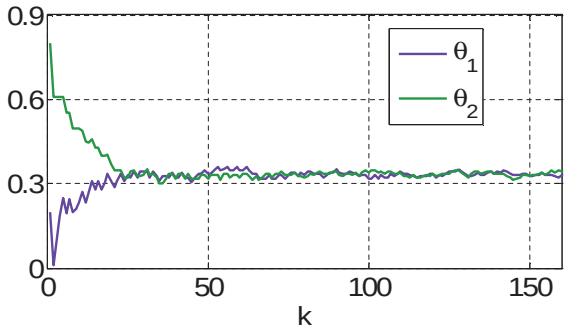

Figure 5: The values of $\theta_{1,k}$ and $\theta_{2,k}$ for $k = 0, 1, \ldots, k_{\max}$ in Case 1

– **Case 2.** Let us now present another situation whose solution can be easily interpreted by simply looking at the considered physical system. Let $\lambda_{1,f} = \lambda_{2,f} = 0.9$, $\lambda_{1,r} = \lambda_{2,r} = 0.1$, $\lambda_{3,f} = 0.1$, and $\lambda_{3,r} = 0.9$. $\theta(1) = (0.1, 0.2)$, and $T = 30$. We ran the algorithm for $k_{max} := 100$.

The values of $\theta_{1,k}$ and $\theta_{2,k}$ for $k = 0, 1, \ldots, k_{\max}$, are shown in Figure 8. As it can be observed both these two variables converge to zero. Consequently, the optimal value of $\theta_3 = 1 - \theta_1 - \theta_2$ is equal to one. This is not surprising because machine M_3 is not operational most of the time, while the other two machines are operational most of the time. Thus, when M_3 is operational (even if the other two machines are simultaneously working) for sure it is better to make M_3 work as much as possible.

The two cases discussed in the example are rather simple, and thus the results obtained are not surprising. Although the algorithm may be applied to more complex problems, whose solution is not apparent a priori, we chose to discuss only these two cases in order to show how the algorithm works and to verify that it converges to a (local) maximum point.

4 Conclusions and future work

The main contribution of this paper is that of applying IPA to Hybrid Petri nets. We focus on a particular net structure representing concurrency and synchronization. We derive the IPA gradient of the throughput with respect to fluid flow parameters at the input places, and then use this information to solve an optimal routing problem in a fork-join system.

Our future work in this framework will be devoted to extend the proposed approach to more general Petri net structures, providing systematic procedures to compute the IPA gradient of certain interesting time functions.

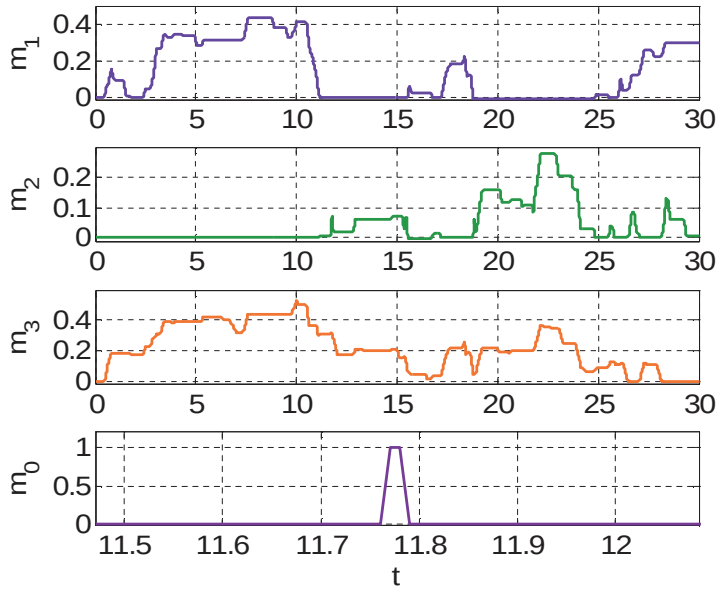

Figure 6: The evolution of m_1, m_2, m_3 and m_0 in Case 1

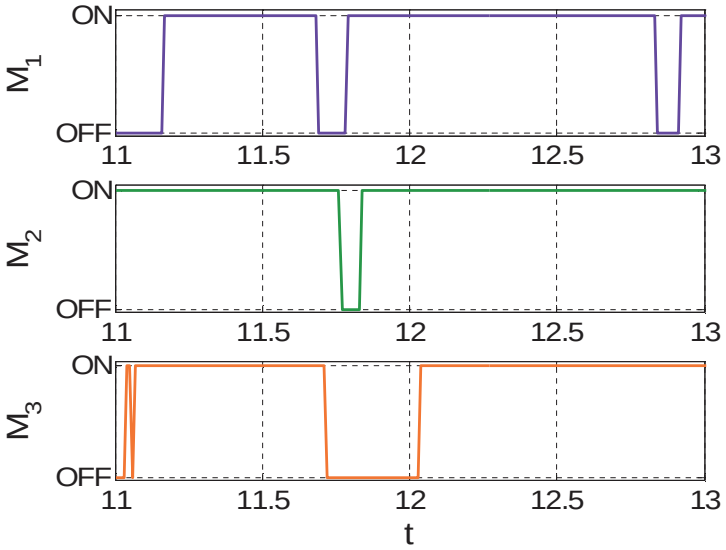

Figure 7: The discrete state of M_1, M_2 and M_3 modeled by transitions t_1, t_2 and t_3, resp., in Case 1

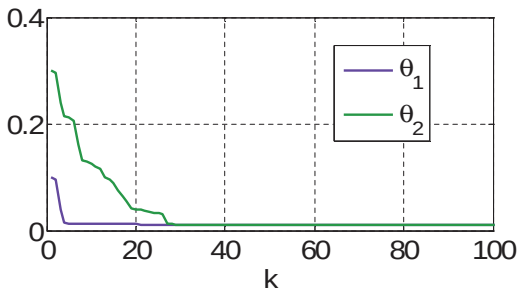

Figure 8: The values of $\theta_{1,k}$ and $\theta_{2,k}$ for $k = 0, 1, \ldots, k_{\max}$ in Case 2

References

[1] D. Anick, D. Mitra, and M.M. Sondhi. Stochastic theory of a data-handling system with multiple sources. 61:1871–1894, 1982.

[2] F. Balduzzi, A. Giua, and G. Menga. First–order hybrid Petri nets: a model for optimization and control. *IEEE Trans. Robotics and Automation*, 16:382–399, 2000.

[3] C.G. Cassandras. Stochastic flow systems: Modeling and sensitivity analysis. *In Stochastic Hybrid Systems: Recent Developments and Research Trends, Eds. C.G. Cassandras and J. Lygeros, CRC Press, New York, New York*, pages 137–165, 2006.

[4] C.G. Cassandras and S. Lafortune. *Introduction to Discrete Event Systems.* Kluwer Academic Publishers, Boston, Massachusetts, 1999.

[5] C.G. Cassandras, Y. Wardi, B. Melamed, G. Sun, and C. Panayiotou. Perturbation analysis for on-line control and optimization of stochastic fluid models. *IEEE Trans. Automatic Control*, 47:1234–1248, 2002.

[6] R. David and H. Alla. *Discrete, continous and hybrid Petri nets.* Springer, Berlin, Heidelberg, 2005.

[7] P. Heidelberger, X.R. Cao, M.A. Zazanis, and R. Suri. Convergence properties of infinitesimal perturbation analysis estimates. *Management Science*, 34:1281–1302, 1988.

[8] J. Hespanha. Stochastic hybrid nodeling of on-off tcp flows. *In Stochastic Hybrid Systems: Recent Developments and Research Trends, Eds. C.G. Cassandras and J. Lygeros, CRC Press, New York, New York*, 34:189–217, 2006.

[9] Y.C. Ho and X.R. Cao. *Perturbation Analysis of Discrete Event Dynamic Systems.* Kluwer Academic Publishers, Boston, Massachusetts, 1991.

[10] Y.C. Ho, X.R. Cao, and C.G. Cassandras. Infinitesimal and finite perturbation analysis for queueing networks. *Automatica*, 19:439–445, 1983.

[11] H.J. Kushner and D.S. Clark. *Stochastic Approximation for Constrained and Unconstrained Systems.* Springer-Verlag, Berlin, 1978.

[12] C.G. Panayiotou, W. Howell, and M.C. Fu. Online traffic light control through gradient estimation using stochastic fluid models. In *Proc. IFAC World Congress*, 2005.

[13] Y. Wardi, B. Melamed, C.G. Cassandras, and C.G. Panayiotou. Ipa gradient estimators in single-node stochastic fluid models. *Journal of Optimization Theory and Applications*, 115:369–406, 2002.

[14] L.A. Zadeh and C.A. Desoer. *Linear Systems Theory: The State Space Approach.* W. Linvill, L.A. Zadeh, and G. Dantzig, Eds., McGraw Hill Series in System Science, New York, NY, 1963.