Mathematical Programming manuscript No.
(will be inserted by the editor)

G. Yin · C. Ion · V. Krishnamurthy

How Does a Stochastic Optimization/Approximation Algorithm Adapt to a Randomly Evolving Optimum/Root with Jump Markov Sample Paths

Dedicated to Professor Boris Polyak on the Occasion of His 70th Birthday

Received: date / Accepted: date

Abstract Stochastic optimization/approximation algorithms are widely used to recursively estimate the optimum of a suitable function or its root under noisy observations when this optimum or root is a constant or evolves randomly according to slowly time-varying continuous sample paths. In comparison, this paper analyzes the asymptotic properties of stochastic optimization/approximation algorithms for recursively estimating the optimum or root when it evolves rapidly with nonsmooth (jump-changing) sample paths. The resulting problem falls into the category of regime-switching stochastic approximation algorithms with two-time scales. Motivated by emerging applications in wireless communications, and system identification, we analyze asymptotic behavior of such algorithms. Our analysis assumes that the noisy

The research of G. Yin was supported in part by the National Science Foundation, and in part by the National Natural Science Foundation of China under #60574069. The research of C. Ion was supported in part by the Wayne State University Rumble Fellowship. The research of V. Krishnamurthy was supported in part by NSERC (Canada).

G. Yin
Department of Mathematics
Wayne State University, Detroit, MI 48202
Tel.: 313-577-2496
Fax: 313-577-7596
E-mail: gyin@math.wayne.edu

C. Ion
Department of Mathematics
Wayne State University, Detroit, MI 48202
E-mail: cion@math.wayne.edu

Vikram Krishnamurthy
Department of Electrical Engineering
University of British Columbia, Vancouver, V6T 1Z4, Canada
E-mail: vikramk@ece.ubc.ca

observations contain a (nonsmooth) jump process modeled by a discrete-time Markov chain whose transition frequency varies much faster than the adaptation rate of the stochastic optimization algorithm. Using stochastic averaging, we prove convergence of the algorithm. Rate of convergence of the algorithm is obtained via bounds on the estimation errors and diffusion approximations. Remarks on improving the convergence rates through iterate averaging, and limit mean dynamics represented by differential inclusions are also presented.

Keywords stochastic approximation, stochastic optimization, nonsmooth-jump component, two-time scale, convergence and rate of convergence.

Mathematics Subject Classification (2000) 62L20, 60J10, 34E05, 90C15.

1 Introduction

1.1 Motivation

This paper considers a class of two-time-scale stochastic optimization and approximation algorithms for estimating a randomly evolving optimum of a suitable function or its root. The randomly evolving optimum/root has nonsmooth (more specifically jump-varying) sample paths modeled as a finite state Markov chain. The motivation for introducing this finite state Markov chain stems from emerging applications in wireless communications (e.g., tracking fast fading channels in cognitive radio systems [9]), financial engineering (e.g., modeling stochastic volatility and other market and economic factors), discrete optimization, system identification, detection, nano technology, and optimization of a function whose noisy measurements contain a jump component; see [1,3,4,11,13,31] among others. In these applications, a Markov chain is used to modulate the regime switching and to model the random environment. Typically, the system under consideration has a number of regimes or configurations nonsmoothly connected, across which the behavior of the system is markedly different.

Due to the small step size used in the recursive computation of the sequence of iterates (parameter estimates), the stochastic optimization and approximation algorithms can be considered as a *slow* dynamical system. That is, the iterates generated by the algorithm evolve slowly with time. Indeed, the step size determines the adaptation rate of the algorithm. In comparison, the Markov chain is a *fast* process – it varies at an order of magnitude faster than the adaptation rate of the stochastic approximation algorithm. We focus on analyzing the behavior of the stochastic approximation/optimization algorithm for such systems.

In general terms, our stochastic optimization problem can be posed as analyzing the tracking behavior of a two-time-scale stochastic approximation algorithm. However unlike standard stochastic approximation problems, there is an additional fast Markov chain parameter in the noisy measurements. Because the stochastic approximation/optimization algorithm (with a small step size) is a slow dynamic system and the parameter jump changes rapidly, it is virtually impossible to track the Markov parameter process using a stochastic approximation method. Thus analyzing the behavior of stochastic approximation and optimization algorithms under this scenario poses a challenging problem. We observe, however, due to the fast variation of the jump-changing parameter, there is no need to track the system at any given instance since it will jump to another (Markov) state within a very short period of time. Instead of tracking the system at any time instant, we suggest to handle it using a different approach. We may treat the Markov chain as another source of noise. As an alternative, we show that a limit system can be obtained in which the switching is averaged out with respect to the stationary measure. Our recommendation is: In lieu of tracking the original system, we can concentrate on estimating the limit system.

Another example that motivates this paper is stochastic optimization problems where noisy observations contain a fast jump component. Such jump components may be used to model sudden changes in the system environment, for example, a Gilbert-Elliott model for a rapid channel fade in a wireless communications system. Suppose we are interested in the following stochastic optimization problem: Minimize a suitable (smooth) deterministic function $EF(x, \xi_n, \theta_n)$ (where E denotes the expectation operator), given noisy observations of the function values $F(x_n, \xi_n, \theta_n)$ (or noisy observations of the gradient $\nabla_x EF(x_n, \xi_n, \theta_n)$) at suitable design points x_n with $n = 0, 1, 2, \dots$ denoting the discrete time. Here $\{\xi_n\}$ is a stationary stochastic process termed "observation noise," $\{\theta_n\}$, independent of $\{\xi_n\}$, is a discrete-time Markov chain taking values in $\{1, 2, \dots, m_0\}$, and having transition matrix $P^\varepsilon = P + \varepsilon Q$ (where P is irreducible and aperiodic, Q is a generator of a continuous-time Markov chain and $\varepsilon > 0$ is a small parameter). It is readily seen that $EF(x, \xi_n, \theta_n) = \sum_{i=1}^{m_0} EF(x, \xi_n, i) P(\theta_n = i)$. The difficulty in using a stochastic approximation algorithm lies in: One needs to track $P(\theta_n = i)$ of the Markov chain at each time instant n. In this paper, we suggest a viable alternative. Instead of computing the optimal estimate (e.g., conditional mean) which evolves rapidly over time, we focus on asymptotic or "near" optimality. In this setting, the instantaneous $P(\theta_n = i)$ is replaced by its averaged, namely the "stationary distribution" of the Markov chain θ_n. We can ignore the detailed variations, and focus on what happens in the limit system. In many practical situations, such an approach will provide us with approximate tracking capability or feasible estimation procedure.

Before proceeding, a remark on the analysis to be presented is in order. Due to the time-varying characteristics and the Markovian jumps, we cannot directly invoke the existing results in the literature of usual stochastic approximation (SA) methods, for example, [17]. Instead, we need to start by working out the stochastic averaging by applying a martingale problem formulation. In this work, we first analyze the convergence of the algorithms. Then we proceed to analyze the rates of convergence, which is handled via the following steps. First an error bound is obtained by use of a Liapunov function. Then the limit of a suitably scaled sequence of the estimation errors is shown to be the solution of a stochastic differential equation. We further demonstrate how the asymptotic convergence rate of the algorithm can be accelerated using minimal window width "iterate averaging" [16]. In the late 1980's and early 1990's, the iterate averaging procedure for accelerating convergence rates of stochastic approximation algorithms was proposed by Polyak [19] (see also [20]) and developed independently by Ruppert [21]. The main idea of their approach is the use of averaging of iterates obtained from a classical stochastic approximation algorithm with slowly varying step sizes. Their work stimulated much of the subsequent research in this area such as [7,22,16,27] among others.

1.2 Perspective

Bayesian Estimation versus Stochastic Approximation. To track a fast jump changing Markovian signal, instead of using a stochastic approximation algorithm (which is what this paper considers), one could use a Bayesian Hidden Markov Model filtering algorithm [8]. Such Bayesian filtering algorithms recursively compute the conditional mean estimate of the state of the underlying Markov chain based on the observation history; see also [12] for recursive (real time) joint state and parameter estimation for HMMs. However, Bayesian filtering algorithms suffer from the twin curses of modeling and dimensionality. That is, exact knowledge of noise distribution and the transition probabilities and state levels of the Markov chain are required and the computational complexity is quadratic in the number of states. For this reason, in many applications, for example, in wireless communications, Bayesian filtering is seldom used. In contrast, stochastic approximation algorithms are widely used although they do not exploit knowledge of the underlying dynamics of the Markov chain to compute the estimates. It is thus of significant interest to analyze the performance of a stochastic approximation algorithm when the underlying parameter is jump-changing rapidly.

Context. To give some perspective on the results in this paper, we briefly discuss two other recent results which also deal with stochastic approximation and optimization algorithms and parameters with finite state Markovian dynamics. If we specialize the SA algorithm to LMS (least mean squares) estimation and tracking, the dynamics of the true parameter considered in this paper (modeled as a Markov chain with transition probability matrix $P+\varepsilon Q$, where P is a transition matrix and Q is a generator of a continuous-time Markov chain) evolves on much faster time scale than the dynamics of the stochastic approximation algorithm, i.e., $\varepsilon = O(\mu^\gamma)$ with $0 < \gamma < 1$. In comparison, the recent papers [29,31] analyze the tracking properties of a stochastic approximation algorithm when the Markov chain dynamics of the true parameter evolves on the same time scale, i.e., $\varepsilon = O(\mu)$, as the stochastic approximation algorithm (with $P = I$ used there). In [31,29], unlike the results here, the limit dynamics of the iterates can be represented by a Markov modulated ordinary differential equation (ODE) and the limit of the scaled sequence of estimation errors satisfies a switching diffusion process. In [30], we address the case where the Markov chain evolves much slower than the dynamics of the stochastic approximation algorithm, i.e., $\varepsilon = o(\mu)$. In that case, the limit dynamics are also captured by an ODE and the limit of a scaled sequence of the estimation errors is a diffusion process. The aforementioned results together with findings of this paper give a characterization of LMS type algorithms when the true parameter changes in accordance with a Markov chain (with different scales).

1.3 Outline

The rest of the paper is arranged as follows. Section 2 presents the precise formulation and the recursive algorithm. Section 3 is devoted to studying the convergence of the algorithm. We derive an associated ordinary differential equation through weak convergence analysis, in which both the observation noise and the Markov chain are averaged out. Also provided is a bound in terms of a Liapunov function, which leads to tightness of the iterates. This tightness further allows us to obtain convergence to the stable point of a limit ordinary differential equation using a stability argument. Section 4 proceeds with the rate of convergence analysis. To ascertain the convergence rate, we first derive a bound on the estimation errors, and then we show that a scaled sequence of the estimation errors converges weakly to a diffusion process through martingale averaging. Section 5 is devoted to iterate averaging. Section 6 provides a case study on an adaptive filtering type algorithm, and concludes the paper with further remarks.

2 Problem Formulation and Stochastic Approximation/Optimization Algorithm

This section presents the problem formulation and the stochastic approximation/optimization algorithm. A main feature of this algorithm is that it includes a discrete-time Markov chain. Compared with the variation of the updates represented by the step size of the algorithm, the Markov chain is rapidly varying. Throughout the paper, we use K to denote a generic positive constant whose values may vary for different appearances. For $z \in \mathbb{R}^{d\times\iota}$ and $d, \iota \geq 1$, z' denotes its transpose.

2.1 Markov Chain θ_n

We will use the following assumption throughout the paper. It characterizes the time-varying parameter as a two-time-scale homogeneous Markov chain with a finite state space.

(A1) Let $\{\theta_n\}$ be a discrete-time Markov chain with finite state space

$$\mathcal{M} = \{1, \ldots, m_0\}, \tag{2.1}$$

and transition probability matrix

$$P^\varepsilon = P + \varepsilon Q, \tag{2.2}$$

where $\varepsilon > 0$ is a small parameter, P is an $m_0 \times m_0$ irreducible and aperiodic transition probability matrix, and $Q = (q_{ij}) \in \mathbb{R}^{m_0 \times m_0}$ is a generator of a continuous-time Markov chain (i.e., Q satisfies $q_{ij} \geq 0$ for $i \neq j$ and $\sum_{j=1}^{m_0} q_{ij} = 0$ for each $i = 1, \ldots, m_0$).

Note that the underlying Markov chain $\{\theta_n\}$ is in fact ε-dependent. Thus it should have been written as $\{\theta_n^\varepsilon\}$. We suppress the ε-dependence for notational simplicity. Also, for simplicity, suppose that the initial distribution $P(\theta_0 = i) = p_{0,i}$ is independent of ε for each $i = 1, \ldots, m_0$, where $p_{0,i} \geq 0$ and $\sum_{i=1}^{m_0} p_{0,i} = 1$.

The irreducibility and aperiodicity of P imply the existence of the associated stationary distribution of the Markov chain associated with the transition matrix P. Denote the stationary distribution by $\nu = (\nu_1, \ldots, \nu_{m_0})$. For the Markov chain θ_n, denote the state probability vector p_n^ε by

$$p_n^\varepsilon = (P(\theta_n = 1), \ldots, P(\theta_n = m_0)) \in \mathbb{R}^{1 \times m_0}, \tag{2.3}$$

and denote the n-step transition probability matrix by $(P^\varepsilon)^n$. For $0 \leq n \leq T/\varepsilon = O(1/\varepsilon)$ and for $T > 0$, it can be shown that p_n^ε converges to the stationary distribution ν, and that the time of the Markov chain θ_n spends in a state $i \in \mathcal{M}$ can be approximated by ν_i in a suitable way. In addition, similar results also hold for the transition matrices. These approximations are formalized in the following lemma.

Lemma 1. *Under conditions* (A1)*, for some* $0 < \lambda < 1$*, and*

(i) *for* $T > 0$*, and* $0 \leq n \leq T/\varepsilon$*,*

$$p_n^\varepsilon = \nu + O(\varepsilon + \lambda^n); \tag{2.4}$$

in addition, for some $0 < n_0 < n$*,*

$$(P^\varepsilon)^{n-n_0} = 1\!\!1_{m_0}\nu \ + \ O(\varepsilon + \lambda^{n-n_0}), \tag{2.5}$$

where in (2.4) *and* (2.5)*, the bounds hold uniformly in* $0 \leq n \leq T/\varepsilon$*;*

(ii) *for all* $n \geq 0$*,*

$$p_n^\varepsilon = \nu + O(\varepsilon^2 n + \varepsilon + \lambda^n). \tag{2.6}$$

Proof. The proof of (i) is a simplified version of that of [32, Theorem 3.11], and the proof of (ii) is that of [32, Proposition 6.6]. □

Remark 1. Note that in statement (i) of Lemma 1, n is chosen to be in the range of $0 \leq n \leq O(1/\varepsilon)$, whereas in (ii), this range is removed. Assertion (ii) will be needed in the diffusion approximation for the rate of convergence study in Section 4.

2.2 Stochastic Approximation/Optimization Algorithm

Let $f(\cdot,\cdot,\cdot) : \mathbb{R}^d \times \mathbb{R}^d \times \mathcal{M} \mapsto \mathbb{R}^d$, $\{\xi_n\}$ be a sequence of $\mathbb{R}^d$-valued noise independent of the Markov chain $\{\theta_n\}$ (more precise conditions will be stated in Section 3.1), $\mu > 0$ be a constant (but small) step size, and $x_n \in \mathbb{R}^d$. Consider a stochastic approximation/optimization algorithm of the form

$$\begin{cases} x_{n+1} = x_n + \mu f(x_n, \xi_n, \theta_n), \\ p^\varepsilon_{n+1} = p^\varepsilon_n P^\varepsilon, \quad p^\varepsilon_0 = p_0. \end{cases} \tag{2.7}$$

The first equation in (2.7) is the stochastic approximation update of the parameter estimate x_n. The second equation, denotes the evolution over time of the state probabilities p^ε_n (2.3) of the Markov chain θ_n. For example, in the context of the stochastic optimization problem outlined in Section 1.1, $f(x_n, \xi_n, \theta_n)$ represents the noisy gradient estimate of $EF(x_n, \xi_n, \theta_n)$. Our main motivation for considering a constant-stepsize algorithm is because we are interested in adaptive optimization, i.e., algorithms that can track a time varying optimal parameter; see [1,15,17] for further motivation.

Throughout the paper, we assume that the step size μ satisfies $\mu \ll \varepsilon$ and that $\varepsilon = \varepsilon(\mu)$ such that $\varepsilon(\mu) \to 0$ as $\mu \to 0$. That is, we assume that the dynamics of the underlying parameter θ_n (the Markov chain with transition probability matrix P^ε) evolve on a time scale that is an order of magnitude faster than the dynamics of the stochastic approximation/optimization algorithm. An example of this two-time-scale behavior is obtained by choosing $\varepsilon = O(\mu^\gamma)$ in (2.7) with $0 < \gamma < 1$.

It is important to note that the numerical implementation of the above algorithm does not require assumption (A1) or knowledge of the explicit dynamics of ξ_n or θ_n. The remainder of this paper analyzes the asymptotic properties of (2.7) – it is in this analysis that we use (A1) and other assumptions listed below.

3 Convergence Analysis

This section is devoted to proving the convergence of the algorithm (2.7). In Section 3.1, we take a continuous-time interpolation of the iterates, and show using weak convergence methods that the limit dynamics satisfy an ordinary differential equation. In Section 3.2, we derive a moment bound on the discrete time iterates x_n via use of a perturbed Liapunov function approach. Finally, using a stability argument, we obtain the convergence to the asymptotically stable point of the ODE.

3.1 Mean Dynamics: ODE Limit

Weak convergence is a generalization to function space of the concept of convergence in distribution of random variables. Establishing weak convergence of the iterates generated by the algorithm (2.7) requires verification of tightness, extraction of a weakly convergent subsequence, and characterization of the limit process. We will prove below that the iterates generated by the algorithm (2.7) converge weakly (in a limiting sense made precise below) to the trajectory of an ODE (ordinary differential equation) [17]. In order to show this weak convergence, it is convenient to switch to continuous time as follows: First, construct a sequence of piecewise constant continuous time trajectories indexed by μ by interpolating the discrete time iterates x_n generated by the stochastic approximation algorithm (2.7) as:

$$x^{\mu}(t) = x_n,\ t \in [\mu n, \mu n + \mu). \tag{3.1}$$

From an electrical engineering point of view, this interpolation is merely equivalent to applying a zero-order hold circuit to the discrete time sequence x_n, resulting in the continuous time trajectory $x^{\mu}(t)$.

Next we make the following assumptions:

(A2) The following conditions hold:

(a) For each ξ and each θ, $f(\cdot, \xi, \theta)$ is a continuous function; there exists $\widehat{f}(\cdot, \cdot)$ such that for each x and each θ, $\widehat{f}(x, \theta) = Ef(x, \xi_n, \theta)$.

(b) $\{\xi_n\}$ is a bounded stationary process taking values in $\mathbb{R}^d$ and being independent of $\{\theta_n\}$; for any $m \geq 0$, as $n \to \infty$,

$$\frac{1}{n} \sum_{k=m}^{m+n-1} E_m f(x, \xi_k, i) \to \widehat{f}(x, i) \ \text{ in probability} \tag{3.2}$$

for each x and each $i \in \mathcal{M}$, where E_m denotes the conditional expectation with respect to $\mathcal{F}_m$, the σ-algebra generated by $\{x_0, \xi_k, \theta_k : k < m, \theta_m\}$.

(c) Define $\overline{f} : \mathbb{R}^d \to \mathbb{R}^d$ as

$$\overline{f}(x) = \sum_{i=1}^{m_0} \widehat{f}(x, i)\nu_i. \tag{3.3}$$

The initial value problem

$$\dot{x} = \overline{f}(x(t)),\ x(0) = x^0 \tag{3.4}$$

has a unique solution for each initial condition x^0.

Assumption (A2) provides certain regularity conditions on the observation noise and the function under consideration. The noise condition (3.2) is easily verified for a ϕ-mixing process since mixing implies ergodicity. Note that with the conditional expectation presence, the condition is even weaker. The noise condition can be further relaxed. For instance, we may assume that $f(x, \xi, \theta) = f_0(x, \widetilde{\xi}, \theta) + f_1(x, \theta)\widehat{\xi}$ such that $\{\widetilde{\xi}_n\}$ is a sequence of bounded noise satisfying the conditions as in (A2), that $f_0(\cdot)$ satisfies the conditions as in (A2), and $f_1(\cdot)$ is a bounded and continuous function for each $i \in \mathcal{M}$, that $\{\widehat{\xi}_n\}$ is a stationary sequence of possibly unbounded noise being independent of $\{\theta_n\}$ and $\{\widetilde{\xi}_n\}$ with $E\widehat{\xi}_n = 0$ and $E|\widehat{\xi}_n|^{2+\gamma} < \infty$ for some $\gamma > 0$, and that for each m,

$$\frac{1}{n}\sum_{k=m}^{m+n-1} E_m \widehat{\xi}_k \to 0 \ \text{ in probability as } \ n \to \infty.$$

Under such conditions, the subsequent convergence analysis carries over. We use the current condition mainly for notational simplicity. To proceed, we present a theorem that gives the mean dynamics of the iterates.

Theorem 1. *Under conditions* (A1) *and* (A2), $x^\mu(\cdot)$ *converges weakly to* $x(\cdot)$*, which is the unique solution of* (3.4)*.*

Remark 2. Its proof is divided into two parts presented in the next two subsections. A pertinent way of carrying out the analysis is to use an N-truncation of $x^\mu(\cdot)$ (see [17]). Since we will use such an approach in the rate of convergence study, we simply omit the truncation step in this section for ease of presentation. Without loss of generality, we assume that $x^\mu(\cdot)$ is bounded throughout the rest of this section.

3.1.1 Tightness of $\{x^\mu(\cdot)\}$

Use $D([0,\infty) : \mathbb{R}^d)$ to denote the space of functions defined on $[0,\infty)$ taking values in $\mathbb{R}^d$ that are right continuous and have left limits endowed with the Skorohod topology; see [17] and the references therein. Then we obtain the following lemma.

Lemma 2. *Under conditions* (A1) *and* (A2), $\{x^\mu(\cdot)\}$ *is tight in* $D([0,\infty) : \mathbb{R}^d)$*.*

Proof. We use the tightness criteria [14, p. 47]. For any $\delta > 0$ and $0 < s \le \delta$,

$$\lim_{\delta\to 0}\limsup_{\mu\to 0} E|x^\mu(t+s) - x^\mu(t)|^2 = 0. \tag{3.5}$$

In fact,

$$x^\mu(t+s) - x^\mu(t) = \mu \sum_{k=t/\mu}^{(t+s)/\mu-1} f(x_k,\xi_k,\theta_k). \tag{3.6}$$

Then by the boundedness of x_n (see Remark 2) and hence the boundedness of $f(x_n,\xi_n,\theta_n)$,

$$\begin{aligned} E|x^\mu(t+s) - x^\mu(t)|^2 &= E\big(\mu \sum_{k=t/\mu}^{(t+s)/\mu-1} f(x_k,\xi_k,\theta_k)\big)'\big(\mu \sum_{j=t/\mu}^{(t+s)/\mu-1} f(x_j,\xi_j,\theta_j)\big) \\ &\le K\mu^2 \left(\frac{t+s}{\mu} - \frac{t}{\mu}\right)^2 = O(s^2). \end{aligned} \tag{3.7}$$

Taking $\limsup_{\mu\to 0}$ followed by $\lim_{\delta\to 0}$, (3.5) is verified. Thus $\{x^\mu(\cdot)\}$ is tight in $D([0,\infty) : \mathbb{R}^d)$. □