SIAM J. CONTROL OPTIM.
Vol. 50, No. 6, pp. 3310–3343

LEAST SQUARES TEMPORAL DIFFERENCE METHODS: AN ANALYSIS UNDER GENERAL CONDITIONS*

HUIZHEN YU†

Abstract. We consider approximate policy evaluation for finite state and action Markov decision processes (MDP) with the least squares temporal difference (LSTD) algorithm, LSTD(λ), in an exploration-enhanced learning context, where policy costs are computed from observations of a Markov chain different from the one corresponding to the policy under evaluation. We establish for the discounted cost criterion that LSTD(λ) converges almost surely under mild, minimal conditions. We also analyze other properties of the iterates involved in the algorithm, including convergence in mean and boundedness. Our analysis draws on theories of both finite space Markov chains and weak Feller Markov chains on a topological space. Our results can be applied to other temporal difference algorithms and MDP models. As examples, we give a convergence analysis of a TD(λ) algorithm and extensions to MDP with compact state and action spaces, as well as a convergence proof of a new LSTD algorithm with state-dependent λ-parameters.

Key words. Markov decision processes, approximate dynamic programming, temporal difference methods, importance sampling, Markov chains

AMS subject classifications. 90C40, 90C39, 65C05

DOI. 10.1137/100807879

1. Introduction.

We consider the following problem of computing expected costs associated with Markov chains. Let $\{i_t\}$ be an irreducible Markov chain on a finite state space $\mathcal{I} = \{1, 2, \ldots, n\}$ with transition probability matrix P. A trajectory of $\{i_t\}$ is observed together with transition costs $g(i_t, i_{t+1})$, $t = 0, 1, \ldots$, where $g : \mathcal{I}^2 \to \Re$. From these observations, we wish to estimate the expected discounted total costs, $E\big[\sum_{t\geq 0} \alpha^t g(i'_t, i'_{t+1}) \mid i'_0 = i\big]$, $i \in \mathcal{I}$, where $\{i'_t\}$ is a Markov chain on $\mathcal{I}$ with transition probability matrix $Q \neq P$. Here $\alpha < 1$ is the discount factor and Q is assumed to be absolutely continuous with respect to P (denoted $Q \prec P$) in the sense that

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

where p_{ij}, q_{ij} are the (i, j)th elements of P, Q, respectively. Aside from the ratios p_{ij}/q_{ij} between their elements, the transition matrices themselves can be unknown.

This problem appears in policy evaluation for discounted cost, finite state, and action Markov decision processes (MDP) in the learning and simulation context, where the model of the MDP is unavailable and policy costs are estimated from data. More specifically, for stationary policies of the MDP, which induce Markov chains on the state and state-action spaces, we estimate their expected discounted costs from observations of actions applied, state transitions, and transition costs incurred. The mechanism that generates the observations is represented by the transition matrix P introduced earlier, with the observation process represented by the Markov chain $\{i_t\}$,

*Received by the editors September 7, 2010; accepted for publication (in revised form) August 21, 2012; published electronically December 12, 2012. This work was supported in part by Academy of Finland grant 118653 (ALGODAN), by the PASCAL Network of Excellence, IST-2002-506778, and by Air Force grant FA9550-10-1-0412. A preliminary, abridged version of this paper appeared at the 27th International Conference on Machine Learning (ICML 2010).
http://www.siam.org/journals/sicon/50-6/80787.html

†Laboratory for Information and Decision Systems (LIDS), Massachusetts Institute of Technology, Cambridge, MA 02139 (janey_yu@mit.edu).

where the space $\mathcal{I}$ will in general correspond to the state-action space of the MDP. A policy for evaluation is represented by the transition matrix Q; if this policy were used to control the MDP, it would induce a Markov chain represented by $\{i'_t\}$. The fact $Q \neq P$ means that we evaluate a policy without actually applying it in the MDP. This is referred to as "off-policy" learning in the terminology of reinforcement learning (in contrast to "on-policy" learning for the case $P = Q$).

There are several motivations for evaluating policies in this seemingly indirect way. The associated computation methods use importance sampling techniques and form an established methodology for reinforcement learning (Sutton and Barto [32]) and simulation-based large-scale dynamic programming in general (Glynn and Iglehart [16]). In the learning context, P corresponds to a policy that we currently use to govern the system, not necessarily for minimizing the cost, but for collecting information about the system. When P is suitably chosen, it allows us to evaluate many policies without having to try them out one by one. It can also balance our need for cost minimization (exploitation) and that for more information (exploration), thus providing a flexible approach to address the exploration-exploitation tradeoff in policy search. In the simulation context, P need not correspond to an actual policy; instead, any sampling mechanism may be used to induce dynamics P (which may not be realizable by any policy) for the purpose of efficient computation (Glynn and Iglehart [16]).

In this paper we analyze a particular approximation algorithm for solving the policy evaluation problem described above. It was proposed in Bertsekas and Yu [8], and it extends the least squares temporal difference (LSTD) algorithm for on-policy learning (Bradtke and Barto [12] and Boyan [11]). We shall refer to the algorithm as the off-policy LSTD or simply as LSTD when there is no confusion. Similarly, to avoid confusion when discussing related earlier works, we will differentiate policy evaluation algorithms into two types, "on-policy" and "off-policy," according to whether they are for $Q = P$ only or for the general case $Q \prec P$. To simplify the language of the discussion, we also adopt some other terms from reinforcement learning: we call the policy associated with Q the "target policy" and the sampling mechanism associated with P the "behavior policy" (although P need not correspond to an actual policy in the simulation context).

The off-policy LSTD algorithm [8] belongs to the family of temporal difference (TD) methods (Sutton [30]; see also the books by Bertsekas and Tsitsiklis [7], Sutton and Barto [32], Bertsekas [4], and Meyn [22]). Beyond the algorithmic level, TD methods share a common approximation framework: solve a projected Bellman equation

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

parametrized by $\lambda \in [0, 1]$. Here $T^{(\lambda)}$ is a λ-weighted multistep Bellman operator that has the cost vector J^* of the target policy as its unique fixed point, $J^* = T^{(\lambda)}(J^*)$, and Π is the projection onto an approximation subspace $\{\Phi r \mid r \in \Re^d\} \subset \Re^n$ (where Φ is an $n \times d$ matrix) with respect to a weighted Euclidean norm. The weights in the projection norm, in the case we consider, will be the steady-state probabilities of the Markov chain $\{i_t\}$ induced by the behavior policy. When the projected Bellman equation (2) is well defined, i.e., has a unique solution Φr^* in the approximation subspace, we use the solution to approximate J^*. There are approximation error bounds (Yu and Bertsekas [37]) and geometric interpretations of the approximation Φr^* as an oblique projection of J^* (Scherrer [29]) in this case. Here, however, we will

not discuss whether the projected Bellman equation is well defined; our interest will be in approximating this equation using sampling and LSTD.

Given λ, the projected Bellman equation (2) has an equivalent low-dimensional representation,

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

where $\bar{b}$ is a d-dimensional vector and $\bar{C}$ a $d \times d$ matrix (whose precise definitions will be given later). The off-policy LSTD(λ) algorithm—λ is an input parameter to the algorithm—uses the observations of states and transition costs, $i_t, g(i_t, i_{t+1}), t = 0, 1, \ldots$, to construct a sequence of equations

$$C_t r + b_t = 0,$$

with the goal of "approaching" in the limit (3). The algorithm takes into account the discrepancies between the behavior policy P and target policy Q 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 [16], Sutton and Barto [32], Precup, Sutton and Dasgupta [25] (one of the first off-policy TD(λ) algorithms), and Ahamed, Borkar, and Juneja [1].

We will analyze the convergence of the off-policy LSTD(λ) algorithm—the convergence of $\{(b_t, C_t)\}$ to $(\bar{b}, \bar{C})$—under the general assumption that P is irreducible and $Q \prec P$. In the earlier work [8], the almost sure convergence of the algorithm was shown for the special case where $\lambda\alpha \max_{(i,j)} \frac{q_{ij}}{p_{ij}} < 1$ (with 0/0 treated as 0), which is a restrictive condition because it requires either $P \approx Q$ or λ to be small. The case of a general value of λ is important in practice. Using a large value of λ not only can improve the quality of the cost approximation obtained from the projected Bellman equation (2), 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, 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, where r_t satisfies $C_t r_t + b_t = 0$, converges to the solution Φr^* of the projected Bellman equation (2) almost surely whenever (2) has a unique solution, and if (2) has multiple solutions, any limit point of $\{\Phi r_t\}$ is one of them.

The line of our analysis is considerably different from those in the literature for similar on-policy TD methods (Tsitsiklis and Van Roy [34], Nedić and Bertsekas [24]) and off-policy TD methods [25, 8]. This is mainly because in the off-policy case with a general value of λ, the auxiliary vectors Z_t (also called "eligibility traces"), calculated by TD algorithms to facilitate iterative computation, can exhibit unboundedness behavior and violate the analytical conditions used in the aforementioned works (see section 3.1, Remark 3.2 for a detailed account). We also do not follow the mean-o.d.e. proof method of stochastic approximation theory (see e.g., Kushner and Yin [17], Borkar [9, 10]), because to apply the method here would require us to verify conditions that are tantamount to the almost sure convergence conclusion we want to establish.

Our approach is to relate the LSTD(λ) iterates to a particular type of Markov chain and resort to the ergodic theory for these chains (Meyn and Tweedie [23],

Meyn [21]). 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 Markov chain $\{i_t\}$. But for proving the almost sure convergence, we did not find such arguments to be sufficient, in contrast with the on-policy LSTD case as analyzed by Meyn [22, Chap. 11.5]. Instead, we will study the Markov chain $\{(i_t, Z_t)\}$ on the topological space $\mathcal{I} \times \Re^d$ and exploit its weak Feller property as well as other properties to establish its ergodicity and the almost sure convergence of $\{b_t\}, \{C_t\}$.

We note that the study of the almost sure convergence of the off-policy LSTD(λ) is not solely of theoretical interest. Various TD algorithms use the same approximations b_t, C_t to build approximating models (e.g., preconditioned TD(λ) in Yao and Liu [35]) or fixed point iterations (e.g., for LSPE(λ), see Bertsekas and Yu [8]; and for scaled versions of LSPE(λ), see Bertsekas [5]). Thus the asymptotic behavior of these algorithms in the off-policy case 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 (e.g., LSPE(λ)), the convergence of $\{b_t\}, \{C_t\}$ on almost every sample path is critical. Moreover, the mode of convergence of the off-policy LSTD(λ) is also relevant for understanding the behavior of algorithms which use stochastic approximation-type iterations to solve projected Bellman equations (3), such as the on-line off-policy TD(λ) algorithm of [8] and the off-policy TD(λ) algorithm of [25]. While these algorithms do not directly compute b_t, C_t, they implicitly depend on the convergence properties of $\{b_t\}, \{C_t\}$. Thus our results and line of analysis are useful also for analyzing various algorithms other than LSTD.

Besides the main results mentioned above, this paper contains several additional results as applications or extensions of the main analysis. These include (i) a convergence analysis of a constrained version of an off-policy TD(λ) algorithm proposed in [8]; (ii) the extension of the analysis of the off-policy LSTD(λ) algorithm to special cases of MDP with compact state and action spaces; and (iii) a convergence proof of a recently proposed LSTD algorithm with state-dependent λ-parameters (Yu and Bertsekas [38]).

The paper is organized as follows. We describe the off-policy LSTD(λ) algorithm and specify notation and definitions in section 2. We present our main convergence results for finite space MDP in section 3. We then give in section 4 the additional results just mentioned. Finally, we discuss other applications of our results and future research in section 5.

2. Preliminaries. In this section we describe the off-policy LSTD(λ) algorithm for approximate policy evaluation and give some definitions in preparation for its convergence analysis. We consider the general policy evaluation problem as introduced at the beginning of this paper, which involves two Markov chains on $\mathcal{I}$, one with transition matrix P (associated with the behavior policy) and the other with transition matrix Q (associated with the target policy). Near the end of this section, we will describe in detail two applications of LSTD in MDP (Examples 2.1 and 2.2), where we will explain the correspondences between the space $\mathcal{I}$, the transition matrices P, Q, and the associated MDP contexts, as mentioned in the introduction. There it will also be seen that in applications, in spite of P and Q being unknown, the ratios between their elements, which are needed in LSTD, are naturally available.

Throughout the paper, we use $\{i_t\}$ to denote the Markov chain with transition matrix P, and we use i or $\bar{i}$ to denote specific states. As mentioned earlier, we require the following condition on P and Q.

Assumption 2.1. The Markov chain $\{i_t\}$ with transition matrix P is irreducible, and $Q \prec P$ in the sense of (1).

Let $J^* \in \Re^n$ be the vector with components $J^*(i)$, where

$$J^*(i) = E\Big[\sum_{t\geq 0} \alpha^t g(i'_t, i'_{t+1}) \mid i'_0 = i\Big]$$

is the expected discounted cost starting from state $i \in \mathcal{I}$ for the Markov chain $\{i'_t\}$ with transition matrix Q. This is the cost vector of the target policy that we wish to compute. By the theory of MDP (see, e.g, [26, 3]), J^* is the unique solution of the Bellman equation

$$J = T(J), \qquad \text{where } \ T(J) = \bar{g} + \alpha QJ \quad \forall J \in \Re^n, \tag{4}$$

and $\bar{g}$ is the vector of expected one-stage costs:

$$\bar{g}(i) = \sum_{j\in\mathcal{I}} q_{ij} g(i,j), \qquad i \in \mathcal{I}.$$

For $\lambda \in [0,1]$, define a multistep Bellman operator by

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

(In particular, $T^{(0)} = T$ and $T^{(1)}(\cdot) \equiv J^*$.) Then J^* is the unique solution of the multistep Bellman equation

$$J^* = T^{(\lambda)}(J^*).$$

We consider approximating J^* by a vector in a subspace $\mathcal{H} \subset \Re^n$ by solving the projected Bellman equation (2),

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

where Π is a projection onto $\mathcal{H}$, to be defined below.

Let Φ be an $n \times d$ matrix whose columns span the approximation subspace $\mathcal{H}$, i.e., $\mathcal{H} = \{\Phi r \,|\, r \in \Re^d\}$. While $\mathcal{H}$ has infinitely many representations, in practice, one often chooses first some matrix Φ based on the understanding of the MDP problem, and Φ then determines the subspace $\mathcal{H}$. Typically, Φ need not be stored because one has access to the function $\phi : \mathcal{I} \to \Re^d$ which maps i to the ith row of Φ; i.e.,

$$\Phi = \begin{bmatrix} \phi(1)' \\ \vdots \\ \phi(n)' \end{bmatrix},$$

where we treat $\phi(i)$ as $d \times 1$ vectors and we use the symbol $'$ to denote transpose. The vectors $\phi(i)$ are often referred to as "features" (of the states and actions of the MDP). Choosing the "feature-mapping" ϕ is extremely important in practice but is beyond the scope of this paper.

We define Π to be the projection onto $\mathcal{H}$ with respect to the weighted Euclidean norm $\|J\|_\xi = \big(\sum_{i\in\mathcal{I}} \xi(i) J(i)^2\big)^{1/2}$, where $\xi(i)$ are the steady-state probabilities of the

Markov chain $\{i_t\}$ and are strictly positive 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 $\xi(i)$ being the diagonal elements. 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

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

To approximate $\bar{b}, \bar{C}$, the off-policy LSTD(λ) algorithm [8, sec. 5.2] computes iteratively vectors b_t and matrices C_t, using the observations $i_t, g(i_t, i_{t+1}), t = 0, 1, \ldots$, generated under the behavior policy. To facilitate iterative computation, the algorithm also computes a third sequence of d-dimensional vectors Z_t. These iterates are defined as follows. With (Z_0, b_0, C_0) being the initial condition, for $t \geq 1$,

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

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

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

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]

If $P = Q$, then all the ratios $\frac{q_{i_{t-1} i_t}}{p_{i_{t-1} i_t}}$ appearing in the above iterates become 1, and the algorithm reduces to the on-policy LSTD algorithm [12, 11]. In general, the ratios $\frac{q_{ij}}{p_{ij}}$ needed for computing the iterates are available in practice, because they depend only on policy parameters and not on the parameter of the MDP model. Moreover, like the matrix Φ, they need not be stored and can be computed on-line. The details will be explained in Examples 2.1 and 2.2 shortly. So, like the on-policy TD algorithms, the off-policy LSTD algorithm is also a model-free method in practice.

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:

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

[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.

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 $\gamma_t \in (0,1]$, $\sum_t \gamma_t = \infty$, $\sum_t \gamma_t^2 < \infty$.

The question of convergence of $\{b_t\}, \{C_t\}$ now 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

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

For the two choices of ψ in (12), we have, respectively, $\Psi = \bar{g}$ or $(\alpha Q - I)\Phi$, $G^* = \bar{b}$ or $\bar{C}$ (cf. (7)).

Before proceeding to convergence analysis, we explain in the rest of this section some details of applications in MDP, which have been left out in our description of the LSTD algorithm so far. (These details will not be relied upon in our analysis.)

Applications of LSTD to Q-Factor and Cost Approximations in MDP

Consider an MDP which has a finite state space D and for each state $s \in D$, a finite set of feasible actions, $U(s)$. From state s with an action $u \in U(s)$, transition to state $\hat{s}$ occurs with probability $p(\hat{s} \mid s, u)$ and incurs cost $c(s, u, \hat{s})$. Let μ^o, μ be two stationary randomized policies: a stationary randomized policy is a function that maps each state s to a probability distribution on $U(s)$, and this distribution is denoted, for policies μ^o and μ, by $\mu^o(\cdot \mid s)$ and $\mu(\cdot \mid s)$, respectively. We apply μ^o in the MDP and observe a sequence of states and actions (s_t, u_t) and transition costs $c(s_t, u_t, s_{t+1}), t = 0, 1, \ldots$, from which we wish to evaluate the α-discounted, expected total costs of policy μ. Thus μ^o is the behavior policy and μ the target policy. We require that for every state s, $\mu(u \mid s) > 0$ implies $\mu^o(u \mid s) > 0$; i.e., possible actions of the target policy μ are also possible actions of the behavior policy μ^o.

There are two common ways to evaluate a policy with learning or simulation: evaluate the costs or evaluate the so-called Q-factors (or state-action values), which are costs associated with initial state-action pairs. The LSTD algorithm has slightly different forms in these two cases, so we describe them separately.

Example 2.1 (*Q-factor approximation*). For all $s \in D, u \in U(s)$, let $V^*(s,u)$ be the expected cost of starting from state s, taking action u, and then applying the policy μ. They are called Q-factors of μ, and in the learning context, they facilitate the computation of an improved policy. The Q-factors uniquely satisfy the Bellman equation: for all $s \in D, u \in U(s)$,

$$V^*(s,u) = \sum_{\hat{s} \in D} p(\hat{s} \mid s, u)\, c(s, u, \hat{s}) + \alpha \sum_{\hat{s} \in D} p(\hat{s} \mid s, u) \sum_{\hat{u} \in U(\hat{s})} \mu(\hat{u} \mid \hat{s})\, V^*(\hat{s}, \hat{u}). \tag{14}$$

We evaluate μ by approximating V^* with LSTD(λ). This Q-factor approximation problem can be cast into our framework with the following correspondences.

The space $\mathcal{I}$ corresponds to the set of state-action pairs, $\{(s,u) \mid s \in D, u \in U(s)\}$, the desired cost vector J^* corresponds to the Q-factors V^* of μ, and the Bellman equation $J^* = T(J^*)$ corresponds to (14). The Markov chain $\{i_t\}$ corresponds to the state-action process $\{(s_t,u_t)\}$ induced by μ^o. For $i,j \in \mathcal{I}$ with their associated state-action pairs being $(s,u),(\hat{s},\hat{u})$, respectively, the transition cost $g(i,j)$ and expected one-stage cost $\bar{g}(i)$ are given by

$$g(i,j) = c(s,u,\hat{s}), \qquad \bar{g}(i) = \sum_{\hat{s}\in D} p(\hat{s} \mid s,u)\, c(s,u,\hat{s}),$$

and the transition probabilities p_{ij}, q_{ij} are given by

$$p_{ij} = p(\hat{s} \mid s,u)\, \mu^o(\hat{u} \mid \hat{s}), \qquad q_{ij} = p(\hat{s} \mid s,u)\, \mu(\hat{u} \mid \hat{s}).$$

Notice that the ratio[2] $\frac{q_{ij}}{p_{ij}} = \frac{\mu(\hat{u}|\hat{s})}{\mu^o(\hat{u}|\hat{s})}$ that appears in (8)–(10) does not depend on the transition probability $p(\hat{s} \mid s,u)$ of the MDP model. Moreover, since they depend only on μ^o and μ, the n^2 terms $\frac{q_{ij}}{p_{ij}}, i,j \in \mathcal{I}$, need not be stored and can be calculated on-line when needed in the LSTD algorithm.

The assumption $Q \prec P$ is satisfied, since $\mu(u \mid s) > 0$ implies $\mu^o(u \mid s) > 0$. The Markov chain $\{i_t\}$ is irreducible if every state-action pair (s,u) is visited infinitely often under μ^o. (More generally, if this is not so, we can let $\mathcal{I}$ be a subset of (s,u) that forms a recurrent class under μ^o.)

Special to the case of Q-factor approximation is that the expected one-stage cost $\bar{g}(i)$, as defined above, does not depend on policy μ. This allows for a simplification in the LSTD(λ) iterates: the updates for b_t 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. (9)). The resulting sequence $\{b_t\}$ is a special case of the sequence $\{G_t\}$ that we will analyze, with the function $\psi(i,j) = g(i,j)$ (cf. (11)). ∎

Example 2.2 (*cost approximation*). With LSTD(λ), one can also approximate the cost vector of μ, which has a lower dimension than the Q-factors. The cost vector/function, denoted again by V^*, uniquely satisfies the Bellman equation: for all $s \in D$,

$$V^*(s) = \sum_{u\in U(s)} \mu(u \mid s) \sum_{\hat{s}\in D} p(\hat{s} \mid s,u)\big(c(s,u,\hat{s}) + \alpha V^*(\hat{s})\big).$$

We approximate V^* by a function of the form $\hat{\phi}(s)'r$, where $\hat{\phi}$ maps states s to $d \times 1$ vectors and $r \in \Re^d$ is a parameter. To obtain a model-free LSTD algorithm for this approximation, we first extend V^* to a larger, action-state space. For each state s, let $\tilde{V}^*(v,s) = V^*(s)$ for all actions $v \in \tilde{U}(s)$, where $\tilde{U}(s)$ is a set of actions defined as $\tilde{U}(s) = \big\{v \mid \exists \tilde{s},\, v \in U(\tilde{s}), \mu^o(v \mid \tilde{s}) p(s \mid \tilde{s}, v) > 0\big\}$ (the set of actions that can be taken by policy μ^o to lead to state s). The function $\tilde{V}^*$ is equivalent to the cost function V^* and uniquely satisfies the Bellman equation: for all $s \in D, v \in \tilde{U}(s)$,

$$\tilde{V}^*(v,s) = \sum_{u\in U(s)} \mu(u \mid s) \sum_{\hat{s}\in D} p(\hat{s} \mid s,u) \left(c(s,u,\hat{s}) + \alpha \tilde{V}^*(u,\hat{s})\right). \tag{15}$$

[2] If $p_{ij} = q_{ij} = 0$, then the ratio q_{ij}/p_{ij} can be defined arbitrarily.

We then consider approximating $\tilde{V}^*$ by $\hat{\phi}(s)'r$ for some parameter $r \in \Re^d$. This approximation problem can be cast into our framework with the following correspondences.

The space $\mathcal{I}$ corresponds to the set of action-state pairs, $\{(v,s) \mid s \in D, v \in \tilde{U}(s)\}$, and J^* and the Bellman equation $J^* = T(J^*)$ correspond to $\tilde{V}^*$ and (15), respectively. The Markov chain $\{i_t\}$ corresponds to the process $\{(u_{t-1}, s_t)\}$ induced by μ^o (the choice of u_{-1} is immaterial). For $i, j \in \mathcal{I}$ with their associated action-state pairs being $(v,s), (u,\hat{s})$, respectively, the transition cost $g(i,j) = c(s,u,\hat{s})$ if $u \in U(s)$, and it can be arbitrarily defined otherwise; and the transition probabilities are given by

$$p_{ij} = \mu^o(u \mid s)p(\hat{s} \mid s,u), \qquad q_{ij} = \mu(u \mid s)p(\hat{s} \mid s,u),$$

where we let $\mu^o(u \mid s) = \mu(u \mid s) = 0$ if $u \notin U(s)$. As in the previous example, here the ratio $\frac{q_{ij}}{p_{ij}} = \frac{\mu(u|s)}{\mu^o(u|s)}$ does not depend on the transition probability $p(\hat{s} \mid s,u)$ of the MDP model. Similarly, we have $Q \prec P$, since $\mu(u \mid s) > 0$ implies $\mu^o(u \mid s) > 0$; and P is irreducible if every state s is visited infinitely often under μ^o. Correspondingly, the off-policy LSTD(λ) iterates (cf. (8)–(10)) become

$$\begin{aligned}
Z_t &= \lambda\alpha\tfrac{\mu(u_{t-1}|s_{t-1})}{\mu^o(u_{t-1}|s_{t-1})} \cdot Z_{t-1} + \hat{\phi}(s_t),\\
b_t &= (1-\gamma_t)b_{t-1} + \gamma_t Z_t \cdot \tfrac{\mu(u_t|s_t)}{\mu^o(u_t|s_t)} \cdot c(s_t,u_t,s_{t+1}),\\
C_t &= (1-\gamma_t)C_{t-1} + \gamma_t Z_t \left(\alpha\tfrac{\mu(u_t|s_t)}{\mu^o(u_t|s_t)} \cdot \hat{\phi}(s_{t+1}) - \hat{\phi}(s_t)\right)'. \qquad \square
\end{aligned}$$

3. Main results. We analyze the convergence of the sequence $\{G_t\}$, defined by (11), 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 related to or needed in the subsequent analysis.

Throughout the paper, let $\|\cdot\|$ denote the norm $\|G\| = \max_{i,j}|G_{ij}|$ if G is a matrix with elements G_{ij}, and the infinity norm $\|G\| = \max_i |G_i|$ if G is a vector with components G_i. Matrix-valued processes (e.g., $\{G_t\}$) or matrix-valued functions will be generally regarded as vector-valued. We let "a.s." stand for "almost surely."

3.1. Some properties of LSTD iterates. We denote by L_ℓ^t the product of ratios of transition probabilities along a segment of the state sequence, $(i_\ell, i_{\ell+1}, \ldots, i_t)$, where $0 \le \ell \le t$:

$$L_\ell^t = \tfrac{q_{i_\ell i_{\ell+1}}}{p_{i_\ell i_{\ell+1}}} \cdot \tfrac{q_{i_{\ell+1} i_{\ell+2}}}{p_{i_{\ell+1} i_{\ell+2}}} \cdots \tfrac{q_{i_{t-1} i_t}}{p_{i_{t-1} i_t}}, \tag{16}$$

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

$$E[\, L_\ell^t \mid i_\ell] = 1. \tag{17}$$

The iterate Z_t given by (8) is

$$Z_t = \beta\tfrac{q_{i_{t-1} i_t}}{p_{i_{t-1} i_t}} \cdot Z_{t-1} + \phi(i_t) = \beta L_{t-1}^t \cdot Z_{t-1} + \phi(i_t), \tag{18}$$

where $\beta = \lambda\alpha$. By unfolding the right-hand side, it can be equivalently expressed as

$$Z_t = \beta^t L_0^t z_0 + \sum_{m=0}^{t-1} \beta^m L_{t-m}^t \phi(i_{t-m}), \tag{19}$$

where $Z_0 = z_0$ is the initial condition.

It is shown in Glynn and Iglehart [16, 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 [27]. In the infinite-horizon case we consider, using the iterative form (18), one can easily construct examples where the second moments of the variables Z_t (or any νth-order moments with $\nu > 1$) are unbounded as t increases. Furthermore, as we will show shortly (Proposition 3.1), under seemingly fairly common situations, Z_t is almost surely unbounded. Thus even for a finite space MDP, the case $P \neq Q$ sharply contrasts with the standard case $P = Q$, where $\{Z_t\}$ is by definition bounded.

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. First, we have the following property, which will be used in the convergence analysis of this and the following sections.

LEMMA 3.1.

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

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

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

(ii) *For each initial condition $Z_0 = z_0$, $\sup_{t\ge 0} E\|Z_t\| \le \max\{\|z_0\|, c\}/(1-\beta)$.*

Proof. Part (i) follows from (17), (18). Part (ii) is a consequence of (i); alternatively, it can be derived from the expression of Z_t in (19): with $\tilde{c} = \max\{\|z_0\|, c\}$,

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

The function V in Lemma 3.1(i) is a stochastic Lyapunov function for the Markov process $\{(i_t, Z_t)\}$ and has powerful implications on the behavior of the process (see [23, 21]). For most of our analysis, however, property (ii) will be sufficient. 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* (18) *with initial conditions $Z_0 = \bar{z}$ and $\hat{Z}_0 = \hat{z}$, respectively, and for the same random variables $\{i_t\}$. Then $Z_t - \hat{Z}_t \overset{a.s.}{\to} 0$.*

Proof. From (18) and, equivalently, (19), we have $Z_t - \hat{Z}_t = \beta^t L_0^t \Delta$, where $\Delta = \bar{z} - \hat{z}$. The sequence of nonnegative scalar random variables $X_t = \beta^t L_0^t, t \ge 0$, satisfies the recursion $X_t = \beta L_{t-1}^t X_{t-1}$ with $X_0 = 1$, and by (17)

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

where $\mathcal{F}_{t-1}$ is the σ-field generated by i_ℓ, $\ell \le t-1$. Hence $\{(X_t, \mathcal{F}_t)\}$ is a nonnegative supermartingale with $EX_0 = 1 < \infty$, and by a martingale convergence theorem (see, e.g., Breiman [13, Theorem 5.14]), $X_t \overset{a.s.}{\to} X$, a nonnegative random variable with

$EX \le \liminf_{t\to\infty} EX_t$. Since $EX_t = \beta^t \to 0$ as $t \to \infty$, $X = 0$ almost surely. Therefore, $X_t \overset{a.s.}{\to} 0$ and $Z_t - \hat{Z}_t \overset{a.s.}{\to} 0$. □

We now demonstrate by construction that in seemingly fairly common situations, Z_t is almost surely unbounded. This suggests that in the case of a general value of λ, it would be unrealistic to assume the boundedness of $\{G_t\}$ by assuming the boundedness of $\{Z_t\}$. Since boundedness of the iterates is often the first step in o.d.e.-based convergence proofs, this result motivates us to use alternative arguments to prove the almost sure convergence of $\{G_t\}$, and, in particular, it leads us to consider $\{(i_t, Z_t)\}$ as a weak Feller Markov chain (section 3.3).

Our construction of unbounded $\{Z_t\}$ is based on a consequence of the extended Borel–Cantelli lemma [13, Problem 5.9, p. 97], given below. In the lemma, the abbreviation "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 $\{X_t, t \ge 0\}$ and Borel-measurable subsets A, B of S, if for all t*

$$\mathbf{P}(\exists \ell, \ell > t, X_\ell \in B \mid X_t, X_{t-1}, \dots, X_0) \ge \delta > 0 \quad \textit{on} \quad \{X_t \in A\} \quad \textit{a.s.},$$

then

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

Our construction is as follows. Denote by $Z_{t,j}$ and $\phi_j(i)$ the jth elements of the vectors Z_t and $\phi(i)$, respectively. Consider a cycle formed by $m \ge 1$ states, $\{\bar{i}_1, \bar{i}_2, \dots, \bar{i}_m, \bar{i}_1\}$, with the following three properties:

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

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

(c) for some $\bar{j}$, the $\bar{j}$th elements of $\phi(\bar{i}_1), \dots, \phi(\bar{i}_m)$ have the same sign and their sum is nonzero:

$$\text{either} \quad \phi_{\bar{j}}(\bar{i}_k) \ge 0 \ \ \forall\, k = 1, \dots, m, \quad \text{with} \ \ \phi_{\bar{j}}(\bar{i}_k) > 0 \text{ for some } k; \tag{20}$$

$$\text{or} \quad \phi_{\bar{j}}(\bar{i}_k) \le 0 \ \ \forall\, k = 1, \dots, m, \quad \text{with} \ \ \phi_{\bar{j}}(\bar{i}_k) < 0 \text{ for some } k. \tag{21}$$

The next proposition shows that if such a cycle exists, then $\{Z_t\}$ is unbounded with probability 1 in almost all natural problems. A nonrestrictive technical condition involved in the proposition will be discussed after the proof.

PROPOSITION 3.1. *Suppose the Markov chain $\{i_t\}$ is irreducible and there exists a cycle of states $\{\bar{i}_1, \bar{i}_2, \dots, \bar{i}_m, \bar{i}_1\}$ with properties* (a)–(c) *above. Let $\bar{j}$ be as in* (c). *Then, the cycle defines a constant ν, which is negative (respectively, positive) if* (20) *(respectively,* (21)*) holds in* (c)*, and if for some neighborhood $\mathcal{O}(\nu)$ of ν, $\mathbf{P}(i_t = \bar{i}_1, Z_{t,\bar{j}} \notin \mathcal{O}(\nu)$ i.o.$) = 1$, then $\mathbf{P}(\sup_{t\ge 0} \|Z_t\| = \infty) = 1$.*

Proof. Denote by $\mathcal{C}$ the set of states $\{\bar{i}_1, \bar{i}_2, \dots, \bar{i}_m\}$ 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 (20).

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

$$z_{t+m,\bar{j}} = \beta^m l_0^m \cdot z_{t,\bar{j}} + \epsilon, \tag{22}$$

where

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

By properties (b) and (c) with (20), we have $\beta^m l_0^m > 1$ and $\epsilon > 0$. Consider the sequence $\{y_\ell\}$ defined by the recursion

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

The iterate y_ℓ corresponds to the value $z_{t+\ell m,\bar{j}}$ (of the $\bar{j}$th component of $Z_{t+\ell m}$) if during the time interval $[t, t+\ell m]$ the chain $\{i_t\}$ would repeat the cycle ℓ times (cf. (22)). Since $\zeta > 1$ and $\epsilon > 0$, simple calculation shows that unless $y_\ell = -\epsilon/(\zeta - 1)$ for all $\ell \ge 0$, $|y_\ell| \to \infty$ 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 any K_1, K_2 with $\eta \le K_1 \le K_2$. Let ℓ be such that $|y_\ell| \ge K_2$ 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

$$\{i_t = \bar{i}_1,\ Z_{t,\bar{j}} \notin (\nu - \eta, \nu + \eta),\ \|Z_t\| \le K_1 \text{ i.o.}\} \subset \{\|Z_t\| \ge K_2 \text{ i.o.}\} \ \text{ a.s.} \tag{23}$$

We now prove $\mathbf{P}\big(\sup_{t\ge 0} \|Z_t\| < \infty\big) = 0$ if there exists a neighborhood $\mathcal{O}(\nu)$ of ν such that $\mathbf{P}(i_t = \bar{i}_1, Z_{t,\bar{j}} \notin \mathcal{O}(\nu) \text{ i.o.}) = 1$. Let us assume $\mathbf{P}\big(\sup_{t\ge 0} \|Z_t\| < \infty\big) \ge \delta > 0$ to derive a contradiction. Define

$$K_1 = \inf_K \Big\{ K \ge 1 \,\Big|\, P\Big(\sup_{t\ge 0} \|Z_t\| \le K \Big) \ge \delta/2 \Big\}, \qquad \mathcal{E} = \big\{ \sup_{t\ge 0} \|Z_t\| \le K_1 \big\}. \tag{24}$$

Then $K_1 < \infty$ and $\mathbf{P}(\mathcal{E}) \ge \delta/2$. Let $\eta \in (0,1)$ be such that $(\nu - \eta, \nu + \eta) \subset \mathcal{O}(\nu)$. Since $\mathbf{P}(i_t = \bar{i}_1, Z_{t,\bar{j}} \notin \mathcal{O}(\nu) \text{ i.o.}) = 1$, we have $\mathbf{P}\big(i_t = \bar{i}_1,\ Z_{t,\bar{j}} \notin (\nu-\eta, \nu+\eta) \text{ i.o.}\big) = 1$. By the definition of the event $\mathcal{E}$, this implies

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

It then follows from (23) that for any $K_2 > K_1$,

$$\mathcal{E} \subset \big\{ \sup_{t\ge 0} \|Z_t\| \ge K_2 \big\} \ \text{ a.s.}$$

Since $\mathbf{P}(\mathcal{E}) \ge \delta/2$, this contradicts the definition of $\mathcal{E}$ in (24). Therefore, we must have $\mathbf{P}\big(\sup_{t\ge 0} \|Z_t\| < \infty\big) = 0$. This completes the proof. ∎

We note that the extra technical condition $\mathbf{P}(i_t = \bar{i}_1, Z_{t,\bar{j}} \notin \mathcal{O}(\nu) \text{ i.o.}) = 1$ in Proposition 3.1 is not restrictive. The opposite case—that on a set with nonnegligible 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. 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 of states with the properties above.[3] Moreover, the sign constraints in property (c) are introduced for the convenience of constructing such examples; they are not necessary for the conclusion of the proposition to hold, as the proof shows.

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 3.4). Note that boundedness of $\{Z_t\}$ is not a necessary condition for the almost sure convergence of $\{G_t\}$. What is necessary is $\gamma_t Z_t \overset{a.s.}{\to} 0$ if the function ψ takes nonzero values; i.e., $\{\lim_{t\to\infty} G_t \text{ exists}\} \subset \{\lim_{t\to\infty} \gamma_t Z_t = 0\}$. (This can be seen from (11) and the fact that γ_t diminishes.) 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 that 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 [13, Theorem 7.43]), that under Assumptions 2.1 and 2.2, for each initial condition (Z_0, G_0),

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

See [36, Prop. 3.1] for details.

Remark 3.2. As we showed, the iterates $\{Z_t\}$ can have different properties in off-policy and on-policy learning. Several earlier works on LSTD or similar TD algorithms have used boundedness properties of $\{Z_t\}$ in their analyses. In the on-policy case, the bounded variance property of $\{Z_t\}$ has been relied upon by the convergence proofs for TD(λ) (Tsitsiklis and Van Roy [34]) and LSTD(λ) (Nedić and Bertsekas [24]). The analyses in [24] and [8] (for off-policy LSTD under an additional condition) also use the boundedness of $\{Z_t\}$, and so does the analysis in [25] for an off-policy TD algorithm, which calculates Z_t only for state trajectories of a predetermined finite length. This is the reason that for the analysis of the off-policy LSTD(λ) algorithm with a general value of λ, we do not follow the approaches in these works.

3.2. Convergence in mean. We show now that G_t converges in mean to G^*. This implies in particular that G_t converges in probability to G^*, and hence that the LSTD(λ) solution Φr_t converges in probability to the solution Φr^* of the projected Bellman equation (2), when the latter exists and is unique. We state the result in a slightly more general context involving a Lipschitz continuous function $h(z, i, j)$, which

[3] Here is one such example. Let $\beta = 0.98$,

$$Q = \begin{bmatrix} 0.2 & 0.8 \\ 0.5 & 0.5 \end{bmatrix}, \qquad P = \begin{bmatrix} 0.45 & 0.55 \\ 0.6 & 0.4 \end{bmatrix}, \qquad \Rightarrow \qquad \begin{bmatrix} \frac{q_{ij}}{p_{ij}} \end{bmatrix} = \begin{bmatrix} 0.44 & 1.45 \\ 0.83 & 1.25 \end{bmatrix}.$$

Let $\Phi' = [\phi(1)\ \phi(2)] = [2\ 1]$ (thus Z_t is one-dimensional). There are several simple cycles of states satisfying the conditions of Proposition 3.1. For example, $\{2, 2\}$ is such a cycle with $\beta \frac{q_{22}}{p_{22}} = 1.225$, $\{1, 2, 1\}$ is another one with $\beta^2 \frac{q_{12}}{p_{12}} \cdot \frac{q_{21}}{p_{21}} = 1.164$, and $\{1, 2, 2, 1\}$ is yet another with $\beta^3 \frac{q_{12}}{p_{12}} \cdot \frac{q_{22}}{p_{22}} \cdot \frac{q_{21}}{p_{21}} = 1.426$. So $\{Z_t\}$ is almost surely unbounded by Proposition 3.1, which agrees with what we observed empirically in simulations. As can be verified, this is also an example in which the variance of Z_t increases to infinity as t increases.

includes as a special case the function $z\psi(i,j)'$ that defines $\{G_t\}$. This is 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.,*

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

Let $\{G_t^h\}$ be a sequence defined by the recursion

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

Then under Assumptions 2.1 *and* 2.2*, there exists a constant vector $G^{h,*}$ (independent of the stepsizes) such that for each initial condition of (Z_0, G_0^h),*

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

Proof. For notational simplicity, we suppress the superscript h in the proof. To prove the convergence of $\{G_t\}$ in mean, we introduce, for each positive integer K, another process $\{(\widetilde{Z}_{t,K}, \widetilde{G}_{t,K})\}$ and apply a law of large numbers for a finite space irreducible Markov chain to $\{\widetilde{G}_{t,K}\}$. We then relate the processes $\{(\widetilde{Z}_{t,K}, \widetilde{G}_{t,K})\}$ to $\{(Z_t, G_t)\}$.

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

$$\widetilde{Z}_{t,K} = \phi(i_t) + \beta L_{t-1}^t \phi(i_{t-1}) + \cdots + \beta^K L_{t-K}^t \phi(i_{t-K}), \quad t > K, \tag{25}$$

$$\widetilde{G}_{t,K} = (1-\gamma_t)\widetilde{G}_{t-1,K} + \gamma_t h\big(\widetilde{Z}_{t,K}, i_t, i_{t+1}\big), \quad t \ge 1. \tag{26}$$

We have, for $t \le K$, $\widetilde{G}_{t,K} = G_t$ because $\widetilde{Z}_{t,K}$ and Z_t coincide. By construction $\{\widetilde{Z}_{t,K}\}$ and $\{\widetilde{G}_{t,K}\}$ lie in some bounded sets depending on K and the initial condition. This is because $\max_i \|\phi(i)\|$ and $L_\ell^{\ell+\tau}, 0 \le \tau \le K, \ell \ge 0$, can be bounded by some (deterministic) constant, so $\sup_{t\ge 0}\|\widetilde{Z}_{t,K}\| \le c_K$ for some constant c_K depending on K and Z_0. Consequently, by the Lipschitz property of h and the assumption $\gamma_t \in (0,1]$ (Assumption 2.2), $\{h\big(\widetilde{Z}_{t,K}, i_t, i_{t+1}\big)\}$ and $\{\widetilde{G}_{t,K}\}$ are also bounded.

The sequence $\{\widetilde{G}_{t,K}\}$ converges almost surely to a constant vector G_K^* independent of the initial condition. This is because, for $t > K$, $h(\widetilde{Z}_{t,K}, i_t, i_{t+1})$ can be viewed as a function of the $K+2$ consecutive states $X_t = (i_{t-K}, i_{t-K+1}, \ldots, i_{t+1})$, and 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 [10, 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 $\{i_t\}$,

$$\widetilde{G}_{t,K} \overset{a.s.}{\to} G_K^*, \qquad \text{where } G_K^* = E_0\big[h(\widetilde{Z}_{k,K}, i_k, i_{k+1})\big] \quad \forall k > K. \tag{27}$$

Clearly, G_K^* does not depend on the values of (Z_0, G_0) and the stepsizes. Since $\sup_{t\ge 0}\|\widetilde{G}_{t,K}\| \le c_K$ for some constant c_K, we also have by the Lebesgue bounded convergence theorem

$$\lim_{t\to\infty} E\big\|\widetilde{G}_{t,K} - G_K^*\big\| = 0. \tag{28}$$

The sequence $\{G_K^*, K \geq 1\}$ converges to some constant vector G^*. To see this, let $K_1 < K_2$, and using the definition of $\widetilde{Z}_{t,K}$ and similar to the proof for Lemma 3.1(ii), we have

$$E_0\|\widetilde{Z}_{k,K_1} - \widetilde{Z}_{k,K_2}\| \leq c\beta^{K_1} \qquad \forall k > K_2,$$

where $c = \max_i \|\phi(i)\|/(1-\beta)$. Then, using the definition of G_K^* in (27) and the Lipschitz property of h, we have, for any $k > K_2$,

$$\begin{aligned}\|G_{K_1}^* - G_{K_2}^*\| &= \big\|E_0\big[h(\widetilde{Z}_{k,K_1}, i_k, i_{k+1}) - h(\widetilde{Z}_{k,K_2}, i_k, i_{k+1})\big]\big\| \\ &\leq M_h\, E_0\big\|\widetilde{Z}_{k,K_1} - \widetilde{Z}_{k,K_2}\big\| \leq cM_h\beta^{K_1}.\end{aligned}$$

This shows that $\{G_K^*\}$ 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 K,

(29)
$$\limsup_{t\to\infty} E\,\|G_t - G^*\| \leq \limsup_{t\to\infty} E\big\|G_t - \widetilde{G}_{t,K}\big\| + \lim_{t\to\infty} E\big\|\widetilde{G}_{t,K} - G_K^*\big\| + \big\|G^* - G_K^*\big\|,$$

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

(30)
$$\|Z_t - \widetilde{Z}_{t,K}\| = 0, \quad t \leq K; \qquad E\|Z_t - \widetilde{Z}_{t,K}\| \leq c\beta^K, \quad t \geq K+1,$$

where $c = \max\{\|Z_0\|, \max_i \|\phi(i)\|\}/(1-\beta)$. By the definition of G_t and $\widetilde{G}_{t,K}$,

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

Therefore, using the triangle inequality, the Lipschitz property of h, and (30), we have

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

which implies under the stepsize condition in Assumption 2.2 that $\limsup_{t\to\infty} E\|G_t - \widetilde{G}_{t,K}\| \leq cM_h\beta^K$, and hence

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

This completes the proof. □

For the special case $h(z,i,j) = z\psi(i,j)'$ and $G_t^h = G_t$, the sequence $\{G_K^{h,*}\}$ given by (27) in the proof and its limit $G^{h,*}$ in the preceding theorem have explicit expressions:

$$G_K^{h,*} = \Phi'\Xi\Big(\sum_{m=0}^{K}\beta^m Q^m\Big)\Psi, \qquad G^{h,*} = \Phi'\Xi\Big(\sum_{m=0}^{\infty}\beta^m Q^m\Big)\Psi.$$

The limit $G^{h,*}$ is G^* given by (13), so the theorem shows that $\{G_t\}$ converges in mean to G^*.

3.3. Almost sure convergence. To study the almost sure convergence of $\{G_t\}$ to G^*, we consider the Markov chain $\{(i_t, Z_t)\}$ 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 $f \in \mathcal{C}_b(S)$, $P_S f \in \mathcal{C}_b(S)$ [23, 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[4] with desirable properties.

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

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 then follows 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 [23, 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. We need the notion of weak convergence of occupation measures.

[4] A Feller chain is not necessarily ψ-irreducible (for the latter notion, see [23]). A simple counterexample in our case is given by setting $\phi(i) = 0$ for all i.

For a Markov chain $\{X_t\}$ on S, the occupation probability measures $\mu_t, t \geq 1$, 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 $\int f\, d\mu_{x,t}$. A sequence of probability measures $\{\mu_t\}$ is said to converge weakly to a probability measure μ if, for all $f \in \mathcal{C}_b(S)$, $\int f\, d\mu_t$ converges to $\int f\, d\mu$ (see [23, Chap. D.5]).

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., the discussion preceding [21, 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 [21, 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 $\{\hat{Z}_t\}$ defined by (18) and initialized with $Z_0 = \bar{z}$ and $\hat{Z}_0 = z$, respectively, and for the same random variables $\{i_t\}$ with $i_0 = \bar{i}$. 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,

$$\lim_{t\to\infty} \frac{1}{t}\sum_{\tau=1}^{t} \big(f(i_\tau, Z_\tau) - f(i_\tau, \hat{Z}_\tau)\big) = 0.$$

Since almost surely $\mu_{x,t} \to \pi_x$ weakly, $\lim_{t\to\infty} \frac{1}{t}\sum_{\tau=1}^{t} f(i_\tau, Z_\tau) = \int f\, d\pi_x$ almost surely. It then follows that, almost surely,

$$\lim_{t\to\infty} \frac{1}{t}\sum_{\tau=1}^{t} f(i_\tau, \hat{Z}_\tau) = \int f\, d\pi_x$$

for all bounded and uniformly continuous functions f, and hence, by [15, Prop. 11.3.3], 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 by [15, Prop. 11.3.2], there exists a bounded Lipschitz continuous 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

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

Therefore, with the initial distribution of (i_0, Z_0) 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, we have that the occupation measures μ_t satisfy that $\{\int h\, d\mu_t\}$ converges $\mathbf{P}_{\tilde{\mu}}$-almost surely to a nondegenerate random variable. On the other hand, since h is Lipschitz, applying Theorem 3.1 with $\gamma_t = 1/(t+1)$ and $G_0^h = 0$, we also have that under $\mathbf{P}_{\tilde{\mu}}$, $\{\int h\, d\mu_t\}$ converges in mean to a constant and therefore has a subsequence converging almost surely to the same constant (which is a degenerate random variable), which is 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., [14, Chap. X, Theorem 2.1]),

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

whereas by the preceding proof we have for each $x \in F$ a set with $\pi(F) = 1$, $\lim_{t\to\infty} \int f\, d\mu_{x,t} = \int f\, d\tilde{\pi}$, $\mathbf{P}_x$-almost surely. 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 3.3. In the preceding proof, we used the conclusion of Theorem 3.1 to show that π_x is the same for all $x \in F$. Alternative arguments can be used at this step for the finite space MDP case, but the preceding proof also applies readily to compact space MDP models that we will consider later. Another entirely different proof based on the theory of e-chains [23] can be found in [36]; however, it is much longer than the one given here.

Remark 3.4. The ergodicity of the chain $\{(i_t, Z_t)\}$ shown by the preceding theorem gives a clear explanation of the unboundedness of $\{Z_t\}$ that we observed in section 3.1, Proposition 3.1: If π does not concentrate its mass on a bounded set of S, then since the sequence of occupation measures converges weakly to π almost surely, $\{Z_t\}$ must be unbounded with probability 1.

Remark 3.5. The preceding theorem also implies that we can obtain a good approximation of $G^{h,*}$ by using modified bounded iterates, such as $\widehat{G}_t = (1-\gamma_t)\widehat{G}_{t-1} + \gamma_t \hat{h}(Z_t, i_t, i_{t+1})$, 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 componentwise to be within $[-K, K]$ for some sufficiently large K. (Then $\widehat{G}_t \overset{a.s.}{\to} E_\pi\big[\hat{h}(Z_0, i_0, i_1)\big]$; see also Theorem 3.3.)

Let E_π denote expectation with respect to $\mathbf{P}_\pi$. To establish the almost sure convergence of $\{G_t\}$, we need to first show 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 that, as $\bar{t} \to \infty$,

$$\frac{1}{\bar{t}} \sum_{t=1}^{\bar{t}} P_S^t(x, \cdot) \overset{\text{weakly}}{\longrightarrow} \pi \qquad \forall x \in S. \tag{31}$$

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

$$E_x[\|Z_t\|] \le c \quad \forall t \ge 0. \tag{32}$$

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. *Let Assumption* 2.1 *hold. Then 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, $\|h(Z_0,i_0,i_1)\| \le M_h\|Z_0\| + \|h(0,i_0,i_1)\|$ for some constant M_h, and, therefore, to prove the proposition, it is sufficient to show that $E_\pi[\|Z_0\|] < \infty$. To this end, consider a sequence of scalars $a_k, k \ge 0$, with

$$a_0 = 0, \quad a_1 \in (0,1], \quad a_{k+1} = a_k + 1, \quad k \ge 1. \tag{33}$$

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

$$O_k = \{z \mid a_k < \|z\| < a_{k+1}\}. \tag{34}$$

It is then sufficient to show that for any such $\{a_k\}$, $\sum_{k=0}^\infty a_{k+1} \cdot \pi\big(\mathcal{I} \times O_k\big) < \infty$.[5]

Fix any initial condition x. Using (32), we have, for all integers $K \ge 0, t \ge 0$,

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

Therefore, for all $K \ge 0, \bar{t} \ge 0$,

$$\frac{1}{\bar{t}} \sum_{t=1}^{\bar{t}} \sum_{k=0}^K a_{k+1} \cdot \mathbf{P}_x(Z_t \in O_k) = \sum_{k=0}^K a_{k+1} \cdot \left(\frac{1}{\bar{t}} \sum_{t=1}^{\bar{t}} \mathbf{P}_x(Z_t \in O_k) \right) \le c+1. \tag{35}$$

Since by construction O_k and $\mathcal{I} \times O_k$ are open sets on $\Re^d$ and S, respectively, by (31) and [23, Theorem D.5.4] we have, for all k,

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

Combining this with (35), we have, for all $K \ge 0$,

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

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

[5] This is because we can choose two sequences $\{a_k^1\}$, $\{a_k^2\}$ as in (33) with $a_1^1 = 1, a_1^2 = 1/2$, for instance, so that the corresponding open sets $O_k^1, O_k^2, k \ge 0$, given by (34) together cover the space of z except for the origin. Then

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

so we can bound $E_\pi[\|Z_0\|]$ by $\sum_{k=0}^\infty a_{k+1}^1 \cdot \pi(\mathcal{I} \times O_k^1) + \sum_{k=0}^\infty a_{k+1}^2 \cdot \pi(\mathcal{I} \times O_k^2)$.

THEOREM 3.3. *Let h and $\{G_t^h\}$ be as defined in Theorem* 3.1*, and let the stepsize in G_t^h be $\gamma_t = 1/(t+1)$. Then, under Assumption* 2.1*, for each initial condition of (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 vector in Theorem* 3.1.

Proof. For each initial condition of (Z_0, G_0^h), by Theorem 3.1, G_t^h converges in mean to $G^{h,*}$ (a constant vector independent of the initial condition and stepsizes), which 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 that G_t^h converges almost 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 Proposition 3.2, $E_\pi\|h(Z_0, i_0, i_1)\| < \infty$. Therefore, applying the strong law of large numbers (see [14, Theorem 2.1] or [23, Theorem 17.1.2]) to the stationary Markov process $\{(i_t, Z_t, i_{t+1})\}$ under $\mathbf{P}_\pi$, we have that $\frac{1}{t}\sum_{k=1}^{t} h(Z_k, i_k, i_{k+1})$ converges $\mathbf{P}_x$-almost surely for each initial $x = (\bar{i}, \bar{z})$ from a set $F \subset S$ with $\pi(F) = 1$. Hence $\{G_t\}$ converges almost surely for each $x \in F$.

For any $\hat{x} = (\bar{i}, \hat{z}) \notin F$, 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 the fact $\pi(F) = 1$ imply $\pi(\{\bar{i}\} \times \Re^d) > 0$.) Corresponding to $\hat{x}$ and $\bar{x}$, consider the two sequences of iterates, $\{(\hat{Z}_t, \hat{G}_t)\}$ and $\{(Z_t, G_t)\}$, defined by the same random variables $\{i_t\}$ with the initial conditions $i_0 = \bar{i}$, $\hat{Z}_0 = \hat{z}$, $Z_0 = \bar{z}$, and $\hat{G}_0 = G_0$. By the Lipschitz property of h,

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

Since $\hat{Z}_t - Z_t \overset{a.s.}{\to} 0$ by Lemma 3.2, we have $\hat{G}_t - G_t \overset{a.s.}{\to} 0$. Then since $\{G_t\}$ converges almost surely as we just proved, so does $\{\hat{G}_t\}$. This shows that $\{G_t\}$ converges $\mathbf{P}_x$-almost surely for each initial condition $x \in S$ and G_0, and hence $G_t \overset{a.s.}{\to} G^*$ for each initial condition of (Z_0, G_0).

Finally, we prove the expression for G^*. By the law of large numbers for stationary processes (see [14, Theorem 2.1] or [23, 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, $E_\pi\big[h(Z_0, i_0, i_1)\big] = E_\pi[G^*] = G^*$. ∎

Remark 3.6. The preceding theorem also implies the convergence $G_t^h \overset{a.s.}{\to} G^{h,*}$ for a stepsize γ_t that is of the order of $1/t$ 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 [17, Chap. 6, Theorem 1.2, and Example 1 of sec. 6.2]. As of 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 [17, 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 [36, Prop. 3.1] (cf. Remark 3.1).

4. Applications and extensions. In this section we discuss applications and extensions of the results of section 3. First, we apply these results to analyze the convergence of an off-policy TD(λ) algorithm (section 4.1). We then extend the analysis of the off-policy LSTD(λ) algorithm from finite space models to compact space models (section 4.2). Finally, we show that the convergence of a recently proposed LSTD algorithm with state-dependent λ-parameters also follows from our results (section 4.3).

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

$$r_t = r_{t-1} + \gamma_t Z_t d_t, \tag{36}$$

where Z_t is as in (18) and d_t is the so-called temporal difference term given by

$$d_t = \tfrac{q_{i_t i_{t+1}}}{p_{i_t i_{t+1}}} \cdot g(i_t, i_{t+1}) + \alpha \tfrac{q_{i_t i_{t+1}}}{p_{i_t i_{t+1}}} \cdot \phi(i_{t+1})' r_{t-1} - \phi(i_t)' r_{t-1}.$$

This algorithm is proposed in [8, sec. 5.3] in the context of approximate solutions of linear equations with TD methods. It bears similarity to the off-policy TD(λ) algorithm of Precup, Sutton, and Dasgupta [25], but the two algorithms also differ in several significant ways. (In particular, they differ in their definitions of Z_t and the projected Bellman equations they aim to solve. Also, they use the observations differently when updating Z_t's: in (36) an infinitely long trajectory of observations is used, whereas in [25] a fixed-length trajectory is used.) Convergence of the algorithm (36) has not been fully analyzed; it was considered only for the range of values of λ for which $\{Z_t\}$ is bounded and $\Pi T^{(\lambda)}$ is a contraction [8]. We now apply the results of section 3.3 and the o.d.e.-based stochastic approximation theory (Kushner and Yin [17, Chap. 6]) to analyze a constrained version of the algorithm.

Introducing the function

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

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 (36) 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 related to the boundedness of $\{r_t\}$ in the 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 H is either a hyperrectangle or a closed ball in $\Re^d$ and $\widehat{\Pi}_H$ is the Euclidean projection onto H.

Let Assumption 2.1 hold. We apply [17, Theorem 6.1.1] to analyze the convergence of the constrained algorithm (38). Since [17] 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 o.d.e." associated with the algorithm (38) is $\dot{r} = \bar{h}(r)$, where the "mean" function $\bar{h} : \Re^d \to \Re^d$ is the continuous function

$$\bar{h}(r) = \bar{C}r + \bar{b},$$

with $\bar{C}, \bar{b}$ defined as in (7). We have, for any fixed r and initial condition of Z_0,

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

by Theorem 3.3. We can bound the function $h(z, i, j; r)$ by

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

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

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

The functions ρ_1 and ρ_2 are Lipschitz continuous in z, and so by Theorem 3.3,

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

The relations (39) and (40) ensure that when γ_t is of the order of $1/t$ 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 [17, Theorem 6.1.1], is satisfied by the various terms as required in the theorem (see [17, Example 6.1, p. 171]).

For the constrained algorithm (38), another condition in [17, Theorem 6.1.1] is

$$\sup_{t\ge 0} E\|h(Z_t, i_t, i_{t+1}; r_{t-1})\| < \infty.$$

It is satisfied because in view of the fact that $\{r_t\}$ lies in the bounded constraint set H, we have $E\|h(Z_t, i_t, i_{t+1}; r_{t-1})\| \le c_1 E\|Z_t\| + c_2$ for some constants c_1, c_2, and by Lemma 3.1 we also have $\sup_{t\ge 0} E\|Z_t\| \le c$ for some constant c. Thus, applying [17, Theorem 6.1.1], we have the following result on the convergence of the constrained off-policy TD(λ) algorithm.

PROPOSITION 4.1. *Let Assumption* 2.1 *hold, and let the stepsize* γ_t *be of the order of* $1/t$ *with* $\frac{\gamma_t - \gamma_{t+1}}{\gamma_t} = O(1/t)$. *Then* $\{r_t\}$ *given by* (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 [8, 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 the boundary-reflecting term is zero at all $r \in H$ and $r_t \overset{a.s.}{\to} r^*$.

Similar to the discussion in Remark 3.6, the question of whether the conclusion of Proposition 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 (39) and (40). (See the discussion in [17, Example 6.1, p. 171].)

4.2. Extension to compact space MDP. In this subsection we extend the convergence analysis of the LSTD algorithm in section 3 from finite space models to compact space models. We focus on the case where $\mathcal{I}$ is a compact metric space, the per-stage cost function is continuous, and the Markov chains associated with the behavior and target policies are both weak Feller Markov chains on $\mathcal{I}$. The results of section 3 then extend directly. The case of more general models is a subject for future research.

Let $\mathcal{I}$ be a compact metric space and $\mathcal{B}(\mathcal{I})$ the Borel σ-field on $\mathcal{I}$. We will still use i or j to denote a point/state in $\mathcal{I}$. Let Q and P be two transition probability kernels on $\big(\mathcal{I}, \mathcal{B}(\mathcal{I})\big)$, represented as

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

where $Q(i, \cdot), P(i, \cdot)$ denote the transition probabilities for state i. As before, we let $\{i_t\}$ denote the Markov chain with transition kernel P. We will later use P_S to denote the transition probability kernel of the Markov 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 all $i \in \mathcal{I}$, $Q(i, \cdot)$ is absolutely continuous with respect to $P(i, \cdot)$. Moreover, there exists a continuous function ζ on $\mathcal{I}^2$ such that $\zeta(i, \cdot)$ is the Radon–Nikodym derivative of $Q(i, \cdot)$ with respect to $P(i, \cdot)$.

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 $\phi_1, \ldots, \phi_d$ are continuous functions on $\mathcal{I}$.

Under Assumption 4.1, the transition probability kernel Q must also have the weak Feller property.[6] So for the target policy, the expected one-stage cost function, $\bar{g}(i) = \int_{\mathcal{I}} Q(i, dy) g(i, y)$, is continuous, in view of Assumption 4.2(i). It then follows that under these assumptions, the cost function J^* of the target policy is continuous and satisfies the Bellman equation

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

as well as the λ-weighted multistep Bellman equation $J = T^{(\lambda)}(J)$, $\lambda \in [0, 1]$, defined as in (5). These Bellman equations are now functional equations. (See, e.g., Bertsekas and Shreve [6] for general space MDP theory.)

4.2.1. The approximation framework and algorithm. 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 $\langle \cdot, \cdot \rangle_\xi$ and the associated seminorm $\| \cdot \|_{2,\xi}$ given, respectively, by

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

[6] By [23, 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 $(Qf)(x) = \int \zeta(x, y) f(y) P(x, dy)$ by Assumption 4.1. Using the continuity of ζ and the weak Feller property of P, 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 on $\mathcal{I}$, Qf is also continuous. So Q has the weak Feller property.

Let $L^2(\mathcal{I},\xi)$ denote the factor space of equivalent classes for the equivalence relation $\sim$ defined by $f \sim \hat{f}$ if and only if $\|f-\hat{f}\|_{2,\xi}=0$. For $f \in \mathcal{L}^2(\mathcal{I},\xi)$, 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

(41)
$$J^\sim = \Pi T^{(\lambda)}(J), \quad J \in \mathcal{H}, \qquad \text{or, equivalently,} \qquad J \in \arg\min_{f\in\mathcal{H}} \|T^{(\lambda)}(J)-f\|^2_{2,\xi},$$

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, the one-stage cost function $\bar{g}$ and any function in $\mathcal{H}$ are bounded under Assumption 4.2, 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 Bellman equation (41) may not have a solution, which we do not discuss here, since our focus is on approximating this equation by samples.) Every $f \in \mathcal{H}$ can be represented as a linear combination of the functions $\phi_1,\dots,\phi_d$. By a direct calculation, a low-dimensional representation of (41) is

$$\bar{C}r + \bar{b} = 0, \qquad r \in \Re^d, \tag{42}$$

where

$$\bar{C} = \begin{bmatrix} \langle \phi_1 \,,\, Q^{(\lambda)}(\alpha Q - I)\phi_1 \rangle_\xi & \cdots & \langle \phi_1 \,,\, Q^{(\lambda)}(\alpha Q - I)\phi_d \rangle_\xi \\ \vdots & \ddots & \vdots \\ \langle \phi_d \,,\, Q^{(\lambda)}(\alpha Q - I)\phi_1 \rangle_\xi & \cdots & \langle \phi_d \,,\, Q^{(\lambda)}(\alpha Q - I)\phi_d \rangle_\xi \end{bmatrix}, \tag{43}$$

$$\bar{b} = \begin{bmatrix} \langle \phi_1 \,,\, Q^{(\lambda)}\bar{g} \rangle_\xi \\ \vdots \\ \langle \phi_d \,,\, Q^{(\lambda)}\bar{g} \rangle_\xi \end{bmatrix}. \tag{44}$$

In the above, $Q^{(\lambda)}$ denotes the weighted sum of m-step transition probability kernels Q^m:

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

(cf. (7)), and it is a linear operator on the space of bounded measurable functions on $\mathcal{I}$.

Let $\beta = \lambda\alpha$, and let $\phi : \mathcal{I} \to \Re^d$ be the function defined by $\phi(i) = (\phi_1(i),\dots,\phi_d(i))$, where we view $\phi(i)$ as a $d \times 1$ vector. The off-policy LSTD(λ) algorithm is similar to the one in the finite space case, except that it involves the Radon–Nikodym derivatives instead of the ratios $\frac{q_{ij}}{p_{ij}}$ (cf. (8)–(10)):

$$Z_t = \beta\,\zeta(i_{t-1},i_t)\cdot Z_{t-1} + \phi(i_t), \tag{45}$$

$$b_t = (1-\gamma_t)b_{t-1} + \gamma_t Z_t\,\zeta(i_t,i_{t+1})\cdot g(i_t,i_{t+1}), \tag{46}$$

$$C_t = (1-\gamma_t)C_{t-1} + \gamma_t Z_t\,(\alpha\zeta(i_t,i_{t+1})\cdot\phi(i_{t+1}) - \phi(i_t))'. \tag{47}$$

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 (41), (42).

As before, to analyze the convergence of LSTD(λ), 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 some $\Re^{\hat{d}}$-valued continuous function on $\mathcal{I}^2$. When ψ is chosen according to (46) or (47), $\{G_t\}$ specializes to $\{b_t\}$ or $\{C_t\}$:

$$G_t = \begin{cases} b_t & \text{if } \psi(i,j) = \zeta(i,j)\cdot g(i,j), \\ C_t & \text{if } \psi(i,j) = \alpha\zeta(i,j)\cdot\phi(j) - \phi(i). \end{cases} \tag{48}$$

Let $\psi_k, k = 1, \ldots, \hat{d}$, denote the kth component function of ψ, and, similar to the finite space case, denote by $\bar{\psi}_k$ the function defined by the following conditional expectations:

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

Then the convergence of $\{b_t\}, \{C_t\}$ to $\bar{b}, \bar{C}$ (given by (43)–(44)), respectively, in any mode, amounts to the convergence of $\{G_t\}$ to

$$G^* = \begin{bmatrix} \langle \phi_1 \,,\, Q^{(\lambda)}\bar{\psi}_1 \rangle_\xi & \cdots & \langle \phi_1 \,,\, Q^{(\lambda)}\bar{\psi}_{\hat{d}} \rangle_\xi \\ \vdots & \ddots & \vdots \\ \langle \phi_d \,,\, Q^{(\lambda)}\bar{\psi}_1 \rangle_\xi & \cdots & \langle \phi_d \,,\, Q^{(\lambda)}\bar{\psi}_{\hat{d}} \rangle_\xi \end{bmatrix}. \tag{49}$$

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, $\ell \le t$, appearing in the analysis of section 3 to be

$$L_\ell^t = \zeta(i_\ell, i_{\ell+1}) \cdot \zeta(i_{\ell+1}, i_{\ell+2}) \cdots \zeta(i_{t-1}, i_t), \tag{50}$$

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

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

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), whereas 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 $\{G_t^h\}$ be a sequence defined by the recursion*

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

with the stepsize sequence $\{\gamma_t\}$ satisfying Assumption 2.2. Then under Assumptions 4.1 and 4.2(ii), there exists a constant vector $G^{h,*}$ (independent of the stepsizes) such that for each initial condition of (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 K the process $\{(\widetilde{Z}_{t,K}, \widetilde{G}_{t,K})\}$ as defined in the proof of Theorem 3.1: $\widetilde{Z}_{t,K} = Z_t$ for $t \le K$; $\widetilde{G}_{0,K} = G_0$; and

$$\widetilde{Z}_{t,K} = \phi(i_t) + \beta L_{t-1}^t \phi(i_{t-1}) + \cdots + \beta^K L_{t-K}^t \phi(i_{t-K}), \quad t > K, \tag{51}$$

$$\widetilde{G}_{t,K} = (1-\gamma_t)\widetilde{G}_{t-1,K} + \gamma_t h\big(\widetilde{Z}_{t,K}, i_t, i_{t+1}\big), \quad t \ge 1. \tag{52}$$

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

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

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

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) $\{\widetilde{G}_{t,K}\}$ is bounded.

(iii) $\{Y_t, t > K\}$ is a Feller chain on a compact metric space (this chain does not depend on $\{G_t\}$), 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 $\{Y_t, t > K\}$ 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 [10, 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 $\{i_t\}$,

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

This is (27) 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 (49).

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 by assumption the space $\mathcal{I}$ is compact and the functions ζ and ϕ are continuous, 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, Lemma 4.1, together with the fact that $\{(i_t, Z_t)\}$ is bounded in probability (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 $\{(i_t, Z_t)\}$. This idea was used by Tsitsiklis and Van Roy [34, (5), p. 682] for analyzing the on-policy TD(λ) algorithm; here we follow the reasoning given in Meyn [22, Chap. 11.5.2], which does not require $\{Z_t\}$ to have bounded variances and is hence more 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 $Y_t = (i_t, i_{t-1}, \ldots)$. We denote by μ_Y the probability distribution of Y_t, which is a probability measure on $\big(\mathcal{I}^\infty, \mathcal{B}(\mathcal{I}^\infty)\big)$ and is the same for all t due to stationarity. For $y \in \mathcal{I}^\infty$ (the space of Y_t), we write y in terms of its components as $(y_0, y_{-1}, \ldots)$. (For example, if $y = (\bar{i}_0, \bar{i}_{-1}, \ldots)$, then $y_0 = \bar{i}_0, y_{-1} = \bar{i}_{-1}, \ldots$.)

Let E_0 denote expectation with respect to $\mathbf{P}^o$. Define functions $\bar{L}_k : \mathcal{I}^{k+1} \to \Re_+$, $k = 0, 1, \ldots$, by

$$\bar{L}_0 \equiv 1, \qquad \bar{L}_k(\bar{i}_0, \ldots, \bar{i}_k) = \zeta(\bar{i}_0, \bar{i}_1) \cdots \zeta(\bar{i}_{k-1}, \bar{i}_k), \quad k \geq 1.$$

Consider $Y_0 = (i_0, i_{-1}, \ldots)$. Since $E_0\big[\bar{L}_k(i_{-k}, \ldots, i_0) \mid i_{-k}\big] = 1$ for all k, we have

$$\sum_{k=0}^{\infty} \beta^k E_0\big[\|\bar{L}_k(i_{-k}, \ldots, i_0) \cdot \phi(i_{-k})\|\big] = \sum_{k=0}^{\infty} \beta^k E_0\big[\|\phi(i_{-k})\|\big] = \frac{1}{1-\beta} \int \|\phi\| \, d\xi < \infty,$$

which is equivalently $\sum_{k=0}^{\infty} \beta^k \int \|\bar{L}_k(y_{-k}, \ldots, y_0) \cdot \phi(y_{-k})\| \, d\mu_Y(y) < \infty$. Then, by a theorem on integration [28, Theorem 1.38, pp. 28–29], we can define an $\Re^d$-valued measurable function f on $\big(\mathcal{I}^\infty, \mathcal{B}(\mathcal{I}^\infty)\big)$ by

$$f(y) = \begin{cases} \sum_{k=0}^{\infty} \beta^k \bar{L}_k(y_{-k}, \ldots, y_0) \cdot \phi(y_{-k}) & \text{if } \ y \in A, \\ 0 & \text{otherwise,} \end{cases} \tag{53}$$

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 definition (53) converges to a vector in $\Re^d$; and f satisfies

$$\int \|f(y)\| \, d\mu_Y(y) < \infty \tag{54}$$

and

$$\int f(y) \, d\mu_Y(y) = \sum_{k=0}^{\infty} \beta^k \int \bar{L}_k(y_{-k}, \ldots, y_0) \cdot \phi(y_{-k}) \, d\mu_Y(y) = \sum_{k=0}^{\infty} \beta^k E_0\left[\phi(i_{-k})\right].$$

Consider now Y_0 and $Y_1 = (i_1, i_0, \ldots) = (i_1, Y_0)$. Let us define $Z_0^o = f(Y_0)$ and define Z_1^o by the same recursion that defines Z_1 with $Z_0 = Z_0^o$ (cf. (45)):

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

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 (53) and the fact that $\bar{L}_{k+1}(\bar{i}_{-k}, \ldots, \bar{i}_0, \bar{i}_1) = \zeta(\bar{i}_0, \bar{i}_1) \cdot \bar{L}_k(\bar{i}_{-k}, \ldots, \bar{i}_0)$ for all k, 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. Since $Z_1^o = \tilde{f}(Y_1)$ and $Z_0^o = f(Y_0)$, this shows that (Y_1, Z_1^o) and (Y_0, Z_0^o) have the same distribution, and hence (i_1, Z_1^o) and (i_0, Z_0^o) have the same distribution, which is an invariant probability measure of the chain $\{(i_t, Z_t)\}$ (with transition probability kernel P_S). Denote this measure by π. Then by (54), $E_\pi[\|Z_0\|] = E_0[\|Z_0^o\|] = \int \|f(y)\|\, d\mu_Y(y) < \infty$. □

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 Proposition 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 that 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. Hence we must have $\pi = \tilde{\pi}$. This shows that π is the unique invariant probability measure of $\{(i_t, Z_t)\}$.

Finally, consider those initial conditions $x = (\bar{i}, \bar{z})$ with $\bar{i} \notin F_{\mathcal{I}}$, so $x \notin F$. Because $\{(i_t, Z_t)\}$ is weak Feller (Lemma 4.1), has a unique invariant probability measure, and 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 [21, 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 Proposition 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 Proposition 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 Proposition* 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 of* (i_0, Z_0, G_0^h) *with* $i_0 = \bar{i} \in A$*,* $G_t^h \overset{a.s.}{\to} G^{h,*}$*, where* $G^{h,*} = E_\pi\big[h(Z_0, i_0, i_1)\big]$ *is the constant vector in Proposition* 4.2*.*

Proof. We argue exactly as in the proof of Theorem 3.3, using Proposition 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 $(i_0, Z_0) = (\bar{i}, \bar{z}) \in F$, where F is a set of full π-measure, and then for each initial condition G_0^h and $(i_0, Z_0) = (\bar{i}, \bar{z})$, where $\bar{i} \in A = \{i \mid (i,z) \in F \text{ for some } z\}$. 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. Unlike in the finite space case, Proposition 4.4 asserts the almost sure convergence of $\{G_t^h\}$ only for the subset of initial conditions with $i_0 \in A$. However, for the rest of the initial conditions, Proposition 4.3 implies that we can use modified bounded iterates to obtain a good approximation of $G^{h,*}$, as noted in Remark 3.5. Thus the conclusions we obtain in this compact space case are practically as strong as those in the finite space case.

The preceding 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 analyze an off-policy TD(λ) algorithm for the compact space model, similar to that in section 4.1.

4.3. Convergence of LSTD with state-dependent $\boldsymbol{\lambda}$-parameters. Recently Yu and Bertsekas [38] proposed the use of generally weighted multistep Bellman equations for approximate policy evaluation, and in that framework they derived an LSTD algorithm which uses, instead of a single parameter λ, state-dependent λ-parameters. We describe this algorithm below and show that its convergence follows as a consequence of the results we gave earlier.

For simplicity, we consider here only the finite space MDP model as given in section 2. We use the notation of that section, in addition to the following. For an operator on $\Re^n$, we use subscript i to denote its ith component mapping. We use $\delta(\cdot)$ to denote the indicator function.

We start with a weighted Bellman equation, which is more general than the Bellman equation $J = T^{(\lambda)}(J)$ associated with TD(λ). Let $\Lambda = \{\lambda_1, \ldots, \lambda_n\}$, where $\lambda_i \in [0,1], i \in \mathcal{I} = \{1, \ldots, n\}$, are given parameters. Define a multistep Bellman operator $T^{(\Lambda)}$ by letting its ith component be the corresponding component of the λ_i-weighted Bellman operator (cf. (5)):

$$T_i^{(\Lambda)} = T_i^{(\lambda_i)}, \qquad i \in \mathcal{I}, \tag{55}$$

or equivalently, expressed explicitly in terms of the Bellman operator T,

$$T_i^{(\Lambda)} = (1-\lambda_i) \sum_{m=0}^{\infty} \lambda_i^m T_i^{m+1} \ \text{ if } \ \lambda_i \in [0,1); \qquad T_i^{(\Lambda)}(\cdot) \equiv J^*(i) \ \text{ if } \ \lambda_i = 1.$$

The cost vector J^* of the target policy is the unique fixed point of $T^{(\Lambda)}$: $J^* = T^{(\Lambda)}(J^*)$. We approximate J^* by solving the projected Bellman equation,

$$J = \Pi T^{(\Lambda)}(J), \tag{56}$$

where the projection Π on $\{\Phi r \mid r \in \Re^d\}$ is as defined in section 2. The approximation properties including error bounds are known when this equation has a unique solution Φr^* [37, 29]. The approximation framework contains the TD(λ) framework and offers more flexibility, since the parameters λ_i can be set for each state independently.

Let $\bar{\lambda}_1, \ldots, \bar{\lambda}_k$ be the distinct values of $\lambda_i, i \in \mathcal{I}$. The LSTD algorithm of [38] for solving (56) is similar to LSTD(λ), except that it computes k sequences

of d-dimensional vector iterates, $Z_t^{(\ell)}, \ell = 1, \ldots, k$, and uses them to define Z_t. In particular, with $(Z_0^{(1)}, \ldots, Z_0^{(k)})$ and (b_0, C_0) being the initial condition, let

$$Z_t^{(\ell)} = \bar{\lambda}_\ell \, \alpha \, \tfrac{q_{i_{t-1} i_t}}{p_{i_{t-1} i_t}} \cdot Z_{t-1}^{(\ell)} + \delta\big(\lambda_{i_t} = \bar{\lambda}_\ell\big) \cdot \phi(i_t), \qquad \ell = 1, \ldots, k, \tag{57}$$

$$Z_t = \sum_{\ell=1}^{k} Z_t^{(\ell)}, \tag{58}$$

$$b_t = (1 - \gamma_t) b_{t-1} + \gamma_t Z_t \cdot \tfrac{q_{i_t i_{t+1}}}{p_{i_t i_{t+1}}} \cdot g(i_t, i_{t+1}), \tag{59}$$

$$C_t = (1 - \gamma_t) C_{t-1} + \gamma_t Z_t \left(\alpha \tfrac{q_{i_t i_{t+1}}}{p_{i_t i_{t+1}}} \cdot \phi(i_{t+1}) - \phi(i_t) \right)'. \tag{60}$$

(The iteration formulas for b_t, C_t are the same as before.) The algorithm is efficient if the number k of distinct values of $\lambda_i, i \in \mathcal{I}$, is relatively small. A solution r_t of the equation $C_t r + b_t = 0$ gives Φr_t as an approximation of J^* at time t.

The convergence of this LSTD algorithm is an immediate consequence of the convergence theorems in section 3, as we show below. The desired "limit" of the equations, $C_t r + b_t = 0, t \geq 0$, is the low-dimensional equivalent of the projected Bellman equation (56):

$$\Phi' \Xi \left(T^{(\Lambda)}(\Phi r) - \Phi r \right) = 0. \tag{61}$$

LEMMA 4.3. *The projected Bellman equation* (61) *can equivalently be written as*

$$\bar{C} r + \bar{b} = 0, \quad \textit{with} \quad \bar{C} = \sum_{\ell=1}^{k} \bar{C}^{(\ell)}, \quad \bar{b} = \sum_{\ell=1}^{k} \bar{b}^{(\ell)},$$

where for $\ell = 1, \ldots, k$, $\bar{C}^{(\ell)}$ *is a* $d \times d$ *matrix and* $\bar{b}^{(\ell)}$ *a* $d \times 1$ *vector, given by*

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

and Φ_ℓ *is an* $n \times d$ *matrix given by*

$$\Phi_\ell' = \Big[\delta(\lambda_1 = \bar{\lambda}_\ell) \cdot \phi(1) \quad \cdots \quad \delta(\lambda_i = \bar{\lambda}_\ell) \cdot \phi(i) \quad \cdots \quad \delta(\lambda_n = \bar{\lambda}_\ell) \cdot \phi(n) \Big].$$

Proof. We have $\Phi = \sum_{\ell=1}^{k} \Phi_\ell$ because $\sum_{\ell=1}^{k} \delta(\lambda_i = \bar{\lambda}_\ell) = 1$ for every $i \in \mathcal{I}$. We also have

$$\Phi_\ell' \, \Xi \left(T^{(\Lambda)}(\Phi r) - \Phi r \right) = \Phi_\ell' \, \Xi \left(T^{(\bar{\lambda}_\ell)}(\Phi r) - \Phi r \right), \qquad \ell = 1, \ldots, k,$$

because the nonzero columns of Φ_ℓ' can only be from those with indices $\{i \in \mathcal{I} \mid \lambda_i = \bar{\lambda}_\ell\}$, and for such i, $T_i^{(\Lambda)} = T_i^{(\bar{\lambda}_\ell)}$ by definition (recall also that Ξ is diagonal). The right-hand side of the above equation is $\bar{C}^{(\ell)} r + \bar{b}^{(\ell)}$ by the definition of $T^{(\bar{\lambda}_\ell)}$. Combining these relations, we obtain that the left-hand side of (61) can be written as

$$\Phi' \, \Xi \left(T^{(\Lambda)}(\Phi r) - \Phi r \right) = \sum_{\ell=1}^{k} \Phi_\ell' \, \Xi \left(T^{(\Lambda)}(\Phi r) - \Phi r \right) = \sum_{\ell=1}^{k} \left(\bar{C}^{(\ell)} r + \bar{b}^{(\ell)} \right),$$

which proves the lemma. ∎

PROPOSITION 4.5. *Let Assumptions* 2.1 *and* 2.2 *hold, and let* $\bar{b}, \bar{C}$ *be as given in Lemma* 4.3. *Then, the sequences* $\{b_t\}, \{C_t\}$ *generated by the LSTD algorithm* (57)–(60) *converge in mean to* $\bar{b}, \bar{C}$*, respectively; and they converge also almost surely if the stepsizes are* $\gamma_t = 1/(t+1)$.

Proof. Since $Z_t = \sum_{\ell=1}^{k} Z_t^{(\ell)}$, using linearity, we can write b_t, C_t equivalently as

$$b_t = \sum_{\ell=1}^{k} b_t^{(\ell)}, \qquad C_t = \sum_{\ell=1}^{k} C_t^{(\ell)},$$

where for $\ell = 1, \ldots, k$, $b_t^{(\ell)}, C_t^{(\ell)}$ are defined by $b_0^{(\ell)} = b_0/k$, $C_t^{(\ell)} = C_0/k$, and

$$b_t^{(\ell)} = (1-\gamma_t) b_{t-1}^{(\ell)} + \gamma_t Z_t^{(\ell)} \cdot \tfrac{q_{i_t i_{t+1}}}{p_{i_t i_{t+1}}} \cdot g(i_t, i_{t+1}),$$
$$C_t^{(\ell)} = (1-\gamma_t) C_{t-1}^{(\ell)} + \gamma_t Z_t^{(\ell)} \left(\alpha \tfrac{q_{i_t i_{t+1}}}{p_{i_t i_{t+1}}} \cdot \phi(i_{t+1}) - \phi(i_t) \right)'.$$

For each $\ell = 1, \ldots, k$, by Theorem 3.1, $\{b_t^{(\ell)}\}$, $\{C_t^{(\ell)}\}$ converge in mean to $\bar{b}^{(\ell)}, \bar{C}^{(\ell)}$, respectively, where $\bar{b}^{(\ell)}, \bar{C}^{(\ell)}$ are as given in Lemma 4.3. When $\gamma_t = 1/(t+1)$, we also have $b_t^{(\ell)} \overset{a.s.}{\to} \bar{b}^{(\ell)}, C_t^{(\ell)} \overset{a.s.}{\to} \bar{C}^{(\ell)}$ by Theorem 3.3. Since $\bar{b} = \sum_{\ell=1}^{k} \bar{b}^{(\ell)}$ and $\bar{C} = \sum_{\ell=1}^{k} \bar{C}^{(\ell)}$ (Lemma 4.3), the proposition follows. ∎

Remark 4.2. The preceding proposition holds for all $\lambda_i \in [0,1], i \in \mathcal{I}$. Similar to [8], if we restrict the λ_i parameters to those with $\alpha\big(\max_i \lambda_i\big) \cdot \max_{(i,j)} \frac{q_{ij}}{p_{ij}} < 1$, then the sequences $\{Z_t^{(\ell)}\}$, $\ell = 1, \ldots k$, are all bounded, and from the ergodicity of the weak Feller chains $\big\{\big(i_t, Z_t^{(\ell)}\big)\big\}$ (Theorem 3.2) and stochastic approximation theory [10, Chap. 6, Lemma 6, Theorem 7, and Cor. 8], it then follows that $b_t \overset{a.s.}{\to} \bar{b}, C_t \overset{a.s.}{\to} \bar{C}$ for all stepsizes γ_t satisfying Assumption 2.2.

Finally, we mention that Sutton [31] and Sutton and Barto [32, Chap. 7.10] discussed another elegant way of using state-dependent λ in the TD algorithm, which corresponds to solving a (generalized) Bellman equation different from the one considered above (see [31] for details). (See Maei and Sutton [20] for a related, two-time-scale gradient-based TD algorithm with function approximation.) Their idea of varying λ can be implemented in LSTD by simply replacing λ with λ_{i_t} in the LSTD iterate (8). The resulting (off-policy) LSTD algorithm solves a projected Bellman equation, which is different from the projected equations discussed in this section and section 2, and its convergence for all $\lambda_i \in [0,1]$ also follows directly from our analyses and convergence results given in section 3.

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 the reader to the discussion at the end of [36]. 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.

As discussed in [8], one can apply TD-type simulation-based methods to approximately solving a linear fixed point equation $x = Ax + b$, where $A = [a_{ij}]$ is an $n \times n$ matrix and b an n-dimensional vector. Compared with policy evaluation in MDP, the main difference is that the substochastic matrix αQ in the Bellman equation (4) is now replaced by an arbitrary matrix A. Then, the TD(λ) approximation framework

and the LSTD(λ) algorithm can be applied for $\lambda \in [0,1]$ such that $\lambda \sum_{j=1}^{n} |a_{ij}| < 1$ for all i or, equivalently, $\lambda|A|$ is a strictly substochastic matrix, where $|A|$ denotes the matrix with elements $|a_{ij}|$. For such values of λ, analogous to the multistep Bellman equation, we can define the λ-weighted multistep fixed point mapping $T^{(\lambda)}$ with $T(x) = Ax + b$ and find an approximate solution of $x = Ax + b$ 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 [8].

Similarly, the analysis given in section 3 extends directly to this context, assuming that P is irreducible and $|A| \prec P$, in addition to $\lambda|A|$ being strictly substochastic. We need only 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 in place of (17), we use the property that

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

for some constant ν. A slightly more general case, where $\lambda \sum_j |a_{ij}| \le 1$ for all i with strict inequality for some i, may be analyzed using a similar approach.

There are several problems deserving further study. One is the convergence of the unconstrained version of the on-line off-policy TD(λ) algorithm [8] for a general value of λ. (In the case $\lambda = 0$, there are several convergent gradient-based off-policy TD variants; see Sutton et al. [33] 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 and incremental policy improvement is carried out at a slower time-scale. One may also try to extend the analysis in this paper to MDP models with a noncompact state-action space and unbounded costs. Finally, while we have focused on the asymptotic convergence of the off-policy LSTD algorithm, its finite-sample properties, such as those considered by Antos, Szepesvári, and Munos [2] and Lazaric, Ghavamzadeh, and Munos [18, 19], and its convergence rate and large deviations properties are also worth studying.

Acknowledgments. I thank Prof. Dimitri Bertsekas, Dr. Dario Gasbarra, and Prof. George Yin for helpful suggestions and discussions. I also thank Prof. Sean Meyn for pointing me to the material on LSTD in his book [22], Prof. Richard Sutton for mentioning his earlier works on TD models, and the anonymous reviewers for their feedback, which helped to improve the presentation of the paper.

REFERENCES

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

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

[3] D. P. Bertsekas, *Dynamic Programming and Optimal Control, Vol.* I, 3rd ed., Athena Scientific, Belmont, MA, 2005.

[4] D. P. Bertsekas, *Dynamic Programming and Optimal Control, Vol.* II, 3rd ed., Athena Scientific, Belmont, MA, 2007.

[5] D. P. Bertsekas, *Temporal difference methods for general projected equations*, IEEE Trans. Automat. Control, 56 (2011), pp. 2128–2139.

[6] D. P. Bertsekas and S. Shreve, *Stochastic Optimal Control: The Discrete Time Case*, Academic Press, New York, 1978.
[7] D. P. Bertsekas and J. N. Tsitsiklis, *Neuro-Dynamic Programming*, Athena Scientific, Belmont, MA, 1996.
[8] D. P. Bertsekas and H. Yu, *Projected equation methods for approximate solution of large linear systems*, J. Comput. Appl. Math., 227 (2009), pp. 27–50.
[9] V. S. Borkar, *Stochastic approximation with 'controlled Markov' noise*, Systems Control Lett., 55 (2006), pp. 139–145.
[10] V. S. Borkar, *Stochastic Approximation: A Dynamic Viewpoint*, Hindustan Book Agency, New Delhi, 2008.
[11] J. A. Boyan, *Least-squares temporal difference learning*, in Proceedings of the 16th International Conference on Machine Learning, Bled, Slovenia, 1999, pp. 49–56.
[12] S. J. Bradtke and A. G. Barto, *Linear least-squares algorithms for temporal difference learning*, Machine Learning, 22 (1996), pp. 33–57.
[13] L. Breiman, *Probability*, Classics Appl. Math. 7, SIAM, Philadelphia, 1992.
[14] J. L. Doob, *Stochastic Processes*, John Wiley & Sons, New York, 1953.
[15] R. M. Dudley, *Real Analysis and Probability*, 2nd ed., Cambridge University Press, New York, 2003.
[16] P. W. Glynn and D. L. Iglehart, *Importance sampling for stochastic simulations*, Management Sci., 35 (1989), pp. 1367–1392.
[17] H. J. Kushner and G. G. Yin, *Stochastic Approximation and Recursive Algorithms and Applications*, 2nd ed., Springer-Verlag, New York, 2003.
[18] A. Lazaric, M. Ghavamzadeh, and R. Munos, *Finite-sample analysis of LSTD*, in Proceedings of the 27th International Conference on Machine Learning, Haifa, Israel, 2010, pp. 615–622.
[19] A. Lazaric, M. Ghavamzadeh, and R. Munos, *Finite-sample analysis of least-squares policy iteration*, J. Mach. Learn. Res., to appear.
[20] H. R. Maei and R. S. Sutton, *GQ(λ): A general gradient algorithm for temporal-difference prediction learning with eligibility traces*, in Proceedings of the 3rd Conference on Artificial General Intelligence, Lugano, Switzerland, 2010, pp. 91–96.
[21] S. P. Meyn, *Ergodic theorems for discrete time stochastic systems using a stochastic Lyapunov function*, SIAM J. Control Optim., 27 (1989), pp. 1409–1439.
[22] S. Meyn, *Control Techniques for Complex Networks*, Cambridge University Press, Cambridge, UK, 2008.
[23] S. Meyn and R. L. Tweedie, *Markov Chains and Stochastic Stability*, 2nd ed., Cambridge University Press, Cambridge, UK, 2009.
[24] A. Nedić and D. P. Bertsekas, *Least squares policy evaluation algorithms with linear function approximation*, Discrete Event Dyn. Syst., 13 (2003), pp. 79–110.
[25] D. Precup, R. S. Sutton, and S. Dasgupta, *Off-policy temporal-difference learning with function approximation*, in Proceedings of the 18th International Conference on Machine Learning, Williamstown, MA, 2001, pp. 417–424.
[26] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*, John Wiley & Sons, New York, 1994.
[27] R. S. Randhawa and S. Juneja, *Combining importance sampling and temporal difference control variates to simulate Markov chains*, ACM Trans. Model. Comput. Simul., 14 (2004), pp. 1–30.
[28] W. Rudin, *Real and Complex Analysis*, McGraw–Hill, New York, 1966.
[29] B. Scherrer, *Should one compute the temporal difference fix point or minimize the Bellman residual? The unified oblique projection view*, in Proceedings of the 27th International Conference on Machine Learning, Haifa, Israel, 2010, pp. 959–966.
[30] R. S. Sutton, *Learning to predict by the methods of temporal differences*, Machine Learning, 3 (1988), pp. 9–44.
[31] R. S. Sutton, *TD models: Modeling the world at a mixture of time scales*, in Proceedings of the 12th International Conference on Machine Learning, Tahoe City, CA, 1995, pp. 531–539.
[32] R. S. Sutton and A. G. Barto, *Reinforcement Learning*, MIT Press, Cambridge, MA, 1998.
[33] 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*, in Proceedings of the 26th International Conference on Machine Learning, Montreal, Canada, 2009, pp. 993–1000.
[34] J. N. Tsitsiklis and B. Van Roy, *An analysis of temporal-difference learning with function approximation*, IEEE Trans. Automat. Control, 42 (1997), pp. 674–690.

[35] H. S. Yao and Z. Q. Liu, *Preconditioned temporal difference learning*, in Proceedings of the 25th International Conference on Machine Learning, Helsinki, Finland, 2008, pp. 1208–1215.

[36] H. Yu, *Convergence of Least Squares Temporal Difference Methods under General Conditions*, Technical report C-2010-1, Department of Computer Science, University of Helsinki, Helsinki, Finland, 2010.

[37] H. Yu and D. P. Bertsekas, *Error bounds for approximations from projected linear equations*, Math. Oper. Res., 35 (2010), pp. 306–329.

[38] H. Yu and D. P. Bertsekas, *Weighted Bellman Equations and Their Applications in Approximate Dynamic Programming*, LIDS technical report 2876, MIT, Cambridge, MA, 2012.