15th Triennial World Congress, Barcelona, Spain

ON THE ALMOST SURE RATE OF CONVERGENCE OF TEMPORAL-DIFFERENCE LEARNING ALGORITHMS

Vladislav B. Tadić *,1

* *Department of Electrical and Electronic Engineering*
The University of Melbourne, Parkville, Victoria 3010, Australia
e-mail: v.tadic@ee.mu.oz.au

Abstract: In this paper, the almost sure rate of convergence of temporal-difference learning algorithms is analyzed. The analysis is carried out for the case of discounted cost function associated with a Markov chain with a finite dimensional state-space. Under mild conditions, it is shown that these algorithms converge at the rate $O(n^{-1/2}(\log\log n)^{1/2})$ almost surely. Since $O(n^{-1/2}(\log\log n)^{1/2})$ characterizes the rate of convergence in the law of iterated logarithm, the obtained results could be considered as the same law for temporal-difference learning algorithms. For the same reason, the obtained convergence rate is probably the least conservative result of this kind. The results are illustrated with examples related to random coefficient autoregression models and $M/G/1$ queues.

Keywords: Machine learning, temporal-difference learning, reinforcement learning, almost sure rate of convergence, law of iterated logarithm, Markov chains, uniform ergodicity, discounted cost function, linear approximation.

1. INTRODUCTION

Temporal-difference learning with function approximation is a recursive parametric method for approximating a cost 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.

The problems of the prediction and approximation of a cost-to-go function associated with a stochastic system modeled 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), 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 and Tsitsiklis, 1996), (Sutton and Barto, 1998) and references cited therein), while their asymptotic properties (almost sure convergence, convergence in mean and probability, convergence of mean) have been analyzed in a great number of papers (see see also (Bertsekas and Tsitsiklis, 1996), (Sutton and Barto, 1998) and references cited therein). Although the existing results provide a good insight into the asymptotic behavior of temporal-difference learning, not much is known about their convergence rate.

[1] This work has been supported by the Australian Research Council (ARC) and the Center of Expertise in Networked Decision Systems (CENDS).

In this paper, the almost sure convergence rate of temporal-difference learning algorithms is analyzed. The analysis is carried out for the case of discounted cost function associated with a Markov chain with a finite dimensional state-space. Under mild conditions, it is shown that these algorithms converge at the rate $O(n^{-1/2}(\log\log n)^{1/2})$ almost surely. Since $O(n^{-1/2}(\log\log n)^{1/2})$ characterizes the rate of convergence in the law of iterated logarithm, the obtained results could be considered as the same law for temporal-difference learning algorithms. For the same reason, the obtained convergence rate is probably the least conservative result of this kind. Furthermore, to the best of the author's knowledge, there does not exist a similar result in the available literature on reinforcement learning. The main results of the paper are illustrated with examples related to random coefficient autoregression models and $M/G/1$ queues. This paper is a continuation of the author's work presented in (Tadić, 2001*b*).

2. MAIN RESULTS

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

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

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

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

$\{\gamma_n\}_{n\geq 1}$ is a sequence of positive reals, while $\alpha \in (0,1)$ and $\lambda \in (0,1]$ are are constants. $c : R^{d'} \times R^{d'} \to R$ and $\phi : R^{d'} \to R^d$ 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.

For $x \in R^{d'}$, let

$$J_*(x) = E\left(\sum_{n=0}^{\infty} \alpha^n c(X_n, X_{n+1}) \middle| X_0 = x\right)$$

(provided that $J_*(\cdot)$ is well-defined). In the context of dynamic programming, $J_*(\cdot)$ is interpreted as a discounted cost function associated with the chain $\{X_n\}_{n\geq 0}$ (for details see e.g., (Bertsekas and Tsitsiklis, 1996)). The task of the algorithm (1) – (3) is to approximate the function $J_*(\cdot)$. It determines the optimal value θ_* of the parameter $\theta \in R^d$ such that the $\theta_*^T\phi(\cdot)$ is the best approximator of $J_*(\cdot)$ among the family $\{\theta^T\phi(\cdot)\}_{\theta\in R^d}$. If $\lambda = 1$ and $\{X_n\}_{n\geq 0}$ has a unique invariant probability measure $\pi(\cdot)$, the algorithm (1) – (3) determines $\theta_* \in R^d$ such that $\theta_*^T\phi(\cdot)$ approximates $J_*(\cdot)$ optimally in the $L^2(\pi)$-sense, i.e., it searches for the minimum of the function $J(\theta) = \int(\theta^T\phi(x) - J_*(x))^2\pi(dx)$, $\theta \in R^d$.

Let $R^+ = (0,\infty)$, $R_0^+ = [0,\infty)$ and $\bar{R}_0^+ = [0,\infty]$, while $\|\cdot\|$ denotes the Euclidean vector norm and the matrix norm induced by the Euclidean vector norm (i.e., $\|A\| = \sup_{\|\theta\|=1}\|A\theta\|$, $A \in R^{d\times d}$). Let $P(x,\cdot)$, $x \in R^{d'}$, and $P_n(x,\cdot)$, $x \in R^{d'}$, be the single and n-th step transition probability kernel of $\{X_n\}_{n\geq 0}$ (respectively), i.e., $P(x,\cdot) = P_1(x,\cdot)$, $\forall x \in R^{d'}$, and

$$\mathcal{P}(X_n \in B | X_0 = x) = P_n(x,B) \ \ w.p.1, \quad \forall B \in \mathcal{B}^{d'},\ \forall x \in R^{d'}\ n \geq 0.$$

In this paper, the almost sure convergence of the algorithm (1) – (3) is analyzed under the following conditions.

A1. $0 < \gamma = \lim_{n\to\infty} n\gamma_n < \infty$ and $\gamma' = \overline{\lim}_{n\to\infty} n|\gamma_n\gamma_{n+1}^{-1} - 1| < \infty$.

A2. $\{X_n\}_{n\geq 0}$ is positive Harris with $\pi(\cdot)$ as a (unique) invariant probability measure.

A3. There exist a constant $p \in (2,\infty)$ and a Borel-measurable function $f : R^{d'} \to R_0^+$ such that $\|\phi(x)\| \leq f(x)$, $\forall x \in R^{d'}$, and

$$\int f^{2p}(x)\pi(dx) < \infty, \tag{4}$$

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

$$\sum_{n=0}^{\infty} \alpha^n (P_n f^p)(x) < \infty, \quad \forall x \in R^{d'}. \tag{5}$$

A4. There exist a constant $q \in (2,\infty)$ and a Borel-measurable function $g : R^{d'} \to R_0^+$ such that

$$\int g^q(x)\pi(dx) < \infty,$$

$$\sum_{n=0}^{\infty} \left\| \int \phi(x')(P_m\phi^T)(x')(P_n - \pi)(x,dx') \right\| \leq g(x), \quad \forall x \in R^{d'},\ m \geq 0, \tag{6}$$

$$\sum_{n=0}^{\infty} \left\| \int \phi(x')(P_m\tilde{c})(x')(P_n - \pi)(x,dx') \right\| \leq g(x), \quad \forall x \in R^{d'},\ m \geq 0, \tag{7}$$

where $\tilde{c}(x) = \int c(x,x')P(x,dx')$.

A5.

$$\int \phi(x)\phi^T(x)\pi(dx) - 2^{-1}\gamma^{-1}(1-\alpha)^{-1}(1-\alpha\lambda)I$$

is positive definite.

Assumption A1 corresponds with the algorithm step size. It is satisfied if $\gamma_n = \gamma n^{-1}$, $n \geq 1$, which is a typical choice for the step step of reinforcement learning algorithms (see e.g., (Bertsekas and Tsitsiklis, 1996), (Sutton and Barto, 1998) and references cited therein).

Assumption A2 is related to the recurrence and stationarity properties of $\{X_n\}_{n\geq 0}$. Assumptions of this type are standard for the asymptotic analysis of temporal-difference learning algorithms (see (Bertsekas and Tsitsiklis, 1996), (Sutton and Barto, 1998) and references cited therein; see also (Tadić, 2001*b*) and (Tsitsiklis and Roy, 1997)).

Assumption A3 corresponds with the growth rate of $c(\cdot,\cdot)$ and $\phi(\cdot)$. It requires these functions not to grow too fast so that their upper bound $f(\cdot)$ satisfies (4) and (5). The role of A3 is to ensure (together with A2) that certain functions of $\{X_n\}_{n\geq 0}$ admit the law of large numbers (see (Tadić, 2001*a*)), as well as to provide that $J_*(\cdot)$ and A_*, b_* (defined in (9) and (10)) are well-defined and finite. A3 is satisfied if $c(\cdot,\cdot)$ and $\phi(\cdot)$ are globally bounded or if $c(\cdot,\cdot)$ 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 also satisfied if $\{X_n\}_{n\geq 0}$ is uniformly ergodic (see Section 3).

Assumption A4 is related to the stability of $\{X_n\}_{n\geq 0}$. Basically, A4 requires $\{X_n\}_{n\geq 0}$ to exhibit sufficient sufficient "degree of stability" (i.e., $P_n(x,\cdot)$, $x \in R^{d'}$, to converge to $\pi(\cdot)$ sufficiently fast) so that (6) and (7) hold. Its role is to ensure that certain Poisson equations have unique solutions (see (Tadić, 2001*a*)). A4 is satisfied under uniform ergodicity conditions (see Section 3) and is typical for the asymptotic analysis of temporal-difference learning algorithms (see (Bertsekas and Tsitsiklis, 1996), (Sutton and Barto, 1998) and references cited therein; see also (Tadić, 2001*b*) and (Tsitsiklis and Roy, 1997)).

Assumption A5 is a "persistancy of excitation" condition. These conditions are typical for the areas of system identification, adaptive control and adaptive signal processing (see e.g., (Goodwin and Sin, 1984), (Solo and Kong, 1995) and references cited therein). A5 requires $\{\phi(X_n)\}_{n\geq 0}$ to sufficiently "rich" with respect to all directions in R^d at the asymptotic steady-state characterized by $\pi(\cdot)$. It is satisfied if $\int \phi(x)\phi^T(x)\pi(dx)$ is positive definite (i.e., $\pi(x : \theta^T\phi(x) \neq 0) > 0$, $\forall x \in R^d$) and γ is sufficiently small. The requirement that $\int \phi(x)\phi^T(x)\pi(dx)$ is positive definite is standard for the asymptotic analysis of temporal-difference learning algorithms (see (Bertsekas and Tsitsiklis, 1996), (Sutton and Barto, 1998) and references cited therein; see also (Tadić, 2001*b*) and (Tsitsiklis and Roy, 1997)).

The main results of the paper are contained in the next theorem.

Theorem 1. Let A1 – A5 hold. Then, there exists a constant $C \in R^+$ such that

$$\overline{\lim_{n\to\infty}}\, n^{1/2}(\log\log n)^{-1/2}\|\theta_n - \theta_*\| \leq C \quad w.p.1, \tag{8}$$

where $\theta_* = -A_*^{-1}b_*$ and

$$A_* = -\int \phi(x)\phi^T(x)\pi(dx) + \alpha(1-\lambda)\sum_{n=0}^{\infty}(\alpha\lambda)^n \int \phi(x)(P_{n+1}\phi^T)(x)\pi(dx), \tag{9}$$

$$b_* = \sum_{n=0}^{\infty}(\alpha\lambda)^n \int \phi(x)(P_n\tilde{c})(x)\pi(dx). \tag{10}$$

Moreover, (8) holds with

$$C = \tilde{C}(1+\tilde{K})(1+\tilde{\lambda}_{min}^{-1}),$$

where $\tilde{C} = c(1+K\lambda_{min}^{-1})(K+L)$, $\tilde{K} = K + 2^{-1}\gamma^{-1}$ and $\tilde{\lambda}_{min} = \lambda_{min} - 2^{-1}\gamma^{-1}$, while λ_{min} is the minimal eigenvalue of $-A_*$, $c = 36(1+\gamma+\gamma')$ and

$$K = 3(1-\alpha\lambda)^{-2}\left(\int f^{2p}(x)\pi(dx)\right)^{1/p},$$

$$L = 2(1-\alpha\lambda)^{-1}\left(\int g^q(x)\pi(dx)\right)^{1/q}.$$

The proof is given in (Tadić, 2001*a*).

Remark. It is important to notice that C does not depend on d and d'. This means that the convergence rate (or learning rate) of the algorithm (1) – (3) is asymptotically independent of the dimension d of the parameter θ and the dimension d' of the state-space of $\{X_n\}_{n\geq 0}$.

Asymptotic behavior of temporal-difference learning algorithms has been considered in a large number of papers (see e.g., (Dayan, 1992), (Dayan and Sejnowski, 1994), (Jaakola *et al.*, 1994), (Sutton, 1988), (Tadić, 2001*b*), (Tsitsiklis and Roy, 1997); see also (Bertsekas and Tsitsiklis, 1996), (Sutton

and Barto, 1998) and references cited therein). Although the existing results provide a good insight into the asymptotic behavior of temporal-difference learning algorithms, not much is known about their rate of convergence. The strongest existing results on their asymptotic behavior are probably contained in (Tsitsiklis and Roy, 1997) (recently, the results of (Tsitsiklis and Roy, 1997) have been extended in (Tadić, 2001*b*)). In comparison with the assumptions adopted in (Tsitsiklis and Roy, 1997), A1 – A5 are just slightly more restrictive: for sufficiently large γ, the assumptions of (Tsitsiklis and Roy, 1997) would be a special case of A1 - A5 if A3 were replaced with the requirement that there exists a Borel-measurable function $f : R^{d'} \to R_0^+$ such that $\int f^2(x)\pi(dx) < \infty$, $\|\phi(x)\| \leq f(x)$, $\forall x \in R^{d'}$, and

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

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

However, only the almost sure convergence of temporal-difference learning algorithms has been analyzed in (Tsitsiklis and Roy, 1997). On the other hand, the results presented in this paper could be thought of as the law of large numbers for these algorithms. Therefore, the results of this paper could be considered as probably the least conservative result on the almost sure rate of convergence of temporal-difference learning algorithms. To the best knowledge of the present author, there does not exist a similar result in the literature on reinforcement learning.

3. SPECIAL CASES

The results of this section correspond with the special cases of A1 – A5 where $\{X_n\}_{n\geq 0}$ is uniformly ergodic and continuously valued (Theorem 2) and where the state-space of $\{X_n\}_{n\geq 0}$ is countable (Theorem 3).

The algorithm (1) – (3) is analyzed now for the case where $\{X_n\}_{n\geq 0}$ is uniformly ergodic. The assumptions under which the analysis is carried out are as follows:

B1. $\{X_n\}_{n\geq 0}$ has a unique invariant probability measure $\pi(\cdot)$.

B2. There exist constants $p, q \in (2, \infty)$ and a Borel-measurable function $f : R^{d'} \to [1, \infty)$ such that $\int f^{pq}(x)\pi(dx) < \infty$, $\|\phi(x)\| \leq f(x)$, $\forall x \in R^{d'}$, and

$$\int |c(x,x')|^p P(x,dx') \leq f^p(x), \quad \forall x \in R^{d'}.$$

B3. There exist constants $M \in [1, \infty)$ and $\rho \in (0, 1)$ such that

$$\left| \int \varphi(x')(P_n - \pi)(x, dx') \right| \leq M\rho^n f^p(x), \quad \forall x \in R^{d'}, \ n \geq 0, \qquad (11)$$

for any Borel-measurable function $\varphi : R^{d'} \to R$ satisfying $0 \leq \varphi(x) \leq f^p(x)$, $\forall x \in R^{d'}$.

Remark. Due to the Jensen inequality,

$$\begin{aligned} &\int f^p(x)\pi(dx) \\ &\leq \left(\int f^{2p}(x)\pi(dx) \right)^{1/2} \\ &\leq \left(\int f^{pq}(x)\pi(dx) \right)^{1/q} < \infty. \end{aligned} \qquad (12)$$

Then, it is clear that the left-hand side of (11) is well-defined.

Theorem 2. Let A1, A5 and B1 – B3 hold. Then, the conclusions of Theorem 1 hold with

$$g(x) = \tilde{L}(1 + f^p(x)), \quad x \in R^{d'},$$

where $\tilde{L} = 4\tilde{M}^2(1 - \rho^{1/r})^{-1}$, while $r = p/(p-2)$ and

$$\tilde{M} = M + \int f^p(x)\pi(dx).$$

The proof is given in (Tadić, 2001*a*).

The algorithm (1) – (3) is now analyzed for the case where the state-space of $\{X_n\}_{n\geq 0}$ is countable. Let Z_0^+ (the set of non-negative integers) be the state-space of $\{X_n\}_{n\geq 0}$. In that case, the functions $c(\cdot,\cdot)$ and $\phi(\cdot)$ (appearing in the algorithm (1) – (3)) map $Z_0^+ \times Z_0^+$ into R and Z_0^+ into R^d, respectively. Let $p_{ij}^n = P(X_n = j | X_0 = i)$ and $p_{ij} = p_{ij}^1$, $n, i, j \geq 0$. The assumptions under which the analysis is carried out are as follows:

C1. $\{X_n\}_{n\geq 0}$ has a unique invariant distribution $\pi = \{\pi_n\}_{n\geq 0}$ (i.e., $\pi_j = \lim_{n\to\infty} p_{ij}^n$, $i, j \geq 0$).

C2. There exist constants $p, q \in (2, \infty)$ and a sequence $\{f_n\}_{n\geq 0}$ from $[1, \infty)$ such that $\|\phi(n)\| \leq f_n$, $n \geq 0$, and

$$\sum_{n=0}^{\infty} f_n^{pq} \pi_n < \infty,$$

$$\sum_{i=0}^{\infty} |c(n,i)|^p p_{ni} \leq f_n^p, \quad n \geq 0.$$

C3. There exist constants $M \in [1, \infty)$ and $\rho \in (0,1)$ such that

$$\left| \sum_{j=0}^{\infty} \varphi_j (p_{ij}^n - \pi_j) \right| \leq M \rho^n f_i^p, \quad n, i \geq 0,$$

for any sequence $\{\varphi_n\}_{n \geq 0}$ satisfying $0 \leq \varphi_n \leq f_n^p$, $n \geq 0$.

C4.

$$\sum_{n=0}^{\infty} \phi(n) \phi^T(n) \pi_n$$
$$-2^{-1} \gamma^{-1} (1-\alpha)^{-1} (1 - \alpha\lambda) I$$

is positive definite.

As an immediate d consequence of Theorem 2, the following result is obtained.

Theorem 3. Let A1 and C1 – C4 hold. Then, the conclusions of Theorem 1 hold with

$$A_* = -\sum_{i=0}^{\infty} \phi(i) \phi^T(i) \pi_i$$
$$+\alpha(1-\lambda) \sum_{n=0}^{\infty} \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} (\alpha\lambda)^n \phi(i) \phi^T(j) p_{ij}^{n+1} \pi_i,$$

$$b_* = \sum_{n=0}^{\infty} \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \sum_{k=0}^{\infty} (\alpha\lambda)^n \phi(i) c(j,k) p_{jk} p_{ij}^n \pi_i,$$

$$K = 3(1-\alpha\lambda)^{-2} \left(\sum_{n=0}^{\infty} f_n^{2p} \pi_n \right)^{1/p},$$

$$L = 2(1-\alpha\lambda)^{-1} \tilde{L} \left(1 + \sum_{n=0}^{\infty} f_n^{pq} \pi_n \right)^{1/q},$$

where $\tilde{L} = 4\tilde{M}^2 (1 - \rho^{1/r})^{-1}$, while $r = p/(p-2)$ and $\tilde{M} = M + \sum_{n=0}^{\infty} f_n^p \pi_n$.

The proof is given in (Tadić, 2001*a*).

4. EXAMPLES

In this section, the main results of the paper are illustrated with examples related to random coefficient autoregression models and $M/G/1$ queues.

An example where $\{X_n\}_{n \geq 0}$ is the state of a random coefficient autoregression model is considered now. Let

$$X_{n+1} = (A + A_{n+1}) X_n + U_{n+1}, \quad n \geq 0,$$

where $A \in R^{d' \times d'}$, X_0 is an $R^{d'}$-valued deterministic variable and $\{A_n\}_{n \geq 0}$, $\{U_n\}_{n \geq 0}$ are jointly independent sequences of $R^{d' \times d'}$- and $R^{d'}$-valued (respectively) i.i.d. random variables.

As an immediate consequence of (Meyn and Tweedie, 1993, Theorem 16.5.1), the following lemma is obtained.

Lemma 1. Suppose that $E\|A_0\|^2 < \infty$ and $E\|U_0\|^2 < \infty$. Moreover, suppose that $E(A_0) = 0$, $E(U_0) = 0$ and the eigenvalues of $A \otimes A + E(A_0 \otimes A_0)$ lie in the interior of the unit circle. Furthermore, suppose that the distribution of (A_0, U_0) has an everywhere positive density with respect to the Lebesgue measure. Then, $\{X_n\}_{n \geq 0}$ has a unique invariant measure $\pi(\cdot)$ and there exist constants $M \in [1, \infty)$, $\rho \in (0,1)$ (depending on the distribution of (A_0, U_0) only) such that $\int \|x\|^2 \pi(dx) < \infty$ and

$$\left\| \int \varphi(x') (P_n - \pi)(x, dx') \right\|$$
$$\leq M \rho^n \|x\|^2, \quad \forall x \in R^{d'}, \; n \geq 0,$$

for any Borel-measurable function $\varphi : R^{d'} \to R$ satisfying $0 \leq \varphi(x) \leq \|x\|^2$, $\forall x \in R^{d'}$.

Then, as a direct consequence of Theorem 2, the following result is obtained.

Corollary 1. Let A1 and the conditions of Lemma 1 hold. Suppose that there exists a constant $N \in R^+$ such that $|c(x,x')| \leq N$ and $\|\phi(x)\| \leq N$, $\forall x, x' \in R^{d'}$. Moreover, suppose that

$$\int \phi(x) \phi^T(x) \pi(dx)$$
$$-2^{-1} \gamma^{-1} (1-\alpha)^{-1} (1 - \alpha\lambda) I$$

is positive definite. Then, there exists a constant $C \in R^+$ (depending on γ, N and the distribution of (A_0, U_0) only) such that (8) is satisfied (θ_* is defined in the statement of Theorem 1).

An example related to $M/G/1$ queues is considered now (the same example has been used in (Tsitsiklis and Roy, 1997)). Let X_{n+1} is the number of customers in a $M/G/1$ queue immediately after the completion of the service of the n-th customer. Let λ be the mean of the interarrival times of the customers in the queue, while $\mu(\cdot)$ is the distribution of their service times (for details on $M/G/1$ queues see e.g., (Asmussen, 1987), (Meyn and Tweedie, 1993) and references cited therein).

Lemma 2. Suppose that there exists a constant $r \in R^+$ such that $\int \exp(rt) \mu(dt) < \infty$. Moreover, suppose that $\int t \mu(dt) < \lambda$. Then, $\{X_n\}_{n \geq 0}$ has a unique invariant distribution $\pi = \{\pi_n\}_{n \geq 0}$ and there exist constants $M \in [1, \infty)$, $\rho \in (0,1)$ and $c \in R^+$ such that $\sum_{n=0}^{\infty} g_n \pi_n < \infty$ and

$$\left|\sum_{j=0}^{\infty}\varphi_j(p_{ij}^n-\pi_j)\right|\leq M\rho^n g_i,\quad n,i\geq 0,$$

for any sequence $\{\varphi_n\}_{n\geq 0}$ satisfying $0\leq\varphi_n\leq g_n$, $n\geq 0$, where $g_n=\exp(cn)$, $n\geq 0$.

Then, as a direct consequence of Theorem 3, the following result is obtained.

Corollary 2. Let A1 and the conditions of Lemma 2 hold. Suppose that there exist constants $N\in[1,\infty)$ and $s\in R^+$ such that $\|\phi(n)\|\leq N(1+n^s)$, $n\geq 0$, and

$$|c(m,n)|\leq N(1+m^s+n^s),\quad m,n\geq 0.$$

Moreover, suppose that

$$\sum_{n=0}^{\infty}\phi(n)\phi^T(n)\pi_n$$
$$-2^{-1}\gamma^{-1}(1-\alpha)^{-1}(1-\alpha\lambda)I$$

is positive definite. Then, there exists a constant $C\in R^+$ (depending on γ, λ, N, s and $\mu(\cdot)$ only) such that (8) is satisfied ($\theta_*=-A_*^{-1}b_*$, while A_*, b_* are defined in the statement of Theorem 3).

5. CONCLUSION

In this paper, the almost sure convergence rate of temporal-difference learning algorithms has been analyzed. The analysis has been carried out for the case of discounted cost function associated with a Markov chain with a finite dimensional state-space. Under mild conditions, it has been shown that these algorithms converge at the rate $O(n^{-1/2}(\log\log n)^{1/2})$ almost surely. Since $O(n^{-1/2}(\log\log n)^{1/2})$ characterizes the rate of convergence in the law of iterated logarithm, the obtained results could be considered as the same law for temporal-difference learning algorithms. For the same reason, the obtained convergence rate is probably the least conservative result of this kind. Furthermore, to the best of the author's knowledge, there does not exist a similar result in the available literature on reinforcement learning. The main results of the paper have been illustrated with examples related to random coefficient autoregression models and $M/G/1$ queues. This paper is a continuation of the author's work presented in (Tadić, 2001*b*).

6. REFERENCES

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

Bertsekas, D. P. and J. N. Tsitsiklis (1996). Neuro-dynamic programming.

Dayan, P. D. (1992). The convergence of $TD(\lambda)$ for general λ. *Machine Learning* **8**, 341–362.

Dayan, P. D. and T. J. Sejnowski (1994). $TD(\lambda)$ converges with probability 1. *Machine Learning* **14**, 295–301.

Goodwin, G. C. and K. S. Sin (1984). *Adaptive Filtering, Prediction and Control*. Prentice Hall.

Jaakola, T., M. I. Jordan and S. P. Singh (1994). On the convergence of stochastic iterative dynamic programming. *Neural Computation* **6**, 1185–1201.

Meyn, S. P. and R. L. Tweedie (1993). *Markov Chains and Stochastic Stability*. Springer Verlag.

Solo, V. and X. Kong (1995). *Adaptive Signal Processing Algorithms: Stability and Performance*. Prentice-Hall.

Sutton, R. S. (1988). Learning to predict by the method of temporal-difference learning. *Machine Learning* **3**, 9–44.

Sutton, R. S. and A. G. Barto (1998). *Reinforcement Learning: An Introduction*.

Tadić, V. B. (2001*a*). On the almost sure rate of convergence of temporal-difference learning algorithms with linear function approximation. Technical report. Department of Electrical and Electronic Engineering, The University of Melbourne.

Tadić, V. B. (2001*b*). On the convergence of temporal-difference learning with linear function approximation. *Machine Learning* **42**, 241–267.

Tsitsiklis, J. N. and B. Van Roy (1997). An analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control* **42**, 674–690.