Strong Points of Weak Convergence:A Study Using RPA Gradient Estimation for Automatic Learning ⋆

Felisa J. Vázquez-Abad [1]

Department of Computer Science and Operations Research
University of Montreal, CANADA
and
Department of Electrical and Electronic Engineering
The University of Melbourne, Parkville, Victoria 3052, AUSTRALIA
+(613) 9344-9197 FAX: +(613) 9344-9188
e-mail: fva@ee.mu.oz.au

Abstract

This paper analyzes the behavior of adaptive control schemes for automatic learning. Estimates of the sensitivities are used in a gradient-based stochastic approximation procedure, in order to drive the process along the steepest descent trajectory in search for the optimum. The learning rates are kept constant for adaptability. For such procedures, convergence can be established in a weak sense. A model problem of a flexible machine is presented, for which the control parameter is a probability vector. We propose a new sensitivity estimator, generalizing the Phantom Rare Perturbation Analysis (RPA) estimator to multi-valued decisions. From the basic properties of the estimators, we build several updating rules based on the weak convergence theory to ensure asymptotic optimality. We illustrate the predicted theoretical behavior with computer simulations. Finally, we present the comparison between the behavior of our proposed scheme with a regenerative one for which we can establish strong convergence. Our results show that weak convergence yields a dramatic improvement in the rate of convergence, in addition to the capability of adaptation, or tracking.

Key words: Stochastic approximation; Gradient Estimation; Rare perturbation Analysis; Discrete event dynamic systems; Adaptive Control

---

⋆ Expanded version of a talk presented at *IFAC-CIS Conference*, Belfort, May 1997.
[1] Supported in part by NSERC-Canada grants # WFA0139015 and FCAR-Québec grant # 93-ER-1654. This research was completed while the author was on leave at the University of Melbourne.

1 Introduction

The purpose of this paper is to illustrate the typical behavior of learning algorithms using stochastic approximations (SA). In particular, we compare the behavior of SA using constant gains with that using decreasing gains. It is well known, as shall be briefly reviewed, that gradient-based SA can be used as a steepest descent algorithm in search for local minima. When gradient based SA is applied to the control parameter of a system to be optimized, the result of the procedure is a (stochastic) adaptive control process. Under some conditions, SA with decreasing gain parameters converge strongly to local optima. The ODE method, introduced by [22] and further developped in [15] was used to show a.s. convergence by [2] and [24] among others. It can also be used to establish weak convergence of SA algorithms with constant gain [16], [18], [35]. In addition, the weak convergence method establishes verifiable conditions that characterize the ensuing behavior of the control processes [32]. We show in a simple example what each of the different assumptions and properties mean and discuss the implications when implementing the method in practice. While strong convergence results are naturally preferable and more reassuring in practice, it is impossible to show strong convergence for a learning algorithm that uses constant gains. If we settle for weak convergence, however, we may be able to track perturbations and adapt to changes in external conditions.

1.1 General Background

This paper presents explicit implementations of gradient estimation into SA, as well as different updating rules, all of which yield the same asymptotic result for SA under decreasing gains. Our focus is to study the short term behavior of the different implementations. In particular, we illustrate weak convergence of the control processes to an ODE by showing sample trajectories and comparing them with the corresponding ones under decreasing gains, which converge a.s. We also discuss the concept of consistency in the average, a condition of the gradient estimators that is used to show weak convergence with the ODE method. A subsection is devoted to illustrate time scales, a theoretical device used in the proofs that can also be used to accelerate convergence by choosing the right implementation. Finally, we present the typical behavior of our learning algorithms for tracking perturbations. We believe that the unique strength of the weak convergence approach lies precisely in that the assumptions and requirements of the algorithms are stated in verifiable terms and the result identifies the asymptotic behavior in terms of the properties of the estimators and the updating rules. Understanding how these conditions affect the tendency of the procedure helps to predict the behaviour of slightly different implementations in terms of asymptotic (sub-)optimality.

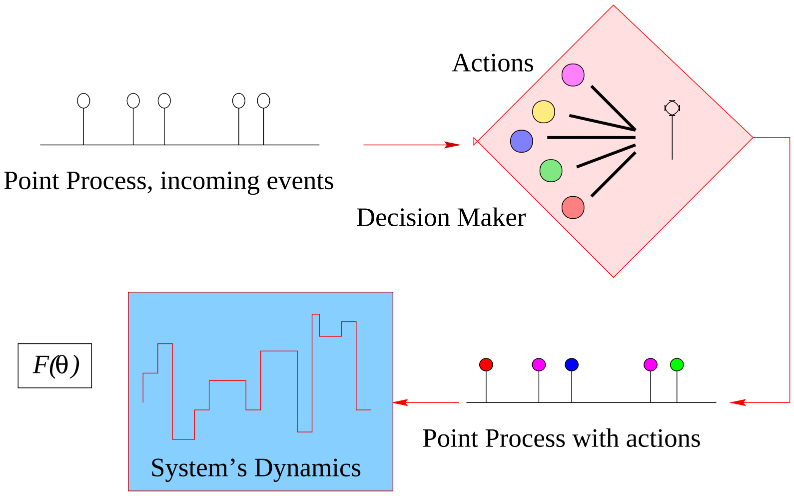

Fig. 1. Controlled DES

Consider a model where $\theta \in \Re^d$ is a control parameter. In Figure 1 we depict such a system, where the control parameter is a probability vector over a set of possible actions of a discrete event system. Suppose that a performance function $F(\theta)$ has been defined, but that there is no closed form expression for it. Instead, is it possible to measure data directly from the physical system, if it evolves in time, or from controlled experiments. Examples of this problem are numerous in applications, where the system is subject to uncertainty. A stochastic approximation procedure is a recursive algorithm of the form:

$$\theta_{n+1} = \theta_n + \epsilon_n Y_n \tag{1}$$

that in some way adjusts the values of the control variable θ according to the observations Y_n that–we hope–reflect a measure of the sensitivity of the performance as a function of θ, as illustrated in Figure 2. The work by Robbins and Monro [28] in the 1950's began a fruitful period for the development of control of stochastic systems. They established that such algorithms behave very much like their deterministic counterpart. Suppose that the gradient $\nabla_\theta F(\theta)$ is well defined and that the function $F(\cdot)$ is convex, so that there is a unique maximum θ^*. If a sequence of random variables $\{Y_n\}$ can be produced, with $E\{Y_n\} = \nabla_\theta F(\theta_n)$ and uniformly bounded variance, then $\theta_n \to \theta^*$, provided that gain parameters satisfy:

$$\epsilon_n > 0, \quad \sum_n \epsilon_n \to \infty, \quad \sum_n \epsilon_n^2 < \infty \tag{2}$$

The condition $\sum \epsilon_n^2 < \infty$ can be weakened (see [20], Section 5.3.2).

Intuitively, if direct measurements of $\nabla_\theta F(\theta)$ corrupted by noise are available, then (1) follows (in the average) the direction of increase of the performance function, while the changes in θ_n become smaller as the optimum is approcahed. Until recently, as we shall mention later on, there were no available methods for measuring derivatives directly from observations of a stochastic process. Kiefer and Wolfowitz [13] proposed the following approach. Use a sequence $c_n \to 0$ to evaluate

$$Y_n = \frac{f(\theta_n + c_n \hat{e}_k) - f(\theta_n - c_n \hat{e}_k)}{2c_n}$$

where $\hat{e}_k$ is a vector with the k-th component set to unity and the remaining to zero, and $f(\theta)$ is a sample or pathwise estimate of the function $F(\theta)$ such that $E[f(\theta)] = F(\theta)$. Their estimators are strongly consistent for the partial derivative $F_k'(\theta)$ w.r.t. θ, in the following sense. If the algorithm does not perform the updates, but rather only produce the measurements $Y_n(\theta)$, then $\lim_{n\to\infty} E[Y_n(\theta)] = F_k'(\theta)$ a.s. It is generally the case that good estimation is achieved only as more samples are used in the construction of each estimate Y_n. This is commonly performed as follows. One simulates or observes the system for a period of time T_n during which the control parameter is kept fixed at value θ_n. At the end of the estimation interval, the parameter θ is updated using (1). It often is necessary that $T_n \to \infty$ in order to achieve strong consistency of the estimators. Under these conditions, the procedure converges strongly to θ^*. In [15] the basic Robbins-Monro procedure is generalized, considering problems subject to constraints and using a truncated version of (1).

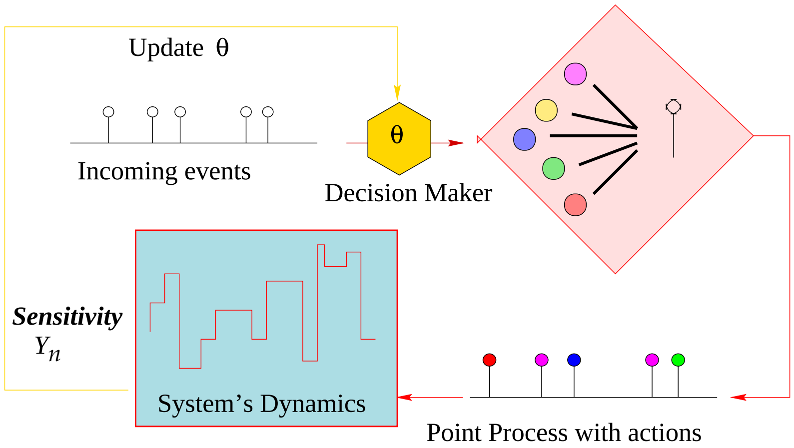

Fig. 2. Controlled DES with Feedback

In the context of stochastic optimization, much progress has been achieved in the past decade, with the increasing interest for constructing single path gradient estimators. Such are the Likelihood Ratio (LR) Method of [27] and [9], the Perturbation Analysis Method (IPA and SPA) of [10], [8] and, more recently, the Rare Perturbation Analysis (RPA) of [4], [30], the Harmonic Gradient (HG) of [12], the Weak Derivative Methods of [26] and the Simultaneous Perturbations (SPSA) Method of [29], some of which have standard implementations using one system trajectory (possibly perturbed) with or without off-line simulations, thus amenable to single path gradient estimation. While many of these gradient estimators need observations of the system as well as detailed information about the system dynamics, some of them can be applied to specific problems in a very robust manner. Notably, SPSA and HG can be used without assuming knowledge of the system's dynamics, but they typically require perturbing the system, which may be undesireable. For specific problems such as routing or threshold control problems, RPA is very robust. In our examples we shall discuss one implementation of RPA that requires only observation of the states along a single unperturbed trajectory.

Most of the work has focused on showing strong convergence for *simulation optimization* of stochastic DES, that is, $\theta_n \to \theta^*$ a.s., as in [1], [5], [6], [7], [11], [21] and [26], among others. Their method of proof refers to the general ODE method in [15], [24], [20], typically under the requirement (2).

Applications of automatic learning in telecommunications, intelligent transportation, flexible manufacturing systems (FMS), surveillance policies and dynamic resource allocation require algorithms capable of performing *adaptive control* under possibly changing environments. The implementation of a learning scheme differs from the common procedure for simulation optimization in that the gain parameters do not decrease, but are kept constant, using:

$$\theta^\epsilon_{n+1} = \theta^\epsilon_n + \epsilon Y^\epsilon_n \tag{3}$$

The requirement that the gain parameter (or "learning rate") does not decrease is essential for the algorithm to be able to adjust and track the optimal control when the underlying processes vary their behavior. While this is desirable for on-line optimization, strong convergence to the optimum cannot be achieved. Such would be the case, for example, when the call patterns in a telecommunications network change, machines fail or deteriorate their processing capabilities in a FMS, etc. An alternative approach has been proposed (see [16], [19], [17], [18], [32], [35]) in which convergence is established in a weak sense. Further, [25] has studied the stationary solution of such control processes.

This paper studies the behavior of (3) when estimates of the sensitivities are used in a gradient-based stochastic approximation procedure. The weak convergence method establishes the conditions under which (3) drives the process along the steepest descent trajectory in search for the optimum. The interpretation of the results of the method is intuitively appealing, but differs considerably from statements such as $\theta_n \to \theta^*$ a.s. Instead of studying a sequence of random variables, the method focuses on the stochastic process defined by the time-varying control parameter. The notions of convergence of a process are, naturally, related to the topology introduced in the appropriate functional space. Roughly speaking, the method establishes the conditions under which the random trajectories approach, as $\epsilon \to 0$, the deterministic trajectories of a companion ODE, the asymptotes of which are its stable points. Therefore, establishing asymptotic optimality of the learning scheme requires identifying such stability points as the optimal control values of the original problem.

1.2 Model Example

Our model example is the following. A single server queue has $V \geq 2$ possible settings or "speeds" under which the service distributions differ. The control parameter θ is the probability vector determining the fraction of customers served at each setting. The performance measure considers the minimization of the stationary average waiting time as well as the operation costs. Ours is an example of the canonical model in [5], [7] and [18], who implement IPA derivative estimators for the sensitivity to the mean service time. In our model, however, the control variable is not scalar, the constraint set requires θ to satisfy the law of total probability, and IPA is not applicable.

1.3 Summary

The aim of this paper is to illustrate how the weak convergence methodology can be used in practice in order to construct learning algorithms, particularly when the control is a probability vector. In so doing, we extend the notion of sensitivity estimation in terms of generalized gradients that take into account the Karush-Kuhn-Tucker (KKT) conditions for optimality. Although truncated algorithms are shown to converge in the limit ($\epsilon \to 0$ for (3) or $n \to \infty$ for (1)), the actual implementations are constructed with positive gain parameters. In our experience, truncations may yield bad behavior when one or more components of θ reach zero, which may be absorbing points of the procedure, even if far from the actual optimum. To solve this practical problem, we introduce the generalized gradients, which makes truncations of (3) unnecessary. In addition, we construct a new estimator for such sensitivities, generalizing the Phantom RPA method to the case of multi-valued decisions and discuss two implementations of our estimators. The predicted theoretical behavior of the updating schemes is illustrated via computer simulations using the regenerative and non-reset versions of the estimators. Finally, the comparison between the behavior of our proposed scheme with a regenerative one for which we can establish strong convergence is discussed.

Section 2 presents a brief review of the weak convergence method, emphasizing the ODE method. The main assumptions on the model and sensitivity estimators are stated in the context of optimization of stochastic DES and the main results for strong and weak convergence are summarized. Section 3 introduces the model example. Section 4 deals with the construction of the "target" ODE that governs the limiting behavior of the learning scheme. Even if the performance measure is unknown in analytical form, the target ODE is constructed such that its stable points are local minima. Section 5 develops the gradient estimators and Section 6 contains the description of the ensuing learning algortihms. Section 7 presents the empirical study of the behavior of the algorithms (3) and (1) under different implementations.

2 Weak Convergence Review

This section briefly presents the basic concepts and notation of the Weak Convergence Theory for stochastic approximation. For details on the model and methods of proof, we refer to [16], [18], [32], [19] and strongly recommend the authoritative reference [20].

2.1 Assumptions

Consider a Discrete Event Driven System (DES) under control. Let $\theta \in \Re^d$ be the control variable. The required assumptions for strong and weak convergence of the SA are now stated.

Fixed θ Process:

Consider a stochastic process $\{\tilde{X}_t(\theta), t \geq 0\}$ whose evolution is determined by a sequence of "events", happening at random times $\{T_i, i \geq 0\}$. Call $\{\mathfrak{F}(t), t \geq 0\}$ the filtration of the process, with $\mathfrak{F}(t)$ the σ-algebra generated by the process up to time t. Assume that T_i is a stopping time (measurable with respect to the history of the process), so that $\{T_i \geq t\} \in \mathfrak{F}(t)$. Calling $X_i(\theta)$ the state of the embedded process plus the residual clocks at time T_i, $\{X_i(\theta)\}$ has a Markovian structure and is generally known as a Generalized Semi-Markov Process (see [8]). Use $\mathfrak{F}_i(\theta)$ to denote the σ-algebra generated by $\{X_j(\theta); j \leq i\}$. Assume that there exists a compact set Θ, such that for any $\theta \in \Theta$, the process $\{X_i(\theta)\}$ regenerates and has a unique, ergodic invariant measure $\tilde{\mu}_\theta(dx)$. The performance measure of interest is of the form:

$$F(\theta) = \int f(x)\tilde{\mu}_\theta(dx) = \lim_{M\to\infty} \frac{1}{M}\sum_{i=1}^{M} f[X_i(\theta)] \tag{4}$$

where $f(x)$ is the "instantaneous" cost at state x, supposed to be continuously differentiable. We shall also assume that F is continuously differentiable. Before proceeding, we introduce two useful definitions.

Definition 1 *A sequence of probability measures $\{\mu_u(\cdot); u \in U\}$ defined on a common, complete, separable metric state space S is said to be* tight *if for every δ there exists a compact set $K_\delta \subset S$ such that $\sup_{u\in U}\mu_u(K_\delta) > 1-\delta$. Accordingly, we say that a sequence of random variables $\{X_i, i = 1, 2\ldots\}$ defined on a common, complete, separable metric state space S is* tight *if for every δ there exists a compact set $K_\delta \subset S$ such that $P\{X_i \notin K_\delta\} \leq \delta$ for all $i = 1, 2, \ldots$*

Tightness is the equivalent of sequential compactness in the sense that if a set of measures is tight, then every sequence has a further weakly convergent subsequence, and the limit is a well defined probability measure (see [3]).

Definition 2 *We say that a sequence of random variables $\{Y_n, n = 0, 1, \ldots\}$*

defined on a common, complete, separable state space S is uniformly integrable *if for every $\alpha > 0$ there exists a compact set K_α such that:*

$$\sup_n \int_{y \notin K_\alpha} y \, P\{Y_n \in dy\} \le \alpha$$

Uniform integrability follows from the weaker condition that $\mathrm{Var}(Y_n)$ be uniformly bounded in n (see [3]).

A *pathwise derivative estimator* is an estimator constructed from the observations or measurements of the system over finite length time intervals, that we shall call the *estimation intervals* of length denoted by Δ_m. Consider the model for a fixed value of the control at θ, where no updates take place, but where a pathwise derivative estimator is computed using the observations contained in the n-th estimation interval, that is:

$$\left\{ i \colon \sum_{m=0}^{n} \Delta_m(\theta) < i \le \sum_{m=0}^{n+1} \Delta_m(\theta) \right\}, \quad \text{where } \bar{\Delta}_n(\theta) = \sum_{m=0}^{n} \Delta_m(\theta)$$

thus obtaining a sequence of estimates. A detailed analysis of weak convergence for general random estimation intervals in [32]. The present work focuses only on the commonly used intervals, namely when $\Delta_m(\theta)$ a constant number M of events of the process $\{X_i(\theta)\}$, or a constant number M of regenerative cycles of the process $\{X_i(\theta)\}$. Thus, both the estimation interval as well as the estimators are functions of the past state values. In particular, the derivative estimator–denoted $Y_n(M,\theta)$, is typically constructed via sample averages of quantities that depend on previous state values. In other words, there exists a $\mathfrak{F}_i(\theta)$-measurable process $Z_i(\theta)$ that accounts for the necessary bookkeeping, that is, $Y_n(M,\theta) = \Pi(Z_i)\mathbf{1}_{\{i=\bar{\Delta}_m(\theta)\}}$ so that the actual estimator is a component (Π denotes here a projection) of the process Z_i at the end of the estimation intervals. In Section 6 and the Apendix, such a process is built in detail for a model example. Call $\xi_i(\theta) = (X_i(\theta), Z_i(\theta))$ the enlarged state space that describes the original process and the derivative estimation process. Notice that the σ-algebra generated by $\{\xi_j(\theta); j \le i\}$ is $\mathfrak{F}_i(\theta)$. Then the event

$\{\Delta_n(\theta) > i\} \subset \mathfrak{F}_i(\theta)$.

Assumption 1 *Consider the fixed-θ process $\{\xi_i(\theta)\}$, for $\theta \in \Theta$. We assume that*

$$P_\theta(\xi, B) = P\{\xi_{i+1}(\theta) \in B | \xi_i(\theta) = \xi\}$$

is weakly continuous in (ξ, θ) and that $\{\xi_i(\theta)\}$ possesses a unique invariant, ergodic measure $\mu_\theta(d\xi)$. In addition, we assume that the set of measures $\{\mu_\theta(\cdot); \theta \in A\}$ is tight for every compact set $A \subset \Theta$.

Recall that $\{X_i(\theta)\}$ has a unique, ergodic invariant measure, by assumption. If $\{Z_i(\theta)\}$ are tight, then every sequence has a further weakly convergent subsequence. Since $Z_i(\theta)$ is a function of the states $\{X_j(\theta); j \leq i\}$ and these possess a unique limiting measure, it follows that the limiting measure of the process $\xi_i(\theta)$ is unique and uniquely determined by $\tilde{\mu}_\theta(\cdot)$.

Definition 3 *Let $Y_n(M, \theta)$ be an estimator over the n-th estimation interval.*

We say that $Y_n(M, \theta)$ is a strongly consistent estimator *of $G(\theta)$ if for any n:*

$$\lim_{M \to \infty} Y_n(M, \theta) = G(\theta) \quad a.s.$$

We say that $Y_n(M, \theta)$ is consistent in the average sense *for $G(\theta)$ if for any M:*

$$\lim_{m \to \infty} E\left\{\frac{1}{m} \sum_{n=0}^{m-1} Y_n(M, \theta)\right\} = G(\theta)$$

It is called strongly consistent in the average sense *if the average inside the brackets above converges a.s.*

Varying θ Process:

We now proceed to the description of the model when (3) is used to update the values of θ at the end of the estimation intervals. When the control variable changes this way, the process ξ_i^ϵ evolves according to the dynamics of the fixed control process at value θ_n^ϵ over the n-th estimation interval of corresponding length Δ_n^ϵ, at the end of which the estimator $Y_n^\epsilon(M)$ is obtained and θ_n^ϵ is updated according to (3). Let $\bar{\Delta}_n^\epsilon = \sum_{m=0}^n \Delta_m^\epsilon$, we denote $\tilde{\theta}_i^\epsilon = \theta_n^\epsilon$ for $i = \bar{\Delta}_n^\epsilon + 1, \ldots, \bar{\Delta}_{n+1}^\epsilon$. This variable keeps the actual value of the control parameter as the process evolves. Denote now by $\mathfrak{F}_i^\epsilon$ the σ-algebra generated by $\{\xi_j^\epsilon, \tilde{\theta}_j^\epsilon; j \leq i\}$. From the Markovian structure, it follows that $P\{\xi_{i+1}^\epsilon \in B | \mathfrak{F}_i^\epsilon\} = P\{\xi_{i+1}^\epsilon \in B | \xi_i^\epsilon, \tilde{\theta}_i^\epsilon\}$ and therefore the process $\{(\xi_i^\epsilon, \tilde{\theta}_i^\epsilon)\}$ is a Markov Decision Process (MDP) with general state space.

Assumption 2 *The random variables $\{Y_n^\epsilon(M)\}$ and $\{\Delta_n^\epsilon\}$ are uniformly integrable.*

Assumption 3 *The sequence $\{(\xi_i^\epsilon, \tilde{\theta}_i^\epsilon); i = 1, 2, \ldots, \epsilon > 0\}$ is tight.*

A more general description can be found in [32], where a more complex system with asynchronous controllers is studied. In that reference, the updating intervals in a decentralized operation do not necessarily coincide with the estimation intervals of the processors that carry out the estimations.

2.2 The Main Results

The main results of strong convergence have been treated in detail in [7] and [5], [21] when IPA derivative estimators are used. The following version of the result considers the most commonly used assumptions on the estimation. We shall apply this result in Section 7.

Theorem 1 *Suppose that $\{Y_n(\theta)\}$ are independent and*

$$E[Y_n(\theta) | \theta_n = \theta] = \nabla_\theta F(\theta) + \beta_n(\theta),$$

that $\sup_\Theta E|Y_n(\theta)|^2 < \infty$, and that $F(\cdot)$ is locally convex and has a unique maximum $\theta^* \in \Theta$. Construct Y_n as an estimator using the information available within the n-th estimation interval, where the control variable is kept at the value θ_n. Update this value at the end of the estimation interval using (1).

In order to ensure that $\theta_n \in \Theta$ with probability 1, use a truncation if necessary. Assume that (2) is satisfied and that

$$\sum_n \epsilon_n |\beta_n| < \infty \quad a.s.$$

Then $\theta_n \to \theta^*$ with probability 1.

This model is known as the "martingale difference noise" model for Y_n [20]. In many cases, when the estimators are strongly consistent, their variance decreases as $1/M$. A common approach is to use increasing update intervals $M(n) \approx n$ and $\epsilon_n \approx 1/n$.

As mentioned before, weak convergence is established in the sense of convergence in distribution of the stochastic process defined by the time-varying control parameters. Following [32], introduce the following control processes:

Definition 4 *The* ladder interpolation process $\vartheta^\epsilon(t)$ *is defined by:*

$$\vartheta^\epsilon(t) = \theta_n^\epsilon \quad t \in [n\epsilon, (n+1)\epsilon)$$

and the natural interpolation process *by:*

$$\tilde{\vartheta}^\epsilon(t) = \tilde{\theta}_i^\epsilon \quad t \in [i\epsilon, (i+1)\epsilon)$$

As we shall see, these processes reflect the common descriptions of learning algorithms, related to the iteration number (that is, in terms of the updates performed) and the actual time (in terms of the event scale of the process). It is customary to present results of convergence as a function of the number of iterations performed. This would be related to the behavior of $\vartheta^\epsilon(\cdot)$. If an updating scheme follows a number M of regenerative cycles to construct the estimators, then the length of the updating intervals depends on the control values themselves and it may be more realistic to study the behavior of $\tilde{\vartheta}^\epsilon(\cdot)$. Naturally, in the case of (1), where $\epsilon_n \to 0$ and $M(n) \to \infty$, the natural interpolation process would stretch the time scale of the corresponding ladder interpolation process. In applications of on-line control, one is interested in the behavior and capability of tracking in real time rather than iteration number. Section 7 illustrates the various time scales of interest with computer simulations.

Assumption 4 *The functions:*

$$G(\theta) = \lim_{m\to\infty} \frac{1}{m} \sum_{n=0}^{m-1} E[Y_n(M,\theta)]$$
$$\Delta(\theta) = \lim_{m\to\infty} \frac{1}{m} \sum_{n=0}^{m-1} E[\Delta_n(\theta)]$$

are continuous and $\sup_{\theta\in\Theta} \Delta(\theta) > 0$.

The following result summarizes the weak convergence approach and follows from the general methodology of [18], as well as from Theorem 4.1 in [20], and Corollary 2 of Theorem 1 in [32].

Theorem 2 *Under Assumptions 1 to 4, if* Y_n^ϵ *is the estimator calculated within the interval* $\bar{\Delta}_n^\epsilon < i \leq \bar{\Delta}_{n+1}^\epsilon$ *and (3) is used to update* θ_n^ϵ *(using a truncation to ensure* $\theta_n^\epsilon \in \Theta$ *if necessary), then* $\vartheta^\epsilon(\cdot)$ *converges in distribution to the deterministic solution of the ODE:*

$$\frac{d}{dt}\vartheta(t) = G[\vartheta(t)] \tag{5}$$

and $\tilde{\vartheta}^\epsilon(\cdot) \Rightarrow \tilde{\vartheta}(\cdot)$*, where* $\tilde{\vartheta}[\delta(t)] = \vartheta(t)$ *and* $\delta(t)$ *satisfies the ODE:*

$$\frac{d[\delta(t)]}{dt} = \Delta[\vartheta(t)].$$

If, in addition, θ^* *is the only stable point of (5), then* $\lim_{t\to\infty} \vartheta(t) = \lim_{t\to\infty} \tilde{\vartheta}(t) = \theta^*$.

In general, $\Delta(\theta) > 1$, for example, $\Delta(\theta) = M$ or M times the average number of customers in a cycle, in the case of finite horizon or regenerative estimation, respectively. That is, the time scale of $\tilde{\vartheta}(\cdot)$ is stretched, making its evolution slower. Recent research has considered the long term behavior of such SA procedures, analyzing the limiting stationary control process [25].

3 A Flexible Machine

Consider a system in which a machine processes items at different speeds $v_k; k = 1, \ldots, V$. Items arrive at the machine following a renewal process $N(t)$, illustrated in Figure 3. Let $\eta_i(\theta)$ be the speed chosen by the machine for the i-th item processed. The service times of consecutive customers $\{S_i\}$ are independent random variables with $\mu_k^{-1} = E[S_i|\eta_i(\theta) = v_k]$, and $\varsigma_k^2 = E[S_i^2|\eta_i(\theta) = v_k]$. The associated operating cost per unit time is c_k. Assume $\mu_k < \mu_j, c_k < c_j$ if $k < j$, thus faster modes of operation are more costly. Although it may be known which speeds are faster in the average, in practice the distribution of the consecutive service times may be unknown: only measurements are available. The problem is to select the speeds of operation of the machines that minimize the waiting time at the lowest cost. We study here the randomized strategy, in which $P\{\eta_i(\theta) = v_k\} = \theta_k$.

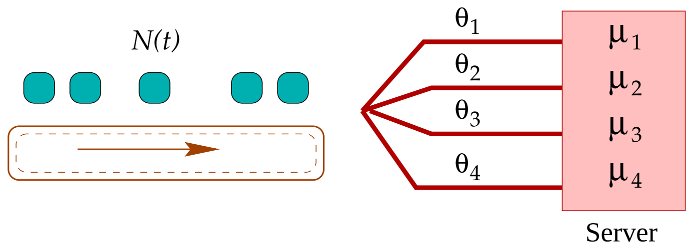

Fig. 3. The Queueing Model

Assume that $\lambda < \mu_V$, ensuring stability of the process for all $\theta \in \Theta$, where

$$\Theta = \left\{ \theta = (\theta_1, \ldots, \theta_V) \in [0,1]^V : \sum_{k=1}^{V} \theta_k = 1 \right\}.$$

Therefore, the limiting probabilities exist, the process is ergodic, and the invariant measure is unique for each θ. The performance function to be minimized is $F(\theta) = W(\theta) + C(\theta)$, where $W(\theta)$ is the stationary average waiting time in the system and $C(\theta)$ is the mean stationary cost rate resulting from the use of the machine. To find the expression for $C(\theta) = \lim_{t\to\infty} \mathcal{C}(t)/t$, where $\mathcal{C}(t)$ is the cumulative operation cost up to time t, define $D(t)$ as the departure process, τ_N the time of the N-th service completion, and $T_k(N)$ as the total time the machine operates at speed v_k, having processed N items. Conditioning on the speed chosen at each of the N items, at time τ_N, one has:

$$E[\mathcal{C}(\tau_N)] = \sum_{k=1}^{V} c_k E[T_k(N)] = N \sum_{k=1}^{V} \frac{c_k \theta_k}{\mu_k}$$

Under stability of the process, the arrival rate is the same as the departure rate, so $\lim_{N\to\infty}(N/\tau_N) = \lim_{t\to\infty}(D(t)/t) = \lambda$ a.s., yielding:

$$C(\theta) = \lim_{t\to\infty} \frac{1}{t} E[D(t)] \sum_{k=1}^{V} \frac{c_k \theta_k}{\mu_k} = \lambda \sum_{k=1}^{V} \frac{c_k \theta_k}{\mu_k} \tag{6}$$

From the structure of the model, $F(\cdot)$ is continuously differentiable. This problem may be stated as a minimization problem of the function $F(\theta)$ under the feasibility constraints:

$$\begin{aligned} &\text{Minimize} && F(\theta) = W(\theta) + C(\theta) \\ &\text{subject to} && h(\theta) = \sum_{i=1}^{V} \theta_i = 1 \end{aligned}$$

4 Limiting ODE

This section builds the "target" ODE and shows that its stable points are local minima. The arguments are general for the optimization of performance functions under randomized multi-valued decisions. Define $F_k'(\theta) = \partial F(\theta)/\partial\theta_k$. Letting θ^* be the vector of optimal values, the theory of Karush-Kuhn-Tucker (KKT) ensures the existence of a real number u such that:

1) For each $k = 1, \ldots, V$
$$\begin{cases} F_k'(\theta^*) - u h_k'(\theta^*) = 0 & \text{if } \theta_k^* > 0 \\ F_k'(\theta^*) - u h_k'(\theta^*) > 0 & \text{if } \theta_k^* = 0 \end{cases}$$
2) $h(\theta^*) = 1$
3) $\theta_k^* \geq 0 \quad \text{for } k = 1, .., V$

Since $h_k'(\theta^*) = 1$, the first condition tells us that either $\theta_k^* = 0$ or $F_k'(\theta^*) = u$, so that $u = \sum_{k=1}^{V} \theta_k^* F_k'(\theta^*)$. Call $\mathcal{G}_k$ the *generalized gradient operator* defined by:

$$\mathcal{G}_k[F(\cdot)] = \frac{\partial F(\cdot)}{\partial \theta_k} - \sum_{j=1}^{V} \theta_j \frac{\partial F(\cdot)}{\partial \theta_j} \tag{7}$$

Then, a value of θ^* such that $\sum_{k=1}^{V} \theta_k^* = 1$ and

$$\mathcal{G}_k[F(\theta^*)] \begin{cases} = 0 & \text{if } \theta_k^* > 0 \\ \geq 0 & \text{if } \theta_k^* = 0 \end{cases} \text{ for } k = 1, .., V \tag{8}$$

satisfies the KKT conditions. Under convexity of $F(\theta)$, if $\bar{\theta}$ satisfies (8), then $\bar{\theta} = \theta^*$. Consider now the following target ODE:

$$\frac{d\vartheta_k(t)}{dt} = -\vartheta_k(t)\mathcal{G}_k[F(\vartheta(t))] \tag{9}$$

and notice that no truncation is necessary: if $\vartheta(0) \in \Theta$, then for any t, $\vartheta(t) \in \Theta$, which follows adding (9) over k and using the definition (7) of the generalized gradient, together with $\vartheta(t) > 0$. That is, if $\sum_{k=1}^{V} \vartheta_k(t) = 1$ for some t, then adding (9) with respect to k, gives:

$$\frac{d}{dt}\sum_{k=1}^{V} \vartheta_k(t) = -\sum_{k=1}^{V} \vartheta_k(t) \frac{\partial F(\vartheta(t))}{\partial \theta_k} + \left(\sum_{k=1}^{V} \vartheta_k(t)\right)\left(\sum_{j=1}^{V} \vartheta_j(t) \frac{\partial F(\vartheta(t))}{\partial \theta_j}\right) = 0$$

thus showing that $\sum_{k=1}^{V} \vartheta_k(t)$ remains constant over time (and equal to unity). On the other hand, notice that from continuity of the RHS of (9), if $\vartheta_k(0) > 0$ then $\vartheta(t)$ can never become negative, since the RHS has a factor proportional to $\vartheta_k(t)$, which means that $\theta_k = 0$ for some k can only be an asymptote of the ODE.

Lemma 3 *If the starting point $\vartheta(0) \in \Theta$ is such that $\vartheta_k(0) > 0$ for each $k = 1, \ldots, V$ and it is not a local maximum of $F(\cdot)$, then the stable points of (9) are local minima. If the function $F(\cdot)$ has a unique minimum, then (9) has an asymptote at $\theta^* = \min_{\theta \in \Theta} F(\theta)$.*

PROOF. We shall verify two properties, namely, that the cost is non-increasing along the trajectory of the ODE, and that the stable points of the ODE are KKT points. The first property is easily verified since

$$\frac{d}{dt}F[\vartheta(t)] = \sum_{k=1}^{V} F'_k[\vartheta(t)]\frac{d\vartheta_k(t)}{dt} = -\left[\sum_{k=1}^{V}\vartheta_k(t)\left(D_k(t) - \bar{D}(t)\right)^2\right] \leq 0 \quad (10)$$

where we have used the fact that (10) is the negative of a variance, with $D_k(t) = F'_k[\vartheta(t)]$ and $\bar{D}(t) = \sum_{k=1}^{V}\vartheta_k(t)D_k(t)$. Since F is bounded below by zero (costs are non-negative) and is non-increasing along the trajectory, then $\lim_{t\to\infty}\frac{d}{dt}F[\vartheta(t)] = 0$ and $F[\vartheta(t)]$ converges to a value F^*, possibly depending upon the initial value $\vartheta(0)$.

The second property stems from the fact that, from (10), if $\bar{\theta} = \lim_{t\to\infty}\vartheta(t)$ is any limit of the ODE, then $\frac{d}{dt}F[\bar{\theta}] = 0$ and each term of the sum in(10) disappears. For each component, this implies that either $\bar{\theta}_k = 0$ or $\lim_{t\to\infty} D_k(t) = \lim_{t\to\infty}\bar{D}(t)$, which may be rewritten as $\mathcal{G}_k[F(\bar{\theta})] = 0$. In order to verify (8), it suffices to show now that if $\bar{\theta}_k = 0$ then the corresponding term $\mathcal{G}_k[F(\bar{\theta})] \geq 0$. By continuity of the generalized gradient, $\mathcal{G}_k[F(\bar{\theta})] = \lim_{t\to\infty}\mathcal{G}_k[F(\vartheta(t))]$. Using (9), since the limit point is $\bar{\theta}_k = 0$, this component must decrease as time increases, thus $\frac{d}{dt}[\vartheta(t)] \leq 0$ for large enough t, implying that $\lim_{t\to\infty}\mathcal{G}_k[F(\vartheta(t))] \geq 0$.

The limiting ODE will therefore have a limit point that satisfies the KKT conditions and is a local minimum. If F has a unique minimum, then the gradient driven process (9) is asymptotically optimal in the sense that the trajectories of $\vartheta(t)$ approach the optimal value as $t \to \infty$. □

5 The Phantom RPA Formula

If the exact values of μ_k are unknown, the gradient of $C(\theta)$ can be strongly consistently estimated via sample averages. Since $F(\theta) = W(\theta) + C(\theta)$, it suffices to build estimators of the generalized gradient in(9) for the waiting time. We shall now propose two phantom RPA estimators, generalizing [4] and [31] for the case of multi-valued decisions. Call:

$$\phi_W(\theta) = E\left[\sum_{i=1}^{\mathbf{N}(\theta)} W_i\right], \quad \phi_N(\theta) = E[\mathbf{N}(\theta)] \tag{11}$$

where $\mathbf{N}(\theta)$ is the number of customers in one busy period of the process at control value θ. Then by the Renewal Theorem, $W(\theta) = \phi_W(\theta)/\phi_N(\theta)$.

5.1 Parallel Phantom Systems

The model can be stated according to the framework of Section 2, as follows. Call A_i the interarrival time between customers i and $i+1$, and let $\{u_i\}$ be a sequence of i.i.d uniform variates. The service time of customer i is

$$S_i(u_i, \eta_i) = \sum_{k=1}^{V} \theta_k G_k^{-1}(u_i) \mathbf{1}_{\{\eta_i = v_k\}}$$

where $G_k(\cdot)$ is the service distribution of given speed v_k. The discrete event process is described via Lindley's equations. If R_i is the total time ith customer remains in the system and W_i its waiting time, then given R_i, Lindley's equations yield:

$$\begin{aligned} W_{i+1} &= \max(0, R_i - A_i) \\ R_{i+1} &= S_{i+1}(u_i, \eta_i) + W_{i+1} \end{aligned}$$

The above equations hold regardless of the way the speeds of the machine are chosen, since we are using the information on the decisions η themselves. Use now the notation $\eta_i(0)$ for the decisions of the *nominal system*, taken according to $P\{\eta_i(0) = v_k\} = \theta_k$, independently of $\{(A_j, u_j); j \le i\}$. Fix the index k to estimate a sensitivity, let $\Delta > 0$ be a small number, and $\tilde{\Delta} \in \Re^V$ be a vector with the value Δ as the k-th component. The other components satisfy $\tilde{\Delta}_l = -p_l \Delta, p_l > 0, l \neq k$ and shall be determined later, according to the two gradient estimators that to be described shortly. Given N customers, we define a parallel phantom system using the same sequence (A_i, u_i) and assigning the decision $\tilde{\eta}_i$ of the i-th customer as follows:

$$P\{\tilde{\eta}_i = v_l | \eta_i(0) = v_l\} = 1, \quad l \neq k$$
$$P\{\tilde{\eta}_i = v_k | \eta_i(0) = v_k\} = 1 - \frac{\Delta}{\theta_k}$$
$$P\{\tilde{\eta}_i = v_l | \eta_i(0) = v_k\} = \frac{p_l \Delta}{\theta_k}, \quad l \neq k$$

Then the decisions of the phantom system satisfy $P\{\tilde{\eta}_i = v_l\} = \theta_l - \tilde{\Delta}_l$, for all $l = 1, \ldots, V$ and $\tilde{\eta}$ has a distribution parametrized by $\tilde{\theta} = \theta - \tilde{\Delta}$, as shown in Figure 4. An item is called a *phantom* if $\tilde{\eta}_i \neq \tilde{\eta}$. Other than the decisions, the phantom system shares common random numbers with the nominal system, since interarrival epochs and service times are the same except for the chosen phantoms. Clearly, the evolution of such a system can be evaluated in parallel to the nominal system using Lindley's equation, as we show in detail in the Appendix.

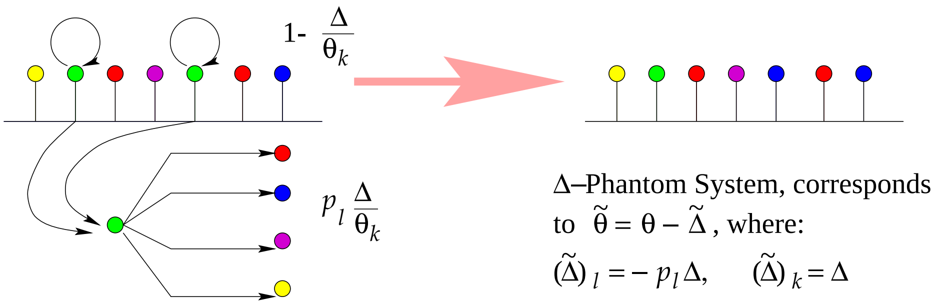

Fig. 4. Parallel Phantom System

For any sequence $\eta = \{\eta_i\}$ of decisions, define the cumulative waiting time over the first M customers as $\varphi_M(\eta)$. We shall be interested in two cases: M a fixed, deterministic number, and $M = \mathbf{N}(\eta)$, the number of customers in one busy period. To distinguish the two cases, use the notation:

$$\varphi_M(\eta) = \sum_{i=1}^{M} W_i, \quad \varphi_{\mathbf{N}}(\eta) = \sum_{i=1}^{\mathbf{N}(\eta)} W_i \tag{12}$$

The finite differences are defined by:

$$D_\Delta(M) = \frac{\varphi_M(\eta(0)) - \varphi_M(\tilde{\eta})}{\Delta} \tag{13}$$

$$D_\Delta(\mathbf{N}) = \frac{\varphi_{\mathbf{N}}(\eta(0)) - \varphi_{\mathbf{N}}(\tilde{\eta})}{\Delta} \tag{14}$$

The RPA formula is based on the fact that the probability that item number i is a phantom is $P\{\tilde{\eta}_i \neq \eta_i(0)\} = P\{\tilde{\eta}_i \neq v_k | \eta_i(0) = v_k\} P\{\eta_i(0) = v_k\} = \Delta$. In addition, the phantom system is constructed by deciding *independently* whether consecutive items will become phantoms or not. Thus, given M customers, the number of phantoms in the sequence $\{\tilde{\eta}_i, i = 1, \ldots, M\}$ follows a Binomial distribution. Therefore,

$$\begin{aligned} E[D_\Delta(M)] = & E\left[M\Delta(1-\Delta)^{M-1}E^{(1)}[D_\Delta(M)]\right] \\ & + E\left[\sum_{m=2}^{M}\binom{M}{m}\Delta^m(1-\Delta)^{M-m}E^{(m)}[D_\Delta(M)]\right] \end{aligned}$$

where $E^{(m)}$ is the expectation w.r.t. $\tilde{\eta}$, conditioning on having exactly m phantoms. As in [31], use the fact that $\sum_{m=2}^{M}\binom{M}{m}\Delta^m(1-\Delta)^{M-m} \leq M^2\Delta^2$ to bound the second term. If $M < \infty$ is a deterministic number, then $\varphi_M(\eta) \leq \sum_{i=1}^{M}\sum_{j=1}^{M} S_i$ for any sequence η of decisions, and since $E[S_i] \leq 1/\mu_V < \infty$, dominated convergence establishes that only the first term above will not vanish in the limit as $\Delta \to 0$.

Let the sequence $\eta(j)$ be defined by $\eta_i(j) = \eta_i(0), i \neq j$, and $P\{\eta_j(j) = v_l | \eta_j(0) = v_k\} = p_l$. That is, it's a phantom sequence that differs from the nominal decisions only in component j (if $\eta_j(0) = v_k$). To shorten notation, call $I_k(j) = \mathbf{1}_{\{\eta_j(0) = v_k\}}$, then:

$$\begin{aligned} \lim_{\Delta\to 0} E[D_\Delta(M)] = & E\left[ME^{(1)}[\varphi_M(\eta(0)) - \varphi_M(\tilde{\eta})]\right] \\ = & E\left[\frac{1}{\theta_k}\sum_{j=1}^{M} I_k(j)\varphi_M(\eta(0))\right] - \sum_{j=1}^{M} E\left[\varphi_M(\eta(j)) | j \in \mathcal{P}_M\right] \end{aligned}$$

where $\mathcal{P}_M = \{1 \leq j \leq M : I_k(j) = 1\}$ is the set of all decisions that assign speed k in the nominal system, and M_k is its cardinality. Therefore $E[M_k\varphi_M(\eta(0))] = \theta_k M$, which justifies the first term above. Now use $\theta_k = P(j \in \mathcal{P}_M)$ and the fact that $E^{(1)}$ is an expectation conditioned on both the number M_k and the fact that there is only one phantom, to get:

$$E\left[\sum_{j=1}^{M}\varphi_M(\eta(j))|j\in\mathcal{P}_M\right]=\frac{1}{\theta_k}E[M_kE^{(1)}\varphi_M(\tilde{\eta})]$$

to justify the second term inside the expectation. We have used here that all the sequences of phantom decisions that differ only in one component with $\eta(0)$ are equiprobable. Finally, rewrite this expression in the form:

$$\lim_{\Delta\to 0}E[D_\Delta(M)]=\frac{1}{\theta_k}E\left\{\sum_{j=1}^{M}I_k(j)\left[\varphi_M(\eta(0))-\varphi_M(\eta(j))\right]\right\} \tag{15}$$

Consider now the finite difference (14). In order to use the dominated convergence, assume that the service distributions are stochastically dominated by a system of possibly random decisions $\bar{\eta}$ such that $S(\eta_i(0),u_i)\le S_i(\bar{\eta},u_i)$ and $S_i(\tilde{\eta},u_i)\le S_i(\bar{\eta},u_i)$. Such is the case if $G_l(s)\ge G_V(s), l<V$, with $\bar{\eta}_i\equiv v_V$. Then the finite difference $D_\Delta(\mathbf{N})$ converges as $\Delta\to 0$ provided that $E[\mathbf{N}(\bar{\eta})^4]<\infty$. This follows from $\varphi(\mathbf{N}(\eta(0)),\eta(0))-\varphi(\mathbf{N}(\tilde{\eta}),\tilde{\eta})\le 2\sum_{i=1}^{\mathbf{N}(\bar{\eta})}W_i(\bar{\eta})$, which implies that $W_i(\bar{\eta})$ is bounded by the length of the busy period in the system with decisions $\bar{\eta}$, obtaining $\Delta E[\mathbf{N}^2(\bar{\eta})\mathbf{N}^2(\bar{\eta})]E[S_i(\bar{\eta})]\to 0$ as $\Delta\to 0$. Following exactly the same arguments as before, the corresponding RPA formula becomes:

$$\lim_{\Delta\to 0}E[D_\Delta(\mathbf{N})]=\frac{1}{\theta_k}E\left\{\sum_{j=1}^{\mathbf{N}(0)}I_k(j)\left[\varphi_{\mathbf{N}}(\eta(0))-\varphi_{\mathbf{N}}(\eta(j))\right]\right\} \tag{16}$$

where the sequences $\eta(j)$ are defined as before.

5.2 The Swapping Phantoms

Our first estimator of the generalized gradient is based on the following observation. Even if the partial derivatives $F_k'(\theta)$ make sense mathematically, expressed as $\lim_{\Delta\to 0}\frac{F(\theta)-F(\tilde{\theta})}{\Delta}$, where $\tilde{\theta}=\theta-\tilde{\Delta}$ and $\tilde{\Delta}_l=0, l\ne k$, the law of total probability is not preserved, because $\sum_{i=1}^{V}\tilde{\theta}_i\ne 1$ and thus the finite difference for $\Delta>0$ does not correspond to a physical process. An alternative is to look for appropriate directional derivatives. Consider $\tilde{\Delta}_l = \theta_l \Delta/(1-\theta_k)$. As detailed in [30], simple algebraic manipulations show that

$$\lim_{\Delta \to 0} \frac{F(\theta) - F(\tilde{\theta})}{\Delta} = \frac{1}{1-\theta_k} \mathcal{G}_k[F(\theta)] \tag{17}$$

Therefore, the estimation problem can be stated in terms of estimating the generalized gradient (7).

Phantom customers swap speeds compared to the corresponding nominal decisions, choosing the remaining ones in their original proportion. Use then (15) or (16) with $p_l = \theta_l/(1-\theta_k)$. This approach defines the corresponding estimators to drive the process towards the limiting ODE (9).

Proposition 4 *Assume that the service times are dominated by $\{S_i(\bar{\eta})\}$, for some random sequence $\bar{\eta}$, and that the dominating queueing process satisfies $E[\mathbf{N}^4(\bar{\eta})] < \infty$. Then:*

$$\theta_k \mathcal{G}_k[W(\theta)] = \lim_{M\to\infty} (1-\theta_k) E\left\{ \sum_{j=1}^{M} I_k(j) \, \frac{[\varphi_M(\eta(0)) - \varphi_M(\eta(j))]}{M} \right\}$$

$$\theta_k \mathcal{G}_k[\phi_W(\theta)] = (1-\theta_k) E\left\{ \sum_{j=1}^{\mathbf{N}(0)} I_k(j) \, [\varphi_{\mathbf{N}}(\eta(0)) - \varphi_{\mathbf{N}}(\eta(j))] \right\}$$

$$\theta_k \mathcal{G}_k[\phi_N(\theta)] = (1-\theta_k) E\left\{ \sum_{j=1}^{\mathbf{N}(0)} I_k(j) \, [\mathbf{N}(0) - \mathbf{N}(j)] \right\}$$

PROOF. The last two statements follow directly from the development of the phantom RPA formula, under the domination assumption. By construction, (15) is an unbiased estimator of

$$\frac{1}{(1-\theta_k)} \mathcal{G}_k \left[\frac{1}{M} \sum_{i=1}^{M} E(W_i | W_0) \right].$$

The convergence as $M \to \infty$ of the r.h.s. follows from the uniform bound $E|\varphi_M(\eta(0)) - \varphi_M(\eta(j))| \le 2E[\sum_{i=1}^{\mathbf{N}_j(\bar{\eta})} W_i(\bar{\eta})] < \infty$, where $\mathbf{N}_j(\bar{\eta})$ is the number of customers in the busy period of the dominating system where customer

j belongs. This fact is a consequence of the construction of the phantom systems: $W_i(0) = W_i(j)$ for all $i \leq j$. When the system $\bar{\eta}$ finishes the busy period where j belongs, both the nominal and the j-th phantom system with decisions $\{\eta(j)\}$ also finish. After this, their evolution is identical, since j is the only decision that is different.

Having established that the limit

$$w_k(\theta) = \lim_{M\to\infty} \frac{\partial}{\partial\theta_k}\left[\frac{1}{M}\sum_{i=1}^{M} E(W_i|W_0)\right]$$

exists a.s. for any initial distribution, we need only to verify that the interchange between the derivative and the limit $M \to \infty$ is valid. Since the invariant measure is unique and ergodic, $\mathcal{G}_k[W(\theta)] < \infty$, and the generalized gradient is a linear operator, this is established if

$$\frac{\partial}{\partial\theta_k}[W(\theta)] = w_k(\theta) = \lim_{M\to\infty} \int \mu_\theta(dx) \frac{\partial}{\partial\theta_k}\left[\frac{1}{M}\sum_{i=1}^{M} E(W_i|W_0 = x)\right]$$

where μ_θ denotes the stationary measure. This result now follows from Theorem 4 of [33], where it is assumed that for each θ, the Markov process $\{W_i\}$ has a unique invariant measure μ_θ, which is true in our case under the uniform stability assumption. Furthermore, $\{\mu_\theta; \theta \in A\}$ is tight for every compact set $A \in \Theta$, as shown in Section 7. The other assumptions in [33] follow from the fact that for our model the n-step transition probability of the queueing process is a polynomial in $(\theta_i, i = 1, \ldots, k)$ of degree n. Therefore the likelihood ratio (w.r.t. θ_k) exists as required, provided θ is in the interior of the feasible set Θ. In addition, under uniform stability these polynomial probabilities converge to the stationary measure exponentially fast, verifying the required rate of convergence of the partial sums and averages. $\square$

The practical problem is the required condition for domination. In our numerical examples, we used Poisson arrivals and uniformly distributed service distributions for each speed. In general, for an $M/G/1$ queue, Takács method (see [14]) gives a functional relationship between the moment generating function of the service time and the one of the number in a busy period. If the m-th moment of the service distribution is bounded, so is $E[\mathbf{N}^m]$, which follows from the above argument, although we omit the details of the proof. In all our simulations, $E[S_i^4] < \infty$.

5.3 The Disappearing Phantoms

The requirement that the system be stable if we always choose the slower speed may be too restrictive, especially when the service distributions are not explicitly known. Naturally, if they were known, the the analytical solution could be evaluated to find the optimal setting without need for adaptive control. In the more general situation, the machine operates and evaluates its own estimates to drive the control towards the optimal value. In such cases, and when the service distributions as well as the input rate may vary, the algorithm should be able to track the optimal control. We propose now a more robust estimator that does not require assigning the services of the phantom customers as $S_i(u_i, v_l)$ with probability $p_l = \theta_l/(1-\theta_k)$.

Consider the alternative model for the original process, where customers of class k arrive according to a renewal process with rate $\theta_k\lambda$. They require an amount of service that has distribution $G_k(\cdot)$. In this model, $\tilde{\theta} = \theta - \tilde{\Delta}$, with $p_l = 0, l \neq k$, meaning that some of the customers, the phantom ones, are not allowed entrance to the machine and thus "disappear" from the system. The corresponding system will have a total incoming rate of $\tilde{\lambda} = \lambda - \Delta\lambda$. In this case, see [30],

$$G_k(\lambda,\theta) = \lim_{\Delta\to 0} \frac{F(\lambda,\theta) - F(\tilde{\lambda},\tilde{\theta})}{\Delta} = \theta_k \mathcal{G}_k[F(\lambda,\theta)] + \lambda\frac{\partial F(\lambda,\theta)}{\partial\lambda} \tag{18}$$

and therefore the generalized gradient can be expressed as:

$$\mathcal{G}_k[F(\lambda,\theta)] = G_k[F(\lambda,\theta)] - \sum_{k=1}^{V} \theta_k G_k[F(\lambda,\theta)] \tag{19}$$

In the case of the disappearing phantoms, Lindley equations can be calculated in parallel to the nominal system using $S_j(j) \equiv S_j(\eta_j(j), u_j) \equiv 0$.

Proposition 5 *Assume that $E[\mathbf{N}(0)^4] < \infty$. Then:*

$$\theta_k G_k[W(\theta)] = \lim_{M\to\infty} E\left\{\sum_{j=1}^{M} I_k(j)\, \frac{[\varphi_M(\eta(0)) - \varphi_M(\eta(j))]}{M}\right\}$$
$$\theta_k G_k[\phi_W(\theta)] = E\left\{\sum_{j=1}^{\mathbf{N}(0)} I_k(j)\, [\varphi_{\mathbf{N}}(\eta(0)) - \varphi_{\mathbf{N}}(\eta(j))]\right\}$$
$$\theta_k G_k[\phi_N(\theta)] = E\left\{\sum_{j=1}^{\mathbf{N}(0)} I_k(j)\, [\mathbf{N}(0) - \mathbf{N}(j)]\right\}$$

PROOF. Since $S_j(j) = 0$, the nominal system dominates all of the possible phantom systems and $\mathbf{N}(\tilde{\eta}) \leq \mathbf{N}(\eta(0))$ a.s., which requires then the condition $E[\mathbf{N}(0)^4] < \infty$. Notice that in this case it is no longer required that $\lambda/\mu_V < 1$, but only that $\lambda < \mu(\theta) = \sum_k \theta_k/\mu_k$. As before, the first statement follows from the a.s. convergence of the derivatives of the finite horizon averages to those of the stationary averages. $\square$

5.4 The RPA Gradient Estimators

The actual estimators of the gradients in (18) via the disappearing phantoms are now presented, using two methods based on the regenerative estimation approach, and a third one called the non-reset version of the estimator [18]. The development for the swapping phantoms is analogous and we omit the details. Let $C_k = \lambda c_k/\mu_k$ be the partial derivative of $C(\theta)$ w.r.t. θ_k and call:

$$\Gamma_k(\theta) = G_k[W(\theta)] + C_k$$

so that, clearly, using (18),

$$\mathcal{G}_k[F(\theta)] = \Gamma_k(\theta) - \sum_{l=1}^{V} \theta_l \Gamma_l(\theta). \tag{20}$$

We shall focus on constructing the estimators of $\Gamma_k(\theta)$. To simplify notation, fix the index k to evaluate the estimator of $\Gamma_k(\theta)$, and call:

$$\gamma_W(\theta) = G_k[\phi_W(\theta)], \quad \gamma_N(\theta) = G_k[\phi_N(\theta)]$$

Taking the appropriate derivative of $W(\theta)$ from (11), we get

$$\Gamma_k(\theta) = \frac{\gamma_W(\theta)}{\phi_N(\theta)} - \frac{\gamma_N(\theta)\phi_W(\theta)}{[\phi_N(\theta)]^2} + C_k$$

We shall first describe the ensuing estimators when a regenerative approach is chosen. Since the nominal system dominates the phantoms, we shall use $\mathbf{N}_m(0)$ to denote the number of customers in the m-th busy period of the phantom system, and $\mathbf{N}_\kappa^{(m)}(j)$ for the κ-th busy period of the j-th phantom system within the m-th nominal system. In order to obtain unbiased estimations of the numerator, we use the approach proposed in [5], [1]. We estimate $\Gamma_k(\theta)$ using $2M$ busy periods, estimating $\gamma_W(\theta)$ and $\phi_W(\theta)$ using the odd numbered BP's, then $\gamma_N(\theta)$ and $\phi_N(\theta)$ in the even ones, so that the expectation of the product is the product of the expectations. The estimators of $\phi_N(\theta)$ and $\phi_W(\theta)$ in the m-th busy period, called $\hat{\phi}_N(m), \hat{\phi}_W(m)$ respectively, are unbiased using the sample averages. The estimators of $\gamma_N(\theta)$ and $\gamma_W(\theta)$ are given respectively by their RPA estimators:

$$\hat{\gamma}_W(m) = \frac{1}{\theta_k} \sum_{j=1}^{\mathbf{N}_m(0)} \sum_{i=j+1}^{\mathbf{N}_m(0)} d_i(j), \quad \hat{\gamma}_N(m) = \frac{1}{\theta_k} \sum_{j=1}^{\mathbf{N}_m(0)} [\mathbf{N}_m(0) - \mathbf{N}_1^{(m)}(j)]$$

where $d_i(j) = W_i(0) - W_i(j) \leq S_j(j)$ a.s. (see the Appendix), therefore:

$$\begin{aligned} \mathrm{Var}\{\hat{\gamma}_W(m)\} &\leq \frac{1}{\theta_k^2} E\left\{ \theta_k^2 \mathbf{N}_m^2(0) E\left[\left(\sum_{i=j+1}^{\mathbf{N}_m(0)} d_i(j) \right)^2 \;\Big|\; \mathbf{N}_m(0) \right] \right\} \\ &\leq \frac{1}{\mu_k} E[\mathbf{N}^4(0)] \equiv \mathbf{V}(\theta) < \infty \end{aligned} \tag{21}$$

which is uniformly bounded for $\theta \in \Theta$, and clearly, $\mathrm{Var}[\hat{\gamma}_N(m)] < \infty$ is also uniformly bounded in Θ. Using Proposition 5 and independence between different busy periods, $E\{\hat{\gamma}_N(2m+1)\hat{\phi}_W(2m)\} = \gamma_N(\theta)\phi_W(\theta)$, for every m. As described, the n-th estimation interval considers M pairs $(2m, 2m+1)$ of busy periods, from $m = nM+1$ up to $m = (n+1)M$. To simplify notation, for any function $f(m)$ call:

$$\sum^{(n)} f(m) = \sum_{m=nM+1}^{(n+1)M} f(m)$$

Method 1: The regenerative estimator $\hat{\Gamma}_n^{(1)}(M,k)$ is given by:

$$\hat{\Gamma}_n^{(1)}(M,k) = \frac{\sum^{(n)} \hat{\gamma}_W(2m)}{\sum^{(n)} \hat{\phi}_N(2m+1)} - \frac{\sum^{(n)} \hat{\gamma}_N(2m+1) \sum^{(n)} \hat{\phi}_W(2m)}{\left[\sum^{(n)} \hat{\phi}_N(2m+1)\right]^2} + C_k \quad (22)$$

For any M and all k, the $\{\hat{\Gamma}_n^{(1)}(M,k), n = 1, 2, \ldots\}$ are i.i.d. and their variance is bounded, therefore they are strongly consistent estimators of $\Gamma_k(\theta)$. However, they are not consistent in the average: for any m, $\frac{1}{m}\sum_{n=0}^{m-1} E[\hat{\Gamma}_n^{(1)}(M,k)] = E[\hat{\Gamma}_1^{(1)}(M,k)]$, but for any fixed M, the estimator $\hat{\Gamma}_1^{(1)}(M,k)$ is biased.

Method 2: As a variant, we consider another regenerative estimator $\hat{\Gamma}_n^{(1)}(M,k)$ similar to $Y_n(N)$, but for which the denominator is replaced by a cumulative version of it. We define

$$\hat{\phi}_N^{(n)}(M) = \frac{1}{(n+1)M} \sum_{m=0}^{(n+1)} \hat{\phi}_N(2m+1)$$

We use now:

$$\hat{\Gamma}_n^{(2)}(M,k) = \frac{\sum^{(n)} \hat{\gamma}_W(2m)}{\hat{\phi}_N^{(n)}(M)} - \frac{\sum^{(n)} \hat{\gamma}_N(2m+1) \sum^{(n)} \hat{\phi}_W(2m)}{\left[\hat{\phi}_N^{(n)}(M)\right]^2} + C_k \quad (23)$$

This estimator is strongly consistent for $\Gamma_k(\theta)$, and also consistent in the average, since for every M, $\hat{\phi}_N^{(n)}(M) \to \phi_N(\theta)$ a.s. as $n \to \infty$, but now $\{\hat{\Gamma}_n^{(2)}(M,k), n = 1, 2, \ldots\}$ are neither independent nor identically distributed.

In practice, Method 1 has been preferred to Method 2 for stochastic optimization, mostly because the sequence of estimates are i.i.d. Both are strongly consistent and that is the required property for (1) to converge towards the optimum. We shall discuss their behavior in Section 7.

The non-reset RPA estimator uses a finite horizon of fixed length M, without resetting the propagation of the perturbations to zero. It uses the information available over the service completions from customer $nM + 1$ to customer $(n + 1)M$. In this case Γ_k is estimated directly, without using the renewal theorem.

Method 3: Consider now:

$$\hat{\Gamma}_n^{(3)}(M,k) = \frac{1}{\theta_k M}\left(\sum_{j=\beta_n}^{nM}\sum_{i=nM+1}^{\alpha_n} d_i(j) + \sum_{j=nM+1}^{(n+1)M}\sum_{i=j+1}^{(n+1)M} d_i(j)\right) + C_k \quad (24)$$

where β_n (respectively, α_n) is the index of the customer that starts (finishes) the nominal busy period where customer $nM + 1$ belongs. The second term corresponds to the (unbiased) estimator of the gradient of the finite horizon average waiting time. The first term accounts for the bookkeeping, and represents the propagation of the perturbation of the previous phantom systems into the current estimation interval $i \geq nM + 1$. It is called the "initial accumulator".

This estimator is both strongly consistent and (strongly) consistent in the average, provided that $E[\mathbf{N}^4(0)] < \infty$. To see this, we use the fact that the double sum in the first term (or accumulator) has expectation bounded by $E[\theta_k \mathbf{N}(0) E[\sum_{j=1}^{\mathbf{N}(0)} S_j(j)|\mathbf{N}(0)]] \leq \theta_k E[\mathbf{N}^2(0]/\mu_k < \infty$. Its variance is bounded by $\mathbf{V}(\theta)$ in (21). Let $N_\theta(M)$ be the number of busy periods totally contained within $[nM+1,(n+1)M]$, then (24) can be written as a sum over the $N_\theta(M)$ busy periods plus the corresponding terms in the first and last busy periods. Therefore, using independence and Wald's identity (conditioning on $N_\theta(M)$):

$$\mathrm{Var}[\hat{\Gamma}_n^{(3)}(M,k)] \leq \frac{1}{M} E\left(\frac{N_\theta(M)}{M}\right)\mathbf{V}(\theta) + 2\frac{\mathbf{V}(\theta)}{M}$$

which implies strong convergence as $M \to \infty$. As for the consistency in the average, it suffices to remark that our estimator is additive, so that for any m,

$$\frac{1}{m} \sum_{n=0}^{m-1} \hat{\Gamma}_n^{(3)}(M,k) = \hat{\Gamma}_0^{(3)}((m-1)M,k)$$

and therefore, for any M, as $m \to \infty$, from Proposition 5 and the fact that the initial accumulator converges to zero a.s. for any initial condition, this latter average converges to $\Gamma_k(\theta)$ a.s.

Finally, using either of the estimators $\hat{\Gamma}_n^{(e)}(M,k)$, $e = 1,2,3$, let:

$$Y_{n,k}^{(e)}(M,\theta) = \left(\hat{\Gamma}_n^{(e)}(M,k) - \sum_{l=1}^{V} \theta_l \hat{\Gamma}_n^{(e)}(M,l) \right) \tag{25}$$

be the estimator of the desired gneralized gradient. Summarizing our previous results, it follows that $\theta_k Y_{n,k}^{(e)}(M,\theta)$ is a strongly consistent estimator of $\theta_k \mathcal{G}_k(\theta)$ for $e = 1,2,3$ and consistent in the average for $e = 2,3$.

6 The Learning Algorithms

We are now ready to go back to where we started. From Lemma 3 the target limiting ODE (9) is asymptotically optimal for this problem. In the following section we shall analyze the behavior of the stochastic approximation using Methods $e = 1,2,3$ (that is, the estimator (25) with either (22), (23) or (24)) and updating each component k of the control variable as:

$$\theta_{n+1,k}^{\epsilon} = \theta_{n,k}^{\epsilon} + \epsilon \theta_{n,k}^{\epsilon} Y_{n,k}^{(e)}(M) \tag{26}$$

where, for Methods 1 and 2, Δ_n^{ϵ} has the total length of the busy periods labeled by $m = nM+1, \ldots, (n+1)M$ and for Method 3, $\Delta_n^{\epsilon} = M$. Theorem 2 will be applied to this procedure and the predicted results will be presented visually.

We shall next compare the performance of the procedures with the more commonly used stochastic approximation using Method 1 and:

$$\theta_{n+1} = \theta_n + \epsilon_n Y^{(1)}_{n,k}(M(n)) \tag{27}$$

where $\epsilon_n \approx 1/n$ and $M(n) \approx n$. According to Theorem 1, $\theta_n \to \theta^*$ a.s.

Remark: Recall that, while the properties of the estimators in Section 2.1 were stated for the fixed θ process, now the notation $Y^{(e)}_{n,k}(M)$ referes to the varying θ process (we have omitted the superscript ϵ for obvious æsthetic reasons).

6.1 A Remark on Truncation

The most commonly used gradient approximation procedures prescribe that the control variables change in the direction of the gradient. It is very common that an individual update results in an unfeasible parameter, having one or more negative components. Truncations have been used (projections into the feasible set) in [18] and [32] (among others) to establish the weak convergence result towards a local (feasible) optimum. While this is a viable tool for mathematical analysis of the limiting behaviour, in practice ϵ is a positive nonvanishing constant and truncations have undesireable behaviour when the controls are probability vectors. For, if an update yields a component larger than unity, how should this value be truncated and the others adjusted so that the resulting vector defines a probability? Clearly, a simple truncation of the vector to the value 1.0 yields a degenerate probability. It is often the case that when a particular component of θ is zero, no more observations are available to estimate the corresponding partial derivative from that time on. In other words, the topology or specification of the system has changed and we can no longer improve the accuracy of the estimation. Such points are typically absorbing points of the Markov Decision Process, even though they may be far from any Karush-Kuhn-Tucker point.

This practical question inspired our introduction of the generalized gradient. We mentioned in Section 4 that (9) needs no further truncation, since its evolution automatically ensures that $\theta(\cdot)$ lies always in the feasible set Θ. Consider now the approximation (26): What ensures that the subsequent θ_n^ϵ are going to be probability vectors in Θ?

In the discussion that follows, we shall omit the indices $(e), M$ to simplify notation. Consider (26) for the updates, with $Y_{n,k}$ defined by (25) with the corresponding current value of θ_n^ϵ. As in Section 4, it is straightforward to show that, given any $\epsilon > 0$, $\sum_k \theta_{n,k}^\epsilon = 1$ remains constant for all n. On the other hand, if the estimators $\{\hat{\Gamma}_n(k)\}$ were uniformly bounded a.s., then for some finite B, $\|Y_{n,k}\| \leq B$ a.s. for all n, k. In this case, choosing $\epsilon \leq B^{-1}$ ensures that $\epsilon Y_{n,k} \geq -1$ a.s, which in turn from (26) implies that $\theta_{n+1,k}^\epsilon \geq 0$. Once again, these two conditions imply that the subsequent values of the control variable lie necessarily inside the feasible set Θ and no further truncation is necessary.

The appropriate (and more reasonable) assumption for the derivative estimators is, however, not a.s. uniform boundedness but *uniform integrability*, which is required by Theorem 2. In order to ensure feasibility, we propose (and implement) the following rule. For a given value of $\epsilon > 0$, use:

$$\epsilon_n = \min\left(\epsilon, \frac{1}{2\max_k \|\hat{\Gamma}_n(k)\|}\right)$$

which ensures that $\epsilon_n Y_{n,k} \geq -1$. Although this seems not be a "constant" step size, we shall now re-write it and show that Theorem 2 is applicable. For a given $\epsilon > 0$, define for the fixed θ process the modified estimator:

$$\tilde{Y}_{n,k} = \begin{cases} Y_{n,k} & \text{if } 2\max_k \|\hat{\Gamma}_n(k)\| < \epsilon^{-1} \\ \dfrac{Y_{n,k}}{2\epsilon \max_k \|\hat{\Gamma}_n(k)\|} & \text{otherwise} \end{cases}$$

so that the stochastic approximation:

$$\theta_{n+1,k}^\epsilon = \theta_{n,k}^\epsilon + \epsilon\, \theta_{n,k}^\epsilon \tilde{Y}_{n,k}$$

will automatically yield feasible controls. This truncates according to a proportional adjustment of all components. Under the assumption that $\hat{\Gamma}_{n,k}$ are uniformly integrable (which we prove later on), $\{Y_{n,k}\}$ in (25) are also uniformly integrable, and since $\|\tilde{Y}_{n,k}\| \leq \|Y_{n,k}\|$, then so are the modified estimators. Consider now the fixed θ process, $\theta \in \Theta$. We now show that the modified estimators are consistent in the average with a vanishing bias:

$$E\left\{\frac{1}{m}\sum_{n=1}^{m}\tilde{Y}_{n,k}\right\} = E\left\{\frac{1}{m}\sum_{n=1}^{m}Y_{n,k}\right\} - \rho_m(\epsilon)$$

with:

$$\rho_m(\epsilon) = E\left\{\frac{1}{m}\sum_{n=1}^{m}Y_{n,k}\left(1-\frac{1}{2\epsilon\max_k\|\hat{\Gamma}_n(k)\|}\right)\mathbf{1}_{\left\{2\max_k\|\hat{\Gamma}_n(k)\|>\epsilon^{-1}\right\}}\right\}$$

Use now:

$$\begin{aligned}\|\rho_m(\epsilon)\| &\leq \frac{1}{m}\sum_{n=1}^{m}E\left\{\|Y_{n,k}\|\mathbf{1}_{\left\{2\max_k\|\hat{\Gamma}_n(k)\|>\epsilon^{-1}\right\}}\right\}\\ &\leq \frac{1}{m}\sum_{n=1}^{m}E\left\{2\max_k\|\hat{\Gamma}_n(k)\|\mathbf{1}_{\left\{2\max_k\|\hat{\Gamma}_n(k)\|>\epsilon^{-1}\right\}}\right\}\end{aligned}$$

which, by uniform integrability, tends to zero uniformly in m as $\epsilon \to 0$.

In all our experiments, although we implemented this rule, the truncation never actually took place. Using similar arguments, it can be shown that (27) can also be modified in this manner without affecting the limiting properties.

6.2 Verification of Assumptions

In the framework of Section 2, we let $\{X_i(\theta)\}$ denote the process that describes the waiting time in terms of customer number. Our underlying probability space is defined by $\omega = \{(A_i, u_i, \eta_i), i = 1, 2, \ldots\}$, so that $X_{i+1}(\theta) = \max[0, (X_i(\theta) + S_i(u_i, \eta_i) - A_i)]$. In order to simplify notation, we focus on Methods 1 and 2 with $M = 1$, but the treatment for Method 3 and any M is similar. The enlarged state is constructed using

$$Z_i(\theta) = (\Psi_i(k), i - \alpha_i, (d_i(j,k); j = \alpha_i, \dots, i-1); k = 1 \dots, V)$$

where α_i is the index of the first customer in the busy period where customer i belongs, $d_i(j,k) = [W_i(0) - W_i(j)]I_k(j)$ and $\Psi_i(k)$ is given by:

$$\Psi_i(k) = \sum_{j=\alpha_i}^{i-1} I_k(j) \sum_{l=j+1}^{i} d_l(j,k) \tag{28}$$

Assumptions 1 and 4 are verified for the process at fixed control value θ. We provide in the Appendix the recursions satisfied by $Z_i(\theta)$, which imply that $P\{\xi_{i+1}(\theta) \in B | \mathfrak{F}_i(\theta)\} = P\{\xi_{i+1}(\theta) \in B | \xi_i(\theta)\}$ and this latter is a linear function of θ, yielding weak continuity. Notice that the enlarged process also regenerates at $W_i(0) = 0$, when the components of Z_i are set to 0. Under stability of $\{X_i(\theta)\}$ for every $\theta \in \Theta$, the number of customers and the length of the busy periods is finite a.s., implying that the invariant measure μ_θ exists, it is ergodic and $\{\mu_\theta; \theta \in A\}$ tight for every compact set $A \subset \Theta$. This verifies Assumption 1.

Assumption 4 follows from $E[\mathbf{N}^4(0)] < \infty$, with $G_k(\theta) = \theta_k \mathcal{G}_k(\theta) + \beta(\theta, M)$ for each component k and $\Delta(t) = E[\mathbf{N}(0)]$ when $M = 1$. For methods 2 and 3, $\beta(\theta, M) = 0$ and for method 1, $\beta(\theta, M) \to 0$ as $M \to \infty$.

Assumptions 2 and 3 refer to the MDP model $\{(\xi_i^\epsilon, \tilde{\theta}_i^\epsilon)\}$. Clearly, when $W_{i+1}(0) = 0$ and an update takes place, the value of $\tilde{\theta}_{i+1}^\epsilon$ is obtained using the current value of $\Psi_i(k), k = 1, \dots, V$. Then the future evolution of the state ξ_j^ϵ depends on the value $\tilde{\theta}_i^\epsilon$. Since $\sup_{\theta \in \Theta} \mathbf{V}(\theta) < \infty$, $\mathrm{Var}\{\Delta_n^\epsilon | \theta_n^\epsilon = \theta\}$ and $\mathrm{Var}\{Y_{n,k}^{(e)}(M) | \theta_n^\epsilon = \theta\}$ are uniformly bounded for $\theta \in \Theta$, therefore they are uniformly integrable, verifying Assumption 2.

Finally, Assumption 4 requires $\{(\xi_i^\epsilon, \tilde{\theta}_i^\epsilon)\}$ to be tight. Tightness of $\tilde{\theta}_i^\epsilon$ follows from boundedness $0 \leq \theta_k \leq 1$. Verifying tightness of ξ_i^ϵ presents a minor difficulty, which was also present when we coded the algorithms: the dimension of Z_i is random and, in principle, unbounded, since it has as many components

$\{d_i(j,k)\}$ as customers present in a busy period. Use an array of fixed but "large" dimension D, and notice that $P\{\mathbf{N}_m(0) > D\}$ can be made uniformly small for any $\theta \in \Theta$, since the latter is a compact set and the process is stable for all $\theta \in \Theta$. Given $\tilde{\theta}_i^\epsilon = \theta$, at the start of the estimation interval, the initial distribution is fixed and independent of θ. Use now the bounds $d_i(j) \leq \bar{S}$ (the upper bound on our uniform distributions), $i - \alpha_i \leq \mathbf{N}_m(0)$ where m is the current BP, $\Psi_i(k) \leq \mathbf{N}_m^2(0)\bar{S}$ and the fact that $W_i(0)$ is a.s. bounded by the length of the current BP. Using our truncation argument, we can now choose constants K_i sufficiently large so that $P\{X_i(\theta) > K_1, \Psi_i > K_2, (i - \alpha_i) > K_3\} < \delta$. This implies tightness of $\{\xi_i^\epsilon | \tilde{\theta}_i^\epsilon = \theta\}$, over one regenerative cycle, where the bounds are independent of θ.

7 Simulation Results

For the purposes of verifying our simulation results, we consider the case $V = 2$. Then $\theta = P\{\eta_i = v_1\}$ becomes scalar, and the process regenerates if $\lambda < \mu(\theta)$. Our model is a M/G/1 queue with uniformly distributed service for each speed. Let $S(k), k = 1, 2$ be the service time of items processed at speed v_k, then $S(k) \sim U[a_k, b_k]$. We used $\lambda = 0.028; a_1 = 33, b_1 = 38, c_1 = 5; a_2 = 3, b_2 = 7, c_2 = 145$. Let $C = \lambda(c_1/\mu_1 - c_2/\mu_2)$ be the derivative of $C(\theta)$. Using the Pollaczek-Kintchine formula:

$$\frac{dF}{d\theta} = \frac{\lambda}{2} \frac{(\varsigma_1^2 - \varsigma_2^2) - \lambda(\varsigma_1^2 \mu_2^{-1} - \varsigma_2^2 \mu_1^{-1})}{\left[1 - \lambda\left(\frac{\theta}{\mu_1} + \frac{1-\theta}{\mu_2}\right)\right]^2} + C \tag{29}$$

where $\varsigma_k^2 = E[S_k^2]$. Figure 5 shows the graph of the performance function, along with the waiting time and cost as functions of the control parameter. The dotted line plots $C(\theta)$, the dashed-dotted line plots $W(\theta)$, and the solid line plots $F(\theta)$.

Clearly, for certain combinations of costs c_1 and c_2, it will be more advantageous to always use one of the two possible speeds, so that θ will either be 0 or

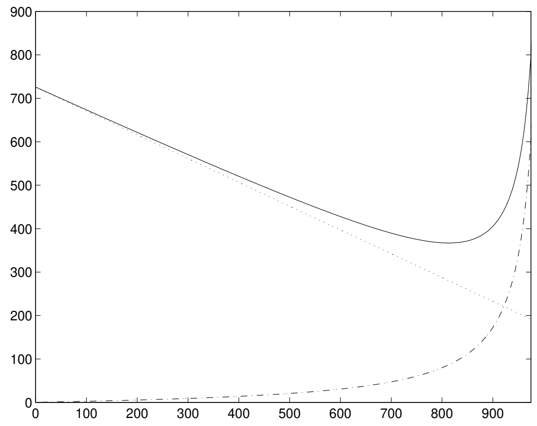

Fig. 5. Behavior of the Cost Functions

1. In our example, $\theta^* = 0.812$ and $F(\theta^*) = 367.01$, but the method can also be used if θ^* lies at the boundaries. In all our simulations, our initial point was $\theta(0) = 0.51$. In all our graphs showing the behavior of $\vartheta^\epsilon(\cdot)$, the solid line indicates the evolution of $\vartheta(\cdot)$, given by (9), and the dotted line indicates the evolution of $\vartheta^\epsilon(\cdot)$.

7.1 Understanding Weak Convergence

In this section we show how the choice of ϵ changes both the convergence rate and the fluctuations of the control process associated to (26) around the limiting ODE. Our aim is to present in a visual manner what the convergence in distribution of a stochastic process to a limiting deterministic process means in practice. Theorem 2 predicts that, as $\epsilon \to 0$, the processes $\vartheta^\epsilon(\cdot)$ approach (in the weak topology) the solid curves in Figure 6, which are the solution of (9). We plotted the values $\tilde{\theta}_i^\epsilon$ against the number of iterations to show the ladder interpolation processes $\vartheta^\epsilon(\cdot)$ The three plots are a sequence of trajectories of the processes $\vartheta^\epsilon(\cdot)$ obtained using Method 3, with $M = 100$, going (from left to right) from $\epsilon = 5 \times 10^{-4}$ to $\epsilon = 5 \times 10^{-5}$ to $\epsilon = 5 \times 10^{-6}$. This is the only difference in the stochastic approximations.

Notice, however, that the iteration numbers increase by a corresponding factor in the plots. Naturally, as one increases ϵ and thus the rate of convergence, one loses accuracy in the stochastic approximation.

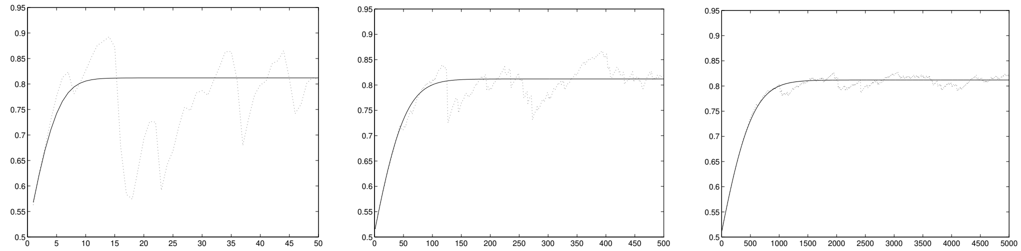

Fig. 6. Weak Convergence to the ODE

7.2 Understanding Consistency In The Average

We show now how closely the different processes converge to the deterministic solution of the target ODE when (26) is used for the updates. Figure 7 shows the control processes with the regenerative estimators, given in (22) (Method 1, to the left) and (23) (Method 2, to the right) respectively. The parameters are identical, i.e. $M = 30$ and $\epsilon = 10^{-6}$.

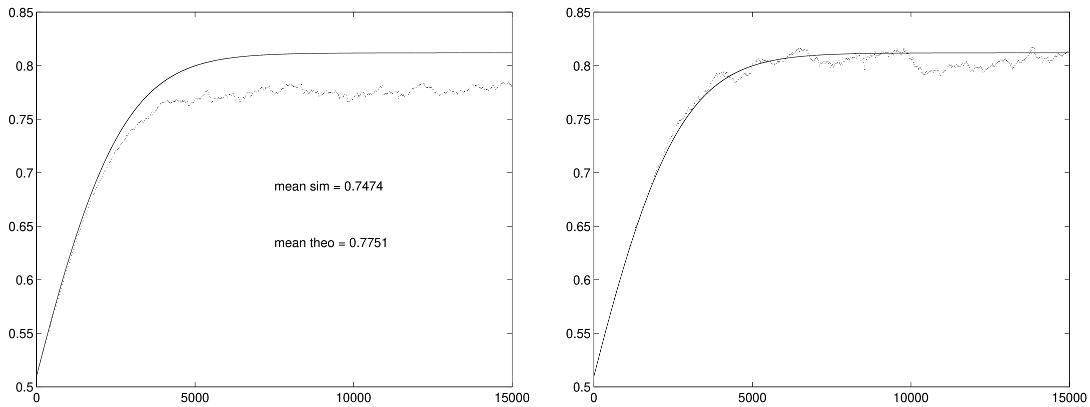

Fig. 7. Ladder Interpolation Process, Methods 1 and 2

The sub-optimality present for Method 1 comes from the fact that the denominator in (22) introduces a bias, not present in Method 2. This bias may be reduced if M becomes large (in practice we needed a value of 5000). Theorem 2 predicts that the stochastic approximations with Methods 2 and 3 converge towards the limiting ODE (9). Figure 8 shows the result using the non-reset estimator (24) of Method 3 in (26), with parameters $M = 50$ and $\epsilon = 10^{-6}$. Its behavior is similar to the one using Method 2, in Figure 7.

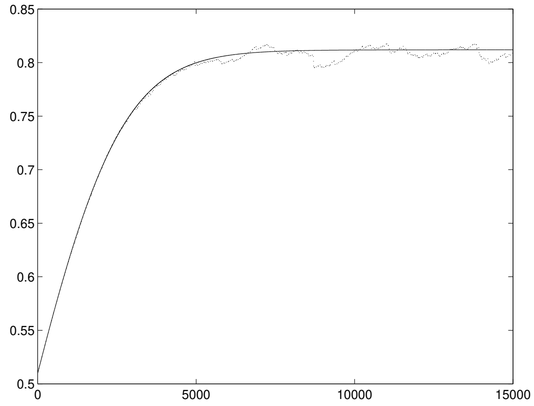

Fig. 8. Ladder Interpolation Process, Method 3

7.3 Understanding Time Scales

We now compare the behavior of the natural interpolation processes using methods 2 and 3 in (26). Both estimators are consistent in the average sense. Figure 7 and Figure 8 indicate that they behave similarly, but the non-reset estimator used less clients ($M = 50$) for its estimation than the regenerative method 2 used in the average: the mean number $\mathbf{N}$ of clients per BP was calculated as the simulation went on, and averaged a little over 5. Multiplying this number by $2M = 60$ gives a rough idea of the number of clients involved in each update interval.

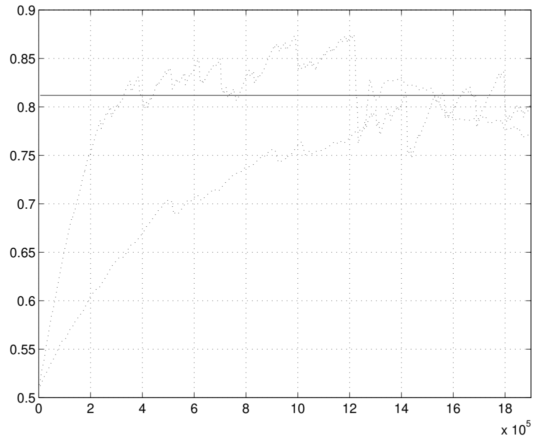

Fig. 9. Natural Interpolation Processes, Methods 2 and 3

We show in Figure 9 their behavior in real time (as given by the actual moments of service completions as simulated), thus using the natural interpolation process $\tilde{\vartheta}^\epsilon(t)$. Convergence to a small vicinity of θ^* may be six times as fast for Method 3, as expected. The solid line is now set at the optimal value θ^*. In this example, it took approximately 3×10^6 units of time for the stochastic approximation using Method 3 to reach a small vicinity of θ^*. Relating this time to $\lambda = 0.028$, we see that the number of items needed is approximately 85 000. Contrast this with the number of customers that it takes to produce a single accurate estimate of the gradient at fixed control value θ. It is common to use as much as 20,000 customers for some estimation methods. In [31] we used the non-reset estimator for a priority queueing system with $M \approx 10,000$ to obtain measurements within a 15% error. The likelihood ratio estimator for the same model of the flexible machine with two speeds uses $10,000$ regenerative cycles at $\theta = 0.7$ to produce estimates within a 20% error (reported also in [31]). We believe that reaching the optimum within a small neighborhood after processing 85,000 customers yields a reasonable performance for adaptive control.

7.4 Strong vs Weak Convergence

In this section we compare the natural interpolation processes corresponding to the stochastic approximations (26) and (27), to illustrate in a visual manner the different meanings of strong and weak convergence . Equation (26) was implemented using Method 3 with $\epsilon = 2 \times 10^{-5}$ and $M = 150$. For (27), we used Method 1 with $M(n) = M(0) + 25m(n)$ and $\epsilon_n = \epsilon_0/m(n)$, where $m(n)$ is the integer part of $n/25$.

Figure 10 shows the corresponding natural interpolation processes as they approach θ^*, represented by a solid line. We used $\epsilon_0 = 10^{-4}$ and $M(0) = 500$ for the plot to the left, while for the plot to the right $M(0) = 400$. Strong convergence indeed ensures a smoother trajectory and less fluctuations, but at the expense of speed, since the adaptive scheme with constant learning rate yields a close approximation to θ^* faster than the time taken by the strongly convergent process.

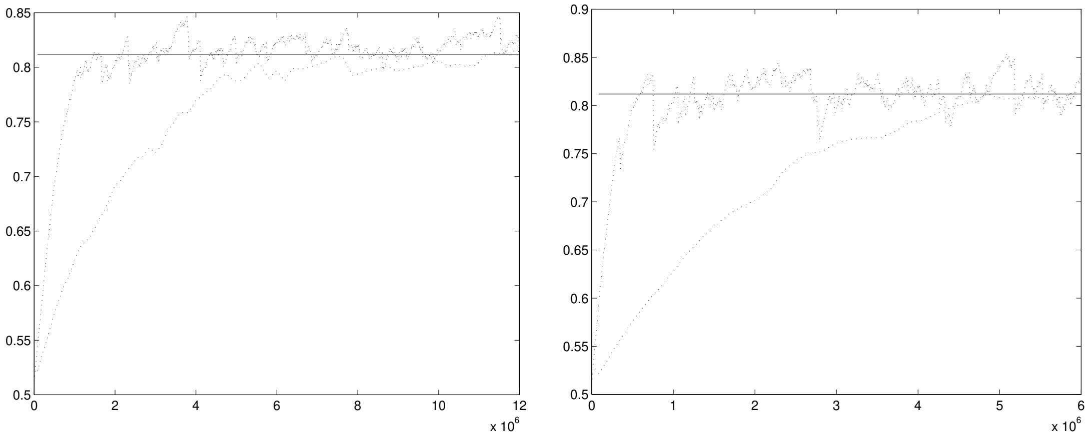

Fig. 10. Comparing Strong and Weak Convergence

7.5 Tracking Perturbations

In this section, to make things more interesting, we introduce a perturbation (the cost c_1 rises up to 18) at a fixed time, so that θ^* suddenly changes from 0.812 to 0.51. In order to compare the capability for adaptation, it seemed unfair to implement a version of (27) in which $M(n)$ is incremented. We used now Method 1 for the updates in (27) where $\epsilon_n = \epsilon_0/m(n)$ with $\epsilon_0 = 5 \times 10^{-4}$, $m(n)$ the integer part of $n/125$ and $M = 5,000$ busy periods (the resulting processes showed a noticeable asymptotic bias, similar to the one present in Figure 7, for smaller values of M). For the adaptive scheme (26), we used Method 3 with $M = 150$ and $5 \times \epsilon = 10^{-6}$. These parameters were chosen so that the non-reset estimator gives a close approximation to θ^*, effectively slowing it down. Even then, Figure 11 clearly shows that (27) cannot effectively track in real time, compared to (26).

We shall now present our tracking experiment using only the adaptive scheme with the non-reset estimators. The time taken to reach within a few percent of θ^* can appreciatively be cut down, if M is reduced and ϵ raised. Figure 12 to the left shows this speed improvement by a factor of 75, when we used $M = 100$ and $\epsilon = 5 \times 10^{-5}$. While convergence seems to slow down right after the perturbation when tracking the new θ^*, contrast this with the plot of the real running cost calculated as the simulation evolved, which is shown

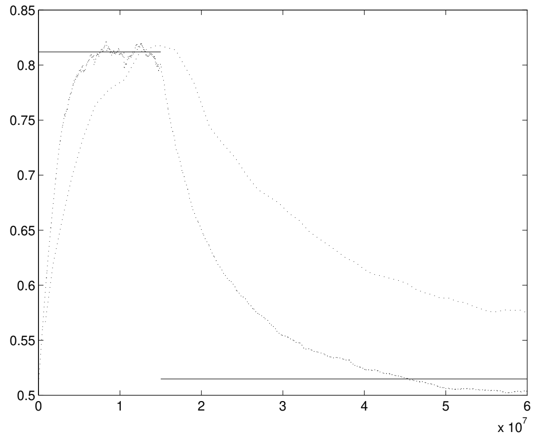

Fig. 11. Comparison for the Tracking Problem

to the right in Figure 12. In this plot, the actual cost was accumulated at each service and divided by the time to create the sample estimate of the cost function. The accumulation is naturally reset to zero at the moment of the perturbation. The solid lines represent the optimal cost values for each of the parameter settings.

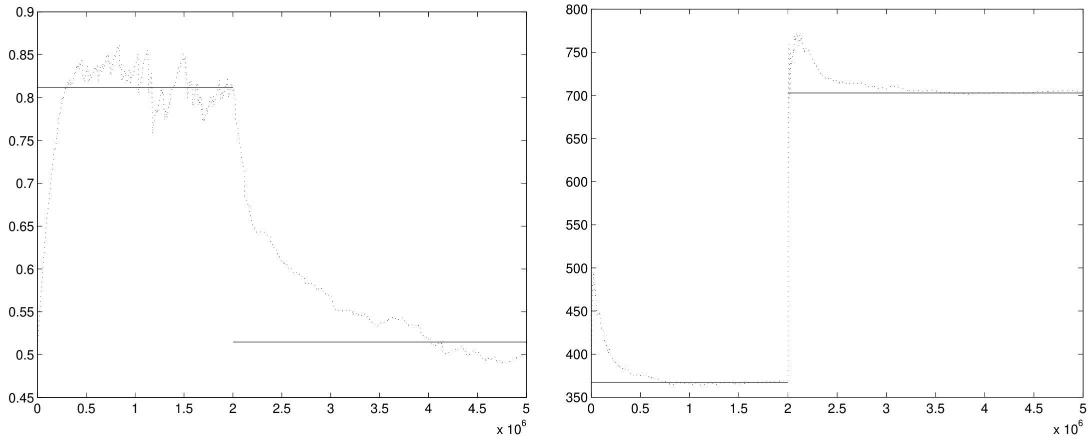

Fig. 12. Tracking Problem: Adaptive Control Process and Running Cost

As one can see, the cost stabilizes itself very fast at the optimal level after the perturbation.

8 Concluding Remarks

Our experiments have been designed to illustrate the strength of the ODE method for establishing convergence, as a tool for the design of intelligent automatic control. While the technical assumptions may be harder to verify than those of Theorem 1 with regenerative estimators, the method allows us to obtain more information. Indeed, it describes the trajectories of the control updates in the space of interpolated processes (as functions of time) and gives us the limiting behavior. We have described how Theorem 2 can be used to construct the limiting process, even if we have no analytical knowledge of the function $G(\cdot)$ in (5). We point out that while our example is purposely simple, it contains the basic features present for adaptive control of stochastic DES. Our estimation procedure via the disappearing phantoms is robust: there is no need to know the service distributions and the distribution of the arrival process. We have consciously avoided the problem of instabilities, using Θ as the whole feasible region. If the stable region is properly contained in the feasible set, then a projection has to be performed in order to avoid instabilities. We have used two types of projections. In the first case, we do nothing and do not update. In general, our estimators tend to "correct" the gradient estimation and promptly perform updates within the stable region. In the second case, we rescale the estimates, effectively reducing the value of ϵ for that update, so that the updated control stays within the stable region. Both schemes work in practice, theoretical stability results for the first case are on-going research.

Adaptability and accuracy have eluded a common solution thus far. *A posteriori*, looking at our experiments, it is rather obvious to try to decrease the learning rate as the optimum is approached, yet be able to reset it and make it larger when perturbations occur. Such changes may be slow rather than abrupt, also. In the deterministic case, a solution for this is to use the information on second derivatives, setting $\epsilon_n \approx [d^2F(\theta_n)/d\theta^2]^{-1}$. Second derivatives can also be estimated, as in [4], but the computational effort may be prohibitive in real time and the asymptotic bias may be hard to control at a given tolerance error, unless we slow down convergence to get accuracy. Other approaches may be more feasible in order to get the best of strong and weak convergence. We believe that the weak convergence method, with its capability for treating non-linear time scales may be useful in the study of adaptive learning rates, which is a challenging open research problem.

Appendix: Computer Algorithms

The computation of the phantom formulas can be presented in a recursive way. We present now these recursions, which are very similar to the way that our programs are coded. The algorithms for both the swapping and the disappearing phantoms can be summarized as follows. Fix k as the index of the gradient being estimated. Let $\beta(j)$ be the index of the customer that finishes simultaneously the busy period of the nominal and the j-th phantom system. Notice that both systems are dominated by the system with decisions $\bar{\eta}$. For the disappearing phantoms, $\bar{\eta} = \eta(0)$. For each $j \geq 0$, let $\mathbf{N}_m(j)$ be the number of customers in the successive BP's of the j-th system that finish within $\{i, \ldots, \beta(j)\}$. Clearly, $W_i(0) = W_i(j), i \leq j$, since j is the only customer phantomized in the j-th phantom system. Also, $W_i(0) = W_i(j), i > \beta(j)$, therefore:

$$\varphi_M(\eta(0)) - \varphi_M(\eta(j)) = \sum_{i=j+1}^{\max(M,\beta(j))} d_i(j,k)$$

where $d_i(j,k) = [W_i(0) - W_i(j)] I_k(j)$. Recall that $I_k(j) = \mathbf{1}_{\{\eta_j(0)=v_k\}}$. Analogously,

$$\varphi_{\mathbf{N}}(\eta(0)) - \varphi_{\mathbf{N}}(\eta(j)) = \sum_{i=j+1}^{\max(\mathbf{N}_1(0),\mathbf{N}_1(j))} d_i(j,k)$$

As explained in Section 6.2, the state has been enlarged to contain $W_{i+1}(0)$ and Z_i, where $Z_i(\theta) = (\Psi_i(k), i - \alpha_i, (d_i(j,k); j = \alpha_i, \ldots, i-1); k = 1 \ldots, V)$ and Ψ is given in (28). The quantity α_i keeps the index of the first customer starting the nominal BP where customer i belongs. We show the recursions for $d_i(j,k)$ for the cases where the initial perturbation $d_{j+1}(j,k)$ is positive in Figure 13 and in the case where it is negative in Figure 14.

Calculation for Finite Horizon: Given the state ξ_i, perform the following recursions to evaluate Z_{i+1}:

(1) $\alpha_{i+1} = \alpha_i \mathbf{1}_{\{W_{i+1}(0)>0\}} + (i+1)\mathbf{1}_{\{W_{i+1}(0)=0\}}$

(2) $d_{i+1}(i,k) = [S_i(0) - S_i(i)]\mathbf{1}_{\{\eta_i = v_k\}}$.

(3) For $j = i - \alpha_{i+1}, \cdots, i-1$ do:

(a) Set $k = \mathbf{1}_{\{S_j(0)=v_k\}}$.

(b) If $d_{j+1}(j,k) > 0$ then

$$d_{i+1}(j,k) = \begin{cases} d_i(j,k) & \text{if } W_{i+1}(0) > d_i(j,k) \\ W_{i+1}(0) & \text{otherwise} \end{cases}$$

(c) If $d_{j+1}(j,k) < 0$ then

$$d_{i+1}(j,k) = \begin{cases} d_i(j,k) & \text{if } W_{i+1}(0) > 0 \\ -W_{i+1}(j) & \text{otherwise} \end{cases}$$

where $W_{i+1}(j) = \max(0, (X_i - A_i) - d_i(j,k))$

(4) Set the partial accumulator to:

$$\Psi_{i+1}(k) = \Psi_i(k) + \sum_{j=\alpha_{i+1}}^{i+1} d_{i+1}(j,k)$$

(5) If $i+1$ is an update moment, use Ψ_{i+1} to build the sensitivity estimator and reset $\Psi_{i+1} = 0$.

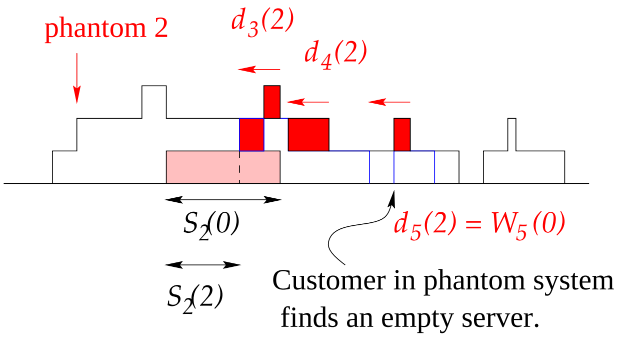

Fig. 13. Propagation of Postive Perturbation

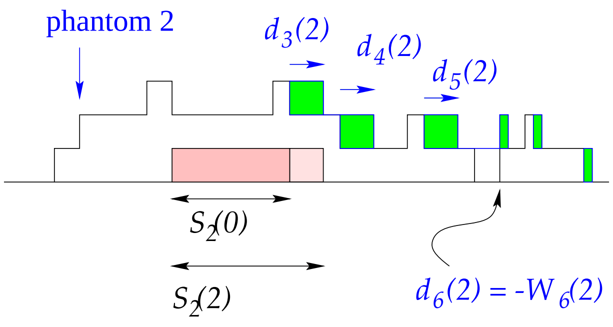

Fig. 14. Propagation of Negative Perturbation

Calculation for Regenerative: Introduce another boolean indicator, $I_i(j); j = 0, i - \alpha_i, \ldots, i-1$ into Z_i that records whether the current BP has finished or not in the j-th phantom system. Given the state ξ_i, perform the following recursions to evaluate Z_{i+1}:

(1) $\alpha_{i+1} = \alpha_i \mathbf{1}_{\{W_{i+1}(0)>0\}} + (i+1)\mathbf{1}_{\{W_{i+1}(0)=0\}}$

(2) $d_{i+1}(i,k) = [S_i(0) - S_i(i)]\mathbf{1}_{\{\eta_i = v_k\}}, \quad I_{i+1}(0) = 0.$

(3) For $i - \alpha_{i+1} \leq j < i$ do:

 (a) Set $k = \mathbf{1}_{\{S_j(0)=v_k\}}$.

 (b) If $d_{j+1}(j,k) > 0$ then

$$I_{i+1}(j) = \begin{cases} \mathbf{1}_{\{W_{i+1}(0)<d_i(j,k)\}} & \text{if } I_i(j) = 0 \\ 1 & \text{otherwise} \end{cases}$$

 and $d_{i+1}(j,k) = d_i(j,k)\, I_{i+1}(j) + W_{i+1}(0)(1 - I_{i+1}(0))$

 (c) If $d_{j+1}(j,k) < 0$ then

$$I_{i+1}(j) = \begin{cases} \mathbf{1}_{\{W_{i+1}(0)=0\}} & \text{if } I_i(j) = 0 \\ 1 & \text{otherwise} \end{cases}$$

 and

$$d_{i+1}(j,k) = d_i(j,k)\, I_{i+1}(j) - W_{i+1}(j)(1 - I_{i+1}),$$

 where $W_{i+1}(j) = \max(0, (X_i - A_i) - d_i(j,k))$

(4) Set the partial accumulator to:

$$\Psi_{i+1}(k) = \Psi_i(k) + \sum_{j=\alpha_{i+1}}^{i+1} d_{i+1}(j,k)$$

(5) If $i+1$ is an update moment, use Ψ_{i+1} to build the sensitivity estimator and reset $\Psi_{i+1} = 0$.

In our programs, we actually did not compute the $d_i(j)$ values, but added them to the partial accumulators in the loops directly, saving computation time and memory.

Acknowledgements

The author wishes to acknowledge the work of Kim Davis and Ianis Queval, who prepared most of the programming code.

References

[1] S. Andraddóttir, Optimization of the Transient and Steady-State Behavior of Discrete Event Systems, *Management Science*, **42** (1995) 717–737.

[2] A. Beneviste, M. Metivier et P. Priouret, *Algorithmes adaptatifs et aproximations stochastiques* (Masson, Paris, 1987).

[3] P. Billingsley, *Convergence of Probability Measures* (John Wiley, New York, 1968).

[4] P. Brémaud and F.J. Vázquez-Abad, On the Pathwise Computation of Derivatives with Respect to the Rate of a Point Process: The Phantom RPA Method, *Queueing Systems*, **10** (1992) 249–270.

[5] E. K. P. Chong and P. J. Ramadge, Optimization of Queues using an Infinitesimal Perturbation Analysis-based Stochastic Algorithm with General Update Times, *SIAM J. on Control and Optimization*, **31**(1993) 698-732.

[6] Y. Ermoliev and A. Gaivoronski, Stochastic quasigradient methods for optimization of discrete event systems, *Annals of Oper Res*, **39** (1992) 1-39.

[7] M. C. Fu, Convergence of the stochastic approximation algorithm for the GI/G/1 queue using infinitesimal perturbation analysis, *J. of Optimization Theory and Applications*, No. 65 (1990) 149–160.

[8] P. Glasserman, *Gradient Estimation via Perturbation Analysis* (Kluwer Academic Publishers, Boston, 1991).

[9] P. Glynn, Likelihood Ratio Gradient Estimation: An Overview, *Proceedings of the 1987 Winter Simulation Conference* (1987) 336-375.

[10] Y.C. Ho and X. Cao, *Perturbation Analysis of Discrete Event Dynamic Systems* (Kluwer Academic Publishers, Boston, 1991).

[11] Y.C. Ho and R. Sreenivas, Ordinal optimization of DEDS, *DEDS, Theory and Appl.* **2** (1992) 61-88.

[12] S. H. Jacobson and L. W. Schruben, A harmonic analysis approach to simulation sensitivity analysis, Technical report, Department of Industrial and Systems Engineering, Virginia Polytechnic Institute and State University(Blacksburg, Virginia, 1994).

[13] J. Kiefer and J. Wolfowitz, Stochastic Estimation of the Maximum of a Regression Function, *Annals of Mathematical Statistics*, 23 (1952) 462-466.

[14] L. Kleinrock, *Queueing Systems, Volume I* (John Wiley & Sons, New York, 1975).

[15] H.J. Kushner and D.S. Clark, *Stochastic Approximation for Constrained and Unconstrained Systems* (Springer Verlag, Berlin and New York, 1978).

[16] H.J. Kushner, *Approximation and Weak Convergence Methods for Random Processes with Applications to Stochastic System Theory* (MIT Press, Cambridge, MA, 1984).

[17] H.J. Kushner and A. Shwartz, Weak convergence and asymptotic properties of adaptive filters with constant gains, *IEEE Trans. on Information Theory*, **IT-30** (1984) 177–182.

[18] H.J. Kushner and F.J Vázquez-Abad, Stochastic Approximation Methods for Systems of Interest Over an Infinite Horizon,*SIAM J. on Control and Optimization*, **34**, No. 2 (1996) 712-756.

[19] H.J. Kushner and G. Yin, Stochastic approximation algorithms for parallel and distributed processing, *Stochastics*, **22** (1987) 219–250.

[20] H.J. Kushner and G. Yin, *Stochastic Approximation and Applications* (Springer Verlag, NewYork, 1997).

[21] P. L'Ecuyer and P.W. Glynn, Stochastic optimization by simulation: Convergence proofs for the GI/G/1 queue, *Management Sci.*, **vol 40** (1994) 1562–1578.

[22] L. Ljung, "Analysis of recursive stochastic algorithms", *IEEE Trans on Automatic Control.*, **AC-22** (1977) 551–575.

[23] A.M. Makowski and A. Shwartz, Stochastic Approximations and Adaptive Control of a Discrete-Time Single Server Network with random Routing, *SIAM J on Control and optim*, **30**, No. 6 (1992) 1476–1506.

[24] M. Metivier and P. Priouret, Applications of a Kushner and Clark Lemma to General Classes of Stochastic Algorithms, *IEEE Trans Infor Theory*, **IT30**, (1984) 140–150.

[25] Miyoshi, N. Weak Stationary Solution of a $G/G/1$ Queue Controlled by IPA-based SA with Constant Stepsize, *Porceddings of the 37th IEEE CDC* (1998), 1716–1721.

[26] G. Ch. Pflug, On line optimization of simulated Markov processes, *Math. of Operations Research* **15** (1990) 381–395.

[27] M. Reiman and A. Weiss, Sensitivity Analysis for Simulations via Likelihood Ratios, *Operations Research*, **37** (1989) 830-844.

[28] H. Robbins and S. Monro, A Stochastic Approximation Method, *Annals of Mathematical Statistics*, 22 (1951) 400–407.

[29] J. C. Spall, Multivariate stochastic approximation using a simultaneous perturbation gradient approximation, *IEEE Tans. on Automatic Control*, 37 (1992) 331–341.

[30] F.J. Vázquez-Abad, The Case of the Swapping and of the Disappearing Phantoms, invited paper *Proceedings of WODES98* (The Institute of Electrical Engineers, Great Brittan, 1998): 102–107. Presentation available at `http://www.ee.mu.oz.au/staff/fva/slides/S_wodes98.pdf`

[31] F.J. Vázquez-Abad and K. Davis, Efficient Implementation of the Phantom RPA Method, with an Application to a Priority Queueing System, Rapport Technique No. 989, Département d'iro, (Université de Montréal, 1995).

[32] F.J. Vázquez-Abad and C.C. Cassandras and V. Julka, Centralized and Decentralized Asynchronous Optimization of Stochastic Discrete Event Systems, *IEEE Trans on Automatic Control*, **vol 43**, No. 5 (1998) 631-655.

[33] F. Vázquez-Abad and H.J. Kushner, Estimation of the derivative of a stationary measure with respect to a control parameter, *Journal of Appl. Prob.*, 29 (1992) 343 - 352.

[34] F.J. Vázquez-Abad and L.G. Mason, Adaptive Decentralized Control Under Non-uniqueness of the Optimal Control, *DEDS, Theory and Applications*, **vol 6**, No. 4 (1996) 323-359.

[35] H. Yan and G. Yin and S.X.C. Lou, Using stochastic approximation to determine threshold values for control of unreliable manufacturing systems, *J. Optimization Theory and Applications*, 83 (1994) 511-539.