Technical Report C-2010-1
Dept. Computer Science
University of Helsinki
Apr 2010

Convergence of Least Squares Temporal Difference Methods Under General Conditions

Huizhen Yu
janey.yu@cs.helsinki.fi

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, in particular, the e-chains.

Contents

1   Introduction

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 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 and Barto [SB98]) and have been suggested as important simulation-based methods for large-scale dynamic programming (Glynn and Iglehart [GI89]).

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] 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 can be written in terms of the transition probabilities of the two Markov chains as

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

We denote this condition by $Q \prec P$.

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

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

With the temporal difference methods (Sutton [Sut88]; see also the books by Bertsekas and Tsitsiklis [BT96], Sutton and Barto [SB98], Bertsekas [Ber07], and Meyn [Mey07]), we 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) lies in the approximation subspace and is used to approximate the cost vector J^* of the target policy.

---

[1]The equivalent MDP on the space of state-action pairs can be described as follows. Consider any two state-action pairs $i = (s, u)$ and $j = (\hat{s}, v)$. Suppose that under action u, a transition from s to $\hat{s}$ occurs with probability $p(\hat{s} \mid s, u)$ and 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 is $g(i, j) = c(s, u, \hat{s})$, and 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 $p(\hat{s} \mid s, u)\mu(v \mid \hat{s})$.

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., [GI89, SB98, PSD01, ABJ06]). The off-policy LSTD(λ) algorithm we will analyze was first given by Bertsekas and Yu [BY09] 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) in Precup et al. [PSD01], as well as to the on-policy LSTD(λ) counterpart (Bradtke and Barto [BB96], Boyan [Boy99]). The algorithm can be described as follows.

Let $(i_0, i_1, \dots)$ 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, i.e.,

$$\Phi' = \begin{bmatrix} \phi(1) & \phi(2) & \cdots & \phi(n) \end{bmatrix}.$$

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 $t \geq 1$,

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

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

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

The vector b_t and matrix C_t aim to approximate the quantities defining the projected Bellman equation (3). A 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 $P = Q$. Then, all the ratios $\frac{q_{i_{t-1} i_t}}{p_{i_{t-1} i_t}}$ appearing above in Z_t and C_t become 1, and the algorithm reduces to the on-policy LSTD algorithm as first given by Bradtke and Barto [BB96] for $\lambda = 0$ and Boyan [Boy99] for $\lambda \in [0, 1]$.

In the off-policy case, a practically important property 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, and do not depend on the state transition probabilities of the original MDP, (as can be seen from Footnote 1). Thus they need not be stored and can be calculated on-line in the algorithm. This fact is well-known and finds use in many existing simulation-based algorithms for MDP.

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 and Bertsekas [NB03]. 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 and Yu [BY09]; 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).

---

[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, in relation to the projected Bellman equation.

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 3.1) can be established using arguments based on finite space Markov chains, while the proof of the almost sure convergence is not so straightforward and finite space Markov chains-based arguments are no longer sufficient. The technical complexity here is partly due to the fact that the sequence $\{Z_t\}$ cannot be bounded a priori. Indeed, the convergence proofs in [NB03, BY09] used the boundedness of $\{Z_t\}$ in the special cases of LSTD(λ), while for the off-policy case and a general value of λ, we can show that in fairly common situations, $\{Z_t\}$ is almost surely unbounded (Prop. 3.2). 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) in Yao and Liu [YL08]) or fixed point mappings (e.g., LSPE(λ), see Bertsekas and Yu [BY09]; and scaled versions of LSPE(λ), see Bertsekas [Ber09]) 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 of 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 and Van Roy [TV97]). We will use the theory of e-chains (see Meyn and Tweedie [MT09]), 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 3.2), 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 3.3). 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 (ordinary differential equation) proof approach: relevant here is the mean-ODE method (see e.g., Kushner and Yin [KY03], Borkar [Bor06, Bor08]), 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 details of proofs/analyses based on finite space Markov chains and topological space Markov chains in Sections 4 and 5, respectively. Finally, we discuss other applications of our results and future research in Section 6.

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

$$T(J) = g + \alpha Q J, \qquad \forall J \in \Re^n.$$

The mapping $T^{(\lambda)}$ in Eq. (3) is defined by

$$T^{(\lambda)} = (1 - \lambda) \sum_{m=0}^{\infty} \lambda^m T^{m+1}, \quad \lambda \in [0, 1);$$

$$T^{(1)}(J) = \lim_{\lambda \to 1} T^{(\lambda)}(J), \qquad \forall J \in \Re^n.$$

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). The convergence of $\{b_t\}, \{C_t\}$ 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

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

Here and in what follows E denotes 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 2.1. *The Markov chain $\{i_t\}$ with transition matrix P is irreducible, and $Q \prec P$ in the sense of Eq. (1).*

Assumption 2.2. *The sequence of stepsizes γ_t is deterministic and satisfies $\gamma_t \in (0, 1]$,*

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

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. In the two following sections we will give detailed proofs.

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 3.1. *Under Assumption 2.1, for each initial condition z_0, $\sup_t E\|Z_t\| \leq \frac{c}{1-\beta}$ where $c = \max\{\|z_0\|, \max_i \|\phi(i)\|\}$. Under Assumptions 2.1 and 2.2, 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 of Markov chains (see Breiman [Bre92, Theorem 7.43]), we can show the following result.

Proposition 3.1. *Under Assumptions 2.1 and 2.2, 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 3.1 and Prop. 3.1(iv) together have the following implication on the convergence of G_t. According to Prop. 3.1(iv), for the event $\mathcal{E} = \{\lim_{t\to\infty} G_t \text{ exists}\}$, 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 3.1 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 3.1. This is summarized as follows.

Corollary 3.1. *Under Assumptions 2.1 and 2.2, for each initial condition (z_0, G_0), either $G_t \overset{a.s.}{\to} G^*$, or $G_t \to G^*$ in probability and with probability 1, on every sample path G_t does not converge.*

In Section 3.2, we will rule out the second case in Cor. 3.1 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. 3.1, contrasting the off-policy case with the standard, on-policy case where $P = Q$. In the latter case, events (i) and (ii) in Prop. 3.1 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,

$$\{\lim_{t\to\infty} G_t \text{ exists}\} \subset \{\lim_{t\to\infty} \gamma_t Z_t = 0\}.$$

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 $t \to \infty$.

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

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

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 [Bre92, Problem 5.9, p. 97] (see Lemma 4.5 in Section 4.3) and the zero-one probability statement for the event $\{\sup_t \|Z_t\| < \infty\}$ in Prop. 3.1(ii).

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:

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

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

$$\beta^m \frac{q_{\bar{i}_1 \bar{i}_2}}{p_{\bar{i}_1 \bar{i}_2}} \frac{q_{\bar{i}_2 \bar{i}_3}}{p_{\bar{i}_2 \bar{i}_3}} \cdots \frac{q_{\bar{i}_m \bar{i}_1}}{p_{\bar{i}_m \bar{i}_1}} > 1; \tag{17}$$

(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.,

$$\text{either} \quad \phi_{\bar{j}}(\bar{i}_k) \geq 0, \quad \forall k = 1, \ldots, m, \quad \text{with} \quad \phi_{\bar{j}}(\bar{i}_k) > 0 \text{ for some } k; \tag{18}$$

$$\text{or} \quad \phi_{\bar{j}}(\bar{i}_k) \leq 0, \quad \forall k = 1, \ldots, m, \quad \text{with} \quad \phi_{\bar{j}}(\bar{i}_k) < 0 \text{ for some } k. \tag{19}$$

Proposition 3.2. *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. 3.2 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 [MT09]. 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:

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

which implies that the difference sequence converges almost surely to zero (Lemma 4.3). Using more careful characterizations of such difference sequences together with the first part of Theorem 3.1, we can establish the various properties needed for applying the law of large numbers (LLN) for e-chains [MT09] 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 will be given in Section 5.

Theorem 3.2. *Under Assumption 2.1, 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 3.3. *Under Assumption 2.1, $G^* = E_\pi[Z_0 \psi(i_0, i_1)']$, and with stepsize $\gamma_t = 1/(t+1)$, for each initial condition (z_0, G_0), $G_t \overset{a.s.}{\to} G^*$.*

Theorem 3.3 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 Based on Finite Space Markov Chains

In this section we prove Theorem 3.1 and Props. 3.1 and 3.2. 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 3.1

The first part of Theorem 3.1 is straightforward to show. By Eq. (10),

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

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 $(\widetilde{Z}_{t,T}, \widetilde{G}_{t,T})$ on the same probability space, and apply the LLN for a finite space irreducible Markov chain to $\widetilde{G}_{t,T}$. We then relate $(\widetilde{Z}_{t,T}, \widetilde{G}_{t,T})$ 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

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

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

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), a standard application of LLN and stochastic approximation theory (see e.g., Borkar [Bor08, Chap. 6, Theorem 7 and Cor. 8]) shows that under the stepsize condition in Assumption 2.2, $\widetilde{G}_{t,T}$ 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 $\{i_t\}$):

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

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 $t \leq T$. For $t \geq T+1$, similarly to bounding $E\|Z_t\|$, we have with $c = \max\{\|z_0\|, \max_i \|\phi(i)\|\}$,

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

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

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

which implies

$$
\|G_t - \widetilde{G}_{t,T}\| \leq (1-\gamma_t)\|G_{t-1} - \widetilde{G}_{t-1,T}\| + \gamma_t\|Z_t - \tilde{Z}_{t,T}\|\|\psi(i_t, i_{t+1})\|.
$$

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

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

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

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

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

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

We now bound $E\|\widetilde{G}_{t,T} - G_T^*\|$. By Eq. (24) $\widetilde{G}_{t,T} - G_T^* \stackrel{a.s.}{\to} 0$. 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, $\|\widetilde{G}_{t,T}\| \leq c_T, \forall t$. Therefore, by the Lebesgue bounded convergence theorem,

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

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

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

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 r.h.s. 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 Prop. 3.1

We will use a zero-one law for tail events of Markov chains. An event $\mathcal{E}$ is called a tail event of a process $\{X_t\}$ if for all positive integers s, $\mathcal{E} \in \sigma(X_t, t \geq s)$, the σ-field generated by $X_t, t \geq s$. (See Breiman [Bre92, Def. 3.10]).

Lemma 4.1 (Zero-one law; see Breiman [Bre92, Theorem 7.43]). *Any tail event of an irreducible and aperiodic finite space Markov chain has probability zero or one.*

We need a form of the zero-one law applicable to events that are "almost" tail events. This is given in Cor. 4.1 after the next lemma.

Lemma 4.2. *Let $X = \{X_t\}$ be a finite space Markov chain on the probability space $(\Omega, \mathcal{F}, \mathbf{P})$. Let $W^* \subset \Omega$ be such that $W^* \in \mathcal{F}$ and $\mathbf{P}(W^*) = 0$. Let $X' = \{X'_t\}$ be the restriction of X to $\Omega \setminus W^*$. Then X' is a Markov chain on the probability space $(\Omega', \mathcal{F}', \mathbf{P}')$, where*

$$\Omega' = \Omega \setminus W^*, \qquad \mathcal{F}' = \{B \setminus W^* \mid B \in \mathcal{F}\}, \qquad \mathbf{P}'(B) = \mathbf{P}(B), \quad \forall B \in \mathcal{F}'.$$

Furthermore, X' has the same state transition probabilities as X.

Proof. Since $W^* \in \mathcal{F}$, $\mathcal{F}' \subset \mathcal{F}$ by construction and $P'(B)$ is well-defined for each $B \in \mathcal{F}'$. Using the fact $W^* \in \mathcal{F}, \mathbf{P}(W^*) = 0$, it is straightforward to verify that $\mathcal{F}'$ is by definition a σ-field of Ω' and $\mathbf{P}'$ by definition a probability measure on $(\Omega', \mathcal{F}')$. So $(\Omega', \mathcal{F}', \mathbf{P}')$ is a well-defined probability space.

To show X' (X restricted to Ω') is a Markov chain on $(\Omega', \mathcal{F}', \mathbf{P}')$, we verify that the conditional probability $\mathbf{P}'(X'_{t+1} = x_{t+1} \mid X'_0, \ldots, X'_t)$ is a function of X'_t and x_{t+1}. For any $s \geq 0$, let

$$\mathcal{E}' = \{\omega \in \Omega' \mid X'_0(\omega) = x_0, \ldots, X'_s(\omega) = x_s\}, \qquad \mathcal{E} = \{\omega \in \Omega \mid X_0(\omega) = x_0, \ldots, X_s(\omega) = x_s\}.$$

We have $\mathcal{E}' \subset \mathcal{E} \subset \mathcal{E}' \cup W^*$, and since $\mathbf{P}(W^*) = 0$, $\mathbf{P}(\mathcal{E}) = \mathbf{P}(\mathcal{E}') = \mathbf{P}'(\mathcal{E}')$. This shows that X on $(\Omega, \mathcal{F}, \mathbf{P})$ and X' on $(\Omega', \mathcal{F}', \mathbf{P}')$ have the same distribution, in particular,

$$\mathbf{P}'(X'_0 = x_0, \ldots, X'_t = x_t) = \mathbf{P}(X_0 = x_0, \ldots, X_t = x_t),$$
$$\mathbf{P}'(X'_0 = x_0, \ldots, X'_t = x_t, X'_{t+1} = x_{t+1}) = \mathbf{P}(X_0 = x_0, \ldots, X_t = x_t, X_{t+1} = x_{t+1}).$$

Hence, for any $(x_0, \ldots, x_t)$ such that $\mathbf{P}'(X'_0 = x_0, \ldots, X'_t = x_t) > 0$,

$$
\begin{aligned}
\mathbf{P}'(X'_{t+1} = x_{t+1} \mid X'_0 = x_0, \ldots, X'_t = x_t) &= \frac{\mathbf{P}'(X'_0 = x_0, \ldots, X'_t = x_t, X'_{t+1} = x_{t+1})}{\mathbf{P}'(X'_0 = x_0, \ldots, X'_t = x_t)} \\
&= \frac{\mathbf{P}(X_0 = x_0, \ldots, X_t = x_t, X_{t+1} = x_{t+1})}{\mathbf{P}(X_0 = x_0, \ldots, X_t = x_t)} \\
&= \mathbf{P}(X_{t+1} = x_{t+1} \mid X_0 = x_0, \ldots, X_t = x_t) \\
&= \mathbf{P}(X_{t+1} = x_{t+1} \mid X_t = x_t),
\end{aligned}
$$

which is a function of x_t and x_{t+1}. This shows X' is a Markov chain and has the same transition probabilities as X. $\square$

Corollary 4.1. *Let X, X' and W^* be as in Lemma 4.2, and furthermore, let X be irreducible and aperiodic. If $\mathcal{E} \in \mathcal{F}$ is such that $\mathcal{E} \setminus W^*$ is a tail event of X', then $\mathbf{P}(\mathcal{E}) = 1$ or 0.*

Proof. By Lemma 4.2, X' is an irreducible and aperiodic Markov chain on $(\Omega', \mathcal{F}', \mathbf{P}')$. Therefore, by Lemma 4.1, the tail event $\mathcal{E}' = \mathcal{E} \setminus W^*$ has either $\mathbf{P}'(\mathcal{E}') = 0$ or $\mathbf{P}'(\mathcal{E}') = 1$. We have $\mathcal{E}' \subset \mathcal{E} \subset \mathcal{E}' \cup W^*$, $\mathbf{P}(W^*) = 0$, and $\mathbf{P}(\mathcal{E}') = \mathbf{P}'(\mathcal{E}')$ by the construction of $\mathbf{P}'$ in Lemma 4.2. This implies $\mathbf{P}(\mathcal{E}) = \mathbf{P}(\mathcal{E}') = \mathbf{P}'(\mathcal{E}')$, so $\mathbf{P}(\mathcal{E})$ is either zero or one. $\square$

We will also use the following two lemmas for bounding iterates.

Lemma 4.3. *Let $Y_t = \beta L^t_{t-1} Y_{t-1}, t \geq 1$ be vector-valued random variables, where $\beta < 1$, $E\|Y_0\| < \infty$, and Y_0 is independent of $i_t, t \geq 1$ conditionally on i_0. Then, $Y_t = \beta^t L^t_0 Y_0 \overset{a.s.}{\to} 0$, and in particular, the sequence of nonnegative scalar random variables $\beta^t L^t_0 \overset{a.s.}{\to} 0$.*

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 any component of Y_t and the nonnegative scalar sequence $X_t = \beta L_{t-1}^t X_{t-1}$ with X_t being the absolute value of the corresponding component of Y_t. For $t \geq 1$, let $\mathcal{F}_{t-1} = \sigma(X_0, i_s, s \leq t-1)$, the σ-field generated by $X_0, i_s, s \leq t-1$. Using the independence assumption on Y_0 and the Markov property of $\{i_t\}$, we have for $t \geq 1$,

$$E[X_t \mid \mathcal{F}_{t-1}] = \beta X_{t-1} \leq X_{t-1},$$

which implies that $\{(X_t, \mathcal{F}_t)\}$ is a nonnegative supermartingale. Since $EX_0 < \infty$, by a martingale congergence theorem (see Breiman [Bre92, Theorem 5.14] and its proof), $X_t \overset{a.s.}{\to} X$, a non-negative random variable with $EX \leq \liminf_{t\to\infty} EX_t$. Since $\beta < 1$, $EX_t = \beta^t EX_0 \to 0$ as $t \to \infty$. Therefore $X = 0$ a.s., implying $X_t \overset{a.s.}{\to} 0$ and $Y_t \overset{a.s.}{\to} 0$. $\square$

Lemma 4.4. *Suppose $\{\gamma_t\}$ satisfies Assumption 2.2 and $\{\delta_t\}$ is a sequence of positive scalars such that for some $\epsilon > 0$*

$$\delta_{t+1} \leq (1 - \gamma_t)\delta_t + \gamma_t \epsilon.$$

Then $\limsup_{t\to\infty} \delta_t \leq \epsilon$.

Proof. This is evident and can be verified using proof by contradiction. $\square$

We now prove Prop. 3.1 by applying the zero-one law for Markov chains. We discuss separately the case of an aperiodic chain and the case of a periodic chain. We give a detailed proof for event (i) and point out the differences in the proof arguments for events (ii)-(iv). Recall the definitions of Z_t, G_t [cf. Eqs. (10), (11)]: with $Z_0 = z_0$,

$$Z_t = \beta L_{t-1}^t \cdot Z_{t-1} + \phi(i_t), \qquad G_t = (1 - \gamma_t)G_{t-1} + \gamma_t Z_t \psi(i_t, i_{t+1})'. \qquad (31)$$

Case A: an aperiodic chain

Event (i): $\mathcal{E} = \{\lim_{t\to\infty} G_t \text{ exists, and } \sup_t \|Z_t\| < \infty\}$. The first step of our proof is to write Z_t and G_t as

$$Z_t = (Z_t - Y_t^0) + Y_t^0, \qquad G_t = (G_t - R_t^0) + R_t^0, \qquad (32)$$

where the sequences $\{Y_t^0\}, \{R_t^0\}$, to be defined shortly, are such that

(a) for all t, $(Z_t - Y_t^0), (G_t - R_t^0)$ are functions of $i_s, s > 0$;

(b) $(Y_t^0, R_t^0) \overset{a.s.}{\to} 0$;

(c) as a consequence of (b) and Eq. (32),

$$\text{on } \Omega \backslash W_0 : \quad G_t \text{ converges and } \sup_t \|Z_t\| < \infty \quad \Leftrightarrow \quad (G_t - R_t^0) \text{ converges and } \sup_t \|Z_t - Y_t^0\| < \infty,$$
$$(33)$$

where W_0 is the set on which $(Y_t^0, R_t^0) \nrightarrow 0$ and $\mathbf{P}(W_0) = 0$.

We define $\{Y_t^0\}, \{R_t^0\}$ as follows:

$$Y_0^0 = Z_0, \qquad\qquad Y_t^0 = \beta L_{t-1}^t \cdot Y_{t-1}^0, \quad t \geq 1; \qquad (34)$$

$$R_0^0 = G_0, \qquad\qquad R_t^0 = (1 - \gamma_t)R_{t-1}^0 + \gamma_t Y_t^0 \psi(i_t, i_{t+1})', \quad t \geq 1. \qquad (35)$$

Then,

$$Z_0 - Y_0^0 = 0, \qquad G_0 - R_0^0 = 0, \qquad (36)$$

and for $t \geq 1$,

$$
\begin{aligned}
Z_t &= \beta L_{t-1}^t \cdot Z_{t-1} + \phi(i_t) \\
&= \beta L_{t-1}^t \cdot \left(Z_{t-1} - Y_{t-1}^0 + Y_{t-1}^0 \right) + \phi(i_t) \\
&= \beta L_{t-1}^t \cdot \left(Z_{t-1} - Y_{t-1}^0 \right) + Y_t^0 + \phi(i_t), \\
G_t &= (1 - \gamma_t) G_{t-1} + \gamma_t Z_t \psi(i_t, i_{t+1})' \\
&= (1 - \gamma_t) \left(G_{t-1} - R_{t-1}^0 + R_{t-1}^0 \right) + \gamma_t (Z_t - Y_t^0 + Y_t^0) \, \psi(i_t, i_{t+1})' \\
&= (1 - \gamma_t) \left(G_{t-1} - R_{t-1}^0 \right) + R_t^0 + \gamma_t (Z_t - Y_t^0) \, \psi(i_t, i_{t+1})',
\end{aligned}
$$

which, by rearranging terms, are equivalent to for $t \geq 1$,

$$
Z_t - Y_t^0 = \beta L_{t-1}^t \cdot \left(Z_{t-1} - Y_{t-1}^0 \right) + \phi(i_t), \tag{37}
$$
$$
G_t - R_t^0 = (1 - \gamma_t) \left(G_{t-1} - R_{t-1}^0 \right) + \gamma_t (Z_t - Y_t^0) \, \psi(i_t, i_{t+1})'. \tag{38}
$$

The processes $(Z_t - Y_t^0), (G_t - R_t^0), t \geq 0$ do not functionally depend on i_0, as can be seen from the fact that the variables $Z_0 - Y_0^0 = 0, G_0 - R_0^0 = 0$, $Z_1 - Y_1^0 = \phi(i_1)$, and $L_{t-1}^t, t \geq 2$ do not functionally depend on i_0. Therefore property (a) above is satisfied.

We now show $(Y_t^0, R_t^0) \overset{a.s.}{\to} 0$. Noticing $E\|Y_0^0\| < \infty$, we apply Lemma 4.3 to $\{Y_t^0, t \geq 0\}$ [cf. Eq. (34)] and obtain $Y_t^0 \overset{a.s.}{\to} 0$. By Eq. (35), $\|R_t^0\|$ satisfies

$$
\|R_t^0\| \leq (1 - \gamma_t)\|R_{t-1}^0\| + \gamma_t c\|Y_t^0\|.
$$

with $c = \max_{i,j} \|\psi(i,j)\|$, a deterministic constant. Applying Lemma 4.4 with $\delta_t = \|R_t^0\|$, the fact $Y_t^0 \overset{a.s.}{\to} 0$ implies $\|R_t^0\| \overset{a.s.}{\to} 0$, equivalently, $R_t^0 \overset{a.s.}{\to} 0$. Thus, $(Y_t^0, R_t^0) \overset{a.s.}{\to} 0$ and property (b) above is satisfied. Consequently Eq. (33) in property (c) follows.

We now apply the preceding argument recursively. Consider the recursions (37)-(38) satisfied by $(Z_t - Y_t^0), (G_t - R_t^0), t \geq 1$. They have the same form as those for $Z_t, G_t, t \geq 0$ [cf. Eq. (31)]. So we can apply the preceding argument to obtain an analogous decomposition of $(Z_t - Y_t^0), (G_t - R_t^0), t \geq 0$ as

$$
Z_t - Y_t^0 = (Z_t - Y_t^0 - Y_t^1) + Y_t^1, \qquad G_t - R_t^0 = (G_t - R_t^0 - R_t^1) + R_t^1,
$$

with

$$
Y_0^1 = 0, \quad R_0^1 = 0; \qquad Y_1^1 = Z_1 - Y_1^0 = \phi(i_1), \quad R_1^1 = G_1 - R_1^0,
$$

and with other desirable properties analogous to properties (a)-(c). Moreover, we can apply recursively the preceding argument to the resulting sequences $\left(Z_t - \sum_{j=0}^{k-1} Y_t^j \right), \left(G_t - \sum_{j=1}^{k-1} R_t^j \right), t \geq k$ for $k = 1, 2, ...,$ (by construction these variables equal zero for $t < k$), and define similarly for each $k \geq 1$, the sequence $\{(Y_t^k, R_t^k), t \geq 0\}$ as follows. We define $Y_t^k = 0, R_t^k = 0$ for $t < k$, and we define $(Y_t^k, R_t^k), t \geq k$ by the recursive formulae in Eqs. (34)-(35), i.e.,

$$
Y_t^k = \beta L_{t-1}^t \cdot Y_{t-1}^k, \qquad R_t^k = (1 - \gamma_t) R_{t-1}^k + \gamma_t Y_t^k \psi(i_t, i_{t+1})', \qquad t \geq k+1,
$$

with the initial variables Y_k^k, R_k^k given by

$$
Y_k^k = Z_k - \sum_{j=0}^{k-1} Y_k^j = \phi(i_k), \qquad R_k^k = G_k - \sum_{j=1}^{k-1} R_k^j.
$$

It is also evident that for each k, the initial variable $Y_k^k = \phi(i_k)$ of the sequence $\{Y_t^k, t \geq k\}$ has finite mean and is independent of $i_s, s > k$ conditionally on i_k. Thus we can apply first Lemma 4.3 to establish $Y_t^k \overset{a.s.}{\to} 0$, and then Lemma 4.4 to show $R_t^k \overset{a.s.}{\to} 0$, as in the preceding analysis. As the final result of this procedure, we obtain decompositions of Z_t, G_t as

$$
Z_t = \left(Z_t - \sum_{j=0}^{k} Y_t^j \right) + \sum_{j=0}^{k} Y_t^j, \qquad G_t = \left(G_t - \sum_{j=0}^{k} R_t^j \right) + \sum_{j=0}^{k} R_t^j, \quad k \geq 0,
$$

with the following properties: for every $k \geq 0$,

(a') for all t, $\left(Z_t - \sum_{j=0}^{k} Y_t^j\right)$, $\left(G_t - \sum_{j=0}^{k} R_t^j\right)$ are functions of $i_s, s > k$;

(b') $(Y_t^k, R_t^k) \overset{a.s.}{\to} 0$;

(c') on $\Omega \setminus \cup_{j=0}^{k} W_j$,

$$G_t \text{ converges and } \sup_t \|Z_t\| < \infty \quad \Leftrightarrow \quad \left(G_t - \sum_{j=0}^{k} R_t^j\right) \text{ converges and } \sup_t \left\|Z_t - \sum_{j=0}^{k} Y_t^j\right\| < \infty,$$

$$(39)$$

where $W_j, j \geq 0$ are sets of zero probability (on which $(Y_t^j, R_t^j) \nrightarrow 0$).

Let $W^* = \cup_{j=0}^{\infty} W_j$. Then, $\mathbf{P}(W^*) = 0$ and for all k,

on $\Omega \setminus W^*$:

$$G_t \text{ converges and } \sup_t \|Z_t\| < \infty \quad \Leftrightarrow \quad \left(G_t - \sum_{j=0}^{k} R_t^j\right) \text{ converges and } \sup_t \left\|Z_t - \sum_{j=0}^{k} Y_t^j\right\| < \infty.$$

$$(40)$$

Now, let $\Omega' = \Omega \setminus W^*$ and consider the event

$$\mathcal{E}' = \mathcal{E} \setminus W^* = \left\{\omega \in \Omega' \,\middle|\, G_t(\omega) \text{ converges, and } \sup_t \|Z_t(\omega)\| < \infty \right\}.$$

Let X' be the Markov chain $X = \{i_t\}$ restricted to Ω'. Lemma 4.2 shows that X' is a Markov chain on the probability space $(\Omega', \mathcal{F}', \mathbf{P}')$ constructed from $(\Omega, \mathcal{F}, \mathbf{P})$ by excluding W^*, as given in the lemma. Equation (40) and property (a') above imply that $\mathcal{E}'$ is a tail event of X'. Therefore by Cor. 4.1, if the original Markov chain X is irreducible and aperiodic, then $\mathbf{P}(\mathcal{E}) = 1$ or 0.

Events (ii)-(iv): We use the same sequences of Y_t^k, R_t^k and sets W_k, W^* constructed earlier. The zero-one probability statements for events (ii)-(iv) can be established using the preceding argument with the following replacements: we replace the equivalence relation in Eqs. (39) and (40) (on $\Omega \setminus \cup_{j=0}^{k} W_j$ and $\Omega \setminus W^*$, respectively) with

- "$\sup_t \|Z_t\| < \infty \quad \Leftrightarrow \quad \sup_t \left\|Z_t - \sum_{j=0}^{k} Y_t^j\right\| < \infty$" for event (ii) $\mathcal{E} = \{\sup_t \|Z_t\| < \infty\}$;

- "$\lim_{t\to\infty} \gamma_t Z_t = 0 \quad \Leftrightarrow \quad \lim_{t\to\infty} \gamma_t \left(Z_t - \sum_{j=0}^{k} Y_t^j\right) = 0$" for event (iii) $\mathcal{E} = \{\lim_{t\to\infty} \gamma_t Z_t = 0\}$;

- "$G_t \text{ converges} \quad \Leftrightarrow \quad \left(G_t - \sum_{j=0}^{k} R_t^j\right) \text{ converges}$" for event (iv) $\mathcal{E} = \{\lim_{t\to\infty} G_t \text{ exists}\}$.

Case B: a periodic chain

Suppose now $\{i_t\}$ is periodic with period d. We first apply the preceding argument to the aperiodic chain $\{X_s^j, s \geq 0\}$, where $X_s^j = (i_{sd+j}, i_{sd+j+1}, \ldots, i_{sd+j+d-1})$, $0 \leq j \leq d-1$.

Event (i): $\mathcal{E} = \{\lim_{t\to\infty} G_t \text{ exists, and } \sup_t \|Z_t\| < \infty\}$. By the proof for the aperiodic case, each of the d events

$$\mathcal{E}_j = \left\{G_{sd+j} \text{ converges, and } \sup_s \|Z_{sd+j}\| < \infty \right\}, \quad 0 \leq j \leq d-1, \tag{41}$$

has $\mathbf{P}(\mathcal{E}_j) = 1$ or 0. Since $\mathcal{E} \subset \cap_{j=0}^{d-1} \mathcal{E}_j$, to show $\mathbf{P}(\mathcal{E}) = 1$ or 0, it is sufficient to show $\cup_{j=0}^{d-1} \mathcal{E}_j \subset \mathcal{E}$. Consider any $\mathcal{E}_j$. From the definition of G_t and Z_t,

$$\|G_t - G_{t-1}\| = \gamma_t \left\|-G_{t-1} + Z_t \psi(i_t, i_{t+1})'\right\|, \qquad \|Z_t\| \leq c(1 + \|Z_{t-1}\|),$$

where c is some deterministic constant. Using the fact $\gamma_t \to 0$, this implies

$$\text{on } \mathcal{E}_j: \quad \|G_{sd+j+1} - G_{sd+j}\| \to 0, \qquad \sup_s \|Z_{sd+j+1}\| < \infty,$$

so $\mathcal{E}_j \subset \mathcal{E}_{(j+1) \bmod d}$ and furthermore, for all $\omega \in \mathcal{E}_j$, the subsequence $\{G_{sd+j+1}(\omega), s \geq 0\}$ converges to the same limit as the subsequence $\{G_{sd+j}(\omega), s \geq 0\}$. This shows that all $\mathcal{E}_j$ are equal, and furthermore, $\cup_{j=0}^{d-1} \mathcal{E}_j \subset \mathcal{E}$. Therefore, $\mathbf{P}(\mathcal{E}) = 1$ or 0.

Events (ii)-(iii): The proof argument for event (ii) $\mathcal{E} = \{\sup_t \|Z_t\| < \infty\}$ in the periodic case is the same as for event (i) except that we replace $\mathcal{E}_j$ in Eq. (41) by

$$\mathcal{E}_j = \left\{ \sup_s \|Z_{sd+j}\| < \infty \right\}, \quad 0 \leq j \leq d-1.$$

A similar argument applies to event (iii) $\mathcal{E} = \{\lim_{t\to\infty} \gamma_t Z_t = 0\}$, in which case we define $\mathcal{E}_j$ to be

$$\mathcal{E}_j = \left\{ \lim_{s\to\infty} \gamma_{sd+j} Z_{sd+j} = 0 \right\}, \quad 0 \leq j \leq d-1,$$

we have $\mathbf{P}(\mathcal{E}_j) = 1$ or 0 by the proof for the aperiodic case, and we show $\cup_{j=0}^{d-1} \mathcal{E}_j \subset \mathcal{E}$ using the fact that $\gamma_t \to 0$ and for some deterministic constants c_1, c_2, c_3,

$$\gamma_t \|Z_t\| \leq c_1 \frac{\gamma_t}{\gamma_{t-1}} \gamma_{t-1} \|Z_{t-1}\| + \gamma_t c_2, \quad \text{and} \quad \frac{\gamma_t}{\gamma_{t-1}} \leq c_3,$$

where the last inequality follows from Assumption 2.2.

Event (iv): $\mathcal{E} = \{\lim_{t\to\infty} G_t \text{ exists}\}$. As shown in Section 3.1, we have

$$\mathcal{E} \subset \widetilde{\mathcal{E}} = \{ \lim_{t\to\infty} \gamma_t Z_t = 0\},$$

where $\widetilde{\mathcal{E}}$ is the event in (iii) and $\mathbf{P}(\widetilde{\mathcal{E}}) = 1$ or 0, as we just proved. We also have by the proof for the aperiodic case that each event

$$\mathcal{E}_j = \left\{ G_{sd+j} \text{ converges} \right\}, \quad 0 \leq j \leq d-1,$$

has $\mathbf{P}(\mathcal{E}_j) = 1$ or 0, and $\mathcal{E} \subset \cap_{j=0}^{d-1} \mathcal{E}_j$. Therefore, to show $\mathbf{P}(\mathcal{E}) = 1$ or 0, it is sufficient to show $\cup_{j=0}^{d-1} \mathcal{E}_j \cap \widetilde{\mathcal{E}} \subset \mathcal{E}$. Consider any $\mathcal{E}_j$. Since

$$\|G_t - G_{t-1}\| \leq \gamma_t \|G_{t-1}\| + c\|\gamma_t Z_t\|,$$

for some deterministic constant c, we have

$$\text{on } \mathcal{E}_j \cap \widetilde{\mathcal{E}}: \quad \|G_{sd+j+1} - G_{sd+j}\| \to 0, \quad \text{as } s \to 0.$$

which implies

$$\mathcal{E}_j \cap \widetilde{\mathcal{E}} \subset \mathcal{E}_{j+1 \bmod d} \cap \widetilde{\mathcal{E}},$$

and furthermore, on $\mathcal{E}_j \cap \widetilde{\mathcal{E}}$, the two subsequences $\{G_{sd+j+1}, s \geq 0\}, \{G_{sd+j}, s \geq 0\}$ converge to the same limit. Repeating this argument for all j, we have that all the sets $\mathcal{E}_j \cap \widetilde{\mathcal{E}}$ are equal, and furthermore, for all $\omega \in \mathcal{E}_j \cap \widetilde{\mathcal{E}}$, the entire sequence $\{G_t(\omega)\}$ converges. This implies $\cup_{j=0}^{d-1} \mathcal{E}_j \cap \widetilde{\mathcal{E}} \subset \mathcal{E}$. The proof is now complete.

4.3 Proof of Prop. 3.2

We will use the following lemma, which is a consequence of the extended Borel-Cantelli lemma [Bre92, Problem 5.9, p. 97].

Lemma 4.5. *For any process $\{X_t, t \geq 0\}$ with X_t taking values in S, and any Borel-measurable subsets A, B of S, if for all t,*

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

then

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

In the above "i.o." stands for "infinitely often," and "a.s." means that the set-inclusion relation holds after excluding a set of zero probability from the expression on the left-hand-side.

We now prove Prop. 3.2. Denote by $\mathcal{C}$ the set of states $\{\bar{i}_1, \bar{i}_2, \ldots, \bar{i}_m\}$ appeared in the cycle configuration. We prove the statement for the case where the cycle satisfies properties (a), (b) and (c) with Eq. (18), that is, $\phi_{\bar{j}}(i)$ is nonnegative for any $i \in \mathcal{C}$ and positive for some $i \in \mathcal{C}$. An identical argument with a change of signs applies to the case where Eq. (18) is replaced by Eq. (19) in property (c).

Suppose at time t, $i_t = \bar{i}_1$ and $Z_t = z_t$. If the chain goes through the cycle of states during the time interval $[t, t+m]$, then a direct calculation shows the value $z_{t+m, \bar{j}}$ of the $\bar{j}$th component of Z_{t+m} would be:

$$z_{t+m, \bar{j}} = \beta^m l_0^m \cdot z_{t, \bar{j}} + \epsilon, \tag{42}$$

where

$$\epsilon = \sum_{k=1}^{m-1} \beta^{m-k} l_k^m \phi_{\bar{j}}(\bar{i}_{k+1}) + \phi_{\bar{j}}(\bar{i}_1), \qquad l_k^m = \frac{q_{\bar{i}_{k+1}\bar{i}_{k+2}}}{p_{\bar{i}_{k+1}\bar{i}_{k+2}}} \frac{q_{\bar{i}_{k+2}\bar{i}_{k+3}}}{p_{\bar{i}_{k+2}\bar{i}_{k+3}}} \cdots \frac{q_{\bar{i}_m \bar{i}_1}}{p_{\bar{i}_m \bar{i}_1}}, \quad 0 \leq k \leq m-1.$$

By properties (b) and (c) with Eq. (18), we have

$$\epsilon > 0, \qquad \beta^m l_0^m > 1.$$

Define $\zeta = \beta^m l_0^m$. Consider the sequence $\{y_s\}$ defined by the recursion

$$y_{s+1} = \zeta y_s + \epsilon, \quad s \geq 0;$$

y_s corresponds to the value $z_{t+sm, \bar{j}}$ if during $[t, t+sm]$ the chain would repeat the cycle s times [cf. Eq. (42)]. Since $\zeta > 1, \epsilon > 0$, simple calculation shows that unless $y_s = -\epsilon/(\zeta - 1)$ for all $s \geq 0$, $|y_s| \to \infty$ as $s \to \infty$.

Let $\nu = -\epsilon/(\zeta - 1) = -\epsilon/(\beta^m l_0^m - 1)$ be the cycle-dependent, negative constant in the statement of the proposition. Consider any $\eta > 0$ and two positive integers K_1, K_2 with $K_1 \leq K_2$. Let s be such that $|y_s| \geq K_2$ for all $y_0 \in [-K_1, K_1], y_0 \notin (\nu - \eta, \nu + \eta)$. By property (a) of the cycle and the Markov property of $\{i_t\}$, whenever $i_t = \bar{i}_1$, conditionally on the history, there is some positive probability δ independent of t to repeat the cycle s times. Therefore, by Lemma 4.5,

$$\{i_t = \bar{i}_1, \ Z_{t, \bar{j}} \notin (\nu - \eta, \nu + \eta), \ \|Z_t\| \leq K_1 \ i.o.\} \subset \{\|Z_t\| \geq K_2 \ i.o.\} \ \ a.s. \tag{43}$$

We now prove Z_t is almost surely unbounded, i.e., $\mathbf{P}\left(\sup_t \|Z_t\| < \infty\right) = 0$. By Prop. 3.1(ii), $\mathbf{P}\left(\sup_t \|Z_t\| < \infty\right) = 0$ or 1, so let us assume $\mathbf{P}\left(\sup_t \|Z_t\| < \infty\right) = 1$ to derive a contradiction. Define

$$K_1 = \mathrm{median}\left(\sup_t \|Z_t\|\right), \qquad \mathcal{E} = \{\sup_t \|Z_t\| \leq K_1\}. \tag{44}$$

Then,

$$K_1 < \infty, \qquad \mathbf{P}(\mathcal{E}) \geq 1/2.$$

Let $\eta > 0$ be such that $(\nu - \eta, \nu + \eta) \subset \mathcal{O}(\nu)$, where $\mathcal{O}(\nu)$ is the neighborhood of ν in the statement of the proposition. By the assumption in the proposition $\mathbf{P}\left(i_t = \bar{i}_1, \ Z_{t, \bar{j}} \notin (\nu - \eta, \nu + \eta) \ i.o.\right) = 1$. Since $\mathcal{E} \subset \{\|Z_t\| \leq K_1 \ i.o.\}$, this implies

$$\mathcal{E} \subset \{i_t = \bar{i}_1, \ Z_{t, \bar{j}} \notin (\nu - \eta, \nu + \eta), \ \|Z_t\| \leq K_1 \ i.o.\} \ \ a.s.$$

It then follows from Eq. (43) that for any $K_2 > K_1$,

$$\mathcal{E} \subset \{\sup_t \|Z_t\| \geq K_2\} \quad a.s.$$

Since $\mathbf{P}(\mathcal{E}) \geq 1/2$, this contradicts the definition of $\mathcal{E}$ in Eq. (44). Therefore $\mathbf{P}\big(\sup_t \|Z_t\| < \infty\big) = 0$. This completes the proof.

5 Details of Analysis Based on e-Chains

In this section we will analyze the properties of the Markov chain $\{(i_t, Z_t)\}$ on the topological space $S = \mathcal{I} \times \Re^{n_r}$ with product topology, where the topology on $\mathcal{I}$ is the discrete topology and that on $\Re^{n_r}$ is the usual topology metrized by $\|\cdot\|$ (equivalently, the Euclidean distance). We will establish Theorem 3.2 on the ergodicity of $\{(i_t, Z_t)\}$ (Section 5.1) and Theorem 3.3 on the almost sure convergence of G_t when the stepsizes are $1/(t+1)$ (Section 5.2), using theories of topological space Markov chains, in particular, the e-chains (Meyn and Tweedie [MT09]).

First, we specify some notation and definitions. In this section, P denotes the transition probability kernel (or transition function) of a Markov chain $\{X_t\}$ on the state space S, i.e.,

$$P = \{P(x, A), x \in S, A \in \mathcal{B}(S)\},$$

where $P(x, \cdot)$ is the conditional probability of X_1 given $X_0 = x$. The k-step transition probability kernel is denoted by P^k. As an operator, P^k maps any bounded measurable function $f : S \to \Re$ to another such function $P^k f$, given by

$$P^k f(x) = \int_S P^k(x, dy) f(y) = E_x\big[f(X_k)\big],$$

where E_x denotes expectation with respect to $\mathbf{P}_x$, the probability distribution of $\{X_t\}$ initialized with $X_0 = x$.

Let $\mathcal{C}_b(S)$, $\mathcal{C}_c(S)$ denote the set of bounded continuous functions on S, the set of continuous functions on S with compact support, respectively. Note that since the space $\mathcal{I}$ of i_t is discrete, $\mathcal{C}_b(S)$ consists of all functions f such that $f(i, z)$ is bounded and continuous in z for each i. Similarly, $\mathcal{C}_c(S)$ consists of all functions f such that for each i, $f(i, z)$ is continuous in z and has compact support on $\Re^{n_r}$, the space of z. Note also that since $\mathcal{I}$ is finite, any $f \in \mathcal{C}_c(S)$ is bounded.

A Markov chain on S is called an *e-chain*, if its transition probability kernel P possesses the equicontinuity property: for each $f \in \mathcal{C}_c(S)$, the family of functions $\{P^t f\}$ is equicontinuous on compact sets. We will show that $\{(i_t, Z_t)\}$ is an e-chain, and furthermore, it has a unique invariant probability measure and almost surely weakly convergent sequences of occupation measures for each initial condition. These are the conclusions of Theorem 3.2, which we will use to prove Theorem 3.3 on the almost sure convergence of G_t. Theorem 3.2 is however stronger than what is needed to prove Theorem 3.3; it can be useful in analyzing the convergence of other incremental variants of the LSTD algorithm in the future.

There is an alternative way to prove Theorem 3.3 using only the existence of an invariant probability measure not the uniqueness. With this proof approach we would use the weak Feller property of $\{(i_t, Z_t)\}$,[3] a property weaker than the e-chain property, and the first part of Theorem 3.1 to establish first the existence of at least one invariant probability measure by applying [MT09, Prop. 12.1.3]. We can then prove Theorem 3.3 using modified versions of the arguments given in Section 5.2.

---

[3] A Markov chain on S is a *weak Feller* chain if its transition kernel P maps $\mathcal{C}_b(S)$ to $\mathcal{C}_b(S)$ [MT09, Prop. 6.1.1(i)]. To see $\{(i_t, Z_t)\}$ is weak Feller, note that by definition the realization of Z_1 given (z_0, i_0, i_1) is a continuous function of z_0 for given i_0, i_1. Denote this function by $Z_1(z_0, i_0, i_1)$. For any $f \in \mathcal{C}_b(S)$, $f(i, z)$ is bounded and continuous in z for each i, and it can be seen that

$$Pf(i, z) = E[f(i_1, Z_1) \mid i_0 = i, Z_0 = z] = \sum_j p_{ij} f\big(j, Z_1(z, i, j)\big)$$

is also bounded and continuous in z for each i, so $Pf \in \mathcal{C}_b(S)$.

We will present the e-chain-based analysis, since it is more thorough and has a broader range of potential applications, as noted earlier.

5.1 Proof of Theorem 3.2

We will prove Theorem 3.2 in a series of propositions. First, we prove $\{(i_t, Z_t)\}$ is an e-chain. In this and the subsequent analysis, we will need to compare multiple realizations of Z_t: $Z_t^1, Z_t^2, \ldots$, corresponding to different initial conditions $\bar{z}_1, \bar{z}_2, \ldots$, respectively, and for the same sample path of $\{i_t\}$. Such comparison is legitimate because Z_t, which we recall is recursively defined by

$$Z_t = \beta L_{t-1}^t \cdot Z_{t-1} + \phi(i_t) \tag{45}$$

[cf. Eqs. (10) and (21)], is a (vector-valued) function of $(i_0, i_1, \ldots, i_t)$ and the initial condition z_0. In such comparison we use $\mathbf{P}$ to denote the probability distribution of the resulting process $\{(i_t, Z_t^1, Z_t^2, \ldots)\}$.

We will need the following lemma, which is a consequence of Lemma 4.3.

Lemma 5.1. *For two initial conditions $(\bar{i}_0, z_0)$ and $(\bar{i}_0, z_0 + \Delta)$, let (i_t, Z_t) and $(i_t, \hat{Z}_t)$ be the corresponding processes, respectively, with the same sample path of $\{i_t\}$. Then $Z_t - \hat{Z}_t = \beta^t L_0^t \Delta$ independently of z_0, and $Z_t - \hat{Z}_t \overset{a.s.}{\to} 0$. Furthermore, for any $a \in (0, 1), \delta > 0$ and $\delta_0 > 0$, there exists a finite integer $\bar{N}_{\delta_0}^{\delta, a}$, such that for all z_0 and Δ with $\|\Delta\| \leq \delta_0$,*

$$\mathbf{P}\big(\|Z_t - \hat{Z}_t\| \leq \delta, \, \forall t \geq \bar{N}_{\delta_0}^{\delta, a}\big) \geq a.$$

Proof. From the evolution of $Z_t, \hat{Z}_t$ given in Eq. (45), we have

$$Z_t - \hat{Z}_t = \beta L_{t-1}^t \cdot \big(Z_{t-1} - \hat{Z}_{t-1}\big), \tag{46}$$

which shows that $Z_t - \hat{Z}_t = \beta^t L_0^t \Delta$ and is independent of z_0 for all t. Furthermore, by Lemma 4.3 $\beta^t L_0^t \overset{a.s.}{\to} 0$ and $Z_t - \hat{Z}_t \overset{a.s.}{\to} 0$.

Consider any $\delta > 0$. Since $\beta^t L_0^t \overset{a.s.}{\to} 0$, for almost every sample path of $\{i_t\}$, i.e., for almost every $\omega \in \Omega$, there exists a finite integer $N(\omega)$ such that

$$\beta^t L_0^t(\omega)\delta_0 \leq \delta, \qquad \forall t \geq N(\omega),$$

which implies that for any Δ with $\|\Delta\| \leq \delta_0$,

$$\|Z_t(\omega) - \hat{Z}_t(\omega)\| = \beta^t L_0^t(\omega)\|\Delta\| \leq \beta^t L_0^t(\omega)\delta_0 \leq \delta, \qquad \forall t \geq N(\omega).$$

Clearly N is a well defined random variable. For $a \in (0, 1)$, let

$$\bar{N}_{\delta_0}^{\delta, a} = \min\{k \mid \mathbf{P}\big(\{\omega : N(\omega) \leq k\}\big) \geq a\}.$$

Then, $\mathbf{P}\{\|Z_t - \hat{Z}_t\| \leq \delta, \, \forall t \geq \bar{N}_{\delta_0}^{\delta, a}\} \geq a$ for all pairs of initial conditions with $\|\Delta\| \leq \delta_0$. $\qquad \square$

Proposition 5.1. *The Markov chain $\{(i_t, Z_t)\}$ is an e-chain.*

Proof. For any $f \in \mathcal{C}_c(S)$ and $\epsilon > 0$, since f is uniformly continuous, there exists $\delta > 0$ such that

$$|f(i, z) - f(i, \hat{z})| \leq \epsilon, \qquad \forall i, z, \hat{z} \, \text{ with } \, \|z - \hat{z}\| \leq \delta. \tag{47}$$

To show $\{P^t f\}$ is equicontinuous with respect to the product topology on S, we consider for each given $\bar{i}_0$, any two initial conditions $x = (\bar{i}_0, z_0)$ and $\hat{x} = (\bar{i}_0, \hat{z}_0)$ with $\|z_0 - \hat{z}_0\| \leq \delta_0$ for some $\delta_0 > 0$, and we bound

$$|P^t f(x) - P^t f(\hat{x})| = |E[f(i_t, Z_t) - f(i_t, \hat{Z}_t)]|,$$

where $\{Z_t\}$ and $\{\hat{Z}_t\}$ are two processes for the same sample path of $\{i_t\}$ and corresponding to the two initial conditions $x, \hat{x}$, respectively.

By Lemma 5.1, for any $a \in (0,1)$ and $\delta_0 > 0$, there exists $\bar{N}_{\delta_0}^{\delta,a}$ such that for all initial conditions $z_0, \hat{z}_0$ with $\|z_0 - \hat{z}_0\| \leq \delta_0$,

$$\mathbf{P}\{\|Z_t - \hat{Z}_t\| \leq \delta\} \geq a, \qquad \forall t \geq \bar{N}_{\delta_0}^{\delta,a}.$$

Let a be sufficiently close to 1 so that $2(1-a)\|f\|_\infty \leq \epsilon$, where $\|f\|_\infty = \sup_{x \in S} |f(x)| < \infty$ (it is finite because f is bounded). Then, since by Eq. (47)

$$|f(i_t, Z_t) - f(i_t, \hat{Z}_t)| \leq \epsilon, \quad \text{on } \{\|Z_t - \hat{Z}_t\| \leq \delta\},$$

we have for all initial conditions $(\bar{i}_0, z_0), (\bar{i}_0, \hat{z}_0)$ with $\|z_0 - \hat{z}_0\| \leq \delta_0$,

$$|E[f(i_t, Z_t) - f(i_t, \hat{Z}_t)]| \leq E[|f(i_t, Z_t) - f(i_t, \hat{Z}_t)|] \leq \epsilon + (1-a)2\|f\|_\infty \leq 2\epsilon, \qquad t \geq \bar{N}_{\delta_0}^{\delta,a}. \quad (48)$$

For $\bar{N}_{\delta_0}^{\delta,a}$ given above we now bound $|E[f(i_t, Z_t) - f(i_t, \hat{Z}_t)]|, t < \bar{N}_{\delta_0}^{\delta,a}$. We have

$$\|Z_t - \hat{Z}_t\| = \beta^t L_0^t \|z_0 - \hat{z}_0\|, \qquad \sup_{t < \bar{N}_{\delta_0}^{\delta,a}} \beta^t L_0^t < c$$

for some positive deterministic constant c. So if $\|z_0 - \hat{z}_0\| \leq s\delta_0 \leq \delta/c$ for some $s \in (0,1]$, then all possible values of $\|Z_t - \hat{Z}_t\|, t < \bar{N}_{\delta_0}^{\delta,a}$ can be enclosed in a ball centered at the origin and with radius δ, and consequently by Eq. (47)

$$|E[f(i_t, Z_t) - f(i_t, \hat{Z}_t)]| \leq E[|f(i_t, Z_t) - f(i_t, \hat{Z}_t)|] \leq \epsilon, \quad t < \bar{N}_{\delta_0}^{\delta,a}, \quad (49)$$

while Eq. (48) certainly holds for all $z_0, \hat{z}_0$ with $\|z_0 - \hat{z}_0\| \leq s\delta_0$ because $s \leq 1$. Thus we have proved that for all $z_0, \hat{z}_0$ with $\|z_0 - \hat{z}_0\| \leq s\delta_0$,

$$|E[f(i_t, Z_t) - f(i_t, \hat{Z}_t)]| \leq 2\epsilon, \quad \forall t.$$

This shows that $\{P^t f\}$ is equicontinuous, so by definition $\{(i_t, Z_t)\}$ is an e-chain. $\square$

Next we prove $\{(i_t, Z_t)\}$ has a unique invariant probability measure. A Markov chain $\{X_t\}$ on S is said to be *bounded in probability* if for each initial condition x and each $\epsilon > 0$, there exists a compact subset $C \subset S$ such that

$$\liminf_{t \to \infty} \mathbf{P}_x(X_t \in C) \geq 1 - \epsilon. \quad (50)$$

This entails $\{X_t\}$ being *bounded in probability on average*: $\liminf_{t \to \infty} \frac{1}{t} \sum_{j=1}^t \mathbf{P}_x(X_j \in C) \geq 1 - \epsilon$, a condition needed in proving that there exists a unique invariant probability measure.

Lemma 5.2. *The Markov chain $\{(i_t, Z_t)\}$ is bounded in probability.*

Proof. Let $x = (\bar{i}_0, z_0)$ be any initial condition. By the first part of Theorem 3.1, for all t, $E_x\|Z_t\| \leq c$ for some constant c depending on z_0. For any $\epsilon > 0$, let K be such that $c/K \leq \epsilon$, and then by Markov's inequality, $\mathbf{P}_x(\|Z_t\| \geq K) \leq c/K \leq \epsilon$ for all t. Therefore

$$\mathbf{P}_x(\|Z_t\| \leq K) \geq 1 - \epsilon, \quad \forall t, \quad (51)$$

and the compact set C in (50) can be chosen to be $\mathcal{I} \times \{z \mid \|z\| \leq K\}$. $\square$

For a Markov chain $\{X_t\}$ on S, by definition a state x^* is *reachable* if for every neighborhood $\mathcal{O}(x^*)$ of x^*,

$$\sum_t \mathbf{P}_y(X_t \in \mathcal{O}(x^*)) > 0, \quad \forall y \in S. \quad (52)$$

For an e-chain which is bounded in probability on average, the existence of a reachable state is necessary and sufficient for the existence of a unique invariant probability measure [MT09, Theorem 18.4.4(i)].

Proposition 5.2. *The Markov chain $\{(i_t, Z_t)\}$ has a reachable state.*

Proof. We proceed in three steps. We consider first a fixed initial condition $(\bar{i}, \bar{z})$, and then initial conditions of the form $(\bar{i}, \bar{z}+\Delta)$, and finally, arbitrary initial conditions. Our proof relies on showing a stronger statement than Eq. (52): there exists a state $x^* = (i^*, z^*)$ such that for every neighborhood $\mathcal{O}(x^*)$ of x^*,

$$\limsup_{t\to\infty} \mathbf{P}_y\big((i_t, Z_t) \in \mathcal{O}(x^*)\big) > 0, \qquad \forall\, y \in S.$$

Since the space of i_t is discrete, this is equivalent to for every neighborhood $\mathcal{O}(z^*)$ of z^*,

$$\limsup_{t\to\infty} \mathbf{P}_y(i_t = i^*, Z_t \in \mathcal{O}(z^*)) > 0, \qquad \forall y \in S.$$

Consider a fixed initial condition $y = (\bar{i}, \bar{z})$. We have shown in the proof of Lemma 5.2 [cf. Eq. (51)] that for any $a \in (0, 1)$, there exists a compact set $C \subset \Re^{n_r}$ such that

$$\mathbf{P}_y(Z_t \in C) \geq a, \quad \forall t. \tag{53}$$

Since $\mathbf{P}_y(Z_t \in C) = \sum_i \mathbf{P}_y(i_t = i, Z_t \in C)$, Eq. (53) implies

$$\sum_i \limsup_{t\to\infty} \mathbf{P}_y(i_t = i, Z_t \in C) \geq \limsup_{t\to\infty} \sum_i \mathbf{P}_y(i_t = i, Z_t \in C) \geq a,$$

so there exists i^* such that

$$\limsup_{t\to\infty} \mathbf{P}_y(i_t = i^*, Z_t \in C) > 0. \tag{54}$$

Fix i^* and we find z^* next.

Let $\{\delta_k, k \geq 0\}$ be a positive sequence with $\delta_k \searrow 0$. We now construct a sequence of closed balls C_k of radius δ_k in $\Re^{n_r}$ and with the following property:

$$\limsup_{t\to\infty} \mathbf{P}_y\big(i_t = i^*, Z_t \in \cap_{j\leq k}C_j \cap C\big) > 0. \tag{55}$$

Starting with $k = 0$, consider the set of all open balls $B_{\delta_k}(z)$ of radius δ_k and centered at z for all $z \in C$. This set $\{B_{\delta_k}(z) \mid z \in C\}$ is an open cover of the compact set C, so it has a finite subcover, in which, because of Eq. (54), there must exist at least one open ball $B_{\delta_k}(\bar{z}_k)$ centered at some point $\bar{z}_k \in C$ and possessing the property

$$\limsup_{t\to\infty} \mathbf{P}_y(i_t = i^*, Z_t \in B_{\delta_k}(\bar{z}_k) \cap C) > 0. \tag{56}$$

Let C_k be the closure of $B_{\delta_k}(\bar{z}_k)$, then C_k satisfies Eq. (55) with $k = 0$. Repeating the above procedure with $\cap_{j\leq k}C_j \cap C$ in place of C, we find a closed ball C_{k+1} of radius δ_{k+1} and centered at some $\bar{z}_{k+1} \in C$ which satisfies Eq. (55). In this manner we construct the sequence of closed balls C_k of radius δ_k and with property (55).

As can be seen from Eq. (55), the compact sets $\cap_{j\leq k}C_j \cap C, k \geq 0$ are all nonempty, and they form a decreasing sequence as $k \to \infty$, so their intersection is nonempty. Let $z^* \in \cap_j C_j \cap C$. Any neighborhood $\mathcal{O}(z^*)$ of z^* contains a closed ball centered at z^* with certain radius $\delta > 0$, and this ball in turn contains C_k for all k with $\delta_k \leq \delta/2$ since $z^* \in \cap_j C_j$. By Eq. (55), this implies for such C_k,

$$\limsup_{t\to\infty} \mathbf{P}_y(i_t = i^*, Z_t \in \mathcal{O}(z^*)) \geq \limsup_{t\to\infty} \mathbf{P}_y(i_t = i^*, Z_t \in C_k) > 0. \tag{57}$$

Thus we obtain for any neighborhood $\mathcal{O}(z^*)$ of z^*,

$$\limsup_{t\to\infty} \mathbf{P}_y(i_t = i^*, Z_t \in \mathcal{O}(z^*)) > 0. \tag{58}$$

We now show (i^*, z^*) is a reachable state by proving the above relation for other initial conditions.

Consider next initial condition $\hat{y} = (\bar{i}, \bar{z} + \Delta)$ for some $\Delta \in \Re^{n_r}$. For any neighborhood $\mathcal{O}(z^*)$, let $\mathcal{O}'(z^*)$ be a smaller neighborhood such that for some $\delta > 0$,

$$\|z_1 - z_2\| \leq \delta, \; z_1 \in \mathcal{O}'(z^*) \quad \Rightarrow \quad z_2 \in \mathcal{O}(z^*). \tag{59}$$

Using Eq. (58), we can define $\epsilon, a \in (0, 1)$ such that

$$\epsilon = \limsup_{t \to \infty} \mathbf{P}_y(i_t = i^*, Z_t \in \mathcal{O}'(z^*)), \qquad a + \epsilon > 1.$$

Consider two processes (i_t, Z_t) and $(i_t, \hat{Z}_t)$ starting with initial conditions $y, \hat{y}$, respectively, and with the same sample path of $\{i_t\}$. Let $\delta_0 = \|\Delta\|$. By Lemma 5.1, for the above a, δ and δ_0, there exists $\bar{N}_{\delta_0}^{\delta,a} < \infty$ such that for all $t \geq \bar{N}_{\delta_0}^{\delta,a}$, $\mathbf{P}(\|Z_t - \hat{Z}_t\| \leq \delta) \geq a$. Since by Eq. (59)

$$\|Z_t - \hat{Z}_t\| \leq \delta, \; Z_t \in \mathcal{O}'(z^*) \quad \Rightarrow \quad \hat{Z}_t \in \mathcal{O}(z^*),$$

we have that for all $t \geq \bar{N}_{\delta_0}^{\delta,a}$,

$$\begin{aligned}
\mathbf{P}(i_t = i^*, \hat{Z}_t \in \mathcal{O}(z^*)) &\geq \mathbf{P}(i_t = i^*, \|Z_t - \hat{Z}_t\| \leq \delta, Z_t \in \mathcal{O}'(z^*)) \\
&\geq \mathbf{P}(i_t = i^*, Z_t \in \mathcal{O}'(z^*)) + \mathbf{P}(\|Z_t - \hat{Z}_t\| \leq \delta) - 1 \\
&\geq \mathbf{P}(i_t = i^*, Z_t \in \mathcal{O}'(z^*)) + a - 1,
\end{aligned}$$

and taking limsup of both sides as $t \to \infty$,

$$\limsup_{t \to \infty} \mathbf{P}(i_t = i^*, \hat{Z}_t \in \mathcal{O}(z^*)) \geq \limsup_{t \to \infty} \mathbf{P}(i_t = i^*, Z_t \in \mathcal{O}'(z^*)) + a - 1 = \epsilon + a - 1 > 0. \tag{60}$$

Thus we have proved that for all initial conditions of the form $\hat{y} = (\bar{i}, \bar{z} + \Delta)$ for some $\Delta \in \Re^{n_r}$,

$$\limsup_{t \to \infty} \mathbf{P}_{\hat{y}}(i_t = i^*, \hat{Z}_t \in \mathcal{O}(z^*)) > 0. \tag{61}$$

Finally, consider an arbitrary initial condition $\tilde{y} = (\tilde{i}, \tilde{z})$. Denote the corresponding process by $(\tilde{i}_t, \tilde{Z}_t)$. Since the Markov chain $\{\tilde{i}_t\}$ is irreducible, there exists a finite time T such that $\mathbf{P}_{\tilde{y}}(\tilde{i}_T = \bar{i}) > 0$. Let $\hat{z}$ be a possible value of $\tilde{Z}_T$, i.e., $\mathbf{P}_{\tilde{y}}(\tilde{i}_T = \bar{i}, \tilde{Z}_T = \hat{z}) > 0$. Denote by (i_t, Z_t) the process with initial condition $(\bar{i}, \hat{z})$. Then, for any neighborhood $\mathcal{O}(z^*)$ of z^*,

$$\mathbf{P}_{\tilde{y}}(\tilde{i}_{t+T} = i^*, \tilde{Z}_{t+T} \in \mathcal{O}(z^*)) \geq \mathbf{P}_{\tilde{y}}(\tilde{i}_T = \bar{i}, \tilde{Z}_T = \hat{z}) \cdot \mathbf{P}_{(\bar{i}, \hat{z})}(i_t = i^*, Z_t \in \mathcal{O}(z^*)),$$

which implies that with $a = \mathbf{P}_{\tilde{y}}(\tilde{i}_T = \bar{i}, \tilde{Z}_T = \hat{z})$,

$$\limsup_{t \to \infty} \mathbf{P}_{\tilde{y}}(\tilde{i}_{t+T} = i^*, \tilde{Z}_{t+T} \in \mathcal{O}(z^*)) \geq a \cdot \limsup_{t \to \infty} \mathbf{P}_{(\bar{i}, \hat{z})}(i_t = i^*, Z_t \in \mathcal{O}(z^*)) > 0, \tag{62}$$

where the second inequality follows from Eq. (61). This completes the proof. $\qquad \square$

By [MT09, Theorem 18.4.4(i)], Props. 5.1 and 5.2 and Lemma 5.2 together imply

Corollary 5.1. *The Markov chain $\{(i_t, Z_t)\}$ has a unique invariant probability measure π.*

By [MT09, Theorems 12.1.1, 18.4.2], Props. 5.1 and 5.2 and Lemma 5.2 together also imply

Corollary 5.2. *For the Markov chain $\{(i_t, Z_t)\}$ and any initial condition $x \in S$, $\frac{1}{T} \sum_{t=1}^{T} P^t(x, \cdot)$ converges weakly to the invariant probability measure π.*

Recall that the occupation probability measures $\mu_t, t \geq 1$ of a Markov chain $\{X_t\}$ on S are defined by

$$\mu_t(A) = \frac{1}{t} \sum_{k=1}^{t} \mathbf{1}_A(X_t), \qquad \forall A \in \mathcal{B}(S),$$

where $\mathbf{1}_A$ denotes the indicator function for a Borel-measurable set A. To show that for the chain $\{(i_t, Z_t)\}$ and each initial condition, the sequence $\{\mu_t\}$ is uniformly tight almost surely, we verify the following geometric drift condition, given by Meyn [Mey89] (see also Meyn and Tweedie [MT09, Theorem 18.5.2]). For the chain $\{(i_t, Z_t)\}$, satisfying this condition implies also that the chain is bounded in probability, which we proved separately earlier.

Lemma 5.3. *The Markov chain $\{X_t\}$, where $X_t = (i_t, Z_t)$, satisfies the following geometric drift condition: There exist a coercive function $V : S \to [1, \infty]$, a compact set $C \subset S$, and constants $\zeta > 0$, $b < \infty$, such that*

$$E_x[V(X_1)] - V(x) \leq -\zeta V(x) + b\mathbf{1}_C(x), \qquad \forall x \in S.$$

Proof. Define $V(x) = \|z\| + 1$ for $x = (i, z)$. Define $C = \mathcal{I} \times D_r$, where $D_r \subset \Re^{n_r}$ is a closed ball of radius r and centered at the origin. The radius r and constants ζ, b are to be specified shortly. Let $c = \max_i \|\phi(i)\|$. Consider any $x = (i_0, z_0) \in S$. Since $Z_1 = \beta L_0^1 \cdot z_0 + \phi(i_1)$, $E\|Z_1\| \leq \beta\|z_0\| + c$. Therefore,

$$
\begin{aligned}
E_x[V(X_1)] - V(x) = E_x[\|Z_1\|] - \|z_0\| &= (\beta - 1)\|z_0\| + c \\
&= -\zeta(\|z_0\| + 1) + (\zeta + \beta - 1)\|z_0\| + \zeta + c \\
&= -\zeta V(x) + (\zeta + \beta - 1)\|z_0\| + \zeta + c. \qquad (63)
\end{aligned}
$$

Let $\zeta, r > 0$ and $b < \infty$ be such that

$$\zeta < 1 - \beta, \qquad r \geq (\zeta + c)/(1 - \beta - \zeta), \qquad b = \zeta + c.$$

Then

$$(\zeta + \beta - 1)\|z_0\| + \zeta + c \ \leq\ \begin{cases} b & \|z_0\| \leq r, \text{ i.e., } x \in C \\ 0 & \|z_0\| > r, \text{ i.e., } x \notin C. \end{cases} \qquad (64)$$

Combining Eqs. (63)-(64) gives the desired inequality for $E_x[V(X_1)] - V(x)$. $\qquad \square$

Lemma 5.3 and Cor. 5.1 together imply that for each initial condition, the sequence $\{\mu_t\}$ of occupation measures of $\{(i_t, Z_t)\}$ converges weakly to π almost surely ([MT09, Theorem 18.5.1(ii)]; see also [MT09, Theorem 18.5.2]). This completes the proof of Theorem 3.2.

5.2 Proof of Theorem 3.3

To establish the almost sure convergence of G_t, we still need to show that $Z_0\psi(i_0, i_1)'$ has finite expectation under the stationary distribution $\mathbf{P}_\pi$ of the Markov chain $\{(i_t, Z_t)\}$ with initial distribution π. We will need Cor. 5.2, which states

$$\frac{1}{T} \sum_{t=1}^{T} P^t(x, \cdot) \overset{\text{weakly}}{\longrightarrow} \pi, \quad \forall x \in S. \qquad (65)$$

Let E_π denote expectation with respect to $\mathbf{P}_\pi$.

Proposition 5.3. $E_\pi\big[\|Z_0\psi(i_0, i_1)'\|\big] < \infty$.

Proof. Since $\|Z_0\psi(i_0,i_1)'\| \leq c\|Z_0\|$ for the deterministic constant $c = \max_{i,j}\|\psi(i,j)\|$, to prove the statement it is sufficient to show $E_\pi[\|Z_0\|] < \infty$. For a given initial condition $x = (\bar{i}_0, z_0)$ and some constant c depending on x, we have by the first part of Theorem 3.1

$$E_x[\|Z_t\|] \leq c, \quad \forall t \geq 0. \tag{66}$$

Consider a sequence of scalars $a_k, k \geq 0$ with

$$a_0 = 0, \quad a_1 \in (0,1], \quad a_{k+1} = a_k + 1, \quad k \geq 1. \tag{67}$$

Define a sequence of disjoint open sets $\{O_k, k \geq 0\}$ on the space of z as

$$O_k = \{z \mid a_k < \|z\| < a_{k+1}\}. \tag{68}$$

We have for all t,

$$\sum_{k=0}^\infty a_k \cdot \mathbf{P}_x(Z_t \in O_k) \leq E_x[\|Z_t\|],$$

and

$$\sum_{k=0}^\infty a_{k+1} \cdot \mathbf{P}_x(Z_t \in O_k) \leq \sum_{k=0}^\infty (a_k+1) \cdot \mathbf{P}_x(Z_t \in O_k) \leq 1 + \sum_{k=0}^\infty a_k \cdot \mathbf{P}_x(Z_t \in O_k).$$

Therefore, by Eq. (66),

$$\sum_{k=0}^\infty a_{k+1} \cdot \mathbf{P}_x(Z_t \in O_k) \leq c+1, \qquad \forall t \geq 0,$$

or equivalently,

$$\sum_{k=0}^K a_{k+1} \cdot \mathbf{P}_x(Z_t \in O_k) \leq c+1, \qquad \forall K \geq 0, \ t \geq 0.$$

It then follows that for all $K \geq 0, T \geq 0$,

$$\frac{1}{T}\sum_{t=1}^T \sum_{k=0}^K a_{k+1} \cdot \mathbf{P}_x(Z_t \in O_k) = \sum_{k=0}^K a_{k+1} \cdot \left(\frac{1}{T}\sum_{t=1}^T \mathbf{P}_x(Z_t \in O_k)\right) \leq c+1. \tag{69}$$

Since by construction O_k and $\mathcal{I}\times O_k$ are open sets on $\Re^{n_r}$ and S, respectively, by Cor. 5.2 [cf. Eq. (65)] and [MT09, Theorem D.5.4] we have for all k,

$$\liminf_{T\to\infty} \frac{1}{T}\sum_{t=1}^T \mathbf{P}_x(Z_t \in O_k) \geq \pi(\mathcal{I} \times O_k).$$

Combining this with Eq. (69), we have for all $K \geq 0$,

$$\sum_{k=0}^K a_{k+1} \cdot \pi(\mathcal{I} \times O_k) \leq \sum_{k=0}^K a_{k+1} \cdot \left(\liminf_{T\to\infty} \frac{1}{T}\sum_{t=1}^T \mathbf{P}_x(Z_t \in O_k)\right)$$

$$\leq \liminf_{T\to\infty} \sum_{k=0}^K a_{k+1} \cdot \left(\frac{1}{T}\sum_{t=1}^T \mathbf{P}_x(Z_t \in O_k)\right)$$

$$\leq c+1.$$

This implies

$$\sum_{k=0}^\infty a_{k+1} \cdot \pi(\mathcal{I} \times O_k) \leq c+1. \tag{70}$$

We now bound $E_\pi[\|Z_0\|]$. Consider two choices $\{a_k^1\}$, $\{a_k^2\}$ of the sequence $\{a_k\}$ in (67) with $a_1^1 = 1, a_1^2 = 1/2$, and denote their corresponding open sets in (68) by O_k^1, O_k^2, respectively. Since the sets $O_k^1, O_k^2, k \geq 0$ together cover the space of z except for the origin, we have

$$\|Z_0\| \leq \|Z_0\| \sum_{k=0}^\infty \left(\mathbf{1}_{O_k^1}(Z_0) + \mathbf{1}_{O_k^2}(Z_0)\right) \leq \sum_{k=0}^\infty \left(a_{k+1}^1 \cdot \mathbf{1}_{O_k^1}(Z_0) + a_{k+1}^2 \cdot \mathbf{1}_{O_k^2}(Z_0)\right),$$

where the second inequality follows from the definition of O_k^1, O_k^2. Therefore,

$$\begin{aligned} E_\pi[\|Z_0\|] &\leq E_\pi\Big[\sum_{k=0}^\infty \left(a_{k+1}^1 \cdot \mathbf{1}_{O_k^1}(Z_0) + a_{k+1}^2 \cdot \mathbf{1}_{O_k^2}(Z_0)\right)\Big] \\ &= \sum_{k=0}^\infty a_{k+1}^1 \cdot \pi(\mathcal{I} \times O_k^1) + \sum_{k=0}^\infty a_{k+1}^2 \cdot \pi(\mathcal{I} \times O_k^2) \\ &\leq 2(c+1), \end{aligned}$$

where the last inequality follows from Eq. (70). This completes the proof. $\qquad\square$

We can now prove Theorem 3.3, which states that with $\gamma_t = 1/(t+1)$, for each initial condition (z_0, G_0), $G_t \overset{a.s.}{\to} G^*$, and $G^* = E_\pi[Z_0\psi(i_0, i_1)']$.

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 $Z_0 = z_0$:

$$G_t = \frac{1}{t+1}\Big(\sum_{k=1}^t Z_k\psi(i_k, i_{k+1})' + G_0\Big), \qquad \widetilde{G}_t = \frac{1}{t+1}\sum_{k=0}^t Z_k\psi(i_k, i_{k+1})'.$$

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. By Prop. 5.3, $E_\pi\|Z_0\psi(i_0, i_1)'\| < \infty$. Therefore, applying the law of large numbers (see [MT09, Theorem 17.1.2]) to the stationary Markov process $\{(i_t, Z_t, i_{t+1})\}$ under $\mathbf{P}_\pi$, it can be seen that for each initial condition $x = (\bar{i}, \bar{z})$ from a measurable set F with $\pi(F) = 1$, $\widetilde{G}_t \overset{a.s.}{\to} G_x$, a random variable (which is a function of x and i_1), and consequently, $G_t \overset{a.s.}{\to} G_x$. But the second part of Theorem 3.1 implies there exists a subsequence $G_{t_k} \overset{a.s.}{\to} G^*$, so G_x is degenerate and $G_x = G^*$ a.s. Hence for any initial condition $x \in F$, $G_t \overset{a.s.}{\to} G^*$.

Furthermore, we have $G^* = E_\pi[Z_0\psi(i_0, i_1)']$. This is because the fact $G_x = G^*$ $\mathbf{P}_x$-a.s. for all $x \in F$ and $\pi(F) = 1$ implies $G^* = E_\pi[G^*] = E_\pi[G_{X_0}]$ where $X_0 = (i_0, Z_0)$, while by the law of large numbers for stationary processes (see [Doo53, Theorem 2.1] or [MT09, Theorem 17.1.2] and its proof), $E_\pi[G_{X_0}] = E_\pi[Z_0\psi(i_0, i_1)']$.

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 $\{(\hat{Z}_t, \hat{G}_t)\}$ and $\{(Z_t, G_t)\}$ corresponding to the two initial conditions $\hat{x} \notin F$ and $\bar{x} \in F$, respectively, and for the same path of $\{i_t\}$. By Lemma 4.3, 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

$$\frac{1}{t+1}\sum_{k=1}^t \beta^k L_0^k \overset{a.s.}{\to} 0. \tag{71}$$

Therefore,

$$\|\hat{G}_t - G_t\| = \Big\|\frac{1}{t+1}\sum_{k=1}^t (\hat{Z}_k - Z_k)\psi(i_k, i_{k+1})'\Big\| \leq c\|\Delta\|\Big(\frac{1}{t+1}\sum_{k=1}^t \beta^k L_0^k\Big),$$

where $c = \max_{i,j} \|\psi(i,j)\|$. By Eq. (71), this implies $\hat{G}_t - G_t \xrightarrow{a.s.} 0$. We have $G_t \xrightarrow{a.s.} G^*$ (because its initial condition $\bar{x} \in F$, as we just proved); therefore $\hat{G}_t \xrightarrow{a.s.} G^*$.

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

6 Discussion

We have analyzed the convergence and boundedness properties of the off-policy LSTD(λ) algorithm for Q-factor approximation in discounted total cost MDP. In this section, we discuss briefly the application of our results in three other contexts: (i) cost approximation (instead of the Q-factor approximation that we considered); (ii) policy evaluation under the average cost criterion; and (iii) approximately solving general linear fixed point equations with TD methods. We then conclude the paper by addressing some topics for future research.

Cost Approximation

We consider first approximating the costs of the target policy in the original MDP. In this case, the off-policy LSTD(λ) algorithm differs slightly form the one for approximating the Q-factors, which was considered in the paper; but it can be cast into a form that fits the framework of our analysis.

More specifically, let $\{(s_0, u_0), (s_1, u_1), \ldots\}$ be a trajectory of state-action pairs generated under the behavior policy. For each state s, let $q(u \mid s)$ and $p(u \mid s)$ denote the probability of taking action u under the target and the behavior policies, respectively, and let $c(s, u, \hat{s})$ be the one-stage cost of transition from s to a successor state $\hat{s}$ under action u. Suppose the approximation subspace is $\{\widehat{\Phi} r \mid r \in \Re^{n_r}\}$, where each row vector of the matrix $\widehat{\Phi}$ represents numerical "features" of some state s and is denoted by $\hat{\phi}(s)'$. Then, for approximating the cost function $J^*(s)$ of the target policy, the LSTD(λ) iterates $Z_t, b_t, C_t, t \geq 1$ can be defined as

$$Z_t = \lambda \alpha \frac{q(u_{t-1} \mid s_{t-1})}{p(u_{t-1} \mid s_{t-1})} \cdot Z_{t-1} + \hat{\phi}(s_t),$$

$$b_t = (1 - \tfrac{1}{t+1}) b_{t-1} + \tfrac{1}{t+1} Z_t \cdot \frac{q(u_t \mid s_t)}{p(u_t \mid s_t)} \cdot c(s_t, u_t, s_{t+1}),$$

$$C_t = (1 - \tfrac{1}{t+1}) C_{t-1} + \tfrac{1}{t+1} Z_t \left(\alpha \frac{q(u_t \mid s_t)}{p(u_t \mid s_t)} \cdot \hat{\phi}(s_{t+1}) - \hat{\phi}(s_t) \right)',$$

with $Z_0 = z_0, b_0$ and C_0 being the initial condition. Note that in the definition of b_t, the transition cost $c(s_t, u_t, s_{t+1})$ is multiplied by the ratio $\frac{q(u_t \mid s_t)}{p(u_t \mid s_t)}$. This is a small difference between the above algorithm and the one given by Eqs. (4)-(6) in Section 1 for Q-factor approximation.

To apply our analysis to the above LSTD(λ) algorithm for cost approximation, again we assume that the behavior policy induces an irreducible Markov chain $\{(s_t, u_t)\}$ on the space of state-action pairs, and that $p(u \mid s) = 0 \Rightarrow q(u \mid s) = 0$ for all states s and actions u. However, we consider the Markov chain $\{i_t, t \geq 0\}$ on the space of *action-state* pairs with $i_t = (u_{t-1}, s_t)$, where u_{-1} is immaterial and can be defined arbitrarily. More precisely, we define the space $\mathcal{I}$ of i_t to be the set of action-state pairs (v, s) such that there exists some state $\tilde{s}$ with $p(v \mid \tilde{s}) p(s \mid \tilde{s}, v) > 0$. It can be seen that with this definition of $\mathcal{I}$, under the behavior policy, $\{i_t\}$ is an irreducible Markov chain on $\mathcal{I}$ if $\{(s_t, u_t)\}$ is irreducible.

We define $\phi(i) = \hat{\phi}(s)$ for a state $i = (v, s) \in \mathcal{I}$ of the Markov chain $\{i_t\}$. For a pair of states $i = (v, s), j = (u, \hat{s}) \in \mathcal{I}$, we define the cost of transition from i to j to be $g(i, j) = c(s, u, \hat{s})$. It can be seen that under the behavior policy, the probability of transition from $i = (v, s)$ to $j = (u, \hat{s})$ is $p(u \mid s) p(\hat{s} \mid s, u)$, whereas under the target policy, it would be $q(u \mid s) p(\hat{s} \mid s, u)$. Therefore $\frac{q_{ij}}{p_{ij}} = \frac{q(u \mid s)}{p(u \mid s)}$, and in particular, $\frac{q_{i_t i_{t+1}}}{p_{i_t i_{t+1}}} = \frac{q(u_t \mid s_t)}{p(u_t \mid s_t)}$, where we define $0/0 = 0$. (This shows also that u_{-1} is immaterial, as we claimed.) We can now cast the above LSTD iterates in the form of the

iterates (Z_t, G_t), which we analyzed in the paper:

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

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

where $\beta = \lambda\alpha$, $\gamma_t = 1/(t+1)$, and

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

The assumption that $p(u \mid s) = 0 \Rightarrow q(u \mid s) = 0$ for all s and u is equivalent to $Q \prec P$. Thus all of our results apply to the cost approximation case.

Policy Evaluation in Average Cost MDP

We consider now approximate policy evaluation in MDP with the average cost criterion and the application of the off-policy LSTD(λ) algorithm in this context. We consider Q-factor approximation; the case of cost approximation is similar, as we just discussed. Assuming that the target policy induces a Markov chain on the state-action space with a single recurrent class, its average cost η^* (a scalar) and differential cost vector J^* together satisfy the Bellman equation

$$J^* = (g - \eta^* e) + QJ^*,$$

where e is the vector of all ones. In the on-policy case, using almost surely convergent on-line estimates of η^*, we can apply TD(λ) to solve a projected multistep Bellman equation

$$J = \Pi T^{(\lambda)}(J), \qquad \text{where} \quad T(J) = (g - \eta^* e) + QJ,$$

and obtain approximate differential costs (Tsitsiklis and Van Roy [TV99]), similar to the discounted case. In the off-policy case, convergent estimates of η^* are not straightforward to obtain by simple averaging or iterative computation. One possibility is to first approximate η^* by some $\hat\eta$ and then solve the projected equation

$$J = \Pi\widehat{T}^{(\lambda)}(J), \qquad \text{where} \quad \widehat{T}(J) = (g - \hat\eta e) + QJ.$$

The latter can be done by the off-policy LSTD(λ) with $\lambda < 1$ and our convergence analysis also applies in this case. The solution of the projected equation, when it exists, approximates the differential cost vector J^*. However, by the average cost MDP theory (see e.g., Puterman [Put94], Bertsekas [Ber07]), unless $\hat\eta = \eta^*$ the corresponding equation without the projection, $J = \widehat{T}^{(\lambda)}(J)$ does not have a solution. This is different from the discounted case.

As to the approximate average cost $\hat\eta$, it can be obtained from either the finite stage costs of the target policy, or by its discounted costs for a discount factor close to 1, based on the well-known relation between the discounted costs and the average cost (see e.g., [Put94, Ber07]). The latter approximation can be computed using the off-policy LSTD(λ) for the discounted problem.

An alternative approach to approximate policy evaluation is to approximate the average cost problem by a discounted one, and to derive from the approximate discounted costs an approximation of the pair (η^*, J^*) simultaneously, based on the relation between the average cost and the discounted problems (see e.g., [Put94, Ber07]). In this case the off-policy LSTD(λ) algorithm we analyzed is certainly applicable.

Linear Fixed Point Equations

We discuss next a direct extension of our analysis to the context of approximately solving a linear fixed point equation

$$x = T(x) = Ax + b$$

with TD methods, as discussed in Bertsekas and Yu [BY09]. Here A is an $n \times n$ matrix and b an n-dimensional vector. Compared with the policy evaluation case, the main difference is that the substochastic matrix αQ in the Bellman equation (2) is replaced by an arbitrary matrix A.

Let $|A|$ be the signless version of A, with the (i,j)th entry being $|a_{ij}|$. Then for $\lambda \in (0,1]$ such that $\lambda|A|$ is strictly substochastic in the sense that

$$\lambda \sum_j |a_{ij}| < 1, \qquad \forall i,$$

we can define analogously the multistep fixed point mapping $T^{(\lambda)}$ involving the matrix $\sum_{k=0}^{\infty} \lambda^k A^k$, and find an approximate solution of $x = T(x)$ by solving $x = \Pi T^{(\lambda)}(x)$ using simulation-based algorithms. In particular, we can treat the indices of rows and columns as states and employ a Markovian row/column sampling scheme described by a transition matrix P, and apply the off-policy LSTD(λ) algorithm with the coefficients αq_{ij} replaced by a_{ij}, as described in [BY09].

Similarly, the analysis in the present paper applies in this more general context, assuming the irreducibility of P and $|A| \prec P$, in addition to $\lambda|A|$ being strictly substochastic. The slight modification we need when bounding various quantities of interest is to replace the ratios $L_{t-1}^t = \frac{a_{i_{t-1}i_t}}{p_{i_{t-1}i_t}}$, now possibly negative, by their absolute values, and to use the fact that

$$E\big[\lambda|L_{t-1}^t| \mid i_{t-1}\big] \leq \nu < 1$$

for some constant ν. A slightly more general case where $\lambda \sum_j |a_{ij}| \leq 1$ for all i and with equality for some but not all i, may be analyzed using a similar approach.

Future Research

We mention some issues in extending our analysis to countable or continuous state space MDP. In this case, we will need additional conditions on the Markov chain $\{i_t\}$ induced by the behavior policy, as well as on the relation between the chain $\{i_t\}$ and the basis functions $\phi(i_t)$ used in the approximation, in order to ensure that the multistep Bellman equation is well defined and that $\{(Z_t, G_t)\}$ behaves properly. This is because the sequence $\{\phi(i_t)\}$ can be unbounded. As a subject for future research, we may consider imposing suitable drift conditions and combine them with the e-chain-based analysis in the present paper.

There are some other problems that deserve future study. One is the convergence of various off-policy TD(λ) algorithm variants for a general value of λ, as mentioned in the introduction. (In the case of $\lambda = 0$, there are several convergent gradient-based off-policy TD variants; see Sutton et al. [SMP+09] and the references therein.) Another is the finite sample properties of these algorithms as well as LSTD, analogous to those considered by Antos et al. [ASM08]. A third one is the question of the almost sure convergence of LSTD(λ) with a general stepsize sequence (possibly random). Such stepsizes are useful particularly when LSTD(λ) is applied to policy evaluation at a faster time-scale in two-time-scale policy iteration schemes with incremental policy improvement at a slower time-scale.

Acknowledgments

References

[ABJ06]   T. P. Ahamed, V. S. Borkar, and S. Juneja, *Adaptive importance sampling technique for Markov chains using stochastic approximation*, Operations Research **54** (2006), 489–504.

[ASM08]   A. Antos, Cs. Szepesvári, and R. Munos, *Learning near-optimal policies with Bellman residual minimization based fitted policy iteration and a single sample path*, Machine Learning **71** (2008), 89–129.

[BB96]   S. J. Bradtke and A. G. Barto, *Linear least-squares algorithms for temporal difference learning*, Machine Learning **22** (1996), no. 2, 33–57.

[Ber07]   D. P. Bertsekas, *Dynamic programming and optimal control*, third ed., vol. II, Athena Scientific, Belmont, MA, 2007.

[Ber09]   ———, *Projected equations, variational inequalities, and temporal difference methods*, LIDS Tech. Report 2808, MIT, 2009, to appear in *IEEE Trans. Automat. Contr.*

[Bor06]   V. S. Borkar, *Stochastic approximation with 'controlled Markov' noise*, Systems Control Lett. **55** (2006), 139–145.

[Bor08]   ———, *Stochastic approximation: A dynamic viewpoint*, Hindustan Book Agency, New Delhi, 2008.

[Boy99]   J. A. Boyan, *Least-squares temporal difference learning*, Proc. The 16th Int. Conf. Machine Learning, 1999, pp. 49–56.

[Bre92]   L. Breiman, *Probability*, SIAM, Philadelphia, PA, 1992, (Originally published by Addison-Wesley, 1968).

[BT96]   D. P. Bertsekas and J. N. Tsitsiklis, *Neuro-dynamic programming*, Athena Scientific, Belmont, MA, 1996.

[BY09]   D. P. Bertsekas and H. Yu, *Projected equation methods for approximate solution of large linear systems*, J. Computational and Applied Mathematics **227** (2009), no. 1, 27–50.

[Doo53]   J. L. Doob, *Stochastic processes*, John Wiley & Sons, New York, 1953.

[GI89]   P. W. Glynn and D. L. Iglehart, *Importance sampling for stochastic simulations*, Management Science **35** (1989), 1367–1392.

[KY03]   H. J. Kushner and G. G. Yin, *Stochastic approximation and recursive algorithms and applications*, 2nd ed., Springer-Verlag, New York, 2003.

[Mey89]   S. Meyn, *Ergodic theorems for discrete time stochastic systems using a stochastic Lyapunov function*, SIAM J. Control Optim. **27** (1989), 1409–1439.

[Mey07]   ———, *Control techniques for complex networks*, Cambridge University Press, Cambdrige, UK, 2007.

[MT09]   S. Meyn and R. L. Tweedie, *Markov chains and stochastic stability*, 2nd ed., Cambridge University Press, Cambdrige, UK, 2009.

[NB03]   A. Nedić and D. P. Bertsekas, *Least squares policy evaluation algorithms with linear function approximation*, Discrete Event Dyn. Syst. **13** (2003), 79–110.

[PSD01]   D. Precup, R. S. Sutton, and S. Dasgupta, *Off-policy temporal-difference learning with function approximation*, Proc. The 18th Int. Conf. Machine Learning, 2001, pp. 417–424.

[Put94]   M. L. Puterman, *Markov decision processes: Discrete stochastic dynamic programming*, John Wiley & Sons, 1994.

[SB98]   R. S. Sutton and A. G. Barto, *Reinforcement learning*, MIT Press, Cambridge, MA, 1998.

[SMP$^+$09] R. S. Sutton, H. R. Maei, D. Precup, S. Bhatnagar, D. Silver, C. Szepesvári, and E. Wiewiora, *Fast gradient-descent methods for temporal-difference learning with linear function approximation*, Proc. The 26th Int. Conf. Machine Learning, 2009.

[Sut88]   R. S. Sutton, *Learning to predict by the methods of temporal differences*, Machine Learning **3** (1988), 9–44.

[TV97]   J. N. Tsitsiklis and B. Van Roy, *An analysis of temporal-difference learning with function approximation*, IEEE Trans. Automat. Contr. **42** (1997), no. 5, 674–690.

[TV99]   _____, *Average cost temporal-difference learning*, Automatica **35** (1999), no. 11, 1799–1808.

[YL08]   H. S. Yao and Z. Q. Liu, *Preconditioned temporal difference learning*, Proc. The 25th Int. Conf. Machine Learning, 2008, pp. 1208–1215.