Proceeding of the 2004 American Control Conference
Boston, Massachusetts June 30 - July 2, 2004

Almost Sure Convergence of Two Time-Scale Stochastic Approximation Algorithms

Vladislav B. Tadić

***Abstract*— The almost sure convergence of two time-scale stochastic approximation algorithms is analyzed under general noise and stability conditions. In the context of the Lyapunov stability, the adopted stability conditions are probably the weakest possible still allowing the almost sure convergence to be shown, while the corresponding noise conditions are the most general ones under which the almost sure convergence analysis can be carried out. The analysis covers the algorithms with additive noise, as well as those with non-additive noise. The algorithms with additive noise are analyzed for the case where the noise is state-dependent. The analysis of the algorithms with non-additive state-dependent noise is carried out for the case where the noise is a Markov chain controlled by the algorithm states, while the algorithms with non-additive exogenous noise are analyzed for the case where the noise is correlated and satisfies strong mixing conditions. The obtained results cover a fairly broad class of highly non-linear two time-scale stochastic approximation algorithms.**

***Index Terms*— Two time-scale stochastic approximation, almost sure convergence, strong mixing conditions, controlled Markov chains, Lyapunov stability, actor-critic algorithms.**

I. Introduction

In this paper, the almost sure convergence of two time-scale stochastic approximation algorithms with decreasing step sizes is analyzed. Generally speaking, stochastic approximation algorithms are sequential non-parametric methods for finding a zero or minimum of a function in the situation where only the noise corrupted observations of the function values are available (see [2], [12] and references cited therein). Two time-scale stochastic approximation algorithms represent one of the most general and complex subclasses of stochastic approximation methods. These algorithms consist of two sub-recursions which are updated with different step sizes (i.e., which evolve on different time scales). The main feature of two time-scale stochastic approximation is that step sizes of one of the sub-recursions (slow one) are considerably smaller than the step size of the another (fast) one. Owing to this feature, stochastic approximation with two scales can be considered as singularly perturbed stochastic difference equations (for more details on singularly perturbed systems see e.g., [14]). During the last five years, two time-scale stochastic approximation algorithms have successfully been applied to several complex problems arising in the ares of reinforcement learning [1], [9], [10], [11], signal processing [6] and admission control in communication networks [5] (to name a few), while their asymptotic properties have thoroughly been analyzed in several papers [4], [9], [10], [11], [17]. Although [4], [9], [10], [11] provide an insight into the almost sure asymptotic properties of two time-scale stochastic approximation, the results presented therein either hold under fairly restrictive conditions or correspond only to the almost sure convergence of subsequences of the algorithm states.

In this paper, the almost sure convergence of two time-scale stochastic approximation algorithms with decreasing step sizes is analyzed under general noise and stability conditions. In the context of the Lyapunov stability, the adopted stability conditions are probably the weakest possible still allowing the almost sure convergence to be shown, while the corresponding noise conditions are the most general ones under which the almost sure convergence analysis can be carried out. The analysis covers the algorithms with additive noise, as well as those with non-additive noise. The algorithms with additive noise are analyzed for the case where the noise is state-dependent. The analysis of the algorithms with non-additive state-dependent noise is carried out for the case where the noise is a Markov chain controlled by the algorithm states, while the algorithms with non-additive exogenous noise are analyzed for the case where the noise is correlated and satisfies strong mixing conditions. The obtained results cover a fairly broad class of highly non-linear two time-scale stochastic approximation algorithms (including the actor-critic learning algorithms introduced in [10], [11]).

II. Algorithms with Additive State-Dependent Noise

The algorithms considered in this section are defined by the following difference equations:

$$x_{n+1} = x_n + \alpha_{n+1} f(x_n, y_n) + \alpha_{n+1} u_{n+1}, \quad n \geq 0, \quad (1)$$

$$y_{n+1} = y_n + \beta_{n+1} g(x_n, y_n) + \beta_{n+1} v_{n+1}, \quad n \geq 0. \quad (2)$$

$\{\alpha_n\}_{n\geq 1}$, $\{\beta_n\}_{n\geq 1}$ are sequences of positive reals, while $f : R^p \times R^q \to R^p$ and $g : R^p \times R^q \to R^q$ are locally Lipschitz continuous functions. x_0 and y_0 are R^p-valued and R^q-valued random variables (respectively) defined on a probability space $(\Omega, \mathcal{F}, P)$, while $\{u_n\}_{n\geq 1}$ and $\{v_n\}_{n\geq 1}$ are R^p-valued and R^q-valued stochastic processes (respectively) defined on the same probability space.

$\{\alpha_n\}_{n\geq 1}$, $\{\beta_n\}_{n\geq 1}$ are the step sizes of the algorithm (1), (2), while $\{u_n\}_{n\geq 1}$, $\{v_n\}_{n\geq 1}$ are considered as the noise in the same algorithm. The analysis (the results of which are presented in this section) is carried out for the case where

Vladislav B. Tadić is with Department of Automatic Control and Systems Engineering, University of Sheffield, Sheffield S1 3JD, United Kingdom, v.tadic@sheffield.ac.uk

0-7803-8335-4/04/$17.00 ©2004 AACC

the noise $\{u_n\}_{n\geq 1}$, $\{v_n\}_{n\geq 1}$ depend on the algorithm states $\{x_n\}_{n\geq 0}$, $\{y_n\}_{n\geq 0}$, i.e.,

$$u_{n+1} = U_{n+1}(x_0, y_0, \ldots, x_n, y_n), \quad n \geq 0, \tag{3}$$

$$v_{n+1} = V_{n+1}(x_0, y_0, \ldots, x_n, y_n), \quad n \geq 0, \tag{4}$$

where $U_n : R^{n(p+q)} \to R^p$ and $V_n : R^{n(p+q)} \to R^q$ are random functions. Since the algorithms with non-additive noise (Sections III, IV) can be represented in the form (1) – (4), the results presented in this section could be considered as a basis for the analysis of the algorithms with non-additive noise.

For $t \in (0, \infty)$, let $a_n(t) = \sup\{j \geq n : \sum_{i=n}^{j-1} \alpha_{i+1} \leq t\}$, $b_n(t) = \sup\{j \geq n : \sum_{i=n}^{j-1} \beta_{i+1} \leq t\}$, $n \geq 1$. The almost sure convergence of the algorithm (1), (2) is analyzed under the following assumptions:

A1: $\lim_{n\to\infty} \alpha_n = \lim_{n\to\infty} \beta_n = \lim_{n\to\infty} \alpha_n \beta_n^{-1} = 0$, $\sum_{n=1}^{\infty} \alpha_n = \sum_{n=1}^{\infty} \beta_n = \infty$.

A2: For all $\rho, t \in [1, \infty)$,

$$\lim_{n\to\infty} \sup_{n\leq j<a_n(t)} \left\| \sum_{i=n}^{j-1} \alpha_{i+1} u_{i+1} \right\| I_{\{\lambda_\rho \geq j\}} = 0 \ w.p.1,$$

$$\lim_{n\to\infty} \sup_{n\leq j<b_n(t)} \left\| \sum_{i=n}^{j-1} \beta_{i+1} v_{i+1} \right\| I_{\{\lambda_\rho \geq j\}} = 0 \ w.p.1,$$

where

$$\lambda_\rho = \inf(\{n \geq 0 : \max\{\|x_n\|, \|y_n\|\} > \rho\} \cup \{\infty\}).$$

A3: There exists a locally Lipschitz continuous function $\psi : R^p \to R^q$ such that $\psi(x)$ is a globally asymptotically stable point of the ODE $dy/dt = g(x, y)$ for all $x \in R^p$.

A4: There exists a differentiable function $u : R^p \to R$ such that:

(i) $\nabla u(\cdot)$ is locally Lipschitz continuous,

(ii) $\dot{u}(x) < 0$ for all $x \in E_*^c$,

(iii) the Lebesgue measure of $u(E_*) \cap u(E_*^c)$ is zero,

where $E_* = \{x \in R^p : f(x, \psi(x)) = 0\}$ and $\dot{u}(x) = \nabla^T u(x) f(x, \psi(x))$.

A1 corresponds to the asymptotic properties of the step sizes $\{\alpha_n\}_{n\geq 1}$, $\{\beta_n\}_{n\geq 1}$, and is standard for the almost sure convergence analysis of two time-scale stochastic approximation algorithms (see e.g. [4]). It holds if $\alpha_n = n^{-a}$, $\beta_n = n^{-b}$, $n \geq 1$, where $a, b \in (0, 1]$ are constants satisfying $a > b$. Moreover, A1 implies that the states $\{x_n\}_{n\geq 0}$ of the recursion (1) evolves on a slower time-scale compared to the states $\{y_n\}_{n\geq 0}$ of the recursion (2).

A2 corresponds to the asymptotic properties of the noise $\{u_n\}_{n\geq 1}$, $\{v_n\}_{n\geq 1}$. It can be considered as a two time-scale generalization of the classical Kushner-Clark noise condition. In the context of single time-scale stochastic approximation, the Kushner-Clark condition is the weakest condition under which the almost sure convergence can be demonstrated (for more details see e.g., [12]). Moreover, under certain (relatively restrictive) stability conditions, the Kushner-Clark condition is necessary and sufficient for the almost sure convergence of single time-scale stochastic approximation algorithms (see e.g., [19]).

A3 and A4 are stability conditions. A3 corresponds to the stability properties of the fast recursion (2) (i.e., to the stability of the family of the ODEs $dy/dt = g(x, y)$, $x \in R^p$) and is standard for the asymptotic analysis of two time-scale stochastic approximation algorithms (see e.g., [4]). A4 is related to the stability properties of the slow recursion (1) (i.e., to the stability of the ODE $dx/dt = f(x, \psi(x))$). Conditions (i), (ii) of A4 require the ODE $dx/dt = f(x, \psi(x))$ to have a global Lyapunov function $u(\cdot)$. In the context of the Lyapunov stability, this requirement represents the weakest condition under which the Lagrange stable solutions of the ODE $dx/dt = f(x, \psi(x))$ converge to the set of zeros of $f(\cdot, \psi(\cdot))$ (i.e., to E_*). On the other hand, condition (iii) of A4 is specific for the almost sure convergence of stochastic approximation algorithms and does not have an interpretation in the context of the Lyapunov stability. Basically, it ensures the Lyapunov function $u(\cdot)$ to admit the following topological property: each closed continuous path starting and ending in E_*^c has a subpath contained in E_*^c along which $u(\cdot)$ does not increase. This property prevents the noise $\{u_n\}_{n\geq 1}$, $\{v_n\}_{n\geq 1}$ from forcing the slowly varying states $\{x_n\}_{n\geq 0}$ to drift from one connected component of E_* to another (which itself ensures $\{x_n\}_{n\geq 0}$ to converge to a connected component of E_*). Condition (iii) of A4 has been introduced in [15], [16] and represents a generalization of the corresponding condition proposed in [7]. It holds if E_* or $u(E_*)$ are countable. It is also satisfied in the following case: $f(x, \psi(x)) = -\nabla J(x)$ for all $x \in R^p$, where $J : R^p \to R$ is a differentiable function satisfying the condition that $\{x \in R^p : \nabla J(x) = 0\}$ is nowhere dense. On the other hand, due to the Morse-Sard theorem (see e.g., [13]), $\{x \in R^p : \nabla J(x) = 0\}$ is nowhere dense if $J(\cdot)$ is p-times differentiable. The case described above is quite common for the two time-scale stochastic approximation appearing in the area of neuro-dynamic programming (see e.g., [10], [11]).

The main results on the almost sure convergence of the algorithm (1), (2) under assumptions A1 – A4 are contained in the next theorem.

Theorem 1: Let A1 – A4 hold. Then, $\lim_{n\to\infty} d(x_n, E_*) = \lim_{n\to\infty} \|y_n - \psi(x_n)\| = 0 \ w.p.1$ on the event $\{\sup_{0\leq n} \|x_n\| < \infty\} \cap \{\sup_{0\leq n} \|y_n\| < \infty\}$.

For the proof, see [18].

Let $\mathcal{F}_0 = \sigma\{x_0, y_0, u_0, v_0\}$ and $\mathcal{F}_n = \mathcal{F}_0 \vee \sigma\{u_n, v_n : n \geq 1\}$, $n \geq 1$. The almost sure convergence of the algorithm (1), (2) is also analyzed under the following assumptions:

B1: $\lim_{n\to\infty} \alpha_n = \lim_{n\to\infty} \beta_n = 0$, $\alpha_n - \alpha_{n+1} = O(\alpha_n^2)$, $\beta_n - \beta_{n+1} = o(\beta_n^2)$, $\alpha_n = O(\beta_n^r)$, $\sum_{n=1}^{\infty} \alpha_n = \sum_{n=1}^{\infty} \beta_n = \infty$, $\sum_{n=1}^{\infty} \alpha_n^s < \infty$, $\sum_{n=1}^{\infty} \beta_n^2 < \infty$, where $r, s \in (1, \infty)$ are constants satisfying $r < 2$, $1/r + 2s \leq 3$.

B2: There exist R^p-valued stochastic processes $\{u_{1,n}\}_{n\geq 1}$, $\{u_{2,n}\}_{n\geq 1}$, $\{u_{3,n}\}_{n\geq 0}$, R^q-valued stochastic processes $\{v_{1,n}\}_{n\geq 1}$, $\{v_{2,n}\}_{n\geq 1}$, $\{v_{3,n}\}_{n\geq 0}$ (defined on

$(\Omega, \mathcal{F}, P)$) and for all $\rho \in [1, \infty)$, there exists a constant $C_\rho \in [1, \infty)$ such that

$$u_{n+1} = u_{1,n+1} + u_{2,n+1} + u_{3,n+1} - u_{3,n}, \quad n \geq 0,$$

$$v_{n+1} = v_{1,n+1} + v_{2,n+1} + v_{3,n+1} - v_{3,n}, \quad n \geq 0,$$

$$E\left(u_{1,n+1} I_{\{\lambda_\rho > n\}} | \mathcal{F}_n\right) = 0 \ w.p.1, \quad n \geq 0,$$

$$E\left(v_{1,n+1} I_{\{\lambda_\rho > n\}} | \mathcal{F}_n\right) = 0 \ w.p.1, \quad n \geq 0,$$

$$\max\{E(\|u_{1,n}\|^2 I_{\{\lambda_\rho \geq n\}}), E(\|v_{1,n}\|^2 I_{\{\lambda_\rho \geq n\}})\} \leq C_\rho, \quad n \geq 1,$$

$$\max\{E(\|u_{2,n}\|^2 I_{\{\lambda_\rho \geq n\}}), E(\|v_{2,n}\|^2 I_{\{\lambda_\rho \geq n\}})\} \leq C_\rho(\alpha_n + \beta_n)^2, \quad n \geq 1,$$

$$\max\{E(\|u_{3,n}\|^2 I_{\{\lambda_\rho \geq n\}}), E(\|v_{3,n}\|^2 I_{\{\lambda_\rho \geq n\}})\} \leq C_\rho, \quad n \geq 1,$$

where

$$\lambda_\rho = \inf(\{n \geq 0 : \max\{\|x_n\|, \|y_n\|\} > \rho\} \cup \{\infty\}).$$

B3: $g(\cdot, \cdot)$ is differentiable and $\nabla_x g(\cdot, \cdot)$, $\nabla_y g(\cdot, \cdot)$ are locally Lipschitz continuous. There exists a differentiable function $\psi : R^p \to R^q$ such that $\nabla \psi(\cdot)$ is locally Lipschitz continuous and $\psi(x)$ is a globally exponentially stable point of the ODE $dy/dt = g(x, y)$ for all $x \in R^p$.

B4: There exists a differentiable function $u : R^p \to R$ such that $\nabla u(\cdot)$ is locally Lipschitz continuous and $\dot{u}(x) < 0$ for all $x \in E_*^c$, where $E_* = \{x \in R^p : f(x, \psi(x)) = 0\}$ and $\dot{u}(x) = \nabla^T u(x) f(x, \psi(x))$.

B1 corresponds to the asymptotic properties of the step sizes $\{\alpha_n\}_{n \geq 1}$, $\{\beta_n\}_{n \geq 1}$. It holds if $\alpha_n = n^{-1}$, $\beta_n = n^{-b}$, $n \geq 1$, where $b \in (1/2, 1)$ is a constant. Moreover, B1 implies that the states $\{x_n\}_{n \geq 0}$ of the recursion (1) evolves on a slower time-scale compared to the states $\{y_n\}_{n \geq 0}$ of the recursion (2).

B2 is a noise condition. Basically, it requires the noise $\{u_n\}_{n \geq 1}$, $\{v_n\}_{n \geq 1}$ to be decomposable as a sum of a martingale-difference sequence ($\{u_{1,n}\}_{n \geq 1}$, $\{v_{1,n}\}_{n \geq 1}$), a vanishing sequence ($\{u_{2,n}\}_{n \geq 1}$, $\{v_{2,n}\}_{n \geq 1}$) and a telescoping sequence ($\{u_{3,n}\}_{n \geq 0}$, $\{v_{3,n}\}_{n \geq 0}$). Compared to A2, B2 is more restrictive. In return, it allows the corresponding stability conditions to be significantly more general (see the comments on B3, B4, next paragraph). Moreover, B2 is still applicable to the analysis of two time-scale stochastic approximation algorithms with non-additive noise and covers several, fairly complex classes of both exogenous and state-dependent non-additive noise (see Sections III, IV). As stability conditions are usually much harder to be verified than noise ones, it is important to demonstrate the almost sure convergence under noise conditions which allow the stability conditions to be the weakest possible and still cover complex classes of exogenous and state-dependent noise.

B3 and B4 are stability conditions. B3 corresponds to the stability properties of the fast recursion (2) (i.e., to the stability of the family of the ODEs $dy/dt = g(x, y)$, $x \in R^p$) and is quite common for the almost sure convergence analysis of two time-scale stochastic approximation algorithms (see e.g., [4]). B4 is related to the stability properties of the slow recursion (1) (i.e., to the stability of the ODE $dx/dt = f(x, \psi(x))$). It requires the ODE $dx/dt = f(x, \psi(x))$ to have a global Lyapunov function $u(\cdot)$. In the context of the Lyapunov stability, this requirement represents the weakest condition under which the Lagrange stable solutions of the ODE $dx/dt = f(x, \psi(x))$ converge to the set of zeros of $f(\cdot, \psi(\cdot))$ (i.e., to E_*). Therefore, B4 can be considered as the weakest stability condition ensuring the almost sure convergence of the slowly varying states $\{x_n\}_{n \geq 0}$.

The main results on the almost sure convergence of the algorithm (1), (2) under assumptions B1 – B4 are contained in the following theorem.

Theorem 2: Let B1 – B4 hold. Then, $\lim_{n \to \infty} d(x_n, E_*) = \lim_{n \to \infty} \|y_n - \psi(x_n)\| = 0 \ w.p.1$ on the event $\{\sup_{0 \leq n} \|x_n\| < \infty\} \cap \{\sup_{0 \leq n} \|y_n\| < \infty\}$.

For the proof, see [18].

The almost sure asymptotic behavior of two time-scale stochastic approximation algorithms with decreasing step sizes has been analyzed in [4], [9], [10], [11]. Although [4], [9], [10], [11] provide an insight into their asymptotic behavior, the results presented therein either hold under fairly restrictive conditions or correspond only to the almost sure convergence of subsequences of the slowly-varying states $\{x_n\}_{n \geq 0}$. In [4], the same results as those of Theorems 1, 2 have been demonstrated under the conditions requiring the noise $\{u_n\}_{n \geq 1}$, $\{u_n\}_{n \geq 1}$ to be martingale-difference sequences and the ODE $dx/dt = f(x, \psi(x))$ to have a globally asymptotically stable point. Obviously, these conditions are one of the simplest special cases of A2 – A4. Moreover, the ODE $dx/dt = f(x, \psi(x))$ almost never has a globally asymptotically stable point in the case of highly non-linear algorithms such as actor-critic learning algorithms introduced and analyzed in [10], [11]. On the other hand, A2 – A4 (as well as B2 – B4) cover a fairly broad class of highly non-linear two time-scale stochastic approximation algorithms (including actor-critic learning algorithms studied in [10], [11]; for details see Section V and [18]) and represent probably the weakest noise and stability conditions under which the almost sure convergence can be demonstrated. In [9] – [11], two time-scale stochastic approximation algorithms have been analyzed under conditions which are similar to B1 – B4. In [10], [11], only the existence of an almost sure convergent subsequence of the slowly-varying states $\{x_n\}_{n \geq 0}$ (i.e., $\underline{\lim}_{n \to \infty} d(x_n, E_*) = 0 \ w.p.1$ on the event where $\{x_n\}_{n \geq 0}$, $\{y_n\}_{n \geq 0}$ are bounded) has been shown, while the results of [9] do not necessary hold under the conditions specified therein (notice that [9, Lemma 4.2] is not correct; otherwise, only the attractivity of E_* would be sufficient for its robustness to the perturbations of the ODE $dx/dt = f(x, \psi(x))$, i.e., E_* could be robust even if it were not stable; this would be completely counter-intuitive and to the best of our knowledge, there is not any similar result

in the literature on the ODE stability).

III. Algorithms with Exogenous Non-Additive Noise

Using the results obtained for the algorithms with additive noise (Section II), the almost sure convergence of the following algorithm is analyzed in this section:

$$x_{n+1} = x_n + \alpha_{n+1} F(x_n, y_n, \xi_{n+1}), \quad n \geq 0, \tag{5}$$

$$y_{n+1} = y_n + \beta_{n+1} G(x_n, y_n, \xi_{n+1}), \quad n \geq 0. \tag{6}$$

$\{\alpha_n\}_{n\geq 1}$, $\{\beta_n\}_{n\geq 1}$ are sequences of positive reals, while $F : R^p \times R^q \times R^r \to R^p$ and $G : R^p \times R^q \times R^r \to R^q$ are Borel-measurable functions. x_0 and y_0 are R^p-valued and R^q-valued random variables (respectively) defined on a probability space $(\Omega, \mathcal{F}, P)$, while $\{\xi_n\}_{n\geq 1}$ is an R^r-valued stochastic process defined on the same probability space.

$\{\alpha_n\}_{n\geq 1}$, $\{\beta_n\}_{n\geq 1}$ are the step sizes of the algorithm (5), (6), while $\{\xi_n\}_{n\geq 1}$ is considered as the (non-additive) noise in the same algorithm. The analysis (the results of which are presented in this section) is carried out for the case where $\{\xi_n\}_{n\geq 1}$ is a sequence of identically distributed random variables which satisfy strong mixing conditions and do not depend on the algorithm states $\{x_n\}_{n\geq 0}$, $\{y_n\}_{n\geq 0}$.

Let $\mathcal{F}_0 = \sigma\{x_0, y_0\}$ and $\mathcal{F}_n = \mathcal{F}_0 \vee \sigma\{\xi_n : n \geq 1\}$, $n \geq 1$. Moreover, for $\rho \in (0, \infty)$, let $B^p_\rho = \{x \in R^p : \|x\| \leq \rho\}$, $B^q_\rho = \{y \in R^q : \|y\| \leq \rho\}$. The algorithm (5), (6) is analyzed under the following assumptions:

C1: $F(\cdot, \cdot, \xi)$ and $G(\cdot, \cdot, \xi)$ are differentiable for all $\xi \in R^r$. For all $\rho \in [1, \infty)$, there exists a Borel-measurable function $\varphi_\rho : R^r \to [1, \infty)$ such that

$$\begin{aligned}
&\max\{\|F(x,y,\xi)\|, \|\nabla_x F(x,y,\xi)\|, \|\nabla_y F(x,y,\xi)\|\} \\
&\leq \varphi_\rho(\xi), \\
&\quad \max\{\, \|\nabla_x F(x',y',\xi) - \nabla_x F(x'',y'',\xi)\|, \\
&\qquad \|\nabla_y F(x',y',\xi) - \nabla_y F(x'',y'',\xi)\|\} \\
&\quad \leq \varphi_\rho(\xi)(\|x' - x''\| + \|y' - y''\|), \\
&\max\{\|G(x,y,\xi)\|, \|\nabla_x G(x,y,\xi)\|, \|\nabla_y G(x,y,\xi)\|\} \\
&\leq \varphi_\rho(\xi), \\
&\quad \max\{\, \|\nabla_x G(x',y',\xi) - \nabla_x G(x'',y'',\xi)\|, \\
&\qquad \|\nabla_y G(x',y',\xi) - \nabla_y G(x'',y'',\xi)\|\} \\
&\quad \leq \varphi_\rho(\xi)(\|x' - x''\| + \|y' - y''\|),
\end{aligned}$$

for all $x, x', x'' \in B^p_\rho$, $y, y', y'' \in B^q_\rho$, $\xi \in R^r$.

C2: There exist a probability measure $\kappa(\cdot)$ defined on $(R^r, \mathcal{B}^r)$, constants $a, b \in (1, \infty)$ and a sequence $\{c_n\}_{n\geq 1}$ of positive reals such that $(r+2)a^{-1} + b^{-1} = 1$, $\sum_{n=1}^\infty c_n^{1/a} < \infty$ and

$$\int \varphi_\rho^{2b}(\xi)\kappa(d\xi) < \infty,$$

$$P(\xi_n \in B) = \kappa(B), \quad n \geq 0,$$

$$E|P(\xi_j \in B|\mathcal{F}_n) - \kappa(B)| \leq c_{j-n}, \quad 0 \leq n \leq j,$$

for all $\rho \in [1, \infty)$, $B \in \mathcal{B}^r$.

Remark: For more details on mixing conditions and situations where they hold, see [8].

The main results on the almost sure convergence of the algorithm (5), (6) under assumptions C1, C2 are presented in the next two theorems.

Theorem 3: Let C1, C2 hold. Suppose that A1 is satisfied and A3, A4 are fulfilled with $f(x,y) = \int F(x,y,\xi)\kappa(d\xi)$, $g(x,y) = \int G(x,y,\xi)\kappa(d\xi)$, $x \in R^p$, $y \in R^q$. Then, $\lim_{n\to\infty} d(x_n, E_*) = \lim_{n\to\infty} \|y_n - \psi(x_n)\| = 0$ $w.p.1$ on the event $\{\sup_{0\leq n} \|x_n\| < \infty\} \cap \{\sup_{0\leq n} \|y_n\| < \infty\}$.

Theorem 4: Let C1, C2 hold. Suppose that B1 is satisfied and B3, B4 are fulfilled with $f(x,y) = \int F(x,y,\xi)\kappa(d\xi)$, $g(x,y) = \int G(x,y,\xi)\kappa(d\xi)$, $x \in R^p$, $y \in R^q$. Then, $\lim_{n\to\infty} d(x_n, E_*) = \lim_{n\to\infty} \|y_n - \psi(x_n)\| = 0$ $w.p.1$ on the event $\{\sup_{0\leq n} \|x_n\| < \infty\} \cap \{\sup_{0\leq n} \|y_n\| < \infty\}$.

For the proofs, see [18].

IV. Algorithms with State-Dependent Non-Additive Noise

Using the results obtained for the algorithms with additive noise (Section II), the almost sure convergence of the following algorithm is analyzed in this section:

$$x_{n+1} = x_n + \alpha_{n+1} F(x_n, y_n, \xi_{n+1}), \quad n \geq 0, \tag{7}$$

$$y_{n+1} = y_n + \beta_{n+1} G(x_n, y_n, \xi_{n+1}), \quad n \geq 0. \tag{8}$$

$\{\alpha_n\}_{n\geq 1}$, $\{\beta_n\}_{n\geq 1}$ are sequences of positive reals, while $F : R^p \times R^q \times R^r \to R^p$ and $G : R^p \times R^q \times R^r \to R^q$ are Borel-measurable functions. x_0 and y_0 are R^p-valued and R^q-valued random variables (respectively) defined on a probability space $(\Omega, \mathcal{F}, P)$, while $\{\xi_n\}_{n\geq 0}$ is an R^r-valued stochastic process defined on the same probability space.

$\{\alpha_n\}_{n\geq 1}$, $\{\beta_n\}_{n\geq 1}$ are the step sizes of the algorithm (7), (8), while $\{\xi_n\}_{n\geq 0}$ is considered as (non-additive) noise in the same algorithm. The analysis (the results of which are presented in this section) is carried out for the case where the noise $\{\xi_n\}_{n\geq 0}$ is a homogeneous Markov chain controlled by the algorithm states $\{x_n\}_{n\geq 0}$, $\{y_n\}_{n\geq 0}$, i.e., for all $x \in R^p, y \in R^q$, there exists a transition probability kernel $\Pi(x, y, \cdot, \cdot)$ such that

$$\begin{aligned}
&P(\xi_{n+1} \in B | x_0, y_0, \xi_0, \ldots, x_n, y_n, \xi_n) \\
&= \Pi(x_n, y_n, \xi_n, B) \; w.p.1, \quad n \geq 0,
\end{aligned}$$

for all $B \in \mathcal{B}^r$.

The algorithm (7), (8) is analyzed under the following assumptions:

D1: There exist Borel-measurable functions $\tilde{F} : R^p \times R^q \times R^r \to R^p$, $\tilde{G} : R^p \times R^q \times R^r \to R^q$ and locally Lipschitz continuous functions $f : R^p \times R^q \to R^p$, $g : R^p \times R^q \to R^q$ such that

$$\int \|\tilde{F}(x,y,\xi')\|\Pi(x,y,\xi,d\xi') < \infty,$$

$$\int \|\tilde{G}(x,y,\xi')\|\Pi(x,y,\xi,d\xi') < \infty,$$

$$F(x,y,\xi)-f(x,y)=\tilde{F}(x,y,\xi)-(\Pi\tilde{F})(x,y,\xi),$$

$$G(x,y,\xi)-g(x,y)=\tilde{G}(x,y,\xi)-(\Pi\tilde{G})(x,y,\xi)$$

for all $x\in R^p$, $y\in R^q$, $\xi\in R^r$, where

$$(\Pi\tilde{F})(x,y,\xi)=\int \tilde{F}(x,y,\xi')\Pi(x,y,\xi,d\xi'),$$

$$(\Pi\tilde{G})(x,y,\xi)=\int \tilde{G}(x,y,\xi')\Pi(x,y,\xi,d\xi').$$

D2: For all $\rho\in[1,\infty)$, there exist Borel-measurable functions $\varphi_\rho,\psi_\rho: R^r\to[1,\infty)$ such that

$$\max\{\|F(x,y,\xi)\|,\|\tilde{F}(x,y,\xi)\|,\|(\Pi\tilde{F})(x,y,\xi)\|\}\le\varphi_\rho(\xi),$$

$$\|(\Pi\tilde{F})(x',y',\xi)-(\Pi\tilde{F})(x'',y'',\xi)\|\le\varphi_\rho(\xi)(\|x'-x''\|+\|y'-y''\|),$$

$$\max\{\|G(x,y,\xi)\|,\|\tilde{G}(x,y,\xi)\|,\|(\Pi\tilde{G})(x,y,\xi)\|\}\le\psi_\rho(\xi),$$

$$\|(\Pi\tilde{G})(x',y',\xi)-(\Pi\tilde{G})(x'',y'',\xi)\|\le\psi_\rho(\xi)(\|x'-x''\|+\|y'-y''\|)$$

for all $x,x',x''\in B^p_\rho$, $y,y',y''\in B^q_\rho$, $\xi\in R^r$.

D3: For all $\rho\in[1,\infty)$, $x\in R^p$, $y\in R^q$, $\xi\in R^r$,

$$\sum_{n=1}^{\infty}\alpha_n^2E\left(\varphi_\rho^2(\xi_n)I_{\{\lambda_\rho\ge n\}}|x_0=x,y_0=y,\xi_0=\xi\right)<\infty,$$

$$\sum_{n=1}^{\infty}\beta_n^2E\left(\psi_\rho^2(\xi_n)I_{\{\lambda_\rho\ge n\}}|x_0=x,y_0=y,\xi_0=\xi\right)<\infty,$$

where

$$\lambda_\rho=\inf(\{n\ge0:\max\{\|x_n\|,\|y_n\|\}>\rho\}\cup\{\infty\}).$$

Remark: D1 – D3 could be considered as a two time-scale extension of the assumptions adopted in [2, Chapter II.3].

The main results on the almost sure convergence of the algorithm (7), (8) under assumptions D1 – D3 are presented in the next two theorems.

Theorem 5: Let A1, D1 – D3 hold. Suppose that A3, A4 are satisfied with $f(\cdot,\cdot)$, $g(\cdot,\cdot)$ introduced in D1. Then, $\lim_{n\to\infty}d(x_n,E_*)=\lim_{n\to\infty}\|y_n-\psi(x_n)\|=0$ $w.p.1$ on the event $\{\sup_{0\le n}\|x_n\|<\infty\}\cap\{\sup_{0\le n}\|y_n\|<\infty\}$.

Theorem 6: Let B1, D1 – D3 hold. Suppose that B3, B4 are satisfied with $f(\cdot,\cdot)$, $g(\cdot,\cdot)$ introduced in D1. Then, $\lim_{n\to\infty}d(x_n,E_*)=\lim_{n\to\infty}\|y_n-\psi(x_n)\|=0$ $w.p.1$ on the event $\{\sup_{0\le n}\|x_n\|<\infty\}\cap\{\sup_{0\le n}\|y_n\|<\infty\}$.

For the proofs, see [18].

V. Actor-Critic Learning

Using the results obtained for the algorithms with non-additive state-dependent noise (Section IV), the almost sure convergence of actor-critic learning algorithms is analyzed in this section. Actor-critic algorithms are a subclass of neuro-dynamic programming (reinforcement) learning algorithms and can be considered as simulation based methods for solving large-scale Markov decision problems.

Let $p(i,k,\cdot)$, $1\le i\le N_a$, $1\le k\le N_b$, be probability distributions on $\{1,\dots,N_a\}$, while $q(x,i,\cdot)$, $1\le i\le N_a$, $x\in R^p$, are probability distributions on $\{1,\dots,N_a\}$ regular in x (i.e., $q(\cdot,i,k)$, $1\le i\le N_a$, $1\le k\le N_b$, are Borel-measurable). Controlled Markov chains with a parameterized stationary randomized policy can be defined as parameterized $\{1,\dots,N_a\}\times\{1,\dots,N_b\}$-valued Markov chains $\{a_n^x,b_n^x\}_{n\ge0}$, $x\in R^p$ (x is the parameter), satisfying the following relations:

$$P(a_{n+1}^x=j|a_0^x,b_0^x\dots,a_n^x,b_n^x)=p(a_n^x,b_n^x,j),\quad 1\le j\le N_a,$$

$$P(b_{n+1}^x=k|a_0^x,b_0^x\dots,a_n^x,b_n^x,a_{n+1}^x)=q(x,a_{n+1}^x,k),\quad 1\le k\le N_b.$$

Let $c:\{1,\dots,N_a\}\times\{1,\dots,N_b\}\to[0,\infty)$, while

$$J_n(x)=E(c(a_n^x,b_n^x)),\quad x\in R^p,\ n\ge0.$$

Average-cost Markov decision problems with a parameterized stationary randomized policy can be defined as the minimization of $\lim_{n\to\infty}J_n(x)$ (provided that $\lim_{n\to\infty}J_n(x)$ is well-defined).

Suppose that $q(\cdot,i,k)$, $1\le i\le N_a$, $x\in R^p$, are differentiable. Let $\phi(\cdot,i,k)$, $1\le i\le N_a$, $x\in R^p$, are Borel-measurable functions mapping R^p into R^q, while

$$\psi(x,i,k)=\frac{\nabla_x q(x,i,k)}{q(x,i,k)},\quad x\in R^p,\ 1\le i\le N_a,\ 1\le k\le N_b.$$

The actor-critic learning algorithms analyzed in this section are defined by the following difference equations:

$$x_{n+1}=x_n-\alpha_{n+1}\psi(x_n,a_{n+1},b_{n+1})\cdot\phi^T(x_n,a_{n+1},b_{n+1})y_n,\quad n\ge0, \tag{9}$$

$$y_{n+1}=y_n-\beta_{n+1}d_{n+1}e_{n+1},\quad n\ge0, \tag{10}$$

$$z_{n+1}=z_n+\beta_{n+1}(c(a_{n+1},b_{n+1})-z_n),\quad n\ge0, \tag{11}$$

$$d_{n+1}=c(a_n,b_n)-z_n+\phi(x_n,a_{n+1},b_{n+1})^Ty_n-\phi(x_n,a_n,b_n)^Ty_n,\quad n\ge0, \tag{12}$$

$$e_{n+1}=(e_n+\phi(x_n,a_{n+1},b_{n+1}))I_{\{i_*\}}(a_{n+1})+\phi(x_n,a_{n+1},b_{n+1})I_{\{i_*\}^c}(a_{n+1}),\quad n\ge0, \tag{13}$$

$$b_{n+1}\sim q(x_n,a_{n+1},\cdot),\quad n\ge0, \tag{14}$$

$$a_{n+1}\sim p(a_n,b_n,\cdot),\quad n\ge0. \tag{15}$$

$\{\alpha_n\}_{n\geq 1}$, $\{\beta_n\}_{n\geq 1}$ are sequences of positive reals, while i_* is defined in assumption $E1$ (below).

Let

$$\tilde{p}(x,i,j) = \sum_{k=1}^{N_a} q(x,i,k)p(i,k,j),\ x \in R^p,\ 1 \leq i,j \leq N_a,$$

while $\tilde{P}(x) = [\tilde{p}(x,i,j)]_{1\leq i,j\leq N_a}$ ($\tilde{P}(x)$ is the transition probability matrix of the Markov chain $\{a_n^x\}_{n\geq 0}$). The algorithm (9) – (15) is analyzed under the following assumptions:

E1: $\phi(\cdot,i,k)$, $\psi(\cdot,i,k)$, $1 \leq i \leq N_a$, $1 \leq k \leq N_b$, are locally Lipschitz continuous.

E2: For all $x \in R^p$, $\{a_n^x, b_n^x\}_{n\geq 0}$ and $\{a_n^x\}_{n\geq 0}$ are irreducible and aperiodic. There exist integers $N \geq 1$, $i_* \in \{1,\ldots,N_a\}$ and a constant $\varepsilon \in (0,\infty)$ such that

$$\sum_{k=1}^{N} [\tilde{P}(\tilde{x}_1)\cdots\tilde{P}(\tilde{x}_k)]_{i,i_*} \geq \varepsilon, \quad 1 \leq i \leq N_a,$$

for all $\tilde{x}_1,\ldots,\tilde{x}_N \in R^p$, where $[\cdot]_{i,j}$ denotes the i,j-th matrix entry.

E3: For all $x \in R^p$,

$$\mathrm{Span}\{[\psi_l(x,1,1)\cdots\psi_l(x,N_a,N_b)]^T : 1 \leq l \leq p\}$$
$$\subseteq \mathrm{Span}\{[\phi_l(x,1,1)\cdots\phi_l(x,N_a,N_b)]^T : 1 \leq l \leq q\},$$

where $\phi_l(x,i,k)$ and $\psi(x,i,k)$ are the l-th components of $\phi(x,i,k)$ and $\psi(x,i,k)$ (respectively).

Let $\pi(x,\cdot)$ be the invariant probability distribution of the Markov chain $\{a_n^x\}_{n\geq 0}$, while

$$J(x) = \sum_{i=1}^{N_a}\sum_{k=1}^{N_b} \pi(x,i)q(x,i,k)c(i,k), \quad x \in R^p.$$

Then, E1 and E2 imply that $J(\cdot)$ is differentiable, $\nabla J(\cdot)$ is locally Lipschitz continuous and $J(x) = \lim_{n\to\infty} J_n(x)$ for all $x \in R^p$. Let $E_* = \{x \in R^p : \nabla J(x) = 0\}$. If $p = N_aN_b$, $x = [x_{11}\cdots x_{N_aN_b}]^T$ and

$$q(x,i,k) = \frac{\exp(x_{ik})}{\sum_{l=1}^{N_b}\exp(x_{il})}, \quad 1 \leq i \leq N_a,\ 1 \leq k \leq N_b,$$

then $J(\cdot)$ is differentiable infinitely many times and the Lebesgue measure of $J(E_*)$ is zero (for more details see the comments on A3, A4 in Section II).

The main results on the almost sure convergence of the algorithm (9) – (15) are presented in the next theorem.

Theorem 7: Let A1 and E1 – E3 hold. Suppose that $q(\cdot,i,k)$, $1 \leq i \leq N_a$, $1 \leq k \leq N_b$, are p times differentiable. Then, $\lim_{n\to\infty} d(x_n,E_*) = 0$ $w.p.1$ on $\{\sup_{0\leq n}\|x_n\| < \infty\} \cap \{\sup_{0\leq n}\|y_n\| < \infty\}$.

Theorem 8: Let B1 and E1 – E3 hold. Then, $\lim_{n\to\infty} d(x_n,E_*) = 0$ $w.p.1$ on $\{\sup_{0\leq n}\|x_n\| < \infty\} \cap \{\sup_{0\leq n}\|y_n\| < \infty\}$.

For the proofs, see [18].

The actor-critic learning algorithms (9) – (15) have been proposed and analyzed in [10], [11]. However, only $\underline{\lim}_{n\to\infty} d(x_n,E_*) = 0$ $w.p.1$ has been demonstrated in [10], [11].

REFERENCES

[1] J. S. Baras and V. S. Borkar, *A learning algorithm for Markov decision processes with adaptive state aggregation*, Proceedings of the 39th IEEE Conference on Decision and Control, 2000.

[2] A. Benveniste, M. Metivier, and P. Priouret, *Adaptive Algorithms and Stochastic Approximation*, Springer Verlag, 1990.

[3] D. P. Bertsekas and J. N. Tsitsiklis, *Neuro-Dynamic Programming*, Athena Scientific, 1996.

[4] V. S. Borkar, "Stochastic approximation with two time scales," *Systems and Control Letters*, vol. 29, pp. 291–294, 1997.

[5] S. Bhatnagar, M. C. Fu, and S. I. Marcus, "Optimal multilevel feedback policies for ABR flow control using two timescale SPSA," *Technical Report TR 99-18*, Institute of systems Research, University of Maryland, 1999.

[6] S. Bhatnagar, M. C. Fu, S. I. Marcus, and S. Bhatnagar, "Randomized difference two-timescale simultaneous perturbation stochastic approximation algorithms for simulation optimization of hidden Markov models," *Technical Report TR 2000-13*, Institute of systems Research, University of Maryland, 2000.

[7] H.-F. Chen and Y.-M. Zhu, "Stochastic approximation with randomly varying truncations," *Scientia Sinica, Ser. A*, vol. 29, 914–926, 1986.

[8] P. Doukhan, *Mixing: Properties and Examples*, Springer Verlag, 1996.

[9] V. R. Konda and V. S. Borkar, "Actor-critic like learning algorithms for Markov decision processes," *SIAM Journal on control and Optimization*, vol. 38, pp. 94–123, 1999.

[10] V. R. Konda, *Actor-Critic Algorithms*, PhD Thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, 2002.

[11] V. R. Konda and J. N. Tsitsiklis, "On actor-critic algorithms," *SIAM Journal on Control and Optimization*, to appear.

[12] H. J. Kushner and G. G. Yin, *Stochastic Approximation Algorithms and Applications*, Springer Verlag, 1997.

[13] J. W. Milnor, *Topology from the Differentiable Point of View*, University of Virginia Press, 1965.

[14] R. E. O'Malley Jr, *Singular Perturbations Methods for Ordinary Differential Equations*, Springer Verlag, 1991.

[15] V. B. Tadić, *Convergence of Stochastic Approximation under General Noise and Stability Conditions*, IEEE Conference on Decision and Control, 1997.

[16] V. B. Tadić, *Asymptotic Analysis of Stochastic Approximation Algorithms under Violated Kushner-Clark Conditions with Applications*, IEEE Conference on Decision and Control, 2000.

[17] V. B. Tadić and S. P. Meyn, "Asymptotic properties of two time-scale stochastic approximation algorithms with constant step sizes," American Control Conference, 2003.

[18] V. B. Tadić, *Asymptotic Analysis of Two Time-Scale Stochastic Approximation Algorithms*, submitted.

[19] I.-J. Wang, E. K. P. Chong, and S. R. Kulkarni, "Equivalent and sufficient conditions on noise sequences for stochastic approximation algorithms," *Advances in Applied Probability*, vol. 28, pp. 784–801, 1996.