On Generalized Bellman Equations and Temporal-Difference Learning*

Huizhen Yu, Ashique Mahmood, and Richard Sutton

RLAI Lab, Department of Computing Science
University of Alberta, Edmonton, Canada

Abstract. We consider off-policy temporal-difference (TD) learning in discounted Markov decision processes, where the goal is to evaluate a policy in a model-free way by using observations of a state process generated without executing the policy. To curb the high variance issue in off-policy TD learning, we propose a new scheme of setting the λ parameters of TD, based on generalized Bellman equations. Our scheme is to set λ according to the eligibility trace iterates calculated in TD, thereby easily keeping these traces in a desired bounded range. Compared to prior works, this scheme is more direct and flexible, and allows much larger λ values for off-policy TD learning with bounded traces. Using Markov chain theory, we prove the ergodicity of the joint state-trace process under nonrestrictive conditions, and we show that associated with our scheme is a generalized Bellman equation (for the policy to be evaluated) that depends on both λ and the unique invariant probability measure of the state-trace process. These results not only lead immediately to a characterization of the convergence behavior of least-squares based implementation of our scheme, but also prepare the ground for further analysis of gradient-based implementations.

Keywords: Markov decision process · policy evaluation · generalized Bellman equation · temporal differences · Markov chain · randomized stopping time

1 Introduction

We consider off-policy temporal-difference (TD) learning in discounted Markov decision processes (MDPs), where the goal is to evaluate a policy in a model-free way by using observations of a state process generated without executing the policy. Off-policy learning is an important part of the reinforcement learning methodology [25] and has been studied in the areas of operations research and machine learning (see e.g., [3,5,6,8,9,10,11,17,18,20,29]). Available algorithms, however, tend to have very high variances due to the use of importance sampling, an issue that limits their applicability in practice. The purpose of this paper is to introduce a new TD learning scheme that can help address this problem.

* This work was supported by a grant from Alberta Innovates—Technology Futures.

Our work is motivated by the recently proposed Retrace [15] and ABQ [12] algorithms, and by the Tree-Backup algorithm [18] that existed earlier. These algorithms, as explained in [12], all try to use the λ-parameters of TD to curb the high variance issue in off-policy learning. In this paper we propose a new scheme of setting the λ-parameters of TD, based on generalized Bellman equations. Our scheme is to set λ according to the eligibility trace iterates calculated in TD, thereby easily keeping those traces in a desired range. Compared to the previous works, this is a direct way to bound the traces in TD, and it is also more flexible, allowing much larger λ values for off-policy learning.

Regarding generalized Bellman equations, they are a powerful tool. In classic MDP theory they have been used in some intricate optimality analyses. Their computational use, however, seems to emerge primarily in the field of reinforcement learning (see [24], [1, Chap. 5.3] and [28] for related early and recent research). Like the earlier works [12,15,18,28,33], our work aims to employ this tool to make off-policy learning more efficient.

Our analyses of the new TD learning scheme will focus on its theoretical side. Using Markov chain theory, we prove the ergodicity of the joint state and trace process under nonrestrictive conditions (see Theorem 2.1), and we show that associated with our scheme is a generalized Bellman equation (for the policy to be evaluated) that depends on both λ and the unique invariant probability measure of the state-trace process (see Theorem 3.1). These results not only lead immediately to a characterization of the convergence behavior of least-squares based implementation of our scheme (see Cor. 2.1 and Remark 3.1), but also prepare the ground for further analysis of gradient-based implementations.

We note that due to space limit, in this paper we can only give the ideas or outlines of our proofs. The full details will be given in the longer version of this paper, which will also include numerical examples that we will not cover here.

The rest of the paper is organized as follows. In Section 2, after a brief background introduction, we present our scheme of TD learning with bounded traces, and we establish the ergodicity of the joint state-trace process. In Section 3 we derive the generalized Bellman equation associated with our scheme.

2 Off-Policy TD Learning with Bounded Traces

2.1 Preliminaries

The off-policy learning problem we consider in this paper concerns two Markov chains on a finite state space $\mathcal{S} = \{1, \ldots, N\}$. The first chain has transition matrix P, and the second P^o. Whatever physical mechanisms that induce the two chains shall be denoted by π and π^o, and referred to as the target policy and behavior policy, respectively. The second Markov chain we can observe; however, it is the system performance for the first Markov chain that we want to evaluate. Specifically, we consider a one-stage reward function $r_\pi : \mathcal{S} \to \Re$ and an associated discounted total reward criterion with state-dependent discount factors $\gamma(s) \in [0, 1], s \in \mathcal{S}$. Let Γ denote the $N \times N$ diagonal matrix with diagonal entries $\gamma(s)$. We assume that P and P^o satisfy the following conditions:

Condition 2.1 (Conditions on the target and behavior policies) (i) *P is such that the inverse $(I - P\Gamma)^{-1}$ exists, and* (ii) *P^o is such that for all $s, s' \in \mathcal{S}$, $P^o_{ss'} = 0 \Rightarrow P_{ss'} = 0$, and moreover, P^o is irreducible.*

The performance of π is defined as the expected discounted total rewards for each initial state $s \in \mathcal{S}$:

$$v_\pi(s) := \mathbb{E}^\pi_s \left[r_\pi(S_0) + \textstyle\sum_{t=1}^\infty \gamma(S_1)\,\gamma(S_2) \cdots \gamma(S_t) \cdot r_\pi(S_t) \right], \tag{2.1}$$

where the notation $\mathbb{E}^\pi_s$ means that the expectation is taken with respect to (w.r.t.) the Markov chain $\{S_t\}$ starting from $S_0 = s$ and induced by π (i.e., with transition matrix P). The function v_π is well-defined under Condition 2.1(i). It is called the *value function* of π, and by standard MDP theory (see e.g., [19]), we can write it in matrix/vector notation as

$$v_\pi = r_\pi + P\Gamma\, v_\pi, \qquad \text{i.e.,} \quad v_\pi = (I - P\Gamma)^{-1} r_\pi.$$

The first equation above is known as the Bellman equation (or dynamic programming equation) for a stationary policy.

We compute an approximation of v_π of the form $v(s) = \phi(s)^\top \theta$, $s \in \mathcal{S}$, where $\theta \in \Re^n$ is a parameter vector and $\phi(s)$ is an n-dimensional feature representation for each state s ($\phi(s), \theta$ are column vectors and $^\top$ stands for transpose). Data available for this computation are:

(i) the Markov chain $\{S_t\}$ with transition matrix P^o generated by π^o, and
(ii) rewards $R_t = r(S_t, S_{t+1})$ associated with state transitions, where the function r relates to $r_\pi(s)$ as $r_\pi(s) = \mathbb{E}^\pi_s[r(s, S_1)]$ for all $s \in \mathcal{S}$.

To find a suitable parameter θ for the approximation $\phi(s)^\top\theta$, we use the off-policy TD learning scheme. Define $\rho(s, s') = P_{ss'}/P^o_{ss'}$ (the importance sampling ratio);[1] denote $\rho_t = \rho(S_t, S_{t+1}), \gamma_t = \gamma(S_t)$. Given an initial $e_0 \in \Re^n$, for each $t \geq 1$, the eligibility trace vector $e_t \in \Re^n$ and the scalar temporal-difference term $\delta_t(v)$ for any approximate value function $v : \mathcal{S} \to \Re$ are calculated according to

$$e_t = \lambda_t\, \gamma_t\, \rho_{t-1}\, e_{t-1} + \phi(S_t), \tag{2.2}$$

$$\delta_t(v) = \rho_t \left(R_t + \gamma_{t+1} v(S_{t+1}) - v(S_t) \right). \tag{2.3}$$

Here $\lambda_t \in [0, 1], t \geq 1$, are important parameters in TD learning, the choice of which we shall elaborate on shortly.

[1] Our problem formulation entails both value function and state-action value function estimation for a stationary policy in the standard MDP context. In these applications, it is the state-action space of the MDP that corresponds to the state space $\mathcal{S}$ here; see [29, Examples 2.1, 2.2] for details. The third application is in a simulation context where P^o corresponds to a simulated system and both P^o, P are known so that the ratio $\rho(s, s')$ is available. Such simulations are useful, for example, in studying system performance under perturbations, and in speeding up the computation when assessing the impacts of events that are rare under the dynamics P.

Using e_t and δ_t, a number of algorithms can be formed to generate a sequence of parameters θ_t for approximate value functions. One such algorithm is LSTD [2,29], which obtains θ_t by solving the linear equation for $\theta \in \Re^n$,

$$\tfrac{1}{t}\sum_{k=0}^{t-1} e_k\,\delta_k(v) = 0, \quad v = \Phi\theta \tag{2.4}$$

(if it admits a solution), where Φ is a matrix with row vectors $\phi(s)^\top, s \in \mathcal{S}$. LSTD updates the equation (2.4) iteratively by incorporating one by one the observation of (S_t, S_{t+1}, R_t) at each state transition. We will discuss primarily this algorithm in the paper, as its behavior can be characterized directly using our subsequent analyses of the joint state-trace process. As mentioned earlier, our analyses will also provide bases for analyzing other gradient-based TD algorithms [9,10] using stochastic approximation theory. However, due to its complexity, this subject is better to be treated separately, not in the present paper.

2.2 Our Choice of $\boldsymbol{\lambda}$

We now come to the choices of λ_t in the trace iterates (2.2). For TD with function approximation, one often lets λ_t be a constant or a function of S_t [23,25,27]. If neither the behavior policy nor the λ_t's are further constrained, $\{e_t\}$ can have unbounded variances and is also unbounded in many natural situations (see e.g., [29, Section 3.1]), and this makes off-policy TD learning challenging.[2] If we let the behavior policy to be close enough to the target policy so that $P^o \approx P$, then variance can be reduced, but it is not a satisfactory solution, for the applicability of off-policy learning would be seriously limited.

Without restricting the behavior policy, the two recent works [12,15] (as well as the closely related early work [18]) exploit state-dependent λ's to control variance. Their choices of λ_t are such that $\lambda_t\rho_{t-1} < 1$ for all t, so that the trace iterates e_t are made bounded, which can help reduce the variance of the iterates.

Our proposal, motivated by these prior works, is to set λ_t according to e_{t-1} directly, so that we can keep e_t in a desired range straightforwardly and at the same time, allow a much larger range of values for the λ-parameters. As a simple example, if we use λ_t to scale the vector $\gamma_t\rho_{t-1}e_{t-1}$ to be within a ball with some given radius, then we keep e_t bounded always.

In the rest of this paper, we shall focus on analyzing the iteration (2.2) with a particular choice of λ_t of the kind just mentioned. We want to be more general than the preceding simple example. However, we also want to retain certain Markovian properties that are very useful for convergence analysis. This leads us to consider λ_t *being a certain function of the previous trace and past states*. More specifically, we will let λ_t be a function of the previous trace and a certain memory state that is a summary of the states observed so far, and the formulation is as follows.

Denote the memory state at time t by y_t. For simplicity, we assume that y_t can only take values from a finite set $\mathcal{M}$, and its evolution is Markovian:

[2] Asymptotic convergence is still ensured, however, for several algorithms [29,30,31], thanks partly to a powerful law of large numbers for stationary processes.

$y_t = g(y_{t-1}, S_t)$ for some given function g. The joint process $\{(S_t, y_t)\}$ is then a simple finite-state Markov chain. Each y_t is a function of $(S_0, \ldots, S_t)$ and y_0. We further require, besides the irreducibility of $\{S_t\}$ (cf. Condition 2.1(ii)), that[3]

Condition 2.2 (Evolution of memory states) *Under the behavior policy π^o, the Markov chain $\{(S_t, y_t)\}$ on $\mathcal{S} \times \mathcal{M}$ has a single recurrent class.*

Thus we let y_t and λ_t evolve as

$$y_t = g(y_{t-1}, S_t), \qquad \lambda_t = \lambda(y_t, e_{t-1}) \tag{2.5}$$

where $\lambda : \mathcal{M} \times \Re^n \to [0, 1]$. We require the function λ to satisfy two conditions.

Condition 2.3 (Conditions for λ) *For some norm $\|\cdot\|$ on $\Re^n$, the following hold for each memory state $y \in \mathcal{M}$:*

(i) *For any $e, e' \in \Re^n$, $\|\lambda(y, e)\, e - \lambda(y, e')\, e'\| \le \|e - e'\|$.*
(ii) *For some constant C_y, $\|\gamma(s')\rho(s, s') \cdot \lambda(y, e)\, e\| \le C_y$ for all possible state transitions (s, s') that can lead to the memory state y.*

In the above, the second condition is to restrict $\{e_t\}$ in a desired range (as it makes $\|e_t\| \le \max_{y \in \mathcal{M}} C_y + \max_{s \in \mathcal{S}} \|\phi(s)\|$). The first condition is to ensure that the traces e_t jointly with (S_t, y_t) form a Markov chain with nice properties (as will be seen in the next subsection).

Consider the simple scaling example mentioned earlier. In this case we can let $y_t = (S_{t-1}, S_t)$, and for each $y = (s, s')$, define $\lambda(y, \cdot)$ to scale back the vector $\gamma(s')\rho(s, s')\, e$ when it is outside the Euclidean ball with radius $C_{ss'}$: $\lambda\big(y, e\big) = 1$ if $\gamma(s')\rho(s, s')\|e\|_2 \le C_{ss'}$; and $\lambda\big(y, e\big) = \frac{C_{ss'}}{\gamma(s')\rho(s,s')\|e\|_2}$ otherwise.

2.3 Ergodicity Result

The properties of the joint state-trace process $\{(S_t, y_t, e_t)\}$ are important for understanding and characterizing the behavior of the proposed TD learning scheme. We study them in this subsection; most importantly, we shall establish the ergodicity of the state-trace process. The result will be useful in convergence analysis of several associated TD algorithms, although in this paper we discuss only the LSTD algorithm. In the next section we will also use the ergodicity result when we relate the LSTD equation (2.4) to a generalized Bellman equation for the target policy, which will then make the meaning of the LSTD solutions clear.

As a side note, one can introduce nonnegative coefficients $i(y)$ for memory states y to weight the state features (similarly to the use of "interest" weights in the ETD algorithm [26]) and update e_t according to

$$e_t = \lambda_t\, \gamma_t\, \rho_{t-1}\, e_{t-1} + i(y_t)\, \phi(S_t). \tag{2.6}$$

The results given below apply to this update rule as well.

Let us start with two basic properties of $\{(S_t, y_t, e_t)\}$ that follow directly from our choice of the λ function:

[3] These conditions are nonrestrictive. If the Markov chains have multiple recurrent classes, each recurrent class can be treated separately using the same arguments.

(i) By Condition 2.3(i), for each y, $\lambda(y,e)e$ is a continuous function of e, and thus e_t depends continuously on e_{t-1}. This, together with the finiteness of $\mathcal{S}\times\mathcal{M}$, ensures that $\{(S_t,y_t,e_t)\}$ is a weak Feller Markov chain.[4]
(ii) Then, by a property of weak Feller Markov chains [14, Theorem 12.1.2(ii)], the boundedness of $\{e_t\}$ ensured by Condition 2.3(ii) implies that $\{(S_t,y_t,e_t)\}$ has at least one invariant probability measure.

The third property, given in the lemma below, concerns the behavior of $\{e_t\}$ for different initial e_0. It is an important implication of Condition 2.3(i) (actually it is our purpose of introducing the condition 2.3(i) in the first place). Due to space limit, we omit the proof, which is similar to the proof of [29, Lemma 3.2].

Lemma 2.1 *Let $\{e_t\}$ and $\{\hat{e}_t\}$ be generated by the iteration (2.2) and (2.5), using the same trajectory of states $\{S_t\}$ and initial y_0, but with different initial e_0 and $\hat{e}_0$, respectively. Then under Conditions 2.1(i) and 2.3(i), $e_t-\hat{e}_t\overset{a.s.}{\to}0$.*

We use the preceding lemma and ergodicity properties of weak Feller Markov chains [13] to prove the ergodicity theorem given below (for lack of space, we again omit the proof). Before stating this result, we note that for $\{(S_t,y_t,e_t)\}$ starting from the initial condition $x=(s,y,e)$, the occupation probability measures $\{\mu_{x,t}\}$ are random probability measures on $\mathcal{S}\times\mathcal{M}\times\Re^n$ given by

$$\mu_{x,t}(D):=\tfrac{1}{t}\textstyle\sum_{k=0}^{t-1}\mathbb{1}\big((S_k,y_k,e_k)\in D\big)$$

for all Borel sets $D\subset\mathcal{S}\times\mathcal{M}\times\Re^n$, where $\mathbb{1}(\cdot)$ is the indicator function. We write $\mathbf{P}_x$ for the probability distribution of $\{(S_t,y_t,e_t)\}$ with initial condition x.

Theorem 2.1 *Under Conditions 2.1-2.3, $\{(S_t,y_t,e_t)\}$ is a weak Feller Markov chain and has a unique invariant probability measure ζ. For each initial condition $(S_0,y_0,e_0)=(s,y,e)=:x$, the occupation probability measures $\{\mu_{x,t}\}$ converge weakly[5] to ζ, $\mathbf{P}_x$-almost surely.*

Let $\mathbb{E}_\zeta$ denote expectation w.r.t. the stationary state-trace process $\{(S_t,y_t,e_t)\}$ with initial distribution ζ. Since the traces and hence the entire process lie in a bounded set under Condition 2.3(ii), the weak convergence of $\{\mu_{x,t}\}$ to ζ implies that the sequence of equations, $\frac{1}{t}\sum_{k=0}^{t-1}e_k\,\delta_k(v)=0$, as given in (2.4) for LSTD, has an asymptotic limit that can be expressed in terms of the stationary state-trace process as follows.

Corollary 2.1 *Let Conditions 2.1-2.3 hold. Then for each initial condition of (S_0,y_0,e_0), almost surely, the first equation in (2.4), viewed as a linear equation in v, tends to[6] the equation $\mathbb{E}_\zeta[\,e_0\,\delta_0(v)]=0$ in the limit as $t\to\infty$.*

[4] This means that for any bounded continuous function f on $\mathcal{S}\times\mathcal{M}\times\Re^n$ (endowed with the usual topology), with $X_t=(S_t,y_t,e_t)$, $\mathbb{E}\big[f(X_1)\mid X_0=x\big]$ is a continuous function of x [14, Prop. 6.1.1].
[5] This means $\int f d\mu_{x,t}\to\int f d\zeta$ as $t\to\infty$, for every bounded continuous function f.
[6] By this we mean that as linear equations in v, the random coefficients in this sequence of equations converge to the corresponding coefficients in the limiting equation.

3 Generalized Bellman Equations

In this section we continue the analysis started in Section 2.3. Our goal is to relate the linear equation $\mathbb{E}_\zeta[\, e_0\, \delta_0(v)] = 0$, the asymptotic limit of the linear equations (2.4) for LSTD as just shown by Cor. 2.1, to a generalized Bellman equation for the target policy π. Then, we can interpret solutions of (2.4) as solutions of approximate versions of that generalized Bellman equation.

To simplify notation in subsequent derivations, we shall use the following shorthand notation: For $k \le m$, denote $S_k^m = (S_k, S_{k+1}, \dots S_m)$, and denote

$$\rho_k^m = \textstyle\prod_{i=k}^m \rho_i, \qquad \lambda_k^m = \prod_{i=k}^m \lambda_i, \qquad \gamma_k^m = \prod_{i=k}^m \gamma_i, \tag{3.1}$$

whereas by convention we treat $\rho_k^m = \lambda_k^m = \gamma_k^m = 1$ if $k > m$.

3.1 Randomized Stopping Times

Consider the Markov chain $\{S_t\}$ induced by the target policy π. Let Condition 2.1(i) hold. Recall that for the value function v_π, we have

$$v_\pi(s) = \mathbb{E}_s^\pi\big[\textstyle\sum_{t=0}^\infty \gamma_1^t\, r_\pi(S_t)\big] \;\text{ (by definition), and }\; v_\pi(s) = r_\pi(s) + \mathbb{E}_s^\pi[\gamma_1 v_\pi(S_1)]$$

for each state s. The second equation is the standard one-step Bellman equation.

To write generalized Bellman equations for π, we need the notion of *randomized stopping times* for $\{S_t\}$. They generalize stopping times for $\{S_t\}$ in that whether to stop at time t depends not only on S_0^t but also on certain random outcomes. A simple example is to toss a coin at each time and stop as soon as the coin lands on heads, regardless of the history S_0^t. (The corresponding Bellman equation is the one associated with TD(λ) for a constant λ.) Of interest here is the general case where the stopping decision does depend on the entire history.

To define a randomized stopping time formally, first, the probability space of $\{S_t\}$ is enlarged to take into account whatever randomization scheme that is used to make the stopping decision. (The enlargement will be problem-dependent, as the next subsection will demonstrate.) Then, on the enlarged space, a randomized stopping time τ for $\{S_t\}$ is by definition a stopping time relative to some increasing sequence of sigma-algebras $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \cdots$, where the sequence $\{\mathcal{F}_t\}$ is such that (i) for all $t \ge 0$, $\mathcal{F}_t \supset \sigma(S_0^t)$ (the sigma-algebra generated by S_0^t), and (ii) w.r.t. $\{\mathcal{F}_t\}$, $\{S_t\}$ remains to be a Markov chain with transition probability P, i.e., $\mathrm{Prob}(S_{t+1} = s \mid \mathcal{F}_t) = P_{S_t s}$. (See [16, Chap. 3.3].)

The advantage of this abstract definition is that it allows us to write Bellman equations in general forms without worrying about the details of the enlarged space which are not important at this point. For notational simplicity, we shall still use $\mathbb{E}^\pi$ to denote expectation for the enlarged probability space and write $\mathbf{P}^\pi$ for the probability measure on that space, when there is no confusion.

If τ is a randomized stopping time for $\{S_t\}$, the strong Markov property [16, Theorem 3.3] allows us to express v_π in terms of $v_\pi(S_\tau)$ and the total discounted

rewards R^τ prior to stopping:

$$\begin{aligned} v_\pi(s) &= \mathbb{E}_s^\pi \left[\textstyle\sum_{t=0}^{\tau-1} \gamma_1^t \, r_\pi(S_t) + \sum_{t=\tau}^{\infty} \gamma_1^\tau \cdot \gamma_{\tau+1}^t \, r_\pi(S_t) \right] \\ &= \mathbb{E}_s^\pi \big[R^\tau + \gamma_1^\tau \, v_\pi(S_\tau) \big], \end{aligned} \tag{3.2}$$

where $R^\tau = \sum_{t=0}^{\tau-1} \gamma_1^t \, r_\pi(S_t)$ for $\tau \in \{0, 1, 2, \ldots\} \cup \{+\infty\}$.[7] We can also write the Bellman equation (3.2) in terms of $\{S_t\}$ only, by taking expectation over τ:

$$v_\pi(s) = \mathbb{E}_s^\pi \left[\textstyle\sum_{t=0}^{\infty} \Big(q_t^+(S_0^t) \cdot \gamma_1^t \, r_\pi(S_t) + q_t(S_0^t) \cdot \gamma_1^t \, v_\pi(S_t) \Big) \right], \tag{3.3}$$

$$\text{where} \quad q_t^+(S_0^t) = \mathbf{P}^\pi(\tau > t \mid S_0^t), \qquad q_t(S_0^t) = \mathbf{P}^\pi(\tau = t \mid S_0^t). \tag{3.4}$$

The r.h.s. of (3.2) or (3.3) defines an associated generalized Bellman operator $T : \Re^N \to \Re^N$ that has several equivalent expressions; e.g.,

$$(Tv)(s) = \mathbb{E}_s^\pi \big[R^\tau + \gamma_1^\tau \, v(S_\tau) \big] = \mathbb{E}_s^\pi \left[\textstyle\sum_{t=0}^{\infty} \Big(q_t^+(S_0^t) \cdot \gamma_1^t \, r_\pi(S_t) + q_t(S_0^t) \cdot \gamma_1^t \, v(S_t) \Big) \right],$$

for all $s \in \mathcal{S}$. Just as in the case of the one-step Bellman operator, the value function v_π is the unique fixed point of T, i.e., the unique solution of $v = Tv$.[8]

3.2 Bellman Equation for the Proposed TD Learning Scheme

With the terminology of randomized stopping times, we are now ready to write down the generalized Bellman equation associated with the TD-learning scheme proposed in Section 2.2. It corresponds to a particular randomized stopping time. We shall first describe this random time, from which a generalized Bellman equation follows as seen in the preceding subsection. That this is indeed the Bellman equation for our TD learning scheme will then be proved.

Consider the Markov chain $\{S_t\}$ *under the target policy* π. We define a randomized stopping time τ for $\{S_t\}$:

- Let $y_t, \lambda_t, e_t, t \geq 1$, evolve according to (2.5) and (2.2).
- Let the initial (S_0, y_0, e_0) be distributed according to ζ, the unique invariant probability measure in Theorem 2.1.
- At time $t \geq 1$, we stop the system with probability $1 - \lambda_t$ if it has not yet been stopped. Let τ be the time when the system stops ($\tau = \infty$ if the system never stops).

To make the dependence on the initial distribution ζ explicit, we write $\mathbf{P}_\zeta^\pi$ for the probability measure of this process.

[7] In the case $\tau = 0$, $R^0 = 0$. In the case $\tau = \infty$, $R^\infty = \sum_{t=0}^{\infty} \gamma_1^t \, r_\pi(S_t)$, and the second term $\gamma_1^\tau \, v_\pi(S_\tau)$ in (3.2) is 0 because $\gamma_1^\infty := \prod_{k=1}^{\infty} \gamma_k = 0$ a.s. under Condition 2.1(i).

[8] To see this, note that the matrix involved in the affine operator T is substochastic and dominated by the substochastic matrix $P\Gamma$ (the matrix in the one-step Bellman operator), whereas in view of Condition 2.1(i), $P\Gamma$ is a linear contraction w.r.t. a weighted sup-norm on $\Re^N$ by nonnegative matrix theory (see also [1, Prop. 2.2]).

Note that by definition λ_t and λ_1^t are functions of the initial (y_0, e_0) and states S_0^t. From how the random time τ is defined, we have for all $t \geq 1$,

$$\mathbf{P}_\zeta^\pi(\tau > t \mid S_0^t, y_0, e_0) = \lambda_1^t =: h_t^+(y_0, e_0, S_0^t), \tag{3.5}$$

$$\mathbf{P}_\zeta^\pi(\tau = t \mid S_0^t, y_0, e_0) = \lambda_1^{t-1}(1-\lambda_t) =: h_t(y_0, e_0, S_0^t), \tag{3.6}$$

and hence

$$q_t^+(S_0^t) := \mathbf{P}_\zeta^\pi(\tau > t \mid S_0^t) = \int h_t^+(y, e, S_0^t)\, \zeta\big(d(y,e) \mid S_0\big), \tag{3.7}$$

$$q_t(S_0^t) := \mathbf{P}_\zeta^\pi(\tau = t \mid S_0^t) = \int h_t(y, e, S_0^t)\, \zeta\big(d(y,e) \mid S_0\big), \tag{3.8}$$

where $\zeta(d(y,e) \mid s)$ is the conditional distribution of (y_0, e_0) given $S_0 = s$, w.r.t. the initial distribution ζ. As before, we can write the generalized Bellman operator T associated with τ in several equivalent forms. Let $\mathbb{E}_\zeta^\pi$ denote expectation under $\mathbf{P}_\zeta^\pi$. Based on (3.2) and (3.5)-(3.6), it is easy to derive that[9] for all $v : \mathcal{S} \to \Re, s \in \mathcal{S}$,

$$(Tv)(s) = \mathbb{E}_\zeta^\pi \left[\textstyle\sum_{t=0}^\infty \lambda_1^t \gamma_1^t\, r_\pi(S_t) + \sum_{t=1}^\infty \lambda_1^{t-1}(1-\lambda_t)\gamma_1^t\, v(S_t) \mid S_0 = s\right]. \tag{3.9}$$

Alternatively, by integrating over (y_0, e_0) and using (3.7)-(3.8), we can write

$$(Tv)(s) = \mathbb{E}_\zeta^\pi \left[\textstyle\sum_{t=0}^\infty \Big(q_t^+(S_0^t)\cdot \gamma_1^t\, r_\pi(S_t) + q_t(S_0^t)\cdot \gamma_1^t\, v(S_t)\Big) \,\Big|\, S_0 = s\right], \tag{3.10}$$

for all $v : \mathcal{S} \to \Re, s \in \mathcal{S}$, where in the case $t = 0$, $q_0^+(\cdot) \equiv 1 = \mathbf{P}_\zeta^\pi(\tau > 0 \mid S_0)$ and $q_0(\cdot) \equiv 0 = \mathbf{P}_\zeta^\pi(\tau = 0 \mid S_0)$ (since $\tau > 0$ by construction).

Comparing the two expressions of T, we remark that the expression (3.9) reflects the role of the λ_t's in determining the stopping time, whereas the expression (3.10), which has eliminated the auxiliary memory states y_t, shows more clearly the dependence of the stopping time on the entire history S_0^t. It can also be seen from the initial distribution ζ that the behavior policy asserts a significant role in determining the Bellman operator T for the target policy. This is in contrast with off-policy TD learning that uses a constant λ, where the behavior policy affects only how one approximates the Bellman equation underlying TD, not the Bellman equation itself.

We now proceed to show how the Bellman equation $v = Tv$ given above relates to the off-policy TD learning scheme in Section 2.2. Some notation is needed. Denote by $\zeta_\mathcal{S}$ the marginal of ζ on $\mathcal{S}$. Note that $\zeta_\mathcal{S}$ coincides with the invariant probability measure of the Markov chain $\{S_t\}$ *induced by the behavior policy*. For two functions v_1, v_2 on $\mathcal{S}$, we write $v_1 \perp_{\zeta_\mathcal{S}} v_2$ if $\sum_{s\in\mathcal{S}} \zeta_\mathcal{S}(s)\, v_1(s)\, v_2(s) = 0$. If $\mathcal{L}$ is a linear subspace of functions on $\mathcal{S}$ and $v \perp_{\zeta_\mathcal{S}} v'$ for all $v' \in \mathcal{L}$, we write $v \perp_{\zeta_\mathcal{S}} \mathcal{L}$. Recall that ϕ is a function that maps each state s to an n-dimensional feature vector. Denote by $\mathcal{L}_\phi$ the subspace spanned by the n component functions of ϕ, which is the space of approximate value functions for our TD learning

[9] Rewrite (3.2) as $v_\pi(s) = \mathbb{E}_s^\pi \left[\sum_{t=0}^\infty \mathbb{1}(\tau > t)\, \gamma_1^t\, r_\pi(S_t) + \sum_{t=0}^\infty \mathbb{1}(\tau = t)\, \gamma_1^t\, v_\pi(S_t)\right]$ and for the tth terms in the r.h.s., take expectation over τ conditioned on (S_0^t, y_0, e_0).

scheme. Recall also that $\mathbb{E}_\zeta$ denotes expectation w.r.t. the *stationary* state-trace process $\{(S_t, y_t, e_t)\}$ under the behavior policy (cf. Theorem 2.1).

Theorem 3.1 *Let Conditions 2.1-2.3 hold. Then as a linear equation in v, $\mathbb{E}_\zeta\big[e_0\,\delta_0(v)\big] = 0$ is equivalently $Tv - v \perp_{\zeta_\mathcal{S}} \mathcal{L}_\phi$, where T is the generalized Bellman operator for π given in (3.9) or (3.10).*

Remark 3.1 (On LSTD) Note that $Tv - v \perp_{\zeta_\mathcal{S}} \mathcal{L}_\phi, v \in \mathcal{L}_\phi$ is a projected version of the generalized Bellman equation $Tv - v = 0$ (projecting the l.h.s. onto the approximation subspace $\mathcal{L}_\phi$ w.r.t. the $\zeta_\mathcal{S}$-weighted Euclidean norm). Theorem 3.1 and Cor. 2.1 together show that this is what LSTD solves in the limit. If this projected Bellman equation admits a unique solution $\bar{v}$, then the approximation error $\bar{v} - v_\pi$ can be characterized as in [22,32].

Proof (outline). We divide the proof into three parts. The first part is more subtle than the other two, which are mostly calculations. Due to space limit, we can only outline the proof here, leaving out the details of some arguments.

(i) We extend the stationary state-trace process to $t = -1, -2, \ldots$ and work with a double-ended stationary process $\{(S_t, y_t, e_t)\}_{-\infty<t<\infty}$ (such a process exists by Kolmogorov's theorem [4, Theorem 12.1.2]). We keep using the notation P_ζ and $\mathbb{E}_\zeta$ for this double-ended stationary Markov chain. Then, by unfolding the iteration (2.2) for e_t backwards in time, we show that[10]

$$e_0 = \phi(S_0) + \textstyle\sum_{t=1}^\infty \lambda_{1-t}^0 \gamma_{1-t}^0 \rho_{-t}^{-1}\, \phi(S_{-t}) \quad P_\zeta\text{-a.s.}, \tag{3.11}$$

or with $\lambda_1^0 = \rho_0^{-1} = 1$ by convention, we can write $e_0 = \sum_{t=0}^\infty \lambda_{1-t}^0 \gamma_{1-t}^0 \rho_{-t}^{-1}\, \phi(S_{-t})$ P_ζ-a.s. The proof of (3.11) uses the stationarity of the process, Condition 2.1(i) and a theorem on integration [21, Theorem 1.38] among others.

(ii) Using the expression (3.11) of e_0, we calculate $\mathbb{E}_\zeta\big[e_0 \cdot \rho_0 f(S_0^1)\big]$ for any bounded measurable function f on $\mathcal{S} \times \mathcal{S}$. In particular, we first obtain

$$\mathbb{E}_\zeta\big[e_0 \cdot \rho_0 f(S_0^1)\big] = \textstyle\sum_{t=0}^\infty \mathbb{E}_\zeta\Big[\phi(S_0) \cdot \mathbb{E}_\zeta\left[\lambda_1^t \gamma_1^t \rho_0^t\, f(S_t^{t+1}) \mid S_0\right]\Big] \tag{3.12}$$

by using (3.11) and the stationarity of the state-trace process. Next we relate the expectations in the summation in (3.12) to expectations w.r.t. the process with probability measure $\mathbf{P}_\zeta^\pi$, which we recall is induced by the target policy π and introduced at the beginning of this subsection. Let $\tilde{\mathbb{E}}_\zeta^\pi$ denote expectation w.r.t. the marginal of $\mathbf{P}_\zeta^\pi$ on the space of $\{(S_t, y_t, e_t)\}_{t\geq 0}$. From the change of measure performed through ρ_0^t, we have

$$\mathbb{E}_\zeta\left[\lambda_1^t \gamma_1^t \rho_0^t\, f(S_t^{t+1}) \mid S_0, y_0, e_0\right] = \tilde{\mathbb{E}}_\zeta^\pi\left[\lambda_1^t \gamma_1^t\, f(S_t^{t+1}) \mid S_0, y_0, e_0\right], \quad t \geq 0. \tag{3.13}$$

Combining this with (3.12) and using the fact that ζ is the marginal distribution of (S_0, y_0, e_0) in both processes, we obtain

$$\mathbb{E}_\zeta\big[e_0 \cdot \rho_0 f(S_0^1)\big] = \textstyle\sum_{t=0}^\infty \tilde{\mathbb{E}}_\zeta^\pi\Big[\phi(S_0) \cdot \tilde{\mathbb{E}}_\zeta^\pi\left[\lambda_1^t \gamma_1^t\, f(S_t^{t+1}) \mid S_0\right]\Big]. \tag{3.14}$$

[10] Recall the shorthand notation (3.1) introduced at the beginning of Section 3.

(iii) We now use (3.14) to calculate $\mathbb{E}_\zeta\big[e_0\,\delta_0(v)\big]$ for a given function v. Recall from (2.3) $\delta_0(v) = \rho_0 \cdot \big(r(S_0^1) + \gamma_1 v(S_1) - v(S_0)\big)$, so we let $f(S_t^{t+1}) = r(S_t^{t+1}) + \gamma_{t+1} v(S_{t+1}) - v(S_t)$ in (3.14). Then a direct calculation shows that[11]

$$\mathbb{E}_\zeta\big[e_0\,\delta_0(v) \mid S_0\big] = \phi(S_0) \cdot \big\{ -v(S_0) + (Tv)(S_0)\big\}. \tag{3.15}$$

Therefore $\mathbb{E}_\zeta\big[e_0\,\delta_0(v)\big] = \sum_{s\in\mathcal{S}} \zeta_{\mathcal{S}}(s)\,\phi(s)\cdot(Tv - v)(s)$, and this shows that $\mathbb{E}_\zeta\big[e_0\,\delta_0(v)\big] = 0$ is equivalent to $Tv - v \perp_{\zeta_{\mathcal{S}}} \mathcal{L}_\phi$. □

Concluding Remark: This completes our analysis of the LSTD algorithm for the proposed TD-learning scheme. To conclude the paper, we note that the preceding results also prepare the ground for analyzing gradient-based algorithms similar to [9,10] in a future work. Specifically, like LSTD, these algorithms would aim to solve the same projected generalized Bellman equation as characterized by Theorem 3.1 (cf. Remark 3.1). Their average dynamics, which is important for analyzing their convergence using the mean ODE approach from stochastic approximation theory [7], can be studied based on the ergodicity result of Theorem 2.1, in essentially the same way as we did in Section 2.3 for the LSTD algorithm.

References

1. Bertsekas, D.P., Tsitsiklis, J.N.: Neuro-Dynamic Programming. Athena Scientific, Belmont, MA (1996)
2. Boyan, J.A.: Least-squares temporal difference learning. In: Proc. The 16th Int. Conf. Machine Learning (ICML) (1999)
3. Dann, C., Neumann, G., Peters, J.: Policy evaluation with temporal differences: A survey and comparison. Journal of Machine Learning Res. 15, 809–883 (2014)
4. Dudley, R.M.: Real Analysis and Probability. Cambridge University Press, Cambridge (2002)
5. Geist, M., Scherrer, B.: Off-policy learning with eligibility traces: A survey. Journal of Machine Learning Res. 15, 289–333 (2014)
6. Glynn, P.W., Iglehart, D.L.: Importance sampling for stochastic simulations. Management Science 35, 1367–1392 (1989)
7. Kushner, H.J., Yin, G.G.: Stochastic Approximation and Recursive Algorithms and Applications. Springer-Verlag, New York, 2nd edn. (2003)
8. Liu, B., Liu, J., Ghavamzadeh, M., Mahadevan, S., Petrik, M.: Finite-sample analysis of proximal gradient TD algorithms. In: The 31st Conf. Uncertainty in Artificial Intelligence (UAI) (2015)
9. Maei, H.R.: Gradient Temporal-Difference Learning Algorithms. Ph.D. thesis, University of Alberta (2011)
10. Mahadevan, S., Liu, B., Thomas, P., Dabney, W., Giguere, S., Jacek, N., Gemp, I., Liu, J.: Proximal reinforcement learning (2014), arXiv:1405.6757

[11] Note $\sum_{t=0}^\infty \tilde{\mathbb{E}}_\zeta^\pi\,\big[\lambda_1^t\gamma_1^t r(S_t^{t+1}) \mid S_0\big] = \sum_{t=0}^\infty \tilde{\mathbb{E}}_\zeta^\pi\big[\lambda_1^t\gamma_1^t\, r_\pi(S_t) \mid S_0\big]$ (since $\tilde{\mathbb{E}}_\zeta^\pi[r(S_t^{t+1}) \mid S_0^t] = r_\pi(S_t)$). By rearranging terms, $\sum_{t=0}^\infty \tilde{\mathbb{E}}_\zeta^\pi\,\big[\lambda_1^t\gamma_1^t\cdot\big(\gamma_{t+1}v(S_{t+1}) - v(S_t)\big) \mid S_0\big]$ equals $-v(S_0) + \sum_{t=1}^\infty \tilde{\mathbb{E}}_\zeta^\pi\,\big[\lambda_1^{t-1}(1-\lambda_t)\cdot\gamma_1^t v(S_t) \mid S_0\big]$. Putting these together and using the expression of T in (3.9), we obtain (3.15).

11. Mahmood, A.R., van Hasselt, H., Sutton, R.S.: Weighted importance sampling for off-policy learning with linear function approximation. In: Advances in Neural Information Processing Systems (NIPS) 27 (2014)
12. Mahmood, A.R., Yu, H., Sutton, R.S.: Multi-step off-policy learning without importance-sampling ratios (2017), arXiv:1702.03006
13. Meyn, S.: Ergodic theorems for discrete time stochastic systems using a stochastic Lyapunov function. SIAM J. Control Optim. 27, 1409–1439 (1989)
14. Meyn, S., Tweedie, R.L.: Markov Chains and Stochastic Stability. Cambridge University Press, Cambridge, 2nd edn. (2009)
15. Munos, R., Stepleton, T., Harutyunyan, A., Bellemare, M.G.: Safe and efficient off-policy reinforcement learning. In: Advances in Neural Information Processing Systems (NIPS) 29 (2016)
16. Nummelin, E.: General Irreducible Markov Chains and Non-Negative Operators. Cambridge University Press, Cambridge (1984)
17. Precup, D., Sutton, R.S., Dasgupta, S.: Off-policy temporal-difference learning with function approximation. In: The 18th Int. Conf. Machine Learning (ICML) (2001)
18. Precup, D., Sutton, R.S., Singh, S.: Eligibility traces for off-policy policy evaluation. In: The 17th Int. Conf. Machine Learning (ICML) (2000)
19. Puterman, M.L.: Markov Decision Processes: Discrete Stochastic Dynamic Programming. John Wiley & Sons, New York (1994)
20. Randhawa, R.S., Juneja, S.: Combining importance sampling and temporal difference control variates to simulate Markov chains. ACM Trans. Modeling and Computer Simulation 14(1), 1–30 (2004)
21. Rudin, W.: Real and Complex Analysis. McGraw-Hill, New York (1966)
22. Scherrer, B.: Should one compute the temporal difference fix point or minimize the Bellman residual? In: The 27th Int. Conf. Machine Learning (ICML) (2010)
23. Sutton, R.S.: Learning to predict by the methods of temporal differences. Machine Learning 3, 9–44 (1988)
24. Sutton, R.S.: TD models: Modeling the world at a mixture of time scales. In: The 12th Int. Conf. Machine Learning (ICML) (1995)
25. Sutton, R.S., Barto, A.G.: Reinforcement Learning. MIT Press, Cambridge, MA (1998)
26. Sutton, R.S., Mahmood, A.R., White, M.: An emphatic approach to the problem of off-policy temporal-difference learning. Journal of Machine Learning Res. 17(73), 1–29 (2016)
27. Tsitsiklis, J.N., Van Roy, B.: An analysis of temporal-difference learning with function approximation. IEEE Trans. Autom. Control 42(5), 674–690 (1997)
28. Ueno, T., Maeda, S., Kawanabe, M., Ishii, S.: Generalized TD learning. Journal of Machine Learning Res. 12, 1977–2020 (2011)
29. Yu, H.: Least squares temporal difference methods: An analysis under general conditions. SIAM J. Control Optim. 50, 3310–3343 (2012)
30. Yu, H.: On convergence of emphatic temporal-difference learning. In: The 28th Ann. Conf. Learning Theory (COLT) (2015), a longer version at arXiv:1506.02582
31. Yu, H.: Weak convergence properties of constrained emphatic temporal-difference learning with constant and slowly diminishing stepsize. Journal of Machine Learning Res. 17(220), 1–58 (2016)
32. Yu, H., Bertsekas, D.P.: Error bounds for approximations from projected linear equations. Math. Oper. Res. 35(2), 306–329 (2010)
33. Yu, H., Bertsekas, D.P.: Weighted Bellman equations and their applications in approximate dynamic programming. LIDS Technical Report 2876, MIT (2012)