Math. Program., Ser. B
DOI 10.1007/s10107-007-0145-1

FULL LENGTH PAPER

How does a stochastic optimization/approximation algorithm adapt to a randomly evolving optimum/root with jump Markov sample paths

G. Yin · C. Ion · V. Krishnamurthy

Received: 17 June 2005 / Accepted: 8 December 2005

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 observations contain a (nonsmooth) jump process modeled by a discrete-time Markov chain whose transition frequency varies much faster than the adaptation rate

Dedicated to Professor Boris Polyak on the occasion of his 70th birthday.

The research of G. Yin was supported in part by the National Science Foundation under DMS-0603287, in part by the National Security Agency under MSPF-068-029, 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 (✉) · C. Ion
Department of Mathematics, Wayne State University, Detroit, MI 48202, USA
e-mail: gyin@math.wayne.edu

C. Ion
e-mail: cion@math.wayne.edu

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

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 [1,3,4,13,12,32] 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 $n = 0, 1, 2, \ldots$ 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 $\{1, 2, \ldots, m_0\}$, 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 "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 [7,16,22,27] 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 [11] 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 [29,32] 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 [29,32], 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. Sect. 2 presents the precise formulation and the recursive algorithm. Sect. 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. Sect. 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. Sect. 5 is devoted to iterate averaging. Sect. 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 $0 \le n \le T/\varepsilon = O(1/\varepsilon)$ 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* $0 \le n \le T/\varepsilon$*,*

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

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

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

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

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

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

Proof The proof of (i) is a simplified version of that of [31, Theorem 3.11], and the proof of (ii) is that of [31, Proposition 6.6]. □

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

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

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_n^ε (2.3) of the Markov chain θ_n. For example, in the context of the stochastic optimization problem outlined in Sect. 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 Sect. 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 Sect. 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 $f(x_n,\xi_n,\theta_n)$,

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

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

3.1.2 Characterization of the limit

Continuation of proof of Theorem 1. Since $\{x^\mu(\cdot)\}$ is tight, by virtue of the Prohorov's theorem, we may extract a weakly convergent subsequence. Select such a subsequence of $x^\mu(\cdot)$ and still index it by μ (for notational simplicity) with the limit denoted by $x(\cdot)$. It can be demonstrated that $x(\cdot)$ has continuous paths w.p.1. By virtue of the Skorohod representation, with a slight abuse of notation, $x^\mu(\cdot)$ converges to the limit process $x(\cdot)$ w.p.1 and the convergence is uniform on any compact interval. We proceed to characterize the limit $x(\cdot)$, which is the solution of a mean ODE.

Recall the assumed boundedness of the iterates x_n; see Remark 2. Choose a sequence of positive integers n_μ such that $n_\mu \to \infty$ as $\mu \to 0$, but $\Delta_\mu = \mu n_\mu \to 0$. Partition

$[t,\ t+s]$ into subintervals, and rewrite the sum in (3.6) as

$$\begin{aligned} x^{\mu}(t+s)-x^{\mu}(t) &= \mu \sum_{k=t/\mu}^{(t+s)/\mu-1} f(x_k,\xi_k,\theta_k) \\ &= \Delta_{\mu}\frac{1}{n_{\mu}} \sum_{k=t/\mu}^{(t+s)/\mu-1} f(x_k,\xi_k,\theta_k) \\ &= \sum_{ln_{\mu}=t/\mu}^{(t+s)/\mu-1} \Delta_{\mu}\frac{1}{n_{\mu}} \sum_{k=ln_{\mu}}^{ln_{\mu}+n_{\mu}-1} f(x_k,\xi_k,\theta_k). \end{aligned}$$

In the above and henceforth, we treat t/μ and $(t+s)/\mu$ as integers without loss of generality since we can always take their integer parts. Thus, we have

$$\begin{aligned} & x^{\mu}(t+s)-x^{\mu}(t) \\ &= \sum_{ln_{\mu}=t/\mu}^{(t+s)/\mu-1} \Delta_{\mu}\frac{1}{n_{\mu}} \sum_{k=ln_{\mu}}^{ln_{\mu}+n_{\mu}-1} f(x_{ln_{\mu}},\xi_k,\theta_k) \\ &\quad + \sum_{ln_{\mu}=t/\mu}^{(t+s)/\mu-1} \Delta_{\mu}\frac{1}{n_{\mu}} \sum_{k=ln_{\mu}}^{ln_{\mu}+n_{\mu}-1} [f(x_k,\xi_k,\theta_k)-f(x_{ln_{\mu}},\xi_k,\theta_k)] \\ &= \sum_{ln_{\mu}=t/\mu}^{(t+s)/\mu-1} \Delta_{\mu}\frac{1}{n_{\mu}} \sum_{k=ln_{\mu}}^{ln_{\mu}+n_{\mu}-1} f(x_{ln_{\mu}},\xi_k,\theta_k)+o(1), \\ &= \sum_{ln_{\mu}=t/\mu}^{(t+s)/\mu-1} \Delta_{\mu}\frac{1}{n_{\mu}} \sum_{i=1}^{m_0} \sum_{k=ln_{\mu}}^{ln_{\mu}+n_{\mu}-1} f(x_{ln_{\mu}},\xi_k,i)[I_{\{\theta_k=i\}}-\nu_i] \\ &\quad + \sum_{ln_{\mu}=t/\mu}^{(t+s)/\mu-1} \Delta_{\mu}\frac{1}{n_{\mu}} \sum_{i=1}^{m_0} \sum_{k=ln_{\mu}}^{ln_{\mu}+n_{\mu}-1} f(x_{ln_{\mu}},\xi_k,i)\nu_i+o(1), \end{aligned} \tag{3.8}$$

since $f(\cdot,\xi,\theta)$ is continuous, where $o(1)\to 0$ in probability uniformly in t. To proceed, we first derive a lemma.

Lemma 3 *For each $i\in\mathcal{M}$,*

$$E\left|\sum_{ln_{\mu}=t/\mu}^{(t+s)/\mu-1} \sum_{k=ln_{\mu}}^{ln_{\mu}+n_{\mu}-1} \Delta_{\mu}\frac{1}{n_{\mu}} f(x_{ln_{\mu}},\xi_k,i)[I_{\{\theta_k=i\}}-\nu_i]\right| \to 0 \ \textit{as}\ \ \mu\longrightarrow 0. \tag{3.9}$$

Proof of Lemma 3 By a partial summation,

$$
\begin{aligned}
&\sum_{ln_\mu=t/\mu}^{(t+s)/\mu-1}\ \sum_{k=ln_\mu}^{ln_\mu+n_\mu-1} \Delta_\mu \frac{1}{n_\mu} f(x_{ln_\mu}, \xi_k, i)[I_{\{\theta_k=i\}} - \nu_i] \\
&= \sum_{ln_\mu=t/\mu}^{(t+s)/\mu-1} \Delta_\mu \frac{1}{n_\mu} f(x_{ln_\mu}, \xi_{ln_\mu+n_\mu-1}, i) \sum_{k=ln_\mu}^{ln_\mu+n_\mu-1} [I_{\{\theta_k=i\}} - \nu_i] \\
&\quad + \sum_{ln_\mu=t/\mu}^{(t+s)/\mu-1}\ \sum_{k=ln_\mu-1}^{ln_\mu+n_\mu-2} \Delta_\mu \frac{1}{n_\mu}[f(x_{ln_\mu}, \xi_k, i) - f(x_{ln_\mu}, \xi_{k+1}, i)] \sum_{k_1=0}^{k} [I_{\{\theta_{k_1}=i\}} - \nu_i].
\end{aligned}
\tag{3.10}
$$

Next, for $k > k_1$ and $i \in \mathcal{M}$,

$$
\begin{aligned}
&E_{ln_\mu}[I_{\{\theta_k=i\}} - \nu_i][I_{\{\theta_{k_1}=i\}} - \nu_i] \\
&\quad = E[I_{\{\theta_k=i,\,\theta_{k_1}=i\}} - \nu_i I_{\{\theta_k=i\}} - \nu_i I_{\{\theta_{k_1}=i\}} + \nu_i^2 \mid \theta_{ln_\mu}] \\
&\quad = \sum_{i_0=1}^{m_0} \left(\nu_i + O(\varepsilon + \lambda^{k_1-ln_\mu})\right)\left(\nu_i + O(\varepsilon + \lambda^{k-k_1})\right) \\
&\qquad - \nu_i \sum_{i_0=1}^{m_0} \left[\left(\nu_i + O(\varepsilon + \lambda^{k-ln_\mu})\right) + \left(\nu_i + O(\varepsilon + \lambda^{k_1-ln_\mu})\right)\right] I_{\{\theta_{ln_\mu}=i_0\}} + \nu_i^2 \\
&\quad = \sum_{i_0=1}^{m_0} O(\varepsilon + \lambda^{k-k_1} + \lambda^{k-ln_\mu}) I_{\{\theta_{ln_\mu}=i_0\}}.
\end{aligned}
$$

Since $\lambda^{k-k_1} > \lambda^{k-ln_\mu}$,

$$
E_{ln_\mu}[I_{\{\theta_k=i\}} - \nu_i][I_{\{\theta_{k_1}=i\}} - \nu_i] = O(\varepsilon + \lambda^{k-k_1}). \tag{3.11}
$$

Similarly, for $k < k_1$ and $i \in \mathcal{M}$,

$$
E_{ln_\mu}[I_{\{\theta_k=i\}} - \nu_i][I_{\{\theta_{k_1}=i\}} - \nu_i] = O(\varepsilon + \lambda^{k_1-k}). \tag{3.12}
$$

Note that

$$
\begin{aligned}
L^{\varepsilon,\mu} &\stackrel{\text{def}}{=} E\left|\sum_{ln_\mu=t/\mu}^{(t+s)/\mu-1} \Delta_\mu \frac{1}{n_\mu} f(x_{ln_\mu}, \xi_{ln_\mu-1}, i) \sum_{k=ln_\mu}^{ln_\mu+n_\mu-1} [I_{\{\theta_k=i\}} - \nu_i]\right| \\
&\le E \sum_{ln_\mu=t/\mu}^{(t+s)/\mu-1} \Delta_\mu \frac{1}{n_\mu} |f(x_{ln_\mu}, \xi_{ln_\mu+n_\mu-1}, i)| \left|\sum_{k=ln_\mu}^{ln_\mu+n_\mu-1} [I_{\{\theta_k=i\}} - \nu_i]\right| \\
&\le K \sum_{ln_\mu=t/\mu}^{(t+s)/\mu-1} \Delta_\mu \frac{1}{n_\mu} E^{1/2} \left|\sum_{k=ln_\mu}^{ln_\mu+n_\mu-1} [I_{\{\theta_k=i\}} - \nu_i]\right|^2.
\end{aligned}
$$

Thus, estimates (3.11) and (3.12), and the assumed boundedness of x_n and $\varepsilon = \varepsilon(\mu)$ (with $\varepsilon(\mu) \to 0$ as $\mu \to 0$) lead to

$$\begin{aligned}
&\widehat{L}^{\varepsilon,\mu} \\
&\le KE \sum_{ln_\mu=t/\mu}^{(t+s)/\mu-1} \Delta_\mu \frac{1}{n_\mu} \left[\sum_{k=ln_\mu}^{ln_\mu+n_\mu-1} \sum_{k_1=ln_\mu}^{ln_\mu+n_\mu-1} E_{ln_\mu}[I_{\{\theta_k=i\}} - \nu_i][I_{\{\theta_{k_1}=i\}} - \nu_i] \right]^{\frac{1}{2}} \\
&\le K \sum_{ln_\mu=t/\mu}^{(t+s)/\mu-1} \frac{\Delta_\mu}{n_\mu} \left[\sum_{k=ln_\mu}^{ln_\mu+n_\mu-1} \sum_{k_1<k} O(\varepsilon + \lambda^{k-k_1}) \sum_{k_1=ln_\mu}^{ln_\mu+n_\mu-1} \sum_{k<k_1} O(\varepsilon + \lambda^{k_1-k}) \right. \\
&\qquad\qquad \left. + \sum_{k_1=ln_\mu}^{ln_\mu+n_\mu-1} O(1) \right]^{\frac{1}{2}} \\
&= \sum_{ln_\mu=t/\mu}^{(t+s)/\mu-1} \frac{\Delta_\mu}{n_\mu} \Big(O(\varepsilon)O(n_\mu^2) + O(n_\mu) \Big)^{1/2} \\
&= \sum_{ln_\mu=t/\mu}^{(t+s)/\mu-1} \Delta_\mu \left[O(\varepsilon^{1/2}) + O\left(\frac{1}{n_\mu^{1/2}} \right) \right] \to 0 \quad \text{as} \quad \mu \to 0.
\end{aligned}$$

This completes the proof of Lemma 3. □

Similarly, it can be shown

$$\begin{aligned}
E\Bigg| &\sum_{ln_\mu=t/\mu}^{(t+s)/\mu-1} \frac{\Delta_\mu}{n_\mu} \sum_{k=ln_\mu-1}^{ln_\mu+n_\mu-2} [f(x_{ln_\mu}, \xi_k, i) - f(x_{ln_\mu}, \xi_{k+1}, i)] \\
&\times \sum_{k_1=ln_\mu}^{k} [I_{\{\theta_{k_1}=i\}} - \nu_i] \Bigg| \to 0, \quad \text{as } \mu \to 0.
\end{aligned} \tag{3.13}$$

What we have demonstrated is that $x^\mu(t+s) - x^\mu(t)$ in (3.8) can be approximated by the terms in its last line. To proceed, for $\eta > 0$ let $\{B_\iota^\eta : \iota \le \iota_\eta\}$ be a finite partition of the range of $x^\mu(l\Delta_\mu)$, and pick out any point $x_\iota^\eta \in B_\iota^{\eta,0}$ in the interior of B_ι^η. It can be shown as in [17],

$$\begin{aligned}
&\sum_{ln_\mu=t/\mu}^{(t+s)/\mu-1} \Delta_\mu \frac{1}{n_\mu} \sum_{i=1}^{m_0} \sum_{k=ln_\mu}^{ln_\mu+n_\mu-1} f(x_{ln_\mu}, \xi_k, i)\nu_i \\
&\quad = \sum_{ln_\mu=t/\mu}^{(t+s)/\mu-1} \Delta_\mu \frac{1}{n_\mu} \sum_{i=1}^{m_0} \sum_{k=ln_\mu}^{ln_\mu+n_\mu-1} f(x^\mu(l\Delta_\mu), \xi_k, i)\nu_i \\
&\quad = \sum_{\iota=1}^{\iota_\eta} \sum_{ln_\mu=t/\mu}^{(t+s)/\mu-1} \Delta_\mu \frac{1}{n_\mu} \sum_{i=1}^{m_0} \sum_{k=ln_\mu}^{ln_\mu+n_\mu-1} f(x_\iota^\eta, \xi_k, i) I_{\{x^\mu(l\Delta_\mu)\in B_\iota^\eta\}} \nu_i + o(1),
\end{aligned} \tag{3.14}$$

where $o(1) \to 0$ in probability uniformly in t. Letting $l\Delta_\mu \to u$ as $\mu \to 0$, using the weak convergence of $x^\mu(l\Delta_\mu)$ to $x(u)$ and the Skorohod representation, and by (3.2), for each $1 \le \iota \le \iota_\eta$,

$$\frac{1}{n_\mu} \sum_{k=ln_\mu}^{ln_\mu+n_\mu-1} f\left(x_\iota^\eta, \xi_k, i\right) I_{\{x^\mu(l\Delta_\mu)\in B_\iota^\eta\}} \to \overline{f}(x(u)) I_{\{x(u)\in B_\iota^\eta\}}.$$

It follows that the limit $x(\cdot)$ satisfies (3.4). The proof of the theorem is concluded. □

Remark 3 Several points are worth of mentioning. First, the continuity of $f(\cdot, \xi, \theta)$ can be relaxed to certain "weak continuity" (e.g., continuity in expectation). Thus, for example, indicator type of functions can be treated. Second, in lieu of (3.2), suppose that there are $\mathbb{F}_i(x)$, set-valued maps that are upper semi-continuous for each x and each $i \in \mathcal{M}$, and that

$$\rho\left[\frac{1}{n}\sum_{k=m}^{m+n-1} E_m f(x, \xi_k, i), \mathbb{F}_i(x)\right] \to 0 \text{ in probability as } \mu \to 0,$$

where $\rho(\cdot, \cdot)$ is the usual distance function defined by

$$\rho(x, G) \stackrel{\text{def}}{=} \text{dist}\,(x, G) = \inf_{y\in G} |x - y|.$$

Then the limit ordinary differential equation is replaced by a differential inclusion in the form

$$\dot{x} \in \sum_{i=1}^{m_0} \mathbb{F}_i(x)\nu_i.$$

3.2 Moment bounds: a perturbed Liapunov function approach

In Sect. 3.1 we studied the behavior of the continuous time interpolated trajectory $x^\mu(t)$ of the discrete time sequence of estimates x_n in the limit as $\mu \to 0$. In this section we study the behavior of the estimated sequence x_n for a small but finite step size μ. In particular, we examine the sequence $\{V(x_n)\}$ to establish a moment bound for the sequence of iterates $\{x_n\}$, where $V(\cdot)$ is a Liapunov function. We show that $EV(x_n) = O(1)$ by a stability argument which requires the following assumption.

(A3) (a) There is a unique x_* satisfying $\widehat{f}(x_*, i) = 0$ for each $i \in \mathcal{M}$.
(b) $f(\cdot, \xi, \theta)$ is twice continuously differentiable for each ξ and each θ, and $f_{xx}(\cdot, \xi, \theta)$ is uniformly bounded.
(c) For each x, each $i_1, i_2, \theta \in \mathcal{M}$, and each ξ,

$$\sum_{j=n}^{\infty} |E_n f(x, \xi_j, \theta) - \widehat{f}(x, \theta)| \le K(1 + V^{1/2}(x)), \tag{3.15}$$

$$\sum_{k\geq j}^{\infty} |Ef(x_*, \xi_j, i_1)f'(x_*, \xi_k, i_2) - E\widehat{f}(x_*, i_1)E\widehat{f}(x_*, i_2)| < \infty, \quad (3.16)$$

$$\sum_{j=n}^{\infty} |E_n \nabla_x [V_x'(x)(f(x, \xi_j, \theta) - \widehat{f}(x, \theta))f(x, \xi, \theta)]| \leq K(1 + V(x)),$$
$$\sum_{j=n}^{\infty} |E_n \nabla_{xx} [V_x'(x)(f(x, \xi_j, \theta) - \widehat{f}(x, \theta))]| \leq K, \quad (3.17)$$

where $\nabla_x b$ and $\nabla_{xx} b$ denote the gradient and the Hessian of b with respect to the variable x, respectively.

(d) There is a three times continuously differentiable Liapunov function $V(\cdot) : \mathbb{R}^d \mapsto \mathbb{R}$ such that
  (i) $V(x) \geq 0$, $V(x) \to \infty$ as $|x| \to \infty$, $|V_{xx}(\cdot)|$ and $|V_{xxx}(\cdot)|$ are uniformly bounded, $|V_x(x)| \leq K\left(1 + V^{1/2}(x)\right)$ and $|f(x, \xi, \theta)|^2 \leq K(1 + V(x))$ for each $\xi \in \mathbb{R}^d$, and $\theta \in \mathcal{M}$.
  (ii) $V_x'(x)\widehat{f}(x, \theta) \leq -\lambda_0 V(x)$ for some $\lambda_0 > 0$, and for each $\theta \in \mathcal{M}$.

Remark 4 By condition (A3) (a) and in view of (3.3), $\overline{f}(x_*) = 0$. Thus, x_* is an asymptotically stable point of (3.4).

The smoothness condition in (A3) (b) implies that $\widehat{f}(\cdot, \theta)$ and $\overline{f}(\cdot)$ are twice continuously differentiable for each θ, and that $\widehat{f}_{xx}(\cdot, \theta)$ and $\overline{f}_{xx}(\cdot)$ are uniformly bounded.

Condition (A3) (c) is a condition on the dependence structure of the noise and the growth of the function $f(x, \xi, \theta)$ and hence $\widehat{f}(x, \theta)$. The dependence is modeled after the mixing process. For example, the sequences being bounded uniform mixing with mixing rate ψ_k satisfying $\sum_{k=0}^{\infty} \psi_k < \infty$ together with the growth of $f(x, \xi, \theta)$ yields (3.15)–(3.17).

Condition (A3) (d) is a stability condition on the limit dynamics of (3.4). This condition has been used frequently in analyzing stochastic approximation type algorithms. Although the existence of the Liapunov function is required, its precise form need not be known. The growth condition on the function $f(x, \xi, \theta)$, implies that

$$|\widehat{f}(x, \theta)|^2 \leq K(1 + V(x)) \text{ and } |\overline{f}(x)|^2 \leq K(1 + V(x)). \quad (3.18)$$

Theorem 2 *Assume* (A1)–(A3). *Then*

$$EV(x_n) = O(1). \quad (3.19)$$

Proof To prove the result, we use the perturbed Liapunov function approach with the Liapunov function $V(\cdot)$ defined in (A3) (d). Direct calculations and assumption (A3)

(d) lead to

$$
\begin{aligned}
&E_n V(x_{n+1}) - V(x_n) \\
&\quad = E_n \mu V_x'(x_n) f(x_n, \xi_n, \theta_n) + E_n \mu^2 \frac{1}{2} f(x_n, \xi_n, \theta_n)' V_{xx}(x_n^+) f(x_n, \xi_n, \theta_n) \\
&\quad \le -\mu\lambda_0 V(x_n) + E_n \mu V_x'(x_n)[f(x_n, \xi_n, \theta_n) - \widehat{f}(x_n, \theta_n)] + O(\mu^2)(1 + V(x_n)),
\end{aligned}
\tag{3.20}
$$

where x_n^+ is on the line segment joining x_n and x_{n+1}. In the above, the second line follows from a Taylor expansion and the use of (2.7). The second term in the last line of (3.20) is an extra term that we wish to get rid of. We proceed to average it out using a perturbation defined by

$$
V_1^\mu(x, \theta, n) \overset{\text{def}}{=} \mu \sum_{j=n}^{\infty} E_n V_x'(x)[f(x, \xi_j, \theta) - \widehat{f}(x, \theta)] \quad \text{for } x \in \mathbb{R}^d, \ \theta \in \mathcal{M}.
$$

Note that $V_1^\mu(x, \xi, \theta)$ is well defined. In fact, using (A3) (d) and (3.15), it is readily seen that for each $(x, \theta) \in \mathbb{R}^d \times \mathcal{M}$,

$$
\left| V_1^\mu(x, \theta, n) \right| \le O(\mu)(V(x) + 1) \tag{3.21}
$$

so the perturbation is small. Moreover,

$$
\begin{aligned}
&E_n V_1^\mu(x_{n+1}, \theta_{n+1}, n+1) - V_1^\mu(x_n, \theta_n, n) \\
&\quad = E_n\big[V_1^\mu(x_{n+1}, \theta_{n+1}, n+1) - V_1^\mu(x_{n+1}, \theta_n, n+1)\big] \\
&\qquad + E_n\big[V_1^\mu(x_{n+1}, \theta_n, n+1) - V_1^\mu(x_n, \theta_n, n+1)\big] \\
&\qquad + E_n V_1^\mu(x_n, \theta_n, n+1) - V_1^\mu(x_n, \theta_n, n).
\end{aligned}
\tag{3.22}
$$

For each $j \ge n$, each x and θ, let

$$
\Gamma(x, \xi_j, \theta) = f(x, \xi_j, \theta) - \widehat{f}(x, \theta). \tag{3.23}
$$

Then, by virtue of the Markov assumption, the structure of the transition probability matrix given by (2.2), and assumptions (A3), (3.15), and (3.23),

$$
\begin{aligned}
&E_n[V_1^\mu(x_{n+1}, \theta_{n+1}, n+1) - V_1^\mu(x_{n+1}, \theta_n, n+1)] \\
&\quad = \mu \sum_{j=n+1}^{\infty} E_n V_x'(x_{n+1})[\Gamma(x_{n+1}, \xi_j, \theta_{n+1}) - \Gamma(x_{n+1}, \xi_j, \theta_n)] \\
&\quad = \mu \sum_{j=n+1}^{\infty} E_n V_x'(x_n)[\Gamma(x_n, \xi_j, \theta_{n+1}) - \Gamma(x_n, \xi_j, \theta_n)] + O(\mu^2)(1 + V(x_n)) \\
&\quad = \mu \sum_{j=n+1}^{\infty} \sum_{i=1}^{m_0} \sum_{l=1}^{m_0} E_n V_x'(x_n)[\Gamma(x_n, \xi_j, l) p_{il}^\varepsilon - \Gamma(x_n, \xi_j, i)] I_{\{\theta_n = i\}} \\
&\qquad\qquad + O(\mu^2)(1 + V(x_n)) \\
&\quad = O(\mu\varepsilon + \mu^2 + \mu)(1 + V(x_n)) = O(\mu)(1 + V(x_n)).
\end{aligned}
\tag{3.24}
$$

Recall that $\mu \ll \varepsilon$. The term on the third line of (3.22) can be estimated using (3.15) and (3.17). In fact,

$$\begin{aligned}
&E_n[V_1^\mu(x_{n+1},\theta_n,n+1) - V_1^\mu(x_n,\theta_n,n+1)] \\
&= \mu \sum_{j=n+1}^{\infty} E_n[V_x'(x_{n+1}) - V_x'(x_n)]\Gamma(x_{n+1},\xi_j,\theta_n) \\
&\quad + \mu \sum_{j=n+1}^{\infty} E_n V_x'(x_n)[\Gamma(x_{n+1},\xi_j,\theta_n) - \Gamma(x_n,\xi_j,\theta_n)] \\
&= O(\mu^2)(1+V(x_n)). \qquad (3.25)
\end{aligned}$$

Moreover, for the last term in (3.22),

$$E_n V_1^\mu(x_n,\theta_n,n+1) - V_1^\mu(x_n,\theta_n,n) = -\mu E_n V_x'(x_n)[f(x_n,\xi_n,\theta_n) - \widehat{f}(x_n,\theta_n)]. \qquad (3.26)$$

Define

$$W(x,n) = V(x) + V_1^\mu(x,\theta,n).$$

Then, using the above estimates, we have

$$\begin{aligned}
E_n W(x_{n+1},n+1) - W(x_n,n) &= -\mu\lambda_0 V(x_n) + O(\mu)(1+V(x_n)) \\
&\le -\mu\lambda_0 W(x_n,n) + O(\mu)(1+W(x_n,n)). \qquad (3.27)
\end{aligned}$$

Therefore,

$$E_n W(x_{n+1},n+1) \le (1-\lambda_0\mu)W(x_n,n) + O(\mu) + O(\mu)W(x_n,n). \qquad (3.28)$$

Taking expectation and iterating on the resulting inequality yield

$$\begin{aligned}
EW(x_{n+1},n+1) &\le (1-\lambda_0\mu)^n EW(x_0,0) + O(\mu)\sum_{j=0}^{n}(1-\lambda_0\mu)^j \\
&\quad + O(\mu)\sum_{j=0}^{n} EW(x_j,j)(1-\lambda_0\mu)^{n-j}. \qquad (3.29)
\end{aligned}$$

For n large enough, $(1-\lambda_0\mu)^n = O(\mu)$, so it is bounded. In addition,

$$O(\mu)\sum_{j=0}^{n}(1-\lambda_0\mu)^j = O(1).$$

Thus,

$$EW(x_{n+1},n+1) \le K + O(\mu)\sum_{j=0}^{n} EW(x_j,j)(1-\lambda_0\mu)^{n-j} \qquad (3.30)$$

for some constant $K > 0$. The Gronwall inequality yields that

$$EW(x_{n+1}, n+1) \le K \exp\left(O(\mu) \sum_{j=0}^{n} (1-\lambda_0 \mu)^j \right) = O(1). \tag{3.31}$$

Replacing $W(x_{n+1}, n+1)$ by $V(x_{n+1})$ and in view of (3.21), we obtain

$$EV(x_{n+1}) = O(1). \tag{3.32}$$

This completes the proof. □

Corollary 1 *Under* (A1)–(A3)*, if the Liapunov function is locally quadratic, i.e., there exists a symmetric and positive definite $d \times d$ matrix A_0 such that*

$$V(x) = (x - x_*)' A_0 (x - x_*) + o(|x - x_*|^2), \tag{3.33}$$

then $\{x_n\}$ is tight. (*Recall x_* is defined in* (A3)(a).)

Remark 5 In the proof of the result, the perturbation $V_1^{\mu}(x, n)$ can be defined somewhat differently. As a result, the estimates can be strengthened and tighter bounds can be obtained. This will be provided in the rate of convergence study. We present the current version of proof to indicate how the Markov chain may be handled in places.

Corollary 2 *Assume conditions of Corollary 1 and suppose $t_\mu \to \infty$ as $\mu \to 0$. Then $x^{\mu}(\cdot + t_\mu)$ converges weakly* (*hence in probability*) *to x_* as $\mu \to 0$.*

Outline of proof. The proof is similar to that of Theorem 1. For any $T > 0$, consider the pair $(x^{\mu}(\cdot + t_\mu), x^{\mu}(\cdot + t_\mu - T))$, and extract a weakly convergent subsequence with limit denoted by $(x(\cdot), x_T(\cdot))$. Clearly, $x(0) = x_T(T)$. Moreover, the set of possible values of $\{x_T(0)\}$ is tight due to Corollary 1. The desired result then follows from the asymptotic stability of the ODE in (3.4).

4 Rate of convergence

This section is divided into two parts. In the first part, we derive moment bounds in terms of a Liapunov function. The second part then proceeds with a local analysis on obtaining a diffusion limit.

4.1 Error bounds

This section is on error bounds for the centered estimation error process $x_n - x_*$ where x_* is the unique asymptotically stable point of (3.4) defined in (A3)(a). The analysis is based on a Liapunov stability argument. The result is again obtained by using the perturbed Liapunov function method. We state this as the next assertion.

Theorem 3 *Choose N_μ to be a positive integer such that*

$$\left(1-\frac{\lambda_0\mu}{2}\right)^{N_\mu} \le K\mu \ \textit{for some} \ \ K>0. \tag{4.1}$$

Under the conditions of Theorem 1, for all $n \ge N_\mu$, $EV(x_n) = O(\mu)$.

Proof Since the proof is similar in spirit to that of Theorem 2, we only point out the main difference. In lieu of $V_1^\mu(x,\theta,n)$, define

$$U^\mu(x,n) = \mu \sum_{j=n}^{\infty} E_n V_x'(x)[f(x,\xi_j,\theta_j) - \widehat{f}(x,\theta_j)].$$

Then it can be shown that by the independence of $\{\xi_n\}$ and $\{\theta_n\}$, and (3.15),

$$\begin{aligned}
|U^\mu(x,n)| &= \left|\mu \sum_{j=n}^{\infty}\sum_{i=1}^{m_0} V_x'(x)[E_n f(x,\xi_j,i) - \widehat{f}(x,i)]E_n I_{\{\theta_j=i\}}\right| \\
&\le \mu \sum_{j=n}^{\infty}\sum_{i=1}^{m_0}\sum_{i_1=1}^{m_0} V_x'(x)|E_n f(x,\xi_j,i) - \widehat{f}(x,i)|P(\theta_j = i|\theta_{n=i_1}) \\
&\le K\mu(1+V(x)).
\end{aligned}$$

Define

$$\widetilde{W}(x,n) = V(x) + U^\mu(x,n).$$

Similar to the proof of Theorem 2, it can be shown that

$$\begin{aligned}
&E_n U^\mu(x_{n+1}, n+1) - U^\mu(x_n,n) \\
&\quad = [E_n U^\mu(x_n,n+1) - U^\mu(x_n,n)] + E_n[U^\mu(x_{n+1},n+1) - U^\mu(x_n,n+1)] \\
&\quad = -\mu V_x'(x_n)[f(x_n,\xi_n,\theta_n) - \widehat{f}(x_n,\theta_n)] + O(\mu)(1+V(x_n)).
\end{aligned}$$

Using the estimates above, we further obtain

$$E_n\widetilde{W}(x_{n+1},n+1) \le (1-\lambda_0\mu)\widetilde{W}(x_n,n) + O(\mu^2)(1+\widetilde{W}(x_n,n)).$$

Then for sufficiently small μ, $1-\lambda_0\mu + O(\mu^2) \le 1-(\lambda_0\mu/2)$, and

$$E\widetilde{W}(x_{n+1},n+1) \le \left(1-\frac{\lambda_0\mu}{2}\right)^n E\widetilde{W}(x_0,0) + O(\mu).$$

The desired result then follows similar as in the last section. □

Remark 6 Theorem 3 provides a bound on the estimation error sequence $\{x_n - x_*\}$, which reveals how the errors vary with respect to the step size μ. To better delineate the evolution of the error dynamics, define a scaled sequence of the estimation errors $\{u_n\}$ by

$$u_n = \frac{x_n - x_*}{\sqrt{\mu}}. \tag{4.2}$$

We proceed with the study on the asymptotics of u_n.

Corollary 3 *Assume conditions of Theorem 3. If the Liapunov function is locally quadratic as given in (3.33), then $\{u_n : n \geq N_\mu\}$ is tight with N_μ given by (4.1).*

4.2 Diffusion limit

We assume the conditions of Corollary 3 are satisfied and define $u^\mu(\cdot)$, the continuous-time interpolation of u_n by

$$u^\mu(t) = u_n, \quad \text{for } t \in [\mu(n - N_\mu), \mu(n - N_\mu) + \mu). \tag{4.3}$$

We aim to derive a diffusion limit for $u^\mu(\cdot)$ as $\mu \to 0$. The scaling factor $\sqrt{\mu}$ together with the asymptotic covariance of the limit process gives us a rate of convergence result. We linearize $f(x, \xi, \theta)$ about x_* and carry out a local analysis. In view of (2.7),

$$u_{n+1} = u_n + \sqrt{\mu} f(x_*, \xi_n, \theta_n) + \mu f_x(x_*, \xi_n, \theta_n) u_n + \mu^{3/2} O(|u_n|^2 | f_{xx}(x_n^*, \xi_n, \theta_n)|), \tag{4.4}$$

where x_n^* is on the line segment joining x_n and x_*. Recall that $\overline{f}(x_*) = 0$ since x_* is the unique asymptotically stable point of (3.4).
Remark: Our aim is to prove that as $\mu \to 0$, the interpolated error $u^\mu(t)$ converges weakly to the diffusion process

$$du(t) = \overline{f}_x(x_*) u(t) dt + \Sigma_w^{1/2} d\widehat{w}. \tag{4.5}$$

where Σ_w is defined in (4.15), (4.13). Before proceeding with the formal, it is worthwhile considering the following intuitive explanation. Consider the first-order time discretization of the stochastic differential equation (4.5) with discretization interval μ equal to the step size of the stochastic approximation algorithm 2.7. The resulting discretized system is

$$u_{n+1} = u_n + \sqrt{\mu} \Sigma_w^{1/2} v_n + \mu \overline{f}_x(x_*) u_n \tag{4.6}$$

where $v_n = \widehat{w}_{n+1} - \widehat{w}_n$ is a discrete-time white Gaussian noise. By comparing (4.6) with (4.4), it is intuitively clear that they are equivalent in distribution. In particular, by the stochastic averaging principle, $f_x(x_*, \xi_n, \theta_n)$ behaves as its average $\overline{f}_x(x_*)$, thus yielding the third term on the right hand side of (4.6). Similarly, by the principle of stochastic averaging, the second term on the right hand side of (4.4) behaves according to the square root of its covariance Σ_w—thus yielding the second time on the right hand side of (4.6). The proof below goes in the reverse direction, i.e., it shows that as the discretization interval $\mu \to 0$, the discrete-time process (4.4) converges weakly (in distribution) to the continuous-time process (4.6) under suitable technical assumptions. The key for this weak convergence of a sequence of discrete-time iterates to a continuous-time process is that the iterates should be well behaved in distribution. This well behavedness is captured by the tightness which roughly speaking states that the process is bounded in probability. The main tool used below to prove the weak convergence is the "martingale problem" formulation as discussed following (4.23).

Up to now, we have assumed only $\mu \ll \varepsilon$. To obtain the desired diffusion limit, we need more information on the step size parameter. In what follows, we assume $\varepsilon = \mu^{\frac{1}{2}+\Delta}$ for some $\Delta > 0$. Owing to definition (4.2), $\{u_n\}$ is not a priori bounded. We use an N-truncation device [17]. Let $N > 0$ be fixed but otherwise arbitrary, $S_N(z) = \{z \in \mathbb{R}^d : |z| \le N\}$ be the sphere with radius N, and $q^N(z)$ be a smooth function satisfying

$$q^N(z) = \begin{cases} 1, & \text{if } |z| \le N; \\ 0, & \text{if } |z| \ge N+1. \end{cases}$$

Note that $q^N(z)$ is smoothly connected between the sphere S_N and S_{N+1}. Now

$$\begin{aligned} u^N_{n+1} &= u^N_n + \mu f_x(x_*, \xi_n, \theta_n) u^N_n q^N\big(u^N_n\big) + \sqrt{\mu} f(x_*, \xi_n, \theta_n) \\ &\quad + O\big(\mu^{3/2} |u^N_n|^2 q^N\big(u^N_n\big)\big). \end{aligned} \tag{4.7}$$

Define $u^{\mu,N}(\cdot)$ as the continuous-time interpolation of u^N_n in $t \in [\mu(n - N_\mu), \mu(n - N_\mu + 1))$. It then follows that

$$\lim_{k_0 \to \infty} \limsup_{\mu \to 0} P\left(\sup_{0 \le t \le T} |u^{\mu,N}(t)| \ge k_0 \right) = 0 \quad \text{for each } T < \infty,$$

and that $u^{\mu,N}(\cdot)$ is a process that is equal to $u^\mu(\cdot)$ up until the first exit from S_N, and hence an N-truncation process of $u^\mu(\cdot)$ [17, p. 284]. To proceed, we work with $\{u^{\mu,N}(\cdot)\}$, and derive its tightness and weak convergence first. Finally, we let $N \to \infty$ to conclude the proof.

Lemma 4 *Under the conditions of Corollary 3, $\{u^{\mu,N}(\cdot)\}$ is tight in $D([0,\infty) : \mathbb{R}^d)$.*

Proof In view of (4.7), for any $\delta > 0$, and $t, s \ge 0$ with $s \le \delta$,

$$\begin{aligned} u^{\mu,N}(t+s) - u^{\mu,N}(t) &= \mu \sum_{k=t/\mu}^{(t+s)/\mu - 1} f_x(x_*, \xi_k, \theta_k) u^N_k q^N\big(u^N_k\big) \\ &\quad + \sqrt{\mu} \sum_{k=t/\mu}^{(t+s)/\mu - 1} f(x_*, \xi_k, \theta_k) + \mu^{3/2} \sum_{k=t/\mu}^{(t+s)/\mu - 1} O\big(|u^N_k|^2 q^N\big(u^N_k\big)\big). \end{aligned} \tag{4.8}$$

Owing to the N-truncation used,

$$E\left|\mu\sum_{k=t/\mu}^{(t+s)/\mu-1} f_x(x_*,\xi_k,\theta_k)u_k^N q^N\left(u_k^N\right)\right|^2 \le Ks^2,$$

and as a result

$$\lim_{\delta\to 0}\limsup_{\mu\to 0} E\left|\mu\sum_{k=t/\mu}^{(t+s)/\mu-1} f_x(x_*,\xi_k,\theta_k)u_k^N q^N\left(u_k^N\right)\right|^2 = 0. \tag{4.9}$$

Moreover, for the last term of (4.8),

$$\lim_{\delta\to 0}\limsup_{\mu\to 0} E\left|\mu^{3/2}\sum_{k=t/\mu}^{(t+s)/\mu-1} O\left(|u_k^N|^2 q^N\left(u_k^N\right)\right)\right|^2 = 0. \tag{4.10}$$

Next, by the independence of $\{\theta_n\}$ and $\{\xi_n\}$, and in view of the mixing type of inequalities (3.15)–(3.17),

$$\begin{aligned}
&E\left|\sqrt{\mu}\sum_{k=t/\mu}^{(t+s)/\mu-1} f(x_*,\xi_n,\theta_n)\right|^2\\
&= \mu\sum_{k=t/\mu}^{(t+s)/\mu-1}\sum_{j=t/\mu}^{(t+s)/\mu-1} Ef'(x_*,\xi_k,\theta_k)f(x_*,\xi_j,\theta_j)\\
&= \mu\sum_{i_1=1}^{m_0}\sum_{i_2=1}^{m_0}\sum_{k=t/\mu}^{(t+s)/\mu-1}\sum_{j=t/\mu}^{(t+s)/\mu-1} Ef'(x_*,\xi_k,i_2)f(x_*,\xi_j,i_1)P(\theta_j=i_1,\theta_k=i_2)\\
&= \mu\sum_{i_1=1}^{m_0}\sum_{i_2=1}^{m_0}\sum_{k=t/\mu}^{(t+s)/\mu-1}\sum_{j=t/\mu}^{(t+s)/\mu-1} E[f'(x_*,\xi_k,i_2)f(x_*,\xi_j,i_1)\\
&\quad -Ef'(x_*,\xi_k,i_2)Ef(x_*,\xi_j,i_1)]P(\theta_j=i_1,\theta_k=i_2)\\
&\le K\mu\left(\frac{t+s}{\mu}-\frac{t}{\mu}\right) = Ks.
\end{aligned}$$

This yields that

$$\lim_{\delta\to 0}\limsup_{\mu\to 0} E\left|\sqrt{\mu}\sum_{k=t/\mu}^{(t+s)/\mu-1} f(x_*,\xi_k,\theta_k)\right|^2 = 0. \tag{4.11}$$

Combining (4.9)–(4.11), we have

$$\lim_{\delta\to 0}\limsup_{\mu\to 0} E\left|u^{\mu,N}(t+s)-u^{\mu,N}(t)\right|^2 = 0,$$

and hence $\{u^{\mu,N}(\cdot)\}$ is tight. □

We proceed to characterize the limit process. Another condition is stated first followed by a lemma concerning a Brownian motion limit. In (A4), convergence to a Brownian motion is assumed. Sufficient conditions (such as $(f(x_*, \xi_k, 1), \ldots, f(x_*, \xi_k, m_0))'$ being a ϕ-mixing process) can be provided. Then it becomes a standard invariance theorem (such as in [2, Theorem 20.1]).

(A4) Define

$$B^{\mu}(t) = \sqrt{\mu} \sum_{k=0}^{t/\mu-1} (f(x_*, \xi_k, 1), \ldots, f(x_*, \xi_k, m_0))'. \tag{4.12}$$

The process $B^{\mu}(\cdot)$ converges weakly to an $\mathbb{R}^{m_0 d}$-valued Brownian motion $B(\cdot)$, whose covariance is $\Sigma_B t$ with $\Sigma_B = (\sigma_{B,ij}, i, j = 1, \ldots, m_0)$ and $\sigma_{B,ij} \in \mathbb{R}^{d\times d}$ given by

$$\begin{aligned}\sigma_{B,ij} &= Ef(x_*, \xi_0, i)f'(x_*, \xi_0, j) + \sum_{k=1}^{\infty} Ef(x_*, \xi_0, i)f'(x_*, \xi_k, j)\\ &\quad + \sum_{k=1}^{\infty} Ef(x_*, \xi_k, i)f'(x_*, \xi_0, j).\end{aligned} \tag{4.13}$$

Moreover, for any $m \geq 0$, as $n \to \infty$,

$$\frac{1}{n}\sum_{k=m}^{m+n-1} E_m f_x(x, \xi_k, i) \to \widehat{f_x}(x, i) \quad \text{in probability for each } i \in \mathcal{M}. \tag{4.14}$$

Lemma 5 *Assume* (A4) *holds. Define*

$$w^{\mu}(t) = \sqrt{\mu}\sum_{i=1}^{m_0}\sum_{k=0}^{t/\mu-1} \nu_i f(x_*, \xi_k, i). \tag{4.15}$$

Then $w^{\mu}(\cdot)$ converges weakly to a Brownian motion $w(\cdot)$, whose covariance is $\Sigma_w t$ with

$$\Sigma_w = \sum_{i=1}^{m_0}\sum_{j=1}^{m_0} \nu_i \nu_j \sigma_{B,ij}. \tag{4.16}$$

Proof The assertion follows from (A4) by noting $w^{\mu}(t) = \nu B^{\mu}(t)$ (in the sense of product of partitioned matrices) with $\nu = (\nu_1, \ldots, \nu_{m_0})$ and then applying the well-known Cramér–Wold device. Thus the proof is completed. □

Lemma 6 *Assume the conditions of Lemma 5 and* (A1). *Define*

$$\widetilde{w}^{\mu}(t) = \sqrt{\mu} \sum_{k=0}^{t/\mu - 1} f(x_*, \xi_k, \theta_k). \tag{4.17}$$

In addition, $\varepsilon = \mu^{\frac{1}{2}+\Delta}$ *for some* $\Delta > 0$. *Then* $\widetilde{w}^{\mu}(\cdot)$ *converges weakly to the Brownian motion* $w(\cdot)$ *given in Lemma 5.*

Proof Since

$$\begin{aligned} \widetilde{w}^{\mu}(t) &= \sqrt{\mu} \sum_{i=1}^{m_0} \sum_{k=0}^{t/\mu-1} f(x_*, \xi_k, i) I_{\{\theta_k = i\}} \\ &= \sqrt{\mu} \sum_{i=1}^{m_0} \sum_{k=0}^{t/\mu-1} f(x_*, \xi_k, i)\nu_i + \sqrt{\mu} \sum_{i=1}^{m_0} \sum_{k=0}^{t/\mu-1} f(x_*, \xi_k, i)[I_{\{\theta_k = i\}} - \nu_i], \end{aligned} \tag{4.18}$$

in view of Lemma 5, we need only show

$$\widetilde{w}^{\mu}(t) = w^{\mu}(t) + o(1),$$

where $o(1) \to 0$ in probability uniformly in $t \in [0, T]$ for any $T > 0$. Working with the last term in (4.18), it suffices to concentrate on a fixed i. In fact, for a fixed $i = 1, \ldots, m_0$, using (3.16),

$$\begin{aligned} &E\left| \sqrt{\mu} \sum_{k=0}^{t/\mu-1} f(x_*, \xi_k, i)[I_{\{\theta_k = i\}} - \nu_i] \right|^2 \\ &= \mu \sum_{k=0}^{t/\mu-1} \sum_{j=0}^{t/\mu-1} Ef'(x_*, \xi_k, i) f(x_*, \xi_j, i) E[I_{\{\theta_k = i\}} - \nu_i][I_{\{\theta_j = i\}} - \nu_i] \\ &= \mu \sum_{k=0}^{t/\mu-1} \sum_{j=0}^{t/\mu-1} Ef'(x_*, \xi_k, i) f(x_*, \xi_j, i) \\ &\quad \times [P(\theta_k = i, \theta_j = i) - \nu_i P(\theta_k = i) - \nu_i P(\theta_j = i) + \nu_i^2] \\ &= \mu \sum_{k=0}^{t/\mu-1} \sum_{j=0}^{t/\mu-1} Ef'(x_*, \xi_k, i) f(x_*, \xi_j, i) O(\varepsilon^2 k + \varepsilon^2 j + \varepsilon + \lambda^k + \lambda^j) \\ &\to 0 \ \text{ as } \ \mu \to 0. \end{aligned} \tag{4.19}$$

In deriving (4.19), we have used the fact $\varepsilon = \mu^{\frac{1}{2}+\Delta}$. □

The last term in (4.8) is asymptotically unimportant. This, combined with Lemma 1 and Lemma 6, yields that

$$u^{\mu,N}(t+s) - u^{\mu,N}(t) = \mu \sum_{i=1}^{m_0} \sum_{k=t/\mu}^{(t+s)/\mu-1} f_x(x_*, \xi_k, i)\nu_i u_k^N q^N\big(u_k^N\big) + \sqrt{\mu}\sum_{i=1}^{m_0} \sum_{k=t/\mu}^{(t+s)/\mu-1} f(x_*, \xi_k, i)\nu_i + o(1), \tag{4.20}$$

where $o(1) \to 0$ in probability as $\mu \to 0$ uniformly in t. Likewise, we also obtain that for l satisfying $t/\mu \le ln_\mu \le (t+s)/\mu - 1$,

$$u^N_{ln_\mu+n_\mu} - u^N_{ln_\mu} = \mu \sum_{i=1}^{m_0} \sum_{k=ln_\mu}^{ln_\mu+n_\mu-1} f_x(x_*, \xi_k, i)\nu_i u_k^N q^N\big(u_k^N\big) + \sqrt{\mu}\sum_{i=1}^{m_0} \sum_{k=ln_\mu}^{ln_\mu+n_\mu-1} f(x_*, \xi_k, i)\nu_i + o(1), \tag{4.21}$$

where $o(1) \to 0$ in probability as $\mu \to 0$.

Since $\{u^{\mu,N}(\cdot)\}$ is tight, we can extract a weakly convergent subsequence by Prohorov's theorem. Still index the subsequence by μ for notational simplicity and denote its limit by $u^N(\cdot)$. We claim that $u^N(\cdot)$ is the unique solution of the martingale problem with operator $\mathcal{L}^N$ defined by

$$\mathcal{L}^N g(u) = g_u'(u)\overline{f}_x(x_*)uq^N(u) + \frac{1}{2}\mathrm{tr}[g_{uu}(u)\Sigma_w], \tag{4.22}$$

for any $g(\cdot) \in C_0^2$ (the collection of C^2 functions with compact support). The proof uses the martingale averaging approach [17].

To verify the desired result, we need only show that for any bounded and continuous function $H(\cdot)$, any $t, s > 0$, any positive integer κ, and any t_j satisfying $0 \le t_j \le t$ with $j \le \kappa$,

$$EH(u^N(t_j) : j \le \kappa)\Big[g\big(u^N(t+s)\big) - g\big(u^N(t)\big) - \int_t^{t+s} \mathcal{L}^N g\big(u^N(\tau)\big)d\tau\Big] = 0. \tag{4.23}$$

To verify (4.23), it suffices to work with the sequence $u^{\mu,N}(\cdot)$ and demonstrate the limit has the desired property.

By the weak convergence and the Skorohod representation,

$$EH\big(u^{\mu,N}(t_j) : j \le \kappa\big)\big[g\big(u^{\mu,N}(t+s)\big) - g\big(u^{\mu,N}(t)\big)\big] \to EH\big(u^N(t_j) : j \le \kappa\big)\big[g\big(u^N(t+s)\big) - g\big(u^N(t)\big)\big] \text{ as } \mu \to 0. \tag{4.24}$$

Using (4.21),

$$
\begin{aligned}
& g\big(u^{\mu,N}(t+s)\big) - g\big(u^{\mu,N}(t)\big) \\
&\quad = \sum_{ln_\mu = t/\mu}^{(t+s)/\mu - 1} \big[g\big(u^N_{ln_\mu + n_\mu}\big) - g\big(u^N_{ln_\mu}\big)\big] \\
&\quad = \sum_{ln_\mu = t/\mu}^{(t+s)/\mu - 1} \Big[g_u\big(u^N_{ln_\mu}\big)\big(u^N_{ln_\mu + n_\mu} - u^N_{ln_\mu}\big) \\
&\qquad + \frac{1}{2}\mathrm{tr}\big(g_{uu}\big(u^N_{ln_\mu}\big)\big(u_{ln_\mu + n_\mu} - u_{ln_\mu}\big)\big(u_{ln_\mu + n_\mu} - u_{ln_\mu}\big)'\big)\Big] + o(1),
\end{aligned} \tag{4.25}
$$

where $o(1) \to 0$ as $\mu \to 0$. Since $u^{\mu,N}(t_j)$, for $j \le \kappa$, are $\mathcal{F}_{ln_\mu}$-measurable, it can be seen that

$$
\begin{aligned}
& \lim_{\mu\to 0} EH\big(u^{\mu,N}(t_j) : j \le \kappa\big)\Big[\sum_{ln_\mu = t/\mu}^{(t+s)/\mu - 1} g_u\big(u^N_{ln_\mu}\big)\big[u^N_{ln_\mu + n_\mu} - u^N_{ln_\mu}\big]\Big] \\
&= \lim_{\mu\to 0} EH\big(u^{\mu,N}(t_j) : j \le \kappa\big)\Big[\mu \sum_{i=1}^{m_0} \sum_{ln_\mu = t/\mu}^{(t+s)/\mu - 1} g_u\big(u^N_{ln_\mu}\big) \\
&\quad \times \sum_{k = ln_\mu}^{ln_\mu + n_\mu - 1} E_{ln_\mu} f_x(x_*, \xi_k, i)\nu_i u^N_{ln_\mu} q^N\big(u^N_{ln_\mu}\big) \\
&\qquad + \sqrt{\mu} \sum_{i=1}^{m_0} \sum_{ln_\mu = t/\mu}^{(t+s)/\mu - 1} g_u\big(u^N_{ln_\mu}\big) q^N\big(u^N\big(u_{ln_\mu}\big)\big) \sum_{k = ln_\mu}^{ln_\mu + n_\mu - 1} E_{ln_\mu} f(x_*, \xi_k, i)\nu_i\Big] \\
&= EH\big(u^N(t_j) : j \le \kappa\big)\Big[\int_t^{t+s} g_u\big(u^N(\tau)\big)\overline{f}_x(x_*)u^N(\tau)q^N\big(u^N(\tau)\big)d\tau\Big].
\end{aligned} \tag{4.26}
$$

Similar to (4.26), we also obtain that

$$
\begin{aligned}
& \lim_{\mu\to 0} EH\big(u^{\mu,N}(t_j) : j \le \kappa\big)\Big[\frac{1}{2} \sum_{ln_\mu = t/\mu}^{(t+s)/\mu - 1} \mathrm{tr}\big[g_{uu}\big(u^N_{ln_\mu}\big)\big(u^N_{ln_\mu + n_\mu} - u^N_{ln_\mu}\big) \\
&\qquad\qquad \times \big(u^N_{ln_\mu + n_\mu} - u^N_{ln_\mu}\big)'\big]\Big] \\
&\quad = EH\big(u^N(t_j) : j \le \kappa\big)\Big[\frac{1}{2}\int_t^{t+s} \mathrm{tr}[g_{uu}(u^N(\tau))\Sigma_w]d\tau\Big].
\end{aligned} \tag{4.27}
$$

Substituting (4.21) into (4.25) and using (4.26) and (4.27) yield

$$
\begin{aligned}
&EH\big(u^{\mu,N}(t_j) : j \le \kappa\big)\big[g\big(u^{\mu,N}(t+s)\big) - g\big(u^{\mu,N}(t)\big)\big] \\
&\quad \to EH\big(u^N(t_j) : j \le \kappa\big)\left[\int_t^{t+s} \mathcal{L}^N g\big(u^N(\tau)\big)d\tau\right] \quad \text{as } \mu \to 0. \qquad (4.28)
\end{aligned}
$$

Combining (4.24) and (4.28), (4.23) is verified. Thus the truncated process $u^N(\cdot)$ is the solution of the martingale problem with the desired operator.

The last step involves showing that as $N \to \infty$, the untruncated process $u^\mu(\cdot)$ converges to $u(\cdot)$ weakly. The argument is similar to that of [17, pp. 282–285]. The basic idea is: Let $P^0(\cdot)$ and $P^N(\cdot)$ be the measures induced by $u(\cdot)$ and $u^N(\cdot)$, respectively. Since

$$
du(t) = \overline{f}_x(x_*)u(t)dt + \Sigma_w^{1/2} d\widehat{w} \qquad (4.29)
$$

is linear, it has a unique solution for each initial condition and $P^0(\cdot)$ is unique. In the above, $\widehat{w}(\cdot)$ is a standard Brownian motion and Σ_w is given by (4.16). For each $T < \infty$ and $t \le T$, $P^0(\cdot)$ agrees with $P^N(\cdot)$ on all Borel subsets of the set of paths in $D^d([0,\infty) : \mathbb{R}^d)$ with values in S_N. Since $P^0(\sup_{t \le T} |u(t)| \le N) \to 1$ as $N \to \infty$, and $u^{\mu,N}(\cdot)$ converges to $u^N(\cdot)$ weakly, we have $u^\mu(\cdot)$ converges weakly to $u(\cdot)$. It turns out that the limit $u(\cdot)$ is the unique solution of the martingale problem with operator

$$
\mathcal{L}g(u) = g_u'(u)\overline{f}_x(x_*)u + \frac{1}{2}\text{tr}[g_{uu}(u)\Sigma_w]. \qquad (4.30)
$$

Equivalently, $u(\cdot)$ is the solution of the stochastic differential equation (4.29). A few details are omitted. We record the result into the following theorem.

Theorem 4 *Assume the conditions of Corollary 1 and Lemma 6 are satisfied, and $\overline{f}_x(x_*)$ is a stable matrix (that is, all of its eigenvalues are on the left-half of the complex plane). Then $u^\mu(\cdot)$ converges weakly to $u(\cdot)$ satisfying the diffusion equation (4.29).*

Remark 7 The solution of (4.29) can be written as

$$
u(t) = \exp(\overline{f}_x(x_*)t)u(0) + \int_0^t \exp(\overline{f}_x(x_*)(t-s))\Sigma^{1/2} d\widehat{w}(s).
$$

Since $\overline{f}_x(x_*)$ is a stable matrix, $u(\cdot)$ is asymptotically stationary with stationary covariance (for $t > 0$) given by

$$
\begin{aligned}
Eu(t)u'(0) &= E\left(\int_{-\infty}^{t} \exp(\overline{f}_x(x_*)(t-s))\Sigma_w^{1/2} d\widehat{w}(s)\right) \\
&\qquad \times \left(\int_{-\infty}^{0} \exp(\overline{f}_x(x_*)s)\Sigma_w^{1/2} d\widehat{w}(s)\right)' \\
&= \exp(\overline{f}_x(x_*)t)\Sigma_0,
\end{aligned}
$$

where

$$
\Sigma_0 = \int_0^\infty \exp(\overline{f}_x(x_*)s)\Sigma_w \exp(\overline{f}_x'(x_*)s)ds.
$$

The scale factor $\sqrt{\mu}$ together with the stationary covariance Σ_0 then provides us with the rate of convergence of the stochastic approximation/optimization algorithm. In fact, in view of Theorem 4, loosely, $x_n - x_*$ is asymptotically normal with mean 0 and covariance $\mu\Sigma_0$.

5 Polyak's iterate averaging

Inspired by the recent work on iterate averaging for stochastic approximation algorithms that was proposed in [19], and developed independently by [21], we consider the iterate averaging algorithms with a fast Markov component in this section. The motivation behind the averaging approach can be traced back to the work of [6] and many subsequent papers on adaptive stochastic approximation. We illustrate the idea in what follows. Consider a decreasing step size stochastic approximation algorithm (without a Markov chain component)

$$
y_{n+1} = y_n + \mu_n F(y_n, \zeta_n),
$$

where $F(\cdot)$ is a suitable function and $\mu_n = O(1/n^\gamma)$ with $0 < \gamma \le 1$. Assuming $\{\zeta_n\}$ is a stationary sequence and y_* is the unique asymptotically stable point of the ODE $\dot{y} = \overline{F}(y)$, under suitable conditions, it can be shown $y_n \to y_*$ in an appropriate sense and $(y_n - y_*)/\sqrt{\mu_n}$ is asymptotically normal. Clearly, $\mu_n = O(1/n)$ yields the best scaling factor among different choice of γ with $0 < \gamma \le 1$. The idea in [6] lies in choosing $\mu_n = \Gamma/n$ and minimizing the resulting asymptotic covariance with respect to the parameter Γ, although that paper dealt with a scalar problem only and no diffusion approximation was involved. It leads to the discovery that the best choice of Γ is $-(\overline{F}_y(y_*))^{-1}$ yielding the minimal variance. It has also been noted this minimal variance in fact, is related to the Cramér–Rao bound and Fisher information [18] leading to an asymptotic efficient procedure. However, the point y_* and most often the function $\overline{F}(\cdot)$ and $\overline{F}_y(\cdot)$ are unknown. To be of any use, one has to estimate the quantity $\overline{F}_y(y_*)$ (a matrix-valued function for multidimensional problems). The adaptive stochastic approximation approach then aims to construct both y_n and the estimate of the gradient $\overline{F}_y(y_*)$ recursively. For multidimensional problems, this is proved to be impractical due to the computational complexity.

As an alternative, the iterate averaging is based on a different idea. First it uses a large step size (larger than $O(1/n)$) to get a sequence of rough estimates. Then a simple arithmetic averaging on the rough estimates is carried out. It has been shown that the iterate-averaging approach leads to asymptotic optimality (the best scaling factor and the minimal variance) and has advantages for various applications. First, its initial approximation uses slowly varying step sizes larger than $O(1/n)$ and enables the iterates to get to a neighborhood of y_* faster than that of a small step-size procedure. Then by averaging the iterates, the resulting sample path possesses the minimal variance.

Our effort in what follows is to prove that the iterate averaging yields asymptotic optimality for the algorithms considered in this paper. Note that the averaging is taken with a window of width $O(1/\mu)$, the so-called minimal window width of averaging (see [17, Chapt. 11]). To analyze the algorithm, we let t_1 and t_2 be nonnegative real numbers that satisfy $t_1 + t_2 = t$, and choose t_μ such that $t_\mu \to \infty$ as $\mu \to 0$. Define

$$X_n = \frac{\mu}{t} \sum_{j=t_\mu - t_2/\mu}^{t_\mu + t_1/\mu - 1} x_j, \quad \widehat{X}_n = \frac{\mu}{t} \sum_{j=t_\mu - t_2/\mu}^{t_\mu + t_1/\mu - 1} (x_j - x_*), \tag{5.1}$$

and

$$\overline{u}^\mu(t) = \frac{\sqrt{\mu}}{t} \sum_{j=t_\mu - t_2/\mu}^{t_\mu + \frac{t_1}{\mu} - 1} (x_j - x_*). \tag{5.2}$$

Using a similar argument as [17, p. 379], we obtain: For each fixed $t > 0$, $\overline{u}^\mu(t)$ converges in distribution to a normal random vector $\overline{u}(t)$ with mean 0 and covariance

$$\frac{1}{t} \overline{f}_x^{-1}(x_*) \Sigma_w (\overline{f}_x^{-1}(x_*))' + O\left(\frac{1}{t^2}\right).$$

Note that $\overline{f}_x^{-1}(x_*) \Sigma_w (\overline{f}_x^{-1}(x_*))'/t$ is the optimal asymptotic covariance in the sense mentioned in the previous paragraphs. Thus iterate averaging over a window of $O(1/\mu)$ when $\varepsilon = O(\mu)$ results in a tracking algorithm with an asymptotically optimal covariance.

Remark 8 In Monte Carlo optimization, one is most concerned with the bias and noise. As observed in [17, p. 374], the iterating averaging does not have effect on the bias for a Kiefer–Wolfowitz type stochastic optimization algorithm, but it does reduce the noise around the bias; see also the related work [5,7], and [24, Sects. 4.5.3 and 6.8], among others.

6 Example and discussion

6.1 A two-time-scale LMS algorithm

This section is devoted to an example of a class of LMS (least mean square) algorithms involving a Markov chain parameter. The motivation for such a study arises from a number of examples in fault diagnosis and change detection (see [1]), target tracking, and econometrics (see [10]) and the references therein. Such problems also arise in emerging applications of adaptive interference suppression in wireless CDMA networks with fast fading and state estimation of hidden Markov models (see [13]).

Let $\{y_n\}$ be a sequence of real-valued signals representing the observations obtained at time n, and $\{\theta_n\}$ be the time-varying true parameter, an $\mathbb{R}^d$-valued random process. Suppose that

$$y_n = \varphi_n' \theta_n + e_n, \tag{6.1}$$

where $\varphi_n \in \mathbb{R}^d$ is the regression vector, θ_n (a vector-valued parameter process) is a Markov chain with state space $\mathcal{M} = \{a_1, \ldots, a_{m_0}\}$ with $a_i \in \mathbb{R}^d$, and $e_n \in \mathbb{R}$ represents the zero mean observation noise. We consider a LMS type algorithm with a constant step size that recursively operates on the observation sequence $\{y_n\}$ to generate a sequence of parameter estimates $\{\widehat{\theta}_n\}$ (or $\{x_n\}$ in the notation of (2.7)) according to

$$\widehat{\theta}_{n+1} = \widehat{\theta}_n + \mu \varphi_n (y_n - \varphi_n' \widehat{\theta}_n), \tag{6.2}$$

where $\mu > 0$ denotes the small constant step size. In a companion paper [30], we have considered a similar problems with a slowly varying Markov chain (slow in the sense of the transitions occurring infrequently, i.e., $\varepsilon \ll \mu$) using an algorithm taken the same form. Nevertheless, fundamentally different, now the parameter process changes much faster than the estimates $\widehat{\theta}_n$ generated (i.e., $\mu \ll \varepsilon$). This difference creates much of the difficulties. One essentially has no hope to track the detailed parameter variations due to the fast-changing dynamics in the parameter process. As is well known that for parameter estimation or tracking, one can only track the time-varying parameter if certain tractability conditions or identifiability conditions are met. Thus, the fast-varying parameter process poses a barrier or limitation that cannot be overcome. Using the results from previous sections, we suggest a viable alternative. Instead of tracking the actual system, we may ignore the detailed variations and consider only the limit system. Note that in the previous sections, the underlying Markov chain was assumed to be a scalar-valued one. In this section, a vector-valued process is under consideration. However, the general approach outlined before can be readily adapted to the current situation. For example, we may treat θ_n as $\Theta(\widetilde{\theta}_n)$ such that $\widetilde{\theta}_n$ is a Markov chain taking values in $\{1, \ldots, m_0\}$ and $\Theta(i) \in \mathbb{R}^d$ for $i = 1, \ldots, m_0$. Using the analysis in the convergence and rate of convergence sections, we obtain the following results. The assertions are provided in what follows although the detailed derivations are omitted.

Assume that the process $\{\varphi_n, e_n\}$ is stationary satisfying $E\varphi_n e_n = 0$ and $E\varphi_n \varphi_n' = A$, and that it is bounded mixing processing with mixing measure ψ_n satisfying $\sum_n \psi_n^{1/2} < \infty$. In addition, assume $\varepsilon = \mu^{\frac{1}{2}+\Delta}$ for some $\Delta > 0$.

Since $\{\varphi_n, e_n\}$ is stationary uniform mixing, it is ergodic. Then for any $m \geq 0$,

$$\frac{1}{n}\sum_{k=m}^{n+m-1} \varphi_k e_k \to 0 \quad \text{and} \quad \frac{1}{n}\sum_{k=m}^{n+m-1} \varphi_k \varphi_k' \to A \quad \text{w.p.1 as } n \to \infty.$$

Furthermore, we obtain:

- Define the continuous-time interpolation $\widehat{\theta}^\mu(t) = \widehat{\theta}_n$ for $t \in [\mu n, \mu n + \mu)$. Then $\widehat{\theta}^\mu(\cdot)$ converges weakly to $\widehat{\theta}(\cdot)$, which is the unique solution of the differential equation

$$\frac{d}{dt}\widehat{\theta}(t) = -A(\widehat{\theta}(t) - \theta_*), \tag{6.3}$$

  where

$$\theta_* = \sum_{i=1}^{m_0} \nu_i a_i.$$

- Suppose $t_\mu \to \infty$ as $\mu \to 0$. Then $\widehat{\theta}^\mu(\cdot + t_\mu)$ converges weakly to θ_*.
- Define $u_n = (\widehat{\theta}_n - \theta_*)/\sqrt{\mu}$. Then $\{u_n : n \geq N_\mu\}$ is tight for N_μ given in (4.1).
- Define

$$w^\mu(t) = \sqrt{\mu} \sum_{k=0}^{t/\mu - 1} \varphi_k e_k.$$

  Then $w^\mu(\cdot)$ converges weakly to $\Sigma^{1/2} w(\cdot)$, where $w(\cdot)$ is a standard Brownian motion, and $\Sigma = E\varphi_0\varphi_0' e_0^2 + \sum_{k=1}^\infty E\varphi_k\varphi_0' e_k e_0 + \sum_{k=1}^\infty E\varphi_0\varphi_k' e_k e_0$. Define $u^\mu(\cdot)$ to be the piecewise constant interpolation of u_n. Then $u^\mu(\cdot)$ converges weakly to $u(\cdot)$, which is the unique solution to the stochastic differential equation

$$du = -Au dt + \Sigma^{1/2} dw. \tag{6.4}$$

- In lieu of the mixing condition on $\{\varphi_n, e_n\}$, assume that for each $m \geq 0$,

$$\frac{1}{n}\sum_{k=m}^{m+n-1} E_m \varphi_k e_k \to 0 \quad \text{in probability as } n \to \infty,$$

  and that there are symmetric positive definite matrices A_1 and A_2 satisfying $A_1 < A_2$ (in the sense of positive definiteness) such that

$$\rho\Big[\frac{1}{n}\sum_{k=m}^{m+n-1} E_m \varphi_k \varphi_k', \mathbb{A}\Big] \to 0 \quad \text{in probability as } n \to \infty,$$

where $\mathbb{A}$ is the set of convex combinations of A_1 and A_2 and $\rho(\cdot)$ is the distance function. Then it can be shown that $\theta^\mu(\cdot)$ converges weakly to $\theta(\cdot)$ such that the limit $\theta(\cdot)$ satisfies the differential inclusion

$$\dot{\theta}(t) \in -\mathbb{A}(\theta(t) - \theta_*).$$

6.2 Discussion and extensions

In this paper, we have developed a class of stochastic approximation and optimization algorithms. One of the main features is the inclusion of a Markov component. Another novel feature is that the Markov chain changes an order of magnitude faster than that of the iterates. Convergence, rates of convergence, and iterate averaging have been examined. A number of issues are worthy of further investigation; several directions may be pursued for subsequent study under the framework on inclusion of a Markov chain that varies faster than the step size used in the iterations.

First, from an applications point of view, the results of this paper yield useful insight into the nature of the use of adaptive filtering algorithms in wireless communication systems. For applications where the communication channel varies much faster than the dynamics of the channel, the results of Sect. 6.1 yield a fundamental limit into how well LMS based channel estimation algorithms can perform. In particular, the above results show that LMS based channel estimation algorithm for estimating the rapidly varying channel amplitude can at best only estimate the invariant distribution of the channel amplitude and not the actual amplitude sample path. For such cases, as suggested in [25], it is appropriate to use alternative measurements such as the phase of the channel which typically varies on a slower time scale. Tracking of fast varying channels is an important ingredient of "Cognitive radio"—we refer the reader to [9].

For the purpose of optimization, the algorithm considered in this paper assumes the noisy gradient estimates being available. When such gradient estimates or pathwise derivatives are not available, we have recourse to the gradient estimates based on noisy cost function value measurements. The results of this paper can readily be extended to such a case where only noisy functional measurements are available. Then, the algorithms involving gradient estimates will be of the Kiefer–Wolfowitz form.

With the gradient-estimate-based stochastic approximation algorithms, considering high dimensional stochastic optimization problems using random directions, nowadays commonly referred to as SPSA (simultaneous perturbation stochastic approximation) [23], is of particular interest. Using a constant step size for the iterates together with another constant step size for the gradient estimates, these step sizes should be incorporated with the perturbation parameter ε in the Markov chain. Constrained algorithms (e.g., projection and truncation algorithms as those studied extensively in [17]) will be beneficial in a wide range of applications. There is also a version using randomly generated truncation bounds to relax the growth condition on the function under consideration [4]. Global stochastic approximation and optimization, and to some extent, global optimization as a whole, remain difficult, especially in improving rates of convergence [28].

Scalar Markov chains are treated in the paper. The results can be extended to vector-valued Markov chains using the same techniques as in this paper. One example is given in Sect. 6. In financial engineering, such an approach is known as calibration. Under the premise that a countable state Markov chain can often be approximated by a finite state Markov chain (in fact under current computation capability, only finite state space cases can be treated), we have focused on Markov chains with a finite state space. Consideration of problems involving countable state space is a worthwhile effort. Stopping rules associated for the algorithms proposed also deserves further study and investigation. A possible approach is to examine the associated stopped random processes; see [26].

In stochastic optimization algorithms, when we deal with nondifferentiable (or nonsmooth) functions, in lieu of limit ordinary differential equations, we have differential inclusions. The use of convex optimization and subgradients naturally arise. It is important to study asymptotic properties for the class of regime-switching stochastic approximation and optimization algorithms with a fast-varying jump component in this context.

The rate of convergence was studied here using diffusion approximation techniques. A worthwhile extension is to conduct a large deviations analysis. It will be advantageous for constrained optimization problems.

Owing to many emerging applications in wireless communications and financial engineering, there are growing interests in studying regime-switching stochastic approximation and optimization algorithms. The analysis in this paper combined with different order of magnitude of step sizes (different combinations of ε and μ as in [30,32]) will open up new domains for a wider range of applications.

Acknowledgments We thank the reviewers and the editors for the detailed comments and suggestions on an early version of the manuscript.

References

1. Benveniste, A., Metivier, M., Priouret, P.: Adaptive algorithms and stochastic approximations. Springer, New York (1990)
2. Billingsley, P.: Convergence of probability measures. Wiley, New York (1968)
3. Buche, R., Kushner, H.J.: Stochastic approximation and user adaptation in a competitive resource sharing system. IEEE Trans. Autom. Control **45**, 844–853 (2000)
4. Chen, H.-F.: Stochastic approximation and its applications. Kluwer Academic, Dordrecht (2002)
5. Chin, D.C., Maryak, J.T.: A cautionary note in iterate averaging in stochastic approximation. In: INFORMS annual meeting pp. 5–8 (1996)
6. Chung, K.L.: On a stochastic approximation method. Ann. Math. Stat. **25**, 463–483 (1954)
7. Dippon, J., Renz, J.: Weighted means in stochastic approximation of minima. SIAM J. Control Optim. **35**, 1811–1827 (1997)
8. Ephraim, Y., Merhav, N.: Hidden Markov processes. IEEE Trans. Inform. Theory **48**, 1518–1569 (2002)
9. Haykin, S.: Cognitive radio: brain empowered wireless communications. IEEE J. Sel. Areas Commun. **23**, 201–220 (2005)
10. Krishnamurthy, V., Rydén, T.: Consistent estimation of linear and non-linear autoregressive models with Markov regime. J. Time Ser. Anal. **19**, 291–308 (1998)
11. Krishnamurthy, V., Yin, G.: Recursive algorithms for estimation of hidden Markov models and autoregressive models with Markov regime. IEEE Trans. Inform. Theory **48**(2), 458–476 (2002)

12. Krishnamurthy, V., Yin, G.: Controlled hidden Markov models for dynamically adapting patch clamp experiment to estimate Nernst potential of single-ion channels. IEEE Trans. Nanobiosci. **5**, 115–125 (2006)
13. Krishnamurthy, V., Wang, X., Yin, G.: Spreading code optimization and adaptation in CDMA via discrete stochastic approximation. IEEE Trans. Inform. Theory **50**, 1927–1949 (2004)
14. Kushner, H.J.: Approximation and weak convergence methods for random processes, with applications to stochastic systems theory. MIT Press, Cambridge (1984)
15. Kushner, H.J., Huang, H.: Asymptotic properties of stochastic approximations with constant coefficients. SIAM J. Control Optim. **19**, 87–105 (1981)
16. Kushner, H.J., Yang, J.: Stochastic approximation with averaging of iterates: optimal asymptotic rate of convergence for general processes. SIAM J. Control Optim. **31**, 1045–1062 (1993)
17. Kushner, H.J., Yin, G.: Stochastic approximation and recursive algorithms and applications. 2nd edn. Springer, New York (2003)
18. Nevel'son, M.B., Khasminskii, R.Z.: Stochastic approximation and recursive estimation. Translation of Math. Monographs, vol. 47. AMS, Providence (1976)
19. Polyak, B.T.: New method of stochastic approximation type. Autom. Remote Control **51**, 937–946 (1990)
20. Polyak, B.T., Juditsky, A.B.: Acceleration of stochastic approximation by averaging. SIAM J. Control Optim. **30**, 838–855 (1992)
21. Ruppert, D.: Stochastic approximation. In: Ghosh, B.K., Sen, P.K. (eds.) Handbook in sequential analysis, pp. 503–529. Marcel Dekker, New York (1991)
22. Schwabe, R.: Stability results for smoothed stochastic approximation procedures. Z. Angew. Math. Mech. **73**, 639–644 (1993)
23. Spall, J.C.: Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. IEEE Trans. Autom. Control **AC-37**, 332–341 (1992)
24. Spall, J.C.: Introduction to stochastic search and optimization: estimation, simulation, and control. Wiley, New York (2003)
25. Tarighat, A., Sayed, A.H.: Least mean-phase adaptive filters with application to communications systems. IEEE Signal Process. Lett. **11**, 220–223 (2004)
26. Yin, G.: A stopping rule for the Robbins–Monro method. J. Optim. Theory Appl. **67**, 151–173 (1990)
27. Yin, G.: On extensions of Polyak's averaging approach to stochastic approximation. Stoch. Stoch. Rep. **36**, 245–264 (1991)
28. Yin, G.: Rates of convergence for a class of global stochastic optimization algorithms. SIAM J. Optim. **10**, 99–120 (1999)
29. Yin, G., Krishnamurthy, V.: Analysis of LMS algorithm for slowly varying Markovian parameters—tracking slow hidden Markov models and adaptive multiuser detection in DS/CDMA. IEEE Trans. Inform. Theory **51**, 2475–2490 (2005)
30. Yin, G., Krishnamurthy, V.: Least mean square algorithms with Markov regime switching limit. IEEE Trans. Autom. Control **50**, 577–593 (2005)
31. Yin, G., Zhang, Q.: Discrete-time Markov chains: two-time-scale methods and applications. Springer, New York (2005)
32. Yin, G., Krishnamurthy, V., Ion, C.: Regime switching stochastic approximation algorithms with application to adaptive discrete stochastic optimization. SIAM J. Optim. **14**, 1187–1215 (2004)