On Convergence of Emphatic Temporal-Difference Learning

Huizhen Yu JANEY.HZYU@GMAIL.COM
Department of Computing Science, University of Alberta, Canada

Editors: Elad Hazan and Peter Grünwald

Abstract

We consider emphatic temporal-difference learning algorithms for policy evaluation in discounted Markov decision processes with finite spaces. Such algorithms were recently proposed by Sutton, Mahmood, and White (2015) as an improved solution to the problem of divergence of off-policy temporal-difference learning with linear function approximation. We present in this paper the first convergence proofs for two emphatic algorithms, ETD(λ) and ELSTD(λ). We prove, under general off-policy conditions, the convergence in L^1 for ELSTD(λ) iterates, and the almost sure convergence of the approximate value functions calculated by both algorithms using a single infinitely long trajectory. Our analysis involves new techniques with applications beyond emphatic algorithms leading, for example, to the first proof that standard TD(λ) also converges under off-policy training for λ sufficiently large.

Keywords: Markov decision processes; approximate policy evaluation; reinforcement learning; temporal difference methods; importance sampling; stochastic approximation; convergence

1. Introduction

We consider discounted finite-spaces Markov decision processes (MDPs) and the problem of learning an approximate value function for a given policy from *off-policy* data, that is, from data due to a different policy. The first policy is called the *target* policy and the second is called the *behavior* policy. For example, one may want to learn value functions for many target policies in parallel from one (exploratory) behavior; this requires off-policy learning.

We focus on temporal-difference (TD) methods with linear function approximation (Sutton, 1988). Such methods are typically convergent when the target and behavior policies are the same (the *on-policy* case), but not in the off-policy case (Tsitsiklis and Van Roy, 1997). This difficulty is intrinsic to sampling states according to an arbitrary policy.[1] Gradient-based or least squares-based approaches have been used to avoid this difficulty.[2]

Recently, Sutton, Mahmood, and White (2015) proposed a new approach to address this issue more directly. They introduced an *emphatic TD(λ)* algorithm, or *ETD(λ)* as we call it here. The approach is related to the early work on episodic off-policy TD(λ) (Precup et al., 2001), and is based on the idea of re-weighting the states when forming the eligibility traces in TD(λ), so that the weights reflect the occupation frequencies of the target policy rather than the behavior policy. The result of this weighting scheme is that the "mean updates" associated with ETD(λ) now involve a negative definite matrix, similar to the convergent on-policy TD algorithms. This is a salient feature of the emphatic TD method.

[1] See the papers (Baird, 1995; Tsitsiklis and Van Roy, 1997; Sutton et al., 2015) and the books (Bertsekas and Tsitsiklis, 1996; Sutton and Barto, 1998) for related examples and discussion.

[2] See e.g., (Maei, 2011; Bertsekas and Yu, 2009; Geist and Scherrer, 2014; Dann et al., 2014).

The purpose of this paper is to investigate the convergence properties of ETD(λ) and its least-squares version, ELSTD(λ). Under general conditions, we show that (see Theorems 2.1, 2.2):

(i) for stepsizes decreasing as $t^{-c}, c \in (1/2, 1]$, the matrix and vector iterates generated by ELSTD(λ) converge in L^1 to the desired limits, which define a projected Bellman equation;

(ii) for stepsizes decreasing as t^{-1}, both algorithms generate approximate value functions that converge almost surely to the desired solution of an associated projected Bellman equation.

These results show that the new emphatic TD algorithms are sound for off-policy learning.

Regarding proof techniques, we note that although the "mean updates" of ETD(λ) involve a negative definite matrix, it is still difficult to directly apply results from stochastic approximation theory to establish rigorously the association between the "mean updates" and the ETD(λ) iterates, thereby obtaining the desired convergence. The stability criterion of (Borkar and Meyn, 2000) (see also (Borkar, 2008, Chap. 3)) and the "natural averaging" argument in (Borkar, 2008, Chap. 6) seem suitable, but they require a certain tightness condition that is hard to verify in the general off-policy learning setting where the variances of the trace iterates can grow to infinity with time.[3] The analysis of (Tsitsiklis and Van Roy, 1997) has a strong condition (Condition (6), p. 683, in particular), which is difficult to satisfy unless the trace iterates are uniformly bounded. But in general, this would impose a strong restriction on the behavior policy (cf. Yu, 2012, Prop. 3.1, Footnote 3, and the discussion in p. 3320-3322).

For regular off-policy LSTD(λ) and TD(λ) (Bertsekas and Yu, 2009), it has been shown by Yu (2012) that the associated joint process of states and trace iterates exhibit useful properties, by which convergence results for LSTD(λ) can be derived. Subsequently, the results can be used to furnish the conditions of a convergence theorem from stochastic approximation theory (Kushner and Yin, 2003) and yield convergence results for TD(λ). In this paper we will take the proof approach used in (Yu, 2012). We note, however, that most of the intermediate results needed in our case require different and more involved proofs, due to the complexity of the emphatic TD method. Furthermore, we will give a new argument to prove the almost sure convergence of ETD(λ), which applies also to the regular off-policy TD(λ) of (Bertsekas and Yu, 2009) for λ near 1. This improves a result of (Yu, 2012), which only dealt with a constrained version of TD(λ) that restricts the iterates to lie in a bounded set.

This paper is organized as follows. In Section 2 we formulate the approximate policy evaluation problem, and we describe the ETD(λ) and ELSTD(λ) algorithms, and the approximate Bellman equations they aim to solve. We also state our main convergence results in this section. In Section 3 we prove our convergence theorem for ELSTD(λ), and prepare results needed for analyzing ETD(λ) with a "mean ODE"[4] method. In Section 4 we prove our convergence theorem for ETD(λ). Due to space limit, several long proofs and related results are given in Appendices A-C. (The full analysis can be found in our arXiv report (2015).)

2. Emphatic TD Algorithms: ETD(λ) and ELSTD(λ)

2.1. A Policy Evaluation Problem in Off-Policy Learning

Let $\mathcal{S} = \{1, \ldots, N\}$ be the state space, and let $\mathcal{A}$ be a finite set of actions. We assume, without loss of generality, that for every state, all actions are feasible. If we take action $a \in \mathcal{A}$ at state $s \in \mathcal{S}$, the

[3] Related examples can be found in (Glynn and Iglehart, 1989; Randhawa and Juneja, 2004; Sutton et al., 2015).

[4] ODE stands for ordinary differential equation.

system moves from state s to state s' with probability $p(s' \mid s, a)$, and we receive a random reward with mean $r(s, a, s')$ and bounded variance, according to a probability distribution $q(\cdot \mid s, a, s')$.

We are interested in evaluating the performance of a given stationary policy[5] π, the target policy, without knowledge of the MDP model. The evaluation is to be done by using just observations of state transitions and rewards, while following a stationary policy $\pi^o \neq \pi$, the behavior policy.

Starting from time $t = 0$, applying π would generate a sequence of rewards $R_0, R_1, \ldots$. The performance of π will be measured in terms of the expected total rewards attained under π up to a random termination time $\tau \geq 1$ that depends on the states in a Markovian way. In particular, if at time $t \geq 1$, the state is s and termination has not occurred yet, then the probability of $\tau = t$ (terminating at time t) is $1 - \gamma(s)$, for a given parameter $\gamma(s) \in [0, 1]$.

Let P_π denote the transition matrix of the Markov chain on $\mathcal{S}$ induced by π. Let Γ denote the $N \times N$ diagonal matrix with diagonal entries $\gamma(s), s \in \mathcal{S}$. Denote by $\pi(a \mid s)$ and $\pi^o(a \mid s)$ the probability of taking action a at state s under the policy π and π^o, respectively.

Assumption 2.1 (Conditions on the target and behavior policies)

(i) *The target policy π is such that $(I - P_\pi \Gamma)^{-1}$ exists (equivalently, termination occurs with probability 1 under π, for any initial state).*

(ii) *The behavior policy π^o induces an irreducible Markov chain on $\mathcal{S}$, and moreover, for all $(s, a) \in \mathcal{S} \times \mathcal{A}$, $\pi^o(a \mid s) > 0$ if $\pi(a \mid s) > 0$.*

Under Assumption 2.1(i), we define the *value function* of the target policy π by $v_\pi : \mathcal{S} \to \mathbb{R}$, $v_\pi(s) = \mathbb{E}_\pi \left[\sum_{t=0}^{\tau-1} R_t \,\middle|\, S_0 = s \right]$, where $\mathbb{E}_\pi$ denotes expectation with respect to the probability distribution of the process of states, actions and rewards, (S_t, A_t, R_t), $t \geq 0$, induced by the policy π. Let r_π be the expected one-stage reward function under π; i.e., $r_\pi(s) = \mathbb{E}_\pi \big[R_0 \mid S_0 = s \big]$ for $s \in \mathcal{S}$. Then the desired function v_π can be seen to satisfy uniquely the Bellman equation[6]

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

2.2. Algorithms

We consider computing v_π with the ETD(λ) algorithm (Sutton et al., 2015) and its least-squares version, ELSTD(λ), using linear function approximation, while following the behavior policy π^o. Let $E \subset \mathbb{R}^N$ be the approximation subspace of dimension n, and let Φ be an $N \times n$ matrix whose columns form a basis of E. The approximation problem is to find a parameter vector $\theta \in \mathbb{R}^n$ such that $v = \Phi\theta \in E$ approximates v_π well.

We express $v = \Phi\theta$ as $v(s) = \phi(s)^\top \theta$, $s \in \mathcal{S}$, where the superscript $^\top$ stands for transpose, and $\phi(s) \in \mathbb{R}^n$ is the transposed s-th row of Φ and represents the "features" of state s. Like standard TD(λ), if a transition (s, s') occurs with reward r', ETD(λ) and ELSTD(λ) use the "temporal difference" term, $r' + \gamma(s')\phi(s')^\top \theta - \phi(s)^\top \theta$, to adjust the parameter θ for the approximate value function. Also like standard TD(λ), these algorithms aim to solve a projected (single-step or multi-step) Bellman equation; but we shall defer the discussion of this until after describing the ETD(λ) algorithm.

[5] A *stationary policy* is a decision rule that specifies the probability of taking action a at state s for every $s \in \mathcal{S}$.

[6] One can verify this Bellman equation directly. It also follows from the standard MDP theory (see e.g., Puterman, 1994), as by definition v_π here can be related to a value function in a discounted MDP where the discount factors depend on state transitions, similar to discounted semi-Markov decision processes.

We focus on a general form of the ETD(λ) algorithm, which uses state-dependent λ values specified by a function $\lambda : \mathcal{S} \to [0,1]$. Inputs to the algorithm are the states, actions and rewards, $\{(S_t, A_t, R_t), t \geq 0\}$, generated under the behavior policy π^o, where R_t is the random reward received upon the transition from state S_t to S_{t+1} with action A_t. The algorithm can access the following functions, in addition to the features $\phi(s)$:

(i) $\gamma : \mathcal{S} \to [0,1]$, which specifies the termination probabilities (or equivalently, the state-dependent discount factors) that define v_π, as described earlier;

(ii) $\lambda : \mathcal{S} \to [0,1]$, which determines the single or multi-step Bellman equation for the algorithm [cf. the subsequent Eqs. (2.5)-(2.6)];

(iii) $\rho : \mathcal{S} \times \mathcal{A} \to \mathbb{R}_+$ given by $\rho(s,a) = \pi(a \mid s)/\pi^o(a \mid s)$ (with $0/0 = 0$), which gives the likelihood ratios for action probabilities that can be used to compensate for sampling states and actions according to the behavior policy π^o instead of the target policy π;

(iv) $i : \mathcal{S} \to \mathbb{R}_+$, which gives the algorithm additional flexibility to weigh states according to the degree of "interest" indicated by $i(s)$.

The ETD(λ) algorithm does the following. For each $t \geq 0$, let $\alpha_t \in (0,1]$ be a stepsize parameter, and to simplify notation, let

$$\rho_t = \rho(S_t, A_t), \qquad \gamma_t = \gamma(S_t), \qquad \lambda_t = \lambda(S_t).$$

ETD(λ) calculates recursively $\theta_t \in \mathbb{R}^n$, $t \geq 0$, according to

$$\theta_{t+1} = \theta_t + \alpha_t \, e_t \cdot \rho_t \left(R_t + \gamma_{t+1} \phi(S_{t+1})^\top \theta_t - \phi(S_t)^\top \theta_t \right), \tag{2.1}$$

where $e_t \in \mathbb{R}^n$ (called the "eligibility trace") is calculated together with two nonnegative scalar iterates (F_t, M_t) according to:[7]

$$F_t = \gamma_t \, \rho_{t-1} \, F_{t-1} + i(S_t), \tag{2.2}$$

$$M_t = \lambda_t \, i(S_t) + (1 - \lambda_t) \, F_t, \tag{2.3}$$

$$e_t = \lambda_t \, \gamma_t \, \rho_{t-1} \, e_{t-1} + M_t \, \phi(S_t). \tag{2.4}$$

For $t = 0$, (e_0, F_0, θ_0) are given as an initial condition of the algorithm.

We recognize that the iteration (2.1) has the same form as standard TD, but the trace e_t is calculated differently, involving an "emphasis" weight M_t on the state S_t, which itself evolves along with the iterate F_t, called the "follow-on" trace. If M_t is always set to 1 regardless of F_t and $i(\cdot)$, then the iteration (2.1) reduces to the standard TD(λ) in the case where γ and λ are constants.

To explain at a high level what ETD(λ) aims to achieve with the weighting scheme (2.2)-(2.4), let us discuss the approximate Bellman equation it aims to solve. Associated with ETD(λ) is a generalized Bellman equation of which v_π is the unique solution (Sutton, 1995):[8]

$$v = r^\lambda_{\pi,\gamma} + P^\lambda_{\pi,\gamma} \, v. \tag{2.5}$$

Here $P^\lambda_{\pi,\gamma}$ is an $N \times N$ substochastic matrix, and $r^\lambda_{\pi,\gamma} \in \mathbb{R}^N$ is a vector of expected total rewards attained by π up to some random time depending on the functions γ and λ, given by

$$P^\lambda_{\pi,\gamma} = I - (I - P_\pi \Gamma \Lambda)^{-1} (I - P_\pi \Gamma), \qquad r^\lambda_{\pi,\gamma} = (I - P_\pi \Gamma \Lambda)^{-1} \, r_\pi, \tag{2.6}$$

[7] For insights about ETD(λ), see (Sutton et al., 2015; Mahmood et al., 2015). Our definition (2.4) of $\{e_t\}$ differs slightly from its original definition, but the two are equivalent; ours appears to be more convenient for our analysis.

[8] For lack of space, we do not explain the details of this Bellman equation, which can be found in the early work (Sutton, 1995; Sutton and Barto, 1998) and the recent work (Sutton et al., 2015).

where Λ is a diagonal matrix with diagonal entries $\lambda(s), s \in \mathcal{S}$. ETD($\lambda$) aims to solve a projected version of the Bellman equation (2.5) (see Sutton et al., 2015):

$$v = \Pi\big(r^{\lambda}_{\pi,\gamma} + P^{\lambda}_{\pi,\gamma} v\big), \quad v \in E, \qquad \Longleftrightarrow \qquad C\theta + b = 0, \quad \theta \in \mathbb{R}^n. \tag{2.7}$$

In the above, Π is the projection onto E with respect to a weighted Euclidean norm or seminorm. The weights that define this norm also define the diagonal entries of a diagonal matrix $\bar{M}$, and are given by

$$diag(\bar{M}) = d^{\top}_{\pi^o,i}(I - P^{\lambda}_{\pi,\gamma})^{-1}, \qquad \text{with} \;\; d_{\pi^o,i} \in \mathbb{R}^N,\, d_{\pi^o,i}(s) = d_{\pi^o}(s) \cdot i(s),\; s \in \mathcal{S}, \tag{2.8}$$

where $d_{\pi^o}(s) > 0$ denotes the steady state probability of state s for the behavior policy π^o, under Assumption 2.1(ii). For the corresponding linear equation in the θ-space in Eq. (2.7),

$$C = -\Phi^{\top} \bar{M}\, (I - P^{\lambda}_{\pi,\gamma})\, \Phi, \qquad\qquad b = \Phi^{\top} \bar{M}\, r^{\lambda}_{\pi,\gamma}. \tag{2.9}$$

Important for the convergence of ETD(λ) is the negative definiteness of C. It can be shown that under Assumption 2.1, C is negative definite whenever C is nonsingular.[9] By comparison, if we set $M_t = 1$ regardless of F_t and $i(\cdot)$, the weights that define the projection norm and $diag(\bar{M})$ would simply become d_{π^o}, the same as in the regular off-policy TD(λ). If we set $M_t = i(s)$, then the weights are given by $d_{\pi^o,i}$. Neither of these cases guarantees C to be negative definite, unless λ is sufficiently close to 1. Having the desirable negative definiteness property of C is one of the motivations for introducing the weighting scheme (2.2)-(2.4) in ETD(λ) (Sutton et al., 2015).

For the convergence analysis in this paper, we shall assume:

Assumption 2.2 (Nonsingularity condition) *The matrix C given in Eq. (2.9) is nonsingular.*

We remark that for ETD(λ) under Assumption 2.1, C is always negative semidefinite (Sutton et al., 2015) (cf. our Prop. C.1, Appendix C), so the nonsingularity condition above is equivalent to C being negative definite, as noted earlier. This condition is fairly mild and allows $i(s) = 0$ for some states s. Specifically, as we prove in Appendix C, Assumption 2.2 is equivalent to a condition on the approximation subspace (Prop. C.2), which requires merely that the set of feature vectors of those states with positive emphasis weights contains n linearly independent vectors (cf. Remark C.2). Moreover, this requirement can be fulfilled easily without knowledge of the model (see Cor. C.1, Remark C.2). We also note that when C is negative definite, the projection Π in Eq. (2.7) is well-defined (with respect to a seminorm if in Eq. (2.8) some diagonal entries of $\bar{M}$ equal zero), the projected Bellman equation (2.7) has a unique solution, and bounds on the approximation error of ETD(λ) can be derived using the approach of Scherrer (2010). (For details of this discussion, see Remark C.1 in Appendix C.)

The ELSTD(λ) algorithm aims to solve the same projected Bellman equation (2.7) as ETD(λ). ELSTD(λ) calculates iteratively an $n \times n$ matrix C_t and a vector $b_t \in \mathbb{R}^n$ according to

$$C_{t+1} = (1 - \alpha_t)\, C_t + \alpha_t\, e_t \cdot \rho_t \big(\gamma_{t+1}\phi(S_{t+1})^{\top} - \phi(S_t)^{\top}\big), \tag{2.10}$$

$$b_{t+1} = (1 - \alpha_t)\, b_t + \alpha_t\, e_t \cdot \rho_t\, R_t, \tag{2.11}$$

where the trace e_t is calculated according to Eqs. (2.2)-(2.4) as in ETD(λ). ELSTD(λ) sets $\theta_t = -C_t^{-1} b_t$, the solution to $C_t\theta + b_t = 0$, when C_t is invertible.

Like ETD(λ), without the weighting scheme (2.2)-(2.4), ELSTD(λ) would reduce essentially to the regular LSTD(λ) (see e.g., (Boyan, 1999; Yu, 2012) for on-policy and off-policy LSTD(λ)).

[9] The negative definiteness of C is proved for positive $i(\cdot)$ under Assumption 2.1 by Sutton et al. (2015), and their result extends to nonnegative $i(\cdot)$, as long as C is nonsingular (see our Prop. C.1 in Appendix C).

2.3. Convergence Results

We analyze ETD(λ) and ELSTD(λ) with diminishing stepsizes. Summarized below are their convergence properties, which we will establish in the rest of this paper. In what follows, we denote by $\|\cdot\|$ the infinity norm for both vectors and matrices (viewed as vectors). For different stepsize conditions, our results will involve different convergence modes: convergence in L^1,[10] in probability, or almost sure (a.s.) convergence (we write $\overset{a.s.}{\to}$ for "converges almost surely"). First, we state a general stepsize condition that we will use.

Assumption 2.3 (Stepsize condition) *The stepsize sequence $\{\alpha_t\}$ is deterministic and eventually nonincreasing, and satisfies $\alpha_t \in (0,1]$, $\sum_t \alpha_t = \infty$, $\sum_t \alpha_t^2 < \infty$.*

Under the above condition we may take $\alpha_t = t^{-c}, c \in (1/2, 1]$. However, stepsizes decreasing as t^{-1} will be required in our almost sure convergence results; some cases will require $\alpha_t = O(1/t)$ with $\frac{\alpha_t - \alpha_{t+1}}{\alpha_t} = O(1/t)$.[11] (For instance, $\alpha_t = c_1/(c_2 + t)$ for some constants $c_1, c_2 > 0$.)

Our results are as follows. Let θ^* denote the desired limit for ETD(λ):

$$\theta^* = -C^{-1}b, \qquad \text{for } C, b \text{ defined by Eq. (2.9) under Assumptions 2.1, 2.2.}$$

Theorem 2.1 (L^1 and almost sure convergence of ELSTD(λ) Iterates)
Under Assumptions 2.1, 2.3, for any given initial (e_0, F_0, C_0, b_0), the sequence $\{(C_t, b_t)\}$ generated by the ELSTD(λ) algorithm (2.10)-(2.11) converges in L^1:

$$\lim_{t\to\infty} \mathbb{E}\big[\|C_t - C\|\big] = 0, \qquad \lim_{t\to\infty} \mathbb{E}\big[\|b_t - b\|\big] = 0.$$

If in addition the stepsize is given by $\alpha_t = 1/(t+1)$, then $C_t \overset{a.s.}{\to} C$, $b_t \overset{a.s.}{\to} b$.

The preceding theorem yields immediately the convergence of the parameter sequence $\{\theta_t\}$ generated by ELSTD(λ):

Corollary 2.1 (Convergence of ELSTD(λ)) *Let Assumptions 2.1-2.3 hold. Let $\{\theta_t\}$ be generated by the ELSTD(λ) algorithm (2.10)-(2.11) as $\theta_t = -C_t^{-1}b_t$. Then for any given initial (e_0, F_0, C_0, b_0), $\{\theta_t\}$ converges to θ^* in probability; if in addition $\alpha_t = 1/(t+1)$, then $\theta_t \overset{a.s.}{\to} \theta^*$.*

Theorem 2.2 (Almost sure convergence of ETD(λ)) *Let Assumptions 2.1-2.3 hold. Let $\{\theta_t\}$ be generated by the ETD(λ) algorithm (2.1) with stepsizes satisfying $\alpha_t = O(1/t)$ and $\frac{\alpha_t - \alpha_{t+1}}{\alpha_t} = O(1/t)$. Then for any given initial (e_0, F_0, θ_0), $\theta_t \overset{a.s.}{\to} \theta^*$.*

Remark 2.1 (On stepsizes) We believe that the range of stepsizes for the a.s. convergence of ELSTD(λ) can be enlarged. If additional conditions on the behavior policy are imposed to restrict the variances of the trace iterates, it should also be possible to enlarge the range of stepsizes for ETD(λ). These topics, as well as the use of random stepsizes, are under active investigation.

Remark 2.2 (On variances) The preceding convergence results hold under almost minimal conditions on the behavior policy (Assumption 2.1(ii)). However, unless we restrict sufficiently the behavior policy (which is difficult to do without knowledge of the model, when $\gamma \not< 1$), the variances of the trace iterates can grow unboundedly (cf. Remark A.1), significantly affecting the speed of convergence. This is a main difficulty in off-policy methods in general. Further research is required to overcome it. For a recent work in this direction, see (Mahmood et al., 2014).

[10] For vector-valued random variables $X, X_t, t \geq 0$, by "$\{X_t\}$ converges to X in L^1" we mean $\mathbb{E}[\|X_t - X\|] \overset{t\to\infty}{\to} 0$.
[11] We write $\delta_t = O(1/t)$ for a scalar sequence $\{\delta_t\}$, if for some $c > 0$, $0 \leq \delta_t \leq c/t$ for all t.

3. Properties of Trace Iterates and Convergence Analysis of ELSTD(λ)

In this section we analyze the trace iterates and convergence properties of ELSTD(λ) iterates. The analysis not only leads to Theorem 2.1 on the convergence of ELSTD(λ), but also prepares the stage for the subsequent ODE-based convergence proof for ETD(λ), by ensuring that "local averaging" gives the desired "mean dynamics," as will be seen in Section 4.

The structure of our analysis will be similar to that of (Yu, 2012) for regular off-policy LSTD(λ), but the proofs at intermediate steps are new and more involved. Due to space limit, we will explain only key proof arguments here and include some proofs in Appendix A. The full details of our analysis can be found in our arXiv report (2015, Appendix A).

3.1. Properties of Trace Iterates

Let $Z_t = (S_t, A_t, e_t, F_t)$ for $t \geq 0$; they form a Markov chain on $\mathcal{S} \times \mathcal{A} \times \mathbb{R}^{n+1}$. First, we observe several important properties of the trace iterates $\{(e_t, F_t)\}$ and the Markov chain $\{Z_t\}$, under Assumption 2.1:

(i) For any given initial (e_0, F_0), $\sup_{t\geq 0} \mathbb{E}\big[\big\|(e_t, F_t)\big\|\big] < \infty$. (See Prop. A.1.)

(ii) Let $\{(e_t, F_t)\}$ and $\{(\hat{e}_t, \hat{F}_t)\}$ be defined by the same recursion (2.2)-(2.4), using the same state and action random variables, but with different initial conditions $(e_0, F_0) \neq (\hat{e}_0, \hat{F}_0)$. Then $F_t - \hat{F}_t \overset{a.s.}{\to} 0$ and $e_t - \hat{e}_t \overset{a.s.}{\to} \mathbf{0}$ (the zero vector in $\mathbb{R}^n$). (See Prop. A.2.)

(iii) We can approximate the traces (e_t, F_t), which depend on the entire history of past states and actions, by similarly defined "truncated traces" $(\tilde{e}_{t,K}, \tilde{F}_{t,K})$ which depend on the most recent $2K$ states and actions only [cf. Eqs. (A.3)-(A.5)]. The expected approximation "error" can be bounded uniformly in t, by a constant L_K which decreases to 0 as $K \to \infty$. (See Prop. A.3.)

(iv) $\{Z_t\}$ is a weak Feller Markov chain[12] and bounded in probability,[13] and hence it has at least one invariant probability measure.[14]

Furthermore, as we will show in Theorem 3.2 below, $\{Z_t\}$ has a *unique* invariant probability measure and is ergodic.

These properties suggest that despite the growing variances, the trace iterates are well-behaved. Figure 1 shows how the convergence results of this section, to be introduced next, will depend on these properties.

3.2. Main Results on L^1 and Almost Sure Convergence

We formulate our convergence results in terms of a general recursion that can be specialized to the ELSTD(λ) iteration. This generality is needed in order to make the results useful for other

[12] A Markov chain $\{X_t\}$ on a metric space is *weak Feller* if $\mathbb{E}[f(X_1) \mid X_0 = x]$ is continuous in x for every bounded continuous function f on the state space (Meyn and Tweedie, 2009, Prop. 6.1.1(i)). Using this and the fact that (e_1, F_1) depends continuously on (e_0, F_0) [cf. Eqs. (2.2)-(2.4)], the weak Feller property of $\{Z_t\}$ can be seen.

[13] A Markov chain $\{X_t\}$ on a topological space is *bounded in probability* if, for each initial state x and each $\epsilon > 0$, there exists a compact subset D of the state space such that $\liminf_{t\to\infty} \mathbf{P}_x(X_t \in D) \geq 1 - \epsilon$, where $\mathbf{P}_x$ denotes the probability of events conditional on $X_0 = x$ (Meyn and Tweedie, 2009, p. 142). In our case, since $\mathcal{S}$ and $\mathcal{A}$ are finite, the property (i) above together with the Markov inequality implies that $\{Z_t\}$ is bounded in probability (cf. Yu, 2012, Lemma 3.4).

[14] By (Meyn and Tweedie, 2009, Theorem 12.1.2(ii)), a weak Feller Markov chain bounded in probability has at least one invariant probability measure. We mention that there is also an alternative, direct proof of the existence of an invariant probability measure for $\{Z_t\}$, which does not rely on the weak Feller property (Yu, 2015, Appendix A.6).

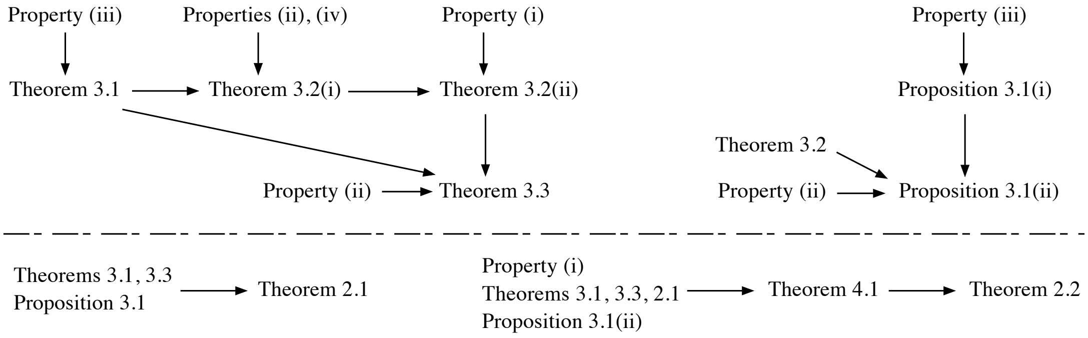

Figure 1: Diagrams showing dependence relations between the results in this paper. "$A \to B$" means A is used in proving B.

proofs, specifically, for proving the uniqueness of the invariant probability measure of $\{Z_t\}$, and for establishing convergence conditions required by an ODE-based analysis for ETD(λ), as those proofs will rely on the convergence properties of certain iterates that are different from ELSTD(λ).

We define the general recursion just mentioned as follows. Denote $y = (e, F)$; thus $y \in \mathbb{R}^{n+1}$. Consider a vector-valued function $h : \mathbb{R}^{n+1} \times \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}^m$ such that $h(y, s, a, s')$ is Lipschitz continuous in y for each (s, a, s'); i.e., there exists some constant L_h such that for any $y, \hat{y} \in \mathbb{R}^{n+1}$,

$$\big\| h(y, s, a, s') - h(\hat{y}, s, a, s') \big\| \leq L_h \| y - \hat{y} \|, \qquad \forall\, (s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}. \tag{3.1}$$

Given h, $\{Z_t\}$ and the stepsizes $\{\alpha_t\}$, we define a recursion as follows:

$$G_{t+1} = (1 - \alpha_t)\, G_t + \alpha_t\, h(Y_t, S_t, A_t, S_{t+1}). \tag{3.2}$$

The ELSTD(λ) iterates C_t and b_t correspond to the following choices of h, respectively:

$$h_1(y, s, a, s') = e \cdot \rho(s, a) \left(\gamma(s') \phi(s')^\top - \phi(s)^\top \right), \quad h_2(y, s, a, s') = e \cdot \rho(s, a)\, r(s, a, s'). \tag{3.3}$$

Here h_1 is matrix-valued (we view it as an $\mathbb{R}^m$-valued function with $m = n \times n$), and h_2 is $\mathbb{R}^n$-valued. As just mentioned, we will also need to consider other choices of h in our proofs later.

We first show that $\{G_t\}$ converges in L^1 to some constant vector. The proof (given in Appendix A.2) exploits the property (iii) of truncated traces mentioned earlier: this property allows us to obtain the desired result by working with simple finite-state Markov chains.

Theorem 3.1 (**L^1-convergence of $\{G_t\}$**) *Let h be a vector-valued function satisfying the Lipschitz condition (3.1), and let $\{G_t\}$ be defined by the recursion (3.2), using the process $\{Z_t\}$. Then under Assumptions 2.1, 2.3, there exists a constant vector G^* (independent of the stepsizes) such that for any given initial $Y_0 = (e_0, F_0)$ and G_0, $\lim_{t\to\infty} \mathbb{E}\big[\|G_t - G^*\|\big] = 0$.*

Next we analyze the a.s. convergence of $\{G_t\}$, by using ergodicity properties of the infinite-space Markov chain $\{Z_t\}$ that we establish first. For each initial condition $Z_0 = z$, define the occupation probability measures $\mu_{z,t}$ for $t \geq 1$, by $\mu_{z,t}(B) = \frac{1}{t}\sum_{k=1}^{t} \mathbb{1}_B(Z_k)$ for any Borel subset B of $\mathcal{S} \times \mathcal{A} \times \mathbb{R}^{n+1}$, where $\mathbb{1}_B$ denotes the indicator function for the set B (i.e., $\mathbb{1}_B(x) = 1$ if $x \in B$, and $\mathbb{1}_B(x) = 0$ otherwise). Let $\mathbb{E}_\mu$ denote expectation with respect to the probability distribution of the process $\{Z_t\}$ with μ as the initial distribution of Z_0.

Theorem 3.2 (Ergodicity of $\{Z_t\}$) *Under Assumption 2.1, the Markov chain $\{Z_t\}$ has a unique invariant probability measure ζ, and moreover, the following hold:*
(i) *For each initial condition $Z_0 = z$, the sequence $\{\mu_{z,t}\}$ of occupation measures converges weakly*[15] *to ζ, almost surely.*
(ii) $\mathbb{E}_\zeta\big[\big\|h(Z_0, S_1)\big\|\big] < \infty$ *for any function h satisfying the Lipschitz condition (3.1).*

The preceding theorem follows from the properties of trace iterates given earlier and Theorem 3.1 (cf. Figure 1). The proof is the same as the corresponding proofs of (Yu, 2012, Theorem 3.2 and Prop. 3.2) for the case of off-policy LSTD. In particular, to prove the uniqueness of the invariant probability measure (which is not as easy to prove as the existence given in the property (iv) earlier), we use the property (ii) and the convergence in L^1 result given in Theorem 3.1.[16]

We can now show that $\{G_t\}$ converges a.s. for stepsize $\alpha_t = 1/(t+1)$, by using the preceding results (cf. Figure 1), together with a strong law of large numbers for stationary processes (Doob, 1953, Chap. X, Theorem 2.1) (see also Meyn and Tweedie, 2009, Theorem 17.1.2). The proof is a verbatim repetition of the proof of (Yu, 2012, Theorem 3.3) and is therefore omitted.

Theorem 3.3 (Almost sure convergence of $\{G_t\}$) *Let h and $\{G_t\}$ be as in Theorem 3.1, and let the stepsize be $\alpha_t = 1/(t+1)$. Then, under Assumption 2.1, for any given initial $Y_0 = (e_0, F_0)$ and G_0, $G_t \overset{a.s.}{\to} G^*$, where $G^* = \mathbb{E}_\zeta\big[h(Y_0, S_0, A_0, S_1)\big]$ is the constant vector in Theorem 3.1.*

Finally, we also need to analyze the cumulative effects of noise in the observed rewards R_t and show that they diminish asymptotically. To this end, consider the following recursion: $W_0 = \mathbf{0}$ and

$$W_{t+1} = (1 - \alpha_t)\, W_t + \alpha_t\, e_t\, \rho_t \cdot \omega_{t+1}, \qquad t \geq 0, \tag{3.4}$$

where $\omega_{t+1} = R_t - r(S_t, A_t, S_{t+1})$ are noise variables.

Proposition 3.1 (Effects of noise in random rewards) *Under Assumptions 2.1, 2.3, for any given initial (e_0, F_0), we have* (i) $\mathbb{E}\big[\|W_t\|\big] \to 0$*; and* (ii) *if, in addition, the stepsize is $\alpha_t = 1/(t+1)$, then $W_t \overset{a.s.}{\to} \mathbf{0}$.*

The proof of the preceding proposition can be found in our arXiv report (2015, Appendix A.4). The proof of part (i) uses the property (iii) of truncated traces, similarly to the proof of Theorem 3.1, and the proof of part (ii) is similar to that of Theorem 3.3 (cf. Figure 1).

The convergence of ELSTD(λ) stated in Theorem 2.1 now follows from the preceding results (cf. Figure 1). Specifically, we calculate the limit G^* in Theorem 3.1 for the two functions h_1, h_2 in Eq. (3.3), which are associated with the ELSTD(λ) iterates $\{C_t\}, \{b_t\}$, respectively, and we show that $G^* = C$ for $h = h_1$ and $G^* = b$ for $h = h_2$. We also write the iterates $\{b_t\}$ equivalently as $b_{t+1} = G_{t+1} + W_{t+1}$ with $h = h_2$ in the definition of $\{G_t\}$. Then, the L^1-convergence part of Theorem 2.1 follows from Theorem 3.1 and Prop. 3.1(i), and the a.s. convergence part of Theorem 2.1 follows from Theorem 3.3 and Prop. 3.1(ii). (For further details, see Appendix A.3 and our arXiv report (2015, Appendix A.5).)

[15] For probability measures $\mu, \mu_t, t \geq 0$, on a metric space X, $\{\mu_t\}$ is said to *converge weakly* to μ if for all bounded continuous functions f on X, $\int f d\mu_t \to \int f d\mu$ as $t \to \infty$ (Dudley, 2002, Chap. 9.3).

[16] Theorem 3.1 is useful here because on the separable metric space $\mathcal{S} \times \mathcal{A} \times \mathbb{R}^{n+1}$, bounded Lipschitz continuous functions are convergence-determining for weak convergence of probability measures (Dudley, 2002, Theorem 11.3.3).

4. Convergence Analysis of ETD(λ)

Recall that ETD(λ) calculates iteratively θ_t, $t \geq 0$, according to

$$\theta_{t+1} = \theta_t + \alpha_t \, e_t \cdot \rho_t \left(R_t + \gamma_{t+1} \phi(S_{t+1})^\top \theta_t - \phi(S_t)^\top \theta_t \right). \tag{4.1}$$

Using the results of Section 3, we can now analyze its convergence by applying a "mean ODE" method from stochastic approximation theory (Kushner and Yin, 2003).

Denoting $\tilde{\omega}_{t+1} = \rho_t \left(R_t - r(S_t, A_t, S_{t+1}) \right)$, let us write the iteration (4.1) equivalently as

$$\theta_{t+1} = \theta_t + \alpha_t \, h(\theta_t, \xi_t) + \alpha_t \, e_t \cdot \tilde{\omega}_{t+1}, \tag{4.2}$$

where $\xi_t = (e_t, S_t, A_t, S_{t+1})$ and $h : \mathbb{R}^n \times \mathbb{R}^n \times \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}^n$ is given by

$$h(\theta, \xi) = e \cdot \rho(s,a) \left(r(s,a,s') + \gamma(s') \, \phi(s')^\top \theta - \phi(s)^\top \theta \right), \quad \text{for } \xi = (e, s, a, s'). \tag{4.3}$$

We will apply (Kushner and Yin, 2003, Theorem 6.1.1) to analyze the convergence of $\{\theta_t\}$ generated by (4.1). The "mean ODE" associated with ETD(λ) (4.1) is

$$\dot{x} = \bar{h}(x), \qquad \text{where } \; \bar{h}(x) = Cx + b. \tag{4.4}$$

When C is negative definite, the above ODE has a unique bounded (constant) solution $x(\cdot) \equiv \theta^* = -C^{-1}b$ on the time interval $(-\infty, +\infty)$, and θ^* is globally asymptotically stable for (4.4) in the sense of Liapunov (cf. Kushner and Clark, 1978, p. 23-24). (A Liapunov function in this case is given by $\|\theta - \theta^*\|_2^2$, where $\| \cdot \|_2$ denotes the Euclidean norm.)

However, the a.s. boundedness of $\{\theta_t\}$ is not easy to prove directly, which has prevented us from getting the desired convergence $\theta_t \overset{a.s.}{\to} \theta^*$ from (Kushner and Yin, 2003, Theorem 6.1.1) directly. For this reason, we analyze first a constrained version of (4.1) and establish its convergence. The result will then help the convergence analysis of the unconstrained algorithm (4.1) in Section 4.2.

4.1. Convergence of Constrained ETD(λ)

Consider the following constrained ETD(λ) algorithm:

$$\theta_{t+1} = \Pi_B \Big(\theta_t + \alpha_t \, h(\theta_t, \xi_t) + \alpha_t \, e_t \cdot \tilde{\omega}_{t+1} \Big), \tag{4.5}$$

where B is a closed ball in $\mathbb{R}^n$ with a sufficiently large radius r: $B = \{ \theta \in \mathbb{R}^n \mid \|\theta\|_2 \leq r \}$, and Π_B is the Euclidean projection onto B. The "mean ODE" associated with the constrained algorithm (4.5) is the projected ODE

$$\dot{x} = \bar{h}(x) + z, \qquad z \in -\mathcal{N}_B(x), \tag{4.6}$$

where $\mathcal{N}_B(x)$ is the normal cone of B at x, and z is the boundary reflection term that cancels out the component of $\bar{h}(x)$ in $\mathcal{N}_B(x)$ and is the "minimal force" needed to keep the solution in B (Kushner and Yin, 2003, Chap. 4.3). The negative definiteness of the matrix C implies that the projected ODE (4.6) has no stationary points other than θ^* if the radius of B is sufficiently large:

Lemma 4.1 *Let $c > 0$ be such that $x^\top C x \leq -c\|x\|_2^2$ for all $x \in \mathbb{R}^n$. Suppose B has a radius $r > \|b\|_2 / c$. Then θ^* lies in the interior of B, and the only solution $x(t), t \in (-\infty, +\infty)$, of the projected ODE (4.6) in B is $x(\cdot) \equiv \theta^*$.*

The proof of Lemma 4.1 is given in Appendix B. We now apply (Kushner and Yin, 2003, Theorem 6.1.1) and Lemma 4.1 to prove the a.s. convergence of the constrained ETD(λ) as stated in the theorem below. The proof is given in Appendix B, and it uses the results of Section 3 to verify the conditions required by (Kushner and Yin, 2003, Theorem 6.1.1).

Theorem 4.1 (Almost sure convergence of constrained ETD(λ)) *Let Assumptions 2.1-2.3 hold. Let $\{\theta_t\}$ be the sequence generated by the constrained ETD(λ) algorithm (4.5) with stepsizes satisfying $\alpha_t = O(1/t)$ and $\frac{\alpha_t - \alpha_{t+1}}{\alpha_t} = O(1/t)$, and with the radius r of B exceeding the threshold given in Lemma 4.1. Then, for any given initial (e_0, F_0, θ_0), $\theta_t \overset{a.s.}{\to} \theta^*$.*

4.2. Convergence of ETD(λ)

We now prove the convergence theorem, Theorem 2.2, for the unconstrained ETD(λ) algorithm by using the convergence of the constrained algorithm we just established. In particular, we shall compare the iterates generated by the unconstrained algorithm with those generated by the constrained one, and show that the difference between them diminishes asymptotically with probability one.

Let $B = \big\{\theta \in \mathbb{R}^n \mid \|\theta\|_2 \leq r\big\}$ with its radius r satisfying the condition of Lemma 4.1. Note that to project θ onto B is simply to scale θ: $\Pi_B \theta = \theta$ if $\|\theta\|_2 \leq r$; and $\Pi_B \theta = r \cdot \theta / \|\theta\|_2$ if $\|\theta\|_2 > r$. More concisely,

$$\Pi_B \theta = \eta\, \theta, \qquad \text{where} \quad \eta = \min\{1, r/\|\theta\|_2\}.$$

To simplify notation, define matrix H_t and vector g_t by

$$H_t = e_t \cdot \rho_t \left(\gamma_{t+1}\, \phi(S_{t+1}) - \phi(S_t)\right)^\top, \qquad g_t = e_t \cdot \rho_t\, R_t.$$

Let us write the constrained algorithm (4.5) equivalently as

$$\tilde{\theta}_{t+1} = (I + \alpha_t\, H_t) \cdot \eta_t\, \tilde{\theta}_t + \alpha_t\, g_t, \tag{4.7}$$

where $\eta_0 = 1$ and $\eta_t = \min\{1, r/\|\tilde{\theta}_t\|_2\}$ for $t \geq 1$. (For $t \geq 1$, $\eta_t\, \tilde{\theta}_t$ corresponds to the projected iterate in (4.5), and $\tilde{\theta}_t$ the iterate just before the projection.) The unconstrained algorithm (4.1) can be equivalently written as

$$\theta_{t+1} = (I + \alpha_t\, H_t) \cdot \theta_t + \alpha_t\, g_t. \tag{4.8}$$

Lemma 4.2 *Under the conditions of Theorem 4.1, for any given initial (e_0, F_0), almost surely, the sequence of matrices, $\prod_{k \geq \bar{t}}^{t} (I + \alpha_k\, H_k)$, $t = \bar{t}, \bar{t}+1, \ldots$, converges to the $n \times n$ zero matrix as $t \to \infty$, for all $\bar{t} \geq 0$.*

Proof It is sufficient to consider a given (arbitrary) vector $y \in \mathbb{R}^n$ and prove that for each initial (e_0, F_0) and each $\bar{t} \geq 0$, $\prod_{k \geq \bar{t}}^{t} (I + \alpha_k\, H_k)\, y \overset{a.s.}{\to} \mathbf{0}$. To this end, consider generating the iterates $\tilde{\theta}_{\bar{t}}, \tilde{\theta}_{\bar{t}+1}, \ldots$, starting from time $\bar{t}$ and $\tilde{\theta}_{\bar{t}} = y$, by using the constrained algorithm (4.7) as follows:

$$\tilde{\theta}_{k+1} = (I + \alpha_k\, H_k) \cdot \eta_k\, \tilde{\theta}_k, \qquad k \geq \bar{t}.$$

In the above, we calculate (e_k, F_k) and H_k as before starting from time 0 and the given initial condition (e_0, F_0), and we have set $g_k = R_k = 0$ for all k. Notice that since the stepsize sequence $\{\alpha_t\}$ satisfies the condition of Theorem 4.1, so does the stepsize sequence, $\alpha_{\bar{t}+1}, \alpha_{\bar{t}+2}, \ldots$. Then, in view of the Markovian property of $\{(S_t, A_t, e_t, F_t)\}$, we can apply Theorem 4.1 to the above iteration starting from time $\bar{t}$ for each possible value of $(e_{\bar{t}}, F_{\bar{t}})$, thereby concluding that for the given (e_0, F_0) and $\bar{t}$, $\tilde{\theta}_t \overset{a.s.}{\to} \mathbf{0}$ (because $R_k = 0$ for all k and the solution to $C\theta = 0$ is $\mathbf{0}$). On the other hand,

$$\tilde{\theta}_{t+1} = \left(\prod_{k \geq \bar{t}}^{t} (I + \alpha_k \, H_k)\right) \cdot \left(\prod_{k \geq \bar{t}}^{t} \eta_k\right) \cdot y. \tag{4.9}$$

Since the solution $\mathbf{0}$ lies in the interior of B, if $\tilde{\theta}_t \to \mathbf{0}$, then $\eta_k = 1$ for all k sufficiently large. Thus the convergence $\tilde{\theta}_t \overset{a.s.}{\to} \mathbf{0}$ implies that as $t \to \infty$, $\prod_{k \geq \bar{t}}^{t} \eta_k$ converges a.s. to a strictly positive number that depends on the sample path and the vector y. Consequently, from Eq. (4.9) and the convergence $\tilde{\theta}_t \overset{a.s.}{\to} \mathbf{0}$, we obtain that $\left(\prod_{k \geq \bar{t}}^{t} (I + \alpha_k \, H_k)\right) y \overset{a.s.}{\to} \mathbf{0}$ as $t \to \infty$. Now this holds for any given vector y, so by letting y be each column of the identity matrix, it follows that as $t \to \infty$, the matrix $\prod_{k \geq \bar{t}}^{t} (I + \alpha_k \, H_k)$ converges a.s. to the zero matrix. ■

Finally, we prove the a.s. convergence of the unconstrained ETD(λ) as stated by Theorem 2.2:

Proof of Theorem 2.2 Let $\{\tilde{\theta}_t\}$ be the iterates generated by the constrained algorithm (4.7) using the same trajectory of states, actions and rewards that are used by the unconstrained algorithm (4.1) to generate $\{\theta_t\}$. By Theorem 4.1 and Lemma 4.2, there exists a set Ω_1 of sample paths such that Ω_1 has probability one and on Ω_1,

$$\tilde{\theta}_t \to \theta^* \qquad \text{and} \qquad \lim_{t \to 0} \prod_{k \geq \bar{t}}^{t} (I + \alpha_k \, H_k) = 0_{n \times n}, \quad \forall \, \bar{t} \geq 0,$$

where $0_{n \times n}$ denotes the $n \times n$ zero matrix. Consider each path in Ω_1. By our choice of the constraint set B, θ^* lies in the interior of B (Lemma 4.1), so the convergence $\tilde{\theta}_t \to \theta^*$ implies the existence of a path-dependent time $t' < \infty$ such that $\eta_k = 1$ for all $k \geq t'$. Then

$$\tilde{\theta}_{k+1} = (I + \alpha_k \, H_k) \cdot \tilde{\theta}_k + \alpha_k \, g_k, \qquad \forall \, k \geq t',$$

and consequently,

$$\theta_{k+1} - \tilde{\theta}_{k+1} = (I + \alpha_k \, H_k) \cdot (\theta_k - \tilde{\theta}_k), \qquad \forall \, k \geq t',$$
$$\theta_{t+1} - \tilde{\theta}_{t+1} = \left(\prod_{k \geq t'}^{t} (I + \alpha_k \, H_k)\right) \cdot (\theta_{t'} - \tilde{\theta}_{t'}), \qquad \forall \, t \geq t'. \tag{4.10}$$

As $t \to \infty$, the matrix $\prod_{k \geq t'}^{t} (I + \alpha_k \, H_k) \to 0_{n \times n}$ for the sample path under consideration. Thus, from Eq. (4.10) we obtain $\theta_t - \tilde{\theta}_t \to \mathbf{0}$; since $\tilde{\theta}_t \to \theta^*$, this implies $\theta_t \to \theta^*$. ■

Remark 4.1 (Almost sure convergence of regular off-policy TD(λ)) If λ is a constant sufficiently close to 1, the matrix associated with the "mean updates" of the regular off-policy TD(λ) algorithm is also negative definite (Bertsekas and Yu, 2009). In that case, (Yu, 2012, Prop. 4.1) established the a.s. convergence but only for a constrained version of the algorithm, similar to our Theorem 4.1. The proofs given in this subsection, combined with (Yu, 2012, Prop. 4.1), can be used to establish the desired a.s. convergence for the unconstrained off-policy TD(λ) in that case.

Acknowledgments

I thank Prof. Richard Sutton for discussions on the ETD algorithm, and Dr. Joseph Modayil, Ashique Mahmood, and several anonymous reviewers for helpful comments on the first version of this paper. This research was supported by a grant from Alberta Innovates – Technology Futures.

References

L. C. Baird. Residual algorithms: Reinforcement learning with function approximation. In *Proc. The 12th Int. Conf. Machine Learning*, pages 30–37, 1995.

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

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

V. S. Borkar. *Stochastic Approximation: A Dynamic Viewpoint*. Cambridge University Press, Cambridge, 2008.

V. S. Borkar and S. P. Meyn. The O.D.E. method for convergence of stochastic approximation and reinforcement learning. *SIAM J. Control Optim.*, 38:447–469, 2000.

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

C. Dann, G. Neumann, and J. Peters. Policy evaluation with temporal differences: A survey and comparison. *Journal of Machine Learning Res.*, 15:809–883, 2014.

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

R. M. Dudley. *Real Analysis and Probability*. Cambridge University Press, Cambridge, 2002.

M. Geist and B. Scherrer. Off-policy learning with eligibility traces: A survey. *Journal of Machine Learning Res.*, 15:289–333, 2014.

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

H. J. Kushner and D. S. Clark. *Stochastic Approximation Methods for Constrained and Unconstrained Systems*. Springer-Verlag, New York, 1978.

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

H. R. Maei. *Gradient Temporal-Difference Learning Algorithms*. PhD thesis, University of Alberta, 2011.

A. R. Mahmood, H. van Hasselt, and R. S. Sutton. Weighted importance sampling for off-policy learning with linear function approximation. In *Proc. Conf. Advances in Neural Information Processing Systems (NIPS) 27*, 2014.

A. R. Mahmood, H. Yu, M. White, and R. S. Sutton. Emphatic temporal-difference learning. In *European Workshops on Reinforcement Learning*, Lille, France, 2015.

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

J. Neveu. *Discrete-Parameter Martingales*. North-Holland, Amsterdam, 1975.

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

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

R. S. Randhawa and S. Juneja. Combining importance sampling and temporal difference control variates to simulate Markov chains. *ACM Trans. Modeling and Computer Simulation*, 14(1): 1–30, 2004.

B. Scherrer. Should one compute the temporal difference fix point or minimize the Bellman residual? The unified oblique projection view. In *Proc. The 27th Int. Conf. Machine Learning*, pages 959–966, 2010.

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

R. S. Sutton. TD models: Modeling the world at a mixture of time scales. In *Proc. The 12th Int. Conf. Machine Learning*, pages 531–539, 1995.

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

R. S. Sutton, A. R. Mahmood, and M. White. An emphatic approach to the problem of off-policy temporal-difference learning, 2015. http://arxiv.org/abs/1503.04269.

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

R. S. Varga. *Matrix Iterative Analysis*. Springer-Verlag, Berlin, 2nd edition, 2000.

H. Yu. Least squares temporal difference methods: An analysis under general conditions. *SIAM J. Control Optim.*, 50:3310–3343, 2012.

H. Yu. On convergence of emphatic temporal-difference larning. Technical report, University of Alberta, 2015. http://arxiv.org/abs/1506.02582.

H. Yu and D. P. Bertsekas. Weighted Bellman equations and their applications in approximate dynamic programming. LIDS Technical Report 2876, MIT, 2012.

Appendix A. Some Proof Details for Section 3

We include in this appendix the following technical results and proofs for Section 3:

(i) formal statements of the properties of trace iterates which we mentioned in Section 3.1 and used in many proofs;

(ii) the proof of Theorem 3.1, which concerns L^1-convergence; and

(iii) the proof of Theorem 2.1 on the convergence of ELSTD(λ), which we outlined at the end of Section 3.2.

We refer the readers to our arXiv report (2015, Appendix A) for the full details of our analysis, some of which have been left out in this paper for lack of space.

A.1. Properties of the Trace Iterates $\{(e_t, F_t)\}$

Throughout this subsection, Assumption 2.1 on the target and behavior policies will be in force and will not be mentioned explicitly. Recall that $\{Z_t\}$ with $Z_t = (S_t, A_t, e_t, F_t)$ denotes the Markov chain on the joint space $\mathcal{S} \times \mathcal{A} \times \mathbb{R}^{n+1}$ of states, actions and traces, and it is a weak Feller Markov chain (cf. Footnote 12). As explained in Section 3.1, since $\mathcal{S}$ and $\mathcal{A}$ are finite, the follow proposition implies that $\{Z_t\}$ is bounded in probability and hence, by its weak Feller property, has at least one invariant probability measure. This proposition is the property (i) mentioned in Section 3.1. We prove it by a direct calculation, the details of which can be found in (Yu, 2015, Appendix A.2).

Proposition A.1 *For any given initial* (e_0, F_0)*,* $\sup_{t \geq 0} \mathbb{E}\big[\big\|(e_t, F_t)\big\|\big] < \infty$.

The following result is the property (ii) mentioned in Section 3.1. It is useful in several proofs; in particular, it is used in proving that $\{Z_t\}$ has a *unique* invariant probability measure.

Let $(\hat{e}_t, \hat{F}_t)$, $t \geq 1$, be defined by the same recursion (2.2)-(2.4) that defines (e_t, F_t), using the same state and action random variables, but with a different initial condition $(\hat{e}_0, \hat{F}_0)$. Let $\mathbf{0}$ denote the zero vector in $\mathbb{R}^n$.

Proposition A.2 *For any two given initial conditions* (e_0, F_0) *and* $(\hat{e}_0, \hat{F}_0)$*,*

$$F_t - \hat{F}_t \overset{a.s.}{\to} 0, \qquad e_t - \hat{e}_t \overset{a.s.}{\to} \mathbf{0}.$$

The proof of the preceding proposition is given in (Yu, 2015, Appendix A.2). The proof uses, among others, convergence theorems for nonnegative supermartingales and random processes (Neveu, 1975).

The property (iii) mentioned in Section 3.1 concerns approximating the trace iterates (e_t, F_t) by truncated traces that depend on a fixed number of the most recent states and actions only. To define the truncated traces, we first express the traces e_t, F_t, by using their definitions (2.2)-(2.4), as

$$F_t = F_0 \cdot \big(\rho_0\gamma_1 \cdots \rho_{t-1}\gamma_t\big) + \sum_{k=1}^{t} i(S_k) \cdot \big(\rho_k\gamma_{k+1} \cdots \rho_{t-1}\gamma_t\big), \tag{A.1}$$

$$e_t = e_0 \cdot \big(\beta_1 \cdots \beta_t\big) + \sum_{k=1}^{t} M_k \cdot \phi(S_k) \cdot \big(\beta_{k+1} \cdots \beta_t\big), \tag{A.2}$$

where $\beta_k = \rho_{k-1}\gamma_k\lambda_k$ (introduced to simplify notation), and

$$M_k = \lambda_k \, i(S_k) + (1 - \lambda_k)\, F_k.$$

We consider now the truncated traces $Y_{t,K} = (\tilde{e}_{t,K}, \tilde{F}_{t,K})$, defined for each integer $K \geq 1$ as

$$Y_{t,K} = (e_t, F_t) \quad \text{for } \ t \leq K,$$

and for $t \geq K + 1$,

$$\tilde{F}_{t,K} = \sum_{k=t-K}^{t} i(S_k) \cdot \big(\rho_k\gamma_{k+1} \cdots \rho_{t-1}\gamma_t\big), \tag{A.3}$$

$$\tilde{M}_{t,K} = \lambda_t \, i(S_t) + (1 - \lambda_t)\tilde{F}_{t,K}, \tag{A.4}$$

$$\tilde{e}_{t,K} = \sum_{k=t-K}^{t} \tilde{M}_{k,K} \cdot \phi(S_k) \cdot \big(\beta_{k+1} \cdots \beta_t\big). \tag{A.5}$$

Denote the original traces by $Y_t = (e_t, F_t)$ (which can be expressed as in Eqs. (A.1)-(A.2)). We have the following result, in which the notation "$L_K \downarrow 0$" means that L_K decreases monotonically to 0 as $K \to \infty$, and in which $Z_0 = (S_0, A_0, e_0, F_0)$ as we recall:

Proposition A.3

(i) *For any given initial $Y_0 = (e_0, F_0)$, there exist constants $L_K, K \geq 1$, with $L_K \downarrow 0$, such that*

$$\mathbb{E}\left[\left\|Y_t - Y_{t,K}\right\|\right] \leq L_K, \qquad \forall\, t \geq 0.$$

(ii) *There exist constants $L_K, K \geq 1$, independent of the given initial value of Z_0, such that $L_K \downarrow 0$ and*

$$\mathbb{E}\left[\left\|Y_{t,K'} - Y_{t,K}\right\|\right] \leq L_K, \qquad \forall\, K' \geq K,\ t > 2K'.$$

The proof of Prop. A.3 can be found in (Yu, 2015, Appendix A.2). We use this proposition subsequently to prove Theorem 3.1: it allows us to work with simple finite-space Markov chains, instead of working with the infinite-space Markov chain $\{Z_t\}$ directly.

Before we proceed further, let us make another remark.

Remark A.1 (On the behavior of trace iterates) From the properties of $\{(e_t, F_t)\}$ given above and the ergodicity of the Markov chain $\{(S_t, A_t, e_t, F_t)\}$ shown in Theorem 3.2, we see that these trace iterates are well-behaved. On the other hand, like in regular off-policy algorithms, these iterates can be unbounded almost surely and their variances can grow to infinity with time. There are no contradictions here. To illustrate this point, let us consider a simple example with just 1 state and 2 actions, $\mathcal{S} = \{1\}, \mathcal{A} = \{a_1, a_2\}$, where all actions result in a self-transition at state 1. Let $\pi(a_1 \,|\, 1) = 1$ for the target policy π, and let $\pi^o(a_1 \,|\, 1) = q < 1$ for the behavior policy π^o. Let the discount factor be a constant $\gamma < 1$. Then for all t,

$$\mathbb{E}[\, \gamma_t^2 \rho_{t-1}^2 \mid \mathcal{F}_{t-1}\,] = \gamma^2/q.$$

Suppose $\gamma^2/q > 1$. Then even with $i(1) = 0$, if $F_0 > 0$, the definition $F_t = \gamma_t \rho_{t-1} F_{t-1}$ implies that

$$\mathbb{E}[\, F_t^2 \,] = \mathbb{E}\big[\, \mathbb{E}[\, \gamma_t^2 \rho_{t-1}^2 \mid \mathcal{F}_{t-1}\,] \cdot F_{t-1}^2 \big] = (\gamma^2/q)^t \cdot F_0^2 \to \infty,$$

yet since $i(1) = 0$, $\{F_t\}$ is also a supermartingale converging to 0 a.s. (Yu, 2015, Lemma A.1). For the case $i(1) > 0$, again $\mathbb{E}[\, F_t^2 \,] \to \infty$ if $\gamma^2/q > 1$, and by (Yu, 2012, Prop. 3.1) the sequence $\{F_t\}$ is almost surely unbounded if $\gamma/q > 1$, yet $\{F_t\}$ is bounded in probability in the sense described by Prop. A.1.

As mentioned earlier in Remark 2.2, it can be desirable to restrict the behavior policy so that the variances of the trace iterates do not grow to infinity. In the simple example above, this can be easily arranged. In the general case, however, if the state-dependent discount factor $\gamma(\cdot)$ can take the value 1 for some states, then without knowledge of the MDP model, to sufficiently restrict the behavior policy seems to be a difficult task.

A.2. Proof of Theorem 3.1

For convenience, we restate Theorem 3.1 here. Recall that the theorem concerns the recursion

$$G_{t+1} = (1 - \alpha_t)\, G_t + \alpha_t\, h(Y_t, S_t, A_t, S_{t+1}),$$

where $Y_t = (e_t, F_t)$, and the function h is Lipschitz continuous in y: for some constant L_h,

$$\big\|h(y,s,a,s') - h(\hat{y},s,a,s')\big\| \le L_h \|y - \hat{y}\|, \quad \forall\, y, \hat{y} \in \mathbb{R}^{n+1},\ \forall\,(s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}.$$

Theorem 3.1 (L^1**-convergence of** $\{G_t\}$) *Let h be a vector-valued function satisfying the Lipschitz condition (3.1), and let $\{G_t\}$ be defined by the recursion (3.2), using the process $\{Z_t\}$. Then under Assumptions 2.1, 2.3, there exists a constant vector G^* (independent of the stepsizes) such that for any given initial $Y_0 = (e_0, F_0)$ and G_0, $\lim_{t\to\infty} \mathbb{E}\big[\|G_t - G^*\|\big] = 0$.*

Proof The proof proceeds in three steps:
(i) For each $K \ge 1$, we consider the truncated traces $Y_{t,K} = (\tilde{e}_{t,K}, \tilde{F}_{t,K})$, $t = 0, 1, \ldots$, defined by Eqs. (A.3)-(A.5). Correspondingly, we define iterates $\tilde{G}_{0,K} = G_0$ and

$$\tilde{G}_{t+1,K} = (1-\alpha_t)\,\tilde{G}_{t,K} + \alpha_t\, h(Y_{t,K}, S_t, A_t, S_{t+1}).$$

For each t, $Y_{t,K}$ is a function of $(S_{t-2K}, A_{t-2K}, \ldots, S_t)$, so $h(Y_{t,K}, S_t, A_t, S_{t+1})$ can be viewed as a function of $X_t = (S_{t-2K}, A_{t-2K}, \ldots, S_{t+1})$, where $\{X_t\}$ is a finite state Markov chain with a single recurrent class by Assumption 2.1(ii). Then, with $\mathbb{E}_0$ denoting the expectation under the stationary distribution of the Markov chain $\{(S_t, A_t)\}$, we have, by a result from stochastic approximation theory (Borkar, 2008, Chap. 6, Theorem 7 and Cor. 8), that under Assumption 2.3 on the stepsizes,

$$\tilde{G}_{t,K} \overset{a.s.}{\to} G^*_K, \qquad \text{where } \ G^*_K = \mathbb{E}_0\big[\,h(Y_{k,K}, S_k, A_k, S_{k+1})\,\big] \ \ \forall\, k > 2K. \tag{A.6}$$

Clearly, the vector G^*_K does not depend on the initial condition (Y_0, G_0) and the stepsizes $\{\alpha_t\}$. Since for all t, $\|\tilde{G}_{t,K}\| \le L$ for some constant $L < \infty$, we also have by the bounded convergence theorem

$$\lim_{t\to\infty} \mathbb{E}\big[\|\tilde{G}_{t,K} - G^*_K\|\big] = 0. \tag{A.7}$$

(ii) We show that as $K \to \infty$, G^*_K converges to some vector G^*. For any $K' > K$, using the Lipschitz property of h and Prop. A.3(ii), we have that for $k > 2K'$,

$$\begin{aligned}\big\|G^*_{K'} - G^*_K\big\| &= \big\|\mathbb{E}_0\big[h\big(Y_{k,K'}, S_k, A_k, S_{k+1}\big) - h\big(Y_{k,K}, S_k, A_k, S_{k+1}\big)\big]\big\| \\ &\le L_h\, \mathbb{E}_0\big[\|Y_{k,K'} - Y_{k,K}\|\big] \le L_h\, L_K,\end{aligned}$$

where L_K is some constant with $L_K \downarrow 0$ as $K \to \infty$. This shows that $\{G^*_K\}$ is a Cauchy sequence and hence converges to some G^*.
(iii) We establish the theorem by bounding the differences between G_t and $\tilde{G}_{t,K}$ for an increasing K. For each K,

$$\limsup_{t\to\infty} \mathbb{E}\big[\|G_t - G^*\|\big] \le \limsup_{t\to\infty} \mathbb{E}\big[\|G_t - \tilde{G}_{t,K}\|\big] + \limsup_{t\to\infty} \mathbb{E}\big[\|\tilde{G}_{t,K} - G^*_K\|\big] + \big\|G^*_K - G^*\big\|.$$

In the right-hand side, the second term equals 0 by Eq. (A.7), and the last term converges to 0 as $K \to \infty$, as we just showed in step (ii). Consider now the first term. Since

$$G_{t+1} - \tilde{G}_{t+1,K} = (1-\alpha_t)\big(G_t - \tilde{G}_{t,K}\big) + \alpha_t\big(h(Y_t, S_t, A_t, S_{t+1}) - h(Y_{t,K}, S_t, A_t, S_{t+1})\big)$$

and $\big\| h(Y_t, S_t, A_t, S_{t+1}) - h(Y_{t,K}, S_t, A_t, S_{t+1}) \big\| \leq L_h \|Y_t - Y_{t,K}\|$ by the Lipschitz property of h, we have

$$\begin{aligned}\mathbb{E}\big[\big\|G_{t+1} - \tilde{G}_{t+1,K}\big\|\big] &\leq (1-\alpha_t)\mathbb{E}\big[\big\|G_t - \tilde{G}_{t,K}\big\|\big] + \alpha_t L_h \mathbb{E}\big[\|Y_t - Y_{t,K}\|\big] \\ &\leq (1-\alpha_t)\mathbb{E}\big[\big\|G_t - \tilde{G}_{t,K}\big\|\big] + \alpha_t L_h L_K, \end{aligned} \tag{A.8}$$

where the second inequality follows from Prop. A.3(i), which gives the constants $L_K, K \geq 1$, with $L_K \downarrow 0$. For each K, in view of Assumption 2.3 on the stepsize, the inequality (A.8) implies that

$$\limsup_{t\to\infty} \mathbb{E}\big[\big\|G_t - \tilde{G}_{t,K}\big\|\big] \leq L_h L_K.$$

Then, since $L_K \downarrow 0$, letting K go to infinity in the right-hand side of the preceding inequality, it follows that $\lim_{t\to\infty} \mathbb{E}\big[\big\|G_t - G^*\big\|\big] = 0$. ■

A.3. Proof of Theorem 2.1 on the Convergence of ELSTD(λ)

The proof proceeds by calculating the limit G^* in Theorem 3.1 for the two functions h_1, h_2 in Eq. (3.3): with $y = (e, F) \in \mathbb{R}^{n+1}$,

$$h_1(y, s, a, s') = e \cdot \rho(s,a) \left(\gamma(s')\phi(s')^\top - \phi(s)^\top\right), \quad h_2(y, s, a, s') = e \cdot \rho(s,a)\, r(s,a,s'),$$

which are associated with the ELSTD(λ) iterates C_t, b_t, respectively. Specifically, based on the proof of Theorem 3.1, we first calculate for each K, the limit G^*_K given in Eq. (A.6), which is associated with the truncated traces $(\tilde{e}_{t,K}, \tilde{F}_{t,K})$. We then take K to ∞ to get the expression of G^* since $G^* = \lim_{K\to\infty} G^*_K$, as shown in the step (ii) of the proof of Theorem 3.1. The details of this calculation are given in our arXiv report (2015, Appendix A.5); Lemma A.4 therein establishes that

$$G^* = C \quad \text{for } h = h_1; \qquad G^* = b \quad \text{for } h = h_2. \tag{A.9}$$

Then with $h = h_1$, Theorem 3.1 yields the L^1-convergence of $\{C_t\}$ to C, and Theorem 3.3 yields $C_t \overset{a.s.}{\to} C$ for stepsizes $\alpha_t = 1/(t+1)$.

For the iterates $\{b_t\}$ [cf. Eq. (2.11)], we also need to take care of the noise in the rewards R_t, by using Prop. 3.1. Specifically, with $W_0 = \mathbf{0}$, let

$$\omega_{t+1} = R_t - r(S_t, A_t, S_{t+1}), \qquad W_{t+1} = (1-\alpha_t)\, W_t + \alpha_t\, e_t\, \rho_t \cdot \omega_{t+1}, \qquad t \geq 0,$$

[cf. Eq. (3.4)]. By definition,

$$b_{t+1} = (1-\alpha_t)\, b_t + \alpha_t\, e_t \cdot \rho_t\, R_t = (1-\alpha_t)\, b_t + \alpha_t\, e_t \cdot \rho_t \left(r(S_t, A_t, S_{t+1}) + \omega_{t+1}\right),$$

so the iteration for $\{b_t\}$ can be equivalently expressed as

$$b_{t+1} = G_{t+1} + W_{t+1},$$

where G_{t+1} is given by the recursion (3.2) with $h = h_2$ and $G_0 = b_0$, and W_{t+1} is as defined above. Then by Theorem 3.1, Eq. (A.9) and Prop. 3.1(i), we have

$$\lim_{t\to\infty} \mathbb{E}\big[\big\|b_t - b\big\|\big] \leq \lim_{t\to\infty} \mathbb{E}\big[\big\|G_t - G^*\big\|\big] + \lim_{t\to\infty} \mathbb{E}\big[\big\|W_t\big\|\big] = 0.$$

This proves the L^1-convergence of $\{b_t\}$ to b. Similarly, its a.s. convergence in the second part of Theorem 2.1 follows from Theorem 3.3, Eq. (A.9) and Prop. 3.1(ii) as

$$G_t \overset{a.s.}{\to} G^* = b \text{ and } W_t \overset{a.s.}{\to} \mathbf{0} \qquad \Longrightarrow \qquad b_t = G_t + W_t \overset{a.s.}{\to} b.$$

This completes the proof of Theorem 2.1.

Appendix B. Proofs for Section 4

In this appendix we prove Lemma 4.1 and Theorem 4.1 for the constrained ETD(λ) algorithm (4.5). We will restate both theorems for convenience.

Recall that the constrained ETD(λ) calculates θ_t, $t \geq 0$, all restricted to be in a closed ball with radius r, $B = \{\theta \in \mathbb{R}^n \mid \|\theta\|_2 \leq r\}$, according to

$$\theta_{t+1} = \Pi_B\Big(\theta_t + \alpha_t\, h(\theta_t, \xi_t) + \alpha_t\, e_t \cdot \tilde{\omega}_{t+1}\Big),$$

where $\tilde{\omega}_{t+1} = \rho_t\big(R_t - r(S_t, A_t, S_{t+1})\big)$ is noise, $\xi_t = (e_t, S_t, A_t, S_{t+1})$, and the function h is given by Eq. (4.3) as

$$h(\theta, \xi) = e \cdot \rho(s, a)\, \big(r(s, a, s') + \gamma(s')\, \phi(s')^\top \theta - \phi(s)^\top \theta\big), \quad \text{for } \xi = (e, s, a, s').$$

The "mean ODE" associated with this algorithm is the projected ODE (4.6):

$$\dot{x} = \bar{h}(x) + z, \qquad z \in -\mathcal{N}_B(x),$$

where $\bar{h}(x) = Cx + b$, $\mathcal{N}_B(x)$ is the normal cone of B at x, and z is the boundary reflection term that keeps the solution in B (Kushner and Yin, 2003). The solution of $\bar{h}(x) = 0$ is denoted θ^*; i.e., $\theta^* = -C^{-1}b$.

Lemma 4.1 *Let $c > 0$ be such that $x^\top C x \leq -c\|x\|_2^2$ for all $x \in \mathbb{R}^n$. Suppose B has a radius $r > \|b\|_2/c$. Then θ^* lies in the interior of B, and the only solution $x(t), t \in (-\infty, +\infty)$, of the projected ODE (4.6) in B is $x(\cdot) \equiv \theta^*$.*

Proof By the definition of θ^*, $C\theta^* + b = 0$. Therefore,

$$0 = \langle \theta^*, C\theta^* + b\rangle = \langle \theta^*, C\theta^*\rangle + \langle \theta^*, b\rangle \leq -c\|\theta^*\|_2^2 + \|b\|_2\|\theta^*\|_2,$$

which implies $\|\theta^*\|_2 \leq b\|_2/c < r$, i.e., θ^* lies in the interior of B.

For a point x on the boundary of B, $\|x\|_2 = r$ and the normal cone $\mathcal{N}_B(x) = \{ax \mid a \geq 0\}$. Since $r > \|b\|_2/c$, we have

$$\langle x, \bar{h}(x)\rangle = \langle x, Cx\rangle + \langle x, b\rangle \leq -c\|x\|_2^2 + \|x\|_2\|b\|_2 = r\,(-c\,r + \|b\|_2) < 0.$$

This shows that for any x on the boundary of B, $\bar{h}(x)$ points inside B and hence at x, the boundary reflection term $z \in -\mathcal{N}_B(x)$ that keeps the solution in B is the zero vector. Consequently, any solution of the projected ODE (4.6) in B is a solution of the ODE (4.4), which is $x(\cdot) \equiv \theta^*$. ■

Next we prove Theorem 4.1.

Theorem 4.1 (Almost sure convergence of constrained ETD(λ)) *Let Assumptions 2.1-2.3 hold. Let $\{\theta_t\}$ be the sequence generated by the constrained ETD(λ) algorithm (4.5) with stepsizes satisfying $\alpha_t = O(1/t)$ and $\frac{\alpha_t - \alpha_{t+1}}{\alpha_t} = O(1/t)$, and with the radius r of B exceeding the threshold given in Lemma 4.1. Then, for any given initial (e_0, F_0, θ_0), $\theta_t \overset{a.s.}{\to} \theta^*$.*

Proof The desired conclusions will follow immediately from (Kushner and Yin, 2003, Theorem 6.1.1) and Lemma 4.1, if we can show that the conditions of (Kushner and Yin, 2003, Theorem 6.1.1) are met. Relevant here are the conditions A.6.1.1-A.6.1.4 and A.6.1.6-A.6.1.7 in (Kushner and Yin, 2003, p. 165). We first adapt these six conditions to our problem, and by using stronger forms of the conditions A.6.1.6-A.6.1.7 given in (Kushner and Yin, 2003, Eq. (6.1.10), p. 166), we obtain the conditions (i)-(vi) below.

The first two conditions are for the functions $h, \bar{h}$ [cf. Eqs. (4.3), (4.4)] and the noise $\{\tilde{\omega}_t\}$:

(i) $\sup_{t \geq 0} \mathbb{E}\big[\| h(\theta_t, \xi_t) + e_t \cdot \tilde{\omega}_{t+1} \|\big] < \infty$.

(ii) $\bar{h}(\theta)$ is continuous, and $h(\theta, \xi)$ is continuous in θ for each ξ.

Condition (i) is satisfied here. Indeed, we have $\sup_{t \geq 0} \mathbb{E}\big[\| h(\theta_t, \xi_t) \|\big] < \infty$, in view of Prop. A.1, the Lipschitz continuity of h in e, and the fact that $\|\theta_t\|_2 \leq r$ for all t by the definition of the constrained algorithm. Since the rewards R_t have bounded variances by assumption and the noise variable $\tilde{\omega}_{t+1} = \rho_t\big(R_t - r(S_t, A_t, S_{t+1})\big)$ by definition, we can bound $\mathbb{E}\big[|\tilde{\omega}_{t+1}| \mid \mathcal{F}_t\big]$ by some constant for all t, where $\mathcal{F}_t = \sigma(S_0, A_0, \ldots, S_{t+1})$, and consequently, we also have $\sup_{t \geq 0} \mathbb{E}\big[\| e_t \cdot \tilde{\omega}_{t+1} \|\big] < \infty$ by Prop. A.1. Hence condition (i) holds. Condition (ii) is also clearly satisfied here.

The four remaining conditions to be introduced are of the same type and relate to the asymptotic rate of change conditions introduced by (Kushner and Clark, 1978). These conditions can guarantee that the effects caused by the noises $\tilde{\omega}_{t+1}$ or by the discrepancies between h and $\bar{h}$ asymptotically "average out" so that the desired convergence can take place.

For any real $T' > 0$, define integer $m(T') = \min\{t \geq 0 \mid \sum_{k=0}^{t} \alpha_k > T'\}$. Conditions (iii)-(vi) below are required to hold for each $a \geq 0$ and some $T > 0$ (here a and T are real numbers):

(iii) For each θ,

$$\lim_{t \to \infty} P\left\{ \sup_{j \geq t} \max_{0 \leq T' \leq T} \left\| \sum_{k=m(jT)}^{m(jT+T')-1} \alpha_k \Big(h(\theta, \xi_k) - \bar{h}(\theta) \Big) \right\| \geq a \right\} = 0. \tag{B.1}$$

(iv)

$$\lim_{t \to \infty} P\left\{ \sup_{j \geq t} \max_{0 \leq T' \leq T} \left\| \sum_{k=m(jT)}^{m(jT+T')-1} \alpha_k \, e_k \cdot \tilde{\omega}_{k+1} \right\| \geq a \right\} = 0. \tag{B.2}$$

(v) There exist nonnegative measurable functions $g_1(\theta), g_2(\xi)$ such that

$$\|h(\theta, \xi)\| \leq g_1(\theta) \, g_2(\xi),$$

where g_1 is bounded on each bounded set of θ, and g_2 satisfies that $\sup_{t \geq 0} \mathbb{E}\big[g_2(\xi_t)\big] < \infty$ and

$$\lim_{t \to \infty} P\left\{ \sup_{j \geq t} \max_{0 \leq T' \leq T} \left| \sum_{k=m(jT)}^{m(jT+T')-1} \alpha_k \Big(g_2(\xi_k) - \mathbb{E}\big[g_2(\xi_k)\big] \Big) \right| \geq a \right\} = 0. \tag{B.3}$$

(vi) There exist nonnegative measurable functions $g_3(\theta), g_4(\xi)$ such that for each θ, θ',

$$\|h(\theta,\xi) - h(\theta',\xi)\| \leq g_3(\theta - \theta')\, g_4(\xi),$$

where g_3 is bounded on each bounded set of θ, with $g_3(\theta) \to 0$ as $\theta \to 0$, and g_4 satisfies that $\sup_{t \geq 0} \mathbb{E}\big[g_4(\xi_t)\big] < \infty$ and

$$\lim_{t\to\infty} P\left\{ \sup_{j \geq t} \max_{0 \leq T' \leq T} \left| \sum_{k=m(jT)}^{m(jT+T')-1} \alpha_k \Big(g_4(\xi_k) - \mathbb{E}\big[g_4(\xi_k)\big] \Big) \right| \geq a \right\} = 0. \tag{B.4}$$

One method given in (Kushner and Yin, 2003, Chap. 6.2, p. 170-171) of verifying the conditions (B.1)-(B.4) above is to show that a strong law of large numbers hold for the processes involved. In particular, let ψ_k represent $h(\theta, \xi_k) - \bar{h}(\theta)$ for condition (iii), $e_k \cdot \tilde{\omega}_{k+1}$ for condition (iv), $g_2(\xi_k) - \mathbb{E}\big[g_2(\xi_k)\big]$ for condition (v), and $g_4(\xi_k) - \mathbb{E}\big[g_4(\xi_k)\big]$ for condition (vi). If

$$\frac{1}{t+1} \sum_{k=0}^{t} \psi_k \overset{a.s.}{\to} 0 \tag{B.5}$$

for the respective $\{\psi_k\}$, then the conditions (B.1)-(B.4) hold for stepsizes satisfying $\alpha_t = O(1/t)$ and $\frac{\alpha_t - \alpha_{t+1}}{\alpha_t} = O(1/t)$ (see Kushner and Yin, 2003, Example 6.1, p. 171).

We now apply the convergence results given earlier in this paper to show that the desired convergence (B.5) holds for the processes involved in conditions (iii)-(vi). In particular, for each fixed θ, the almost sure convergence part of Theorem 2.1 implies that

$$\frac{1}{t+1} \sum_{k=0}^{t} h(\theta, \xi_k) \overset{a.s.}{\to} \mathbb{E}_\zeta\big[h(\theta, \xi_0)\big] = \bar{h}(\theta).$$

Thus, condition (iii) holds, as just discussed. By Prop. 3.1(ii), $\frac{1}{t+1}\sum_{k=0}^{t} e_k \cdot \tilde{\omega}_{k+1} \overset{a.s.}{\to} \mathbf{0}$, so condition (iv) is also met.

We verify now conditions (v)-(vi). For condition (v), we take $g_1(\theta) = \|\theta\| + 1$, and we bound the function h by

$$\|h(\theta,\xi)\| \leq \big(\|\theta\| + 1\big)\, g_2(\xi), \qquad \text{where } \ g_2(\xi) = L\|e\|,$$

and $L > 0$ is some constant. (This bound can be verified directly using the expression of h and the fact that the sets $\mathcal{S}$ and $\mathcal{A}$ are finite.) Similarly, for condition (vi), we take $g_3(\theta) = \|\theta\|$, and we bound the change in $h(\theta, \xi)$ in terms of the change in θ as follows: for any $\theta, \theta' \in \mathbb{R}^n$,

$$\big\|h(\theta,\xi) - h(\theta',\xi)\big\| \leq \|\theta - \theta'\|\, g_4(\xi), \qquad \text{where } \ g_4(\xi) = L'\|e\|,$$

and $L' > 0$ is some constant. Now the functions g_2, g_4 are Lipschitz continuous in e. Hence, for $j = 2, 4$, it follows from Prop. A.1 that $\sup_{t\geq 0} \mathbb{E}\big[g_j(\xi_t)\big] < \infty$, and it follows from Theorems 3.3 and 3.1 that

$$\frac{1}{t+1} \sum_{k=0}^{t} g_j(\xi_k) \overset{a.s.}{\to} \mathbb{E}_\zeta\big[g_j(\xi_0)\big], \quad \text{and} \quad \frac{1}{t+1} \sum_{k=0}^{t} \mathbb{E}\big[g_j(\xi_k)\big] \to \mathbb{E}_\zeta\big[g_j(\xi_0)\big], \quad \text{as } t \to \infty.$$

The preceding two relations imply the desired convergence:

$$\frac{1}{t+1} \sum_{k=0}^{t} \Big(g_j(\xi_k) - \mathbb{E}\big[g_j(\xi_k)\big] \Big) \overset{a.s.}{\to} 0, \qquad j = 2, 4.$$

This shows that conditions (v)-(vi) are met.

The theorem now follows by combining (Kushner and Yin, 2003, Theorem 6.1.1) with the characterization of the solution of the projected ODE (4.6) given by Lemma 4.1, using the fact that under Assumptions 2.1 and 2.2, the matrix C is negative definite (Prop. C.1). ■

Appendix C. Negative Definiteness of the Matrix C

In this appendix we prove a necessary and sufficient condition (Prop. C.2 below) for the matrix C associated with ETD(λ) to be negative definite. Recall from Eqs. (2.8)-(2.9) that

$$C = -\Phi^\top \bar{M} (I - P^{\lambda}_{\pi,\gamma}) \, \Phi$$

where Φ is the feature matrix with full column rank, $P^{\lambda}_{\pi,\gamma}$ is a substochatic matrix, and $\bar{M}$ is a nonnegative diagonal matrix with its diagonal, $diag(\bar{M})$, given by

$$diag(\bar{M}) = d^\top_{\pi^o,i} (I - P^{\lambda}_{\pi,\gamma})^{-1}, \qquad d^\top_{\pi^o,i} = \big(d_{\pi^o}(1)\, i(1), \ \ldots, \ d_{\pi^o}(N)\, i(N)\big).$$

Here Assumption 2.1 is in force and ensures that $(I - P^{\lambda}_{\pi,\gamma})^{-1}$ exists and $d_{\pi^o}(s) > 0$ for all $s \in \mathcal{S}$.

The negative definiteness of C is important for the a.s. convergence of ETD(λ). It is known to hold if $i(s) > 0$ for all $s \in \mathcal{S}$ (Sutton et al., 2015). In general, C is always negative semidefinite for nonnegative $i(\cdot)$, and thus C is negative definite whenever it is nonsingular.

In what follows, we first include a proof of the fact just mentioned, for completeness (see Prop. C.1). We then give explicitly a condition on the approximation subspace which we will prove to be equivalent to the nonsingularity/negative definiteness of C (Prop. C.2). We also show, by specializing this subspace condition, that if those states s of interest (i.e., $i(s) > 0$) are represented by features $\phi(s)$ that are rich enough, then C can be made negative definite, without knowledge of the model (See Cor. C.1, Remark C.2). In addition, we discuss the connection of this subspace condition to seminorm projections, and show that when C is nonsingular, the ETD(λ) solution can be viewed as the solution of a projected Bellman equation involving a seminorm projection (see Remark C.1).

C.1. Preliminaries

First, recall that the matrix C is said to be *negative definite* if there exists $c > 0$ such that

$$y^\top C y \leq -c \, \|y\|_2^2, \qquad \forall \, y \in \mathbb{R}^n,$$

and *negative semidefinite* if $c = 0$ in the preceding inequality. The negative definiteness of C is equivalent to that of the symmetric matrix

$$C + C^\top = -\Phi^\top \Big(\bar{M}(I - P^{\lambda}_{\pi,\gamma}) + (I - P^{\lambda}_{\pi,\gamma})^\top \bar{M} \Big) \, \Phi.$$

Similarly to (Sutton, 1988; Sutton et al., 2015), our analysis will focus on the $N \times N$ symmetric matrix

$$G = \bar{M}(I - Q) + (I - Q)^\top \bar{M}$$

for the substochastic matrix $Q = P^\lambda_{\pi,\gamma}$ and the nonnegative diagonal matrix $\bar{M}$ as given above. We will use a theorem from (Varga, 2000, Cor. 1.22, p. 23), according to which a symmetric real matrix with positive diagonal entries is positive definite if it is strictly diagonally dominant or irreducibly diagonally dominant. Note that by definition, G is *irreducibly diagonally dominant* if G is irreducible[17] and satisfies the following diagonally dominant conditions for every row of G, with strict inequality holding for at least one row:

$$|G_{ss}| \geq \sum_{\bar{s} \neq s} |G_{s\bar{s}}|, \qquad s = 1, \ldots, N,$$

whereas G is *strictly diagonally dominant* if it satisfies the above inequalities strictly for all rows.

We now give a proof of the fact about the relation between the nonsingularity and the negative definiteness of C mentioned at the beginning. This result is due to (Sutton et al., 2015).

Regarding notation, in the proofs below, for $v \in \mathbb{R}^N$, we write $v(s)$ for the s-th entry of v, and for an expression H that results in a vector in $\mathbb{R}^N$, we write $(H)(s)$ for the s-th entry of that vector. For an expression H that results in an $N \times N$ matrix, we write $[H]_{s\bar{s}}$ for its $(s, \bar{s})$-th element. We write $\mathbf{0}$ for a zero vector in any Euclidean space.

Proposition C.1 *Let Assumption 2.1 hold. Then, C is negative definite if C is nonsingular.*

Proof We show first that if $i(s) > 0$ for all $s \in \mathcal{S}$, then G is strictly diagonally dominant, and hence positive definite; and that if $i(s) \geq 0$ for all $s \in \mathcal{S}$, then G is positive semidefinite.

Let $\mathcal{J} = \{s \in \mathcal{S} \mid i(s) = 0\}$. Suppose $\mathcal{J} = \emptyset$. By definition $\bar{M}_{ss} = \big(d^\top_{\pi^o,i}(I - Q)^{-1}\big)(s)$. Using this together with the fact that Q is substochastic, by a direct calculation as in (Sutton et al., 2015), we have that for each $s \in \mathcal{S}$,

$$\begin{aligned} G_{ss} - \sum_{\bar{s} \neq s} |G_{s\bar{s}}| &= \bar{M}_{ss} \cdot \left(1 - \sum_{\bar{s}=1}^{N} Q_{s\bar{s}}\right) + \sum_{\bar{s}=1}^{N} \bar{M}_{\bar{s}\bar{s}} \cdot \big[I - Q\big]_{\bar{s}s} && \text{(C.1)} \\ &\geq 0 + \big(d^\top_{\pi^o,i}(I - Q)^{-1} \cdot (I - Q)\big)(s) \\ &= 0 + d_{\pi^o,i}(s) && \text{(C.2)} \\ &> 0, \end{aligned}$$

where in the last strict inequality, we used the fact that $i(s) > 0$ implies $d_{\pi^o,i}(s) > 0$ under Assumption 2.1(ii). This shows that G is strictly diagonally dominant with positive diagonal entries, and hence positive definite by (Varga, 2000, Cor. 1.22).

Consider now the case $\mathcal{J} \neq \emptyset$. For all $s \in \mathcal{J}$, perturb $i(s)$ to $\delta > 0$, and denote by G_δ the matrix G corresponding to the perturbed $i(\cdot)$. Then G_δ is positive definite by the preceding proof. So for the original G, by continuity, $G = \lim_{\delta \to 0} G_\delta$ is positive semidefinite. It then follows that the matrix $\Phi^\top G \Phi = -C - C^\top$ is positive semidefinite. Hence C is negative semidefinite; but C is nonsingular by assumption, so C must be negative definite. ∎

[17] A symmetric matrix G is *irreducible* if it corresponds to a connected (undirected) graph when the indices are viewed as the nodes of the graph, and the nonzero entries of G are viewed as edges of the graph.

C.2. Main Result

We now give the main result of this section. It expresses the nonsingularity condition on C explicitly in terms of a condition on the approximation subspace E (the column space of Φ).

Proposition C.2 *Let Assumption 2.1 hold, and let $\mathcal{J}_0 = \{s \in \mathcal{S} \mid \bar{M}_{ss} = 0\}$. Suppose the approximation subspace $E \subset \mathbb{R}^N$ is such that*

$$v \in E \;\; \textit{and} \;\; v(s) = 0, \, \forall \, s \notin \mathcal{J}_0 \qquad \Longrightarrow \qquad v = \mathbf{0}. \tag{C.3}$$

Then the matrix C is negative definite. Furthermore, C is nonsingular if and only if the condition (C.3) holds.

The corollary below gives a sufficient condition (C.4) for C being negative definite, which can be fulfilled without knowledge of the model, as we will elaborate in Remark C.2. This corollary is a direct consequence of the preceding proposition, and follows from the observation that since $i(s) > 0$ implies $\bar{M}_{ss} > 0$, the condition (C.4) implies the condition (C.3) in Prop. C.2.

Corollary C.1 *Let Assumption 2.1 hold, and let $\mathcal{J} = \{s \in \mathcal{S} \mid i(s) = 0\}$. Suppose the approximation subspace $E \subset \mathbb{R}^N$ is such that*

$$v \in E \;\; \textit{and} \;\; v(s) = 0, \, \forall \, s \notin \mathcal{J} \qquad \Longrightarrow \qquad v = \mathbf{0}. \tag{C.4}$$

Then the matrix C is negative definite.

We now proceed to prove Prop. C.2. Roughly speaking, the method of proof is to decompose the matrix G into irreducible diagonal blocks and use, among others, the theorem (Varga, 2000, Cor. 1.22, p. 23) on irreducibly diagonally dominant matrices mentioned earlier.

In the two technical lemmas that follow, we let the matrix G and the nonnegative diagonal matrix $\bar{M}$ take a slightly more general form:

$$G = \bar{M}(I - Q) + \big(\bar{M}(I - Q)\big)^\top, \qquad \mathit{diag}(\bar{M}) = d_{\pi^o,i}^\top (I - Q)^{-1},$$

where Q is a substochastic matrix (not necessarily $P^{\lambda}_{\pi,\gamma}$), and $d_{\pi^o,i}$ is a nonnegative vector (for notational simplicity, we keep using $d_{\pi^o,i}$ instead of introducing new notation).

Lemma C.1 *Suppose the matrix $(I - Q)$ is invertible. Then the s-th diagonal entry $\bar{M}_{ss} = 0$ if and only if the s-th row and s-th column of G contain all zeros.*

Proof We have $G = \bar{M}(I - Q) + \big(\bar{M}(I - Q)\big)^\top$. Suppose s is a state with $\bar{M}_{ss} \neq 0$. Then the s-th row of the matrix $M(I - Q)$ is nonzero (because the s-th row of $I - Q$ is nonzero, given that $(I - Q)^{-1}$ exists). The nonzero entries of this row cannot be canceled out by the corresponding entries from the s-th row of $\big(\bar{M}(I - Q)\big)^\top$, because Q is a substochastic matrix and $\bar{M}$ is nonnegative. Therefore, the s-th row of G must also be nonzero. This proves the "if" part.

For the "only if" part, suppose s is a state with $\bar{M}_{ss} = 0$. Then the s-th row of the matrix $\bar{M}(I - Q)$ contains all zeros, so, since $G = \bar{M}(I - Q) + \big(\bar{M}(I - Q)\big)^\top$ and is symmetric, to prove the "only if" part, we only need to show that the s-th column of $\bar{M}(I - Q)$ is a zero column. We prove this by contradiction.

Suppose for some state $\bar{s} \neq s$, the $(\bar{s}, s)$-entry of the matrix $\bar{M}(I - Q)$ is nonzero. Then using the definition of $\bar{M}_{\bar{s}\bar{s}}$, this entry can be expressed as

$$M_{\bar{s}\bar{s}} \cdot \big[I - Q\big]_{\bar{s}s} = -\big(d_{\pi^o,i}^\top (I - Q)^{-1}\big)(\bar{s}) \cdot Q_{\bar{s}s} \neq 0,$$

which, in view of the equality $(I - Q)^{-1} = \sum_{k \geq 0} Q^k$ and the nonnegativity of Q, implies that

$$\big(d_{\pi^o,i}^\top Q^k\big)(\bar{s}) \cdot Q_{\bar{s}s} > 0 \quad \text{for some } k \geq 0.$$

This in turn implies that for the state s,

$$\big(d_{\pi^o,i}^\top Q^k\big)(s) > 0 \quad \text{for some } k \geq 0,$$

and hence

$$\bar{M}_{ss} = \big(d_{\pi^o,i}^\top (I - Q)^{-1}\big)(s) \geq \big(d_{\pi^o,i}^\top Q^k\big)(s) > 0,$$

contradicting the assumption $\bar{M}_{ss} = 0$. Thus the s-th column of $\bar{M}(I - Q)$ must be a zero column. ■

Lemma C.2 *Suppose that the matrix $(I - Q)$ is invertible and the matrix G is irreducible. Then the diagonal entries of $\bar{M}$ must be positive, and G is irreducibly diagonally dominant with positive diagonal entries, and hence positive definite.*

Proof If s is a state with $\bar{M}_{ss} = 0$, by Lemma C.1, the s-th row and s-th column of G would contain all zeros, which cannot happen if G is irreducible. Thus $\bar{M}_{ss} > 0$ for all $s \in \mathcal{S}$.

We have calculated in the proof of Prop. C.1 [cf. Eqs. (C.1)-(C.2)] that for nonnegative $i(\cdot)$,

$$G_{ss} - \sum_{\bar{s} \neq s} |G_{s\bar{s}}| = \bar{M}_{ss} \cdot \left(1 - \sum_{\bar{s}=1}^{N} Q_{s\bar{s}}\right) + \sum_{\bar{s}=1}^{N} \bar{M}_{\bar{s}\bar{s}} \cdot \big[I - Q\big]_{\bar{s}s} \geq 0$$

for all rows s. The strict inequality $G_{ss} - \sum_{\bar{s} \neq s} |G_{s\bar{s}}| > 0$ must hold for some s. To see this, note that the invertibility of $(I - Q)$ implies that $1 - \sum_{\bar{s}=1}^{N} Q_{s\bar{s}} > 0$ for some s, which together with $\bar{M}_{ss} > 0$ implies that the first term in the right-hand side above, $\bar{M}_{ss} \cdot \Big(1 - \sum_{\bar{s}=1}^{N} Q_{s\bar{s}}\Big)$, must be positive for at least one row s, whereas the second term in the right-hand side above equals $d_{\pi^o,i}(s) \geq 0$ [cf. Eqs. (C.1)-(C.2)]. Since G is irreducible by assumption, this proves that G is irreducibly diagonally dominant.

Finally, since Q is substochastic and $(I - Q)^{-1}$ exists, the diagonals of $I - Q$ must be positive. The diagonals of $\bar{M}$ are also positive, as proved earlier. Thus the diagonal entries $G_{ss} > 0$ for all rows s. It then follows from (Varga, 2000, Cor. 1.22) that G is positive definite. ■

We are now ready to prove Prop. C.2. Regarding notation, in the proof, if $G_1, G_2, \ldots, G_L$ are L square matrices (which can have different sizes), we will write $diag\big(G_1, G_2, \ldots, G_L\big)$ for the block-diagonal matrix that has G_k as its k-th diagonal block. However, for a single square matrix G_1, we will keep using $diag(G_1)$ to mean the diagonal of G_1.

Proof of Prop. C.2 By Assumption 2.1(i), $(I - P_\pi \Gamma)^{-1}$ exists, which implies that for the substochastic matrix $Q = P_{\pi,\gamma}^{\lambda}$ [cf. Eq. (2.6)], $(I - Q)^{-1}$ also exists. So the matrices $\bar{M}$, C and G are well defined. By reordering the states if necessary, we can arrange G into a block-diagonal matrix with L blocks,

$$G = diag\Big(G^{(1)},\, \dots,\, G^{(L-1)},\, G^{(L)}\Big) \tag{C.5}$$

such that:

(i) for each $\ell = 1, \dots, L-1$, the ℓth-block $G^{(\ell)}$ is irreducible; and

(ii) the L-th block $G^{(L)}$ is a zero matrix (if G does not have a zero block, we will treat $G^{(L)}$ as a matrix of size zero, and this will not affect the proof below).

Note that by Lemma C.1, the row/column indices associated with the zero block $G^{(L)}$ are exactly those in the set

$$\mathcal{J}_0 = \{s \in \mathcal{S} \mid \bar{M}_{ss} = 0\}.$$

Since the condition (C.3) rules out the case $\mathcal{J}_0 = \mathcal{S}$, G cannot be a zero matrix, so it must have at least one irreducible block.

We prove next that the matrix Q has the following structure, matching the block-diagonal structure of G:

$$Q = \begin{bmatrix} Q^{(1)} & & & & \\ & Q^{(2)} & & & \\ & & \diagdown & & \\ & & & Q^{(L-1)} & \\ * & * & \cdots & * & * \end{bmatrix} \tag{C.6}$$

where the blocks $Q^{(\ell)}$, $\ell \leq L-1$, on the diagonal correspond to the blocks $G^{(\ell)}$, $\ell \leq L-1$, on the diagonal of G, the unmarked blocks contain all zeros, and the $*$-blocks can have both zeros and positive entries.

To prove Eq. (C.6) by contradiction, suppose it does not hold. This means that there must exist two states $s \neq \bar{s}$ with $Q_{s\bar{s}} > 0$, but the entry $Q_{s\bar{s}}$ lies inside an unmarked block of the matrix on the right-hand side of Eq. (C.6). This position of $Q_{s\bar{s}}$ implies $G_{s\bar{s}} = 0$, which is possible only if $\bar{M}_{ss} = 0$ (otherwise, $Q_{s\bar{s}} \neq 0$ would force $G_{s\bar{s}} \neq 0$). But if $\bar{M}_{ss} = 0$, $s \in \mathcal{J}_0$, which is the set of indices associated with the last zero block, as shown earlier. Consequently, the entry $Q_{s\bar{s}}$ cannot lie inside an unmarked block as we assumed. This contradiction proves that Eq. (C.6) must hold.

From the structure of Q shown in (C.6), it follows that $(I-Q)^{-1}$ has the same structure:

$$\big(I-Q\big)^{-1} = \begin{bmatrix} \big(I-Q^{(1)}\big)^{-1} & & & & \\ & \big(I-Q^{(2)}\big)^{-1} & & & \\ & & \diagdown & & \\ & & & \big(I-Q^{(L-1)}\big)^{-1} & \\ * & * & \cdots & * & * \end{bmatrix}. \tag{C.7}$$

Since $G = \bar{M}(I-Q) + (I-Q)^\top \bar{M}$, Eqs. (C.5), (C.6) and (C.7) together imply that for each $\ell \leq L-1$, the matrix $G^{(\ell)}$ can be expressed as

$$G^{(\ell)} = \bar{M}^{(\ell)}\big(I - Q^{(\ell)}\big) + \big(I - Q^{(\ell)}\big)^\top \bar{M}^{(\ell)},$$

where $\bar{M}^{(\ell)}$ is the ℓ-th diagonal block in the corresponding decomposition of $\bar{M}$ as

$$\bar{M} = diag\big(\bar{M}^{(1)}, \dots, \bar{M}^{(L)}\big),$$

and if we decompose the vector $d_{\pi^o,i}$ similarly as $d_{\pi^o,i} = \big(d^{(1)}_{\pi^o,i}, \ldots, d^{(L)}_{\pi^o,i}\big)$, then for each $\ell \leq L-1$, the diagonal block $\bar{M}^{(\ell)}$ has its diagonal entries given by

$$diag\big(\bar{M}^{(\ell)}\big) = \big(d^{(\ell)}_{\pi^o,i}\big)^\top \big(I - Q^{(\ell)}\big)^{-1}, \qquad \ell \leq L-1.$$

In the above expression, we also used the fact $d^{(L)}_{\pi^o,i} = \mathbf{0}$, which is implied by $\bar{M}^{(L)}$ being a zero matrix (which we showed at the beginning of this proof).[18]

We now apply Lemma C.2 to each irreducible block $G^{(\ell)}$, $\ell \leq L-1$ (with $\bar{M} = \bar{M}^{(\ell)}$ and $Q = Q^{(\ell)}$, a substochastic matrix). This yields that each of these $G^{(\ell)}$ is positive definite, and consequently, the block-diagonal matrix

$$\hat{G} = diag\Big(G^{(1)}, \ldots, G^{(L-1)}\Big)$$

is positive definite.

Finally, we prove the statement of the proposition. For the block-diagonal decomposition of G as $G = diag(\hat{G}, G^{(L)})$, write a point $y \in \mathbb{R}^N$ correspondingly as $y = (y_1, y_0)$. I.e., the indices of the components of y_0 are those in $\mathcal{J}_0 = \{s \in \mathcal{S} \mid \bar{M}_{ss} = 0\}$, and the dimension of y_1 is $\hat{N} = N - |\mathcal{J}_0|$.

Since $\hat{G}$ is positive definite, there exists some $c > 0$ such that

$$y_1{}^\top \hat{G}\, y_1 \geq c\, \|y_1\|_2^2, \qquad \forall\, y_1 \in \mathbb{R}^{\hat{N}}. \tag{C.8}$$

Consider a point $y = (y_1, y_0) \in E$ with $y_1 = \mathbf{0}$. Then $y_0 = \mathbf{0}$ by the assumption (C.3). Since E is a subspace, this implies that there exists some constant $\delta > 0$ such that

$$\inf_{y \in E,\, \|y\|_2 = 1} \|y_1\|_2 \geq \delta. \tag{C.9}$$

Using Eqs. (C.8)-(C.9), we have

$$\inf_{y \in E,\, \|y\|_2 = 1} y^\top G\, y = \inf_{y \in E,\, \|y\|_2 = 1} y_1{}^\top \hat{G}\, y_1 \geq \inf_{y \in E,\, \|y\|_2 = 1} c\, \|y_1\|_2^2 \geq c\, \delta^2 > 0. \tag{C.10}$$

Since E is the column space of Φ and Φ has linearly independent columns by definition, the inequality (C.10) establishes that the matrix $\Phi^\top G \Phi = -C - C^\top$ is positive definite, and consequently, C is negative definite.

The preceding proof also shows that C is nonsingular if the condition (C.3) holds. To complete the proof, let us assume that the condition (C.3) does not hold and show that C must be singular. Let $y = (y_1, y_0) \in E$ be such that $y_1 = \mathbf{0}$ and $y_0 \neq \mathbf{0}$. Then since $G^{(L)}$ is a zero block, $y^\top G\, y = 0$, which implies that the matrix $\Phi^\top G \Phi = -C - C^\top$ is singular. If C were nonsingular, then by Prop. C.1, $-C - C^\top$ would be positive definite and hence nonsingular, a contradiction. Therefore, C must be singular. ■

Finally, we make two remarks on the conditions (C.3) and (C.4) in Prop. C.2 and Cor. C.1.

[18] Using the expression $(I - Q)^{-1} = \sum_{k \geq 0} Q^k$, it can be seen from the definition of $\bar{M}_{ss}$ that $\bar{M}_{ss} \geq d_{\pi^o,i}(s)$. Therefore, $\bar{M}_{ss} = 0$ implies that $d_{\pi^o,i}(s) = 0$.

Remark C.1 (Seminorm projection) Using seminorm projections to formulate the projected Bellman equations associated with TD methods is introduced in (Yu and Bertsekas, 2012). There, conditions of the form (C.3) or (C.4) are used to define a projection on the approximation subspace with respect to a seminorm. We can use this formulation here to interpret the solution of ETD(λ) and ELSTD(λ). Specifically, define a weighted Euclidean seminorm $\| \cdot \|_{\bar{M}}$ on $\mathbb{R}^N$, using $diag(\bar{M})$ as the weights, as

$$\|v\|^2_{\bar{M}} = \textstyle\sum_{s \in \mathcal{S}} \bar{M}_{ss} \cdot v(s)^2.$$

Condition (C.3) ensures that the projection $\Pi_{\bar{M}}$ onto E with respect to the seminorm $\| \cdot \|_{\bar{M}}$ is well-defined and has the matrix representation

$$\Pi_{\bar{M}} = \Phi \left(\Phi^\top \bar{M} \Phi \right)^{-1} \Phi^\top \bar{M}$$

(cf. Yu and Bertsekas, 2012, Sec. 2.1). So by Prop. C.2 and the convergence results of this paper, when C is nonsingular, ETD(λ) and ELSTD(λ) solve in the limit the projected Bellman equation

$$v = \Pi_{\bar{M}} \big(r^\lambda_{\pi,\gamma} + P^\lambda_{\pi,\gamma} \, v \big).$$

The relation between the solution $v = \Phi\theta^*$ of this equation and the desired value function v_π, in particular, the approximation error, can be analyzed then, using the oblique projection viewpoint (Scherrer, 2010) (for details, see also (Yu and Bertsekas, 2012)).

Remark C.2 (Equivalent conditions in terms of features) The condition (C.3) can be paraphrased in terms of the features $\phi(s)$ as follows:

$$\forall \, s \in \mathcal{S} \text{ with } \bar{M}_{ss} = 0, \quad \phi(s) \in span\big\{ \phi(\bar{s}) \,\big|\, \bar{s} \in \mathcal{S} \text{ and } \bar{M}_{\bar{s}\bar{s}} > 0 \big\}; \tag{C.11}$$

namely, from those states with positive emphasis weights $\bar{M}_{\bar{s}\bar{s}} > 0$, n linearly independent feature vectors can be found. Similarly, the condition (C.4) can be paraphrased as:

$$\forall \, s \in \mathcal{S} \ \text{ with } i(s) = 0, \quad \phi(s) \in span\big\{ \phi(\bar{s}) \,\big|\, \bar{s} \in \mathcal{S} \text{ and } i(\bar{s}) > 0 \big\}; \tag{C.12}$$

namely, from the states with positive interest weights, n linearly independent feature vectors can be found. This shows that even without knowing P_π and $\bar{M}$, by designing a rich enough set of features for states of interest beforehand, we can ensure the sufficient condition (C.4) for the nonsingularity and negative definiteness of the matrix C.

Conditions like (C.11), (C.12) [or equivalently, (C.3), (C.4)] are naturally satisfied in the case where the approximate values of the policy π at certain states (e.g., those states s with $\bar{M}_{ss} = 0$ or $i(s) = 0$) are interpolated or extrapolated from the approximate values of π at some other "representative" states, based on the "proximity" of the former states to the representative ones.