Available online at www.sciencedirect.com

SCIENCE DIRECT®

Systems & Control Letters 50 (2003) 95–102

SYSTEMS & CONTROL LETTERS

www.elsevier.com/locate/sysconle

Linear stochastic approximation driven by slowly varying Markov chains

Vijay R. Konda*, John N. Tsitsiklis

Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, USA

Received 5 June 2002; received in revised form 30 December 2002; accepted 23 February 2003

Abstract

We study a linear stochastic approximation algorithm that arises in the context of reinforcement learning. The algorithm employs a decreasing step-size, and is driven by Markov noise with time-varying statistics. We show that under suitable conditions, the algorithm can track the changes in the statistics of the Markov noise, as long as these changes are slower than the rate at which the step-size of the algorithm goes to zero.

Keywords: Stochastic approximation; Adaptive algorithms; Reinforcement learning

0. Introduction

The convergence of stochastic approximation algorithms driven by ergodic noise sequences has been extensively studied in the literature [1,5,9,10]. Typical convergence results show that the iterates converge to a stationary point of the vector field obtained by averaging the update direction with respect to the statistics of the driving noise.

In some applications, the statistics of the driving noise change with time. In such cases, the point to which the algorithm would converge if the noise were held stationary, also changes with time. In this context, a meaningful objective is to have the stochastic approximation algorithm "track" this changing point closely, after an initial transient period. Algorithms of this type are often called "adaptive", as they adapt themselves to the changing environment (for a textbook account of adaptive algorithms, see [1]).

The tracking ability of adaptive algorithms has been analyzed in several contexts [6,14]. For example, it is known that the usual constant step-size stochastic approximation algorithms can "adapt" to changes in the statistics of the driving noise that are "slow" relative to the step-size of the algorithm. However, the tracking ability of decreasing step-size stochastic approximation algorithms has not been studied before, because of the nature of the assumptions that would be required: for the changes in the statistics of the driving noise to remain slow relative to the decreasing step-size, these changes should become progressively slower. Such an assumption would be too restrictive and unnatural in applications where the noise is exogenously generated.

* Corresponding author.
E-mail addresses: konda@alum.mit.edu (V.R. Konda), jnt@mit.edu (J.N. Tsitsiklis).

0167-6911/03/$ - see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/S0167-6911(03)00132-4

In some contexts, however, the statistics of the driving noise depend only on a parameter that is deliberately changed by the user at a rate which is slower than the natural time scale of the stochastic approximation algorithm. In such cases, it becomes meaningful to study the tracking ability of decreasing step-size stochastic approximation. In this paper, we focus on linear iterations and establish that when the update direction and the statistics of the noise in the update direction depend on a "slowly" changing parameter, the algorithm can track such changes in a strong sense to be explained later. To the best of our knowledge, this is the first result on the tracking ability of "adaptive" algorithms with decreasing step-sizes. Similar results are possible for methods involving nonlinear iterations under certain stability conditions (see, e.g. [4] for stability conditions in a single time scale setting with a simpler noise model), but this direction is not pursued here.

The rest of the paper is organized as follows. In Section 1, we motivate the linear algorithms considered within the context of reinforcement learning. In Section 2, we state the main result of this paper. The last three sections are devoted to the proof of this result.

1. Motivation

The motivation for the analysis of the linear iterations considered in this paper comes from simulation-based ("reinforcement learning") methods for Markov decision problems [2,13], such as Temporal Difference (TD) learning [12], that are used to approximate the value function under a given *fixed* policy.

More precisely, consider an ergodic finite-state Markov chain $\{X_k\}$, with transition probabilities $p(y\,|\,x)$. Let $g(x)$ be a one-stage reward function, and let α be its steady-state expectation. TD learning methods consider value functions that are linear in a prescribed set of basis functions $\phi^i(x)$,

$$\hat{V}(x;r) = \sum_i r^i \phi^i(x)$$

and adjust the vector r of free parameters r^i, to obtain a sequence of approximations to the solution $V(\cdot)$ of the Poisson equation

$$V(x) = g(x) - \alpha + \sum_y p(y\,|\,x)V(y).$$

The parameters α_k and r_k, which denote the estimate of average reward and the parameters of the approximate value function, are updated recursively as

$$\alpha_{k+1} = \alpha_k + \gamma_k(g(X_k) - \alpha_k),$$

$$r_{k+1} = r_k + \gamma_k(g(X_k) - \alpha_k + r_k'\phi(X_{k+1}) - r_k'\phi(X_k))Z_k,$$

where

$$Z_k = \sum_{l=0}^{k} \lambda^{k-l}\phi(X_l)$$

is a so-called eligibility trace, and λ is a constant with $0 \leqslant \lambda < 1$.

Observe that the update equation for the vector (α_k, r_k) is of the form

$$r_{k+1} = r_k + \gamma_k(h(Y_{k+1}) - G(Y_{k+1})r_k),$$

where $Y_{k+1} = (X_k, X_{k+1}, Z_k)$ is a Markov chain whose transition probabilities are not affected by the parameters being updated. However, in the context of optimization over a parametric family of Markov chains, both the basis functions and the transition probabilities of the Markov chain $\{X_k\}$ depend on a policy parameter θ that is constantly changing. Therefore, in this context, the update takes the form

$$r_{k+1} = r_k + \gamma_k(h_{\theta_k}(Y_{k+1}) - G_{\theta_k}(Y_{k+1})r_k),$$

where θ_k denotes the value of θ at time k, and where Y_{k+1} is generated from Y_k using the transition probabilities corresponding to θ_k. If θ_k is held constant at θ, the algorithm would generally converge to some $\bar{r}(\theta)$. We would like to prove that even if θ_k moves "slowly", then r_k tracks $\bar{r}(\theta_k)$, in the sense that $|r_k - \bar{r}(\theta_k)|$ goes to zero. A result of this type, as developed in the next section, is necessary for analyzing the convergence properties of "actor-critic" algorithms, which combine temporal difference learning and slowly changing policies [7,8].

2. Main result

Consider a stochastic process $\{Y_k\}$ taking values in a Polish (complete, separable, metric) space $\mathbb{Y}$, endowed with its Borel σ-field. Let $\{P_\theta(y, \mathrm{d}\bar{y});\ \theta \in \mathbb{R}^n\}$ be a parameterized family of transition kernels on $\mathbb{Y}$. Consider the following update equations for a vector $r \in \mathbb{R}^m$ and a parameter $\theta \in \mathbb{R}^n$:

$$r_{k+1} = r_k + \gamma_k(h_{\theta_k}(Y_{k+1}) - G_{\theta_k}(Y_{k+1})r_k) + \gamma_k \xi_{k+1} r_k,$$

$$\theta_{k+1} = \theta_k + \beta_k H_{k+1}. \qquad (1)$$

In the above iteration, $\{h_\theta(\cdot), G_\theta(\cdot)\colon\ \theta \in \mathbb{R}^n\}$ is a parameterized family of m-vector valued and $m \times m$-matrix valued measurable functions on $\mathbb{Y}$. Also, H_k is a random process that drives the changes in the parameter θ_k, which in turn affects the linear stochastic approximation updates of r_k. We now continue with our assumptions.

Assumption 1. The step-size sequence $\{\gamma_k\}$ is deterministic, non-increasing, and satisfies

$$\sum_k \gamma_k = \infty, \qquad \sum_k \gamma_k^2 < \infty.$$

Let $\mathscr{F}_k$ be the σ-field generated by $\{Y_l, H_l, r_l, \theta_l\colon\ l \leqslant k\}$.

Assumption 2. For every measurable set $A \subset \mathbb{Y}$,

$$\mathbf{P}(Y_{k+1} \in A | \mathscr{F}_k) = \mathbf{P}(Y_{k+1} \in A | Y_k, \theta_k) = P_{\theta_k}(Y_k, A).$$

For any measurable function f on $\mathbb{Y}$, let $P_\theta f$ denote the measurable function $y \mapsto \int P_\theta(y, \mathrm{d}\bar{y}) f(\bar{y})$. Also, for any vector r, let $|r|$ denote its Euclidean norm.

Assumption 3 (Existence and properties of solutions to the Poisson equation). For each θ, there exist functions $\bar{h}(\theta) \in \mathbb{R}^m$, $\bar{G}(\theta) \in \mathbb{R}^{m\times m}$, $\hat{h}_\theta : \mathbb{Y} \to \mathbb{R}^m$, and $\hat{G}_\theta : \mathbb{Y} \to \mathbb{R}^{m\times m}$ that satisfy the following:

1. For each $y \in \mathbb{Y}$,
   $$\hat{h}_\theta(y) = h_\theta(y) - \bar{h}(\theta) + (P_\theta \hat{h}_\theta)(y),$$
   $$\hat{G}_\theta(y) = G_\theta(y) - \bar{G}(\theta) + (P_\theta \hat{G}_\theta)(y).$$
2. For some constant C and for all θ, we have
   $$\max(|\bar{h}(\theta)|, |\bar{G}(\theta)|) \leqslant C.$$
3. For any $d > 0$, there exists $C_d > 0$ such that
   $$\sup_k \mathbf{E}[|f_{\theta_k}(Y_k)|^d] \leqslant C_d,$$
   where $f_\theta(\cdot)$ represents any of the functions $\hat{h}_\theta(\cdot), h_\theta(\cdot), \hat{G}_\theta(\cdot), G_\theta(\cdot)$.
4. For some constant $C > 0$, and for all $\theta, \bar{\theta} \in \mathbb{R}^n$,
   $$\max(|\bar{h}(\theta) - \bar{h}(\bar{\theta})|, |\bar{G}(\theta) - \bar{G}(\bar{\theta})|) \leqslant C|\theta - \bar{\theta}|.$$
5. There exists a positive measurable function $C(\cdot)$ on $\mathbb{Y}$ such that for each $d > 0$,
   $$\sup_k \mathbf{E}[C(Y_k)^d] < \infty,$$
   and
   $$|(P_\theta f_\theta)(y) - (P_{\bar{\theta}} f_{\bar{\theta}})(y)| \leqslant C(y)|\theta - \bar{\theta}|,$$
   where $f_\theta(\cdot)$ is any of the functions $\hat{h}_\theta(\cdot)$ and $\hat{G}_\theta(\cdot)$.

Assumption 4 (Slowly changing environment). The random process $\{H_k\}$ satisfies

$$\sup_k \mathbf{E}[|H_k|^d] < \infty$$

for all $d > 0$. Furthermore, the sequence $\{\beta_k\}$ is deterministic and satisfies

$$\sum_k \left(\frac{\beta_k}{\gamma_k}\right)^d < \infty,$$

for some $d > 0$.

Assumption 5. The sequence $\{\xi_k\}$ is a $m \times m$-matrix valued $\mathscr{F}_k$-martingale difference, with bounded moments i.e.,

$$\mathbf{E}[\xi_{k+1} | \mathscr{F}_k] = 0, \qquad \sup_k \mathbf{E}[|\xi_k|^d] < \infty \quad \forall d > 0.$$

Assumption 6 (Uniform positive definiteness). There exists some $a > 0$ such that for all $r \in \mathbb{R}^m$ and $\theta \in \mathbb{R}^n$:

$$r' \bar{G}(\theta) r \geqslant a|r|^2.$$

Our main result is the following.

Theorem 7. *If Assumptions* 1–6 *are satisfied, then*

$$\lim_k |r_k - \bar{G}(\theta_k)^{-1} \bar{h}(\theta_k)| = 0.$$

When θ_k is held fixed at some value θ^*, for all k, our result states that r_k converges to $\bar{G}(\theta^*)^{-1}\bar{h}(\theta^*)$, which is a special case of Theorem 17 in p. 239 of [1]. In fact Assumptions 1 and 2 are the counterparts of Assumptions A.1 and A.2 of [1], and Assumption 3 is the counterpart of Assumptions A.3 and A.4 in [1]. Several sufficient conditions for this assumption are presented in [1] when the state space $\mathbb{Y}$ of the process Y_k is a subset of a Euclidean space. When $\mathbb{Y}$ is Polish, these conditions can be generalized using the techniques of [11].

There are two differences between our assumptions and those in [1]. A minor difference is that [1] considers vector-valued processes Y_k, whereas we consider more general processes Y_k. Accordingly, the bounds in [1] are stated in terms of $|Y_k|$, whereas our bounds are stated in terms of some positive functions of Y_k. The second difference, which is the more significant one, is that θ_k is changing, albeit slowly. For this reason, we need to use a different proof technique. Our proof combines various techniques used in [1,3,4]. In the next section, we present an overview of the proof and the intuition behind it.

3. Overview of the proof

We note that the sequence $\hat{\rho}_k = \bar{G}(\theta_k)r_k - \bar{h}(\theta_k)$ satisfies the iteration:

$$\hat{\rho}_{k+1} = \hat{\rho}_k - \gamma_k \bar{G}(\theta_{k+1})\hat{\rho}_k + \gamma_k \varepsilon^{(1)}_{k+1} + \gamma_k \varepsilon^{(2)}_{k+1},$$

where

$$\begin{aligned}\varepsilon^{(1)}_{k+1} &= \bar{G}(\theta_{k+1})(h_{\theta_k}(Y_{k+1}) - \bar{h}(\theta_k))\\ &\quad - \bar{G}(\theta_{k+1})(G_{\theta_k}(Y_{k+1}) - \bar{G}(\theta_k))r_k\\ &\quad + \bar{G}(\theta_{k+1})\xi_{k+1}r_k,\\ \varepsilon^{(2)}_{k+1} &= \frac{1}{\gamma_k}((\bar{G}(\theta_{k+1}) - \bar{G}(\theta_k))r_k\\ &\quad - (\bar{h}(\theta_{k+1}) - \bar{h}(\theta_k)))\end{aligned}$$

as can be verified by some simple algebra.

Assumption 3 implies that the vector $h_\theta(Y_{k+1})$ and the matrix $G_\theta(Y_{k+1})$ have expected values $\bar{h}(\theta)$ and $\bar{G}(\theta)$ respectively under the steady-state distribution of the time-homogeneous Markov chain Y_k with transition kernel P_θ. Therefore, we expect that the effect of the error term $\varepsilon^{(1)}_{k+1}$ can be “averaged out” in the long term. Similarly, since θ_k is changing very slowly with respect to the step-size γ_k, we expect that $\varepsilon^{(2)}_{k+1}$ goes to zero. The proof consists of showing that the terms $\varepsilon^{(i)}_{k+1}$, $i = 1, 2$, are inconsequential in the limit, and by observing that the sequence $\{\hat{\rho}_k\}$ converges to zero when these terms are absent.

We formalize this intuition in the next two sections. Note that the error terms are affine in r_k, and therefore can be very large if r_k is large. A key step in the proof is to show boundedness of the iterates r_k, and this is the subject of the next section. The main result is then proved in the last section. The approach and techniques used here are inspired by more general techniques developed in [1,4].

4. Proof of boundedness

Note that the difference between two successive iterates at time k is of the order γ_k, which goes to zero as k goes to infinity. Therefore, to study the asymptotic behavior of the sequence $\{r_k\}$, we need to focus on a subsequence $\{r_{k_j}\}$, where the sequence of non-negative integers $\{k_j\}$ is defined by

$$k_0 = 0, \qquad k_{j+1} = \min\left\{ k > k_j \;\middle|\; \sum_{l=k_j}^{k-1} \gamma_k > T \right\}.$$

Here, T is a positive constant that will be held fixed throughout the rest of the paper. The sequence $\{k_j\}$ is chosen so that any two successive elements are sufficiently apart, resulting in a non-trivial difference between r_{k_j+1} and r_{k_j}. To obtain a relation between r_{k_j+1} and r_{k_j}, we define a sequence $\{\hat{r}^j_k\}$ by

$$\hat{r}^j_k = r_k / \max(1, |r_{k_j}|) \qquad \text{for } k \geqslant k_j.$$

Note that $|\hat{r}^j_k| \leqslant 1$ for all j. Furthermore, $\hat{r}^j_k$ is $\mathscr{F}_k$-adapted and satisfies

$$\hat{r}^j_{k+1} = \hat{r}^j_k + \gamma_k \left(\frac{\bar{h}(\theta_k)}{\max(1, |r_{k_j}|)} - \bar{G}(\theta_k)\hat{r}^j_k \right) + \gamma_k \hat{\varepsilon}^j_{k+1},$$

$$k \geqslant k_j,$$

where for $k \geqslant k_j$, the term

$$\hat{\varepsilon}^j_{k+1} = \left(\frac{h_{\theta_k}(Y_{k+1}) - \bar{h}(\theta_k)}{\max(1, |r_{k_j}|)} - (G_{\theta_k}(Y_{k+1}) - \bar{G}(\theta_k))\hat{r}^j_k \right) + \xi_{k+1}\hat{r}^j_k,$$

can be viewed as perturbation noise. Similarly, for each j, define a sequence $\{r_k^j\}$ as follows:

$$r_{k_j}^j = \hat{r}_{k_j}^j,$$

$$r_{k+1}^j = r_k^j + \gamma_k \left(\frac{\bar{h}(\theta_k)}{\max(1, |r_{k_j}|)} - \bar{G}(\theta_k) r_k^j \right), \quad k \geqslant k_j.$$

Note that the iteration satisfied by r_k^j is the same as that of $\hat{r}_k^j$, except that the perturbation noise is not present in it. We will show that the perturbation noise is negligible, and that $\hat{r}_k^j$ tracks r_k^j, in the sense that

$$\lim_j \max_{k_j \leqslant k \leqslant k_{j+1}} |\hat{r}_k^j - r_k^j| = 0, \quad \text{w.p.1.}$$

The next lemma provides bounds on the perturbation noise. It involves the stopping times $\tau_j^{(1)}(C)$, which are defined as follows: given some constant $C > 1$, let

$$\tau_j^{(1)}(C) = \min\{k \geqslant k_j \colon |\hat{r}_k^j| \geqslant C\}.$$

The stopping time $\tau_j^{(1)}(C)$ is the first time the (random) sequence $\{\hat{r}_k^j\}$ exits a ball of radius C. (We will often use the simpler notation $\tau_j^{(1)}$, if the value of C is clear from the context.) The following lemma derives bounds on the "effect" of the perturbation noise $\hat{\varepsilon}_k^j$ before time $\tau_j^{(1)}$.

Lemma 8. *For any given $C > 0$, there exists a constant $C_1 > 0$ such that for all j, we have*

$$\mathbf{E}\left[\max_{k_j < k \leqslant \tau_j^{(1)} \wedge k_{j+1}} \left| \sum_{l=k_j+1}^{k} \gamma_l \hat{\varepsilon}_{l+1}^j \right|^2 \right] \leqslant C_1 \sum_{k=k_j+1}^{k_{j+1}} \gamma_k^2.$$

Proof. We only outline the proof as this result is similar to Proposition 7, in [1, pp. 228]. The main difference is that this proposition considers the sum $\sum_{l=0}^{k-1} \gamma_l \hat{\varepsilon}_{l+1}^j$, whereas we are interested in the sum $\sum_{l=k_j+1}^{k} \gamma_l \hat{\varepsilon}_{l+1}^j$. The difference in the initial limit of the summation does not affect the proof. However, the difference in the final limit of the summation could affect the proof, because $\hat{r}_k^j I\{\tau_j^{(1)} = k\}$ is not bounded. However, we note that the proof in [1] still goes through, as long as $\hat{r}_k^j I\{\tau_j^{(1)} = k\}$ has bounded moments. To see that $\hat{r}_k^j I\{\tau_j^{(1)} = k\}$ has bounded moments, we observe that it is affine in $\hat{r}_{k-1}^j I\{\tau_j^{(1)} = k\}$ (which is bounded), with the coefficients having bounded moments.

The outline of the rest of the proof is as follows. Consider a fixed j, and suppress the superscript j to simplify notation. Note that the perturbation noise $\hat{\varepsilon}_{l+1}$ is of the form

$$F_{\theta_l}(\hat{r}_l; Y_{l+1}) - \bar{F}_{\theta_l}(\hat{r}_l) + \xi_{l+1}\hat{r}_l,$$

where $\bar{F}_\theta(r)$ is the steady-state expectation of $F_\theta(r; \bar{Y}_l)$, and where $\bar{Y}_l$ is a Markov chain with transition kernel P_θ. Using Assumption 3, it is easy to see that for each θ, r there exists a solution $\hat{F}_\theta(r; y)$ to the Poisson equation:

$$\hat{F}_\theta(r; y) = F_\theta(r; y) - \bar{F}_\theta(r) + (P_\theta \hat{F}_\theta)(r; y).$$

The perturbation noise can be expressed in terms of $\hat{F}$ as follows:

$$\begin{aligned}
\hat{\varepsilon}_{l+1} &= \xi_{l+1}\hat{r}_l + F_{\theta_l}(\hat{r}_l; Y_{l+1}) - \bar{F}_{\theta_l}(\hat{r}_l) \\
&= \xi_{l+1}\hat{r}_l + (\hat{F}_{\theta_l}(\hat{r}_l; Y_{l+1}) - (P_{\theta_l}\hat{F}_{\theta_l})(\hat{r}_l; Y_{l+1}) \\
&= (\xi_{l+1}\hat{r}_l + \hat{F}_{\theta_l}(\hat{r}_l; Y_{l+1}) - (P_{\theta_l}\hat{F}_{\theta_l})(\hat{r}_l; Y_l)) \\
&\quad + ((P_{\theta_{l-1}}\hat{F}_{\theta_{l-1}})(\hat{r}_{l-1}; Y_l) - (P_{\theta_l}\hat{F}_{\theta_l})(\hat{r}_l; Y_{l+1})) \\
&\quad + ((P_{\theta_l}\hat{F}_{\theta_l})(\hat{r}_l; Y_l) - (P_{\theta_l}\hat{F}_{\theta_l})(\hat{r}_{l-1}; Y_l)) \\
&\quad + ((P_{\theta_l}\hat{F}_{\theta_l})(\hat{r}_{l-1}; Y_l) - (P_{\theta_{l-1}}\hat{F}_{\theta_{l-1}})(\hat{r}_{l-1}; Y_l)).
\end{aligned}$$

To prove the lemma, it is sufficient to show that the desired inequality holds when $\hat{\varepsilon}_l^j$ is replaced by each of the terms on the right-hand side of the above equation. Indeed, the first term is a martingale difference with bounded second moment and the second term is the summand in a telescoping series. The last two terms are of the order $\mathrm{O}(|\hat{r}_l - \hat{r}_{l-1}|)$ and $\mathrm{O}(|\theta_{l+1} - \theta_l|)$, respectively. Using these observations, it is easily shown that each of these terms satisfies the desired inequality. □

Lemma 8 indicates that as long as $\hat{r}_k$ is bounded, the perturbation noise remains negligible. In the next lemma, we prove that the sequence $\{\hat{r}_k^j\}$ closely approximates r_k^j.

Lemma 9. $\mathrm{Lim}_j \max_{k_j \leqslant k \leqslant k_{j+1}} |\hat{r}_k^j - r_k^j| = 0$, *w.p.*1.

Proof. Since $\bar{G}(\cdot)$ is bounded, there exists a constant D such that

$$|\hat{r}_{k+1}^j - r_{k+1}^j| \leqslant D \sum_{l=k_j}^{k} \gamma_l |\hat{r}_l^j - r_l^j| + \left| \sum_{l=k_j}^{k} \gamma_l \hat{\varepsilon}_{l+1} \right|$$

for every $k \geqslant k_j$. Using the discrete Gronwall inequality, [1] it is easy to see that

$$\max_{k_j \leqslant k \leqslant k_{j+1} \wedge \tau_j^{(1)}} |\hat{r}_{k+1}^j - r_{k+1}^j|$$

$$\leqslant e^{DT'} \max_{k_j \leqslant k \leqslant k_{j+1} \wedge \tau_j^{(1)}} \left| \sum_{l=k_j}^{k} \gamma_l \hat{\varepsilon}_{l+1} \right|,$$

where $T' = T + \gamma_{k_{j+1}}$. Therefore, Lemma 8 and the Chebyshev inequality imply that

$$\mathbf{P}\left(\max_{k_j \leqslant k \leqslant k_{j+1} \wedge \tau_j^{(1)}} |\hat{r}_{k+1}^j - r_{k+1}^j| \geqslant \delta \right) \leqslant \frac{C_1}{\delta^2} \sum_{l=k_j}^{k_{j+1}-1} \gamma_l^2$$

for some $C_1 > 0$ that depends on the constant C in the definition of the stopping time $\tau_j^{(1)}(C)$. Consider the stopping time $\tau_j^{(2)}(\delta)$ defined by

$$\tau_j^{(2)}(\delta) = \min\{k \geqslant k_j\colon |\hat{r}_k^j - r_k^j| \geqslant \delta\},$$

which is the first time that the sequence $\{\hat{r}_k^j\}$ exits from a tube of radius δ around the sequence $\{r_k^j\}$. To prove the lemma, we need bounds on $\mathbf{P}(\tau_j^{(2)} \leqslant k_{j+1})$, whereas we only have bounds on $\mathbf{P}(\tau_j^{(2)} \leqslant k_{j+1} \wedge \tau_j^{(1)})$. Therefore, we need to relate the stopping times $\tau_j^{(1)}$ and $\tau_j^{(2)}$. To do this, note that

$$\sup_j \max_{k_j \leqslant k} |r_k^j| \leqslant C$$

for some constant $C > 1$, because $\bar{h}(\cdot)$ and $\bar{G}(\cdot)$ are bounded, and the function $\bar{G}(\cdot)$ satisfies Assumption 6. At time $k = \tau_j^{(1)}(C + \delta)$, we have $|\hat{r}_k^j| \geqslant C + \delta$ and $|r_k^j| \leqslant C$, which implies that $|\hat{r}_k^j - r_k^j| \geqslant \delta$, i.e.,

$$\tau_j^{(2)}(\delta) \leqslant \tau_j^{(1)}(C + \delta).$$

Therefore,

$$\mathbf{P}\left(\max_{k_j \leqslant k \leqslant k_{j+1}} |\hat{r}_{k+1}^j - r_{k+1}^j| \geqslant \delta \right)$$

$$= \mathbf{P}(\tau_j^{(2)}(\delta) \leqslant k_{j+1})$$

$$= \mathbf{P}(\tau_j^{(2)}(\delta) \leqslant k_{j+1} \wedge \tau_j^{(1)}(C + \delta))$$

$$\leqslant \frac{C_1}{\delta^2} \sum_{l=k_j}^{k_{j+1}-1} \gamma_l^2.$$

The result follows from the summability of the series $\sum_k \gamma_k^2$ and the Borel–Cantelli lemma. □

Now we are ready to prove boundedness, using the following simple results.

Lemma 10. *Suppose that* $0 \leqslant \lambda < 1$ *and that* $\{a_k\}$, $\{\delta_k\}$ *are non-negative sequences that satisfy* $a_{k+1} \leqslant \lambda a_k + \delta_k$.

(1) *If* $\sup_k \delta_k < \infty$, *then* $\sup_k a_k < \infty$.
(2) *If* $\delta_k \to 0$, *then* $a_k \to 0$.

Lemma 11. *If an* $m \times m$ *matrix* G *satisfies*

$$r'Gr \geqslant \delta |r|^2 \quad \forall r \in \mathbb{R}^m,$$

then for sufficiently small $\gamma > 0$,

$$|(I - \gamma G)r| \leqslant (1 - \tfrac{1}{2}\gamma\delta)|r|.$$

Lemma 12. $\sup_k |r_k| < \infty$, *w.p.*1.

Proof. Since $\bar{h}(\cdot)$ is bounded, Assumption 6 and Lemma 11 imply the following. There exists a constant C such that, for j sufficiently large, and $k \geqslant k_j$,

$$|r_{k+1}^j| \leqslant \left(1 - \tfrac{1}{2}\gamma_k a\right) |r_k^j| + \gamma_k \frac{C}{\max(1, |r_{k_j}|)}.$$

Using the inequality $1 - x \leqslant e^{-x}$, we have

$$|r_{k+1}^j| \leqslant e^{-\left(\frac{1}{2} a \sum_{l=k_j}^{k} \gamma_l\right)} |r_{k_j}^j| + \left(\sum_{l=k_j}^{k} \gamma_l \right) \frac{C}{\max(1, |r_{k_j}|)}. \quad (2)$$

[1] For a non-negative sequence $\{\gamma_k\}$ and a constant $B > 0$, let $\{b_k\}$ be a sequence satisfying $b_0 = 0$ and

$$b_{k+1} \leqslant \sum_{l=0}^{k} b_k \gamma_k + B \qquad \forall k.$$

Then, for every k, we have

$$b_{k+1} \leqslant B \exp\left(\sum_{l=0}^{k} \gamma_l \right).$$

This, along with Lemma 9, implies

$$\frac{|r_{k_{j+1}}|}{\max(1,|r_{k_j}|)} \leqslant \mathrm{e}^{-aT/2}\frac{|r_{k_j}|}{\max(1,|r_{k_j}|)} + T\frac{C}{\max(1,|r_{k_j}|)} + \delta_j,$$

where $\delta_j \to 0$, w.p.1. Multiplying both sides by $\max(1,|r_{k_j}|)$ and using the fact that it is less than $1+|r_{k_j}|$, we have

$$|r_{k_{j+1}}| \leqslant (\mathrm{e}^{-aT/2}+\delta_j)|r_{k_j}| + CT + \delta_j.$$

Since $\mathrm{e}^{-aT} < 1$ and $\delta_j \to 0$, w.p.1, it follows from Lemma 10(a) that

$$\sup_j |r_{k_j}| < \infty, \quad \text{w.p.1.}$$

Recall that $r^j_{k_j} = \hat{r}^j_{k_j}$ is bounded. Then, using Eq. (2) it follows that

$$\sup_{j,k\geqslant k_j} |r^j_k| < \infty \quad \text{w.p.1.}$$

The boundedness of $|r_k|$ now follows from the observation:

$$\begin{aligned}\sup_k |r_k| &= \sup_j (\max(1,|r_{k_j}|)\cdot \max_{k_j\leqslant k\leqslant k_{j+1}} |\hat{r}^j_k|)\\ &\leqslant \sup_j \{(1+|r_{k_j}|)(\max_{k_j\leqslant k\leqslant k_{j+1}}|r^j_k| \\ &\quad + \max_{k_j\leqslant k\leqslant k_{j+1}} |r^j_k - \hat{r}^j_k|)\},\end{aligned}$$

the boundedness of r^j_k and r_{k_j}, and Lemma 9. □

5. Proof of Theorem 7

To prove Theorem 7, we consider the sequence $\hat{\rho}_k = \bar{G}(\theta_k)r_k - \bar{h}(\theta_k)$. As noted in Section 3, this sequence evolves according to the iteration:

$$\hat{\rho}_{k+1} = \hat{\rho}_k - \gamma_k\bar{G}(\theta_{k+1})\hat{\rho}_k + \gamma_k\varepsilon^{(1)}_{k+1} + \gamma_k\varepsilon^{(2)}_{k+1},$$

where

$$\begin{aligned}\varepsilon^{(1)}_{k+1} &= \bar{G}(\theta_{k+1})(h_{\theta_k}(Y_{k+1}) - \bar{h}(\theta_k)) - \bar{G}(\theta_{k+1})\\ &\quad\times(G_{\theta_k}(Y_{k+1}) - \bar{G}(\theta_k))r_k + \bar{G}(\theta_{k+1})\xi_{k+1}r_k,\end{aligned}$$

$$\begin{aligned}\varepsilon^{(2)}_{k+1} &= \frac{1}{\gamma_k}((\bar{G}(\theta_{k+1}) - \bar{G}(\theta_k))r_k\\ &\quad -(\bar{h}(\theta_{k+1}) - \bar{h}(\theta_k))).\end{aligned}$$

Lemma 13. *$\sum_k \gamma_k\varepsilon^{(1)}_{k+1}$ converges w.p.1.*

Proof. The proof relies on Assumptions 1, 3, and 5, and is omitted because it is very similar to the proof of Lemma 2 in [2, pp. 224]. □

Lemma 14. *$\lim_k \varepsilon^{(2)}_k = 0$, w.p.1.*

Proof. Using Assumption 3(4) and the update equation for θ_k, we have

$$\begin{aligned}|\varepsilon^{(2)}_{k+1}| &\leqslant \frac{C}{\gamma_k}|\theta_{k+1}-\theta_k|(|r_k|+1)\\ &= C\frac{\beta_k}{\gamma_k}|H_k|(|r_k|+1).\end{aligned}$$

Since $\{H_k\}$ has bounded moments, the second part of Assumption (4) yields

$$\mathbf{E}\left[\sum_k\left(\frac{\beta_k}{\gamma_k}\right)^d |H_k|^d\right] < \infty$$

for some $d>0$. Therefore, $(\beta_k/\gamma_k)H_k$ converges to zero, w.p.1. The result follows from the boundedness of $\{r_k\}$. □

Recall the notation k_j from the previous section. For each j, define ρ^j_k, for $k \geqslant k_j$ as follows:

$$\rho^j_{k+1} = (I - \gamma_k\bar{G}(\theta_k))\rho^j_k, \quad \rho^j_{k_j} = \hat{\rho}_{k_j}.$$

Lemma 15. *$\lim_j \max_{k_j\leqslant k\leqslant k_{j+1}} |\hat{\rho}_k - \rho^j_k| = 0$, w.p.1.*

Proof. Since $\bar{G}(\theta_k)$ is bounded, there exists some C such that for every j and $k\geqslant k_j$,

$$|\hat{\rho}_k - \rho^j_k| \leqslant C\sum_{l=k_j}^{k-1}\gamma_l|\hat{\rho}_l - \rho^j_l| + \left|\sum_{l=k_j}^{k-1}\gamma_l(\varepsilon^{(1)}_{l+1} + \varepsilon^{(2)}_{l+1})\right|.$$

Using the discrete Gronwall inequality, it can be seen that

$$\begin{aligned}&\max_{k_j\leqslant k\leqslant k_{j+1}} |\hat{\rho}_k - \rho^j_k|\\ &\quad\leqslant \mathrm{e}^{CT}\max_{k_j\leqslant k\leqslant k_{j+1}}\left|\sum_{l=k_j}^{k-1}\gamma_l(\varepsilon^{(1)}_{l+1}+\varepsilon^{(2)}_{l+1})\right|\\ &\quad\leqslant \mathrm{e}^{CT}\sup_{k\geqslant k_j}\left|\sum_{l=k_j}^{k-1}\gamma_l\varepsilon^{(1)}_{l+1}\right| + \mathrm{e}^{CT}T\sup_{k\geqslant k_j}|\varepsilon^{(2)}_{k+1}|.\end{aligned}$$

The result follows from the previous two lemmas. □

We are now ready to complete the proof of Theorem 7. Using Assumption 6 and Lemma 11, we have

$$|\rho^j_{k_{j+1}}| \leqslant \mathrm{e}^{-aT/2}|\rho^j_{k_j}|.$$

Therefore Lemma 15 implies that

$$|\hat{\rho}_{k_{j+1}}| \leqslant \mathrm{e}^{-aT/2}|\hat{\rho}_{k_j}| + \delta_j,$$

where $\delta_j \to 0$ w.p.1. Using Lemma 10(b), it follows that $|\bar{G}(\theta_{k_j})r_{k_j} - \bar{h}(\theta_{k_j})| = |\hat{\rho}_{k_j}|$ converges to zero. Using arguments similar to the closing arguments in the proof of Lemma 12, we finally conclude that

$$\lim_k |\bar{G}(\theta_k)r_k - \bar{h}(\theta_k)| = 0, \quad \text{w.p.1.} \qquad \square$$

Acknowledgements

This research was partially supported by the National Science Foundation under contract ECS-9873451, and by the AFOSR under contract F49620-99-1-0320.

References

[1] A. Benveniste, M. Metivier, P. Priouret, Adaptive Algorithms and Stochastic Approximations, Springer, Berlin, 1990.

[2] D.P. Bertsekas, J.N. Tsitsiklis, Neuro-Dynamic Programming, Athena Scientific, Belmont, MA, 1996.

[3] V.S. Borkar, Stochastic approximation with two time scales, Systems Control Lett. 29 (1997) 291–294.

[4] V.S. Borkar, S.P. Meyn, The o.d.e. method for convergence of stochastic approximation and reinforcement learning, SIAM J. Control Optim. 38 (2) (2000) 447–469.

[5] M. Duflo, Random Iterative Models, Springer, New York, 1997.

[6] E. Eweda, O. Machi, Convergence of an adaptive linear estimation algorithm, IEEE Trans. Automat. Control AC-29 (2) (1984) 119–127.

[7] V.R. Konda, Actor-critic algorithms, Ph.D. Thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, 2002.

[8] V.R. Konda, J.N. Tsitsiklis, On actor-critic algorithms, SIAM J. Control Optim., to appear.

[9] H.H. Kushner, Asymptotic behavior of stochastic approximation and large deviations, IEEE Trans. Automat. Control 29 (1984) 984–990.

[10] H.J. Kushner, G.G. Yin, Stochastic Approximation Algorithms and Applications, Springer, New York, 1997.

[11] S.P. Meyn, R.L. Tweedie, Markov Chains and Stochastic Stability, Springer, London, 1993.

[12] R.S. Sutton, Learning to predict by the methods of temporal differences, Mach. Learning 3 (1988) 9–44.

[13] R. Sutton, A. Barto, Reinforcement Learning: An Introduction, MIT Press, Cambridge, MA, 1998.

[14] B. Widrow, J. McCool, M.G. Larimore, C.R. Johnson, Stationary and non-stationary learning characteristics of the LMS adaptive filter, Proc. IEEE 64 (8) (1976) 1151–1161.