TECHNICAL RESEARCH REPORT

On Stochastic Approximations Driven by Sample Averages:
Convergence Results via the ODE Method

by J.D. Bartusek, A.M. Makowski

CSHCN T.R. 94-8
(ISR T.R. 94-4)

CENTER FOR SATELLITE AND HYBRID COMMUNICATION NETWORKS

ON STOCHASTIC APPROXIMATIONS DRIVEN BY SAMPLE AVERAGES: CONVERGENCE RESULTS VIA THE ODE METHOD

by

John D. Bartusek[1] and Armand M. Makowski[2]

ABSTRACT

We consider a class of algorithms which arise naturally in problems of on–line parametric optimization for discrete event dynamical systems, e.g., queueing systems and Petri net models. With Θ a closed convex subset of $\mathbb{R}^p$, the projected stochastic approximation algorithms of interest are of the form

$$\theta_0 \in \Theta, \quad \theta_{n+1} = \Pi_\Theta \left\{ \theta_n + a_{n+1} g\left(\theta_n, Y_{n+1}\right) \right\} \qquad n = 0, 1, \ldots$$

with

$$Y_{n+1} = \frac{1}{\ell_{n+1}} \sum_{\ell=1}^{\ell_{n+1}} f(\theta_n, X_{n+1,\ell}) \qquad n = 0, 1, \ldots$$

for non–random integer $\ell_{n+1} \uparrow \infty$, state process $\{X_{n+1,\ell},\ \ell = 1, \ldots\}$ taking values in some Borel subset E of $\mathbb{R}^s$, and Borel mappings $f : \Theta \times E \to \mathbb{R}^d$ and $g : \Theta \times \mathbb{R}^d \to \mathbb{R}^p$. The statistics of the sequence $\{X_{n+1,\ell},\ \ell = 1, \ldots\}$ are fully determined by the parameter value θ_n and the final state X_{n,ℓ_n} reached in the previous evaluation interval, and the standard conditions $a_n \downarrow 0$ and $\sum_{n=0}^{\infty} a_{n+1} = \infty$ are satisfied. We develop a general framework for investigating the a.s. convergence of the iterate sequence $\{\theta_n,\ n = 0, 1, \ldots\}$. In particular, we show how such convergence results can be obtained by means of the ordinary differential equation (ODE) method under a condition of exponential convergence for the random variables $\{g(\theta_n, Y_{n+1}) - h(\theta_n),\ n = 0, 1, \ldots\}$. We relate this condition of exponential convergence to uniform Large Deviations upper bounds for the collection of probability measures $\{\mathbf{P}_{\theta,x},\ \theta \in \Theta,\ x \in E\}$, with $\mathbf{P}_{\theta,x}$ denoting the probability measure on the set of system trajectories when starting in state x under parameter value θ. This upper bound is uniform in both the parameter θ and the initial condition x. To demonstrate the applicability of the results obtained herein, we specialize them to two specific classes of state processes, namely sequences of i.i.d. random variables and finite state time–homogeneous Markov chains. In both cases, we identify simple (and checkable) conditions that ensure the validity of a uniform Large Deviations upper bound.

---

[1] Electrical Engineering Department and Institute for Systems Research, University of Maryland, College Park, MD 20742.

[2] Electrical Engineering Department and Institute for Systems Research, University of Maryland, College Park, MD 20742. The work of this author was supported partially through NSF Grants NSFD CDR-88-03012

1. INTRODUCTION

In many contexts it is necessary to find parameter values $\theta^\star$ which satisfy a nonlinear equation of the form

$$h(\theta) = 0, \quad \theta \in \Theta \tag{1.1}$$

for some mapping $h : \Theta \to I\!R^p$ defined on a subset Θ of $I\!R^p$. A most typical example arises when minimizing a performance measure $J : \Theta \to I\!R$, a task often equivalent to setting the gradient of J to zero. At other times it is desirable to maintain system performance at some prespecified level $J^\star$, and this points to solving (1.1) with $h(\theta) = J(\theta) - J^\star$.

The overwhelming majority of methods for solving (1.1) are recursive in nature, and produce a sequence of iterates $\{\theta_n, \; n = 0, 1, \ldots\}$ which eventually converge to the desired value(s) $\theta^\star$: Starting with an initial guess θ_0, the $(n+1)^{rst}$ iterate θ_{n+1} is computed on the basis of the previous iterate θ_n and of past values of h, say $h(\theta_i)$, $i = 1, 2, \ldots, n$, (and in some cases, of derivatives of h at these points).

Unfortunately, it is often the case that h is not directly available, because either its functional form is unknown, or its evaluation is computationally prohibitve. To remedy this difficulty, Robbins and Monro [24] proposed the class of algorithms called stochastic approximations. In their simplest form, such algorithms deal with the unconstrained case (i.e., $\Theta = I\!R^p$) and produce a sequence of iterates $\{\theta_n, \; n = 0, 1, \ldots\}$ through the recursion

$$\theta_0 \in I\!R^p, \quad \theta_{n+1} = \theta_n + a_{n+1} Z_{n+1} \qquad n = 0, 1, \ldots (1.2)$$

for some $I\!R^p$-valued "driving" process $\{Z_n, \; n = 0, 1, \ldots\}$ and sequence of stepsizes $\{a_{n+1}, \; n = 0, 1, \ldots\}$ which satisfy the standard conditions $a_n \downarrow 0$ and $\sum_{n=0}^{\infty} a_{n+1} = \infty$.

It is customary to view Z_{n+1} as an approximation to $h(\theta_n)$. In their original paper, Robbins and Monro generated the random variables (rvs) $\{Z_n, \; n = 0, 1, \ldots\}$ according to

$$\mathbf{P}[Z_{n+1} \in B | Z_0, \ldots, Z_n] = \mu_{\theta_n}(B) \qquad n = 0, 1, \ldots (1.3)$$

for some family of probability measures $\{\mu_\theta, \; \theta \in I\!R^p\}$ on $I\!R^p$ such that

$$h(\theta) = \int_{R^p} z \mu_\theta(dz), \quad \theta \in \Theta. \tag{1.4}$$

The key issue in the study of algorithms such as (1.2) (and variations thereof) is concerned with the convergence of the iterates $\{\theta_n, \; n = 0, 1, \ldots\}$ to the desired value $\theta^\star$ in some mode of convergence.

Over the years, increasingly more complex applications have lead to the use of *projected* versions of the stochastic approximation scheme (1.2) which take the form

$$\theta_0 \in \Theta, \quad \theta_{n+1} = \Pi_\Theta \{\theta_n + a_{n+1} Z_{n+1}\} \qquad n = 0, 1, \ldots (1.5)$$

with Π_Θ denoting nearest-point projection on Θ. It also became necessary to consider versions of (1.5) which are driven by processes $\{Z_n,\ n = 0,1,\ldots\}$ with a more general statistical structure than (1.3). For instance, several authors [17,20,22,28] have considered both (1.2) and (1.5) when

$$Z_{n+1} = g(\theta_n, Y_{n+1}) \qquad n = 0,1,\ldots (1.6)$$

for some Borel mapping $g : \Theta \times I\!R^d \to I\!R^p$ and $I\!R^d$-valued process $\{Y_n,\ n = 0,1,\ldots\}$ which is Markov in the sense that

$$\mathbf{P}[Y_{n+1} \in B | \theta_i, Y_i, i = 0,1,\ldots,n] = \int_B K(\theta_n; Y_n, dy) \qquad n = 0,1,\ldots$$

for some family of transition kernels $\{K(\theta; x, dy),\ \theta \in \Theta,\ x \in I\!R^d\}$ on $I\!R^d$.

In this paper, we are concerned with the convergence properties of yet another class of (projected) stochastic approximations which arise naturally (but not exclusively) in problems of on-line parametric optimization for discrete event dynamical systems, e.g., queueing systems and Petri net models [1]. These algorithms are driven by *sample averages* defined on well-structured state processes and operate at two different times scales, with state transitions occuring more frequently than parameter updates. For *non*-random integers $\{\ell_{n+1}, n = 0,1,\ldots\}$, the stochastic approximations of interest are of the form (1.5)–(1.6) with

$$Y_{n+1} = \frac{1}{\ell_{n+1}} \sum_{\ell=1}^{\ell_{n+1}} f(\theta_n, X_{n+1,\ell}) \qquad n = 0,1,\ldots (1.7)$$

for a state process $\{X_{n+1,\ell},\ \ell = 1,\ldots\}$ taking values in a Borel subset E of $I\!R^s$ and Borel mapping $f : \Theta \times E \to I\!R^d$. In words, with iterate θ_n just returned by the algorithm, we observe or simulate the state process with the understanding that the statistics of the sequence $\{X_{n+1,\ell},\ \ell = 1,\ldots\}$ are fully determined by the parameter value θ_n and the final state reached in the previous evaluation interval, i.e., X_{n,ℓ_n}. After ℓ_{n+1} transitions, the sample average (1.7) is computed, and the algorithmic step is then completed by returning iterate θ_{n+1} according to (1.5)–(1.6).

Whenever such algorithms arise, we can invariably write $h(\theta) = g(\theta, F(\theta))$ for some *known* mapping $g : \Theta \times I\!R^d \to I\!R^p$ and some quantity $F(\theta)$ which is obtainable only through observation or simulation of the state process at operating point θ. Fortunately, it is often the case that

$$F(\theta) = \lim_{L \uparrow \infty} \frac{1}{L} \sum_{\ell=1}^{L} f(\theta, \xi_\ell) \quad \mathbf{P}_{\theta,x} - a.s. \qquad (1.8)$$

where $\{\xi_\ell,\ \ell = 0,1,\ldots\}$ is a generic E-valued random sequence modelling the time evolution of the system, and $\mathbf{P}_{\theta,x}$ denotes the probability measure on the set of system trajectories when starting in state x under parameter value θ. This suggests that for ℓ_{n+1} large, under appropriate conditions on g, the rvs Y_{n+1} (given by (1.7)) and $g(\theta_n, Y_{n+1})$ can be viewed as good approximations to $F(\theta_n)$ and $h(\theta_n)$, respectively. Therefore, if the deterministic algorithm

$$\theta_0 \in \Theta, \quad \theta_{n+1} = \Pi_\Theta \{\theta_n + a_{n+1} h(\theta_n)\} \qquad n = 0, 1, \ldots (1.9)$$

converges to some $\theta^\star$, then we should expect the stochastic version (1.5)–(1.7) to also converge, say almost surely, to the same point as the size of the sampling window grows unbounded.

Instances of algorithms (1.5)–(1.7) have appeared previously in the literature: In [21], for a tandem pair of $M/M/1$ queues, Meketon presents a level crossing application where the objective is to steer the steady state number of customers in system to a nominal value. A simple iterative scheme is proposed for tuning the service rate at the first server. The resulting stochastic approximation is driven by a sample average estimate constructed over a finite observation window; encouraging experimental results are presented but no proof of convergence is offered. Another important application occurs when optimizing the long–run average performance functional associated with a parametrized Generalized Semi–Markov Process (GSMP) [1,11,13]. In that context, $g(\theta_n, Y_{n+1})$ is an approximation to the gradient of $J(\theta)$ at $\theta = \theta_n$, and several methods are now available for generating such approximations, the two most often used methods being Infinitesimal Perturbation Analysis (IPA) [11,13,27] and the Likelihood Ratio (LR) method [1,12,23]. In many instances the gradient estimates are of the form (1.6)–(1.7) for some appropriately chosen process [11, Chap. 8], and (1.8) then simply expresses the strong consistency of the estimates.

Before describing the contributions of this paper, we briefly review some of the existing results on stochastic approximations driven by sample averages. In [29], with a non–standard notion of convergence, Wardi showed convergence for such an algorithm which seeks to minimize $J(\theta) \equiv f(W(\theta))$ for continuously differentiable function $f : \mathbb{R} \to \mathbb{R}$ and mean waiting time $W(\theta)$ in $GI/GI/1$ queues. Chong and Ramadge [2], and Fu [8] consider the a.s. convergence of a class of IPA–based stochastic gradient algorithms which update only at the completion epochs of busy periods in a single server queue. In these two studies, the number of state transitions ℓ_{n+1} is *random* and determined by the underlying process, rather than prespecified as in Wardi's and herein. As is well known [1,3], stochastic algorithms which tie updates to the regenerative cycles of the underlying state process can have undesirable properties, e.g., slow convergence, if the regenerative cycles tend to be long (as would be the case in queueing systems operating near critical load). In the $GI/GI/1$ context, Chong and Ramadge [3] address this issue by allowing updates after every transition of the state process. Dupuis and Simha [6] studied similar schemes with a prespecified sequence $\{\ell_{n+1}, n = 0, 1, \ldots\}$ tending to infinity when the rvs $\{X_{n+1,\ell},\ \ell = 1, \ldots\}$ are i.i.d. and the stepsizes are *constant*, i.e., $a_{n+1} \equiv a$, $n = 0, 1, \ldots$

In this paper we develop a *general framework* for investigating the a.s. convergence of the iterate sequence $\{\theta_n,\ n = 0, 1, \ldots\}$ generated by (1.5)–(1.7). We start essentially with no structural assumptions on the probability measures $\{\mathbf{P}_{\theta,x},\ \theta \in \Theta,\ x \in E\}$ governing the statistical behavior of the state process; it is only assumed that a law of large numbers such as (1.8) is in effect. Our focus is on charting a sequence of basic steps to help establish a.s. convergence; these steps point to a set of technical conditions that need to be verified in each specific application.

Our framework relies on the ordinary differential equation (ODE) method [16,22], which in most of its forms proceeds in two separate steps. The first step relies on the Kushner–Clark Lemma [16, Thm. 5.3.1, p. 191] to identify a deterministic ODE, the stability properties of which determine the limit points of $\{\theta_n,\ n=0,1,\ldots\}$. The second step, which is probabilistic in nature and depends on the algorithm, involves showing that asymptotically (in the mode of convergence of interest) the output sequence of the original algorithm behaves like the solution to the ODE. Although general conditions are given in [16] for successfully completing this last step, these conditions are not readily checkable in terms of the model data. Nevertheless, we first show here that this second step is determined by the *exponential* convergence to zero of the rvs $\{g(\theta_n, Y_{n+1}) - h(\theta_n),\ n=0,1,\ldots\}$, i.e., for every $\varepsilon > 0$, the convergence

$$\lim_{n\to\infty} \mathbf{P}[||g(\theta_n, Y_{n+1}) - h(\theta_n)|| \geq \varepsilon] = 0 \tag{1.10}$$

takes place exponentially fast (with respect to the sequence of sample durations $\{\ell_{n+1},\ n = 0,1,\ldots\}$). This exponential convergence viewpoint was already implicit in the work of Dupuis and Simha [6].

Going one step further, we give more explicit conditions to ensure this exponential convergence. As in [6], we do so by invoking a *uniform* Large Deviations upper bound for the collection of probability measures $\{\mathbf{P}_{\theta,x},\ \theta\in\Theta,\ x\in E\}$. Here, this upper bound is uniform in both the parameter θ *and* the initial condition x, and with some functional $I : I\!R^d \to [0,\infty]$, takes the form

$$\limsup_{L\to\infty} \frac{1}{L}\log \sup_{\theta\in\Theta, x\in E} \mathbf{P}_{\theta,x}[\frac{1}{L}\sum_{\ell=1}^{L} f(\theta,\xi_\ell) - F(\theta)\in C] \leq -I(C) \tag{1.11}$$

for every closed subset C of $I\!R^d$. We are able to provide checkable conditions to ensure that (1.11) holds. The approach for doing this is in the spirit of the Gärtner–Ellis Theorem [7, Thm. II.2., p. 3]; in fact, we broaden the applicability of the ideas of Dupuis and Simha to very general classes of dependent state processes.

To demonstrate the applicability of the results obtained herein, we specialize them to two specific classes of state processes. In the first class, the successive states form a sequence of i.i.d. rvs as in [6]. In the second class, the state sequence is a finite state time–homogeneous Markov chain; this is an important class of processes which is often used in applications. In both cases, we identify simple (and checkable) conditions that ensure the validity of a uniform Large Deviations upper bound.

The paper is organized as follows: In Section 2 we introduce the basic building blocks that we use in Section 3 to formally define the class of stochastic approximations investigated here. The basic convergence result is stated as Theorem 1 in Section 4. Next, exponential convergence is shown in Section 5 to be the key condition for establishing a.s. convergence via the ODE method. In turn, this condition of exponential convergence is related in Section 6 to the existence of a uniform large deviations upper bounds. Conditions to ensure such uniform large deviations upper bounds are derived in Section 7. Several specific situations are treated in Sections 8 and 9, namely, the cases where the process driving the sample averages is modelled by a sequence of i.i.d. rvs and by a finite–state Markov chain; in all cases, we give concrete conditions for uniform large deviations upper bounds to exist. We close with Section 10 where the results obtained so far are reviewed, and various extensions and open problems are discussed. Two technical results have been relegated to the appendices.

Finally, a few words on the notation used throughout the paper: The set of all real (resp. non–negative real) numbers is denoted by $\mathbb{R}$ (resp. $\mathbb{R}_+$). For any set E endowed with a topology, measurability is always taken to mean Borel measurability and the corresponding Borel σ–field, i.e., the smallest σ–field on E generated by the open sets of the topology, is denoted by $\mathcal{B}(E)$. Moreover, the infimum over an empty set is taken to be ∞ by convention. An element v of $\mathbb{R}^n$ is always understood as a column vector and its transpose is denoted by v'. For elements v and w of some $\mathbb{R}^n$, we write $\langle v, w\rangle$ for their usual scalar product, so that $\|v\| \equiv \sqrt{\langle v, v\rangle}$ denotes the Euclidean norm of v; the dependence on the dimension will always be omitted from the notation as it is clear from the context.

2. THE BASIC INGREDIENTS

Before defining the stochastic approximation procedures considered here, we devote this section to introducing the basic building blocks used in the formal definitions of Section 3: Throughout the discussion, p, s and d are fixed positive integers. We assume given a closed convex subset Θ of $\mathbb{R}^p$, and a Borel subset E of $\mathbb{R}^s$. Furthermore, let $f : \Theta \times E \to \mathbb{R}^d$ and $g : \Theta \times \mathbb{R}^d \to \mathbb{R}^p$ denote fixed Borel mappings. Additional assumptions will be imposed in due time.

We consider two sequences $\{a_{n+1},\ n = 0, 1, \ldots\}$ and $\{\ell_{n+1},\ n = 0, 1, \ldots\}$ which take values in $\mathbb{R}_+$ and $\mathbb{N}$, respectively. The following assumptions are enforced:

(A) The $\mathbb{R}_+$–valued sequence $\{a_{n+1},\ n = 0, 1, \ldots\}$ is monotone decreasing with $a_n \downarrow 0$ ($n \uparrow \infty$), under the usual divergence condition $\sum_{n=0}^{\infty} a_{n+1} = \infty$.

(L) The $\mathbb{N}$–valued sequence $\{\ell_{n+1},\ n = 0, 1, \ldots\}$ is monotone increasing and for all $\beta > 0$ satisfies the condition

$$\sum_{n=0}^{\infty} \exp(-\beta \ell_{n+1}) < \infty. \tag{2.1}$$

Condition **(L)** implies $\ell_n \uparrow \infty$ as $n \to \infty$ but the reverse implication is not always true: Indeed, in the case $\ell_n = \lceil \log n \rceil$, $n = 1, 2, \ldots$, we see that (2.1) fails for $0 < \beta \leq 1$ since then $\sum n^{-\beta} = \infty$.

Let E^∞ be the infinite cartesian product of E with itself, and denote by $\mathcal{B}(E^\infty)$ the standard σ–field on E^∞. We write a generic element ξ of E^∞ as $\xi = (x, x_1, \ldots)$ where $x, x_1, \ldots$ are all elements of E. The coordinate process $\{\xi_\ell,\ \ell = 0, 1, \ldots\}$ is then simply defined by

$$\xi_0(\xi) \equiv x, \quad \xi_\ell(\xi) \equiv x_\ell, \quad \xi \in E^\infty. \qquad \ell = 1, \ldots (2.2)$$

We postulate the existence of a family $\{\mathbf{P}_{\theta,x},\ \theta \in \Theta,\ x \in E\}$ of probability measures on $\mathcal{B}(E^\infty)$ such that

$$\mathbf{P}_{\theta,x}[\xi_0 = x] = 1, \quad \theta \in \Theta,\ x \in E. \tag{2.3}$$

For technical reasons, we need to assume a measurable functional dependence in θ and x:

(P1) For every $L = 1, 2, \ldots$, the mapping $\Theta \times E \to I\!R : (\theta, x) \to \mathbf{P}_{\theta,x}[\xi_\ell \in B_\ell,\ \ell = 1, \ldots, L]$ is Borel measurable for all possible choices of Borel subsets $B_1, \ldots, B_L$ in $\mathcal{B}(E)$.

We also assume that a strong law of large numbers is in effect:

(P2) There exists a Borel mapping $F : \Theta \to I\!R^d$ such that for all θ in Θ and x in E, we have

$$\lim_{L \uparrow \infty} \frac{1}{L} \sum_{\ell=1}^{L} f(\theta, \xi_\ell) = F(\theta) \quad \mathbf{P}_{\theta,x}-a.s. \tag{2.4}$$

3. MODEL AND ASSUMPTIONS

In order to define the stochastic approximation procedures, we start with a sample space Ω equipped with a σ–field of events $\mathcal{F}$. The measurable space $(\Omega, \mathcal{F})$ is assumed large enough to carry a double array of E–valued rvs $\{X_{n,\ell},\ \ell = 1, \ldots, \ell_n;\ n = 0, 1, \ldots\}$ where we use the convention $\ell_0 = 1$. We define the Θ–valued rvs $\{\theta_n,\ n = 0, 1, \ldots\}$ through the recursion

$$\theta_0 \in \Theta, \quad \theta_{n+1} = \Pi_\Theta\left\{\theta_n + a_{n+1} g(\theta_n, Y_{n+1})\right\} \qquad n = 0, 1, \ldots (3.1)$$

where we use the notation

$$Y_{n+1} \equiv \frac{1}{l_{n+1}} \sum_{\ell=1}^{\ell_{n+1}} f(\theta_n, X_{n+1,\ell}). \qquad n = 0, 1, \ldots (3.2)$$

In (3.1), Π_Θ denotes the nearest-point projection operator on the set Θ; it is well defined since Θ is assumed closed and convex.

Next we introduce the filtration $\{\mathcal{F}_n,\ n = 0, 1, \ldots\}$ on $(\Omega, \mathcal{F})$ by setting

$$\begin{aligned}\mathcal{F}_n \equiv& \sigma\{\theta_m, X_{m,\ell},\ \ell = 1, \ldots, \ell_m,\ m = 0, 1, \ldots, n\} \\ =& \sigma\{\theta_0; X_{m,\ell},\ \ell = 1, \ldots, \ell_m,\ m = 0, 1, \ldots, n\}\end{aligned} \qquad n = 0, 1, \ldots (3.3)$$

where the equality follows from the fact that the rvs θ_m, $m = 1, 2, \ldots, n$, are fully determined by the rvs θ_0, $X_{0,1}$, and $X_{m+1,\ell}$, $\ell = 0, 1, \ldots, \ell_{m+1}$, $m = 1, \ldots, n-1$.

Finally, given a probability measure ν on $\mathcal{B}(\Theta \times E)$, we postulate the existence of a probability measure $\mathbf{P}$ on $(\Omega, \mathcal{F})$ satisfying

$$\mathbf{P}[\theta_0 \in B, X_{0,1} \in B_1] = \nu(B \times B_1), \quad B \in \mathcal{B}(\Theta),\ B_1 \in \mathcal{B}(E) \tag{3.4}$$

and

$$\mathbf{P}[X_{n+1,\ell} \in B_\ell,\ \ell = 1,\ldots,\ell_{n+1} | \mathcal{F}_n] = \mathbf{P}_{\theta_n, X_{n,\ell_n}}[\xi_\ell \in B_\ell,\ \ell = 1,\ldots,\ell_{n+1}],$$

$$B_\ell \in \mathcal{B}(E),\ \ell = 1,\ldots,\ell_{n+1},\ n = 1,\ldots \quad (3.5)$$

Note that **(P1)** is needed in order to make sense of this last requirement. The existence of such a set-up is readily justified by invoking the Daniell–Kolmogorov consistency theorem [19, p. 94] on $\Theta \times E \times E^\infty$ in the usual manner.

4. THE CONVERGENCE RESULTS

The presentation of the main convergence result is simplified by the introduction of the following notation: Setting

$$h(\theta) \equiv g(\theta, F(\theta)), \quad \theta \in \Theta \tag{4.1}$$

we define the $\mathbb{R}^p$–valued rvs $\{\gamma_{n+1},\ n = 0,1,\ldots\}$ by

$$\gamma_{n+1} \equiv g(\theta_n, Y_{n+1}) - h(\theta_n) \qquad n = 0,1,\ldots \quad (4.2)$$

so that the recursion (3.1) now becomes

$$\theta_0 \in \Theta, \quad \theta_{n+1} = \Pi_\Theta\left\{\theta_n + a_{n+1}h(\theta_n) + a_{n+1}\gamma_{n+1}\right\}. \qquad n = 0,1,\ldots \quad (4.3)$$

The relevant assumptions concerning these quantities are the following:

(H) The mapping $h : \Theta \to \mathbb{R}^p$ is continuous.

(E) The $\mathbb{R}^p$–valued rvs $\{\gamma_{n+1},\ n = 0,1,\ldots\}$ converge exponentially to the zero vector, in the sense that for every $\varepsilon > 0$, there exist a finite integer $n(\varepsilon)$ and a positive constant $K(\varepsilon)$ such that

$$\mathbf{P}[\|\gamma_{n+1}\| \geq \varepsilon] \leq \exp\left(-\ell_{n+1}K(\varepsilon)\right), \quad n \geq n(\varepsilon). \tag{4.4}$$

Sufficient conditions for **(E)** are provided in Section 6 and follow from the availability of uniform large deviations upper bounds.

With the projection operator Π_Θ, we associate the transformation $\bar{\Pi}_\Theta : \Theta \times \mathbb{R}^p \to \mathbb{R}^p$ given by

$$\bar{\Pi}_\Theta(\theta, v) \equiv \lim_{\Delta \downarrow 0} \frac{\Pi_\Theta\{\theta + \Delta v\} - \theta}{\Delta}, \quad \theta \in \Theta,\ v \in \mathbb{R}^p. \tag{4.5}$$

The limiting ODE corresponding to (4.3) is

$$\theta(0) \in \Theta, \quad \frac{d\theta}{dt}(t) = \bar{\Pi}_\Theta[\theta(t), h(\theta(t))], \quad t \geq 0. \tag{4.6}$$

The unconstrained case corresponds to $\Theta \equiv \mathbb{R}^p$, in which case the recursion (3.1) reduces to

$$\theta_0 \in \Theta, \quad \theta_{n+1} = \theta_n + a_{n+1}h(\theta_n) + a_{n+1}\gamma_{n+1} \qquad n = 0,1,\ldots \quad (4.7)$$

and the limiting ODE corresponding to (4.7) becomes

$$\theta(0) \in \mathbb{R}^p, \quad \frac{d\theta}{dt}(t) = h(\theta(t)), \quad t \geq 0. \tag{4.8}$$

The basic convergence result for the scheme (3.1) is contained in Theorem 1; a proof is given in the next section.

Theorem 1. *Consider the stochastic approximation scheme (4.3) under assumptions* **(A)**, **(L)**, **(P1)–(P2)**, **(H)** *and* **(E)**. *Let $\theta^\star$ be a point in the interior Θ^o which is a locally asymptotically stable solution to (4.6), and let $\mathcal{A}(\theta^\star)$ denote its domain of attraction. Assume the following conditions hold:*

(i): The $\mathbb{R}^p$–valued random variables $\{\theta_n, n = 0, 1, \ldots\}$ are bounded with probability one, i.e.

$$\mathbf{P}[\sup_n ||\theta_n|| < \infty] = 1; \tag{4.9}$$

(ii): There exists a compact set $K \subseteq \mathcal{A}(\theta^\star)$ such that

$$\mathbf{P}[\theta_n \in K \ i.o.]. \tag{4.10}$$

Then $\lim_{n\to\infty} \theta_n = \theta^\star \quad \mathbf{P}-a.s.$

In some cases, it can be difficult to validate conditions (4.9) and (4.10). There is, however, one situation which naturally occurs in practice where (4.9) is automatically satisfied, namely when Θ is a *compact* subset of $\mathbb{R}^p$. Furthermore, (4.10) is automatically satisfied when $\mathcal{A}(\theta^\star) = \Theta$.

5. A PROOF OF THEOREM 1

Given the gain sequence $\{a_{n+1},\ n = 0, 1, \ldots\}$, we define the sequence of times $\{t_n,\ n = 0, 1, \ldots\}$ by

$$t_0 \equiv 0, \quad t_{n+1} \equiv \sum_{i=0}^{n} a_{i+1}, \qquad n = 0, 1, \ldots \tag{5.1}$$

and set

$$m(t) \equiv \max\{n \in \mathbb{N} : t_n \leq t\}, \quad t \geq 0. \tag{5.2}$$

Theorem 1 is a simple consequence of the so-called ODE method as developed by Kushner and Clark [16, Thm. 5.3.1, p. 191] once we observe the following lemma.

Lemma 3. *Assume condition* **(A)**, **(L)** *and* **(E)** *to be enforced. For every $T > 0$ and $\varepsilon > 0$, we have*

$$\lim_{n\to\infty} \mathbf{P}\left[\sup_{j\geq n} \max_{0\leq t\leq T} \left\| \sum_{i=m(jT)}^{m(jT+t)-1} a_{i+1}\gamma_{i+1} \right\| \geq \varepsilon\right] = 0. \tag{5.3}$$

Proof. Fix $T > 0$ and $\varepsilon > 0$. We readily observe that

$$\begin{aligned}
&\mathbf{P}\left[\sup_{j\geq n}\max_{0\leq t\leq T}\left\|\sum_{i=m(jT)}^{m(jT+t)-1} a_{i+1}\gamma_{i+1}\right\|\geq\varepsilon\right] \\
&\leq \mathbf{P}\left[\sup_{j\geq n}\max_{0\leq t\leq T}\sum_{i=m(jT)}^{m(jT+t)-1} a_{i+1}\|\gamma_{i+1}\|\geq\varepsilon\right] \\
&\leq \mathbf{P}\left[\sup_{j\geq n}\sum_{i=m(jT)}^{m(jT+T)-1} a_{i+1}\,\|\gamma_{i+1}\|\geq\varepsilon\right] \\
&\leq \sum_{j=n}^{\infty}\mathbf{P}\left[\sum_{i=m(jT)}^{m(jT+T)-1} a_{i+1}\,\|\gamma_{i+1}\|\geq\varepsilon\right] \\
&\leq \sum_{j=n}^{\infty}\mathbf{P}\left[\max_{m(jT)\leq i<m(jT+T)}\|\gamma_{i+1}\|\cdot\sum_{i=m(jT)}^{m(jT+T)-1} a_{i+1}\geq\varepsilon\right]. \qquad (5.4)
\end{aligned}$$

It is plain from (5.1)–(5.2) that $t_{m(jT+T)} \leq jT + T$ and $t_{m(jT)+1} = t_{m(jT)} + a_{m(jT)+1} \geq jT$ for all $j = 0, 1, \ldots$. Therefore, we have

$$\begin{aligned}
\sum_{i=m(jT)}^{m(jT+T)-1} a_{i+1} &= t_{m(jT+T)} - t_{m(jT)} \\
&\leq (jT+T) - (jT - a_{m(jT)+1}), \\
&\leq T + a_1 \qquad j = 0, 1, \ldots (5.5)
\end{aligned}$$

since the gain sequence $\{a_{n+1},\ n = 0, 1, \ldots\}$ is monotone decreasing. Combining (5.4) and (5.5), with $\varepsilon' = \frac{\varepsilon}{T+a_1}$, we get

$$\begin{aligned}
\sum_{j=n}^{\infty}\mathbf{P}\left[\max_{m(jT)\leq i<m(jT+T)}\|\gamma_{i+1}\|\cdot\sum_{i=m(jT)}^{m(jT+T)-1} a_{i+1}\geq\varepsilon\right] &\leq \sum_{j=n}^{\infty}\mathbf{P}\left[\max_{m(jT)\leq i<m(jT+T)}\|\gamma_{i+1}\|\geq\varepsilon'\right] \\
&\leq \sum_{j=n}^{\infty}\sum_{i=m(jT)}^{m(jT+T)-1}\mathbf{P}\left[\|\gamma_{i+1}\|\geq\varepsilon'\right] \\
&= \sum_{i=m(nT)}^{\infty}\mathbf{P}\left[\|\gamma_{i+1}\|\geq\varepsilon'\right]. \qquad (5.6)
\end{aligned}$$

Next, upon invoking the exponential convergence condition **(E)**, we can assert the existence of a finite integer $n(\varepsilon')$ and of a positive constant $K(\varepsilon')$ such that

$$\mathbf{P}[\|\gamma_{i+1}\|\geq\varepsilon'] \leq \exp\left(-\ell_{i+1}K(\varepsilon')\right), \quad i \geq n(\varepsilon'). \qquad (5.7)$$

Finally we select a finite integer $n^\star$ such that $m(n^\star T) \geq n(\varepsilon')$; such a selection is always possible since $\lim_{n\uparrow\infty} m(nT) = \infty$ by virtue of (**A**). For all $n \geq n^\star$, we easily conclude from (5.6) and (5.7) that

$$\mathbf{P}\left[\sup_{j\geq n}\max_{0\leq t\leq T}\left\|\sum_{i=m(jT)}^{m(jT+t)-1} a_{i+1}\gamma_{i+1}\right\| \geq \varepsilon\right] \leq \sum_{i=m(nT)}^{\infty} \exp\left(-\ell_{i+1}K(\varepsilon')\right) \tag{5.8}$$

and the convergence (5.3) is now an immediate consequence of (5.8) and of the summability condition (**L**) since $\lim_{n\uparrow\infty} m(nT) = \infty$. ■

6. SUFFICIENT CONDITIONS FOR (E)

A sufficient condition for (**E**) can be derived from uniform Large Deviations upper bounds as we now show. First a few definitions: With the coordinate process $\{\xi_\ell,\ \ell = 0,1,\ldots\}$ defined on the measurable space $(E^\infty, \mathcal{B}(E^\infty))$, we write

$$\bar{S}_L(\theta) \equiv \frac{1}{L}\sum_{\ell=1}^{L} f(\theta,\xi_\ell), \quad \theta \in \Theta. \qquad L = 1,\ldots \tag{6.1}$$

Since condition (**P2**) can be rephrased as $\lim_{L\to\infty} \bar{S}_L(\theta) = F(\theta)$ $\mathbf{P}_{\theta,x}$–a.s., the rate of convergence implied by (**E**) thus suggests that the law of large numbers associated with the sample averages (6.1) be complemented by a Large Deviations upper bound. This is essentially the content of condition (**U1**):

(**U1**) The collection of probability measures $\{\mathbf{P}_{\theta,x},\ \theta \in \Theta,\ x \in E\}$ satisfies a *uniform* Large Deviations upper bound principle with respect to (the sample averages associated with) f if there exists a closed convex function $I : \mathbb{R}^d \to [0,\infty]$ such that

$$\limsup_{L\to\infty} \frac{1}{L}\log \sup_{\theta\in\Theta, x\in E} \mathbf{P}_{\theta,x}[\bar{S}_L(\theta) - F(\theta) \in C] \leq -\inf_{z\in C} I(z) \tag{6.2}$$

for every closed subset C of $\mathbb{R}^d$.

We refer to I as the rate functional associated with this uniform Large Deviations upper bound principle. By itself, condition (**U1**) is not sufficient for (**E**), so we supplement (**U1**) by imposing additional conditions (**U2**)–(**U3**) on the rate functional I:

(**U2**) The rate function I in (**U1**) is level compact, i.e., the set $\{z \in \mathbb{R}^d : I(z) \leq r\}$ is compact for all $r \geq 0$; and

(**U3**) The rate function I in (**U1**) has the property that $I(z) = 0$ if and only if $z = 0$.

In a brief but necessary interlude, we pause to establish the following consequence of (**U1**)–(**U3**).

Lemma 4. *Assume* (**U2**)–(**U3**) *to hold for some closed convex rate function* $I : \mathbb{R}^d \to [0,+\infty]$. *Then, for every* $\delta > 0$, *we have*

$$K(\delta) \equiv \inf_{z\in C_\delta} I(z) > 0 \tag{6.3}$$

where $C_\delta \equiv \{z \in I\!R^d : ||z|| \geq \delta\}$.

Proof. We need only consider the case $0 \leq K(\delta) < \infty$, for otherwise (6.3) trivially holds. Therefore, there exists an C_δ-valued sequence $\{z_n, n = 1, 2, \ldots\}$ such that the values $\{I(z_n), n = 1, 2, \ldots\}$ are non-increasing with $\lim_{n\to\infty} I(z_n) = K(\delta)$. Hence, for every $\eta > 0$, there exists a positive integer $n(\eta)$ such that

$$K(\delta) \leq I(z_n) \leq K(\delta) + \eta, \quad n \geq n(\eta). \tag{6.4}$$

Invoking the level-compactness condition **(U2)**, we conclude from (6.4) that a convergent C_δ-valued subsequence $\{z_{n_j}, j = 1, 2, \ldots\}$ can be extracted from $\{z_n, n \geq n(\eta)\}$. If $z^\star$ denotes the limit of this convergent subsequence, then $z^\star$ necessarily belongs to the closed set C_δ so that $z^\star \neq 0$. By the lower semicontinuity of I, we see that

$$K(\delta) = \lim_{j\uparrow\infty} I(z_{n_j}) \geq I(z^\star) > 0 \tag{6.5}$$

with the strict positivity following from **(U3)** since $z^\star \neq 0$. ■

We also need some additional conditions on the mapping g.

(G) The mapping $I\!R^d \to I\!R^p : z \to g(\theta, z + F(\theta))$ is continuous at $z = 0$ uniformly in θ, i.e., for every $\varepsilon > 0$, there exists $\delta(\varepsilon) > 0$ with the property that if $||z|| < \delta(\varepsilon)$, then

$$\sup_{\theta\in\Theta} ||g(\theta, z + F(\theta)) - g(\theta, F(\theta))|| < \varepsilon. \tag{6.6}$$

In many situations of interest, the mapping g takes the form

$$g(\theta, x) = \gamma(x), \quad \theta \in \Theta, \ x \in I\!R^d \tag{6.7}$$

for some Borel mapping $\gamma : I\!R^d \to I\!R^p$. In such cases, condition **(G)** is guaranteed by requiring that γ be uniformly continuous on $I\!R^d$. This latter requirement is satisfied when γ is Lipschitz continuous, a condition obviously met for the frequent choice $\gamma(x) \equiv x$.

Theorem 5. *Assume conditions* **(U1)–(U3)** *and* **(G)** *to hold. Then the rvs* $\{\gamma_{n+1},\ n = 0, 1, \ldots\}$ *satisfy condition* **(E)**.

Proof. Fix $\varepsilon > 0$, and for each $L = 1, 2, \ldots$, set

$$G_L(\theta, x) \equiv \mathbf{P}_{\theta,x}[\|g(\theta, \bar{S}_L(\theta)) - h(\theta)\| \geq \varepsilon], \quad \theta \in \Theta, \ x \in E. \tag{6.8}$$

From the definition (4.2) we readily observe that

$$\begin{aligned}\mathbf{P}[||\gamma_{n+1}|| \geq \varepsilon] &= \mathbf{E}[\mathbf{P}[||g(\theta_n, Y_{n+1}) - h(\theta_n)|| \geq \varepsilon|\mathcal{F}_n]] \\ &= \mathbf{E}[G_{\ell_{n+1}}(\theta_n, X_{n,\ell_n})] \qquad n = 0, 1, \ldots\end{aligned} \tag{6.9}$$

where in the last equality we have made use of the requirement (3.4) on $\mathbf{P}$.

By the uniform continuity condition **(G)**, there exists $\delta(\varepsilon) > 0$ such that (6.6) holds whenever $||z|| < \delta(\varepsilon)$. Hence, for each θ in Θ, the event $[||g(\theta, \bar{S}_L(\theta)) - h(\theta)|| \geq \varepsilon]$ is contained in $[||\bar{S}_L(\theta) - F(\theta)|| \geq \delta(\varepsilon)]$. Therfore, with the notation of Lemma 4, we conclude that

$$G_L(\theta, x) \leq \mathbf{P}_{\theta,x}[\bar{S}_L(\theta) - F(\theta) \in C_{\delta(\varepsilon)}], \quad \theta \in \Theta, \ x \in E \qquad L = 1, 2, \ldots \tag{6.10}$$

Next we pick η in the interval $(0, K(\delta(\varepsilon)))$ which is non–empty due to Lemma 4. Under condition **(U1)** if $K(\delta(\varepsilon))$ is finite, then there exists a finite integer $L(\eta)$ such that

$$\sup_{\theta\in\Theta, x\in E} \mathbf{P}_{\theta,x}[\bar{S}_L(\theta) - F(\theta) \in C_{\delta(\varepsilon)}] \leq e^{-L(K(\delta(\varepsilon))-\eta)}, \quad L \geq L(\eta), \tag{6.11}$$

while if $K(\delta(\varepsilon)) = \infty$, then for every $R > 0$, there exists a finite integer $L(R)$ such that

$$\sup_{\theta\in\Theta, x\in E} \mathbf{P}_{\theta,x}[\bar{S}_L(\theta) - F(\theta) \in C_{\delta(\varepsilon)}] \leq e^{-LR}, \quad L \geq L(R). \tag{6.12}$$

In any event, either from (6.11) or (6.12), we can assert the existence of a finite integer $L^\star$ and of a strictly positive constant $K^\star$ such that

$$\sup_{\theta\in\Theta, x\in E} \mathbf{P}_{\theta,x}[\bar{S}_L(\theta) - F(\theta) \in C_{\delta(\varepsilon)}] \leq e^{-LK^\star}, \quad L \geq L^\star. \tag{6.13}$$

Using this information in (6.10), we readily conclude from (6.9) that **(E)** indeed holds. ■

From the proof of Theorem 5 we see that the law of large numbers **(P2)** automatically holds under conditions **(U1)**–**(U3)**. This is a simple consequence of the bound (6.13) and of the Borel–Cantelli Lemma.

7. SUFFICIENT CONDITIONS FOR (U1)–(U3)

In this section, we develop a uniform large deviations upper bound for a parameterized sequence of dependent rvs. This result generalizes a similar result recently obtained by Dupuis and Simha [6] for i.i.d. rvs.

For $L = 1, 2, \ldots$, we define

$$c_L(t, \theta, x) \equiv \frac{1}{L} \log \mathbf{E}_{\theta,x}[\exp\left(\langle t, L\bar{S}_L(\theta) - LF(\theta)\rangle\right)], \quad t \in I\!R^d, \ \theta \in \Theta, \ x \in E \tag{7.1}$$

and

$$c_L(t) \equiv \sup_{\theta\in\Theta, x\in E} c_L(t, \theta, x), \quad t \in I\!R^d. \tag{7.2}$$

As in [7], we require the following assumptions **(C1)**–**(C2)**:

(C1) For all t in $I\!R^d$, the limit $c(t) \equiv \lim_{L\to\infty} c_L(t)$ exists where we allow $+\infty$ both as a limit value and as an element in the sequence $\{c_L(t), \ L = 1, 2, \ldots\}$.

(C2) The mapping $c : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is a closed convex function whose effective domain $\mathcal{D}(c) \equiv \{t \in \mathbb{R}^d : c(t) < \infty\}$ has a non–empty interior containing the point $t = 0$.

The Legendre–Fenchel transform of c is the closed convex mapping $I : \mathbb{R}^d \to [0, +\infty]$ defined by

$$I(z) \equiv \sup_{t \in \mathbb{R}^d} \{\langle t, z\rangle - c(t)\}, \quad z \in \mathbb{R}^d, \tag{7.3}$$

and for notational convenience, we write

$$I(A) \equiv \inf_{z \in A} I(z), \quad A \subseteq \mathbb{R}^d. \tag{7.4}$$

The first result of this section shows that the conditions **(C1)**–**(C2)** are sufficient to ensure **(U1)**. The proof, which is available in Appendix A, is similar to that given by Dupuis and Simha in the i.i.d. case discussed in [6].

Theorem 6. *Assume* **(P1)**–**(P2)** *and* **(C1)**–**(C2)** *to hold. Then, for any closed subset C of $\mathbb{R}^d$, the inequality*

$$\limsup_{L \to \infty} \frac{1}{L} \log \sup_{\theta \in \Theta, x \in E} \mathbf{P}_{\theta,x}\left[\bar{S}_L(\theta) - F(\theta) \in C\right] \leq -I(C) \tag{7.5}$$

holds.

We are now in a position to give a set of sufficient conditions for **(U1)**–**(U2)** to hold.

Lemma 7. *Assume conditions* **(P1)**–**(P2)** *and* **(C1)**–**(C2)** *to hold. Then the collection of probability measures $\{\mathbf{P}_{\theta,x}, \theta \in \Theta, x \in E\}$ satisfies the uniform Large Deviations upper bound condition* **(U1)**. *The corresponding rate functional I is given by (7.3), and satisfies* **(U2)**.

Proof. That condition **(U1)** holds is immediate from Theorem 6 since the Legendre-Fenchel transform I given by (7.3) is a closed convex mapping.

Next, we show that I given by (7.3) is indeed level–compact. We do so by slightly modifying the arguments of Theorem V.1, Part (f) in [7, pp. 6–7]: For $r \geq 0$, we consider the level set $K_r \equiv \{z \in \mathbb{R}^d : I(z) \leq r\}$ which is closed by the lower semicontinuity of c. From the definition of I, we see that

$$\langle t, z\rangle \leq I(z) + c(t) \leq r + c(t), \quad t \in \mathbb{R}^d, \ z \in K_r \tag{7.6}$$

and therefore, for each $R > 0$, we get

$$\sup_{\|t\| \leq R} \langle t, z\rangle \leq r + \sup_{\|t\| \leq R} c(t), \quad z \in K_r. \tag{7.7}$$

In view of **(C2)**, we can choose R such that the closed ball $B_R \equiv \{z \in \mathbb{R}^d : \|z\| \leq R\}$ is contained in the effective domain $\mathcal{D}(c)$, in which case c is continuous on B_R. Therefore, by standard results from real analysis, we can assert that

$$\sup_{\|t\| \leq R} |c(t)| \equiv A < \infty \quad \text{and} \quad \sup_{\|t\| \leq R} \langle t, z\rangle = R\,\|z\|\,. \tag{7.8}$$

Combining (7.7) and (7.8), we find $||z|| \leq R^{-1}(r+A)$ for all z in K_r, and the level set K_r is thus compact since it is closed and bounded. ■

We address next the crucial condition **(U3)** on the rate functional I. We do so in two steps; the first step being contained in the next lemma and the second step appearing in Theorem 9.

Lemma 8. *Assume* **(P1)–(P2)** *and* **(C1)–(C2)** *to hold. If* $z = 0$, *then* $I(z) = 0$, *in which case* $I(\mathbb{R}^d) = 0$.

Proof. In order to show that $z = 0$ implies $I(z) = 0$, we proceed by contradiction. Assuming $I(0) > 0$, we claim that $\varepsilon > 0$ can always be selected small enough so that $I(B_\varepsilon) > 0$, where $B_\varepsilon \equiv \{z \in \mathbb{R}^d : ||z|| \leq \varepsilon\}$. Indeed, recall [25, Thm. 10.1, p. 82] that the convex function I is continuous on the interior of $\mathcal{D}(I)$ (which contains the origin $z = 0$). By choosing ε small enough, we can ensure that B_ε is contained in the interior of $\mathcal{D}(I)$, and that $I(z) > 0$ for all z in B_ε, this last fact by continuity under the assumption $I(0) > 0$. Continuity over the compact set B_ε yields $0 < I(B_\varepsilon) < \infty$, and by Theorem 6, for $0 < \eta < I(B_\varepsilon)$ there exists a finite integer $L^\star(\eta)$ such that

$$\sup_{\theta\in\Theta, x\in E} \mathbf{P}_{\theta,x}[\bar{S}_L(\theta) - F(\theta) \in B_\varepsilon] \leq e^{-L(I(B_\varepsilon)-\eta)}, \quad L \geq L^\star(\eta). \tag{7.9}$$

Now, taking the limit in (7.9), we conclude that

$$\lim_{L\to\infty} \mathbf{P}_{\theta,x}[\bar{S}_L(\theta) - F(\theta) \in B_\varepsilon] = 0, \quad \theta \in \Theta,\ x \in E, \tag{7.10}$$

or equivalently, that the sample averages (6.1) do not converge in probability, thus not a.s. This conclusion is in direct contradiction with **(P2)** and the assumption $I(0) > 0$ cannot hold. Thus $I(0) = 0$, and we get $I(\mathbb{R}^d) = 0$ from the fact that $I(z) \geq 0$ for all z in $\mathbb{R}^d$. ■

In order to show that $I(z) = 0$ implies $z = 0$, we need an additional condition on the function c defined in **(C1)–(C2)**.

(C3) The function c is (Fréchet–) differentiable at $t = 0$, i.e., its gradient $\nabla c(t)$ exists at $t = 0$, with $\nabla c(0) = 0$.

We are now ready to conclude with the main result of this section:

Theorem 9. *Under* **(P1)–(P2)**, **(C1)–(C3)**, *the conditions* **(U1)–(U3)** *hold.*

Proof. Combining Theorem 6 with Lemmas 7 and 8, we see that all of **(U1)–(U3)** hold *except* for the property that I achieves its global minimum at the unique point $z = 0$, but this follows directly from [7, Thm. V.1 (g), pp. 6–7] under **(C3)**. ■

8. THE I.I.D. CASE

We refer to the *i.i.d.* case as the situation characterized by some collection $\{\mu_\theta,\ \theta \in \Theta\}$ of probability measures on $(E, \mathcal{B}(E))$ such that for Borel subsets $B_1, \ldots, B_L$ in $\mathcal{B}(E)$,

$$\mathbf{P}_{\theta,x}[\xi_\ell \in B_\ell,\ \ell = 1, \ldots, L] = \prod_{\ell=1}^{L} \mu_\theta(B_\ell) \qquad L = 1, \ldots (8.1)$$

for all θ in Θ and x in E. Assumption **(P1)** is satisfied by requiring that the collection $\{\mu_\theta,\ \theta \in \Theta\}$ be measurable in the sense that for every Borel subset B in $\mathcal{B}(E)$, the mapping $\theta \to \mu_\theta(B)$ is Borel measurable. The validity of **(P2)** is guaranteed by the strong law of large numbers for i.i.d. sequences provided the moment condition

$$\int_E |f(\theta,x)| d\mu_\theta(x) < \infty, \quad \theta \in \Theta \tag{8.2}$$

holds, in which case we have

$$F(\theta) = \int_E f(\theta,x) d\mu_\theta(x), \quad \theta \in \Theta. \tag{8.3}$$

The Borel measurability of F then follows readily from the Borel measurability of f by standard arguments.

With (8.1), the definition (7.1) yields

$$c_L(t,\theta,x) = \log \int_E e^{\langle t, f(\theta,x) - F(\theta)\rangle} d\mu_\theta(x), \quad t \in I\!R^d,\ \theta \in \Theta,\ x \in E \tag{8.4}$$

for all $L = 1, 2, \ldots$, and **(C1)** holds in the form

$$\begin{aligned} c(t) &= \lim_{L\to\infty} \sup_{\theta\in\Theta, x\in E} c_L(t,\theta,x) \\ &= \sup_{\theta\in\Theta} \log \int_E e^{\langle t, f(\theta,x) - F(\theta)\rangle} d\mu_\theta(x), \quad t \in I\!R^d. \end{aligned} \tag{8.5}$$

For each θ in Θ and each x in E, the mapping $t \to c_L(t,\theta,x)$ given by (8.4) is convex by Hölder's inequality [5, Lemma 2.2.31, p. 37] and is lower semicontinuous by Fatou's Lemma. Since for proper convex functions, closedness is equivalent to lower semicontinuity [25, Thm. 7.1, pp. 51-52], the mapping $t \to c_L(t,\theta,x)$ is therefore closed and convex. That c is closed and convex follows from the fact that both convexity [25, Thm. 10.8, p. 90] and closedness are preserved under the supremum operation. The other conditions within **(C2)**–**(C3)** can be investigated in specific instances.

As a simple example we consider the case when for each θ in Θ, the measure μ_θ is a Gaussian measure on $I\!R^d$ with mean $m(\theta)$ and covariance matrix $\Sigma(\theta)$. If $f(\theta,x) \equiv x$, then

$$c(t) = \frac{1}{2} \sup_{\theta\in\Theta} \langle t, \Sigma(\theta) t\rangle, \quad t \in I\!R^d. \tag{8.4}$$

It is then easy to see that **(C2)** holds and that $\nabla c(t)$ exists at $t = 0$ (with $\nabla c(0) = 0$) if there exists a symmetric positive semi–definite matrix Σ such that $\Sigma(\theta) \le \Sigma$ for all θ in Θ (where inequalities are with respect to the usual ordering on the cone of symmetric positive semi–definite matrices).

9. THE MARKOV CASE

In the *Markovian* case, we assume the existence of a collection $\{K_\theta, \theta \in \Theta\}$ of measurable transition kernels $E \times \mathcal{B}(E) \to [0,1]$ such that

$$\mathbf{P}_{\theta,x}[\xi_{L+1} \in B | \xi_\ell, \ \ell = 0,1,\ldots,L] = K_\theta(\xi_L; B), \quad B \in \mathcal{B}(E) \qquad L = 0,1,\ldots (9.1)$$

for all θ in Θ and x in E. Condition **(P1)** follows by requiring that for each x in E and each Borel subset B in $\mathcal{B}(E)$, the mapping $\theta \to K_\theta(x_L; B)$ is Borel measurable on Θ. Condition **(P2)** is guaranteed by imposing some ergodicity conditions on the Markov chains with transition kernels $\{K_\theta, \theta \in \Theta\}$.

Of particular interest for applications are the models involving *finite* state Markov chains. We develop this important case by finding explicit conditions on the one–step transition probabilities which ensure the various conditions discussed so far. The set–up is as follows: The state space E is a finite set, say with s elements. As in [5,7], we identify E with the canonical basis $\{e_1, \ldots, e_s\}$ of $\mathbb{R}^s$, i.e., $\langle e_x, e_y \rangle = \delta_{xy}$, $x, y = 1, \ldots, s$; the notation x and e_x, $x = 1, \ldots, s$, is used interchangeably. For each θ in Θ, with the transition kernel K_θ we associate the $s \times s$ stochastic matrix $P(\theta) \equiv (P_\theta(x,y))$ whose entries are defined by

$$P_\theta(x,y) \equiv K_\theta(x; \{y\}), \quad x, y \in E. \tag{9.2}$$

In short, under each of the measures $\mathbf{P}_{\theta,x}$, the rvs $\{\xi_\ell, \ \ell = 0,1,\ldots\}$ form a time–homogeneous Markov chain with one–step transition matrix P_θ.

Next, given the mapping $f : \Theta \times E \to \mathbb{R}^d$, we seek to evaluate the corresponding quantities (7.1)–(7.2). Fixing t in $\mathbb{R}^d$, we define the $s \times s$ matrices $\{\Pi_{t,\theta}, \theta \in \Theta\}$ by

$$\Pi_{t,\theta}(x,y) \equiv P_\theta(x,y) e^{\langle t, f(\theta,y) - F(\theta) \rangle}, \quad \theta \in \Theta, \ x, y \in E. \tag{9.3}$$

As in [5, pp. 58–61] we have

$$c_\ell(t,\theta,x) = \frac{1}{\ell} \log \left\langle e_x, \Pi^\ell_{t,\theta} e \right\rangle, \quad x \in E, \theta \in \Theta \qquad \ell = 1,2,\ldots (9.4)$$

where e is the element $(1,\ldots,1)$ of $\mathbb{R}^d$. Armed with this notation, we can now turn to the main results of this section. We begin with an auxiliary result of a technical nature:

Lemma 10. *Consider the family of finite state space Markov chains with one–step transition matrices* $\{P(\theta), \ \theta \in \Theta\}$. *Suppose that the following conditions are enforced:*

(i): For each θ *in* Θ, *the one–step transition matrix is irreducible and aperiodic; and*

(ii): For each x *and* y *in* E, *the mappings* $\theta \to P_\theta(x,y)$ *and* $\theta \to f(\theta, x)$ *are continuous on* Θ.

Then for each t *in* $\mathbb{R}^d$, *the following statements are true:*

1. *For each* θ *in* Θ, *the non–negative matrix* $\Pi_{t,\theta}$ *is irreducible and primitive; its spectral radius* $\rho(\Pi_{t,\theta})$ *coincides with the largest positive eigenvalue of* $\Pi_{t,\theta}$ *which always has*

multiplicity one, and the eigenvector $u(\Pi_{t,\theta})$ *corresponding to* $\rho(\Pi_{t,\theta})$ *can be selected such that*

$$m_{t,\theta} \equiv \min_i u_i(\Pi_{t,\theta}) > 0 \quad \textit{and} \quad \langle e, u(\Pi_{t,\theta})\rangle = 1; \tag{9.5}$$

2. *The mappings* $\theta \to \rho(\Pi_{t,\theta})$ *and* $\theta \to u(\Pi_{t,\theta})$ *are continuous on* Θ.

Proof. (Claim **1.**) Fix t in $I\!R^d$ and θ in Θ. Since the exponential factors entering the definition (9.3) are strictly positive, it is plain from (ii) that the non-negative matrix $\Pi_{t,\theta}$ is irreducible and primitive [9, Thm. 8, p. 80], and most of Claim **1** is now a simple rephrasing of the Perron–Frobenius theorem [14, Thm. 2.2, p. 545]. The existence of an eigenvector satisfying the normalization condition in (9.5) follows from the positivity condition in (9.5) and the scalability property of eigenvectors.

(Claim **2.**) Fix t in $I\!R^d$. For each θ in Θ, the stochastic matrix $P(\theta)$ is ergodic by virtue of (i), and therefore admits a unique invariant probability vector $\pi(\theta)$, i.e., $\pi(\theta)' = \pi(\theta)'P(\theta)$ and $\langle e, \pi(\theta)\rangle = 1$; we also have $F(\theta) = \sum_x \pi_x(\theta) f(\theta, x)$ by the Ergodic Theorem for Markov Chains [4, Thm. 2, p. 92]. With this in mind, we note that the continuity assumption (ii) on $\theta \to P(\theta)$ implies the continuity of $\theta \to \pi(\theta)$ since $\rho(\Pi_{t,\theta})$ has multiplicity one for all θ in Θ [15, p. 110]. Therefore $\theta \to F(\theta)$ is also continuous by the continuity assumption (ii) on f. In short, from (9.3) and assumption (ii) we conclude that the matrix–valued mapping $\theta \to \Pi_{t,\theta}$ is (entrywise) continuous on Θ, whence the mapping $\theta \to \rho(\Pi_{t,\theta})$ is continuous since each eigenvalue is a continuous mapping on the space of square matrices [18, p. 225]. It is now a simple matter to see that the mapping $\theta \to u(\Pi_{t,\theta})$ is continuous: Indeed, for each θ in Θ, the conditions

$$[\Pi_{t,\theta} - \rho(\Pi_{t,\theta})I_s]\, u = 0 \quad \text{and} \quad \langle e, u\rangle = 1 \tag{9.6}$$

uniquely determine the eigenvector $u(\Pi_{t,\theta})$ since $\rho(\Pi_{t,\theta})$ has multiplicity one. Using this characterization, we can now establish the desired continuity by adapting the arguments of [20, p. 39]; another argument is available in [15, p. 110]. ■

The validity of the conditions **(C1)**–**(C3)** is now discussed:

Theorem 11. *Consider the family of finite state space Markov chains with one–step transition matrices* $\{P(\theta),\ \theta \in \Theta\}$ *under the assumptions (i)–(ii) of Lemma 10. If the parameter set* Θ *is a compact subset of* $I\!R^p$, *then conditions* **(C1)**–**(C3)** *hold with*

$$c(t) \equiv \lim_{\ell\to\infty} \sup_{\theta\in\Theta, x\in E} c_\ell(t,\theta,x) = \sup_{\theta\in\Theta} \log\rho(\Pi_{t,\theta}), \quad t \in I\!R^d. \tag{9.7}$$

Proof. (Condition **(C1)**) Fix t in $I\!R^d$. As pointed out in the proof of Theorem 3.1.2 of [5, p. 60], we have

$$c(t,\theta,x) \equiv \lim_{l\to\infty} c_\ell(t,\theta,x) = \log\rho(\Pi_{t,\theta}), \quad \theta \in \Theta,\ x \in E \tag{9.8}$$

so that

$$\sup_{\theta \in \Theta} \log \rho(\Pi_{t,\theta}) \leq \liminf_{\ell \to \infty} c_\ell(t) \tag{9.9}$$

by invoking the definition (7.1)–(7.2). The conclusion (9.7) (including the existence of the limit) will follow if we can establish that

$$\limsup_{\ell \to \infty} c_\ell(t) \leq \sup_{\theta \in \Theta} \log \rho(\Pi_{t,\theta}). \tag{9.10}$$

To do so, we fix θ in Θ and x in E. In the notation of Lemma 10, $u(\Pi_{t,\theta})$ is the eigenvector of $\Pi_{t,\theta}$ associated with the eigenvalue $\rho(\Pi_{t,\theta})$ such that (9.6) holds. Using the representation (9.5), we readily get

$$\begin{aligned}
c_\ell(t,\theta,x) &= \frac{1}{\ell} \log \langle e_x, \Pi_{t,\theta}^\ell e \rangle \\
&\leq \frac{1}{\ell} \log \left\langle e_x, \Pi_{t,\theta}^\ell \frac{u(\Pi_{t,\theta})}{m_{t,\theta}} \right\rangle \\
&= \frac{1}{\ell} \log \left\langle e_x, \rho(\Pi_{t,\theta})^\ell \frac{u(\Pi_{t,\theta})}{m_{t,\theta}} \right\rangle \\
&\leq \log \rho(\Pi_{t,\theta}) + \frac{1}{\ell} \log \frac{\langle e_x, u(\Pi_{t,\theta}) \rangle}{m_{t,\theta}} \\
&\leq \log \rho(\Pi_{t,\theta}) - \frac{1}{\ell} \log m_{t,\theta}. \qquad \ell = 1, 2, \ldots
\end{aligned} \tag{9.11}$$

Next, upon taking the supremum in (9.11), we see that

$$c_\ell(t) = \sup_{\theta \in \Theta, x \in E} c_\ell(t,\theta,x) \leq \sup_{\theta \in \Theta} \log \rho(\Pi_{t,\theta}) - \frac{1}{\ell} \log \left\{ \inf_{\theta \in \Theta} m_{t,\theta} \right\} \qquad \ell = 1, 2, \ldots \tag{9.12}$$

and the desired inequality (9.10) follows provided (9.7) can be strengthened to read $\inf_{\theta \in \Theta} m_{t,\theta} > 0$, or equivalently, $\min_i \inf_{\theta \in \Theta} u_i(\Pi_{t,\theta}) > 0$. This last condition is now an immediate consequence of the continuity result of Lemma 10 under the compactness condition on Θ.

(Condition **(C2)**) A careful inspection of the proof of (9.7) reveals that in fact we have shown

$$c(t) = \lim_{\ell \to \infty} \sup_{\theta \in \Theta, x \in E} c_\ell(t,\theta,x) = \sup_{\theta \in \Theta, x \in E} \lim_{\ell \to \infty} c_\ell(t,\theta,x), \quad t \in \mathbb{R}^d \tag{9.13}$$

With this in mind, fix θ in Θ and x in E: For each $\ell = 1, 2, \ldots$, the mapping $t \to c_\ell(t,\theta,x)$ is convex (as can be seen by standard arguments [5, Lemma 2.3.9, p. 46] using Hölder's inequality). Therefore, the mapping $t \to c(t,\theta,x)$ is also convex since the pointwise limit of convex mappings is convex [25, Thm. 10.8, p. 90]. Hence, by (9.13), the mapping c is also convex since convexity is preserved under the supremum operation [25, Thm. 5.5, p. 35]. Next, it is plain from (9.8) and (9.13) that $\mathcal{D}(c) = \mathbb{R}^d$ since $0 < \sup_\theta \rho(\Pi_{t,\theta}) < \infty$ by the continuity result of Lemma 10 under the compactness condition on Θ. Therefore, c is continuous throughout $\mathbb{R}^d$, thus *a fortiori* closed.

(Condition **(C3)**) We need to establish that the mapping c is differentiable at $t = 0$ with $\nabla c(0) = 0$. We do so in three steps:

Step 1 – Fix θ in Θ and observe from Jensen's inequality that

$$c_\ell(t,\theta,x) \geq \langle t, \mathbf{E}_{x,\theta}\left[\bar{S}_\ell(\theta)\right] - F(\theta)\rangle, \quad t \in I\!R^d,\ x \in E. \qquad \ell = 1,2,\ldots \tag{9.14}$$

It also follows from assumption (ii) of Lemma 10 that **(P2)** holds, whence $\lim_{\ell\to\infty} \mathbf{E}_{x,\theta}[\bar{S}_\ell(\theta)] = F(\theta)$ via the Bounded Convergence Theorem. Taking the limit in (9.14) and using this last limit result, we get $c(t,\theta,x) \geq 0$ for all t in $I\!R^d$ and x in E. Therefore, since $c(0,\theta,x) = 0$, we conclude that

$$\inf_{t\in I\!R^d} c(t,\theta,x) = c(0,\theta,x) = 0, \quad t \in I\!R^d,\ x \in E. \tag{9.15}$$

Step 2 – Now, for any direction v in $I\!R^d$, the mapping $\lambda \to \Pi_{\lambda v,\theta}$ is entrywise analytic on $I\!R$. Hence, the mapping $\lambda \to \rho(\Pi_{\lambda v,\theta})$ is differentiable on $I\!R$, since in fact analytic on $I\!R$ [18, Thm. 7.7.1, p. 241] as the largest eigenvalue of $\Pi_{\lambda v,\theta}$ is guaranteed to be of multiplicity one by the Perron–Frobenius theory. Thus from differentiability and (9.15) we readily see by standard arguments that

$$\lim_{\lambda\to 0} D_v(\lambda;\theta) = 0 \tag{9.16}$$

where

$$D_v(\lambda;\theta) \equiv \frac{c(\lambda v,\theta,x)}{\lambda}, \quad \lambda \neq 0. \tag{9.17}$$

In particular, the convex mapping $t \to c(t;\theta,x)$ is Gâteaux–differentiable at $t = 0$ along any direction; its differentiability at $t = 0$ now follows from Theorem 25.2 of Rockafellar [25, p. 244].

Step 3 – It follows from convexity that $\lambda \to D_v(\lambda;\theta)$ is non–decreasing on $(0,\infty)$. Moreover, since $\theta \to \rho(\Pi_{\lambda v,\theta})$ is continuous on Θ for each $\lambda \neq 0$, we see that $\theta \to D_v(\lambda,\theta)$ is also continuous on Θ for each $\lambda > 0$. Therefore, starting with a decreasing sequence $\{\lambda_n, n = 0,1,\ldots\}$ such that $\lambda_n \downarrow 0$ as $n \to \infty$, we see from (9.16) that $\lim_n D_v(\lambda_n,\theta) = 0$ monotonically for each θ in Θ. By Dini's Theorem [26, p. 195], this last convergence is taking place uniformly on the compact set Θ, i.e., for every $\varepsilon > 0$, there exists a finite integer $N(\varepsilon)$ such that

$$\sup_{\theta\in\Theta} |D_v(\lambda_n,\theta)| < \varepsilon, \quad n \geq N(\varepsilon). \tag{9.18}$$

Therefore, combining (9.17) and (9.18), we find that

$$\lim_{n\to\infty} \frac{c(\lambda_n v)}{\lambda_n} = \lim_{n\to\infty} \sup_{\theta\in\Theta} D_v(\lambda_n,\theta) = 0 \tag{9.19}$$

or equivalently, the mapping c is Gâteaux–differentiable at $t = 0$ along all directions, and c is indeed Frèchet–differentiable at $t = 0$ by virtue of Theorem 25.2 of [25, p. 244]. ∎

10. CLOSING COMMENTS

In this paper we have shown how the ODE method can be used to establish the a.s. convergence of a class of stochastic approximations driven by sample averages. The pivotal step of the analysis lies in the exponential convergence (**E**) of the "error" terms $\{\gamma_{n+1},\ n = 0, 1, \ldots\}$. Building on the ideas of Dupuis and Simha, we related this property to uniform large deviations upper bounds for the collections of probability measures $\{\mathbf{P}_{\theta,x},\ \theta \in \Theta,\ x \in E\}$. As we take stock in the results obtained thus far, it is plain that they are somewhat preliminary with more to be done in order to expand the domain of applicability of the theory. Below we briefly discuss some possible directions for future inquiry.

a. So far uniform large deviations upper bounds have been established only for sequences of i.i.d. rvs and for finite state Markov chains. Although this last class of models covers already many situations of interest in applications, other classes of parametric stochastic processes need to be investigated, including Markov chains on *countable* state spaces, and discrete–time semi–Markov processes among others. However, it should be noted that the non–finiteness of the state space E will in general preclude the existence of large deviations upper bounds which are uniform in the initial condition, *unless* additional constraints are imposed on the parametric family $\{\mathbf{P}_{\theta,x},\ \theta \in \Theta,\ x \in E\}$. That this is so, even in the Markov case, can already be seen from the results discussed under a uniformity assumption in [5, pp. 249–252]. The correesponding conditions for the parametric family $\{\mathbf{P}_{\theta,x},\ \theta \in \Theta,\ x \in E\}$ may be unduly restrictive to be useful in practice. One way to remedy this diificulty consists in modifying the basic algorithm (3.1)–(3.5) by giving up the requirement that the initial condition $X_{n+1,0}$ at the begining of the $(n+1)^{rst}$ evaluation interval coincides with the final position X_{n,ℓ_n} reached at the end of the previous interval. Instead, with K denoting a *compact* subset of the state space E, we ask only that given $\mathcal{F}_n$, the initial condition $X_{n+1,0}$ be chosen according to some fixed distribution with support on K. For this modified algorithm, which is perhaps more realistic from an implementation viewpoint, we need only establish large deviations upper bounds which are uniform in the parameter θ (in Θ) and in the initial condition x in K (rather than in all of E). This is clearly a less daunting task.

b. The condition (**G**) imposed on the mapping g is quite severe, and may not be satisfied by the mapping g appearing in specific stochastic gradient estimates. Additional work, probably tailored to specific applications, needs to be undertaken in order to relax condition (**G**).

c. Also of great interest are situations where the window sizes $\{\ell_{n+1},\ n = 0, 1, \ldots\}$ are *random*, possibly driven by the observed state process itself, as was the case in earlier work [3,8,12] on stochastic approximations driven by sample averages. It remains to be seen whether the framework developed here can be suitably modified to handle these situations.

d. In some applications the state process driving the sample averages (3.2) is one that evolves in continuous time, rather than in discrete time as was assumed in this paper. In the continuous–time context, the probability measures $\{\mathbf{P}_{\theta,x},\ \theta \in \Theta,\ x \in E\}$ describing the statistical behavior of the state process are probability measures which are defined on well–structured subsets of $E^{\mathbb{R}_+}$, the space of all E–valued mappings defined on $\mathbb{R}_+$; typical examples would include $C[\mathbb{R}_+; \mathbb{R}^s]$ or $D[\mathbb{R}_+; E]$ for some countable set E. Technical details aside, if $\{\xi_t,\ t \geq 0\}$ now denotes a generic E–valued mapping describing the time evolution of the system. then **(P2)** is now replaced by

$$\lim_{t\to\infty} \frac{1}{t} \int_0^t f(\theta, \xi_s) ds \quad \mathbf{P}_{\theta,x} - a.s. \tag{10.1}$$

for every θ in Θ and x in E. For some given monotone increasing sequence of *deterministic* times $\{t_n,\ n = 1, 2, \ldots\}$ such that $t_n \uparrow \infty$, we define

$$Y_{n+1} \equiv \frac{1}{t_{n+1}} \int_0^{t_{n+1}} f(\theta, X_{n+1,s}) ds \qquad n = 0, 1, \ldots (10.2)$$

where $\{X_{n+1,s},\ 0 \leq s \leq t_{n+1}\}$ denotes the piece of the state process which is observed during the $(n+1)^{rst}$ observation interval, and whose statistics are determined by the parameter value θ_n and the initial condition X_{n,t_n}. We can now consider the algorithm (3.1) with (3.2) replaced by (10.2); the framework developed in this paper still applies, but now requires that uniform large deviations upper bounds for continuous time processes be established.

e. As the approach relies heavily on large deviations arguments, this requires the finiteness of certain exponential moments, thus leading naturally to the condition **(L)** on the window sizes $\{\ell_{n+1},\ n = 0, 1, \ldots\}$. Of course such a condition is dictated by the technique adopted here, and is certainly far from necessary as we now show through an example. We shall see that in some cases only finite second order moments suffice in order to yield a.s. convergence, and this in the absence of condition **(L)**, provided an additional condition is imposed on the gain sequence $\{a_{n+1},\ n = 0, 1, \ldots\}$, namely

$$\sum_{n=0}^{\infty} a_{n+1}^2 < \infty. \tag{10.3}$$

To develop this point, we consider an unconstrained scheme (i.e. $\Theta = \mathbb{R}^p$) with $p = d = s$, and $g(\theta, x) = f(\theta, x) = x$ for all θ and x in $\mathbb{R}^p$, so that (1.5)–(1.7) takes the form

$$\theta_0 \in \mathbb{R}^p, \quad \theta_{n+1} = \theta_n + a_{n+1} \frac{1}{\ell_{n+1}} \sum_{\ell=1}^{\ell_{n+1}} X_{n+1,\ell}. \qquad n = 0, 1, \ldots (10.4)$$

We put ourselves in the *i.i.d.* case of Section 8 with the additional assumption that for each θ in $\mathbb{R}^p$, the probability measure μ_θ has finite mean $m(\theta)$ and covariance matrix $\Sigma(\theta)$. We assume that $m(\theta) \neq 0$ except for $\theta = \theta^\star$; we take $\theta^\star = 0$ for the sake of convenience. By following an argument of Gladyshev [10], we get the following result whose proof is available in Appendix B.

Proposition 12. *Under the foregoing assumptions on the probability measures $\{\mu_\theta,\ \theta \in \mathbb{R}^p\}$, we further assume the conditions*

$$\sup_{\varepsilon^{-1} < ||\theta|| < \varepsilon} \langle \theta, m(\theta) \rangle < 0, \quad \varepsilon \in (0, 1) \tag{10.5}$$

and

$$||m(\theta)||^2 + Tr(\Sigma(\theta)) \le K(1 + ||\theta||^2), \quad \theta \in I\!R^p \tag{10.6}$$

for positive constant K. If the gain sequence $\{a_{n+1},\ n = 0, 1, \ldots\}$ satisfies both (**A**) *and (10.3), then* $\lim_{n\to\infty} \theta_n = 0$ **P**–*a.s. without any additional condition on the window size sequence* $\{\ell_{n+1},\ n = 0, 1, \ldots\}$.

It is plain under the i.i.d. assumption that there is no loss of generality in taking $f(\theta, x) \equiv x$ for all θ and x in $I\!R^p$. Moreover, projected versions of the algorithm can in principle be addressed by arguments similar to the ones given by Chong and Ramadge [2, Appendix A, p. 365]. Therefore, in the i.i.d. case with linear g, Proposition 12 (and its variants) suggest conditions for a.s. convergence which are similar to those given for the standard Robbins–Monro scheme (without averaging), and probably weaker than the ones developed in this paper so that the framework developed here then seems to provide little improvement, if any. However, the situation is quite different when g is nonlinear; the martingale arguments break down even in the i.i.d. case and the framework discussed in this paper now leads to conditions for a.s. convergence.

APPENDICES

A. A PROOF OF THEOREM 6

Let C be a closed subset of $I\!R^d$. If $I(C) = 0$, then (7.5) automatically holds. Hence, we need only establish (7.5) when $0 < I(C)$, in which case two cases emerge, namely $0 < I(C) < \infty$ and $I(C) = \infty$.

Case 1 – If $0 < I(C) < \infty$, then ε can be selected in the interval $(0, I(C))$. By Gärtner's covering lemma, there exist r distinct non–zero points $t_1, \ldots, t_r$ in $\mathcal{D}(c)$ such that

$$C \subset \bigcup_{i=1}^{r} H_+(t_i, I(C) - \varepsilon) \tag{A.1}$$

where $H_+(t, \alpha) \equiv \{z \in I\!R^d : \langle t, z\rangle - c(t) \ge \alpha\}$.

The integer r and the points $t_1, \ldots, t_r$ depend on both ε and C, but *not* on θ and x. For each $i = 1, \ldots, r$, the point t_i belongs to $\mathcal{D}(c)$, so that $c(t_i)$ is finite and $c_L(t_i)$ is therefore also *finite* for L large enough, say $L \ge L'$ – it is plain that L' can be chosen the same for all $i = 1, \ldots, r$.

Fix θ in Θ, x in E and $L \ge L'$. With these facts in mind, we readily see from (A.1) that

$$\begin{aligned}
\mathbf{P}_{\theta,x}[\bar{S}_L(\theta) - F(\theta) \in C] &\le \sum_{i=1}^{r} \mathbf{P}_{\theta,x}\left[\langle t_i, \bar{S}_L(\theta) - F(\theta)\rangle - c(t_i) \ge I(C) - \varepsilon\right] \\
&= \sum_{i=1}^{r} \mathbf{P}_{\theta,x}[\exp(\langle t_i, L\bar{S}_L(\theta) - LF(\theta)\rangle) \ge \exp\left(L(c(t_i) + I(C) - \varepsilon)\right)] \\
&\le \sum_{i=1}^{r} \mathbf{E}_{\theta,x}[\exp\left(\langle t_i, L\bar{S}_L(\theta) - LF(\theta)\rangle\right)] \exp\left(-L(c(t_i) + I(C) - \varepsilon)\right)
\end{aligned}$$

$$= \sum_{i=1}^{r} \exp\left(Lc_L(t_i,\theta,x)\right) \exp\left(-L(c(t_i) + I(C) - \varepsilon)\right)$$

$$= \sum_{i=1}^{r} \exp\left(L(c_L(t_i,\theta,x) - c(t_i))\right) \exp\left(-L(I(C) - \varepsilon)\right)$$

$$\leq \sum_{i=1}^{r} \exp(L(c_L(t_i) - c(t_i))) \exp\left(-L(I(C) - \varepsilon)\right). \qquad (A.2)$$

The last inequality follows from the fact that $c_L(t,\theta,x) \leq c_L(t)$ for all θ in Θ, x in E and t in $I\!R^d$.

Since $\lim_{L\to\infty} c_L(t_i) = c(t_i)$, $i = 1,\ldots,r$, we can find for every $\delta > 0$, an integer $L^\star = L^\star(\delta)$ such that $L^\star \geq L'$ and $|c_L(t_i) - c(t_i)| < \delta$, $i = 1,\ldots,r$, whenever $L \geq L^\star$, and therefore

$$\sup_{i=1,\ldots,r} \exp(L(c_L(t_i) - c(t_i))) \leq \exp(L\delta), \quad L \geq L^\star. \qquad (A.3)$$

Using this last fact, we conclude from (A.2) that

$$\mathbf{P}_{\theta,x}[\bar{S}_L(\theta) - F(\theta) \in C] \leq r \exp\left(-L(I(C) - \varepsilon - \delta)\right), \quad L \geq L^\star \qquad (A.4)$$

whence

$$\sup_{\theta\in\Theta, x\in E} \mathbf{P}_{\theta,x}\left[\bar{S}_L(\theta) - F(\theta) \in C\right] \leq r \exp\left(-L(I(C) - \varepsilon - \delta)\right), \quad L \geq L^\star \qquad (A.5)$$

since the integer r and the points $t_1, t_2, \cdots, t_r$ depend on the set C and on the chosen ε, and the integer $L^\star$ depends on C, ε and the chosen $\delta > 0$. It then follows that

$$\limsup_{L\to\infty} \frac{1}{L} \log \sup_{\theta\in\Theta, x\in E} \mathbf{P}_{\theta,x}\left[\bar{S}_L(\theta) - F(\theta) \in C\right] \leq -(I(C) - \varepsilon - \delta) \qquad (A.6)$$

and (7.5) now follows since (A.6) holds for all ε in the interval $(0, I(C))$ and for all $\delta > 0$.

Case 2 – If $I(C) = \infty$, then fix $R > 0$ and by Gärtner's covering lemma, there again exist r distinct non-zero points $t_1,\ldots,t_r$ in $\mathcal{D}(c)$ such that

$$C \subset \bigcup_{i=1}^{r} H_+(t_i, R). \qquad (A.7)$$

The integer r and the points $t_1,\ldots,t_r$ depend on both R and C, but *not* on θ and x. For each $i = 1,\ldots,r$, the point t_i belongs to $\mathcal{D}(c)$, so that $c(t_i)$ is finite and $c_L(t_i)$ is therefore also finite for L large enough, say $L \geq L''$ – it is again plain that L'' can be chosen the same for all $i = 1,\ldots,r$.

Fix θ in Θ and $L \geq L''$. By the same arguments as the one leading to (A.2), this time with the help of (A.7), we get

$$\mathbf{P}_{\theta,x}[\bar{S}_L(\theta) - F(\theta) \in A] \leq \sum_{i=1}^{r} \mathbf{P}_{\theta,x}\left[\langle t_i, \bar{S}_L(\theta) - F(\theta)\rangle - c(t_i) \geq R\right]$$

$$\leq \sum_{i=1}^{r} \exp(L(c_L(t_i) - c(t_i))) \exp\left(-L(R)\right). \qquad (A.8)$$

Since $\lim_{L\to\infty} c_L(t_i) = c(t_i)$, $i = 1, \ldots, r$, we can find for every $\delta > 0$, an integer $L^\star = L^\star(\delta)$ such that $L^\star \geq L''$ and $|c_L(t_i) - c(t_i)| < \delta$, $i = 1, \ldots, r$, whenever $L \geq L^\star$, and as in **Case 1**, we can conclude from (A.8) that

$$\sup_{\theta\in\Theta, x\in E} \mathbf{P}_{\theta,x}\left[\bar{S}_L(\theta) - F(\theta) \in C\right] \leq r\exp\left(-L(R-\delta)\right), \quad L \geq L^\star. \tag{A.9}$$

Therefore,

$$\limsup_{L\to\infty} \frac{1}{L} \log \sup_{\theta\in\Theta, x\in E} \mathbf{P}_{\theta,x}\left[\bar{S}_L(\theta) - F(\theta) \in C\right] \leq -(R-\delta) \tag{A.10}$$

and with R and δ being arbitrary, we conclude that (7.5) holds in the form

$$\limsup_{L\to\infty} \frac{1}{L} \log \sup_{\theta\in\Theta, x\in E} \mathbf{P}_{\theta,x}\left[\bar{S}_L(\theta) - F(\theta) \in C\right] = -\infty. \tag{A.11}$$

■

B. A PROOF OF PROPOSITION 12

In the proof, there is no loss of generality in asssuming θ_0 to be non–random, as we do from now on. We begin by writing (10.4) in the form

$$\theta_0 \in I\!R^p, \quad \theta_{n+1} = \theta_n + a_{n+1}\{m(\theta_n) + \gamma_{n+1}\}, \qquad n = 0, 1, \ldots (B.1)$$

and by noting that under the i.i.d. assumption, we have

$$\mathbf{E}[\gamma_{n+1}|\mathcal{F}_n] = 0 \quad \text{and} \quad \mathbf{E}[||\gamma_{n+1}||^2|\mathcal{F}_n] = \frac{1}{\ell_{n+1}} Tr[\Sigma(\theta_n)]. \qquad n = 0, 1, \ldots (B.2)$$

With the notation $R(\theta) \equiv \langle \theta, m(\theta)\rangle$ for all θ in $I\!R^p$, we readily get from (B.2) that

$$\begin{aligned}
\mathbf{E}[||\theta_{n+1}||^2|\mathcal{F}_n] =& ||\theta_n||^2 + 2a_{n+1}R(\theta_n) + a_{n+1}^2\left[||m(\theta_n)||^2 + \frac{1}{\ell_{n+1}} Tr[\Sigma(\theta_n)]\right] && (B.3)\\
\leq& ||\theta_n||^2 + a_{n+1}^2 K(1 + ||\theta_n||^2) + \frac{a_{n+1}^2}{\ell_{n+1}} K(1 + ||\theta_n||^2) && (B.4)\\
\leq& (1 + 2Ka_{n+1}^2)||\theta_n||^2 + 2Ka_{n+1}^2 && n = 0, 1, \ldots (B.5)
\end{aligned}$$

where in (B.4) we used (10.5)–(10.6). Next we introduce the integrable rvs $\{M_n,\ n = 0, 1, \ldots\}$ by setting

$$M_0 \equiv ||\theta_0||^2, \quad M_{n+1} \equiv \alpha_{n+1}||\theta_{n+1}||^2 - \beta_{n+1} \qquad n = 0, 1, \ldots (B.6)$$

with

$$\alpha_{n+1} \equiv \prod_{i=0}^{n} \left(1 + 2Ka_{i+1}^2\right)^{-1} \quad \text{and} \quad \beta_{n+1} \equiv \sum_{i=0}^{n} 2Ka_{i+1}^2 \alpha_{i+1}. \qquad n = 0, 1, \ldots (B.7)$$

We observe that (B.5) is equivalent to the supermartingale property

$$\mathbf{E}[M_{n+1}|\mathcal{F}_n] \leq M_n \quad \mathbf{P} - a.s. \qquad n = 0, 1, \ldots (B.8)$$

so that

$$\sup_n \mathbf{E}[M_n] \leq ||\theta_0||^2. \qquad (B.9)$$

We also note the easy bounds

$$A \leq \alpha_{n+1} \leq 1 \quad \text{and} \quad 0 \leq \beta_{n+1} \leq B \qquad n = 0, 1, \ldots (B.10)$$

where

$$A \equiv \exp[-2K \sum_{i=0}^{\infty} a_{i+1}^2] \quad \text{and} \quad B \equiv \lim_{n\to\infty} \beta_n; \qquad (B.11)$$

from (10.3) we see that $0 < A \leq 1$ and $B < \infty$.

From (B.6), with $\alpha_0 = 1$ and $\beta_0 = 0$, we readily obtain the inequalities

$$M_n + \beta_n \geq A||\theta_n||^2 \quad \text{and} \quad |M_n| \leq \alpha_n||\theta_n||^2 + \beta_n. \qquad n = 0, 1, \ldots (B.12)$$

Using (B.10)–(B.11) we conclude from (B.9) and the first inequality in (B.12) that $\sup_n \mathbf{E}[||\theta_n||^2] < \infty$, whence $\sup_n \mathbf{E}[|M_n|] < \infty$ by the second part of (B.12). Therefore, by the basic martingale convergence theorem [14, Thm. 5.1., p. 278], the supermartingale $\{M_n,\ n = 0, 1, \ldots\}$ converges $\mathbf{P}$–a.s. to a finite rv, and so does also the sequence $\{||\theta_n||^2,\ n = 0, 1, \ldots\}$.

It remains to show that $\lim_{n\to\infty} ||\theta_n||^2 = 0$ $\mathbf{P}$–a.s. To do this, we take expectations on both sides of (B.3) and get

$$\mathbf{E}[||\theta_{n+1}||^2] = \mathbf{E}[||\theta_n||^2] + 2a_{n+1}\mathbf{E}[R(\theta_n)] + a_{n+1}^2\mathbf{E}\left[||m(\theta_n)||^2 + \frac{1}{\ell_{n+1}}Tr[\Sigma(\theta_n)]\right]. \qquad n = 0, 1, \ldots (B.13)$$

After adding these relations for $k = 0, 1, \ldots, n$ and cancelling appropriate terms, we are then left with the relation

$$\begin{aligned} &\mathbf{E}[||\theta_{n+1}||^2] \qquad n = 0, 1, \ldots (B.14) \\ &= ||\theta_0||^2 + \sum_{k=0}^{n} a_{k+1}\mathbf{E}[R(\theta_k)] + \sum_{k=0}^{n} a_{k+1}^2\mathbf{E}\left[\mathbf{E}[||m(\theta_k)||^2] + \frac{1}{\ell_{k+1}}Tr[\Sigma(\theta_k)\right]. \end{aligned}$$

Upon using the inequality (10.6) and the bound $\sup_n \mathbf{E}[||\theta_n||^2] < \infty$ obtained earlier, we easily conclude from (B.14) that

$$0 \leq -\sum_{k=0}^{\infty} a_{k+1}\mathbf{E}[R(\theta_k)] < \infty. \qquad (B.15)$$

Therefore, $\lim_{k\to\infty} a_{k+1}\mathbf{E}[R(\theta_k)] = 0$ and under **(A)** a simple argument by contradiction shows that we must necessarily have $\liminf_{k\to\infty} \mathbf{E}[R(\theta_k)] = 0$. In other words, along a subsequence, say $\{n_j;\ j=1,2,\ldots\}$, we have $\lim_{j\to\infty} \mathbf{E}[R(\theta_{n_j})] = 0$, whence $\lim_{j\to\infty} R(\theta_{n_j}) = 0$ in probability (under $\mathbf{P}$). Consequently, along a further subsequence, still denoted $\{n_j,\ j=1,2,\ldots\}$, we have $\lim_{j\to\infty} R(\theta_{n_j}) = 0$ $\mathbf{P}$–a.s. Using this last fact in conjunction with (10.5) readily yields $\lim_{j\to\infty} \theta_{n_j} = 0$ $\mathbf{P}$–a.s. and the desired conclusion now follows. ■

REFERENCES

[1] C. Cassandras, *Discrete Event Systems: Modeling and Performance Analysis*, Irwin , Boston (MA), (1993).

[2] E. Chong and P. Ramadge, "Convergence of recursive optimization algorithms using infinitesimal perturbation analysis estimates," *Journal of Discrete Event Dynamical Systems* **1** (1992), pp. 339–372.

[3] E. Chong and P. Ramadge, "Optimization of queues using an infitesimal perturbation analysis–based stochastic algorithm with general update times," *SIAM Journal of Control and Optimization* **31** (1993), pp. 1–35.

[4] K.L. Chung, *Markov Chains with Stationary Transition Probabilities*, Springer-Verlag, Berlin, (1967).

[5] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, Jones and Bartlett Publishers, Boston (MA) (1993).

[6] P. Dupuis and R. Simha, "On sampling controlled stochastic approximation, *IEEE Transactions on Automatic Control* **AC–36** (1991), pp. 915–924.

[7] R. C. Ellis, "Large deviations for a general class of random vectors," *Annals of Probability* **12**, (1984), pp. 1–12.

[8] M. Fu, "Convergence of a stochastic approximation algorithms for the $GI/G/1$ queue using infinitesimal perturbation analysis," *Journal of Optimization Theory and Applications* **65** (1990), pp. 149–160.

[9] F. R. Gantmacher, *The Theory of Matrices*, Volume 2, Chelsea Publishing Co., New York (NY) (1960).

[10] E. G. Gladyshev, "On stochastic approximation," *Theo. Prob. Appl.* **10** (1965), pp. 275–278.

[11] P. Glasserman, *Gradient Estimation Via Perturbation Analysis*, Kluwer Academic Press, Boston (MA) (1991).

[12] P. Glynn, "Likelihood ratio gradient estimation: An overview," *Proceedings of 1987 Winter Simulation Conference*, November 1987, Atlanta (GA), pp. 366-375.

[13] Y.–C. Ho, *Perturbation Analysis for Discrete Event Systems*, Kluwer Academic Press, Boston (MA) (1991).

[14] S. Karlin and H. M. Taylor, *A First Course in Stochastic Processes*, Second Edition, Academic Press, New York (NY) (1975).

[15] T. Kato, *Perturbation Theory for Linear Operators*, First Edition, Springer-Verlag, New York (NY) (1966).

[16] H.J. Kushner and D. S. Clark, *Stochastic Approximation for Constrained and Unconstrained Systems*, Applied Mathematical Sciences **26**, Springer–Verlag, Berlin (1978).

[17] H.J. Kushner and A. Shwartz, "An invariant measure approach to the convergence of stochastic approximations with state–dependent noise," *SIAM Journal of Control and Optimization* **21** (1983), pp. 1–35.

[18] P. Lancaster, *Theory of Matrices*, Academic Press, New York (NY) (1969).

[19] M. Loéve, *Probability Theory I*, Fourth Edition, Springer–Verlag, Berlin (1977).

[20] D.-J. Ma, A. M. Makowski and A. Schwartz "Stochastic approximations for finite state markov chains," Stochastic Processes and their Applications **35**, (1990), pp. 27–45.

[21] M.S. Meketon, "Optimization in simulation: A survey of recent results," *Proceedings of the 1987 Winter Simulation Conference*, November 1987, Atlanta (GA), pp. 58–67.

[22] M. Metivier and P. Priouret, "Applications of a Kushner and Clark lemma to general classes of stochastic algorithms," *IEEE Transactions on Information Theory* **30** (1984), pp. 140–150.

[23] M. Reiman and A. Weiss, "Sensitivity analysis via likelihood ratio," *Operations Research* **22** (1989), pp. 830-844.

[24] H. Robbins and S. Monro, "A stochastic approximation method," *Annals of Mathematical Statistics* **22**(1951), pp. 400–407.

[25] R.T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton (NJ) (1970).

[26] H.L. Royden, *Real Analysis*, Third Edition, MacMillan Publishing Company, New York (NY) (1988).

[27] R. Suri, "Perturbation analysis: the state of the art and research issues explained via the $GI/G/1$ queue," *Proceedings of the IEEE*, **77** (1989) pp. 114–137.

[28] A. Shwartz, *Convergence of Stochastic Approximations: the Invariant Measure Approach*, Ph.D. Thesis, Division of Engineering, Brown University, Providence (RI) (1982).

[29] Y. Wardi, "Simulation-based stochastic algorithm for optimizing GI/G/1 queues," Ben Gurion University of the Negev, Beer Sheva (Israel), Manuscript (1988).