Convergence of Least Squares Temporal Difference Methods Under General Conditions

Huizhen Yu JANEY.YU@CS.HELSINKI.FI
Department of Computer Science, University of Helsinki, Finland

Abstract

We consider approximate policy evaluation for finite state and action Markov decision processes (MDP) in the off-policy learning context and with the simulation-based least squares temporal difference algorithm, LSTD(λ). We establish for the discounted cost criterion that the off-policy LSTD(λ) converges almost surely under mild, minimal conditions. We also analyze other convergence and boundedness properties of the iterates involved in the algorithm, and based on them, we suggest a modification in its practical implementation. Our analysis uses theories of both finite space Markov chains and Markov chains on topological spaces.

1. Overview

We consider approximate policy evaluation for finite state and action Markov decision processes (MDP) in an exploration-enhanced learning context, called "off-policy" learning. In this context, we employ a certain policy called the "behavior policy" to adequately explore the state and action space, and using the observations of costs and transitions generated under the behavior policy, we may approximately evaluate any suitable "target policy" of interest. This differs from the standard policy evaluation case – "on-policy" learning – where the behavior policy always coincides with the policy to be evaluated. The dichotomy between the off-policy and on-policy learning stems from the exploration-exploitation tradeoff in practical model-free/simulation-based methods for policy search. With their flexibility, off-policy methods form an important part of the model-free learning methodology (Sutton & Barto, 1998) and have been suggested as important simulation-based methods for large-scale dynamic programming (Glynn & Iglehart, 1989).

The algorithm we consider in this paper, the off-policy least squares temporal difference (TD) algorithm, LSTD(λ), is one of the exploration-enhanced methods for policy evaluation. More specifically, we consider discounted total cost problems with discount factor $\alpha < 1$. We evaluate the so-called Q-factors of the target policy, which are essential for policy iteration, and which are simply the costs of the policy in an equivalent MDP that has as its states the joint state-action pairs of the original MDP[1] [see e.g., (Bertsekas & Tsitsiklis, 1996)]. This MDP will be the focus of our discussion, and we will refer to Q-factors as costs for simplicity. Let $\mathcal{I} = \{1, 2, \ldots, n\}$ be the set of state-action pairs indexed by integers from 1 to n. We assume that the behavior policy induces an irreducible Markov chain on the space $\mathcal{I}$ of state-action pairs with transition matrix P, and that the target policy we aim to evaluate would induce a Markov chain with transition matrix Q. We require naturally that for all states, possible actions of the target policy are also possible actions of the behavior policy. This condition, denoted $Q \prec P$, can be written as

$$p_{ij} = 0 \quad \Rightarrow \quad q_{ij} = 0, \quad i, j \in \mathcal{I}. \tag{1}$$

Let g be the vector of expected one-stage costs $g(i)$ under the target policy. The cost J^* of the target policy satisfies the Bellman equation

$$J^* = g + \alpha Q J^*. \tag{2}$$

With the temporal difference methods (Sutton, 1988) [see also the books (Bertsekas & Tsitsiklis, 1996; Sutton & Barto, 1998; Bertsekas, 2007; Meyn, 2007)], we

Appearing in *Proceedings of the 27th International Conference on Machine Learning*, Haifa, Israel, 2010. Copyright 2010 by the author(s)/owner(s).

[1] The equivalent MDP on the space of state-action pairs can be defined as follows. Consider any two state-action pairs $i = (s, u)$ and $j = (\hat{s}, v)$. Suppose a transition from s to $\hat{s}$ under action u incurs the cost $c(s, u, \hat{s})$ in the original MDP. Then the cost of transition from i to j in the equivalent MDP can be defined as $g(i, j) = c(s, u, \hat{s})$. The transition probability from i to j under a policy which takes action v at state $\hat{s}$ with probability $\mu(v \mid \hat{s})$ is given by $p(\hat{s} \mid s, u)\mu(v \mid \hat{s})$.

approximate J^* by the solution of a projected multistep Bellman equation

$$J = \Pi T^{(\lambda)}(J) \tag{3}$$

involving a multistep Bellman operator $T^{(\lambda)}$ parametrized by $\lambda \in [0,1]$, whose exact form will be given later. Here Π is a linear operator of projection onto an approximation subspace $\{\Phi r \mid r \in \Re^{n_r}\} \subset \Re^n$ with respect to a weighted Euclidean norm, where Φ is an $n \times n_r$ matrix whose columns span the approximation subspace and whose rows are often called "features" of states/actions. In the case considered here, we take the weights in the projection norm to be the steady-state probabilities of the Markov chain induced by the behavior policy. The projected Bellman equation (3) is equivalent to a low dimensional equation on $\Re^{n_r}$, and its solution Φr^* (when it exists) is used to approximate the cost J^* of the target policy.

The off-policy LSTD(λ) algorithm that we will analyze aims to construct the low-dimensional equivalent of the projected equation (3) by using observations generated under the behavior policy. The algorithm takes into account the discrepancies between the behavior and the target policies by properly weighting the observations. The technique is based on importance sampling, which is widely used in dynamic programming and reinforcement learning contexts [see e.g., (Glynn & Iglehart, 1989; Sutton & Barto, 1998; Precup et al., 2001; Ahamed et al., 2006)]. The off-policy LSTD(λ) algorithm we will analyze was first given by (Bertsekas & Yu, 2009) in the general context of approximating solutions of linear systems of equations. The form of the algorithm bears similarities to other off-policy TD(λ) algorithms, e.g., the episodic off-policy TD(λ) (Precup et al., 2001), as well as to the on-policy LSTD(λ) counterpart (Bradtke & Barto, 1996; Boyan, 1999). The algorithm can be described as follows.

Let $(i_0, i_1, \ldots)$ be an infinitely long state trajectory of the Markov chain with transition matrix P, generated under the behavior policy. Let $\phi(i)$ denote the transpose of the ith row vector of matrix Φ, and let $g(i,j)$ be the per-stage cost of transition from state i to j. The off-policy LSTD(λ) method computes low-dimensional vector iterates Z_t, b_t and matrix iterates C_t as follows: with (z_0, b_0, C_0) being the initial condition, for $t \geq 1$,

$$Z_t = \lambda\alpha \tfrac{q_{i_{t-1}i_t}}{p_{i_{t-1}i_t}} \cdot Z_{t-1} + \phi(i_t), \tag{4}$$

$$b_t = (1 - \tfrac{1}{t+1}) b_{t-1} + \tfrac{1}{t+1} Z_t g(i_t, i_{t+1}), \tag{5}$$

$$C_t = (1 - \tfrac{1}{t+1}) C_{t-1} + \tfrac{1}{t+1} Z_t \left(\alpha \tfrac{q_{i_t i_{t+1}}}{p_{i_t i_{t+1}}} \cdot \phi(i_{t+1}) - \phi(i_t) \right)'. \tag{6}$$

The vector b_t and matrix C_t aim to approximate the quantities defining the projected Bellman equation (3). The solution r_t of the equation $C_t r + b_t = 0$ is used to give Φr_t as an approximation of J^* at time t.[2]

The on-policy case corresponds to the special case $P = Q$. There, all the ratios $\frac{q_{i_{t-1}i_t}}{p_{i_{t-1}i_t}}$ appeared above in Z_t and C_t become 1, and the algorithm reduces to the on-policy LSTD algorithm as first given by (Bradtke & Barto, 1996) for $\lambda = 0$ and (Boyan, 1999) for $\lambda \in [0,1]$.

In the off-policy case, a property of practical importance is that the ratios $\frac{q_{ij}}{p_{ij}}$ are determined by the ratios between the action probabilities of the target and the behavior policies (as can be seen from Footnote 1); therefore, they need not be stored and can be calculated on-line in the algorithm.

A full convergence analysis of the off-policy LSTD(λ) algorithm does not exist in the literature, to our knowledge. The almost sure convergence of the algorithm (i.e., convergence with probability one) in special cases has been studied. A proof for the on-policy case can be found in (Nedić & Bertsekas, 2003). A proof for the off-policy case under the assumption that $\lambda\alpha \max_{i,j} \frac{q_{ij}}{p_{ij}} < 1$ (with 0/0 treated as 0) is given in (Bertsekas & Yu, 2009); this covers the on-policy case as well as the off-policy LSTD(λ) for λ close or equal to 0, but for a general value of λ, the condition is too stringent on either the target or the behavior policy. Note that the case with a general value of λ is important in practice, because using a large value of λ not only improves the quality of the approximation from the projected Bellman equation, but also avoids potential pathologies regarding the existence of solution of the equation (as λ approaches 1, $\Pi T^{(\lambda)}$ becomes a contraction mapping, ensuring the existence of a unique solution).

In this work, we establish the almost sure convergence of the sequences $\{b_t\}, \{C_t\}$, as well as their convergence in the first mean, under the general conditions given at the beginning, namely, the irreducibility of P and $Q \prec P$. Our results imply that the off-policy LSTD(λ) solution Φr_t converges to the solution Φr^* of the projected Bellman equation (3) almost surely, whenever Eq. (3) has a unique solution (if (3) has multiple solutions, then any limit point of $\{\Phi r_t\}$ is a solution of it.) As will be seen later, the convergence of $\{b_t\}, \{C_t\}$ in the first mean (Theorem 1) can be established using arguments based on finite space Markov chains, while the proof of the almost sure

[2] In this paper we do not discuss the exceptional case where $C_t r + b_t = 0$ does not have a solution. Our focus will be on the asymptotic properties of the sequence of equations $C_t r + b_t = 0$ themselves.

convergence is not so straightforward and finite space Markov chains-based arguments are no longer sufficient. In contrast to the relative simplicity of the on-policy case, the technical complexity here is partly due to the fact that the sequence $\{Z_t\}$ cannot be bounded a priori. Indeed, we can show that for the off-policy case, in fairly common situations, $\{Z_t\}$ is almost surely unbounded (Prop. 2). Neither does it seem likely that without imposing extra conditions, the sequence of Z_t can have bounded variance. Nevertheless, these do not preclude the almost sure convergence of the off-policy LSTD(λ) algorithm, as we will show.

It is worth mentioning that the study of the almost sure convergence of the off-policy LSTD(λ) is not solely of theoretic interest. Various TD algorithms other than LSTD(λ) use the same approximations b_t, C_t to build approximating models [e.g., preconditioned TD(λ) (Yao & Liu, 2008)] or fixed point mappings [e.g., LSPE(λ), see (Bertsekas & Yu, 2009); and (Bertsekas, 2009)] needed in the algorithms. Therefore in the off-policy case, the asymptotic behavior of these algorithms on a sample path depends on the mode of convergence of $\{b_t\}, \{C_t\}$, and so does the interpretation of the approximate solutions generated by these algorithms. For algorithms whose convergence relies on the contraction property of mappings, (for instance, LSPE(λ)), the almost sure convergence of $\{b_t\}, \{C_t\}$ on every sample path is critical. Furthermore, the mode of convergence of the off-policy LSTD(λ) is also relevant for understanding the behavior of other off-policy TD algorithms, e.g., the non-episodic off-policy TD(λ) and episodic off-policy TD(λ) with very long episodes, which, although not computing directly b_t, C_t, implicitly depend on the convergence properties of $\{b_t\}, \{C_t\}$.

To establish the almost sure convergence of $\{b_t\}, \{C_t\}$, we will study the Markov chain $\{(i_t, Z_t)\}$ on the topological space $\mathcal{I} \times \Re^{n_r}$. Again, the lack of boundedness condition on Z_t makes it difficult to argue the existence of an invariant probability measure by constructing explicitly the form of Z_t for a stationary Markov chain $\{(i_t, Z_t)\}$ in the limit, as can be done in the on-policy case (Tsitsiklis & Van Roy, 1997). We will use the theory of e-chains (Meyn & Tweedie, 2009), which concerns topological space Markov chains with equicontinuous transition kernels, to establish two main results: (i) the Markov chain $\{(i_t, Z_t)\}$ has a unique invariant probability measure and is ergodic (Theorem 2), and (ii) the almost sure convergence of $\{b_t\}, \{C_t\}$ (and hence the almost sure convergence of the off-policy LSTD(λ) algorithm) (Theorem 3). The first ergodicity result is indeed stronger than what is needed to show (ii); but it sheds light on the nature of the TD iterates and provides a basis for analyzing other off-policy TD(λ) algorithms in the future.

Let us also mention the ODE proof approach: relevant here is the mean-ODE method [see e.g., (Kushner & Yin, 2003; Borkar, 2008)], which, however, requires the verification of conditions that in our case would be tantamount to the almost sure convergence conclusion we want to establish.

The paper is organized as follows. We specify notation and definitions in Section 2. We present our main results and outline their key proof arguments in Section 3. We then give proof details for two of the main theorems, namely, Theorems 1 and 3, in Section 4. Complete proofs and further discussion can be found in (Yu, 2010).

2. Notation and Specifications

The projected Bellman equation (3) associated with TD(λ) methods is a projected version of a multistep Bellman equation parametrized by $\lambda \in [0, 1]$. In particular, let T be the Bellman operator $T(J) = g + \alpha Q J$ for all $J \in \Re^n$. The mapping $T^{(\lambda)}$ in Eq. (3) is defined by

$$T^{(\lambda)} = (1-\lambda) \sum_{m=0}^{\infty} \lambda^m T^{m+1}, \quad \lambda \in [0, 1);$$
$$T^{(1)}(J) = \lim_{\lambda \to 1} T^{(\lambda)}(J), \quad \forall J \in \Re^n.$$

Let Ξ_p denote the diagonal matrix with the diagonal elements being the steady-state probabilities of the Markov chain with transition matrix P, induced by the behavior policy. Equation (3) is equivalent to the low dimensional equation on $\Re^{n_r}$,

$$\Phi' \Xi_p \Phi r = \Phi' \Xi_p T^{(\lambda)}(\Phi r)$$
$$= \Phi' \Xi_p \sum_{m=0}^{\infty} \lambda^m (\alpha Q)^m \big(g + (1-\lambda)\alpha Q \Phi r\big).$$

By rearranging terms, it can be written as

$$\bar{C} r + \bar{b} = 0, \tag{7}$$

where $\bar{b}$ is an $n_r \times 1$ vector and $\bar{C}$ an $n_r \times n_r$ matrix, given by

$$\bar{b} = \Phi' \Xi_p \sum_{m=0}^{\infty} \lambda^m (\alpha Q)^m g, \tag{8}$$

$$\bar{C} = \Phi' \Xi_p \sum_{m=0}^{\infty} \lambda^m (\alpha Q)^m (\alpha Q - I) \Phi. \tag{9}$$

The iterates b_t, C_t in the off-policy LSTD(λ) [Eqs. (5), (6)] aim to approximate $\bar{b}, \bar{C}$ respectively, (which define the projected equation (7) and equivalently (3)).

Their convergence to $\bar{b}, \bar{C}$, respectively, in any relevant mode, is what we want to show.

In the rest of the paper, we use i_t to denote the random state variable at time t and $\bar{i}$ or i^* to denote specific states. To simplify notation, we denote $\beta = \lambda\alpha$ and study iterates of the form

$$Z_t = \beta \frac{q_{i_{t-1}i_t}}{p_{i_{t-1}i_t}} \cdot Z_{t-1} + \phi(i_t), \tag{10}$$

$$G_t = (1-\gamma_t)G_{t-1} + \gamma_t Z_t \psi(i_t, i_{t+1})', \tag{11}$$

with $\beta < 1$, (z_0, G_0) being the initial condition, and $\{\gamma_t\}$ being a stepsize sequence. The correspondence between iterates G_t and the vectors b_t and matrices C_t in LSTD(λ) [cf. Eqs. (5), (6)] is as follows: with $\gamma_t = 1/(t+1)$,

$$G_t = \begin{cases} b_t, & \text{if } \psi(i_t, i_{t+1}) = g(i_t, i_{t+1}), \\ C_t, & \text{if } \psi(i_t, i_{t+1}) = \alpha \frac{q_{i_t i_{t+1}}}{p_{i_t i_{t+1}}} \cdot \phi(i_{t+1}) - \phi(i_t). \end{cases} \tag{12}$$

We want to show that G_t converges, in any relevant mode, to the constant vector/matrix

$$G^* = \Phi' \Xi_p \Big(\sum_{m=0}^{\infty} \beta^m Q^m \Big) \Psi, \tag{13}$$

where the vector/matrix Ψ is given row-wisely by

$$\Psi = \begin{bmatrix} \bar{\psi}(1)' \\ \bar{\psi}(2)' \\ \cdots \\ \bar{\psi}(n)' \end{bmatrix}, \quad \text{with} \quad \bar{\psi}(i) = E\big[\psi(i_0, i_1) \mid i_0 = i\big].$$

Here and in what follows E denotes the expectation with respect to the distribution of the Markov chain $\{i_t\}$ with transition matrix P. As can be seen, corresponding to the two choices of ψ in the expression of G_t [Eq. (12)], Ψ equals g or $(\alpha Q - I)\Phi$, and G^* equals $\bar{b}$ or $\bar{C}$, respectively [cf. Eqs. (8)-(9)].

We make two assumptions, one on the transition matrices P and Q, as mentioned at the beginning of Section 1, and the other on the stepsize sequence.

Assumption 1. *The Markov chain $\{i_t\}$ with transition matrix P is irreducible, and $Q \prec P$ in the sense of Eq. (1).*

Assumption 2. *The sequence of stepsizes γ_t is deterministic and satisfies $\gamma_t \in (0, 1]$,*

$$\sum_t \gamma_t = \infty, \quad \sum_t \gamma_t^2 < \infty, \quad \limsup_{t\to\infty} \frac{\gamma_t}{\gamma_{t-1}} < \infty. \tag{14}$$

Such sequences of γ_t include $1/t$, $t^{-\nu}, \nu \in (0, 1]$, for instance. When conclusions hold for a specific sequence $\{\gamma_t\}$, such as $\gamma_t = 1/t$, we will state them explicitly.

3. Main Results

We pursue separately two lines of analysis, one based on properties of the finite space Markov chain $\{i_t\}$ and the other based on properties of the topological space Markov chain $\{(i_t, Z_t)\}$. In this section we overview our main results and outline key proof arguments.

Throughout the paper, let $\|\cdot\|$ denote the F-norm $\|V\| = \max_{i,j} |V_{ij}|$ for a matrix V, and the infinity norm $\|V\| = \max_i |V_i|$ for a vector V, in particular, $\|V\| = |V|$ for a scalar V. Let "a.s." stand for almost sure convergence.

3.1. Analysis Based on Finite Space Markov Chains

First, it is not difficult to show that G_t converges in mean. This implies immediately that the LSTD(λ) solution r_t converges in probability to the solution r^* of Eq. (7) when the latter exists and is unique.

Theorem 1. *Under Assumption 1, for each initial condition z_0, $\sup_t E\|Z_t\| \le \frac{c}{1-\beta}$ where $c = \max\{\|z_0\|, \max_i \|\phi(i)\|\}$. Under Assumptions 1 and 2, for each initial condition (z_0, G_0),*

$$\lim_{t\to\infty} E\|G_t - G^*\| = 0.$$

Next, based essentially on a zero-one law for tail events[3] of Markov chains [see (Breiman, 1992), Theorem 7.43], we can show the following result.

Proposition 1. *Under Assumptions 1 and 2, for each initial condition (z_0, G_0) and any $\mathcal{E}$ of the following events, either $\mathbf{P}(\mathcal{E}) = 0$ or $\mathbf{P}(\mathcal{E}) = 1$:*
(i) $\mathcal{E} = \{\lim_{t\to\infty} G_t$ exists, and $\sup_t \|Z_t\| < \infty\}$;
(ii) $\mathcal{E} = \{\sup_t \|Z_t\| < \infty\}$;
(iii) $\mathcal{E} = \{\lim_{t\to\infty} \gamma_t Z_t = 0\}$;
(iv) $\mathcal{E} = \{\lim_{t\to\infty} G_t$ exists$\}$.

Theorem 1 and Prop. 1(iv) together have the following implication on the convergence of G_t. According to Prop. 1(iv), for the event $\mathcal{E} = \{\lim_{t\to\infty} G_t$ exists$\}$, we have $\mathbf{P}(\mathcal{E}) = 1$ or 0. Suppose $\mathbf{P}(\mathcal{E}) = 1$. Then $G_t \overset{a.s.}{\to} G$ for some random variable G. Since Theorem 1 implies $G_t \to G^*$ in probability, which further implies the convergence of a subsequence $G_{t_k} \overset{a.s.}{\to} G^*$, we must have $G = G^*$ a.s.; therefore $G_t \overset{a.s.}{\to} G^*$. Suppose now $\mathbf{P}(\mathcal{E}) = 0$. Then we only have the convergence of G_t to G^* in probability implied by Theorem 1, and with probability 1, on every sample path G_t does not converge. In Section 3.2, we will rule out the sec-

[3] An event is called a tail event of a process $\{X_t\}$ if it is determined by $X_t, t \ge k$ for any k [see e.g., (Breiman, 1992), Def. 3.10].

ond case for the stepsize sequence $\gamma_t = 1/(t+1)$, using the line of analysis based on the Markov chain $\{(i_t, Z_t)\}$.

We discuss other implications of Prop. 1, contrasting the off-policy case with the standard, on-policy case where $P = Q$. In the latter case, events (i) and (ii) in Prop. 1 both have probability one; event (ii) – the boundedness of Z_t – is true by the definition of Z_t. By contrast, in the off-policy case, under seemingly fairly common situations (as we show below), Z_t is almost surely unbounded, and consequently, events (i) and (ii) have probability zero. While the unboundedness of Z_t may sound disquieting, note that it is $\gamma_t Z_t \overset{a.s.}{\to} 0$, the event shown in (iii), and not the boundedness of Z_t, that is necessary for the almost sure convergence of G_t. In other words, $\{\lim_{t\to\infty} G_t \text{ exists}\} \subset \{\lim_{t\to\infty} \gamma_t Z_t = 0\}$.[4]

For practical implementation, however, the unboundedness of Z_t can be unwieldy. This suggests that in practice, instead of iterating Z_t directly, we equivalently iterate $\gamma_t Z_t$ via

$$\gamma_t Z_t = \beta \tfrac{q_{i_{t-1} i_t}}{p_{i_{t-1} i_t}} \cdot \tfrac{\gamma_t}{\gamma_{t-1}} \cdot (\gamma_{t-1} Z_{t-1}) + \gamma_t \phi(i_t), \quad (15)$$

whenever the magnitude of Z_t becomes intolerably large. That $\gamma_t Z_t \overset{a.s.}{\to} 0$ when $\gamma_t = 1/(t+1)$ will be implied by the almost sure convergence of G_t we later establish.

We now demonstrate by construction that in seemingly fairly common situations, Z_t is almost surely unbounded. Our construction is based on a consequence of the extended Borel-Cantelli lemma [(Breiman, 1992), Problem 5.9, p. 97], which says that for any process $\{X_t, t \geq 0\}$ with X_t taking values in S, and any measurable subsets A, B of S, if for all t,

$$\mathbf{P}\left(\exists s, s > t, X_s \in B \mid X_t, X_{t-1}, \ldots, X_0\right) \geq \delta > 0$$

on $\{X_t \in A\}$, then

$$\{X_t \in A \text{ } i.o.\} \subset \{X_t \in B \text{ } i.o.\} \quad a.s.$$

Here, "i.o." stands for "infinitely often," and "a.s." means that the set-inclusion relation holds after excluding a set of zero probability. In our context, this result together with the zero-one probability statement for the event $\{\sup_t \|Z_t\| < \infty\}$ in Prop. 1(ii) has the following implications.

[4] This can be seen from the fact that

$$G_t - G_{t-1} = -\gamma_t G_{t-1} + \gamma_t Z_t \psi(i_t, i_{t+1})',$$

and $\gamma_t \to 0$ as $t \to \infty$.

Denote by $Z_{t,j}$ and $\phi_j(i_t)$ the jth element of the vector Z_t and $\phi(i_t)$, respectively. Consider a cycle configuration of states $(\bar{i}_1, \bar{i}_2, \ldots, \bar{i}_m, \bar{i}_1)$ with the following three properties:

(a) it occurs with positive probability:

$$p_{\bar{i}_1 \bar{i}_2} p_{\bar{i}_2 \bar{i}_3} \cdots p_{\bar{i}_m \bar{i}_1} > 0; \quad (16)$$

(b) it has an amplifying effect in the sense that

$$\beta^m \tfrac{q_{\bar{i}_1 \bar{i}_2}}{p_{\bar{i}_1 \bar{i}_2}} \tfrac{q_{\bar{i}_2 \bar{i}_3}}{p_{\bar{i}_2 \bar{i}_3}} \cdots \tfrac{q_{\bar{i}_m \bar{i}_1}}{p_{\bar{i}_m \bar{i}_1}} > 1; \quad (17)$$

(c) for some $\bar{j}$, the $\bar{j}$th elements of $\phi(\bar{i}_1), \ldots, \phi(\bar{i}_m)$ have the same sign and their sum is non-zero: i.e., either for all $k = 1, \ldots, m$,

$$\phi_{\bar{j}}(\bar{i}_k) \geq 0, \quad \text{with} \quad \phi_{\bar{j}}(\bar{i}_k) > 0 \text{ for some } k; \quad (18)$$

or for all $k = 1, \ldots, m$,

$$\phi_{\bar{j}}(\bar{i}_k) \leq 0, \quad \text{with} \quad \phi_{\bar{j}}(\bar{i}_k) < 0 \text{ for some } k. \quad (19)$$

Proposition 2. *Suppose there exists a cycle configuration of states $(\bar{i}_1, \bar{i}_2, \ldots, \bar{i}_m, \bar{i}_1)$ possessing properties (a)-(c) above, and $\bar{j}$ is as in (c). Then there exists a constant ν, which depends on the cycle and is negative (respectively, positive) if Eq. (18) (respectively, Eq. (19)) holds in (c), and if for some neighborhood $\mathcal{O}(\nu)$ of ν, $\mathbf{P}(i_t = \bar{i}_1, Z_{t,\bar{j}} \notin \mathcal{O}(\nu) \text{ } i.o.) = 1$, then $\mathbf{P}(\sup_t \|Z_t\| = \infty) = 1$.*

We remark that the extra technical condition $\mathbf{P}(i_t = \bar{i}_1, Z_{t,\bar{j}} \notin \mathcal{O}(\nu) \text{ } i.o.) = 1$ in Prop. 2 is nonrestrictive. The opposite case – that on a set with non-negligible probability, $Z_{t,\bar{j}}$ eventually always lies arbitrarily close to ν whenver $i_t = \bar{i}_1$ – seems unlikely to occur except in highly contrived examples.

3.2. Analysis Based on Topological Space Markov Chains

To establish the almost sure convergence of G_t to G^*, we consider the Markov chain $\{(i_t, Z_t), t \geq 0\}$ on the topological space $S = \mathcal{I} \times \Re^{n_r}$ with product topology (discrete topology on $\mathcal{I}$ and usual topology on $\Re^{n_r}$). We show that $\{(i_t, Z_t)\}$ can be related to a type of Markov chains, called e-chains, whose transition kernel functions possess a certain equicontinuity property (Meyn & Tweedie, 2009). Central to our proof is the analysis of the differences in the processes $\{Z_t\}$ for different initial conditions z_0 and the same sample path of $\{i_t\}$. As can already be seen from Eq. (10), for two such processes $\{Z_t\}, \{\hat{Z}_t\}$ with initial conditions $z_0, \hat{z}_0$, respectively, their differences satisfy the simple recursion:

$$Z_t - \hat{Z}_t = \beta \tfrac{q_{i_{t-1} i_t}}{p_{i_{t-1} i_t}} \cdot (Z_{t-1} - \hat{Z}_{t-1}), \quad (20)$$

which implies that the difference sequence converges almost surely to zero (Lemma 1). Using more careful characterizations of such difference sequences together with the first part of Theorem 1, we can establish the various properties needed for applying the law of large numbers (LLN) for e-chains (Meyn & Tweedie, 2009) and show that the chain $\{(i_t, Z_t)\}$ is ergodic.

Our conclusions are summarized in the following two theorems. (Definitions of related terminologies and detailed analysis can be found in (Yu, 2010).)

Theorem 2. *Under Assumption 1, the Markov chain $\{(i_t, Z_t)\}$ is an e-chain with a unique invariant probability measure π, and almost surely, for each initial condition, the sequence of occupation measures $\{\mu_t\}$ on S converges weakly to π, where μ_t is defined by*

$$\mu_t(A) = \frac{1}{t}\sum_{k=1}^{t} \mathbf{1}_A(i_k, Z_k)$$

for all Borel-measurable subsets A of S, and $\mathbf{1}_A$ denotes the indicator function for the set A.

Let E_π denote expectation with respect to the stationary distribution $\mathbf{P}_\pi$ of the Markov chain $\{(i_t, Z_t)\}$ with initial distribution π.

Theorem 3. *Under Assumption 1, $G^* = E_\pi[Z_0\psi(i_0, i_1)']$, and with stepsize $\gamma_t = 1/(t+1)$, for each initial condition (z_0, G_0), $G_t \overset{a.s.}{\to} G^*$.*

Theorem 3 implies that for each initial condition, the sequence $\{\Phi r_t\}$ computed by the off-policy LSTD(λ) algorithm converges almost surely to the solution Φr^* of the projected Bellman equation (3) when the latter exists and is unique.

4. Details of Analysis

Due to space limit, we give the proof for Theorem 1 on the convergence of LSTD(λ) iterates b_t, C_t in the first mean, and a partial proof for Theorem 3 on the almost sure convergence of LSTD(λ). Complete proofs for all theorems in Section 3 can be found in (Yu, 2010).

We denote by L_s^t the product of ratios of transition probabilities along a segment of the state trajectory, $(i_s, i_{s+1}, \ldots, i_t)$:

$$L_s^t = \frac{q_{i_s i_{s+1}}}{p_{i_s i_{s+1}}} \cdot \frac{q_{i_{s+1} i_{s+2}}}{p_{i_{s+1} i_{s+2}}} \cdots \frac{q_{i_{t-1} i_t}}{p_{i_{t-1} i_t}}. \tag{21}$$

Define $L_t^t = 1$. Note that for $s \le s' \le t$, $L_s^{s'} L_{s'}^t = L_s^t$ and

$$E[\, L_s^t \mid i_s] = 1.$$

4.1. Proof of Theorem 1

To show the first part of Theorem 1, we have by Eqs. (10) and (21),

$$Z_t = \beta^t L_0^t z_0 + \sum_{m=0}^{t-1} \beta^m L_{t-m}^t \phi(i_{t-m}),$$

so, with $c = \max\{\|z_0\|, \max_i \|\phi(i)\|\}$,

$$E\|Z_t\| \le c\, E\Big[\beta^t L_0^t + \sum_{m=0}^{t-1} \beta^m L_{t-m}^t\Big]$$
$$= c \sum_{m=0}^{t} \beta^m \le \frac{c}{1-\beta}.$$

To prove the second part of theorem on the convergence of G_t to G^* in the first mean, we first consider another process $(\tilde{Z}_{t,T}, \widetilde{G}_{t,T})$ on the same probability space, and apply the LLN for a finite space irreducible Markov chain to $\widetilde{G}_{t,T}$. We then relate $(\tilde{Z}_{t,T}, \widetilde{G}_{t,T})$ to (Z_t, G_t). In particular, for a positive integer T, define

$$\tilde{Z}_{t,T} = Z_t, \quad t \le T; \qquad \widetilde{G}_{0,T} = G_0,$$

and define for $t > T$,

$$\tilde{Z}_{t,T} = \phi(i_t) + \beta L_{t-1}^t \phi(i_{t-1}) + \cdots + \beta^T L_{t-T}^t \cdot \phi(i_{t-T}); \tag{22}$$

$$\widetilde{G}_{t,T} = (1-\gamma_t)\widetilde{G}_{t-1,T} + \gamma_t \tilde{Z}_{t,T}\, \psi(i_t, i_{t+1})', \quad t \ge 1. \tag{23}$$

Note that for $t \le T$, $\widetilde{G}_{t,T} = G_t$ because $\tilde{Z}_{t,T}$ and Z_t coincide.

It is straightforward to show $\{\widetilde{G}_{t,T}\}$ converges almost surely to a constant G_T^* related to G^*. By construction $\{\tilde{Z}_{t,T}\}$ is bounded. Furthermore, if we consider the finite space Markov chain $\{X_t\}$ with $X_t = (i_{t-T}, i_{t-T+1}, \ldots, i_t, i_{t+1})$, then for $t > T$, $\tilde{Z}_{t,T}\psi(i_t, i_{t+1})'$ is a function of X_t. Denote this function by f. Since $\widetilde{G}_{T,T}$ takes values in a finite set (whose size depends on T), an application of LLN and stochastic approximation theory (see e.g., (Borkar, 2008), Chap. 6, Theorem 7 and Cor. 8] shows that under the stepsize condition in Assumption 2, $\widetilde{G}_{t,T}$ converges a.s. to $E_0[f(X_{T+1})]$, the expectation of $f(X_{T+1})$ under the stationary distribution of the Markov chain $\{X_t\}$ (equivalently, that of the chain $\{i_t\}$):

$$\widetilde{G}_{t,T} \overset{a.s.}{\to} G_T^* = E_0\big[\tilde{Z}_{T+1,T}\, \psi(i_{T+1}, i_{T+2})'\big]$$
$$= \Phi' \Xi_p \Big(\sum_{m=0}^{T} \beta^m Q^m\Big)\Psi. \tag{24}$$

We now relate (Z_t, G_t) to $(\tilde{Z}_{t,T}, \widetilde{G}_{t,T})$. First we bound $E\|Z_t - \tilde{Z}_{t,T}\|$. By definition $\|Z_t - \tilde{Z}_{t,T}\| = 0$ for $t \le T$. For $t \ge T+1$, similarly to bounding $E\|Z_t\|$, we have with $c = \max\{\|z_0\|, \max_i \|\phi(i)\|\}$,

$$E\|Z_t - \tilde{Z}_{t,T}\| = E\Big\|\beta^t L_0^t z_0 + \sum_{m=T+1}^{t-1} \beta^m L_{t-m}^t \phi(i_{t-m})\Big\| \le c\, E\Big[\sum_{m=T+1}^{t} \beta^m L_{t-m}^t\Big] \le \frac{c\beta^{T+1}}{1-\beta}. \tag{25}$$

Next we bound $E\|G_t - \widetilde{G}_{t,T}\|$. By the definition of G_t and $\widetilde{G}_{t,T}$,

$$G_t - \widetilde{G}_{t,T} = (1-\gamma_t)\big(G_{t-1} - \widetilde{G}_{t-1,T}\big) + \gamma_t\big(Z_t - \tilde{Z}_{t,T}\big)\psi(i_t, i_{t+1})'.$$

Consequently, with $c = \max_{i,j} \|\psi(i,j)\|$,

$$\begin{aligned} E\|G_t - \widetilde{G}_{t,T}\| &\le (1-\gamma_t) E\|G_{t-1} - \widetilde{G}_{t-1,T}\| + \gamma_t c\, E\|Z_t - \tilde{Z}_{t,T}\| \\ &\le (1-\gamma_t) E\|G_{t-1} - \widetilde{G}_{t-1,T}\| + \gamma_t \epsilon_T, \end{aligned} \tag{26}$$

where the last inequality follows from Eq. (25), and for some constant c,

$$\epsilon_T = c\beta^{T+1}/(1-\beta) \to 0, \quad \text{as} \quad T \to \infty. \tag{27}$$

Since $\gamma_t \in (0,1]$ and $\|G_t - \widetilde{G}_{t,T}\| = 0$ for $t \le T$, Eq. (26) implies

$$\sup_t E\|G_t - \widetilde{G}_{t,T}\| \le \epsilon_T. \tag{28}$$

We now bound $E\big\|\widetilde{G}_{t,T} - G_T^*\big\|$. By Eq. (24) $\widetilde{G}_{t,T} - G_T^* \overset{a.s.}{\to} 0$. By the construction of $\widetilde{G}_{t,T}$ and the fact $\gamma_t \in (0,1]$, for some deterministic constant c_T depending on T, $\|\widetilde{G}_{t,T}\| \le c_T, \forall t$. Therefore, by the Lebesgue bounded convergence theorem,

$$\lim_{t\to\infty} E\big\|\widetilde{G}_{t,T} - G_T^*\big\| = 0. \tag{29}$$

Combining Eqs. (28) and (29), we have

$$\begin{aligned} \limsup_{t\to\infty} E\,\|G_t - G^*\| &\le \limsup_{t\to\infty} E\big\|G_t - \widetilde{G}_{t,T}\big\| + \lim_{t\to\infty} E\big\|\widetilde{G}_{t,T} - G_T^*\big\| + \big\|G^* - G_T^*\big\| \\ &\le \epsilon_T + 0 + \tilde{\epsilon}_T, \end{aligned} \tag{30}$$

where $\tilde{\epsilon}_T = \|G^* - G_T^*\|$, and $\tilde{\epsilon}_T \to 0$ as $T \to \infty$, as can be seen from the definition of G^* and G_T^*, Eqs. (13) and (24). Letting T go to ∞ in the right-hand-side of (30) and using also Eq. (27), it follows that $\limsup_{t\to\infty} E\,\|G_t - G^*\| = 0$. This completes the proof.

4.2. Proof of Theorem 3

We need the following lemma, which also plays a key role in establishing Theorem 2.

Lemma 1. *Let $Y_t = \beta L_{t-1}^t Y_{t-1}$ with $Y_0 = y_0 \in \Re^m$ and $\beta < 1$. Then, the sequence of nonnegative scalar random variables $\beta^t L_0^t \overset{a.s.}{\to} 0$, and $Y_t = \beta^t L_0^t y_0 \overset{a.s.}{\to} 0$.*

Proof. From the definition of Y_t and L_s^t [cf. Eq. (21)], $Y_t = \beta^t L_{t-1}^t L_{t-2}^{t-1} \cdots L_0^1 y_0 = \beta^t L_0^t y_0$. Consider the nonnegative sequence $X_t = \beta^t L_0^t$ with $X_0 = 1$. We have

$$X_t = \beta L_{t-1}^t X_{t-1}, \;\Rightarrow\; E\big[X_t \mid \mathcal{F}_{t-1}\big] = \beta X_{t-1} \le X_{t-1},$$

where $\mathcal{F}_{t-1} = \sigma(i_s, s \le t-1)$ is the σ-field generated by $i_s, s \le t-1$. This implies that $\{(X_t, \mathcal{F}_t)\}$ is a nonnegative supermartingale. Since $EX_0 = 1 < \infty$, by a martingale congergence theorem [see (Breiman, 1992), Theorem 5.14 and its proof], $X_t \overset{a.s.}{\to} X$, a nonnegative random variable with $EX \le \liminf_{t\to\infty} EX_t$. Since $\beta < 1$, $EX_t = \beta^t \to 0$ as $t \to \infty$. Therefore $X = 0$ a.s., implying $X_t \overset{a.s.}{\to} 0$ and $Y_t \overset{a.s.}{\to} 0$. □

By Theorem 2 the Markov chain $\{(i_t, Z_t)\}$ has a unique invariant probability measure π. It can be shown that under the stationary distribution $\mathbf{P}_\pi$ of $\{(i_t, Z_t)\}$ with initial distribution π, $E_\pi\|Z_0\psi(i_0, i_1)'\| < \infty$ [(Yu, 2010), Prop. 5.3]. We can now prove Theorem 3, which states that with $\gamma_t = 1/(t+1)$, for each initial condition (z_0, G_0), $G_t \overset{a.s.}{\to} G^* = E_\pi\big[Z_0\psi(i_0, i_1)'\big]$.

Fix G_0, and consider an initial condition (z_0, G_0) for any z_0. Consider the sequence $\{G_t\}$ corresponding to $\gamma_t = 1/(t+1)$, and a related sequence $\{\widetilde{G}_t\}$ given below, with $Z_0 = z_0$:

$$G_t = \frac{1}{t+1}\Big(\sum_{k=1}^{t} Z_k\psi(i_k, i_{k+1})' + G_0\Big),$$

$$\widetilde{G}_t = \frac{1}{t+1}\sum_{k=0}^{t} Z_k\psi(i_k, i_{k+1})'.$$

Since $G_0/(t+1) \to 0$ and $Z_0\psi(i_0, i_1)'/(t+1) \to 0$ as $t \to \infty$, the convergence of $\{G_t\}$ on a sample path is equivalent to that of $\{\widetilde{G}_t\}$, which does not depend on G_0. Since $E_\pi\|Z_0\psi(i_0, i_1)'\| < \infty$, applying LLN [see (Meyn & Tweedie, 2009), Theorem 17.1.2; (Doob, 1953), Theorem 2.1] to the stationary Markov process $\{(i_t, Z_t, i_{t+1})\}$ under $\mathbf{P}_\pi$, and using also the second part of Theorem 1, it can be shown that for each initial condition $x = (\bar{i}, \bar{z})$ from a measurable set F with $\pi(F) = 1$, $\widetilde{G}_t \overset{a.s.}{\to} G^*$, and $G^* = E_\pi\big[Z_0\psi(i_0, i_1)'\big]$. Hence $G_t \overset{a.s.}{\to} G^*$ for initial condition $x \in F$.

We now show for any initial condition $\hat{x} \notin F$, the corresponding $\hat{G}_t$ also converges almost surely to G^*.

Let $\hat{x} = (\bar{i}, \hat{z})$. Since $\{i_t\}$ is irreducible, $\pi(\{\bar{i}\} \times \Re^{n_r}) > 0$. We also have $\pi(F) = 1$, so there exists $\bar{x} = (\bar{i}, \bar{z}) \in F$ for some $\bar{z} \in \Re^{n_r}$. Let $\Delta = \hat{z} - z$. Consider the two processes $(\hat{Z}_t, \hat{G}_t)$ and (Z_t, G_t) corresponding to the two initial conditions $\hat{x} \notin F, \bar{x} \in F$, respectively, and for the same path of $\{i_t\}$. By Lemma 1, we have

$$\hat{Z}_t - Z_t = \beta^t L_0^t \Delta, \qquad \beta^t L_0^t \overset{a.s.}{\to} 0.$$

The second relation implies also $\frac{1}{t+1}\sum_{k=1}^{t} \beta^k L_0^k \overset{a.s.}{\to} 0$. Therefore, with $c = \max_{i,j} \|\psi(i,j)\|$, we have

$$\begin{aligned}\|\hat{G}_t - G_t\| &= \Big\| \frac{1}{t+1} \sum_{k=1}^{t} \big(\hat{Z}_k - Z_k\big)\psi(i_k, i_{k+1})' \Big\| \\ &\le c\|\Delta\| \Big(\frac{1}{t+1} \sum_{k=1}^{t} \beta^k L_0^k \Big) \overset{a.s.}{\to} 0.\end{aligned}$$

We also have $G_t \overset{a.s.}{\to} G^*$ (because its initial condition $\bar{x} \in F$); therefore $\hat{G}_t \overset{a.s.}{\to} G^*$.

Thus for any initial condition $(\bar{i}, \bar{z})$ and G_0, $G_t \overset{a.s.}{\to} G^*$. Since the space of i_0 is finite, this implies for any initial distribution of i_0 and initial $(\bar{z}, G_0)$, $G_t \overset{a.s.}{\to} G^*$. The proof is complete.

Acknowledgments

I thank Prof. Dimitri Bertsekas, Dr. Dario Gasbarra, and the anonymous reviewers for their helpful feedback. This work is supported in part by Academy of Finland Grant 118653 (ALGODAN) and by PASCAL IST-2002-506778.

References

Ahamed, T. P., Borkar, V. S., and Juneja, S. Adaptive importance sampling technique for Markov chains using stochastic approximation. *Operations Research*, 54:489–504, 2006.

Bertsekas, D. P. *Dynamic Programming and Optimal Control*, volume II. Athena Scientific, Belmont, MA, third edition, 2007.

Bertsekas, D. P. Projected equations, variational inequalities, and temporal difference methods. *IEEE Trans. Automat. Contr.*, 2009. to appear.

Bertsekas, D. P. and Tsitsiklis, J. N. *Neuro-Dynamic Programming*. Athena Scientific, Belmont, MA, 1996.

Bertsekas, D. P. and Yu, H. Projected equation methods for approximate solution of large linear systems. *J. Computational and Applied Mathematics*, 227(1): 27–50, 2009.

Borkar, V. S. *Stochastic Approximation: A Dynamic Viewpoint*. Hindustan Book Agency, New Delhi, 2008.

Boyan, J. A. Least-squares temporal difference learning. In *Proc. the 16th ICML*, pp. 49–56, 1999.

Bradtke, S. J. and Barto, A. G. Linear least-squares algorithms for temporal difference learning. *Machine Learning*, 22(2):33–57, 1996.

Breiman, L. *Probability*. SIAM, Philadelphia, PA, 1992.

Doob, J. L. *Stochastic Processes*. John Wiley & Sons, New York, 1953.

Glynn, P. W. and Iglehart, D. L. Importance sampling for stochastic simulations. *Management Science*, 35: 1367–1392, 1989.

Kushner, H. J. and Yin, G. G. *Stochastic Approximation and Recursive Algorithms and Applications*. Springer-Verlag, New York, 2nd edition, 2003.

Meyn, S. *Control Techniques for Complex Networks*. Cambridge University Press, Cambdrige, UK, 2007.

Meyn, S. and Tweedie, R. L. *Markov Chains and Stochastic Stability*. Cambridge University Press, Cambdrige, UK, 2nd edition, 2009.

Nedić, A. and Bertsekas, D. P. Least squares policy evaluation algorithms with linear function approximation. *Discrete Event Dyn. Syst.*, 13:79–110, 2003.

Precup, D., Sutton, R. S., and Dasgupta, S. Off-policy temporal-difference learning with function approximation. In *Proc. the 18th ICML*, pp. 417–424, 2001.

Sutton, R. S. Learning to predict by the methods of temporal differences. *Machine Learning*, 3:9–44, 1988.

Sutton, R. S. and Barto, A. G. *Reinforcement Learning*. MIT Press, Cambridge, MA, 1998.

Tsitsiklis, J. N. and Van Roy, B. An analysis of temporal-difference learning with function approximation. *IEEE Trans. Automat. Contr.*, 42(5):674–690, 1997.

Yao, H. S. and Liu, Z. Q. Preconditioned temporal difference learning. In *Proc. the 25th ICML*, pp. 1208–1215, 2008.

Yu, H. Convergence of least squares temporal difference methods under general conditions. Tech. Report C-2010-1, Dept. CS, Univ. of Helsinki, 2010.