Technical Report C-2010-39
Dept. Computer Science
University of Helsinki
Sep 2010

Least Squares Temporal Difference Methods: An Analysis 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) with the least squares temporal difference algorithm, LSTD(λ), in an exploration-enhanced off-policy learning context. 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. Our analysis draws on theories of both finite space Markov chains and weak Feller Markov chains on topological spaces. Our results can be applied to other temporal difference algorithms and MDP models. As examples, we give a convergence analysis of an off-policy TD(λ) algorithm and extensions to MDP with compact action and state spaces.

Keywords: Markov decision processes, approximate dynamic programming, temporal difference methods, importance sampling, Markov chains

*This technical report is a revised and extended version of the technical report C-2010-1. It contains simplified and improved proofs, as well as extensions of some of the earlier results.

Contents

1 Introduction

We consider approximate policy evaluation for Markov decision processes (MDP) in an exploration-enhanced learning context, commonly referred to as "off-policy" learning in the terminology of reinforcement learning. In this context, we employ a certain policy called the "behavior policy" to adequately explore the state and action spaces, and using the observations of costs and transitions generated under the behavior policy, we may approximately evaluate any suitable "target policy" of interest. Off-policy learning differs from "on-policy" learning – the standard policy evaluation, where the behavior policy always coincides with the policy to be evaluated. The dichotomy between the two stems from the exploration-exploitation tradeoff in practical model-free/simulation-based methods for policy search. With their flexibility, methods for off-policy learning form an important part of the model-free reinforcement learning methodology (Sutton and Barto [SB98]). They have also been suggested as an important class of importance-sampling based techniques (Glynn and Iglehart [GI89]) in the broad context of simulation-based methods for large-scale dynamic programming. In this context, any sampling mechanism may play the role of the behavior policy, inducing system dynamics that may not be realizable under any policy, for the purpose of efficient policy evaluation.

We focus primarily on finite state and action MDP, and we consider discounted total cost problems with discount factor $\alpha < 1$. When the MDP model is unavailable or when simulation is involved, there are two common approaches to evaluating a stationary target policy: evaluating its costs, and evaluating its so-called Q-factors, which are expected total discounted costs associated with initial state-action pairs. In either case, the function to be evaluated can be viewed as the cost function of the policy on a finite space $\mathcal{I} = \{1, 2, \ldots, n\}$, on which the policy induces a homogeneous Markov chain, and the goal is to solve a corresponding Bellman equation on $\mathcal{I}$ satisfied by the cost function. The Bellman equation in matrix notation has the form

$$J = \bar{g} + \alpha Q J, \qquad J \in \Re^n, \tag{1}$$

where $\bar{g}$ is the vector of expected one-stage costs and Q the transition matrix of the Markov chain on $\mathcal{I}$ associated with the target policy. The cost vector J^* of the target policy is the unique solution of the Bellman equation.

Our focus will be on a particular algorithm for policy evaluation with function approximation and exploration-enhancements, which will be referred to in this paper as the off-policy least squares temporal difference (LSTD) algorithm. It is a counterpart of the on-policy LSTD algorithm for policy evaluation (Bradtke and Barto [BB96], Boyan [Boy99]), and it was first given by Bertsekas and Yu [BY09] in the general context of approximating solutions of linear systems of equations. It belongs to the family of temporal difference (TD) methods (Sutton [Sut88]; see also the books by Bertsekas and Tsitsiklis [BT96], Sutton and Barto [SB98], Bertsekas [Ber07], and Meyn [Mey07]). Beyond the algorithmic level, TD methods share a common approximation framework which involves multistep Bellman equations and projected equations. In this framework, we consider a projected version of a multistep Bellman equation parametrized by $\lambda \in [0, 1]$,

$$J = \Pi T^{(\lambda)}(J), \tag{2}$$

where $T^{(\lambda)}$ is a multistep Bellman operator associated with the target policy and parametrized by $\lambda \in [0, 1]$, whose exact form will be given later, and Π is the projection onto an approximation subspace $\{\Phi r \mid r \in \Re^d\} \subset \Re^n$. The projection here is with respect to a weighted Euclidean norm. The weights in the projection norm, in the off-policy case that we consider, are the only quantities related to the behavior policy; they are the steady-state probabilities of the Markov chain induced by the behavior policy. When the projected equation (2) is well defined, i.e., has a unique solution Φr^* in the approximation subspace, we use the solution to approximate the cost vector J^* of the target policy. There are general approximation error bounds (Yu and Bertsekas [YB10]) and geometric interpretations of the approximation (Scherrer [Sch10]) in this case. Our interest in this paper, however, will not be in whether the projected Bellman equation is well defined, but rather in the approximation of the equation using sampling and the off-policy LSTD(λ) algorithm.

For any given λ, the projected Bellman equation (2) is equivalent to a low dimensional linear equation on $\Re^d$, which may be written as

$$\bar{C}r + \bar{b} = 0, \qquad r \in \Re^d, \tag{3}$$

where $\bar{b}$ is a d-dimensional vector and $\bar{C}$ a $d \times d$ matrix. The precise definitions of $\bar{b}, \bar{C}$ will be given later. The off-policy LSTD(λ) algorithm that we will analyze constructs a sequence of equations

$$C_t r + b_t = 0, \qquad t \geq 1,$$

using observations generated under the behavior policy, with the goal of "approaching" in the limit Eq. (3), the low dimensional representation of (2). The algorithm takes into account the discrepancies between the behavior and the target policies by properly weighting the observations. The technique is based on importance sampling, which is widely used in dynamic programming and reinforcement learning contexts; see e.g., Glynn and Iglehart [GI89], Sutton and Barto [SB98], Precup et al. [PSD01], (which is one of the first off-policy TD(λ) algorithms), and Ahamed et al. [ABJ06].

The assumptions underlying the off-policy LSTD(λ) algorithm are that every state (in the case of cost approximation) or state-action pair (in the case of Q-factor approximation) is visited infinitely often under the behavior policy, and for every state, possible actions of the target policy are also possible actions of the behavior policy. These are natural, minimal requirements for off-policy learning. In terms of transition probabilities, the assumptions can be expressed as follows. Let $P = [p_{ij}]$ be the transition matrix of the Markov chain on $\mathcal{I}$ induced by the behavior policy. We require that this Markov chain is irreducible, and that the transition matrix $Q = [q_{ij}]$ associated with the target policy is absolutely continuous with respect to P in the sense that

$$p_{ij} = 0 \quad \Rightarrow \quad q_{ij} = 0, \quad i, j \in \mathcal{I}. \tag{4}$$

We denote the latter condition by $Q \prec P$.

In this paper we analyze the convergence of the off-policy LSTD(λ) algorithm – the convergence of $\{(b_t, C_t)\}$ to $(\bar{b}, \bar{C})$ – for all $\lambda \in [0, 1]$ under the general conditions given above. Prior to our work, the almost sure convergence of the algorithm (i.e., convergence with probability one) in special cases has been studied. A proof under the additional 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 additional condition is technically convenient because it guarantees the boundedness of a key sequence in the algorithm (the sequence $\{Z_t\}$ defined in Section 2 and to be mentioned below), but it is restrictive. It either requires the behavior policy to be very similar to the target policy, or restricts λ to be close to 0, while the case of a general value of λ is important in practice. Using a large value of λ can not only improve the quality of the cost approximation obtained from the projected Bellman equation, but can also avoid 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).

As the main results of this paper, we establish for all $\lambda \in [0, 1]$, the almost sure convergence of the sequences $\{b_t\}, \{C_t\}$, as well as their convergence in the first mean, under the assumptions of the irreducibility of P and $Q \prec P$. These results imply in particular that the off-policy LSTD(λ) solution Φr_t converges to the solution Φr^* of the projected Bellman equation (2) almost surely, whenever Eq. (2) has a unique solution, and if (2) has multiple solutions, any limit point of $\{\Phi r_t\}$ is one of them.

On the technical side, the line of our analysis is considerably different from those in the literature for similar type TD algorithms. In an iterative form, the off-policy LSTD(λ) looks very close to the on-policy LSTD(λ) counterpart (Bradtke and Barto [BB96], Boyan [Boy99]), and also bears similarities to the on-policy TD(λ) (Sutton [Sut88], Tsitsiklis and Van Roy [TV97]) and the off-policy TD(λ) given in Precup et al. [PSD01]. When $\lambda > 0$, to facilitate iterative computation, all the algorithms calculate iteratively an auxiliary sequence of vectors Z_t, (sometimes called the "eligibility traces"), where each Z_t is a function of the entire set of past observations up to the time t. However, in the off-policy case, without restricting the value of λ, the sequence $\{Z_t\}$ is not necessarily bounded, and neither does it necessarily have uniformly bounded variances. Indeed, we will show in the paper that in fairly common situations, $\{Z_t\}$ is almost surely unbounded. It is also not difficult to construct examples where $\{Z_t\}$ has unbounded variances or unbounded νth order moments with $\nu > 1$. In the on-policy case, the bounded variance property of $\{Z_t\}$ has been relied on by the convergence proofs for TD(λ) (Tsitsiklis and Van Roy [TV97]) and LSTD(λ) (Nedić and Bertsekas [NB03]). The analyses in [NB03, BY09] also use the boundedness of $\{Z_t\}$, so does [PSD01], which calculates Z_t only for state trajectories of a predetermined finite length. Therefore for the convergence analysis in the off-policy case with a general value of λ, we do not follow the approaches in these works, and instead we will relate the off-policy LSTD(λ) iterates to particular type of Markov chains and resort to the ergodic theory for these chains [MT09, Mey89].

Let us also mention a proof approach from stochastic approximation theory, the mean-o.d.e. method (see e.g., Kushner and Yin [KY03], Borkar [Bor06, Bor08]). It requires the verification of conditions that in our case would be tantamount to the almost sure convergence conclusion we want to establish.

As we will show, the convergence of $\{b_t\}, \{C_t\}$ in the first mean can be established using arguments based on the ergodicity of the finite space Markov chain $\{i_t\}$ on $\mathcal{I}$ induced by the behavior policy. But for proving their almost sure convergence, we did not find such arguments to be sufficient, in contrast with the on-policy LSTD case as analyzed by Meyn [Mey07, Chap. 11.5]. Instead, we will study the Markov chain $\{(i_t, Z_t)\}$ on the topological space $\mathcal{I} \times \Re^d$. We will exploit the weak Feller property of the chain $\{(i_t, Z_t)\}$, as well as its other properties, to establish two results: (i) the Markov chain $\{(i_t, Z_t)\}$ has a unique invariant probability measure and is ergodic (in the sense of weak convergence of occupation measures), and (ii) the sequences $\{b_t\}, \{C_t\}$ converge almost surely to $\bar{b}, \bar{C}$, respectively, (and hence the off-policy LSTD(λ) algorithm also converges almost surely).

We note 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) need the same approximations b_t, C_t to build approximating models (e.g., preconditioned TD(λ) in Yao and Liu [YL08]) or fixed point iterations (e.g., LSPE(λ), see Bertsekas and Yu [BY09]; and scaled versions of LSPE(λ), see Bertsekas [Ber09]). 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. Moreover, the mode of convergence of the off-policy LSTD(λ) is also relevant for understanding the behavior of other off-policy TD algorithms which use stochastic approximation type iterations to solve projected Bellman equations (3), for instance, the on-line off-policy TD(λ) algorithm of [BY09], and the off-policy TD(λ) algorithm of [PSD01] in the case where it uses very long trajectories to update Z_t. Although these algorithms do not directly compute approximations b_t, C_t, they implicitly depend on the convergence properties of $\{b_t\}, \{C_t\}$. Thus our results and our line of analysis are useful also for analyzing various off-policy TD algorithms other than LSTD.

Besides the main results mentioned above, this paper contains some additional results. In particular, we will combine our convergence results with stochastic approximation theory to prove the convergence of a constrained version of an on-line off-policy TD(λ) algorithm proposed in [BY09]. We will also extend our results to special cases of MDP with compact state and action spaces.

The paper is organized as follows. We specify notation and definitions in Section 2. We present our main convergence results for the off-policy LSTD(λ) algorithm in finite space MDP in Section 3. We then give in Section 4 additional results on the convergence of a constrained off-policy TD(λ) algorithm and the extension of our analysis to MDP with compact spaces. Finally, we discuss other applications of our results and future research in Section 5.

2 Notation and Background

We consider stationary randomized target and behavior policies and the evaluation of a target policy by using observations of transitions and costs generated under the behavior policy. For notational simplicity, let $\mathcal{I} = \{1, \ldots, n\}$ denote a certain set of state and action pairs, on which it is assumed that the behavior and the target policies induce Markov chains with transition matrices P and Q, respectively. Our discussion will be centered on these two chains. Their particular forms differ slightly for Q-factor approximation and cost approximation (see Examples 2.1, 2.2), and will not be central to our analysis. Throughout the paper, we use $\{i_t\}$ to denote the Markov chain with transition matrix P, and use i or $\bar{i}$ to denote specific states. We assume the following condition on P and Q, as mentioned in the introduction.

Assumption 2.1. *The Markov chain $\{i_t\}$ with transition matrix P is irreducible, and $Q \prec P$ in the sense of Eq. (4).*

By the standard MDP theory (see Bertsekas [Ber05], Puterman [Put94]), the cost function J^* of the policy associated with transition matrix Q satisfies the Bellman equation

$$J = T(J), \qquad \text{where } \; T(J) = \bar{g} + \alpha Q J, \quad \forall J \in \Re^n,$$

and $\bar{g}$ is the vector of expected one-stage costs under that policy. We define a multistep Bellman operator parametrized by $\lambda \in [0, 1]$ by

$$T^{(\lambda)} = (1 - \lambda) \sum_{m=0}^{\infty} \lambda^m T^{m+1}, \quad \lambda \in [0, 1); \qquad T^{(1)}(J) = \lim_{\lambda \to 1} T^{(\lambda)}(J), \qquad \forall J \in \Re^n. \tag{5}$$

($T^{(0)} = T$ in particular.) It appears in the projected Bellman equation (2), $J = \Pi T^{(\lambda)}(J)$, associated with the TD(λ) methods.

We approximate J^* by a vector in a subspace of $\Re^n$, which has a representation $\{\Phi r \,|\, r \in \Re^d\}$ for some $n \times d$ matrix Φ whose columns span the approximation subspace. While any of such representations is mathematically equivalent, in practice, often some subspace-determining matrix Φ is first chosen based on one's understanding of the problem at hand. Typically Φ need not be stored because one has access to the function ϕ which maps $i \in \mathcal{I}$ to the ith row of Φ. The vectors $\phi(i)$ are often referred to as "features" of states/actions and are treated here as $d \times 1$ vectors, so Φ can be expressed in terms of $\phi(i)$ as $\Phi' = \begin{bmatrix} \phi(1) & \phi(2) & \cdots & \phi(n) \end{bmatrix}$, while the components of the function ϕ span the approximation subspace. Choosing the "feature-mapping" ϕ is extremely important in practice but is beyond the scope of this paper.

We define the projection Π onto the approximation subspace to be with respect to a weighted Euclidean norm. The weights in the norm are the steady-state probabilities of the Markov chain with transition matrix P, and are well defined under our irreducibility assumption on P. To derive a low-dimensional representation of the projected Bellman equation (2) in terms of r, let Ξ denote the diagonal matrix with the diagonal elements being these steady-state probabilities. Equation (2) is equivalent to

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

and by rearranging terms, it can be written as

$$\bar{C} r + \bar{b} = 0, \tag{6}$$

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

$$\bar{b} = \Phi' \Xi \sum_{m=0}^{\infty} \lambda^m (\alpha Q)^m \bar{g}, \qquad \bar{C} = \Phi' \Xi \sum_{m=0}^{\infty} \lambda^m (\alpha Q)^m (\alpha Q - I) \Phi. \tag{7}$$

The off-policy LSTD(λ) algorithm [BY09, Sec. 5.2] computes iteratively vectors b_t and matrices C_t, using observations generated under the policy associated with transition matrix P. The vector b_t and matrix C_t aim to approximate the quantities $\bar{b}$ and $\bar{C}$ (respectively), which define the projected Bellman equation (6), equivalently (2). To facilitate iterative computation, the algorithm also computes a third sequence of d-dimensional vectors Z_t. These iterates are defined as follows. Let $g(i,j)$ denote the one-stage cost function of transition from i to j, which relates to the expected one-stage cost $\bar{g}(i)$ by $\bar{g}(i) = \sum_{j\in\mathcal{I}} q_{ij} g(i,j)$. With (z_0, b_0, C_0) being the initial condition, for $t \geq 1$,

$$Z_t = \lambda\alpha \tfrac{q_{i_{t-1}i_t}}{p_{i_{t-1}i_t}} \cdot Z_{t-1} + \phi(i_t), \tag{8}$$

$$b_t = (1-\gamma_t) b_{t-1} + \gamma_t Z_t \cdot \tfrac{q_{i_t i_{t+1}}}{p_{i_t i_{t+1}}} \cdot g(i_t, i_{t+1}), \tag{9}$$

$$C_t = (1-\gamma_t) C_{t-1} + \gamma_t Z_t \Big(\alpha \tfrac{q_{i_t i_{t+1}}}{p_{i_t i_{t+1}}} \cdot \phi(i_{t+1}) - \phi(i_t)\Big)'. \tag{10}$$

Here $\{\gamma_t\}$ is a stepsize sequence with $\gamma_t \in (0,1]$, and typically $\gamma_t = 1/(t+1)$ in practice. 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.[1]

In the standard on-policy case where $P = Q$, 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 with the typical stepsize $\gamma_t = 1/(t+1)$ reduces to the on-policy LSTD algorithm as first given by Bradtke and Barto [BB96] for $\lambda = 0$ and Boyan [Boy99] for $\lambda \in [0,1]$.

We are interested in whether $\{b_t\}, \{C_t\}$ converge to $\bar{b}, \bar{C}$ respectively, in some mode (in mean, with probability one, or in probability). As the two sequences $\{b_t\}$ and $\{C_t\}$ have the same iterative structure, we can consider just one sequence in a more general form to simplify notation:

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

with (z_0, G_0) being the initial condition. The sequence $\{G_t\}$ specializes to $\{b_t\}$ or $\{C_t\}$ with particular choices of the (vector-valued) function $\psi(i,j)$:

$$G_t = \begin{cases} b_t & \text{if } \ \psi(i,j) = \frac{q_{ij}}{p_{ij}} \cdot g(i,j), \\ C_t & \text{if } \ \psi(i,j) = \alpha \frac{q_{ij}}{p_{ij}} \cdot \phi(j) - \phi(i). \end{cases} \tag{12}$$

We will consider stepsize sequences $\{\gamma_t\}$ that satisfy the following condition. Such sequences include $\gamma_t = t^{-\nu}, \nu \in (0.5, 1]$, for example. When conclusions hold for a specific sequence $\{\gamma_t\}$, such as $\gamma_t = 1/t$, we will state them explicitly.

Assumption 2.2. *The sequence of stepsizes γ_t is deterministic and eventually nonincreasing, and satisfies $\gamma_t \in (0,1]$, $\sum_t \gamma_t = \infty$, $\sum_t \gamma_t^2 < \infty$.*

With this notation, the question of convergence of $\{b_t\}, \{C_t\}$ amounts to that of the convergence of $\{G_t\}$, in any mode, to the constant vector/matrix

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

where $\beta = \lambda\alpha$ and the vector/matrix Ψ is given in terms of its rows by

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

[1] 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, as mentioned in the introduction.

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)], $\Psi = \bar{g}$ or $(\alpha Q - I)\Phi$, and $G^* = \bar{b}$ or $\bar{C}$, respectively [cf. Eq. (7)].

Before proceeding to convergence analysis, we provide below specific details relating the above framework to practical implementations of the algorithm for Q-factor and cost approximations in the model-free learning context. These details will not be relied on in our analysis.

Example 2.1 (Q-factor approximation)**.** Suppose in the MDP, transition from state s to state $\hat{s}$ occurs according to the probability $p(\hat{s} \mid s, u)$ when taking an action u that is admissible at s, and the transition incurs cost $c(s, u, \hat{s})$, where c is a function of the transition and action. The Q-factor of a policy for each initial state and action pair (s, u) is the expected cost of first taking action u at the state s and then following the policy. For approximating Q-factors of the target policy, we let $\mathcal{I}$ correspond to the set of state-action pairs, and let the chain $\{i_t\}$ correspond to the process $\{(s_t, u_t)\}$ of states and actions induced by the behavior policy, with $i_t \sim (s_t, u_t)$, where "$\sim$" indicates association. For two state-action pairs $i \sim (s, u)$, $j \sim (\hat{s}, \hat{u})$, the probability of transition from i to j under a policy which takes action $\hat{u}$ at state $\hat{s}$ with probability $\mu(\hat{u} \mid \hat{s})$ is naturally given by $p(\hat{s} \mid s, u)\mu(\hat{u} \mid \hat{s})$. The transition matrices P and Q associated with the behavior and target policies are defined in this way. We can set the one-stage transition costs $g(i, j)$ and the corresponding expected one-stage costs $\bar{g}(i)$ to be

$$g(i,j) = c(s,u,\hat{s}), \qquad \bar{g}(i) = \sum_{\hat{s}} p(\hat{s} \mid s,u)\, c(s,u,\hat{s}), \qquad i,j \in \mathcal{I} \ \text{ with } \ i \sim (s,u), \ j \sim (\hat{s},\hat{u}).$$

By definition both $g(i, j)$ and $\bar{g}(i)$ do not depend on policies, which is special to the Q-factor evaluation scenario. Correspondingly, the updates for b_t in the off-policy LSTD(λ) algorithm can be simplified to

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

omitting the term $\frac{q_{i_t i_{t+1}}}{p_{i_t i_{t+1}}}$ before $g(i_t, i_{t+1})$ [cf. Eq. (9)]. The resulting sequence $\{b_t\}$ is a special case of the sequence $\{G_t\}$ given by Eq. (11) that we will analyze, with the function $\psi(i, j) = g(i, j)$.

In the model-free learning context, it is practically important that the ratios $\frac{q_{ij}}{p_{ij}}$ are functionally independent of the state transition dynamics $p(\hat{s} \mid s, u)$ of the MDP; they are equal to the ratios between the corresponding action probabilities of the target and the behavior policies, as can be seen from the above model description. Thus the n^2 terms $\frac{q_{ij}}{p_{ij}}$ need not be stored and can be calculated on-line in the off-policy LSTD(λ) algorithm. This is a well-known fact and finds use in many existing simulation-based algorithms for MDP. □

Example 2.2 (Cost approximation)**.** Let the MDP be as in the preceding example, and let $\{(s_t, u_t)\}$ be the process of states and actions induced by the behavior policy. Suppose we want to approximate the cost vector of the target policy in the MDP by a vector of the form $\hat{\phi}(s)'r$, where $\hat{\phi}$ maps states s to $d \times 1$ vectors. Then, given initial (z_0, b_0, C_0), the LSTD(λ) iterates can be defined as

$$Z_t = \lambda\alpha \tfrac{\mu(u_{t-1}|s_{t-1})}{\mu^o(u_{t-1}|s_{t-1})} \cdot Z_{t-1} + \hat{\phi}(s_t), \tag{14}$$

$$b_t = (1 - \gamma_t)b_{t-1} + \gamma_t Z_t \cdot \tfrac{\mu(u_t|s_t)}{\mu^o(u_t|s_t)} \cdot c(s_t, u_t, s_{t+1}), \tag{15}$$

$$C_t = (1 - \gamma_t)C_{t-1} + \gamma_t Z_t \left(\alpha \tfrac{\mu(u_t|s_t)}{\mu^o(u_t|s_t)} \cdot \hat{\phi}(s_{t+1}) - \hat{\phi}(s_t)\right)', \tag{16}$$

where $\mu(\cdot \mid s)$ and $\mu^o(\cdot \mid s)$ denote the conditional probabilities over actions at state s under the target and behavior policies, respectively, and it is required that $\mu(\cdot \mid s)$ is absolutely continuous with respect to $\mu^o(\cdot \mid s)$, i.e., $\mu^o(u \mid s) = 0 \Rightarrow \mu(u \mid s) = 0$. The above iterates can be cast in the form given by Eqs. (8)-(10) as follows.

We consider the Markov chain $\{i_t\}$ with $i_t \sim (u_{t-1}, s_t)$ (where "$\sim$" indicates association and the choice of u_{-1} is immaterial). We assume that every state s can be visited infinitely often under the behavior policy, and we let $\mathcal{I}$ be the set of action-state pairs (v, s) such that s is accessible from some state $\tilde{s}$ by taking action v under the behavior policy, i.e., $\mu^o(v \mid \tilde{s})p(s \mid \tilde{s}, v) > 0$. For any $i, j \in \mathcal{I}$ with $i \sim (v, s)$, $j \sim (u, \hat{s})$, let $\phi(i) = \hat{\phi}(s)$, and let the cost of transition from i to j be $g(i, j) = c(s, u, \hat{s})$. For the above i, j, the probability of transition from i to j under the target or behavior policy is $\mu(u \mid s)p(\hat{s} \mid s, u)$ or $\mu^o(u \mid s)p(\hat{s} \mid s, u)$, respectively. This defines the transition matrices P and Q. In particular, it can be seen that $\frac{q_{ij}}{p_{ij}} = \frac{\mu(u|s)}{\mu^o(u|s)}$ for the above i, j, and $\frac{q_{i_t i_{t+1}}}{p_{i_t i_{t+1}}} = \frac{\mu(u_t|s_t)}{\mu^o(u_t|s_t)}$, (where we define $0/0 = 0$). The off-policy LSTD(λ) algorithm for cost approximation given by Eqs. (14)-(16) then takes exactly the same form as the algorithm given by Eqs. (8)-(10). □

3 Main Results

We analyze the convergence of $\{G_t\}$ in mean and with probability one. For the former, we will use properties of the finite space Markov chain $\{i_t\}$, and for the latter, those of the topological space Markov chain $\{(i_t, Z_t)\}$. Along with the convergence results, we will establish an ergodic theorem for $\{(i_t, Z_t)\}$. We start by listing several properties of the iterates $\{Z_t\}$, which will be either related to or needed in the subsequent analysis.

Throughout the paper, let $\|\cdot\|$ denote the 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 surely.

3.1 Some Properties of Iterates

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

$$L_\ell^t = \frac{q_{i_\ell i_{\ell+1}}}{p_{i_\ell i_{\ell+1}}} \cdot \frac{q_{i_{\ell+1} i_{\ell+2}}}{p_{i_{\ell+1} i_{\ell+2}}} \cdots \frac{q_{i_{t-1} i_t}}{p_{i_{t-1} i_t}}. \tag{17}$$

Define $L_t^t = 1$. We have for $\ell \le \ell' \le t$, $L_\ell^{\ell'} L_{\ell'}^t = L_\ell^t$ and since $Q \prec P$ under Assumption 2.1,

$$E[\, L_\ell^t \mid i_\ell] = 1. \tag{18}$$

Let $\beta = \lambda\alpha$. The iterates Z_t can be expressed as

$$Z_t = \beta \frac{q_{i_{t-1} i_t}}{p_{i_{t-1} i_t}} \cdot Z_{t-1} + \phi(i_t) = \beta L_{t-1}^t \cdot Z_{t-1} + \phi(i_t), \tag{19}$$

and by unfolding the right-hand side,

$$Z_t = \beta^t L_0^t z_0 + \sum_{m=0}^{t-1} \beta^m L_{t-m}^t \phi(i_{t-m}). \tag{20}$$

It is shown in Glynn and Iglehart [GI89, Prop. 5] that L_0^τ can have infinite variance, where τ is the first entrance time of a certain state. It is also known in this setting that the estimator of the total cost up to time τ, $L_0^\tau \sum_{\ell=0}^{\tau-1} g(i_\ell, i_{\ell+1})$, can have infinite variance; this is shown by Randhawa and Juneja [RJ04]. In the infinite-horizon case we consider, using the iterative form (19) of Z_t, one can easily construct examples of Z_t having unbounded second moments, or unbounded νth order moments with $\nu > 1$, as t increases. Furthermore, as we show below (Prop. 3.1), under seemingly fairly common situations, Z_t is almost surely unbounded. Thus even for a finite space MDP, the case $P \ne Q$ sharply contrasts the standard case where $P = Q$ and $\{Z_t\}$ is bounded by definition.

On the other hand, the iterates Z_t exhibit a number of "good" properties indicating that the process $\{Z_t\}$ is well-behaved for all values of λ. The two properties below will be used in the convergence analysis of the present and the next sections, where some additional properties of the process $\{(i_t, Z_t)\}$ will be discussed.

Lemma 3.1.

(i) The Markov chain $\{(i_t, Z_t)\}$ satisfies the drift condition,

$$E\big[V(i_t, Z_t) \mid i_{t-1}, Z_{t-1}\big] \leq \beta V(i_{t-1}, Z_{t-1}) + c$$

for the deterministic constant $c = \max_i \|\phi(i)\|$ and non-negative function $V(i, z) = \|z\|$.

(ii) For each initial condition z_0, $\sup_t E\|Z_t\| \leq \max\{\|z_0\|, c\}/(1-\beta)$.

Proof. The statement in (i) follows from Eqs. (18) and (19). The statement in (ii) is a consequence of (i). Alternatively, it can be derived from the expression of Z_t in Eq. (20): with $\tilde{c} = \max\{\|z_0\|, c\}$,

$$E\|Z_t\| \leq \tilde{c}\, E\Big[\beta^t L_0^t + \sum_{m=0}^{t-1} \beta^m L_{t-m}^t\Big] \leq \tilde{c} \sum_{m=0}^{\infty} \beta^m \leq \tilde{c}/(1-\beta). \qquad \square$$

The function V is a stochastic Lyapunov function for the Markov process $\{(i_t, Z_t)\}$, and has powerful implications on its behavior (see [MT09, Mey89]), beyond the property (ii) above, which will however be sufficient for most of our analysis. The next property will be used to establish, among others, the uniqueness of the invariant probability measure of the process $\{(i_t, Z_t)\}$.

Lemma 3.2. *Let $\{Z_t\}$ and $\{\hat{Z}_t\}$ be defined by Eq. (19) with initial conditions $\bar{z}$ and $\bar{z} + \Delta$, respectively, and for the same sample path of $\{i_t\}$. Then $Z_t - \hat{Z}_t \overset{a.s.}{\to} 0$.*

Proof. From Eq. (19) and equivalently, Eq. (20), we have $Z_t - \hat{Z}_t = \beta^t L_0^t \Delta$, independent of $\bar{z}$ for all t. The sequence of nonnegative scalar random variables $X_t = \beta^t L_0^t, t \geq 0$ satisfies the recursion $X_t = \beta L_{t-1}^t X_{t-1}$ with $X_0 = 1$, and by Eq. (18)

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

where $\mathcal{F}_{t-1}$ is the σ-field generated by i_ℓ, $\ell \leq t-1$. Hence $\{(X_t, \mathcal{F}_t)\}$ is a nonnegative supermartingale with $EX_0 = 1 < \infty$. By a martingale congergence theorem (see e.g., 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 $EX_t = \beta^t \to 0$ as $t \to \infty$, $X = 0$ a.s. Hence $X_t \overset{a.s.}{\to} 0$ and $Z_t - \hat{Z}_t \overset{a.s.}{\to} 0$. $\square$

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], given below, (in which "i.o." stands for "infinitely often," and "a.s." attached to a set-inclusion relation means that the relation holds after excluding a set of probability zero from the sample space).

Lemma 3.3. *Let S be a topological space. For any S-valued process $\{X_t, t \geq 0\}$ and Borel-measurable subsets A, B of S, if for all t,*

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

then

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

We have the following result. Denote by $Z_{t,j}$ and $\phi_j(i_t)$ the jth elements of the vectors Z_t and $\phi(i_t)$, respectively. Consider a cycle of states $\{\bar{\imath}_1, \bar{\imath}_2, \ldots, \bar{\imath}_m, \bar{\imath}_1\} \subset \mathcal{I}$ with the following three properties:

(a) it occurs with positive probability: $p_{\bar{\imath}_1\bar{\imath}_2} p_{\bar{\imath}_2\bar{\imath}_3} \cdots p_{\bar{\imath}_m\bar{\imath}_1} > 0$;

(b) it has an amplifying effect in the sense that $\beta^m \frac{q_{\bar{\imath}_1\bar{\imath}_2}}{p_{\bar{\imath}_1\bar{\imath}_2}} \frac{q_{\bar{\imath}_2\bar{\imath}_3}}{p_{\bar{\imath}_2\bar{\imath}_3}} \cdots \frac{q_{\bar{\imath}_m\bar{\imath}_1}}{p_{\bar{\imath}_m\bar{\imath}_1}} > 1$;

(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:

$$\text{either} \quad \phi_{\bar{j}}(\bar{i}_k) \geq 0, \quad \forall k = 1, \ldots, m, \quad \text{with} \quad \phi_{\bar{j}}(\bar{i}_k) > 0 \text{ for some } k; \tag{21}$$

$$\text{or} \quad \phi_{\bar{j}}(\bar{i}_k) \leq 0, \quad \forall k = 1, \ldots, m, \quad \text{with} \quad \phi_{\bar{j}}(\bar{i}_k) < 0 \text{ for some } k. \tag{22}$$

The next proposition shows that if such a cycle exists, then $\{Z_t\}$ is unbounded with probability 1, in almost all natural problems. The latter qualification relates to a nonrestrictive technical condition in the proposition and will be discussed after the proof. Simple examples with almost surely unbounded $\{Z_t\}$ can be obtained by letting z_0 and $\phi(i), i \in \mathcal{I}$, all be nonnegative and constructing a cycle as above. The phenomenon of unbounded $\{Z_t\}$ can be better understood from the viewpoint of the ergodic behavior of the Markov process $\{(i_t, Z_t)\}$, to be discussed in Section 3.3 (Remark 3.3).

Proposition 3.1. *Suppose the Markov chain* $\{i_t\}$ *is irreducible, there exists a cycle 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. (21) (respectively, Eq. (22)) 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* $\mathbf{P}(\sup_t \|Z_t\| = \infty) = 1$.

Proof. Denote by $\mathcal{C}$ the set of states $\{\bar{i}_1, \bar{i}_2, \ldots, \bar{i}_m\}$ in the cycle. By symmetry, it is sufficient to prove the statement for the case where the cycle satisfies properties (a), (b) and (c) with Eq. (21).

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

$$z_{t+m,\bar{j}} = \beta^m l_0^m \cdot z_{t,\bar{j}} + \epsilon, \tag{23}$$

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. (21), we have $\epsilon > 0$ and $\beta^m l_0^m > 1$. Consider the sequence $\{y_\ell\}$ defined by the recursion

$$y_{\ell+1} = \zeta y_\ell + \epsilon, \quad \ell \geq 0, \qquad \text{where} \quad \zeta = \beta^m l_0^m > 1;$$

y_ℓ corresponds to the value $z_{t+\ell m,\bar{j}}$ if during $[t, t+\ell m]$ the chain $\{i_t\}$ would repeat the cycle ℓ times [cf. Eq. (23)]. Since $\zeta > 1$ and $\epsilon > 0$, simple calculation shows that unless $y_\ell = -\epsilon/(\zeta - 1)$ for all $\ell \geq 0$, $|y_\ell| \to \infty$ as $\ell \to \infty$.

Let $\nu = -\epsilon/(\zeta-1) = -\epsilon/(\beta^m l_0^m - 1)$ be the 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 ℓ be such that $|y_\ell| \geq K_2$ 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 ℓ times. Therefore, applying Lemma 3.3 with $X_t = (i_t, Z_t)$, we have

$$\{i_t = \bar{i}_1, \ Z_{t,\bar{j}} \notin (\nu - \eta, \nu + \eta), \ \|Z_t\| \leq K_1 \text{ i.o.}\} \subset \{\|Z_t\| \geq K_2 \text{ i.o.}\} \ \text{ a.s.} \tag{24}$$

We now prove $\mathbf{P}\big(\sup_t \|Z_t\| < \infty\big) = 0$. Let us assume $\mathbf{P}\big(\sup_t \|Z_t\| < \infty\big) \geq \delta > 0$ to derive a contradiction. Define

$$K_1 = \inf_K \Big\{K \mid P\big(\sup_t \|Z_t\| \leq K\big) \geq \delta/2\Big\}, \qquad \mathcal{E} = \{\sup_t \|Z_t\| \leq K_1\}. \tag{25}$$

Then $K_1 < \infty$ and $\mathbf{P}(\mathcal{E}) \geq \delta/2$. 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 of the proposition, $\mathbf{P}\big(i_t = \bar{i}_1, \ Z_{t,\bar{j}} \notin (\nu - \eta, \nu + \eta) \text{ i.o.}\big) = 1$, and by the definition of $\mathcal{E}$, this implies

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

It then follows from Eq. (24) that for any $K_2 > K_1$,

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

Since $\mathbf{P}(\mathcal{E}) \geq \delta/2$, this contradicts the definition of $\mathcal{E}$ in Eq. (25). Therefore $\mathbf{P}\big(\sup_t \|Z_t\| < \infty\big) = 0$. This completes the proof. □

We remark that the extra technical condition $\mathbf{P}(i_t = \bar{i}_1, Z_{t,\bar{j}} \notin \mathcal{O}(\nu)\ i.o.) = 1$ in Prop. 3.1 is not restrictive. The opposite case – that on a set with non-negligible probability, $Z_{t,\bar{j}}$ eventually always lies arbitrarily close to ν whenever $i_t = \bar{i}_1$ – seems unlikely to occur except in highly contrived examples. Thus the proposition shows that in the case of a general value of λ, we cannot claim directly the boundedness of $\{G_t\}$, which is often the first step in convergence proofs, by assuming the boundedness of $\{Z_t\}$ unrealistically.

On the other hand, although the unboundedness of Z_t may sound disquieting, it is $\gamma_t Z_t \overset{a.s.}{\to} 0$ 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 Eq. (11) and the fact that $\lim_{t\to\infty} \gamma_t = 0$.) 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. For practical implementation, if $\|Z_t\|$ becomes intolerably large, we can equivalently iterate $\gamma_t Z_t$ via

$$\gamma_t Z_t = \beta L_{t-1}^t \cdot \tfrac{\gamma_t}{\gamma_{t-1}} \cdot (\gamma_{t-1} Z_{t-1}) + \gamma_t \phi(i_t),$$

instead of iterating Z_t directly. Similarly, we can also choose scalars $a_t, t \geq 1$, dynamically to keep $a_t Z_t$ in a desirable range, iterate $a_t Z_t$ instead of Z_t, and use $\frac{\gamma_t}{a_t}(a_t Z_t)$ in the update of G_t.

Remark 3.1. It can also be shown, using essentially a zero-one law for tail events of Markov chains (see [Bre92, Theorem 7.43]), that under Assumptions 2.1 and 2.2, for each initial condition (z_0, G_0),

$$\mathbf{P}\big(\sup_t \|Z_t\| < \infty\big) = 1 \text{ or } 0, \qquad \mathbf{P}\big(\lim_{t\to\infty} \gamma_t Z_t = 0\big) = 1 \text{ or } 0.$$

See [Yu10, Prop. 3.1] for details. □

3.2 Convergence in Mean

We show now that G_t converges in mean to G^*. This implies that G_t converges in probability to G^*, and hence that the LSTD(λ) solution r_t converges in probability to the solution r^* of Eq. (6) when the latter exists and is unique. We state the result in a slightly more general context involving a Lipschitz continuous function $h(z, i, j)$ in place of $z\psi(i, j)'$, to prepare also for the subsequent almost sure convergence analysis in Sections 3.3 and 4.1.

Theorem 3.1. *Let $h(z, i, j)$ be a vector-valued function on $\Re^d \times \mathcal{I}^2$ which is Lipschitz continuous in z with Lipschitz constant M_h, i.e.,*

$$\|h(z, i, j) - h(\hat{z}, i, j)\| \leq M_h \|z - \hat{z}\|, \qquad \forall\, z, \hat{z} \in \Re^d,\ i, j \in \mathcal{I}.$$

Let

$$G_t^h = (1 - \gamma_t) G_{t-1}^h + \gamma_t h(Z_t, i_t, i_{t+1}).$$

Then under Assumptions 2.1 and 2.2, there exists a constant $G^{h,*}$ such that for each initial condition (z_0, G_0),

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

Proof. For notational simplicity, we suppress the superscript h in the proof. First, we introduce another process $(\widetilde{Z}_{t,T}, \widetilde{G}_{t,T})$ on the same probability space, and apply an 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).

For a positive integer T, define $\widetilde{Z}_{t,T} = Z_t$ for $t \leq T$ and $\widetilde{G}_{0,T} = G_0$, and define

$$\widetilde{Z}_{t,T} = \phi(i_t) + \beta L_{t-1}^t \phi(i_{t-1}) + \cdots + \beta^T L_{t-T}^t \phi(i_{t-T}), \quad t > T; \tag{26}$$

$$\widetilde{G}_{t,T} = (1-\gamma_t)\widetilde{G}_{t-1,T} + \gamma_t h\big(\widetilde{Z}_{t,T}, i_t, i_{t+1}\big), \quad t \geq 1. \tag{27}$$

Then for $t \leq T$, $\widetilde{G}_{t,T} = G_t$ because $\widetilde{Z}_{t,T}$ and Z_t coincide. By construction $\{\widetilde{Z}_{t,T}\}$ and $\{\widetilde{G}_{t,T}\}$ are bounded. This is because $\max_i \|\phi(i)\|$ and $L_\ell^{\ell+\tau}, 0 \leq \tau \leq T, \ell \geq 0$, can be bounded by some deterministic constant, so $\sup_t \|\widetilde{Z}_{t,T}\| \leq c_T$ for some deterministic constant c_T depending on T. Consequently, by the Lipschitz property of h and the assumption $\gamma_t \in (0,1]$ (Assumption 2.2), $\{h\big(\widetilde{Z}_{t,T}, i_t, i_{t+1}\big)\}$ and $\{\widetilde{G}_{t,T}\}$ are also bounded.

The sequence $\{\widetilde{G}_{t,T}\}$ converges almost surely to a constant G_T^* independent of the initial condition. This is because for $t > T$, $h(\widetilde{Z}_{t,T}, i_t, i_{t+1})$ can be viewed as a function of the $T+2$ consecutive states $X_t = (i_{t-T}, i_{t-T+1}, \ldots, i_{t+1})$, while under Assumption 2.1, $\{X_t\}$ is a finite space Markov chain with a single recurrent class. Thus, an application of the result in stochastic approximation theory given in Borkar [Bor08, Chap. 6, Theorem 7 and Cor. 8] shows that under the stepsize condition in Assumption 2.2, with E_0 denoting expectation under the stationary distribution of the Markov chain $\{i_t\}$,

$$\widetilde{G}_{t,T} \overset{a.s.}{\to} G_T^*, \qquad \text{where } \; G_T^* = E_0\big[h(\widetilde{Z}_{k,T}, i_k, i_{k+1})\big], \quad \forall k > T. \tag{28}$$

Clearly, G_T^* does not depend on (z_0, G_0). Since $\sup_t \|\widetilde{G}_{t,T}\| \leq c_T$ for some deterministic constant c_T, we also have by the Lebesgue bounded convergence theorem

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

The sequence $\{G_T^*, T \geq 1\}$ converges to some constant G^*. To see this, consider any $T_1 < T_2$. Using the definition of $\widetilde{Z}_{t,T}$ and arguing similar to the proof for Lemma 3.1(ii), we have

$$E_0\|\widetilde{Z}_{k,T_1} - \widetilde{Z}_{k,T_2}\| \leq c\beta^{T_1}, \qquad \forall k > T_2,$$

where $c = \max_i \|\phi(i)\|/(1-\beta)$. Therefore, using the definition of G_T^* in Eq. (28) and the Lipschitz property of h, we have for any $k > T_2$,

$$\begin{aligned}\|G_{T_1}^* - G_{T_2}^*\| &= \big\|E_0\big[h(\widetilde{Z}_{k,T_1}, i_k, i_{k+1}) - h(\widetilde{Z}_{k,T_2}, i_k, i_{k+1})\big]\big\| \\ &\leq M_h\, E_0\big\|\widetilde{Z}_{k,T_1} - \widetilde{Z}_{k,T_2}\big\| \leq cM_h\beta^{T_1}.\end{aligned}$$

This shows that $\{G_T^*\}$ is a Cauchy sequence and therefore converges to some constant G^*.

We now show $\lim_{t\to\infty} E\,\|G_t - G^*\| = 0$. Since for each T,

$$\limsup_{t\to\infty} E\,\|G_t - G^*\| \leq \limsup_{t\to\infty} E\big\|G_t - \widetilde{G}_{t,T}\big\| + \lim_{t\to\infty} E\big\|\widetilde{G}_{t,T} - G_T^*\big\| + \big\|G^* - G_T^*\big\|, \tag{30}$$

and by the preceding proof, $\lim_{t\to\infty} E\big\|\widetilde{G}_{t,T} - G_T^*\big\| = 0$ and $\lim_{T\to\infty}\big\|G^* - G_T^*\big\| = 0$, it suffices to show $\lim_{T\to\infty}\limsup_{t\to\infty} E\big\|G_t - \widetilde{G}_{t,T}\big\| = 0$. Using the definition of $\widetilde{Z}_{t,T}$ and arguing similar to the proof of Lemma 3.1(ii), we have

$$\|Z_t - \widetilde{Z}_{t,T}\| = 0, \quad t \leq T; \qquad E\|Z_t - \widetilde{Z}_{t,T}\| \leq c\beta^T, \quad t \geq T+1, \tag{31}$$

where $c = \max\{\|z_0\|, \max_i \|\phi(i)\|\}/(1-\beta)$. By the definition of G_t and $\widetilde{G}_{t,T}$,

$$G_t - \widetilde{G}_{t,T} = (1-\gamma_t)\big(G_{t-1} - \widetilde{G}_{t-1,T}\big) + \gamma_t\big(h(Z_t, i_t, i_{t+1}) - h(\widetilde{Z}_{t,T}, i_t, i_{t+1})\big).$$

Therefore, using the triangle inequality, the Lipschitz property of h and Eq. (31), we have

$$\begin{aligned} E\|G_t - \widetilde{G}_{t,T}\| &\leq (1-\gamma_t)E\|G_{t-1} - \widetilde{G}_{t-1,T}\| + \gamma_t E\|h(Z_t, i_t, i_{t+1}) - h(\widetilde{Z}_{t,T}, i_t, i_{t+1})\| \\ &\leq (1-\gamma_t)E\|G_{t-1} - \widetilde{G}_{t-1,T}\| + \gamma_t M_h E\|Z_t - \widetilde{Z}_{t,T}\| \\ &\leq (1-\gamma_t)E\|G_{t-1} - \widetilde{G}_{t-1,T}\| + \gamma_t c M_h \beta^T, \end{aligned}$$

which implies under the stepsize condition in Assumption 2.2,

$$\lim_{T\to\infty} \limsup_{t\to\infty} E\|G_t - \widetilde{G}_{t,T}\| \leq \lim_{T\to\infty} cM_h\beta^T = 0.$$

This completes the proof. □

For the case $h(z,i,j) = z\psi(i,j)'$, $G_T^{h,*}$ given in Eq. (28) has an explicit expression:

$$G_T^{h,*} = \Phi'\Xi\Big(\sum_{m=0}^{T} \beta^m Q^m\Big)\Psi,$$

from which it can be seen that the limit $G^{h,*}$ of $\{G_T^{h,*}\}$ is G^* given by Eq. (13).

3.3 Almost Sure Convergence

To study 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 $S = \mathcal{I} \times \Re^d$ with product topology (discrete topology on $\mathcal{I}$ and usual topology on $\Re^d$). We view S also as a metric space (with the usual metric consistent with the topology). We will establish an ergodic theorem for $\{(i_t, Z_t)\}$ (Theorem 3.2) and the almost sure convergence of $\{G_t\}$ when the stepsize is $\gamma_t = 1/(t+1)$ (Theorem 3.3). The latter will imply that the sequence $\{\Phi r_t\}$ computed by the off-policy LSTD(λ) algorithm with the same stepsizes converges almost surely to the solution Φr^* of the projected Bellman equation (2) when the latter exists and is unique.

First, we specify some notation and definitions for topological space Markov chains in general. Let P_S denote the transition probability kernel of a Markov chain $\{X_t\}$ on the state space S, i.e.,

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

where $P_S(x,\cdot)$ is the conditional probability of X_1 given $X_0 = x$, and $\mathcal{B}(S)$ denotes the Borel σ-field on S. The k-step transition probability kernel is denoted by P_S^k. As an operator, P_S^k maps any bounded Borel-measurable function $f : S \to \Re$ to another such function $P_S^k f$, given by

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

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

Let $\mathcal{C}_b(S)$ denote the set of bounded continuous functions on S. A Markov chain on S is a *weak Feller* chain (or simply, a Feller chain) if for all $f \in \mathcal{C}_b(S)$, $P_S f \in \mathcal{C}_b(S)$ [MT09, Prop. 6.1.1(i)]. A Markov chain $\{X_t\}$ on S is said to be bounded in probability, if for each initial state x and each $\epsilon > 0$, there exists a compact subset $C \subset S$ such that $\liminf_{t\to\infty} \mathbf{P}_x(X_t \in C) \geq 1 - \epsilon$.

We now relate $\{(i_t, Z_t)\}$ to a Feller chain with desirable properties.[2]

Lemma 3.4. *The Markov chain $\{(i_t, Z_t)\}$ is weak Feller and bounded in probability, therefore has at least one invariant probability measure.*

[2] A Feller chain is not necessarily ψ-irreducible (for the latter notion, see [MT09]). A simple counterexample in our case is given by setting $\phi(i) = 0$ for all i.

Proof. Since $Z_1 = \beta \frac{q_{i_0 i_1}}{p_{i_0 i_1}} \cdot z_0 + \phi(i_1)$, Z_1 is a function of (z_0, i_0, i_1); denote this function by $Z_1(z_0, i_0, i_1)$. It is continuous in z_0 for given (i_0, i_1). Since the space $\mathcal{I}$ is discrete, for any $f \in \mathcal{C}_b(S)$, $f(i, z)$ is bounded and continuous in z for each i. It can be seen that

$$(P_S f)(i, z) = E[f(i_1, Z_1) \mid i_0 = i, Z_0 = z] = \sum_{j \in \mathcal{I}} p_{ij} f\big(j, Z_1(z, i, j)\big)$$

is also bounded and continuous in z for each i, so $P_S f \in \mathcal{C}_b(S)$ and the chain $\{(i_t, Z_t)\}$ is weak Feller. Lemma 3.1 together with Markov's inequality implies that for each initial condition $x = (\bar{i}, \bar{z})$ and some constant c_x, $\mathbf{P}_x(\|Z_t\| \leq K) \geq 1 - c_x/K$ for all $t \geq 0$. Since $\mathcal{I}$ is compact, this shows that the chain $\{(i_t, Z_t)\}$ is bounded in probability. By [MT09, Prop. 12.1.3], a weak Feller chain that is bounded in probability has at least one invariant probability measure. □

We now show that the invariant probability measure of $\{(i_t, Z_t)\}$ is unique and the chain is ergodic. Recall that the occupation probability measures $\mu_t, t \geq 1$ of a Markov chain $\{X_t\}$ on S are defined by

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

where $\mathbf{1}_A$ denotes the indicator function for a Borel-measurable set $A \subset S$. For an initial condition $x \in S$, we use $\{\mu_{x,t}\}$ to denote the occupation measure sequence, and we note that for any Borel-measurable function f on S, the expression $\frac{1}{t} \sum_{k=1}^{t} f(X_k)$ is equivalent to $\int f(y) \mu_{x,t}(dy)$, or $\int f \, d\mu_{x,t}$.

Theorem 3.2. *Under Assumption 2.1, the Markov chain $\{(i_t, Z_t)\}$ has a unique invariant probability measure π, and for each initial condition $x = (i, z)$, almost surely, the sequence of occupation measures $\{\mu_{x,t}\}$ converges weakly to π.*

Proof. Since $\{(i_t, Z_t)\}$ has an invariant probability measure π, it follows by a strong law of large numbers for stationary Markov chains (see e.g., discussion preceding [Mey89, Prop. 4.1]) that for each $x = (\bar{i}, \bar{z})$ from a set $F \subset S$ with full π-measure, almost surely $\{\mu_{x,t}\}$ converges weakly to some probability measure π_x on S that is a function of x. (Since $\{(i_t, Z_t)\}$ is weak Feller, these π_x must also be invariant probability measures [Mey89, Prop. 4.1]; but this fact will not be used in our proof.)

We show first that corresponding to $x = (\bar{i}, \bar{z}) \in F$, for each $\hat{x} = (\bar{i}, z)$, almost surely $\{\mu_{\hat{x},t}\}$ converges weakly to π_x, so in particular, π_x does not depend on $\bar{z}$. To this end, consider the processes $\{Z_t\}$ and $\{\hat{Z}_t\}$ initialized with x and $\hat{x}$, respectively, and for the same sample path of $\{i_t\}$. By Lemma 3.2, $Z_t - \hat{Z}_t \overset{a.s.}{\to} 0$. Therefore, almost surely, for all bounded and uniformly continuous functions f on S, $\lim_{t \to \infty} \big(f(i_t, Z_t) - f(i_t, \hat{Z}_t)\big) = 0$, and consequently,

$$\lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \big(f(i_t, Z_t) - f(i_t, \hat{Z}_t)\big) = 0.$$

Since almost surely $\mu_{x,t} \to \pi_x$ weakly, we have almost surely, $\lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} f(i_t, Z_t) = \int f \, d\pi_x$ for all the above f. It then follows that almost surely,

$$\lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} f(i_t, \hat{Z}_t) = \int f \, d\pi_x$$

for all bounded and uniformly continuous functions f, and hence, by [MT09, Prop. D.5.1], almost surely $\mu_{\hat{x},t} \to \pi_x$ weakly.

We now show that π_x is the same for all $x \in F$. Suppose this is not true: there exist states $x = (\bar{i}, \bar{z}), \hat{x} = (\hat{i}, \hat{z}) \in F$ with $\pi_x \neq \pi_{\hat{x}}$. Then, since S is a metric space, by [Dud03, Prop. 11.3.2] there exists a bounded Lipschitz function h on S such that

$$\int h \, d\pi_x \neq \int h \, d\pi_{\hat{x}}.$$

For any z, by the weak convergence $\mu_{(\bar{i},z),t} \to \pi_x$ and $\mu_{(\hat{i},z),t} \to \pi_{\hat{x}}$ just proved, we have

$$\lim_{t\to\infty} \int h \, d\mu_{(\bar{i},z),t} = \int h \, d\pi_x, \quad \mathbf{P}_{(\bar{i},z)}\text{-}a.s.; \qquad \lim_{t\to\infty} \int h \, d\mu_{(\hat{i},z),t} = \int h \, d\pi_{\hat{x}}, \quad \mathbf{P}_{(\hat{i},z)}\text{-}a.s.$$

Therefore, with the initial distribution being $\tilde{\mu} = \frac{1}{2}\delta_{(\bar{i},z)} + \frac{1}{2}\delta_{(\hat{i},z)}$, where δ_x denotes the Dirac probability measure, $\{\int h \, d\mu_t\}$ converges $\mathbf{P}_{\tilde{\mu}}$-almost surely to a non-degenerate random variable. On the other hand, since h is Lipschitz, applying Theorem 3.1 with $\gamma_t = 1/(t+1)$ and $G_0 = 0$, we have that under $\mathbf{P}_{\tilde{\mu}}$, $\{\int h \, d\mu_t\}$ converges in mean to a constant and therefore has a subsequence converge almost surely to the same constant, a contradiction. Thus π_x must be the same for all $x \in F$; denote this probability measure by $\tilde{\pi}$.

We now show $\pi = \tilde{\pi}$. Consider any bounded and continuous function f on S. By the strong law of large numbers for stationary processes (see e.g., [Doo53, Chap. X, Theorem 2.1]),

$$E_\pi \Big[\lim_{t\to\infty} \int f \, d\mu_{X_0,t} \Big] = E_\pi \big[f(X_0) \big],$$

while by the preceding proof we have for each $x \in F$, a set with $\pi(F) = 1$, $\lim_{t\to\infty} \int f \, d\mu_{x,t} = \int f \, d\tilde{\pi}$, $\mathbf{P}_x$-$a.s.$ Therefore

$$\int f \, d\tilde{\pi} = E_\pi \Big[\lim_{t\to\infty} \int f \, d\mu_{X_0,t} \Big] = E_\pi \big[f(X_0) \big] = \int f \, d\pi.$$

This shows $\pi = \tilde{\pi}$.

Finally, suppose there exists another invariant probability measure $\tilde{\pi}$. Then, the preceding conclusions apply also to $\tilde{\pi}$ and some set $\tilde{F} \subset S$ with $\tilde{\pi}(\tilde{F}) = 1$. On the other hand, clearly the marginals of π and $\tilde{\pi}$ on $\mathcal{I}$ must coincide with the unique invariant probability of the irreducible chain $\{i_t\}$, so using the fact $\pi(F) = \tilde{\pi}(\tilde{F}) = 1$, we have that for any state $\bar{i}$, there exist $\bar{z}, \tilde{z}$ such that $(\bar{i}, \bar{z}) \in F$ and $(\bar{i}, \tilde{z}) \in \tilde{F}$. Then, by the preceding proof, with initial condition $x = (\bar{i}, z)$ for any z, almost surely, $\mu_{x,t} \to \tilde{\pi}$ and $\mu_{x,t} \to \pi$ weakly. Hence $\pi = \tilde{\pi}$ and the chain has a unique invariant probability measure. □

Remark 3.2. In the above proof, we used the conclusion of Theorem 3.1 to show that π_x is the same for all $x \in F$. We may avoid this reliance by using alternative arguments at this step for the finite space MDP case, but the above proof applies readily also to compact space MDP models that we will consider later. Another entirely different proof based on the theory of e-chains [MT09] can be found in [Yu10]; however, it is much longer than the one given here. □

Remark 3.3. The ergodicity of the chain $\{(i_t, Z_t)\}$ shown by the preceding theorem gives a clear explanation to the unboundedness of $\{Z_t\}$ that we observed in Section 3.1, Prop. 3.1: If the total mass of π does not concentrate on a bounded set of S, then because the sequence of occupation measures converges weakly to π almost surely, $\{Z_t\}$ must be unbounded with probability 1. □

Remark 3.4. The preceding theorem also implies that we can obtain a good approximation of $G^{h,*}$ by using modified bounded iterates, such as $\widehat{G}^h_t = (1-\gamma_t)\widehat{G}^h_{t-1} + \gamma_t \hat{h}(Z_t, i_t, i_{t+1})$, where $\gamma_t = 1/(t+1)$ and $\hat{h}(Z_t, i_t, i_{t+1})$ is $h(Z_t, i_t, i_{t+1})$ truncated component-wise to be within $[-K, K]$ for some sufficiently large K. □

Let E_π denote expectation with respect to $\mathbf{P}_\pi$. To establish the almost sure convergence of $\{G_t\}$, we need to show first that $E_\pi\big[\|Z_0\psi(i_0,i_1)'\|\big] < \infty$. Here we prove it using the following two facts. First, Theorem 3.2 implies

$$\frac{1}{T}\sum_{t=1}^{T} P_S^t(x,\cdot) \overset{\text{weakly}}{\longrightarrow} \pi, \qquad \forall x \in S. \tag{32}$$

Second, by Lemma 3.1, for some constant c depending on the initial condition x,

$$E_x[\|Z_t\|] \leq c, \quad \forall t \geq 0. \tag{33}$$

As in the preceding subsection, we state the result in slightly more general terms for all functions Lipschitz continuous in z, which will be useful later in analyzing the convergence of other TD(λ) algorithms.

Proposition 3.2. *Under Assumption 2.1, for any (vector-valued) function $h(z,i,j)$ on $\Re^d \times \mathcal{I}^2$ that is Lipschitz continuous in z, $E_\pi\big[\|h(Z_0,i_0,i_1)\|\big] < \infty$.*

Proof. By the Lipschitz property of h, $\|h(Z_0,i_0,i_1)\| \leq M_h\|Z_0\| + \|h(0,i_0,i_1)\|$ for some constant M_h, therefore, to prove the result, it is sufficient to show $E_\pi[\|Z_0\|] < \infty$. To this end, consider a sequence of scalars $a_k, k \geq 0$ with

$$a_0 = 0, \quad a_1 \in (0,1], \quad a_{k+1} = a_k + 1, \quad k \geq 1. \tag{34}$$

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

$$O_k = \{z \mid a_k < \|z\| < a_{k+1}\}. \tag{35}$$

It is then sufficient to show that for any such $\{a_k\}$, $\sum_{k=0}^{\infty} a_{k+1}\cdot\pi\big(\mathcal{I}\times O_k\big) < \infty$.[3]

Fix any initial condition x. Using Eq. (33), we have for all integers $K \geq 0, t \geq 0$,

$$\sum_{k=0}^{K} a_{k+1}\cdot\mathbf{P}_x(Z_t\in O_k) \leq 1 + \sum_{k=0}^{K} a_k\cdot\mathbf{P}_x(Z_t\in O_k) \leq 1 + E_x\big[\|Z_t\|\big] \leq c+1.$$

Therefore for all $K \geq 0, T \geq 0$,

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{k=0}^{K} a_{k+1}\cdot\mathbf{P}_x(Z_t\in O_k) = \sum_{k=0}^{K} a_{k+1}\cdot\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{P}_x(Z_t\in O_k)\right) \leq c+1. \tag{36}$$

Since by construction O_k and $\mathcal{I}\times O_k$ are open sets on $\Re^d$ and S, respectively, by Eq. (32) and [MT09, Theorem D.5.4] we have for all k,

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

[3] This is because we can choose two sequences $\{a_k^1\}$, $\{a_k^2\}$ as in (34) with $a_1^1 = 1, a_1^2 = 1/2$, for instance, such that the corresponding open sets $O_k^1, O_k^2, k \geq 0$ given by (35) together cover the space of z except for the origin. Then

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

so we can bound $E_\pi[\|Z_0\|]$ by

$$E_\pi[\|Z_0\|] \leq E_\pi\Big[\sum_{k=0}^{\infty}\big(a_{k+1}^1\cdot\mathbf{1}_{O_k^1}(Z_0) + a_{k+1}^2\cdot\mathbf{1}_{O_k^2}(Z_0)\big)\Big] = \sum_{k=0}^{\infty} a_{k+1}^1\cdot\pi(\mathcal{I}\times O_k^1) + \sum_{k=0}^{\infty} a_{k+1}^2\cdot\pi(\mathcal{I}\times O_k^2).$$

Combining this with Eq. (36), we have for all $K \geq 0$,

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

and therefore $\sum_{k=0}^{\infty} a_{k+1} \cdot \pi\big(\mathcal{I} \times O_k\big) \leq c+1$. This completes the proof. □

Theorem 3.3. *Assume the conditions and notation of Theorem 3.1 and let* $\gamma_t = 1/(t+1)$. *Then, for each initial condition* (z_0, G_0^h), $G_t^h \overset{a.s.}{\to} G^{h,*}$, *where* $G^{h,*} = E_\pi\big[h(Z_0, i_0, i_1)\big]$ *is the constant in Theorem 3.1.*

Proof. For each initial (z_0, G_0^h), by Theorem 3.1, G_t^h converges in mean to $G^{h,*}$, a constant independent of the initial condition. This further implies the convergence of a subsequence $G_{t_k}^h \overset{a.s.}{\to} G^{h,*}$, so in order to show $G_t^h \overset{a.s.}{\to} G^{h,*}$, it is sufficient to show G_t^h converges alsmost surely. For simplicity, in the rest of the proof we suppress the superscript h. With $\gamma_t = 1/(1+t)$,

$$G_t = \frac{1}{t+1} \Big(\sum_{k=1}^{t} h(Z_k, i_k, i_{k+1}) + G_0 \Big);$$

it is clear that on a sample path, the convergence of $\{G_t\}$ is equivalent to that of the sequence $\{\frac{1}{t}\sum_{k=1}^{t} h(Z_k, i_k, i_{k+1})\}$.

By Prop. 3.2, $E_\pi\|h(Z_0, i_0, i_1)\| < \infty$. Therefore, applying the strong law of large numbers (see [Doo53, Theorem 2.1] or [MT09, Theorem 17.1.2]) to the stationary Markov process $\{(i_t, Z_t, i_{t+1})\}$ under $\mathbf{P}_\pi$, we have $\frac{1}{t}\sum_{k=1}^{t} h(Z_k, i_k, i_{k+1})$ converges $\mathbf{P}_x$-almost surely for each initial $x = (\bar{i}, \bar{z})$ from a set $F \subset S$ with $\pi(F) = 1$. So G_t converges almost surely for each $x \in F$.

For any initial condition $\hat{x} = (\bar{i}, \hat{z}) \notin F$, let $\bar{x} = (\bar{i}, \bar{z}) \in F$ for some $\bar{z} \in \Re^d$. (Such $\bar{x}$ exists because the irreducibility of $\{i_t\}$ and $\pi(F) = 1$ imply $\pi(\{\bar{i}\} \times \Re^d) > 0$.) 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, with $\hat{G}_0 = G_0$, and for the same path of $\{i_t\}$. By the Lipschitz property of h,

$$\|\hat{G}_t - G_t\| = \Big\| \frac{1}{t+1} \sum_{k=1}^{t} \big(h(\hat{Z}_k, i_k, i_{k+1}) - h(Z_k, i_k, i_{k+1}) \big) \Big\| \leq \frac{M_h}{t+1} \sum_{k=1}^{t} \|\hat{Z}_k - Z_k\|.$$

Since $\hat{Z}_t - Z_t \overset{a.s.}{\to} 0$ by Lemma 3.2, we have $\hat{G}_t - G_t \overset{a.s.}{\to} 0$; since G_t converges almost surely, so is $\hat{G}_t$. Thus $\{G_t\}$ converges $\mathbf{P}_x$-almost surely for each initial condition $x = (\bar{i}, \bar{z})$ and G_0, implying $G_t \overset{a.s.}{\to} G^*$ for each initial condition (z_0, G_0).

Finally, we prove the expression for G^*. By the law of large numbers for stationary processes (see [Doo53, Theorem 2.1] or [MT09, Theorem 17.1.2]), we have $E_\pi[\lim_{t\to\infty} G_t] = E_\pi\big[h(Z_0, i_0, i_1)\big]$. Therefore, $G^* = E_\pi[G^*] = E_\pi\big[h(Z_0, i_0, i_1)\big]$. □

Remark 3.5. The conclusion of the above theorem also implies the convergence $G_t^h \overset{a.s.}{\to} G^{h,*}$ for a stepsize γ_t that is of order $O(1/t)$ and satisfies $\frac{\gamma_t - \gamma_{t+1}}{\gamma_t} = O(1/t)$, (such as $\gamma_t = \frac{c_1}{c_2 + t}$ for some constants c_1, c_2). This can be shown using Theorems 3.1 and 3.3 together with stochastic approximation theory [KY03, Chap. 6, Theorem 1.2 and Example 1 of Sec. 6.2]. As yet we do not have a full answer to the question of whether $G_t^h \overset{a.s.}{\to} G^{h,*}$ for a stepsize sequence that decreases at a rate slower than $1/t$. This question is closely connected to the rate of convergence of $\frac{1}{t}\sum_{k=1}^{t} h(Z_k, i_k, i_{k+1})$ to $G^{h,*}$. In particular, suppose it holds that $\frac{1}{t^{\bar{\nu}}} \sum_{k=1}^{t} \big(h(Z_k, i_k, i_{k+1}) - G^{h,*} \big) \overset{a.s.}{\to} 0$ for some $\bar{\nu} \in (0.5, 1]$, then using stochastic approximation theory [KY03, Chap. 6], we can show that $G_t^h \overset{a.s.}{\to} G^{h,*}$ for the stepsizes $\gamma_t = (t+1)^{-\nu}$, $\nu \in [\bar{\nu}, 1]$. We also note that for a general stepsize sequence satisfying Assumption 2.2, it can be shown that $\{G_t\}$ converges with probability zero or one [Yu10, Prop. 3.1] (cf. Remark 3.1). □

4 Applications and Extensions

In this section we apply the results of Section 3 to analyze the convergence of an off-policy TD(λ) algorithm, and we also extend the convergence analysis of the off-policy LSTD(λ) algorithm for finite space MDP to MDP with compact action and state spaces.

4.1 Convergence of an Off-Policy TD(λ) Algorithm

We consider an off-policy TD(λ) algorithm which aims to solve the projected Bellman equation (6) with stochastic approximation type iterations. It has the same form as the standard, on-policy TD(λ) algorithm, and it is given by

$$r_t = r_{t-1} + \gamma_t Z_t d_t,$$

where Z_t is as in Eq. (19), and d_t is the so-called temporal difference term given by

$$d_t = L_t^{t+1} g(i_t, i_{t+1}) + \alpha L_t^{t+1} \phi(i_{t+1})' r_{t-1} - \phi(i_t)' r_{t-1}.$$

This algorithm is proposed in [BY09, Sec. 5.3] in the context of approximate solutions of linear equations with TD methods. It bears similarity to the off-policy TD(λ) [PSD01], but differs from the latter in a considerable way. (In particular, it differs from the latter in the definitions of Z_t and the projected Bellman equation, as well as in using an infinitely long trajectory of observations instead of a fixed-length trajectory to update Z_t's.) Convergence of the algorithm has not been fully analyzed. We now apply the results of Section 3.3 and the o.d.e.-based stochastic approximation theory [KY03, Chap. 6] to analyze a constrained version of the algorithm.

Introducing the function

$$h(z, i, j; r) = z\, \psi_1(i,j)' r + \psi_2(i,j) \tag{37}$$

with $\psi_1(i,j) = \alpha \frac{q_{ij}}{p_{ij}} \phi(j) - \phi(i)$ and $\psi_2(i,j) = \frac{q_{ij}}{p_{ij}} g(i,j)$, we may write the off-policy TD(λ) algorithm equivalently as

$$r_t = r_{t-1} + \gamma_t h(Z_t, i_t, i_{t+1}; r_{t-1}).$$

To avoid the technical difficulty regarding the boundedness of $\{r_t\}$ in the above unconstrained algorithm, we consider its constrained version

$$r_t = \widehat{\Pi}_H \big[r_{t-1} + \gamma_t h(Z_t, i_t, i_{t+1}; r_{t-1}) \big], \tag{38}$$

where $\widehat{\Pi}_H$ is the projection onto some compact convex set $H \subset \Re^d$.

We apply [KY03, Theorem 6.1.1] to analyze the convergence of this algorithm. Since [KY03] is a standard reference on stochastic approximation, we do not repeat here the theorem and its long list of conditions, nor do we verify the conditions one by one for the TD(λ) algorithm, as some of them obviously hold. We will point out only the key arguments in the analysis.

The "mean" function involved in the mean o.d.e. is the continuous function $\bar{h}(r)$ given by

$$\bar{h}(r) = \bar{C} r + \bar{b}, \qquad r \in \Re^d,$$

with $\bar{C}, \bar{b}$ defined as in Eq. (7). For any fixed r, by Theorem 3.3, for each initial z_0,

$$\frac{1}{t} \sum_{k=1}^{t} h(Z_k, i_k, i_{k+1}; r) \overset{a.s.}{\to} \bar{h}(r). \tag{39}$$

We can bound the function $h(z, i, j; r)$ by

$$\|h(z, i, j; r)\| \leq (\|r\| + 1)\rho_1(z, i, j), \qquad \text{where} \;\; \rho_1(z, i, j) = d\,\big\|z\,\psi_1(i, j)'\big\| + \big\|\psi_2(i, j)\big\|,$$

and bound the change in $h(z, i, j; r)$ in terms of the change in r by

$$\|h(z, i, j; \bar{r}) - h(z, i, j; \hat{r})\| \leq \|\bar{r} - \hat{r}\|\rho_2(z, i, j), \qquad \text{where} \;\; \rho_2(z, i, j) = d\,\big\|z\,\psi_1(i, j)'\big\|.$$

The functions ρ_1 and ρ_2 are Lipschitz continuous in z, so by Theorem 3.3, for each initial z_0,

$$\frac{1}{t}\sum_{k=1}^{t} \rho_j(Z_k, i_k, i_{k+1}) \overset{a.s.}{\to} E_\pi\big[\rho_j(Z_0, i_0, i_1)\big], \qquad j = 1, 2. \tag{40}$$

From Eqs. (39) and (40) it follows that when $\gamma_t = O(1/t)$ with $\frac{\gamma_t - \gamma_{t+1}}{\gamma_t} = O(1/t)$, the asymptotic rate of change condition (the Kushner-Clark condition), which is the main condition in [KY03, Theorem 6.1.1], is satisfied by the various terms as required in the theorem (see [KY03, Example 6.1, p. 171]).

For the constrained algorithm (38), another condition in [KY03, Theorem 6.1.1] is

$$\sup_t E\|h(Z_t, i_t, i_{t+1}; r_{t-1})\| < \infty.$$

It is satisfied because with $\{r_t\}$ confined in the compact set H, $E\|h(Z_t, i_t, i_{t+1}; r_{t-1})\| \leq c_1 E\|Z_t\| + c_2$ for some constants c_1, c_2, while by Lemma 3.1 $\sup_t E\|Z_t\| \leq c$ for some constant c depending on the initial z_0. Hence, applying [KY03, Theorem 6.1.1], we have the convergence of the constrained off-policy TD(λ) algorithm.

Proposition 4.1. *Let the stepsize γ_t satisfy $\gamma_t = O(1/t)$ and $\frac{\gamma_t - \gamma_{t+1}}{\gamma_t} = O(1/t)$. Then $\{r_t\}$ given by Eq. (38) converges almost surely to some limit set of the o.d.e.:*

$$\dot{r} = \bar{h}(r) + z \qquad \textit{for some } z \in -N_H(r),$$

where $N_H(r)$ is the normal cone of H at the point $r \in H$, and z is the boundary-reflecting term to keep the o.d.e. solution in H.

As shown in [BY09, Props. 3 and 5], when λ is sufficiently close to 1, the mapping $\Pi T^{(\lambda)}$ becomes a contraction, and correspondingly, with Φ having full rank, the matrix $\bar{C}$ in $\bar{h}(r)$ is negative definite. In that case, if the unique solution r^* of $\bar{h}(r) = 0$ lies in H, and if H is a closed ball centered at the origin with sufficiently large radius, then, using the negative definiteness of $\bar{C}$, it can be shown that no points r on the boundary of H can be stationary for the above o.d.e., so $r_t \overset{a.s.}{\to} r^*$.

Similar to the discussion in Remark 3.5, the question of whether the conclusion of Prop. 4.1 holds for a stepsize sequence that decreases at a rate slower than $1/t$ is closely connected to the rate of the convergence in Eqs. (39) and (40). (See the discussion in [KY03, Example 6.1, p. 171].)

4.2 Extension to Compact Space MDP

We now extend the convergence analysis of the off-policy LSTD(λ) algorithm in Section 3 for finite space MDP models to MDP with a compact state and action space $\mathcal{I}$. In particular, we focus on the case where $\mathcal{I}$ is a compact metric space, the per-stage cost function is continuous, and both the behavior and the target policies induce weak Feller Markov chains on $\mathcal{I}$. The results of Section 3 then extend directly. The case of more general compact space MDP models is a subject for future research.

Let Q and P denote the transition probability kernels of the Markov chains on $\mathcal{I}$ induced by the target and behavior policies, respectively, i.e.,

$$Q = \big\{Q(i, A), i \in \mathcal{I}, A \in \mathcal{B}(\mathcal{I})\big\}, \qquad P = \big\{P(i, A), i \in \mathcal{I}, A \in \mathcal{B}(\mathcal{I})\big\}.$$

Abusing notation, we still let $\{i_t\}$ denote the compact space Markov chain with transition kernel P. We will later use P_S to denote the transition probability kernel of the chain $\{(i_t, Z_t)\}$. We impose the following conditions on P, Q, the per-stage costs and the approximation subspace.

Assumption 4.1.

(i) The Markov chain $\{i_t\}$ is weak Feller and has a unique invariant probability measure ξ.

(ii) For each $i \in \mathcal{I}$, the conditional probability $Q(i,\cdot)$ is absolutely continuous with respect to $P(i,\cdot)$, with $\zeta(i,\cdot)$ being one version of the Radon-Nikodym derivative. The function ζ is continuous on $\mathcal{I}^2$.

Assumption 4.2.

(i) The per-stage transition cost $g(i,j)$ is a continuous function on $\mathcal{I}^2$.

(ii) The approximation subspace $\mathcal{H}$ is the linear span of $\{\phi_1,\ldots,\phi_d\}$, where $\phi = (\phi_1,\ldots,\phi_d)$ is an $\Re^d$-valued continuous function on $\mathcal{I}$.

4.2.1 The Approximation Framework and Algorithm

Assumption 4.1 implies that the transition probability kernel Q must also have the weak Feller property.[4] Then, with a continuous per-stage transition cost function under Assumption 4.2(i), the cost function J^* of the policy associated with Q is continuous. It satisfies the Bellman equation

$$J = T(J), \qquad \text{where} \;\; T(J) = \bar{g} + \alpha Q J,$$

and $\bar{g}$ is the expected one-stage cost function; and it also satisfies the multistep Bellman equation $J = T^{(\lambda)} J$, $\lambda \in [0,1]$ defined as in Eq. (5), all of which are now functional equations. (See e.g., Bertsekas and Shreve [BS78] for general space MDP theory.)

In the TD approximation framework, we consider the set of continuous functions as a subset of the larger space $\mathcal{L}^2(\mathcal{I},\xi) = \{f \mid f : \mathcal{I} \to \Re, \int f^2(x)\xi(dx) < \infty\}$ with semi-inner product $\langle\cdot,\cdot\rangle$ and the associated seminorm $\|\cdot\|_{2,\xi}$ given, respectively, by

$$\langle f, \hat{f}\rangle = \int f(x)\hat{f}(x)\,\xi(dx), \qquad \|f\|_{2,\xi}^2 = \langle f, f\rangle, \qquad f,\, \hat{f} \in \mathcal{L}^2(\mathcal{I},\xi).$$

For $\mathcal{L}^2(\mathcal{I},\xi)$, denote by $L^2(\mathcal{I},\xi)$ the factor space of equivalent classes (corresponding to the equivalence relation $\sim$ defined by $f \sim \hat{f}$ if and only if $\|f - \hat{f}\|_{2,\xi} = 0$). For any $f \in \mathcal{L}^2(\mathcal{I},\xi)$, let $f^\sim$ denote its equivalent class in $L^2(\mathcal{I},\xi)$, and let $\mathcal{H}^\sim$ denote the subspace of equivalent classes of f, $f \in \mathcal{H}$. We consider the projected multistep Bellman equation

$$J^\sim = \Pi T^{(\lambda)}(J), \quad J \in \mathcal{H}, \qquad \Leftrightarrow \qquad J = \operatorname*{arg\,min}_{f\in\mathcal{H}} \|T^{(\lambda)}J - f\|_{2,\xi}^2, \tag{41}$$

where $\Pi : L^2(\mathcal{I},\xi) \to L^2(\mathcal{I},\xi)$ is the projection onto $\mathcal{H}^\sim$ with respect to the $\|\cdot\|_{2,\xi}$-norm. Since $\mathcal{I}$ is compact, Assumption 4.2 implies the boundedness of the one-stage cost function $\bar{g}$ as well as the boundedness of any function $f \in \mathcal{H}$, so for any $J \in \mathcal{H}$, $T^{(\lambda)}(J) \in \mathcal{L}^2(\mathcal{I},\xi)$ and $\Pi T^{(\lambda)}(J)$ is well defined. The projected equation (41) may not have a solution; however, this case will not be discussed here, since our focus is on the approximation of the equation by samples. By a direct calculation, a low-dimensional representation of (41) is now given by

$$\bar{C}r + \bar{b} = 0, \qquad r \in \Re^d,$$

[4] By [MT09, Prop. 6.1.1(i)], Q is weak Feller if $Qf \in \mathcal{C}_b(\mathcal{I})$ for all $f \in \mathcal{C}_b(\mathcal{I})$. We have $(Qf)(x) = \int \zeta(x,y)f(y)P(x,dy)$. Using the continuity of ζ and the weak Feller property of P under Assumption 4.1, and using also the fact that a continuous function on a compact space is bounded and uniformly continuous, it can be verified that for any continuous function f, Qf is also continuous. So Q has the weak Feller property.

where

$$\bar{C} = \begin{bmatrix} \langle \phi_1 \,,\, Q^{(\lambda)}(\alpha Q - I)\phi_1 \rangle & \cdots & \langle \phi_1 \,,\, Q^{(\lambda)}(\alpha Q - I)\phi_d \rangle \\ \vdots & \ddots & \vdots \\ \langle \phi_d \,,\, Q^{(\lambda)}(\alpha Q - I)\phi_1 \rangle & \cdots & \langle \phi_d \,,\, Q^{(\lambda)}(\alpha Q - I)\phi_d \rangle \end{bmatrix}, \qquad \bar{b} = \begin{bmatrix} \langle \phi_1 \,,\, Q^{(\lambda)}\bar{g} \rangle \\ \vdots \\ \langle \phi_d \,,\, Q^{(\lambda)}\bar{g} \rangle \end{bmatrix},$$

and $Q^{(\lambda)}$ in the above is defined by the weighted sum of m-step transition probability kernels Q^m:

$$Q^{(\lambda)} = \sum_{m=0}^{\infty} (\lambda\alpha)^m Q^m$$

[cf. Eq. (7)], and it is an operator on the space of measurable functions on $\mathcal{I}$.

The off-policy LSTD(λ) algorithm takes the same form as the one in the finite space case, but has the Radon-Nikodym derivative $\zeta(i, j)$ in place of the ratios $\frac{q_{ij}}{p_{ij}}$ [cf. Eqs. (8)-(10)]:

$$Z_t = \beta\, \zeta(i_{t-1}, i_t) \cdot Z_{t-1} + \phi(i_t), \tag{42}$$
$$b_t = (1 - \gamma_t) b_{t-1} + \gamma_t Z_t\, \zeta(i_t, i_{t+1}) \cdot g(i_t, i_{t+1}), \tag{43}$$
$$C_t = (1 - \gamma_t) C_{t-1} + \gamma_t Z_t \left(\alpha \zeta(i_t, i_{t+1}) \cdot \phi(i_{t+1}) - \phi(i_t) \right)', \tag{44}$$

where $\beta = \lambda\alpha$ and $\phi(i) = (\phi_1(i), \ldots, \phi_d(i))$ is viewed as a $d \times 1$ vector. The goal is again to use sample-based approximations (b_t, C_t) to estimate $(\bar{b}, \bar{C})$, which define the projected Bellman equation. As before, we will study the iterates Z_t and

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

where ψ is a real-valued (corresponding to b_t) or $\Re^d$-valued (corresponding to C_t) continuous function on $\mathcal{I}^2$. In particular, it can be seen from Eqs. (43)-(44) that depending on the choice of ψ, $\{G_t\}$ specializes to $\{b_t\}$ or $\{C_t\}$:

$$G_t = \begin{cases} b_t & \text{if } \psi(i, j) = \zeta(i, j) \cdot g(i, j), \\ C_t & \text{if } \psi(i, j) = \alpha\zeta(i, j) \cdot \phi(j) - \phi(i). \end{cases} \tag{45}$$

We write ψ in terms of its components as $(\psi_1, \ldots, \psi_m)$, for $m = 1$ or d. The convergence of $\{b_t\}, \{C_t\}$ to $\bar{b}, \bar{C}$, respectively, in any mode, amounts to the convergence of $\{G_t\}$ to

$$G^* = \begin{bmatrix} \langle \phi_1 \,,\, Q^{(\lambda)}\bar{\psi}_1 \rangle & \cdots & \langle \phi_1 \,,\, Q^{(\lambda)}\bar{\psi}_m \rangle \\ \vdots & \ddots & \vdots \\ \langle \phi_d \,,\, Q^{(\lambda)}\bar{\psi}_1 \rangle & \cdots & \langle \phi_d \,,\, Q^{(\lambda)}\bar{\psi}_m \rangle \end{bmatrix}, \tag{46}$$

where $\bar{\psi}_j$ is defined to be the mean of the jth component of ψ, as in the finite space case:

$$\bar{\psi}_j(i) = E\big[\psi_j(i_0, i_1) \mid i_0 = i\big], \qquad i \in \mathcal{I}.$$

4.2.2 Convergence Analysis

We now show the convergence of $\{G_t\}$ to G^* in mean and with probability one under Assumptions 4.1 and 4.2 and proper conditions on the stepsizes γ_t. First, we redefine L_ℓ^t, $\ell < t$ appearing in the analysis of Section 3 to be

$$L_\ell^t = \zeta(i_\ell, i_{\ell+1}) \cdot \zeta(i_{\ell+1}, i_{\ell+2}) \cdots \zeta(i_{t-1}, i_t), \tag{47}$$

and define $L_t^t = 1$. Under Assumption 4.1(ii), we have as in the finite space case,

$$E[L_\ell^t \mid i_\ell] = 1 \quad a.s.$$

The conclusions of Lemmas 3.1 and 3.2 continue to hold in the compact space case considered here. In particular, for Lemma 3.1 to hold, it is sufficient that $\|\phi(i)\|$ is uniformly bounded on $\mathcal{I}$, which is implied by Assumption 4.2(ii), while Lemma 3.2 holds by the definition of Z_t, requiring no extra conditions. We can now extend the convergence analysis of Sections 3.2 and 3.3 straightforwardly, using most of the proofs given there.

Extending Theorem 3.1, we have the convergence of $\{G_t\}$ in mean stated in slightly more general terms as follows.

Proposition 4.2. *Let $h(z,i,j)$ be a vector-valued continuous function on $\Re^d \times \mathcal{I}^2$ which is Lipschitz continuous in z uniformly with respect to (i,j). Let*

$$G_t^h = (1-\gamma_t)G_{t-1}^h + \gamma_t h(Z_t, i_t, i_{t+1})$$

with the stepsize sequence $\{\gamma_t\}$ satisfying Assumption 2.2. Then under Assumptions 4.1 and 4.2(ii), there exists a constant $G^{h,*}$ such that for each initial condition (z_0, G_0),

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

Proof. The proof is almost the same as that of Theorem 3.1. Suppressing the superscript h for simplicity, we first consider for a positive integer T, the process $\{(\widetilde{Z}_{t,T}, \widetilde{G}_{t,T})\}$ as defined in the proof of Theorem 3.1: $\widetilde{Z}_{t,T} = Z_t$ for $t \le T$; $\widetilde{G}_{0,T} = G_0$; and

$$\widetilde{Z}_{t,T} = \phi(i_t) + \beta L_{t-1}^t \phi(i_{t-1}) + \cdots + \beta^T L_{t-T}^t \phi(i_{t-T}), \quad t > T; \tag{48}$$

$$\widetilde{G}_{t,T} = (1-\gamma_t)\widetilde{G}_{t-1,T} + \gamma_t h\big(\widetilde{Z}_{t,T}, i_t, i_{t+1}\big), \quad t \ge 1. \tag{49}$$

By Assumptions 4.1(ii) and 4.2(ii), ζ and ϕ are uniformly bounded on their domains. Consequently, $\{\|\widetilde{Z}_{t,T}\|\}$ can be bounded by some deterministic constant depending on T, and so are $\{\|h(\widetilde{Z}_{t,T}, i_t, i_{t+1})\|\}$ and $\{\|\widetilde{G}_{t,T}\|\}$ because of the boundedness of h on compact sets and the assumption $\gamma_t \in (0,1]$ (Assumption 2.2).

We then show that $\{\widetilde{G}_{t,T}\}$ converges almost surely to a constant G_T^* independent of the initial condition. To this end, we view $h(\widetilde{Z}_{t,T}, i_t, i_{t+1})$ as a function of $X_t = (i_{t-T}, i_{t-T+1}, \ldots, i_{t+1})$ for $t > T$, and we write it as $\hat{h}(X_t)$. Let $Y_t = (Y_{1,t}, Y_{2,t}) = (X_t, \hat{h}(X_t)), t > T$. We can write the iteration for $\widetilde{G}_{t,T}$, $t > T$ as

$$\widetilde{G}_{t,T} = \widetilde{G}_{t-1,T} + \gamma_t f\big(Y_t, \widetilde{G}_{t-1,T}\big),$$

where the function f is given by $f(y,G) = y_2 - G$ for $y = (y_1, y_2)$. Then we have the following facts:

(i) f is continuous in (y,G) and Lipschitz in G uniformly with respect to y.

(ii) $\{\widetilde{G}_{t,T}\}$ is bounded.

(iii) $\{Y_t, t > T\}$ is a Feller chain on a compact metric space which is independent of the initial G_0, and moreover, it has a unique invariant probability measure. This follows from Assumption 4.1(i) and the continuity of h: since $\{i_t\}$ is a Feller chain on a compact metric space, $\{X_t\}$ is also a Feller chain, which together with $\hat{h}$ being continuous implies that $\{Y_t, t > T\}$ is also weak Feller. The unique invariant probability measure of the latter chain is clearly determined by that of $\{i_t\}$.

Using these facts, we can apply the result of Borkar [Bor08, Chap. 6, Lemma 6, Theorem 7 and Cor. 8] to obtain that with E_0 denoting expectation under the stationary distribution of the Markov chain $\{i_t\}$,

$$\widetilde{G}_{t,T} \overset{a.s.}{\to} G_T^*, \qquad \text{where } \; G_T^* = E_0\big[h(\widetilde{Z}_{k,T}, i_k, i_{k+1})\big], \quad \forall k > T.$$

This is Eq. (28) in the proof of Theorem 3.1. We then apply the rest of the latter proof. □

The sequence $\{G_t\}$ is a special case of the sequence $\{G_t^h\}$ in the proposition, with the function h given by $h(z,i,j) = z\psi(i,j)'$. In this case, similar to the derivation given after the proof of Theorem 3.1, it can be shown that $G^{h,*} = G^*$ given in Eq. (46).

We now proceed to show the ergodicity of $\{(i_t, Z_t)\}$ and the almost sure convergence of $\{G_t\}$, extending Theorems 3.2 and 3.3. In what follows, we use P_S to denote the transition probability kernel of the Markov chain $\{(i_t, Z_t)\}$ on the metric space $S = \mathcal{I} \times \Re^d$.

Since ζ and ϕ are continuous functions under our assumptions, it can be verified directly that if $\{i_t\}$ is weak Feller, then $\{(i_t, Z_t)\}$ is also weak Feller. We state this as a lemma, omitting the proof.

Lemma 4.1. *Under Assumptions 4.1 and 4.2(ii), the Markov chain $\{(i_t, Z_t)\}$ is weak Feller.*

As in the finite space case, this together with the boundedness in probability of $\{(i_t, Z_t)\}$ indicated by Lemma 3.1(ii) implies that $\{(i_t, Z_t)\}$ has at least one invariant probability measure π. But we will now give an alternative way of reasoning for this, which is much more general and does not rely on which type of chain $\{i_t\}$ is or whether ϕ is bounded. The argument is based on constructing directly a stationary process $\{(i_t, Z_t)\}$, and it was used by Tsitsiklis and Van Roy [TV97, Eq. (5), p. 682] for analyzing the on-policy TD(λ) algorithm. Here we follow the reasoning given in Meyn [Mey07, Chap. 11.5, p. 520] for analyzing the on-policy LSTD algorithm, which is more general than the argument given in the former work and suitable for our case.

Lemma 4.2. *If $\{i_t\}$ has a unique invariant probability measure ξ and ϕ is Borel-measurable with $\int \|\phi\| d\xi < \infty$, then the Markov chain $\{(i_t, Z_t)\}$ has at least one invariant probability measure π with $E_\pi[\|Z_0\|] < \infty$.*

Proof. Consider a double-ended stationary Markov chain $\{i_t, -\infty < t < \infty\}$ with transition probability kernel P and probability distribution $\mathbf{P}^o$. Let $Y_t = (i_t, i_{t-1}, \ldots)$. Due to stationarity, for all t, the probability distributions of Y_t are the same, which is a measure on $\big(\mathcal{I}^\infty, \mathcal{B}(\mathcal{I}^\infty)\big)$ and will be denoted by μ_Y. We will consider in particular Y_0 and Y_1. For $y \in \mathcal{I}^\infty$, the space of Y_t, we write y in terms of its components as $(y_0, y_{-1}, \ldots)$. So corresponding to a realization of Y_0 given by $y = (\bar{i}_0, \bar{i}_{-1}, \ldots)$, $y_0 = \bar{i}_0, y_{-1} = \bar{i}_{-1}, \ldots$, for example.

Denote by E_0 expectation with respect to $\mathbf{P}^o$. We write $L_\ell^m, \ell \leq m$ given by Eq. (47) as $L(i_\ell, i_{\ell+1}, \ldots, i_m)$ to make the dependence on the i_t's explicit. We have

$$\sum_{k=0}^{\infty} \beta^k E_0\big[\|L(i_{-k}, \ldots, i_0) \cdot \phi(i_{-k})\|\big] = \sum_{k=0}^{\infty} \beta^k E_0\big[\|\phi(i_{-k})\|\big] < \infty,$$

which is equivalent to

$$\sum_{k=0}^{\infty} \beta^k \int \|L(y_{-k}, \ldots, y_0) \cdot \phi(y_{-k})\| \, d\mu_Y(y) < \infty.$$

Therefore by a theorem on integration [Rud66, Theorem 1.38, p. 28-29], we can define an $\Re^d$-valued measurable function on $\big(\mathcal{I}^\infty, \mathcal{B}(\mathcal{I}^\infty)\big)$ by

$$f(y) = \begin{cases} \sum_{k=0}^{\infty} \beta^k L(y_{-k}, \ldots, y_0) \cdot \phi(y_{-k}) & \text{if } y \in A; \\ 0 & \text{otherwise,} \end{cases} \tag{50}$$

where A is a measurable subset of $\mathcal{I}^\infty$ such that $\mu_Y(A) = 1$ and for all $y \in A$, the series appearing in the first case of the above definition converges to a vector in $\Re^d$; and f satisfies

$$\int \|f(y)\| \, d\mu_Y(y) < \infty \qquad \text{and} \qquad \int f(y) \, d\mu_Y(y) = E_0\big[f(Y_0)\big] = \sum_{k=0}^{\infty} \beta^k E_0\big[L_{-k}^0 \phi(i_{-k})\big]. \tag{51}$$

Let $Z_0^o = f(Y_0)$, and define Z_1^o by the recursion that defines Z_1 with $z_0 = Z_0^o$:

$$Z_0^o = f(Y_0), \qquad Z_1^o = \tilde{f}(Y_1) \stackrel{def}{=} \beta\zeta(i_0, i_1) \cdot f(Y_0) + \phi(i_1).$$

Then $\{(i_0, Z_0^o), (i_1, Z_1^o)\}$ is a Markov chain with transition probability kernel P_S. Consider the two functions f and $\tilde{f}$. By the definition of f in Eq. (50) and the fact that $L(y_{\ell_1}, \dots, y_{\ell_2}) \cdot L(y_{\ell_2}, \dots, y_{\ell_3}) = L(y_{\ell_1}, \dots, y_{\ell_3})$ for $\ell_1 \leq \ell_2 \leq \ell_3$, we have

$$\tilde{f}(y) = f(y), \qquad \forall\, y \in A \cap (\mathcal{I} \times A).$$

Since $\mathbf{P}^o(Y_0 \in A) = \mu_Y(A) = 1$ implies $\mu_Y(\mathcal{I} \times A) = \mathbf{P}^o\big(Y_1 = (i_1, Y_0) \in \mathcal{I} \times A\big) = 1$, we have $\mu_Y\big(A \cap (\mathcal{I} \times A)\big) = 1$. So $\tilde{f}$ and f can differ only on the set $\big(A \cap (\mathcal{I} \times A)\big)^c$, which has μ_Y-measure zero. As they define Z_1^o and Z_0^o, respectively, this shows that (Y_0, Z_0^o) and (Y_1, Z_1^o) have the same distribution, and hence that (i_0, Z_0^o) and (i_1, Z_1^0) have the same distribution, which is an invariant probability measure of the chain $\{(i_t, Z_t)\}$. Denote the latter by π. We have by Eq. (51), $E_\pi[\|Z_0\|] = E_0[\|Z_0^o\|] = \int \|f(y)\|\, d\mu_Y(y) < \infty$. □

The following proposition parallels Theorem 3.2 and shows that the chain $\{(i_t, Z_t)\}$ has a unique invariant probability measure and is ergodic.

Proposition 4.3. *Under Assumptions 4.1 and 4.2(ii), the Markov chain $\{(i_t, Z_t)\}$ has a unique invariant probability measure π, and for each initial condition x, almost surely, the sequence of occupation measures $\{\mu_{x,t}\}$ converges weakly to π.*

Proof. Let π be any invariant probability measure of $\{(i_t, Z_t)\}$, the existence of which follows from Lemma 4.2. First, we argue exactly as in the proof of Theorem 3.2, using Prop. 4.2 in place of Theorem 3.1, to establish that there exists a subset F of S with $\pi(F) = 1$, and for each initial condition $x = (\bar{i}, z)$ such that $(\bar{i}, \bar{z}) \in F$ for some $\bar{z}$, $\{\mu_{x,t}\}$ converges weakly to π, $\mathbf{P}_x$-almost surely.

Next we show π is unique. Suppose $\tilde{\pi}$ is another invariant probability measure. Then the preceding conclusion holds for a set $\tilde{F}$ with full $\tilde{\pi}$-measure. On the other hand, π and $\tilde{\pi}$ must have their marginals on $\mathcal{I}$ coincide with ξ, the unique invariant probability measure of the chain $\{i_t\}$. Let $F_{\mathcal{I}} = \{i \mid (i, z) \in F \text{ for some } z\}$ and define $\tilde{F}_{\mathcal{I}}$ similarly as the projection of $\tilde{F}$ on $\mathcal{I}$. The fact $\pi(F) = \tilde{\pi}(\tilde{F}) = 1$ implies $\xi(F_{\mathcal{I}}) = \xi(\tilde{F}_{\mathcal{I}}) = 1$, so $F_{\mathcal{I}} \cap \tilde{F}_{\mathcal{I}} \neq \emptyset$ and there exists a state $\bar{i}$ with $(\bar{i}, \bar{z}) \in F$ and $(\bar{i}, \hat{z}) \in \tilde{F}$ for some $\bar{z}, \hat{z}$. Then, by the preceding proof, for any initial condition $x = (\bar{i}, z)$ with $z \in \Re^d$, $\mu_{x,t} \to \pi$ and $\mu_{x,t} \to \tilde{\pi}$ weakly, $\mathbf{P}_x$-almost surely. This shows $\pi = \tilde{\pi}$ and π is the unique invariant probability measure.

Finally, consider initial conditions $x = (\bar{i}, \bar{z})$ with $\bar{i} \notin F_{\mathcal{I}}$. Because $\{(i_t, Z_t)\}$ is weak Feller (Lemma 3.4), has a unique invariant probability measure, and also satisfies the drift condition given in Lemma 3.1(i) with the stochastic Lyapunov function $V(i, z) = \|z\|$, which is nonnegative, continuous and coercive on S, we have the almost sure weak convergence of $\{\mu_{x,t}\}$ to π also for each $x \notin F$ by [Mey89, Props. 3.2, 4.2]. This completes the proof. □

Let us use E_π to denote also the expectation with respect to the stationary distribution of $\{(i_t, Z_t)\}$. Similar to the proof of Prop. 3.2, it can be seen that the conclusion $E_\pi[\|Z_0\|] < \infty$ of Lemma 4.2 implies that $E_\pi[\|h(Z_0, i_0, i_1)\|] < \infty$ for all functions h satisfying the conditions in Prop. 4.2, that is, all vector-valued continuous functions $h(z, i, j)$ that are Lipschitz continuous in z uniformly with respect to (i, j). Thus we can extend Theorem 3.3 as follows.

Proposition 4.4. *Let h and $\{G_t^h\}$ be as defined in Prop. 4.2. Let the stepsize in G_t^h be $\gamma_t = 1/(t+1)$. Then, under Assumptions 4.1 and 4.2(ii), there exists a set $A \subset \mathcal{I}$ with $\xi(A) = 1$, where ξ is the unique invariant probability measure of $\{i_t\}$, such that for each initial condition $(\bar{i}_0, z_0, G_0^h)$ with $\bar{i}_0 \in A$, $G_t^h \stackrel{a.s.}{\to} G^{h,*}$, where $G^{h,*} = E_\pi\big[h(Z_0, i_0, i_1)\big]$ is the constant in Prop. 4.2.*

Proof. We argue exactly as in the proof of Theorem 3.3, using Prop. 4.2 in place of Theorem 3.1, to establish the convergence of $\{G_t^h\}$ to $G^{h,*}$, first for each initial condition G_0^h and $x = (\bar{i}_0, z_0) \in F$, where F is a set of full π-measure, and then for each initial condition G_0^h and $x = (\bar{i}_0, z_0)$ where $\bar{i}_0 \in A = \{i \mid (i, z) \in F \text{ for some } z\}$. Since the marginal of π on $\mathcal{I}$ coincides with ξ and $\pi(F) = 1$, the set A, being the projection of F on $\mathcal{I}$, has measure 1 under ξ. The proof of the expression of $G^{h,*}$ is the same as that in Theorem 3.3. □

Remark 4.1. The conclusions of Props. 4.4 and 4.3 are stronger than what we can obtain by just applying the strong law of large numbers for the stationary process $\{(i_t, Z_t)\}$, without using its Feller property and the weak convergence result of Prop. 4.2. In the latter case, what we can claim directly is only that $\{G_t^h\}$ converges almost surely for the stepsize $\gamma_t = 1/(t + 1)$ and each initial condition as in Prop. 4.4.

Unlike in the finite space case, Prop. 4.4 asserts the almost sure convergence of $\{G_t^h\}$ only for the subset of initial conditions with $\bar{i}_0 \in A$. However, for the rest of the initial conditions, Prop. 4.3 implies that we can use modified bounded iterates to obtain a good approximation of $G^{h,*}$, as noted in Remark 3.4. Thus the conclusions we obtain in this compact space case are practically as strong as those in the finite space case. □

The above theorems apply to the off-policy LSTD(λ) iterates $\{G_t\}$ with the function h being $h(z, i, j) = z\psi(i, j)'$. They can also be applied to analyzing an off-policy TD(λ) algorithm for the compact space MDP model, similar to that in Section 4.1.

5 Discussion

While we have focused on the discounted total cost problems, the off-policy LSTD(λ) algorithm and the analysis given in the paper can be applied to average cost problems if a reliable estimate of the average cost of the target policy is available. For details we refer to the discussion at the end of [Yu10]. Here we mention briefly the application of the results of Section 3 in a related, non-MDP context of approximate solutions of linear fixed point equations. We then conclude the paper by addressing some topics for future research.

Consider approximately solving a linear fixed point equation

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

where $A = [a_{ij}]$ is an $n \times n$ matrix and b an n-dimensional vector. We may apply the TD methods, as discussed in Bertsekas and Yu [BY09]. Compared with policy evaluation in MDP, the main difference is that the substochastic matrix αQ in the Bellman equation (1) is now replaced by an arbitrary matrix A.

In particular, the TD(λ) approximation framework and algorithms can be applied for $\lambda \in [0, 1]$ such that $\lambda \sum_{j=1}^n |a_{ij}| < 1$ for all i. If we let $|A|$ be the signless version of A, with the (i, j)th entry being $|a_{ij}|$, then the latter condition on λ is equivalent to $\lambda|A|$ being a strictly substochastic matrix. For the above λ, analogous to the multistep Bellman equation, we can define the parametrized multistep fixed point mapping $T^{(\lambda)}$ involving the matrix $\sum_{k=0}^\infty \lambda^k A^k$. We can then 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 row/column indices of the matrix A as states, 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 given in Section 3 extends directly to this context, assuming the irreducibility of P and $|A| \prec P$, in addition to $\lambda|A|$ being strictly substochastic. We only need a slight modification in the analysis: when bounding various quantities of interest, we 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 we use the property

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

for some constant ν in place of Eq. (18). 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.

There are many problems deserving further study. One is the almost sure convergence of the unconstrained version of the on-line off-policy TD(λ) algorithm [BY09] for a general value of λ. (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 almost sure convergence of LSTD(λ) with a general stepsize sequence, possibly random; such stepsizes are useful particularly in two-time-scale policy iteration schemes, where LSTD(λ) is applied to policy evaluation at a faster time-scale, while incremental policy improvement is carried out at a slower time-scale. Another subject for future research is to extend the analysis in this paper to MDP models with a non-compact state-action space and unbounded costs. Finally, while we have focused on analyzing the asymptotic properties of the off-policy LSTD algorithm, its finite-sample properties such as those considered by Antos et al. [ASM08] and Lazaric et al. [LGM10] are also worth studying.

Acknowledgments

I thank Prof. Dimitri Bertsekas, Dr. Dario Gasbarra and Prof. George Yin for helpful suggestions and discussion, and Prof. Sean Meyn for pointing me to the material on LSTD in his book [Mey07]. A preliminary version of this paper appeared at the 27th International Conference on Machine Learning (ICML 2010). I thank the anonymous reviewers of ICML for their helpful feedback. This work is supported in part by Academy of Finland Grant 118653 (ALGODAN) and by the PASCAL Network of Excellence, IST-2002-506778.

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. Szepesv́ari, 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.

[Ber05] D. P. Bertsekas, *Dynamic programming and optimal control*, third ed., vol. I, Athena Scientific, Belmont, MA, 2005.

[Ber07] ———, *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).

[BS78] D. P. Bertsekas and S. Shreve, *Stochastic optimal control: The discrete time case*, Academic Press, 1978.

[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.

[Dud03] R. M. Dudley, *Real analysis and probability*, 2nd ed., Cambridge University Press, New York, 2003.

[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.

[LGM10] A. Lazaric, M. Ghavamzadeh, and R. Munos, *Finite-sample analysis of LSTD*, Proc. The 27th Int. Conf. Machine Learning, 2010.

[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.

[RJ04] R. S. Randhawa and S. Juneja, *Combining importance sampling and temporal difference control variates to simulate Markov chains*, ACM Trans. Modeling and Computer Simulation **14** (2004), no. 1, 1–30.

[Rud66] W. Rudin, *Real and complex analysis*, McGraw-Hill, Inc., New York, 1966.

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

[Sch10] B. Scherrer, *Should one compute the temporal difference fix point or minimize the Bellman residual? The unified oblique projection view*, Proc. The 27th Int. Conf. Machine Learning, 2010.

[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.

[YB10] H. Yu and D. P. Bertsekas, *Error bounds for approximations from projected linear equations*, Mathematics of Operations Research **35** (2010), no. 2, 306–329.

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

[Yu10] H. Yu, *Convergence of least squares temporal difference methods under general conditions*, Tech. Report C-2010-1, Dept. Computer Science, University of Helsinki, 2010.

Appendix: A Numerical Example

In this appendix, we use a simple 2-state example to illustrate the unboundedness of $\{Z_t\}$ and the convergence behavior of the LSTD(λ) algorithm for different stepsize sequences.

We let $\beta = 0.98$,

$$Q = \begin{bmatrix} 0.2 & 0.8 \\ 0.5 & 0.5 \end{bmatrix}, \qquad P = \begin{bmatrix} 0.45 & 0.55 \\ 0.6 & 0.4 \end{bmatrix},$$

$$\Phi' = [\phi(1) \;\; \phi(2)] = [2 \;\; 1], \qquad \psi(i,j) = 1, \;\; i,j \in \{1,2\}.$$

Thus Z_t, G_t are one-dimensional and

$$\left[\tfrac{q_{ij}}{p_{ij}}\right] = \begin{bmatrix} 0.44 & 1.45 \\ 0.83 & 1.25 \end{bmatrix}.$$

There are several simple cycles of states satisfying the conditions of Prop. 3.1. For example, $\{2,2\}$ is such a cycle with $\beta \frac{q_{22}}{p_{22}} = 1.225 > 1$, $\{1,2,1\}$ is another one with $\beta^2 \frac{q_{12}}{p_{12}} \cdot \frac{q_{21}}{p_{21}} = 1.164 > 1$, and $\{1,2,2,1\}$ is yet another with $\beta^3 \frac{q_{12}}{p_{12}} \cdot \frac{q_{22}}{p_{22}} \cdot \frac{q_{21}}{p_{21}} = 1.426 > 1$. So $\{Z_t\}$ is almost surely unbounded (cf. Prop. 3.1 and the discussion preceding it). This phenomenon is demonstrated by a simulation run shown in the figure below, where the maximal values of $\|Z_t\|$ in intervals of length $C = 5 \times 10^6$ are plotted.

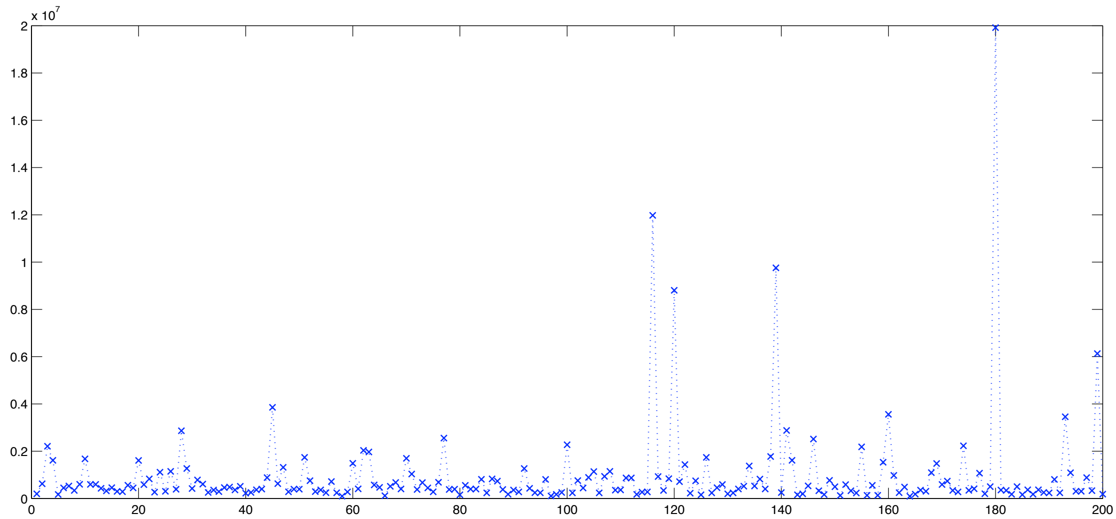

Figure 1: $\{Z_t\}$ from a simulation run. Y-axis: $\max_{(k-1)C<t\le kC} \|Z_t\|$ where $C = 5 \times 10^6$; X-axis: k.

For this example, it can be verified also that the variance of Z_t increases to infinity as t increases. In the next figure, we compare the behavior of G_t for stepsizes γ_t that decrease at different rates.

We plotted the values of $\{G_t\}$ in a simulation run for t in the time interval $(k-1)C < t \le kC$ with $k = 200$ and $C = 5 \times 10^6$ as in the previous figure. The horizontal axis shows $t - (k-1)C$.

For $\gamma_t = O(1/t), O(1/t^{0.95})$ and $O(1/t^{0.9})$, the corresponding $\{G_t\}$ is converging to G^*, while for $\gamma_t = O(1/t^{0.8})$ and $O(1/t^{0.7})$, the corresponding $\{G_t\}$ seems to converge to G^* not almost surely, but only weakly, as demonstrated by its oscillation around G^*. These simulation results seem to confirm that almost sure convergence of $\{G_t\}$ may occur only for those stepsizes that decrease at a rate much faster than $t^{-0.5}$. (Compare with Theorem 3.3, Remark 3.5 and Theorem 3.1.)

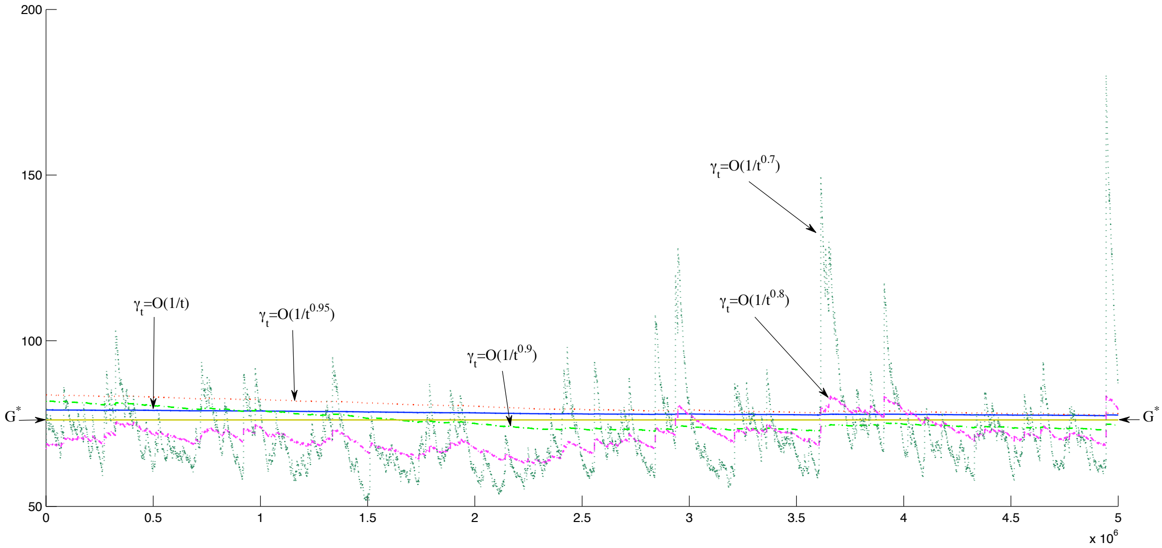

Figure 2: Behavior of $\{G_t\}$ for different stepsizes γ_t.