Machine Learning, 42, 241–267, 2001

On the Convergence of Temporal-Difference Learning with Linear Function Approximation

VLADISLAV TADIĆ v.tadic@ee.mu.oz.au
Department of Electrical and Electronic Engineering, The University of Melbourne, Parkville, Victoria 3010, Australia

Editor: Michael Jordan

Abstract. The asymptotic properties of temporal-difference learning algorithms with linear function approximation are analyzed in this paper. The analysis is carried out in the context of the approximation of a discounted cost-to-go function associated with an uncontrolled Markov chain with an uncountable finite-dimensional state-space. Under mild conditions, the almost sure convergence of temporal-difference learning algorithms with linear function approximation is established and an upper bound for their asymptotic approximation error is determined. The obtained results are a generalization and extension of the existing results related to the asymptotic behavior of temporal-difference learning. Moreover, they cover cases to which the existing results cannot be applied, while the adopted assumptions seem to be the weakest possible under which the almost sure convergence of temporal-difference learning algorithms is still possible to be demonstrated.

Keywords: temporal-difference learning, reinforcement learning, neuro-dynamic programming, almost sure convergence, Markov chains, positive Harris recurrence

1. Introduction

The asymptotic properties of temporal-difference learning algorithms with linear function approximation are considered in this paper. Temporal-difference learning with function approximation is a recursive parametric method for approximating a cost-to-go function associated with a Markov chain. Algorithms of this type aim at determining the optimal value of the approximator parameter by using only the available observations of the underlying chain. Basically, they update the approximator parameter whenever a new observation of the underlying chain is available trying to minimize the approximation error. Temporal-difference learning with function approximation represents an extension of the classical temporal-difference learning algorithms (see e.g., Sutton, 1988) and has extensively been analyzed in Tsitsiklis and Van Roy (1997). As opposed to temporal-difference learning with function approximation, the classical temporal-difference learning algorithms are only capable of predicting the value of the cost-to-go function.

The problems of the prediction and approximation of a cost-to-go function associated with a stochastic system modelled as a Markov chain appear in the areas such as automatic control and time-series analysis. Among several methods proposed for solving these problems (e.g., Monte Carlo methods in statistics and maximum likelihood methods in automatic control; see e.g., Kumar and Varaiya, 1986), temporal-difference learning is probably the most general. Moreover, it is efficient and simple to be implemented. Due to their excellent

performances, temporal-difference learning algorithms have found a wide range of application (for details see e.g., Bertsekas & Tsitsiklis, 1996; Sutton & Barto, 1998 and references cited therein), while their asymptotic properties (almost sure convergence, convergence in mean and probability, convergence of mean and rate of convergence) have been analyzed in a great number of papers (see Dayan, 1992; Dayan & Sejnowski, 1994; Jaakola, Jordan, & Singh, 1994; Sutton, 1988; Tsitsiklis & Van Roy, 1997; see also Bertsekas & Tsitsiklis, 1996; Sutton & Barto, 1998 and references cited therein). Although the existing results provide a good insight into the asymptotic behavior of temporal-difference learning algorithms, they practically cover only the case where the chain is geometrically ergodic and are constrained to the case where the state-space of the underlying chain is countable. However, this is too restrictive for applications such as the prediction and approximation of a cost-to-go function associated with Markov chains appearing in the areas of queuing theory and time-series analysis.

In this paper, the almost sure convergence and asymptotic approximation error of temporal-difference learning algorithms with linear function approximation are analyzed. The analysis is carried out in the context of the approximation of a discounted cost-to-go function associated with an uncontrolled Markov chain with an uncountable finite-dimensional state-space. The results of this paper are a generalization and extension of those presented in Tsitsiklis and Van Roy (1997). Moreover, they cover cases to which the previous results on temporal-difference learning cannot be applied, while the adopted assumptions seem to be the weakest possible under which the almost sure convergence can be demonstrated.

The paper is organized as follows. In Section 2, temporal-difference learning algorithms with linear function approximation are formally defined and the assumptions under which their analysis is carried out are introduced. A detailed comparison of the adopted assumptions with those of Tsitsiklis and Van Roy (1997) (as well as with the assumptions used in other papers related to the asymptotic behavior of temporal-difference learning algorithms) is also given in Section 2. The main results are presented in Section 3, where the almost sure convergence of temporal-difference learning algorithms with linear function approximation is established and an upper bound for their asymptotic approximation error is determined. In Section 4, these results are illustrated by an example related to the queueing theory and not covered by the previous results on temporal-difference learning. Subsidiary results which are of crucial importance for obtaining the main ones are provided in Appendix.

2. Algorithm and assumptions

Temporal-difference learning algorithms with linear function approximation are defined by the following difference equations:

$$\theta_{n+1} = \theta_n + \gamma_{n+1}\delta_{n+1}\varepsilon_{n+1}, \quad n \geq 0, \tag{1}$$

$$\delta_{n+1} = g(X_n, X_{n+1}) + \alpha\theta_n^T\phi(X_{n+1}) - \theta_n^T\phi(X_n), \quad n \geq 0, \tag{2}$$

$$\varepsilon_{n+1} = \sum_{i=0}^{n}(\alpha\lambda)^{n-i}\phi(X_i), \quad n \geq 0. \tag{3}$$

$\{\gamma_n\}_{n\geq 1}$ is a sequence of positive reals, while $\alpha \in (0, 1)$, $\lambda \in [0, 1]$ are constants. $\phi : R^{d'} \to R^d$ and $g : R^{d'} \times R^{d'} \to R$ are Borel-measurable functions. θ_0 is an R^d-valued random variable defined on a probability space $(\Omega, \mathcal{F}, \mathcal{P})$, while $\{X_n\}_{n\geq 0}$ is an $R^{d'}$-valued homogeneous Markov chain defined on the same probability space. In the case of temporal-difference learning algorithms with general, non-linear function approximation, the Eqs. (2) and (3) are replaced by the following ones:

$$\begin{aligned} \delta_{n+1} &= g(X_n, X_{n+1}) + \alpha f(\theta_n, X_{n+1}) - f(\theta_n, X_n), \quad n \geq 0, \\ \varepsilon_{n+1} &= \sum_{i=0}^{n} (\alpha\lambda)^{n-i} \nabla_\theta f(\theta_i, X_i), \quad n \geq 0, \end{aligned} \tag{4}$$

where $f : R^d \times R^{d'} \to R$ is a Borel-measurable function which is differentiable in the first argument. In order to get a practically implementable algorithm, (4) should be rewritten in the following way:

$$\varepsilon_{n+1} = \alpha\lambda\varepsilon_n + \nabla_\theta f(\theta_n, X_n), \quad n \geq 0,$$

where $\varepsilon_0 = 0$. Let

$$f_*(x) = E\left(\sum_{n=0}^{\infty} \alpha^n g(X_n, X_{n+1}) \,\middle|\, X_0 = x \right), \quad x \in R^{d'}$$

(provided that $f_*(\cdot)$ is well-defined and finite). In the context of dynamic programming, $f_*(\cdot)$ is interpreted as a discounted cost-to-go function associated with the Markov chain $\{X_n\}_{n\geq 0}$ (for details see e.g., Bertsekas, 1976). The task of the algorithm (1)–(3) is to approximate the function $f_*(\cdot)$. It determines the optimal value θ_* of the approximator parameter $\theta \in R^d$ such that $\theta_*^T \phi(\cdot)$ (i.e., $f(\theta_*, \cdot)$ in the case of non-linear approximation) is the best approximator of $f_*(\cdot)$ among $\{\theta^T \phi(\cdot)\}_{\theta \in R^d}$ (i.e., among $\{f(\theta, \cdot)\}_{\theta \in R^d}$ in the case of non-linear approximation). If $\lambda = 1$ and if $\{X_n\}_{n\geq 0}$ has a unique invariant measure $\mu(\cdot)$, the algorithm (1)–(3) determines $\theta_* \in R^d$ such that $\theta_*^T \phi(\cdot)$ approximates $f_*(\cdot)$ optimally in the $L^2(\mu)$-sense, i.e., it looks for the minimum of the function $J_*(\theta) = \int (\theta^T \phi(x) - f_*(x))^2 \mu(dx)$, $\theta \in R^d$ (see Theorems 1 and 2).

Throughout the paper, the following notation is used. R^+ and R_0^+ are the sets of positive and non-negative reals (respectively). $\|\cdot\|$ denotes both the Euclidean vector norm and the matrix norm induced by the Euclidean vector norm (i.e., $\|A\| = \sup_{\|\theta\|=1} \|A\theta\|$ for $A \in R^{d\times d}$). $P(x, \cdot)$, $x \in R^{d'}$, is the transition probability of $\{X_n\}_{n\geq 0}$ (i.e., $\mathcal{P}(X_{n+1} \in B \mid X_n) = P(X_n, B)$ w.p.1, $\forall B \in \mathcal{B}^{d'}$, $n \geq 0$), while

$$\tilde{g}(x) = \int g(x, x') P(x, dx'), \quad x \in R^{d'}$$

(provided that $\tilde{g}(\cdot)$ is well-defined and finite). For $x \in R^{d'}$ and $B \in \mathcal{B}^{d'}$, let $P_0(x, B) = I_B(x)$ ($I_B(\cdot)$ stands for the indicator function of B) and

$$P_{n+1}(x, B) = \int P(x', B) P_n(x, dx'), \quad n \geq 0.$$

For $t \in R^+$, let $\eta(n, t) = \sup\{j \geq n : \sum_{i=n}^{j-1} \gamma_{i+1} \leq t\}$, $n \geq 0$.

In this paper, the algorithm (1)–(3) is analyzed under the following assumptions:

A1. $\{n\gamma_n\}_{n\geq 1}$ converges and $0 < \lim_{n\to\infty} n\gamma_n < \infty$.
A2. $\{X_n\}_{n\geq 0}$ has a unique invariant probability measure $\mu(\cdot)$.
A3. There exists a Borel-measurable function $\psi : R^{d'} \to R_0^+$ such that

$$\|\phi(x)\| \leq \psi(x), \quad \forall x \in R^{d'}, \tag{5}$$

$$\int g^2(x, x') P(x, dx') \leq \psi^2(x), \quad \forall x \in R^{d'},$$

$$\sum_{n=0}^{\infty} \alpha^n (P_n \psi^2)(x) < \infty, \quad \forall x \in R^{d'}, \tag{6}$$

$$\int \psi^2(x) \mu(dx) < \infty, \tag{7}$$

where $(P_n\psi^2)(x) = \int \psi^2(x') P_n(x, dx')$.

A4.

$$\overline{\lim_{n\to\infty}} n^{-1} \sum_{i=0}^{n-1} g^2(X_i, X_{i+1}) < \infty \quad \text{w.p.1}, \tag{8}$$

$$\lim_{n\to\infty} n^{-1} \sum_{i=0}^{n-1} \phi(X_i) g(X_{i+j}, X_{i+j+1}) = \int \phi(x)(P_j \tilde{g})(x) \mu(dx) \quad \text{w.p.1}, \ j \geq 0, \tag{9}$$

$$\lim_{n\to\infty} n^{-1} \sum_{i=0}^{n-1} \phi(X_i) \phi^T(X_{i+j}) = \int \phi(x)(P_j \phi^T)(x) \mu(dx) \quad \text{w.p.1}, \ j \geq 0, \tag{10}$$

where $(P_j\tilde{g})(x) = \int \tilde{g}(x') P_j(x, dx')$ and $(P_j\phi^T)(x) = \int \phi^T(x') P_j(x, dx')$.
A5. $\int \phi(x)\phi^T(x)\mu(dx)$ is positive definite.

Remark. Due to the Jensen inequality and A3,

$$\int |g(x, x')| P(x, dx') \leq \psi(x), \quad \forall x \in R^{d'},$$

$$\int\int g^2(x, x') P(x, dx') \mu(dx) \leq \int \psi^2(x) \mu(dx) < \infty, \tag{11}$$

Consequently, $\tilde{g}(\cdot)$ is well-defined and finite, as well as

$$|\tilde{g}(x)| \leq \psi(x), \quad \forall x \in R^{d'}. \tag{12}$$

Then, A3, implies

$$\begin{aligned}
\int \|\phi(x)(P_n\tilde{g})(x)\|\mu(dx) &\leq \int \psi(x)(P_n\psi)(x)\mu(dx) \\
&\leq \left(\int \psi^2(x)\mu(dx)\right)^{1/2}\left(\int (P_n\psi^2)(x)\mu(dx)\right)^{1/2} \\
&= \int \psi^2(x)\mu(dx) < \infty, \quad n \geq 0, \qquad (13) \\
\int \|\phi(x)(P_n\phi^T)(x)\|\mu(dx) &\leq \int \psi(x)(P_n\psi)(x)\mu(dx) \\
&\leq \left(\int \psi^2(x)\mu(dx)\right)^{1/2}\left(\int (P_n\psi^2)(x)\mu(dx)\right)^{1/2} \\
&= \int \psi^2(x)\mu(dx) < \infty, \quad n \geq 0. \qquad (14)
\end{aligned}$$

Therefore, $\int \phi(x)\phi^T(x)\mu(dx)$ and the right-hand sides of (9) and (10) are well-defined and finite.

Assumption A1 is satisfied if $\gamma_n = n^{-1}$, $n \geq 1$, which is a typical choice for the stepsize of stochastic approximation algorithms (see e.g., Ljung, Pflug, & Walk, 1992). It implies that $\eta(n,t)$, $n \geq 0$, are well-defined and finite for all $t \in R^+$, as well as that

$$\sum_{i=n}^{\eta(n,t)-1} \gamma_{i+1} \leq t < \sum_{i=n}^{\eta(n,t)} \gamma_{i+1}; \quad \forall t \in R^+,\ n \geq 0, \tag{15}$$

$$\lim_{n\to\infty} \sum_{i=n}^{\eta(n,t)} \gamma_{i+1} = t, \quad \forall t \in R^+. \tag{16}$$

Assumption A2 requires $\{X_n\}_{n\geq 0}$ to exhibit an asymptotic stationarity. It is satisfied if $\{X_n\}_{n\geq 0}$ is positive Harris (see e.g., Meyn & Tweedie, 1993, Chapter 10). Assumptions of this type are standard for the analyses of temporal-difference learning algorithms, as well as in the analyses of stochastic approximation algorithms operating in a Markovian environment (see Benvensite, Metivier, & Priouret, 1990, Part II; Bertsekas & Tsitsiklis, 1996 and references cited therein).

Assumption A3 corresponds to the growth rate of $g(\cdot,\cdot)$ and $\phi(\cdot)$. It requires them not to grow too fast so that their upper bound $\psi(\cdot)$ satisfies (6) and (7). The role of A3 is to ensure that $f_*(\cdot)$, A and b (introduced in (20) and (21)) are well-defined and finite. Assumption A3 is satisfied if $g(\cdot,\cdot)$ and $\phi(\cdot)$ are globally bounded or if $g(\cdot,\cdot)$ and $\phi(\cdot)$ are locally bounded and there exists a constant $K \in R^+$ such that $\|X_n\| \leq K$ w.p.1, $n \geq 0$. It is important to notice that A3 allows $\{(P_n\psi^2)(x)\}_{n\geq 0}$ to grow exponentially as $n \to \infty$ for any $x \in R^{d'}$.

Assumption A4 requires $\{X_n\}_{n\geq 0}$ to exhibit certain "degree of stability". The role of A4 is to provide that $\{A_n\}_{n\geq 1}$ and $\{b_n\}_{n\geq 1}$ (defined in (18) and (19)) converge to A and b, respectively. As A3 implies (11), (13) and (14), it can easily be deduced from Lemma 9 (given in Appendix) that (9), (10) and

$$\lim_{n\to\infty} n^{-1}\sum_{i=0}^{n-1} g^2(X_i, X_{i+1}) = \int\int g^2(x, x')P(x, dx')\mu(dx) < \infty \quad \text{w.p.1}$$

are satisfied if A3 holds and if $\{X_n\}_{n\geq 0}$ is positive Harris (for the definition and further details see e.g., Meyn and Tweedie, 1993). It should be emphasized that A4 represents one of the weakest sample-path properties related to the stability of $\{X_n\}_{n\geq 0}$.

Assumption A5 is a "persistency of excitation" condition. These conditions are typical for the areas of system identification, adaptive control and adaptive signal processing (see e.g., Chen & Guo, 1991; Solo & Kong, 1995). Assumption A5 requires $\{\phi(X_n)\}_{n\geq 0}$ to be sufficiently rich with respect to all directions in R^d at the asymptotic steady-state characterized by the invariant measure $\mu(\cdot)$, i.e., it demands that $\mu(x : \theta^T\phi(x) \neq 0) = 1$, $\forall\theta \in R^d$. If $\{X_n\}_{n\geq 0}$ has a finite state-space $\{x_1, \ldots, x_m\}$, A5 is implied by the requirement that $\mu(x = x_i) > 0$, $1 \leq i \leq m$, and that $[\phi(x_1)\cdots\phi(x_m)]$ is a full row-rank matrix. Without A5, only the almost sure convergence of $\{\Pi\theta_n\}_{n\geq 0}$ could be demonstrated, where Π is the projection operator onto the space spanned by the rows of $\int \phi(x)\phi^T(x)\mu(dx)$.

The asymptotic properties of temporal-difference learning algorithms have been considered in several papers (Dayan 1992; Dayan & Sejnowski, 1994; Jaakola, Jordan, & Singh, 1994; Sutton, 1988; Tsitsiklis & Van Roy, 1997; see also Bertsekas & Tsitsiklis, 1996; Sutton & Barto, 1998 and references cited therein). Among these papers, Tsitsiklis and Van Roy (1997) contains probably the strongest results. The results of this paper are a generalization and extension of those presented in Tsitsiklis and Van Roy (1997). Due to the fact that $\{X_n\}_{n\geq 0}$ is positive Harris if it is irreducible, aperiodic and positive and if its state-space is countable, Lemma 9 (given in Appendix) directly implies that A2–A5 are just a special case of Assumptions 1–3 adopted in Tsitsiklis and Van Roy (1997). It is particularly important to emphasize that Assumption 4 of Tsitsiklis and Van Roy (1997) is not necessary that A2–A5 hold. In other words, using the results of this paper, the convergence of the algorithm (1)–(3) can be shown under only Assumptions 1–3 of Tsitsiklis and Van Roy (1997) (note that this is possibly only if the state-space of $\{X_n\}_{n\geq 0}$ is countable; otherwise, irreducibility, aperiodicity and positiveness are not sufficient for the positive Harris recurrence). On the other hand, Assumption 4 is the most restrictive among the assumptions adopted in Tsitsiklis and Van Roy (1997). It practically covers only the case where $\{X_n\}_{n\geq 0}$ is geometrically ergodic and implies that there exist a constant $C \in R^+$ and a Borel-measurable function $\psi : R^{d'} \to R_0^+$ such that (5) and the following relation hold:

$$(P_n\psi^p)(x) \leq C\psi^p(x); \quad \forall x \in R^{d'}, \quad \forall p \in [1, \infty), \quad n \geq 0. \tag{17}$$

However, this is too restrictive for applications such as the prediction and approximation of a cost-to-go function associated with Markov chains appearing in the areas of queueing theory and time-series analysis (note that (17) implies $\|\!| P_n(x, \cdot) |\!\|_{\psi^p} \leq C, \forall p \in [1, \infty), n \geq 0$, and

that typically $|||P_n(x,\cdot)|||_{\psi^p} \to \infty$ as $n \to \infty$ in the cases where $\{X_n\}_{n\geq 0}$ is not geometrically ergodic; for the definition of $|||\cdot|||_{\psi^p}$ see e.g., Meyn and Tweedie, 1993). As opposed to Tsitsiklis and Van Roy (1997), the assumptions of this paper require $\{X_n\}_{n\geq 0}$ only to satisfy certain laws of large numbers, allow $\{(P_n\psi^2)(x)\}_{n\geq 0}$ to grow exponentially as $n \to \infty$ for any $x \in R^{d'}$ and cover the case where $\{X_n\}_{n\geq 0}$ is positive Harris (note that laws of large numbers are probably the weakest sample-path properties related to the stability of Markov chains; also note that $\{X_n\}_{n\geq 0}$ is geometrically ergodic if and only if $\{X_n\}_{n\geq 0}$ is positive Harris and if there exists a constant $\rho \in (1,\infty)$ such that $\sum_{n=0}^{\infty} \rho^n \|(P_n - \mu)(x,\cdot)\| < \infty$, $\forall x \in R^{d'}$, where $\|\cdot\|$ denotes the total variation of a signed measure; for further details see e.g., Meyn and Tweedie, 1993). Due to this, A1–A5 cover a broader class of Markov chains of practical interest than the previous results on temporal-difference learning. The area of queueing theory is particularly rich in the examples of Markov chains satisfying A1–A5 and not being covered by the assumptions under which the previous results on temporal-difference learning have been obtained. Such an example related to the waiting times of GI/G/1 queue is provided in Section 4, while Dai (1995) gives directions how A1–A5 can be verified in the context of queueing networks.

Besides the fact that A1–A5 are more general than those of Tsitsiklis and Van Roy (1997) and include cases not covered by the results presented therein, they seem to be the weakest conditions under which the almost sure convergence of the algorithm (1)–(3) is still possible to be shown. The rationale for this comes out from the fact that stochastic approximation algorithms in general converge if and only if their noise satisfies a law of large numbers (see Clark, 1984; Kulkarni & Horn, 1996; Wang, Chong, & Kulkarni, 1996. See also the note at the end of the paper) and from the fact that (8)–(10) themselves express laws of large numbers for functionals of $\{X_n\}_{n\geq 0}$. It is also important to emphasize the methodological differences between the analyses carried out in Tsitsiklis and Van Roy (1997) and here. The results presented in Tsitsiklis and Van Roy (1997) (as well as in other papers related to the convergence of temporal-difference learning) have been obtained by using the general approach to the asymptotic analysis of stochastic approximation algorithms based on martingale convergence arguments and the Poisson equation (for details see Benvensite, Metivier, & Priouret, 1990, Part II). However, A1–A5 do not guarantee that there exist unique Borel-measurable functions $U : R^{d+2d'} \to R^{d\times d}$ and $v : R^{d+2d'} \to R^d$ satisfying the following Poisson equations:

$$U(z) - \int U(z')\Pi(z,dz') = A(z) - A, \quad \forall z \in R^{d+2d'},$$
$$v(z) - \int v(z')\Pi(z,dz') = b(z) - b, \quad \forall z \in R^{d+2d'},$$

where

$$\Pi(z,B) = \int I_B(x',x'',\alpha\lambda y + \phi(x'))P(x',dx''),$$
$$A(z) = y(\alpha\phi(x') - \phi(x)),$$
$$b(z) = yg(x,x'),$$

for $B \in \mathcal{B}^{d+2d'}$, $x, x' \in R^{d'}$, $y \in R^d$ and $z = (x, x', y)$ ($I_B(\cdot)$ denotes the indicator function of B), while A and b are defined in (20) and (21) (given in Section 3). As this is of a crucial importance for the analysis carried out in Tsitsiklis and Van Roy (1997), the approach used therein is completely inapplicable to the asymptotic analysis of temporal-difference learning algorithms under the assumptions A1–A5. Instead, the algorithm (1)–(3) is analyzed in this paper by using the approach which is based on the ideas standing behind the results presented in Tadić (1997) and which is closer to the ODE methodology (see e.g., Kushner & Clark, 1978).

3. Convergence analysis

The main results are presented in this section. These results are contained in Theorems 1 and 2. In Theorem 1, the almost sure convergence of the algorithm (1)–(3) is demonstrated. In Theorem 2, an interpretation of the algorithm limit is provided and an upper bound for the asymptotic approximation error is determined in terms of α, λ and the error of the $L^2(\mu)$-optimal linear approximation of $f_*(\cdot)$. Lemmas 1–5, as well as Lemmas 7–9 (given in Appendix) are prerequisites for Theorems 1 and 2.

Throughout this section, the following notation is used. Let

$$A_{n+1} = \sum_{i=0}^{n} (\alpha\lambda)^{n-i} \phi(X_i)(\alpha\phi(X_{n+1}) - \phi(X_n))^T, \quad n \geq 0, \tag{18}$$

$$b_{n+1} = \sum_{i=0}^{n} (\alpha\lambda)^{n-i} \phi(X_i) g(X_n, X_{n+1}), \quad n \geq 0, \tag{19}$$

$$A = -\int \phi(x)\phi^T(x)\mu(dx) + \alpha(1-\lambda)\sum_{n=0}^{\infty}(\alpha\lambda)^n \int \phi(x)(P_{n+1}\phi^T)(x)\mu(dx), \tag{20}$$

$$b = \sum_{n=0}^{\infty}(\alpha\lambda)^n \int \phi(x)(P_n\tilde{g})(x)\mu(dx), \tag{21}$$

while $\theta_* = -A^{-1}b$ (provided that A, b and θ_* are well-defined and finite). Then, the algorithm (1)–(3) can be rewritten as follows:

$$\theta_{n+1} = \theta_n + \gamma_{n+1}(A_{n+1}\theta_n + b_{n+1}), \quad n \geq 0. \tag{22}$$

Let $\vartheta_n = \theta_n - \theta_*$, $U_{nn} = I$ and $V_{nn} = 0$, $n \geq 0$, while

$$\begin{aligned}
U_{nj} &= (I + \gamma_j A_j)\cdots(I + \gamma_{n+1}A_{n+1}), \quad 0 \leq n < j, \\
V_{nj} &= \sum_{i=n+1}^{j} U_{ij}\gamma_i(A_i\theta_* + b_i), \quad 0 \leq n < j, \\
e_n(t) &= 2\vartheta_n^T\big(U_{n,\eta(n,t)} - I - tA\big)\vartheta_n + \vartheta_n^T U_{n,\eta(n,t)}V_{n,\eta(n,t)} \\
&\quad + \big\|(U_{n,\eta(n,t)} - I)\vartheta_n\big\|^2; \quad t \in R^+, \quad n \geq 0
\end{aligned}$$

(I denotes the $d \times d$ unit matrix). Then, it is straightforward to verify that

$$\vartheta_j = U_{nj}\vartheta_n + V_{nj}, \quad 0 \le n \le j, \tag{23}$$

$$\left\|\vartheta_{\eta(n,t)}\right\|^2 = \|\vartheta_n\|^2 + 2t\vartheta_n^T A\vartheta_n + e_n(t); \quad \forall t \in R^+, \ \ n \ge 0. \tag{24}$$

In the next lemma, it is shown that A, b and θ_* are well-defined and finite. The proof is based on similar ideas as the corresponding result of Tsitsiklis and Van Roy (1997).

Lemma 1. *Let A*2, *A*3 *and A*5 *hold. Then, $f_*(\cdot)$, A, b and θ_* are well-defined and finite. Moreover, A is negative definite and*

$$f_*(x) = \sum_{n=0}^{\infty} \alpha^n (P_n\tilde{g})(x), \quad \forall x \in R^{d'}. \tag{25}$$

Proof: Due to (12)–(14),

$$\begin{aligned}
&|(P_n\tilde{g})(x)| \le (P_n\psi)(x) \le 1 + (P_n\psi^2)(x); \quad \forall x \in R^{d'}, \ \ n \ge 0,\\
&E(|g(X_n, X_{n+1})| \mid X_0 = x) = \int\int |g(x', x'')| P(x', dx'') P_n(x, dx')\\
&\quad \le (P_n\psi)(x) \le 1 + (P_n\psi^2)(x) < \infty; \quad \forall x \in R^{d'}, \ \ n \ge 0,\\
&\sum_{n=0}^{\infty} (\alpha\lambda)^n \int \|\phi(x)(P_{n+1}\phi^T)(x)\| \mu(dx) \le (1-\alpha\lambda)^{-1} \int \psi^2(x)\mu(dx) < \infty,\\
&\sum_{n=0}^{\infty} (\alpha\lambda)^n \int \|\phi(x)(P_n\tilde{g})(x)\| \mu(dx) \le (1-\alpha\lambda)^{-1} \int \psi^2(x)\mu(dx) < \infty.
\end{aligned}$$

Consequently,

$$\begin{aligned}
E(g(X_n, X_{n+1}) \mid X_0 = x) &= \int\int g(x', x'') P(x', dx'') P_n(x, dx')\\
&= (P_n\tilde{g})(x); \quad \forall x \in R^{d'}, \ \ n \ge 0,\\
\sum_{n=0}^{\infty} \alpha^n |(P_n\tilde{g})(x)| &\le (1-\alpha)^{-1} + \sum_{n=0}^{\infty} \alpha^n (P_n\psi^2)(x) < \infty, \quad \forall x \in R^{d'}.
\end{aligned}$$

Then, it is obvious that $f_*(\cdot)$, A and b are well-defined and finite, as well as that (25) holds. On the other hand, owing to the Jensen inequality,

$$\begin{aligned}
\int (\theta^T (P_n\phi)(x))^2 \mu(dx) &\le \int\int (\theta^T \phi(x'))^2 P_n(x, dx')\mu(dx)\\
&= \int (\theta^T \phi(x))^2 \mu(dx); \quad \forall \theta \in R^d, \ \ n \ge 0.
\end{aligned}$$

Therefore,

$$\left|\int \theta^T\phi(x)(P_n\phi^T)(x)\theta\mu(dx)\right| \le \left(\int(\theta^T\phi(x))^2\mu(dx)\right)^{1/2} \times\left(\int(\theta^T(P_n\phi)(x))^2\mu(dx)\right)^{1/2} \le \int(\theta^T\phi(x))^2\mu(dx);\quad \forall\theta\in R^d,\ \ n\ge 0.$$

Consequently,

$$\begin{aligned}\theta^T A\theta &= -\int(\theta^T\phi(x))^2\mu(dx)+\alpha(1-\lambda)\sum_{n=0}^{\infty}(\alpha\lambda)^n\int\theta^T\phi(x)(P_{n+1}\phi^T)(x)\theta\mu(dx)\\ &\le -\left(1-\alpha(1-\lambda)\sum_{n=0}^{\infty}(\alpha\lambda)^n\right)\int(\theta^T\phi(x))^2\mu(dx)\\ &= -(1-\alpha)(1-\alpha\lambda)^{-1}\theta^T\left(\int\phi(x)\phi^T(x)\mu(dx)\right)\theta,\quad \forall\theta\in R^d.\end{aligned}$$

Then, it is obvious that A is negative definite, as well as that θ_* is well-defined and finite. □

In the next lemma, it is demonstrated that $\{A_n\}_{n\ge1}$ and $\{b_n\}_{n\ge1}$ satisfy the law of large numbers. The proof is essentially based on Lemmas 8 and 9 (given in Appendix). Among the prerequisites of Theorem 1, the results presented in Lemma 2 are probably the most important.

Lemma 2. *Let A*1–*A*4 *hold. Then,*

$$\lim_{n\to\infty} n^{-1}\sum_{i=1}^{n} A_i = A \quad \text{w.p.1}, \tag{26}$$

$$\lim_{n\to\infty} n^{-1}\sum_{i=1}^{n} b_i = b \quad \text{w.p.1}. \tag{27}$$

Moreover, there exist non-negative random variables K' and K'' defined on $(\Omega,\mathcal{F},\mathcal{P})$ such that

$$\lim_{n\to\infty} n^{-1}\sum_{i=0}^{n-1}\sum_{j=0}^{i}(\alpha\lambda)^{i-j}\|\phi(X_i)\|\|\phi(X_j)\| = K' \quad \text{w.p.1}, \tag{28}$$

$$\lim_{n\to\infty} n^{-1}\sum_{i=0}^{n-1}\sum_{j=0}^{i+1}(\alpha\lambda)^{i-j}\|\phi(X_{i+1})\|\|\phi(X_j)\| = K'' \quad \text{w.p.1}. \tag{29}$$

Proof: Due to A3, A4 and Lemma 8,

$$\lim_{n\to\infty} n^{-1}\sum_{i=0}^{n-1}\|\phi(X_i)\|^2$$
$$= \int \|\phi(x)\|^2\mu(dx) \le \int \psi^2(x)\mu(dx) < \infty \quad \text{w.p.1}, \tag{30}$$
$$\lim_{n\to\infty} n^{-1}\sum_{i=0}^{n-1}\sum_{j=0}^{i}(\alpha\lambda)^{i-j}\phi(X_j)g(X_i, X_{i+1})$$
$$= \sum_{n=0}^{\infty}(\alpha\lambda)^n \int \phi(x)(P_n\tilde{g})(x)\mu(dx) \quad \text{w.p.1},$$
$$\lim_{n\to\infty} n^{-1}\sum_{i=0}^{n-1}\sum_{j=0}^{i}(\alpha\lambda)^{i-j}\phi(X_j)\phi^T(X_i)$$
$$= \sum_{n=0}^{\infty}(\alpha\lambda)^n \int \phi(x)(P_n\phi^T)(x)\mu(dx) \quad \text{w.p.1},$$
$$\lim_{n\to\infty} n^{-1}\sum_{i=0}^{n-1}\sum_{j=0}^{i}(\alpha\lambda)^{i-j}\phi(X_j)\phi^T(X_{i+1})$$
$$= \sum_{n=0}^{\infty}(\alpha\lambda)^n \int \phi(x)(P_{n+1}\phi^T)(x)\mu(dx) \quad \text{w.p.1}$$

(in order to get (30), set $j = 0$ in (10)), wherefrom (26) and (27) follow. On the other hand, Lemma 8 and (30) imply that there exists a non-negative random variable K' such that (28) holds. Let $K'' = (\alpha\lambda)^{-1}K'$. Then, it can easily be deduced that (29) holds, too. □

The asymptotic properties of $\{U_{nj}\}_{0\le n\le j}$ and $\{V_{nj}\}_{0\le n\le j}$ are dealt with in the next lemma. The proof is essentially based on the results of Lemma 2.

Lemma 3. *Let A1–A4 hold. Then, there exist $N_0 \in \mathcal{F}$ and a positive random variable L defined on $(\Omega, \mathcal{F}, \mathcal{P})$ such that $\mathcal{P}(N_0) = 0$ and such that the following relations hold on N_0^c:*

$$\overline{\lim_{n\to\infty}} \sup_{n\le j\le\eta(n,t)} \|U_{nj} - I\| \le Lt, \quad \forall t \in (0,1), \tag{31}$$
$$\overline{\lim_{n\to\infty}} \|U_{n,\eta(n,t)} - I - tA\| \le Lt^2, \quad \forall t \in (0,1), \tag{32}$$
$$\overline{\lim_{n\to\infty}} \sup_{n\le j\le\eta(n,t)} \|V_{nj}\| = 0, \quad \forall t \in (0,1). \tag{33}$$

Proof: Let $K = K' + K''$ (K' and K'' are defined in the statement of Lemma 2). Due to Lemmas 2 and 7, there exists $N_0 \in \mathcal{F}$ such that $\mathcal{P}(N_0) = 0$ and such that the following

relations hold on N_0^c:

$$\lim_{n\to\infty} \sup_{n\le j\le \eta(n,t)} \left\| \sum_{i=n}^{j} \gamma_{i+1}(A_{i+1} - A) \right\| = 0, \quad \forall t \in R^+, \tag{34}$$

$$\lim_{n\to\infty} \sup_{n\le j\le \eta(n,t)} \left\| \sum_{i=n}^{j} \gamma_{i+1}(b_{i+1} - b) \right\| = 0, \quad \forall t \in R^+, \tag{35}$$

$$\overline{\lim_{n\to\infty}} \sum_{i=n}^{\eta(n,t)} \gamma_{i+1} \sum_{j=0}^{i} (\alpha\lambda)^{i-j} \|\phi(X_i)\| \|\phi(X_j)\| \le K't, \quad \forall t \in R^+, \tag{36}$$

$$\overline{\lim_{n\to\infty}} \sum_{i=n}^{\eta(n,t)} \gamma_{i+1} \sum_{j=0}^{i+1} (\alpha\lambda)^{i-j} \|\phi(X_{i+1})\| \|\phi(X_j)\| \le K''t, \quad \forall t \in R^+. \tag{37}$$

Since

$$\|A_{n+1}\| \le \sum_{i=0}^{n} (\alpha\lambda)^{n-i} \|\phi(X_n)\| \|\phi(X_i)\| + \sum_{i=0}^{n+1} (\alpha\lambda)^{n-i} \|\phi(X_{n+1})\| \|\phi(X_i)\|, \quad n \ge 0,$$

it follows from (36) and (37) that the following relation also holds on N_0^c:

$$\overline{\lim_{n\to\infty}} \sum_{i=n}^{\eta(n,t)} \gamma_{i+1} \|A_{i+1}\| \le Kt, \quad \forall t \in R^+. \tag{38}$$

Let $L = \|A\| + K^2 \exp(K)$. Let ω be an arbitrary sample from N_0^c (for the sake of notational simplicity, ω does not explicitly appear in the relations and expressions which follow in the proof). It is straightforward to verify that

$$U_{nj} = I + \sum_{i=1}^{j-n} \sum_{n<m_1<\cdots<m_i\le j} \gamma_{m_i} A_{m_i} \cdots \gamma_{m_1} A_{m_1}, \quad 0 \le n \le j,$$

$$V_{nj} = \sum_{i=n+1}^{j-1} U_{i+1,\,j} \gamma_{i+1} A_{i+1} \sum_{k=n}^{i-1} \gamma_{k+1}(A_{k+1}\theta_* + b_{k+1}) + \sum_{i=n}^{j-1} \gamma_{i+1}(A_{i+1}\theta_* + b_{i+1}), \quad 0 \le n \le j,$$

$$\prod_{i=n+1}^{j} (1 + \gamma_i \|A_i\|) = 1 + \sum_{i=1}^{j-n} \sum_{n<m_1<\cdots<m_i\le j} \gamma_{m_1} \left\|A_{m_1}\right\| \cdots \gamma_{m_i} \left\|A_{m_i}\right\|, \quad 0 \le n \le j.$$

Consequently,

$$\left\| U_{nj} - I - \sum_{i=n}^{j-1} \gamma_{i+1} A_{i+1} \right\| \le \sum_{i=2}^{j-n} \sum_{n<m_1<\cdots<m_i\le j} \gamma_{m_1} \left\|A_{m_1}\right\| \cdots \gamma_{m_i} \left\|A_{m_i}\right\|$$

$$\le \sum_{i=2}^{\eta(n,t)-n+1} \sum_{n<m_1<\cdots<m_i\le \eta(n,t)+1} \gamma_{m_1} \left\|A_{m_1}\right\| \cdots \gamma_{m_i} \left\|A_{m_i}\right\|$$

$$= \prod_{i=n+1}^{\eta(n,t)+1} (1+\gamma_i\|A_i\|) - 1 - \sum_{i=n}^{\eta(n,t)} \gamma_{i+1}\|A_{i+1}\|$$

$$\le \left(1 + (\eta(n,t)-n+1)^{-1} \sum_{i=n}^{\eta(n,t)} \gamma_{i+1}\|A_{i+1}\|\right)^{\eta(n,t)-n+1}$$

$$- 1 - \sum_{i=n}^{\eta(n,t)} \gamma_{i+1}\|A_{i+1}\|; \quad \forall t \in R^+,$$

$$0 \le n \le j \le \eta(n,t),$$

$$\|V_{nj}\| \le \left(1 + \sup_{n\le k\le l\le\eta(n,t)} \|U_{kl}\| \sum_{i=n}^{\eta(n,t)} \gamma_{i+1}\|A_{i+1}\|\right)$$

$$\cdot \sup_{n\le k\le\eta(n,t)} \left\| \sum_{i=n}^{k} \gamma_{i+1}(A_{i+1}\theta_* + b_{i+1}) \right\|; \quad \forall t \in R^+, \ \ 0 \le n \le j \le \eta(n,t). \tag{39}$$

Therefore and owing to (38),

$$\overline{\lim_{n\to\infty}} \sup_{n\le j\le\eta(n,t)} \left\| U_{nj} - I - \sum_{i=n}^{j-1} \gamma_{i+1}A_{i+1} \right\| \le f(Kt), \quad \forall t \in R^+, \tag{40}$$

where $f(t) = \exp(t) - t - 1$, $t \in R$ (note that $\lim_{n\to\infty}(\eta(n,t)-n) = \infty$, $\forall t \in R^+$, and that $f(\cdot)$ is increasing on R_0^+). On the other hand,

$$\left\| \sum_{i=n}^{j} \gamma_{i+1}(A_{i+1}\theta_* + b_{i+1}) \right\| \le \left\| \sum_{i=n}^{j} \gamma_{i+1}(A_{i+1} - A) \right\| \|\theta_*\|$$

$$+ \left\| \sum_{i=n}^{j} \gamma_{i+1}(b_{i+1} - b) \right\|, \quad 0 \le n \le j,$$

$$\|U_{nj} - I\| \le \sup_{n\le j\le\eta(n,t)} \left\| U_{nj} - I - \sum_{i=n}^{j-1} \gamma_{i+1}A_{i+1} \right\|$$

$$+ \sup_{n\le j\le\eta(n,t)} \left\| \sum_{i=n}^{j} \gamma_{i+1}(A_{i+1} - A) \right\| + t\|A\|;$$

$$\forall t \in R^+, \ \ 0 \le n \le j \le \eta(n,t), \tag{41}$$

$$\|U_{ij}\| \le 1 + \sup_{n\le k\le l\le\eta(k,t)} \|U_{kl} - I\|; \quad \forall t \in R^+,$$

$$0 \le n \le i \le j \le \eta(n,t), \tag{42}$$

$$\|U_{n,\eta(n,t)} - I - tA\| \le \left\| U_{n,\eta(n,t)} - I - \sum_{i=n}^{\eta(n,t)-1} \gamma_{i+1} A_{i+1} \right\| + \left\| \sum_{i=n}^{\eta(n,t)-1} \gamma_{i+1}(A_{i+1} - A) \right\| + \|A\| \left(t - \sum_{i=n}^{\eta(n,t)-1} \gamma_{i+1} \right); \quad \forall t \in R^+, \ \ n \ge 0 \tag{43}$$

(for obtaining (41) use (15)). Due to (16) and (40)–(43),

$$\overline{\lim_{n\to\infty}} \sup_{n\le j\le\eta(n,t)} \left\| \sum_{i=n}^{j} \gamma_{i+1}(A_{i+1}\theta_* + b_{i+1}) \right\| = 0, \quad \forall t \in R^+, \tag{44}$$

$$\overline{\lim_{n\to\infty}} \sup_{n\le j\le\eta(n,t)} \|U_{nj} - I\| \le \|A\| t + f(Kt), \quad \forall t \in R^+, \tag{45}$$

$$\overline{\lim_{n\to\infty}} \sup_{n\le i\le j\le\eta(n,t)} \|U_{ij}\| < \infty, \quad \forall t \in R^+, \tag{46}$$

$$\overline{\lim_{n\to\infty}} \|U_{n,\eta(n,t)} - I - tA\| \le f(Kt), \quad \forall t \in R^+. \tag{47}$$

Since $f(Kt) \le K^2 t^2 \exp(Kt)$, $\forall t \in R^+$, it can easily be deduced from (38), (39) and ((44)–(47) that (31)–(33) hold. This completes the proof. □

The almost sure boundedness of $\{\vartheta_n\}_{n\ge}$ is shown in the next lemma. The proof essentially relies on the results of Lemma 3.

Lemma 4. *Let A*1–*A*5 *hold. Then,* $\sup_{0\le n} \|\vartheta_n\| < \infty$ *on* N_0^c (N_0 *is defined in the statement of Lemma* 3).

Proof: Let $\lambda_{\min}$ and $\lambda_{\max}$ be the minimal and maximal eigenvalue of $-A$ (respectively). Let ω be an arbitrary sample from N_0^c (for the sake of notational simplicity, ω does not explicitly appear in the relations and expressions which follow in the proof). Let $\tau = \min\{1, \lambda_{\max}^{-1}, 4^{-1}L^{-1}\lambda_{\min}\}$ (L is defined in the statement of Lemma 3) and $\rho = 1 - 2^{-1}\lambda_{\min}\tau$. Obviously, $0 < \rho < 1$, while Lemma 3 implies that there exists $K \in R^+$ (depending on ω) such that

$$\sup_{n\le j\le\eta(n,t)} \max\{\|U_{nj}\|, \|V_{nj}\|\} \le K, \quad n \ge 0. \tag{48}$$

Let $n_0 = 0$ and $n_{k+1} = \eta(n_k, \tau)$, $k \ge 0$. Due to Lemma 3, there exists $k_0 \ge 0$ (depending on ω) such that

$$\left\| U_{n_k,n_{k+1}} - I - \tau A \right\| \le 2L\tau^2, \quad k \ge k_0. \tag{49}$$

Since A is negative definite (due to Lemma 1) and $\tau\lambda_{\max} < 1$, $I + \tau A$ is non-negative definite. Consequently, $\|I + \tau A\| = 1 - \lambda_{\min}\tau$. Therefore and owing to (49),

$$\left\|U_{n_k,n_{k+1}}\right\| \le \left\|U_{n_k,n_{k+1}} - I - \tau A\right\| + \|I + \tau A\| \le \rho, \quad k \ge k_0$$

(note that $1 - \lambda_{\min}\tau + 2L\tau^2 \le \rho$). Then, (23) and (48) yield

$$\left\|\vartheta_{n_{k+1}}\right\| \le \rho\left\|\vartheta_{n_k}\right\| + K, \quad k \ge 0,$$

wherefrom $\sup_{0\le k} \|\vartheta_{n_k}\| < \infty$ follows. As

$$\|\vartheta_j\| \le K\left\|\vartheta_{n_k}\right\| + K; \quad n_k \le j \le n_{k+1}, \quad k \ge 0$$

(due to (23) and (48)), it is obvious that $\sup_{0\le n} \|\vartheta_n\| < \infty$. This completes the proof. □

In the next lemma, the asymptotic behavior of $\{\vartheta_j - \vartheta_n\}_{0\le n\le j}$ and $\{e_n(t)\}_{n\ge 0}$, $t \in R^+$, is dealt with. The proof is based on the results of Lemmas 3 and 4.

Lemma 5. *Let A*1–*A*5 *hold. Then, there exists a positive random variable M defined on* $(\Omega, \mathcal{F}, \mathcal{P})$ *such that the following relations hold on* N_0^c (N_0 *is defined in the statement of Lemma* 3):

$$\lim_{n\to\infty} \|\vartheta_{n+1} - \vartheta_n\| = 0, \tag{50}$$

$$\overline{\lim_{n\to\infty}} \sup_{n\le j\le\eta(n,t)} \|\vartheta_j - \vartheta_n\| \le Mt, \quad \forall t \in (0, 1), \tag{51}$$

$$\overline{\lim_{n\to\infty}} \|e_n(t)\| \le Mt^2, \quad \forall t \in (0, 1). \tag{52}$$

Proof: Due to Lemma 4, there exists a non-negative random variable K defined on $(\Omega, \mathcal{F}, \mathcal{P})$ such that $\|\vartheta_n\| \le K$, $n \ge 0$, on N_0^c. Since

$$\|\vartheta_j - \vartheta_n\| \le K\|U_{nj} - I\| + \|V_{nj}\|, \quad 0 \le n \le j,$$

$$|e_n(t)| \le 2K^2\left\|U_{n,\eta(n,t)} - I - tA\right\| + K^2\left\|U_{n,\eta(n,t)} - I\right\|^2 + K\left\|U_{n,\eta(n,t)}\right\|\left\|V_{n,\eta(n,t)}\right\|; \quad \forall t \in (0, 1), \quad n \ge 0,$$

on N_0^c (due to (23)), Lemma 3 implies that (50)–(52) hold N_0^c (note that $\sup_{0\le n} \|U_{n,\eta(n,t)}\| < \infty$, $\forall t \in (0, 1)$, on N_0^c). □

The almost sure convergence of the algorithm (1)–(3) is demonstrated in the next theorem. The proof is based on the similar ideas as the results of Tadić (1997).

Theorem 1. *Let A*1–*A*2 *hold. Then,* $\lim_{n\to\infty} \theta_n = \theta_*$ *w.p.*1.

Proof: Obviously, it is sufficient to show that $\lim_{n\to\infty}\vartheta_n = 0$ on N_0^c (N_0 is defined in the statement of Lemma 3). Let $\lambda_{\min}$ and $\lambda_{\max}$ be the minimal and maximal eigenvalue of $-A$ (respectively). Let ω be an arbitrary sample from N_0^c (for the sake of notational simplicity, ω does not explicitly appear in the relations and expressions which follow in the proof).

Since $\vartheta_n^T A\vartheta_n \le -\lambda_{\min}\|\vartheta_n\|^2$, $n \ge 0$, (24) implies

$$\left\|\vartheta_{\eta(n,t)}\right\|^2 \le \|\vartheta_n\|^2 - 2\lambda_{\min}t\|\vartheta_n\|^2 + e_n(t); \quad \forall t \in R^+, \quad n \ge 0. \tag{53}$$

Now, let us show that $\underline{\lim}_{n\to\infty}\|\vartheta_n\| = 0$. Suppose the opposite. Then, there exist $\delta \in R^+$ and $n_0 \ge 0$ (both depending on ω) such that $\|\vartheta_n\| \ge \delta$, $n \ge n_0$. Therefore and owing to (53),

$$\left\|\vartheta_{\eta(n,t)}\right\|^2 \le \|\vartheta_n\|^2 - 2\lambda_{\min}\delta^2 t + e_n(t); \quad \forall t \in R^+, \quad n \ge n_0. \tag{54}$$

Due to Lemma 5 and (54),

$$\underline{\lim_{n\to\infty}}\,\|\vartheta_n\|^2 \le \underline{\lim_{n\to\infty}}\left\|\vartheta_{\eta(n,t)}\right\|^2 \le \underline{\lim_{n\to\infty}}\,\|\vartheta_n\|^2 - 2\lambda_{\min}\delta^2 t + Mt^2, \quad \forall t \in (0,1).$$

However, this is impossible, since $-2\lambda_{\min}\delta^2 t + Mt^2 < 0$, $\forall t \in (0, 2\lambda_{\min}\delta^2 M^{-1})$.

Now, let us suppose that $\overline{\lim}_{n\to\infty}\|\vartheta_n\| > 0$. Then, there exists $\varepsilon \in R^+$ (depending on ω) such that $\overline{\lim}_{n\to\infty}\|\vartheta_n\| > 2\varepsilon$. Let $m_0'' = \inf\{n \ge 0 : \|\vartheta_n\| \le \varepsilon\}$, while $m_k' = \inf\{n \ge m_k'' : \|\vartheta_n\| \ge 2\varepsilon\}$, $m_k = \sup\{n \le m_k' : \|\vartheta_n\| \le \varepsilon\}$ and $m_{k+1}'' = \inf\{n \ge m_k' : \|\vartheta_n\| \le \varepsilon\}$, $k \ge 0$. Obviously, $\{m_k\}_{k\ge0}$, $\{m_k'\}_{k\ge0}$ and $\{m_k''\}_{k\ge0}$ are well-defined, as well as $m_k < m_k' < m_{k+1}$, $k \ge 0$, and

$$\begin{aligned} &\left\|\vartheta_{m_k}\right\| \le \varepsilon, \left\|\vartheta_{m_k+1}\right\| > \varepsilon, \left\|\vartheta_{m_k'}\right\| \ge 2\varepsilon; \quad k \ge 0, \\ &\|\vartheta_n\| \ge \varepsilon; \quad m_k < n \le m_k', \quad k \ge 0. \end{aligned} \tag{55}$$

Consequently,

$$\begin{aligned} &\varepsilon \le \left\|\vartheta_{m_k'}\right\| - \left\|\vartheta_{m_k}\right\| \le \left\|\vartheta_{m_k'} - \vartheta_{m_k}\right\|, \quad k \ge 0, \\ &0 \le \varepsilon - \left\|\vartheta_{m_k}\right\| < \left\|\vartheta_{m_k+1}\right\| - \left\|\vartheta_{m_k}\right\| \le \left\|\vartheta_{m_k+1} - \vartheta_{m_k}\right\|, \quad k \ge 0. \end{aligned} \tag{56}$$

Therefore and owing to Lemma 5,

$$\lim_{k\to\infty}\left\|\vartheta_{m_k}\right\| = \varepsilon. \tag{57}$$

Now, let us show that

$$\tau' = \underline{\lim_{k\to\infty}}\sum_{i=m_k}^{m_k'-1}\gamma_{i+1} > 0.$$

Suppose the opposite. Then, there exists a subsequence $\{\tilde{m}_k, \tilde{m}'_k\}_{k\geq 0}$ of $\{m_k, m'_k\}_{k\geq 0}$ such that

$$\lim_{k\to\infty} \sum_{i=\tilde{m}_k}^{\tilde{m}'_k-1} \gamma_{i+1} = 0.$$

Consequently, for all $t \in (0, 1)$, there exists $\tilde{k}_0(t) \geq 0$ (also depending on ω) such that

$$\sum_{i=\tilde{m}_k}^{\tilde{m}'_k-1} \gamma_{i+1} \leq t, \quad k \geq \tilde{k}_0(t). \tag{58}$$

Therefore, $\tilde{m}'_k \leq \eta(\tilde{m}_k, t)$, $\forall t \in (0, 1)$, $k \geq \tilde{k}_0(t)$. Then, Lemma 5 and (56) imply

$$\varepsilon \leq \overline{\lim_{k\to\infty}} \left\| \vartheta_{\tilde{m}'_k} - \vartheta_{\tilde{m}_k} \right\| \leq Mt, \quad \forall t \in (0, 1).$$

However, this is impossible, since the limit process $t \to 0+$ yields $\varepsilon \leq 0$. Hence, $\tau' > 0$.

Let $\tau = \min\{1, 2^{-1}\tau'\}$. Then, for all $t \in (0, \tau)$, there exists $k_0(t) \geq 0$ (also depending on ω) such that $\gamma_{m_k+1} \leq t$, $k \geq k_0(t)$, and

$$\sum_{i=m_k}^{m'_k-1} \gamma_{i+1} > t, \quad k \geq k_0(t).$$

Therefore, $m_k < \eta(m_k, t) \leq m'_k$, $\forall t \in (0, \tau)$, $k \geq k_0(t)$, which, together with (55), implies

$$\left\| \vartheta_{\eta(m_k,t)} \right\| \geq \varepsilon; \quad \forall t \in (0, \tau), \quad k \geq k_0(t).$$

Then, Lemma 5, (54), and (57) yield

$$\varepsilon^2 \leq \overline{\lim_{k\to\infty}} \left\| \vartheta_{\eta(m_k,t)} \right\|^2 \leq \varepsilon^2 - 2\lambda_{\min}\varepsilon^2 t + Mt^2, \quad \forall t \in (0, \tau).$$

However, this is impossible, since $-2\lambda_{\min}\varepsilon^2 t + Mt^2 < 0$, $\forall t \in (0, 2\lambda_{\min}\varepsilon^2 M^{-1})$. Hence, $\lim_{n\to\infty} \|\vartheta_n\| = 0$. This completes the proof. □

An interpretation of the almost sure limit θ_* of $\{\theta_n\}_{n\geq 0}$ is provided in the next theorem. Namely, an upper bound for the error of the approximation of $f_*(\cdot)$ by $\theta_*^T \phi(\cdot)$ is determined in the terms of α, λ and the error of the $L^2(\mu)$-optimal linear approximation of $f_*(\cdot)$. The proof is based on similar ideas as the corresponding result of Tsitsiklis and Van Roy (1997).

Theorem 2. *Let A*2, *A*3 *and A*5 *hold. Then,*

$$\int (\theta_*^T \phi(x) - f_*(x))^2 \mu(dx)$$
$$\leq (1-\alpha)^{-2}(1-\alpha\lambda)^2 \inf_{\theta\in R^d} \int (\theta^T \phi(x) - f_*(x))^2 \mu(dx). \tag{59}$$

Proof: Due to (12) and (14),

$$|(P_n\tilde{g})(x)| \le (P_n\psi)(x) \le 1 + (P_n\psi^2)(x); \quad \forall x \in R^{d'}, \ \ n \ge 0,$$

$$\sum_{n=0}^{\infty}(\alpha\lambda)^n \int \|\phi(x)(P_n\tilde{g})(x)\|\mu(dx) \le (1-\alpha\lambda)^{-1}\int \psi^2(x)\mu(dx) < \infty, \tag{60}$$

$$\sum_{n=0}^{\infty}(\alpha\lambda)^n \int \|\phi(x)(P_{n+1}\phi^T)(x)\|\mu(dx) \le (1-\alpha\lambda)^{-1}\int \psi^2(x)\mu(dx) < \infty. \tag{61}$$

Therefore and owing to Lemma 1,

$$|f_*(x)| \le \sum_{n=0}^{\infty}\alpha^n(P_n\psi)(x), \quad \forall x \in R^{d'},$$

$$\sum_{n=0}^{\infty}\alpha^n \int |(P_n\tilde{g})(x')|P_j(x,dx')$$

$$\le \sum_{n=0}^{\infty}\alpha^n(1+(P_{n+j}\psi^2))(x)$$

$$\le (1-\alpha)^{-1} + \alpha^{-j}\sum_{n=0}^{\infty}\alpha^n(P_n\psi^2)(x) < \infty; \quad \forall x \in R^{d'}, \ \ j \ge 0, \tag{62}$$

$$\sum_{n=0}^{\infty}\sum_{i=1}^{\infty}\alpha^{n+i}\lambda^n|(P_{n+i}\tilde{g})(x)|$$

$$\le \sum_{n=0}^{\infty}\sum_{i=0}^{\infty}\alpha^{n+i}\lambda^n(1+(P_{n+i}\psi^2)(x))$$

$$\le (1-\alpha)^{-1}(1-\alpha\lambda)^{-1} + (1-\lambda)^{-1}\sum_{n=0}^{\infty}\alpha^n(P_n\psi^2)(x) < \infty, \quad \forall x \in R^{d'}, \tag{63}$$

$$\int \|\phi(x)(P_n f_*)(x)\|\mu(dx)$$

$$\le \sum_{i=0}^{\infty}\alpha^i \int \psi(x)(P_{n+i}\psi)(x)\mu(dx)$$

$$\le \sum_{i=0}^{\infty}\alpha^i\left(\int \psi^2(x)\mu(dx)\right)^{1/2}\left(\int (P_{n+i}\psi^2)(x)\mu(dx)\right)^{1/2}$$

$$= (1-\alpha)^{-1}\int \psi^2(x)\mu(dx) < \infty, \quad n \ge 0. \tag{64}$$

According to Lemma 1 and (62),

$$(P_n f_*)(x) = \sum_{i=0}^{\infty}\alpha^i(P_{n+i}\tilde{g})(x); \quad \forall x \in R^{d'}, \ \ n \ge 0.$$

Then, (63) yields

$$\begin{aligned}
&\alpha(1-\lambda)\sum_{n=0}^{\infty}(\alpha\lambda)^n(P_{n+1}f_*)(x)\\
&\quad=(1-\lambda)\sum_{n=0}^{\infty}\sum_{i=1}^{\infty}\alpha^{n+i}\lambda^n(P_{n+i}\tilde{g})(x)\\
&\quad=(1-\lambda)\sum_{n=0}^{\infty}\alpha^n(P_n\tilde{g})(x)\sum_{i=0}^{n-1}\lambda^i\\
&\quad=\sum_{n=1}^{\infty}\alpha^n(P_n\tilde{g})(x)-\sum_{n=1}^{\infty}(\alpha\lambda)^n(P_n\tilde{g})(x)\\
&\quad=f_*(x)-\sum_{n=0}^{\infty}(\alpha\lambda)^n(P_n\tilde{g})(x),\quad \forall x\in R^{d'}. \qquad (65)
\end{aligned}$$

On the other hand, (60), (61), (64) and (65) imply

$$A=-\int\phi(x)\phi^T(x)\mu(dx)+\alpha(1-\lambda)\int\phi(x)\left(\sum_{n=0}^{\infty}(\alpha\lambda)^n(P_{n+1}\phi^T)(x)\right)\mu(dx), \qquad (66)$$

$$\begin{aligned}
b&=\int\phi(x)\left(\sum_{n=0}^{\infty}(\alpha\lambda)^n(P_n\tilde{g})(x)\right)\mu(dx)\\
&=-\alpha(1-\lambda)\int\phi(x)\left(\sum_{n=0}^{\infty}(\alpha\lambda)^n(P_{n+1}f_*)(x)\right)\mu(dx)+\int\phi(x)f_*(x)\mu(dx). \qquad (67)
\end{aligned}$$

Since $A\theta_*+b=0$, (66) and (67) yield

$$\left(\int\phi(x)\phi^T(x)\mu(dx)\right)\theta_*=\int\phi(x)(f_*(x)+h(x))\mu(dx),$$

where

$$h(x)=\alpha(1-\lambda)\sum_{n=0}^{\infty}(\alpha\lambda)^n(\theta_*^T(P_{n+1}\phi)(x)-(P_{n+1}f_*)(x)),\quad x\in R^{d'}.$$

Let

$$\tilde{\theta}_*=\left(\int\phi(x)\phi^T(x)\mu(dx)\right)^{-1}\int\phi(x)f_*(x)\mu(dx)$$

(due to A5 and (64), $\tilde{\theta}_*$ is well-defined and finite). Then, it is straightforward to verify that

$$\begin{aligned} &\int (\tilde{\theta}_*^T \phi(x) - f_*(x))^2 \mu(dx) = \inf_{\theta \in R^d} \int (\theta^T \phi(x) - f_*(x))^2 \mu(dx), \\ &\left(\int \phi(x) \phi^T(x) \mu(dx) \right) (\theta_* - \tilde{\theta}_*) = \int \phi(x) h(x) \mu(dx). \end{aligned} \tag{68}$$

Consequently,

$$\begin{aligned} &\int ((\theta_* - \tilde{\theta}_*)^T \phi(x) - h(x))^2 \mu(dx) \\ &\quad = \int ((\theta_* - \tilde{\theta}_*)^T \phi(x))^2 \mu(dx) + \int h^2(x) \mu(dx) - 2 \int (\theta_* - \tilde{\theta}_*)^T \phi(x) h(x) \mu(dx) \\ &\quad = - \int ((\theta_* - \tilde{\theta}_*)^T \phi(x))^2 \mu(dx) + \int h^2(x) \mu(dx). \end{aligned} \tag{69}$$

Since

$$\begin{aligned} \int (\theta_*^T (P_n \phi)(x) - (P_n f_*)(x))^2 \mu(dx) &\le \int \int (\theta_*^T \phi(x') - f_*(x'))^2 P_n(x, dx') \mu(dx) \\ &= \int (\theta_*^T \phi(x) - f_*(x))^2 \mu(dx), \quad n \ge 0 \end{aligned}$$

(due to the Jensen inequality), it follows from the Minkowski inequality that

$$\begin{aligned} &\left(\int h^2(x) \mu(dx) \right)^{1/2} \\ &\quad \le \alpha(1 - \lambda) \sum_{n=0}^{\infty} (\alpha\lambda)^n \left(\int (\theta_*^T (P_{n+1} \phi)(x) - (P_{n+1} f_*)(x))^2 \mu(dx) \right)^{1/2} \\ &\quad \le \alpha(1 - \lambda)(1 - \alpha\lambda)^{-1} \left(\int (\theta_*^T \phi(x) - f_*(x))^2 \mu(dx) \right)^{1/2}. \end{aligned}$$

Then, (69) implies

$$\begin{aligned} \int ((\theta_* - \tilde{\theta}_*)^T \phi(x))^2 \mu(dx) \le \int h^2(x) \mu(dx) &\le \alpha^2 (1 - \lambda)^2 (1 - \alpha\lambda)^{-2} \\ &\times \int (\theta_*^T \phi(x) - f_*(x))^2 \mu(dx). \end{aligned}$$

Therefore and owing to the Minkowski inequality,

$$\begin{aligned}
&\left(\int (\theta_*^T\phi(x) - f_*(x))^2\mu(dx)\right)^{1/2} \\
&\quad\le \left(\int ((\theta_* - \tilde{\theta}_*)^T\phi(x))^2\mu(dx)\right)^{1/2} + \left(\int (\tilde{\theta}_*^T\phi(x) - f_*(x))^2\mu(dx)\right)^{1/2} \\
&\quad\le \alpha(1-\lambda)(1-\alpha\lambda)^{-1}\left(\int (\theta_*^T\phi(x) - f_*(x))^2\mu(dx)\right)^{1/2} \\
&\qquad + \left(\int (\tilde{\theta}_*^T\phi(x) - f_*(x))^2\mu(dx)\right)^{1/2}.
\end{aligned}$$

Consequently,

$$\int (\theta_*^T\phi(x) - f_*(x))^2\mu(dx) \le (1-\alpha)^{-2}(1-\alpha\lambda)^2 \int (\tilde{\theta}_*^T\phi(x) - f_*(x))^2\mu(dx),$$

wherefrom (59) follows by (68). □

4. Example

The purpose of this section is to illustrate the assumptions A1–A5 and to show that they can be applied to Markov chains of practical interest which are not covered by Tsitsiklis and Van Roy (1997). The example considered in this section is related to the queueing theory.

Let $\{U_n\}_{n\ge1}$ and $\{V_n\}_{n\ge1}$ be i.i.d. R^+-valued random processes defined on a probability space $(\Omega, \mathcal{F}, \mathcal{P})$, while U and V are R^+-valued random variables defined on the same probability space and having the same probability measures as $\{U_n\}_{n\ge1}$ and $\{V_n\}_{n\ge1}$, respectively. Let $\{U_n\}_{n\ge1}$ and $\{V_n\}_{n\ge1}$ be mutually independent, while $X_0 = 0$ and

$$X_{n+1} = (X_n + U_{n+1} - V_{n+1})_+, \quad n \ge 0,$$

where $t_+ = \max\{t, 0\}$, $t \in R$. Then, it can easily be deduced that $\{X_n\}_{n\ge0}$ is a homogeneous Markov chain with the following transition probability:

$$P(x, B) = E\big(I_{B\cap R_0^+}(x + U - V)\big);\ x \in R, \quad B \in \mathcal{B}.$$

In the context of the queueing theory, X_n represents the waiting time of the n-th customer served in GI/G/1 queue with the first in – first out service discipline, while U_{n+1} is the service time of the n-th customer and V_{n+1} is the interarrival time between n-th and $(n+1)$-th customer (for details see e.g., Asmussen, 1987).

Lemma 6. *Let $E(U) < E(V)$ and $E(U^{2p+1}) < \infty$, where $p \in [1, \infty)$ is a constant. Then, $\{X_n\}_{n\ge0}$ is positive Harris with an invariant probability measure $\mu(\cdot)$ satisfying*

$\int x^{2p}\mu(dx) < \infty$. *Moreover, there exists a constant* $K \in R^+$ *such that*

$$E\left(X_n^{2p} \mid X_0 = x\right) \leq K(1 + x^{2p})n^{2p}; \quad \forall x \in R, \ \ n \geq 0. \tag{70}$$

Proof: Due to Asmussen (1987, Theorem XI.2.2), $\{X_n\}_{n\geq 0}$ is positive Harris, while Asmussen (1987, Theorem VIII.2.1) implies $\int x^{2p}\mu(dx) < \infty$. Since

$$|X_n|^{2p} \leq \left(|X_0| + \sum_{i=1}^{n} U_i\right)^{2p} \leq 4^p X_0^{2p} + 4^p \left(\sum_{i=1}^{n} U_i\right)^{2p}, \quad n \geq 1,$$

$$E\left(\sum_{i=1}^{n} U_i\right)^{2p} \leq n^{2p-1} \sum_{i=1}^{n} E\left(U_i^{2p}\right) = n^{2p} E(U^{2p}) < \infty, \quad n \geq 1$$

(due to the Jensen inequality), it can easily be deduced that there exists a constant $K \in R^+$ such that (70) □

Theorem 3. *Let* $E(U) < E(V)$ *and* $E(U^{2p+1}) < \infty$, *where* $p \in [1, \infty)$ *is a constant. Suppose that* $\int \phi(x)\phi^T(x)\mu(dx)$ *is positive definite* ($\mu(\cdot)$ *is the invariant probability measure of* $\{X_n\}_{n\geq 0}$) *and that there exists a constant* $L \in R^+$ *such that*

$$|g(x, x')| \leq L(1 + |x|^p + |x'|^p), \quad \forall x, x' \in R,$$
$$\|\phi(x)\| \leq L(1 + |x|^p), \quad \forall x \in R.$$

Then, $\lim_{n\to\infty} \theta_n = \theta_*$ *w.p.*1 ($\{\theta_n\}_{n\geq 0}$ *is generated by the algorithm* (1)–(3), *while* $\theta_* = -A^{-1}b$, *where* A *and* b *are defined in* (20) *and* (21)).

Proof: Let $\psi(x) = 2^{p+1}L(1 + |x|^p + E(U^{2p}))$, $x \in R$. Since

$$\int g^2(x, x')P(x, dx') \leq 3L^2(1 + x^{2p} + E((x + U - V)_+^{2p}))$$
$$\leq 4^{p+1}L^2(1 + x^{2p} + E(U^{2p})), \quad \forall x \in R,$$

it can easily be deduced using Lemmas 6 and 9 that the conditions of Theorem 1 hold. This completes the proof. □

To the best of the present author's knowledge, there are no results on the geometric ergodicity of the waiting times of GI/G/1 queue under the conditions of Lemma 6 and Theorem 3. Therefore, it seems that the assumptions of Tsitsiklis and Van Roy (1997) are not likely to cover the example presented in this section. Moreover, it is straightforward to extend the results of Theorem 3 to the case of GI/G/m, $m > 1$, queues. Furthermore, Dai (1995) gives directions how A1–A5 can be verified in the context of queueing networks.

5. Conclusion

The asymptotic properties of temporal-difference learning algorithms with linear function approximation have been analyzed in this paper. The analysis has been carried out in the context of the approximation of a discounted cost-to-go function associated with an uncontrolled Markov chain with an uncountable finite-dimensional state-space. Under mild conditions and using entirely different arguments than those which the previous results are based on, the almost sure convergence of temporal-difference learning algorithms with linear function approximation has been established and an upper bound for their asymptotic approximation error has been determined. Moreover, the obtained results have been illustrated by an example related to the queueing theory and not covered by the previous results on temporal-difference learning.

The results of this paper are a generalization and extension of those presented in Tsitsiklis and Van Roy (1997). In comparison with the assumptions adopted in Tsitsiklis and Van Roy (1997), the assumptions of this paper are more general and cover a significantly broader class of Markov chains of practical interest. The assumptions used in this paper allow the chain to be positive Harris, while the analysis carried out in Tsitsiklis and Van Roy (1997) practically covers only the case where the underlying chain is geometrically ergodic. Furthermore, the assumptions adopted here seem to be the weakest possible under which the almost sure convergence of temporal-difference learning algorithms with linear function approximation is still possible to be demonstrated.

Appendix

Lemma 7. *Let A*1 *hold. Let* $\{x_n\}_{n\geq 0}$ *be a sequence of reals satisfying*

$$\lim_{n\to\infty} n^{-1}\sum_{i=0}^{n-1} x_i = x,$$

where $x \in R$. *Then,*

$$\lim_{n\to\infty} \sup_{n\leq j\leq \eta(n,t)} \left|\sum_{i=n}^{j} \gamma_{i+1}(x_{i+1}-x)\right| = 0, \quad \forall t \in R^+, \tag{71}$$

$$\overline{\lim_{n\to\infty}} \sum_{i=n}^{\eta(n,t)} \gamma_{i+1}x_{i+1} \leq t|x|, \quad \forall t \in R^+. \tag{72}$$

Proof: Due to A1, there exists a constant $c \in R^+$ such that

$$n\gamma_{n+i} \leq c, n\left|\gamma_n\gamma_{n+1}^{-1} - 1\right| \leq c; \quad n \geq 1, \quad i \geq 0. \tag{73}$$

Let

$$u_n = \sum_{i=0}^{n-1}(x_i - x), \quad n \geq 1.$$

Then,

$$\sum_{i=n}^{j}\gamma_{i+1}(x_{i+1}-x)$$
$$=(j+1)\gamma_{j+1}u_{j+1}-n\gamma_n u_n+\sum_{i=n}^{j}i\big(\gamma_i\gamma_{i+1}^{-1}-1\big)\gamma_{i+1}u_i,\quad 1\le n\le j.$$

Therefore and owing to (15) and (73),

$$\left|\sum_{i=n}^{j}\gamma_{i+1}(x_{i+1}-x)\right|\le c(2+t+cn^{-1})\sup_{n\le i}|u_i|;\quad \forall t\in R^+,\ 1\le n\le j\le\eta(n,t),$$
$$\sum_{i=n}^{j}\gamma_{i+1}x_{i+1}\le(t+cn^{-1})|x|+c(2+t+cn^{-1})\sup_{n\le i}|u_i|;$$
$$\forall t\in R^+,\ \ 1\le n\le j\le\eta(n,t),$$

wherefrom (71) and (72) follow. □

Lemma 8. *Let $\alpha\in(0,1)$ be a constant, while $\{x_n\}_{n\ge0}$, $\{y_n\}_{n\ge0}$ and $\{z_n\}_{n\ge0}$ are sequences of reals satisfying*

$$\overline{\lim_{n\to\infty}}\ n^{-1}\sum_{i=0}^{n-1}\big(x_i^2+y_i^2\big)<\infty,$$
$$\lim_{n\to\infty}n^{-1}\sum_{i=0}^{n-1}x_{i+j}y_i=z_j,\quad j\ge0.$$

Then, $\sum_{n=0}^{\infty}\alpha^n|z_n|<\infty$ and

$$\lim_{n\to\infty}n^{-1}\sum_{i=0}^{n-1}\sum_{j=0}^{i}\alpha^{i-j}x_iy_j=\sum_{n=0}^{\infty}\alpha^n z_n.\tag{74}$$

Proof: Let

$$u_n=n^{-1}\sum_{i=0}^{n-1}x_i^2,\quad n\ge1,$$
$$v_n=n^{-1}\sum_{i=0}^{n-1}y_i^2,\quad n\ge1,$$
$$z_{jn}=n^{-1}\sum_{i=0}^{n-1}x_{i+j}y_i,\quad n\ge1,\ \ j\ge0.$$

Obviously, there exists a constant $c \in R^+$ such that $\max\{u_n, v_n\} \le c$, $n \ge 1$. Therefore,

$$|z_{jn}| \le n^{-1}\left(\sum_{i=0}^{n-1} x_{i+j}^2\right)^{1/2}\left(\sum_{i=0}^{n-1} y_i^2\right)^{1/2}$$
$$\le (1 + jn^{-1})^{1/2} u_{n+j}^{1/2} v_n^{1/2} \le c(1 + jn^{-1})^{1/2}; \quad n \ge 1, \quad j \ge 0.$$

Consequently,

$$|(1 - in^{-1})z_{i,n-i}| \le c(1 - in^{-1})^{1/2}, \quad 0 \le i < n, \tag{75}$$
$$|z_n| \le c, \quad n \ge 0, \tag{76}$$

wherefrom $\sum_{n=0}^{\infty} \alpha^n |z_n| < \infty$ follows. On the other hand,

$$n^{-1}\sum_{i=0}^{n-1}\sum_{j=0}^{i} \alpha^{i-j} x_i y_j - \sum_{i=0}^{\infty} \alpha^i z_i$$
$$= \sum_{i=0}^{n-1} \alpha^i (1 - in^{-1}) z_{i,n-i} - \sum_{i=0}^{\infty} \alpha^i z_i$$
$$= \sum_{i=0}^{k-1} \alpha^i (1 - in^{-1})(z_{i,n-i} - z_i) + \sum_{i=0}^{k-1} \alpha^i in^{-1} z_i$$
$$+ \sum_{i=k}^{n-1} \alpha^i (1 - in^{-1}) z_{i,n-i} - \sum_{i=k}^{\infty} \alpha^i z_i, \quad 1 \le k \le n. \tag{77}$$

Due to (75)–(77),

$$\left| n^{-1}\sum_{i=0}^{n-1}\sum_{j=0}^{i} \alpha^{i-j} x_i y_j - \sum_{i=0}^{\infty} \alpha^i z_i \right|$$
$$\le \sum_{i=0}^{k-1} |z_{i,n-i} - z_i| + 2c(1 - \alpha)^{-1}\alpha^k + ck^2 n^{-1}, \quad 1 \le k \le n.$$

Consequently,

$$\overline{\lim_{n\to\infty}} \left| n^{-1}\sum_{i=0}^{n-1}\sum_{j=0}^{i} \alpha^{i-j} x_i y_j - \sum_{i=0}^{\infty} \alpha^i z_i \right| \le 2c(1 - \alpha)^{-1}\alpha^k, \quad k \ge 1,$$

wherefrom (74) results by the limit process $k \to \infty$. □

Lemma 9. *Let $\{X_n\}_{n\ge 0}$ be an R^d-valued homogeneous Markov chain defined on the probability space $(\Omega, \mathcal{F}, \mathcal{P})$ and having a unique invariant probability measure $\mu(\cdot)$. Let*

$P(x,\cdot)$, $x \in R^d$, be the transition probability of $\{X_n\}_{n\geq 0}$, while $f : R^{d(j+1)} \to R$ is a Borel-measurable function satisfying

$$\int\int \cdots \int |f(x_0, x_1, \ldots, x_j)| P(x_{j-1}, dx_j) \cdots P(x_0, dx_1)\mu(dx_0) < \infty \tag{78}$$

($j \geq 0$). If $\{X_n\}_{n\geq 0}$ is positive Harris, then

$$\lim_{n\to\infty} n^{-1} \sum_{i=0}^{n-1} f(X_i, \ldots, X_{i+j}) = \int\int \cdots \int f(x_0, x_1, \ldots, x_j) P(x_{j-1}, dx_j) \cdots P(x_0, dx_1)\mu(dx_0) \quad \text{w.p.1}\,. \tag{79}$$

Proof: Let $\Lambda \in \mathcal{F}$ be the event where (79) holds. Due to Meyn and Tweedie (1993, Theorem 17.1.2) and (78), there exist $x \in R^d$ such that $\mathcal{P}(\Lambda \mid X_0 = x) = 1$. Therefore and owing to Meyn and Tweedie (1993, Theorem 17.1.7), $\mathcal{P}(\Lambda \mid X_0 = x) = 1$, $\forall x \in R^d$, wherefrom $\mathcal{P}(\Lambda) = 1$ follows. This completes the proof. □

Note

In Clark (1984), Kulkarni and Horn (1996), and Wang, Chong, and Kulkarni (1996) the following general result has been established. Let

$$\theta_{n+1} = \theta_n + \gamma_{n+1} h(\theta_n) + \gamma_{n+1}\xi_{n+1}, \quad n \geq 0,$$

where $\{\gamma_n\}_{n\geq 1}$ is a sequence of positive reals satisfying A1, $h : R^d \to R^d$ is a continuous function fullfiling $h(\theta_*) = 0$ and $(\theta - \theta_*)^T h(\theta) < 0$, $\forall \theta \in R^d \backslash \{\theta_*\}$, while $\theta_* \in R^d$ is a deterministic vector and $\{\xi_n\}_{n\geq 1}$ is an R^d-valued random process. Then, $\lim_{n\to\infty} \theta_n = \theta_*$ w.p.1 only if

$$\lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} \xi_i = 0 \quad \text{w.p.1.}$$

Using this result and (22) (given in Section 3), it can easily be deduced that the algorithm (1)–(3) converges w.p.1 only if

$$\lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} (A_i\theta_* + b_i) = A\theta_* + b \quad \text{w.p.1,} \tag{80}$$

where $\{A_n\}_{n\geq 1}$, $\{b_n\}_{n\geq 1}$, A and b are defined in (18)–(21) (given in Section 3), while $\theta_* = A^{-1}b$. Although A4 is only a sufficient condition for (80) (see Lemma 2), in this context it is hard (if possible at all) to imagine any weaker condition leading to (80).

References

Asmussen, S. (1987). *Applied Probability and Queues*. New York: Wiley.

Benveniste, A., Metivier, M., Priouret, P. (1990). *Adaptive Algorithms and Stochastic Approximation*. Berlin: Springer Verlag.

Bertsekas, D. P. (1976). *Dynamic Programming and Optimal Control*. New York: Academic Press.
Bertsekas, D. P. & Tsitsiklis, J. N. (1996). *Neuro-Dynamic Programming*. Belmont, MA: Athena Scientific.
Clark, D. S. (1984). Necessary and sufficient conditions for the Robbins-Monro method. *Stochastic Processes and their Applications, 17*, 359–367.
Chen, H. F. & Guo, L. (1991). *Identification and Stochastic Adaptive Control*. Basel: Birkhäuser Verlag.
Dai, J. G. (1995). On positive Harris recurrence of multiclass queueing networks: A unified approach via fluid limit models. *Annals of Applied Probability, 5*, 49–77.
Dayan, P. D. (1992). The convergence of $TD(\lambda)$ for general λ. *Machine Learning, 8*, 341–362.
Dayan, P. D. & Sejnowski, T. J. (1994). $TD(\lambda)$ converges with probability 1. *Machine Learning, 14*, 295–301.
Jaakola, T., Jordan, M. I., & Singh, S. P. (1994). On the convergence of stochastic iterative dynamic programming algorithms. *Neural Computation, 6*, 1185–1201.
Kulkarni, S. R. & Horn, C. S. (1996). An alternative proof for convergence of stochastic approximation algorithms. *IEEE Transactions of Automatic Control, 41*, 419–424.
Kumar, P. R. & Varaiya, P. (1986). *Stochastic Systems: Estimation, Identification and Adaptive Control*. Englewood Cliffs, NJ: Prentice Hall.
Kushner, H. J. & Clark, D. S. (1978). *Stochastic Approximation Methods for Constrained and Unconstrained Systems*. Berlin: Springer Verlag.
Ljung, L., Pflug, G., & Walk, H. (1992). *Stochastic Approximation and Optimization of Random Systems*. Basel: Birkhäuser Verlag.
Meyn, S. P. & Tweedie, R. L. (1993). *Markov Chains and Stochastic Stability*. Berlin: Springer-Verlag.
Solo, V. & Kong, X. (1995). *Adaptive Signal Processing Algorithms: Stability and Performance*. Englewood Cliffs, NJ: Prentice Hall.
Sutton, R. S. (1988). Learning to predict by the methods of temporal-differences. *Machine Learning, 3*, 9–44.
Sutton, R. S. & Barto, A. G. (1998). *Reinforcement Learning: An Introduction.* Cambridge, MA: MIT Press.
Tadić, V. (1997). Convergence of stochastic approximation under general noise and stability conditions. In *Proceedings of the 36 IEEE Conference on Decision and Control*.
Tsitsiklis, J. N. & Van Roy, B. (1997). An analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control, 42*, 674–690.
Wang, I.-J., Chong, E. K. P., & Kulkarni, S. R. (1996). Equivalent and sufficient conditions on noise sequences for stochastic approximation algorithms. *Advances in Applied Probability, 28*, 784–801.

Received March 30, 1999
Revised April 21, 2000
Accepted May 9, 2000
Final manuscript