Journal of Machine Learning Research 23 (2022) 1-59 Submitted 8/21; Revised 4/22; Published 5/22

Truncated Emphatic Temporal Difference Methods for Prediction and Control

Shangtong Zhang SHANGTONG.ZHANG@CS.OX.AC.UK
Department of Computer Science
University of Oxford
Wolfson Building, Parks Rd, Oxford, OX1 3QD, UK

Shimon Whiteson SHIMON.WHITESON@CS.OX.AC.UK
Department of Computer Science
University of Oxford
Wolfson Building, Parks Rd, Oxford, OX1 3QD, UK

Editor: Marc Bellemare

Abstract

Emphatic Temporal Difference (TD) methods are a class of off-policy Reinforcement Learning (RL) methods involving the use of followon traces. Despite the theoretical success of emphatic TD methods in addressing the notorious deadly triad of off-policy RL, there are still two open problems. First, followon traces typically suffer from large variance, making them hard to use in practice. Second, though Yu (2015) confirms the asymptotic convergence of some emphatic TD methods for prediction problems, there is still no finite sample analysis for any emphatic TD method for prediction, much less control. In this paper, we address those two open problems simultaneously via using *truncated followon traces* in emphatic TD methods. Unlike the original followon traces, which depend on all previous history, truncated followon traces depend on only finite history, reducing variance and enabling the finite sample analysis of our proposed emphatic TD methods for both prediction and control.

Keywords: off-policy learning, emphatic methods, finite sample analysis, reinforcement learning, approximate value iteration

1. Introduction

Off-policy learning, where an agent learns a policy of interest (target policy) while following a different policy (behavior policy), is arguably one of the most important techniques in Reinforcement Learning (RL, Sutton and Barto 2018). Off-policy learning can improve the sample efficiency (Lin, 1992; Sutton et al., 2011) and safety (Dulac-Arnold et al., 2019) of RL algorithms. However, they can be unstable if combined with *function approximation* and *bootstrapping*, two arguably indispensable ingredients for RL algorithms to work at scale. This instability is known as the notorious deadly triad (Chapter 11 of Sutton and Barto 2018).

Emphatic Temporal Difference methods are a class of off-policy Temporal Difference (TD, Sutton 1988) methods first proposed by Sutton et al. (2016) to address the deadly triad. Compared with gradient TD methods (Sutton et al., 2008, 2009), another class of off-policy TD methods that address the deadly triad, emphatic TD (ETD) methods usually have better asymptotic performance guarantees (Kolter, 2011; Hallak et al., 2016). The key idea of emphatic TD methods is the *followon* trace, a recursively computed scalar depending on all previous history that reweights the naive off-policy TD (Chapter 11.1 of Sutton and Barto 2018) updates, first introduced in ETD(λ) (Sutton et al., 2016). In addition to ETD(λ), variants have been proposed such as ETD(λ, β) (Hallak et al., 2016), which allows for additional bias-variance tradeoff, and NETD (Jiang et al., 2021), which copes with multi-step TD methods like VTrace (Espeholt et al., 2018). Emphatic TD methods have enjoyed both theoretical and empirical success. For example, Yu (2015) confirms the asymptotic convergence of ETD(λ) under general conditions; Jiang et al. (2021) demonstrate state-of-the-art performance of several NETD-based algorithms in certain Arcade Learning Environment (Bellemare et al., 2013) settings. Nonetheless, there are still two open problems.

1. The followon trace can have infinite variance, as demonstrated by Sutton et al. (2016). As a result, Sutton and Barto (2018) report that though ETD(λ) is proven to be convergent, in Baird's counterexample (Baird, 1995), a commonly used benchmark for testing off-policy RL algorithms, "*it is nigh impossible to get consistent results in computational experiments*" (Chapter 11.9 of Sutton and Barto 2018) for ETD(λ). To lower the variance introduced by the followon trace, Hallak et al. (2016) introduce an additional hyperparameter β for bias-variance trade-off in computing the followon trace, resulting in ETD(λ, β). When β is sufficiently small, Hallak et al. (2016) prove that the variance of the followon trace is bounded. Hallak et al. (2016), however, also require β to be sufficiently large such that the expected update of ETD(λ, β) is contractive, which plays a key role in bounding the performance of the fixed point of ETD(λ, β). Unfortunately, there is no guarantee that such a β (i.e., a β that is both sufficiently small and sufficiently large) always exists. Later on, Zhang et al. (2020b) propose to *learn* the expectation of the followon trace directly by employing a second function approximator and use the *learned* followon trace to reweight the naive off-policy TD updates. However, little can be said about the quality of the learned followon trace. It, therefore, remains an open problem to design a theoretically grounded method to reduce the variance introduced by the followon trace.

2. Twenty years after the seminal work Tsitsiklis and Roy (1996) confirming the asymptotic convergence of TD(λ), finite sample analysis of TD methods were obtained for both prediction (Dalal et al., 2018; Lakshminarayanan and Szepesvári, 2018; Bhandari et al., 2018; Srikant and Ying, 2019) and control (Zou et al., 2019). Though Yu (2015) confirms the asymptotic convergence of ETD(λ), we still do not have finite sample analysis for any emphatic TD method even for prediction problems, much less control.

In this paper, we address these two problems simultaneously by using *truncated* followon traces instead of the original followon trace in Section 4 for prediction problems and extend the results to control problems in Sections 5 and 6. Truncated traces are introduced by Yu (2012, 2015, 2017) as an intermediate mathematical tool in proofs to understand the asymptotic behavior of some least-square TD methods (e.g., off-policy LSTD(λ) in Yu 2012, emphatic LSTD(λ) in Yu 2015) and gradient TD methods (e.g., GTD(λ) in Sutton et al.

2009) for prediction. In this paper, we instead use truncated followon traces *algorithmically* as a tool for variance reduction for both prediction and control. Whereas the original followon trace depends on all previous history, the truncated followon trace depends on only *finite* history. Consequently, the variance of truncated followon traces is immediately bounded. We refer to emphatic TD methods that involve this truncated followon traces as *truncated emphatic TD methods*. Moreover, we show that under certain conditions on their length, truncated followon traces maintain all the desirable properties of the original followon trace, enabling us to analyse truncated emphatic TD methods both asymptotically and non-asymptotically, for both prediction and control.

In particular, this paper makes the following contributions. First, we propose the Truncated Emphatic TD algorithm for off-policy prediction and provide both asymptotic and nonasymptotic convergence analysis. This is the first finite sample analysis for emphatic TD methods. Second, we propose the Truncated Emphatic Expected SARSA algorithm for off-policy control and provide both asymptotic and nonasymptotic analysis. This is the first emphatic TD algorithm for off-policy control. Third, we empirically study truncated emphatic TD methods in both synthetic Markov Decision Processes (MDPs) and nonsynthetic control problems, confirming their efficacy in practice.

2. Background

In this paper, all vectors are column. A matrix M (not necessarily symmetric) is said to be positive definite (p.d.) if there exists a constant $\lambda > 0$ such that $x^\top M x \geq \lambda x^\top x$ holds for any x. It is well known that M is p.d. if and only if $M + M^\top$ is p.d. M is negative definite (n.d.) if and only if $-M$ is p.d. For a vector x and a p.d. matrix M, we use $\|x\|_M \doteq \sqrt{x^\top M x}$ to denote the vector norm induced by M. We also use $\|\cdot\|_M$ to denote the corresponding induced matrix norm. We use $\|\cdot\|$ as shorthand for $\|\cdot\|_I$ where I is the identity matrix, i.e., $\|\cdot\|$ is the standard ℓ_2-norm. We use $diag(x)$ to denote a diagonal matrix whose diagonal entry is x and write $\|\cdot\|_x$ as shorthand for $\|\cdot\|_{diag(x)}$ when $diag(x)$ is p.d. We use $\|\cdot\|_\infty$ and $\|\cdot\|_1$ to denote the standard infinity norm and ℓ_1-norm respectively. We use $\langle \cdot, \cdot \rangle$ to denote the inner product in Euclidean spaces, i.e., $\langle x, y \rangle \doteq x^\top y$. We use functions and vectors interchangeably when it does not confuse, e.g., if f is a function from $\mathcal{S}$ to $\mathbb{R}$, we also use f to denote a vector in $\mathbb{R}^{|\mathcal{S}|}$, whose s-th element is $f(s)$.

We consider an infinite horizon MDP with a finite state space $\mathcal{S}$, a finite action space $\mathcal{A}$, a reward function $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, a transition kernel $p : \mathcal{S} \times \mathcal{S} \times \mathcal{A} \to [0, 1]$, an initial state distribution $p_0 : \mathcal{S} \to [0, 1]$, and a discount factor $\gamma \in [0, 1)$. At time step 0, an initial state S_0 is sampled according to p_0. At time step t, an agent at a state S_t takes an action A_t according to $\pi(\cdot|S_t)$, where $\pi : \mathcal{A} \times \mathcal{S} \to [0, 1]$ is the policy being followed by the agent. The agent then receives a reward $R_{t+1} \doteq r(S_t, A_t)$ and proceeds to a successor state S_{t+1} sampled from $p(\cdot|S_t, A_t)$.

The return at time step t is defined as

$$G_t \doteq \sum_{i=1}^{\infty} \gamma^{i-1} R_{t+i},$$

which allows us to define the state value and action value functions respectively as

$$v_\pi(s) \doteq \mathbb{E}\left[G_t | S_t = s, \pi, p\right],$$
$$q_\pi(s, a) \doteq \mathbb{E}\left[G_t | S_t = s, A_t = a, \pi, p\right].$$

The value function v_π is the unique fixed point of the Bellman operator $\mathcal{T}_\pi$:

$$\mathcal{T}_\pi v \doteq r_\pi + \gamma P_\pi v,$$

where $r_\pi \in \mathbb{R}^{|\mathcal{S}|}$ is the reward vector induced by the policy π, i.e., $r_\pi(s) \doteq \sum_a \pi(a|s) r(s, a)$. Prediction and control are two fundamental problems in RL.

2.1 Prediction

The goal of prediction is to estimate the value function of a given policy π, perhaps with the help of parameterized function approximation. In this paper, we consider linear function approximation and assume access to a feature function $x : \mathcal{S} \to \mathbb{R}^K$, which maps a state into a K-dimensional numerical feature. We then use $x(s)^\top w$ as our estimate for $v_\pi(s)$, where $w \in \mathbb{R}^K$ is the parameter vector to be learned. Arguably, one of the most important methods for prediction is TD, which updates w iteratively as

$$\begin{aligned} w_{t+1} &\doteq w_t + \alpha_t (R_{t+1} + \gamma x_{t+1}^\top w_t - x_t^\top w_t) x_t \\ &= w_t + \alpha_t \underbrace{x_t (\gamma x_{t+1}^\top - x_t^\top)}_{M_t} w_t + \alpha_t R_{t+1} x_t, \end{aligned} \tag{1}$$

where $\{\alpha_t\}$ is a sequence of learning rates and x_t is shorthand for $x(S_t)$. The expectation of the update matrix M_t w.r.t. d_π, the invariant state distribution of the chain induced by π, is

$$M \doteq \mathbb{E}_{S_t \sim d_\pi, A_t \sim \pi(\cdot|S_t), S_{t+1} \sim p(\cdot|S_t, A_t)}[M_t] = X^\top D_\pi (\gamma P_\pi - I) X,$$

where $X \in \mathbb{R}^{|\mathcal{S}| \times K}$ is the feature matrix whose s-th row is $x(s)^\top$, $D_\pi \doteq diag(d_\pi)$, and $P_\pi \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{S}|}$ is the state transition matrix under the policy π, i.e.,

$$P_\pi(s, s') \doteq \sum_a \pi(a|s) p(s'|s, a).$$

Tsitsiklis and Roy (1996) prove that M is n.d. under mild conditions. Consequently, standard Ordinary Differential Equation (ODE) based convergence results (e.g., Theorem 2 of Tsitsiklis and Roy 1996, Proposition 4.8[1] of Bertsekas and Tsitsiklis 1996) can be used to show that the iterates $\{w_t\}$ generated by (1) converge almost surely (a.s.).

So far we have focused on the on-policy setting, where the policy to be evaluated is the same as the policy used for action selection during interaction with the environment. In the off-policy setting, those two policies can, however, be different, allowing extra flexibility. We use π to denote the policy to be evaluated (target policy) and μ to denote the policy used for action selection (behavior policy).

[1] For completeness, we include this proposition as Theorem 13 in Section A.1.

Since the action selection is performed according to μ instead of π (i.e., $A_t \sim \mu(\cdot|S_t), R_{t+1} \doteq r(S_t, A_t), S_{t+1} \sim p(\cdot|S_t, A_t)$), we can reweight the update made in (1) by the importance sampling ratio $\rho_t \doteq \frac{\pi(A_t|S_t)}{\mu(A_t|S_t)}$, yielding the following off-policy TD updates:

$$\begin{aligned} w_{t+1} &\doteq w_t + \alpha_t \rho_t (R_{t+1} + \gamma x_{t+1}^\top w_t - x_t^\top w_t) x_t \\ &= w_t + \alpha_t \underbrace{\rho_t x_t (\gamma x_{t+1}^\top - x_t^\top)}_{M_t} w_t + \alpha_t \rho_t R_{t+1} x_t. \end{aligned} \tag{2}$$

The expectation of the update matrix M_t in (2) w.r.t. d_μ, the invariant state distribution of the chain induced by μ, is then

$$M \doteq \mathbb{E}_{S_t \sim d_\mu, A_t \sim \mu(\cdot|S_t), p(\cdot|S_t, A_t)}[M_t] = X^\top D_\mu (\gamma P_\pi - I) X,$$

where $D_\mu \doteq diag(d_\mu)$. Unfortunately, this M is not guaranteed to be n.d. and the possible divergence of (2) is well documented in Baird's counterexample (Baird, 1995).

Sutton et al. (2016) propose ETD(λ) to address this divergence issue. In its simplest form with $\lambda = 0$, ETD(0) further reweights the update in (2) by the *followon trace* F_t:

$$\begin{aligned} F_t &\doteq i(S_t) + \gamma \rho_{t-1} F_{t-1}, \\ w_{t+1} &\doteq w_t + \alpha_t \rho_t F_t (R_{t+1} + \gamma x_{t+1}^\top w_t - x_t^\top w_t) x_t \\ &= w_t + \alpha_t \underbrace{\rho_t F_t x_t (\gamma x_{t+1}^\top - x_t^\top)}_{M_t} w_t + \alpha_t \rho_t F_t R_{t+1} x_t. \end{aligned} \tag{3}$$

where $i : \mathcal{S} \to (0, +\infty)$ is the *interest* function representing the user's preference for different states. The motivation for introducing F_t is to ensure that the corresponding limiting update $M \doteq \lim_{t\to\infty} \mathbb{E}[M_t]$, assuming the limit exists for now, is n.d. such that standard ODE-based convergent results (e.g., Theorem 13) can be used to show convergence. In fact, Sutton et al. (2016) show

$$M = X^\top D_f (\gamma P_\pi - I) X,$$

where $D_f \doteq diag(f)$ with

$$f \doteq (I - \gamma P_\pi^\top)^{-1} D_\mu i. \tag{4}$$

Sutton et al. (2016) prove that this M is n.d. and the convergence of ETD(λ) is later on established by Yu (2015). It is worth mentioning that one important step in computing this M is to show

$$\lim_{t\to\infty} \mathbb{E}\left[F_t | S_t = s\right] = d_\mu(s)^{-1} f(s). \tag{5}$$

2.2 Control

The goal for control is to find an optimal policy π^* such that $v_{\pi^*}(s) \geq v_\pi(s)$ holds for any π and s. Though there can be more than one optimal policy, all of them share the same optimal value function, which is referred to as v_*. One classical approach for finding v_*

is *value iteration* (see, e.g., Puterman 2014). Given an arbitrary vector $v \in \mathbb{R}^{|\mathcal{S}|}$, value iteration updates v iteratively as

$$v_{k+1} \doteq \mathcal{T}_{\pi_{v_k}} v_k,$$

where we use π_{v_k} to denote the greedy policy w.r.t. v_k. Let

$$q_{v_k}(s, a) \doteq r(s, a) + \gamma \sum_{s'} p(s'|s, a) v_{\pi_k}(s'),$$

then at a state s, π_{v_k} selects an action action greedily w.r.t. $q_{v_k}(s, \cdot)$. It is well known (see, e.g., Puterman 2014) that $\lim_{k\to\infty} v_k = v_*$.

With function approximation, we have $v_k \doteq X w_k$, where w_k is the parameter at the k-th iteration. When doing value iteration, however, $\mathcal{T}_{\pi_{v_k}} v_k$ may not lie in the column space of X. Consequently, an additional projection operator is used to project $\mathcal{T}_{\pi_{v_k}} v_k$ back to the column space of X, yielding *approximate value iteration* (De Farias and Van Roy, 2000), which updates v_k as

$$v_{k+1} \doteq \mathcal{H}(v_k) \doteq \Pi_{d_{\pi_{v_k}}} \mathcal{T}_{\pi_{v_k}} v_k, \tag{6}$$

where

$$\Pi_{d_{\pi_{v_k}}} y \doteq X \arg\min_w \|Xw - y\|^2_{d_{\pi_{v_k}}}$$

is the projection operator to the column space of X w.r.t. to the norm induced by the invariant state distribution $d_{\pi_{v_k}}$ under the current policy π_{v_k}. Unfortunately, if π_v is greedy w.r.t. v, De Farias and Van Roy (2000) show that the approximate value iteration operator $\mathcal{H}$ does not necessarily have a fixed point. However, if the policy π_v is continuous in v, e.g., π_v is a softmax policy such that

$$\pi_v(a|s) \doteq \frac{\exp\left(r(s, a) + \gamma \sum_{s'} p(s'|s, a) v(s')\right)}{\sum_{s_0, a_0} \exp\big(r(s_0, a_0) + \gamma \sum_{s_1} p(s_1|s_0, a_0) v(s_1)\big)}, \tag{7}$$

De Farias and Van Roy (2000) show that there exists at least one w_* such that

$$X w_* = \mathcal{H}(X w_*).$$

In RL, one way to implement approximate value iteration incrementally is SARSA (Rummery and Niranjan, 1994), which updates w iteratively as

$$w_{t+1} \doteq w_t + \alpha_t (R_{t+1} + \gamma x(S_{t+1}, A_{t+1})^\top w_t - x(S_t, A_t)^\top w_t) x(S_t, A_t), \tag{8}$$

where we have overloaded x as a function from $\mathcal{S} \times \mathcal{A}$ to $\mathbb{R}^K$ to denote the state-action feature. We then use $x(s, a)^\top w$ as our estimate for the action value function. In the above SARSA update, actions are selected such that $A_t \sim \pi_{w_{t-1}}(\cdot|S_t)$, where $\pi_{w_{t-1}}$ denotes a policy depending on the action value estimate $x(s, a)^\top w_{t-1}$, e.g., a softmax policy

$$\pi_{w_{t-1}}(a|s) \doteq \frac{\exp\big(x(s, a)^\top w_{t-1}\big)}{\sum_{s_0, a_0} \exp\left(x(s_0, a_0)^\top w_{t-1}\right)}.$$

Melo et al. (2008) and Zou et al. (2019) provide asymptotic convergence analysis and finite sample analysis of SARSA respectively, under mild conditions.

3. Open Problems in Emphatic TD Methods

In this section, we discuss in detail two open problems of emphatic TD methods.

First, though ETD(λ) is proven to be convergent, the large variance of F_t makes it hard to use directly. There are several attempts to address this variance. Hallak et al. (2016) propose to replace F_t with $F_{t,\beta}$, which is computed recursively as

$$F_{t,\beta} \doteq i(S_t) + \beta\rho_{t-1}F_{t-1,\beta}, \tag{9}$$

where $\beta \in (0,1)$ is an additional hyperparameter. The resulting ETD(λ, β) then updates $\{w_t\}$ iteratively as

$$w_{t+1} \doteq w_t + \alpha_t F_{t,\beta}\rho_t(R_{t+1} + \gamma x_{t+1}^\top w_t - x_t^\top w_t)x_t. \tag{10}$$

Theorem 1 of Hallak et al. (2016) states that there exists a problem-dependent constant β_{upper} such that $\beta \leq \beta_{\text{upper}}$ implies that the variance of $F_{t,\beta}$ is bounded. Further, Proposition 1 of Hallak et al. (2016) states that there exists a problem-dependent constant β_{lower} such that $\beta \geq \beta_{\text{lower}}$ implies that the expected update corresponding to (10) is contractive, which plays a key role in bounding the performance of the fixed point of (10), assuming (10) converges. Unfortunately, there is no guarantee that $\beta_{\text{lower}} \leq \beta_{\text{upper}}$ always holds, i.e, the desired β does not always exist. Zhang et al. (2020b) instead propose to use a second function approximator to learn the expectation of the followon trace directly. For example, let $x(s)^\top\theta$ be the estimate for the expectation of the followon trace; Zhang et al. (2020b) replace F_t in the ETD update (3) by $x(S_t)^\top\theta$ and update w iteratively as

$$w_{t+1} \doteq w_t + \alpha_t(x_t^\top\theta)\rho_t(R_{t+1} + \gamma x_{t+1}^\top w_t - x_t^\top w_t)x_t.$$

Zhang et al. (2020b) use Gradient Emphasis Learning (GEM) to learn θ. GEM shares the same idea as gradient TD methods. Though Zhang et al. (2020b) confirm the convergence of GEM, like any off-policy gradient TD method, little can be said about the quality of its solution, i.e., the error $\left|x(s)^\top\theta - \lim_{t\to\infty}\mathbb{E}\left[F_t|S_t = s\right]\right|$ can be arbitrarily large as long as the feature matrix X cannot perfectly represent the expected followon trace (Kolter, 2011). Jiang et al. (2021) propose to clip the importance sampling ratio ρ_t when computing the followon trace F_t to reduce variance. However, nothing can be said about the convergence of the resulting algorithm due to the bias introduced by clipping. Despite these attempts, it remains an open problem to reduce the variance of emphatic TD methods introduced by the followon trace in a theoretically grounded way.

Second, the analysis of ETD(λ) in Yu (2015) is only asymptotic. So far no finite sample analysis is available for any emphatic TD method. The finite sample analysis of TD(λ) in Bhandari et al. (2018) cannot be easily extended to ETD(λ). Key to the finite sample analysis of TD(λ) is Lemma 17 of Bhandari et al. (2018), which establishes the boundedness of the eligibility trace used in on-policy TD(λ). This immediately implies the boundedness of the second moments of the eligibility trace, which is a key bound for error terms. However, such boundedness cannot be expected for the followon trace in ETD(λ) since Sutton et al. (2016) already show that the variance of the followon trace can be unbounded. It thus remains an open problem to provide a finite sample analysis for emphatic TD methods for prediction problems. For control problems, the policy usually changes every time step (cf. the SARSA algorithm (8)). Consequently, the induced chain is not stationary. The asymptotic convergence analysis for ETD(λ) in Yu (2015), however, relies on the strong law of large numbers on *stationary* chains. It thus remains unclear whether the asymptotic convergence analysis of Yu (2015) can be extended to the control setting. Hence, providing a finite sample analysis for emphatic TD methods for control problems is even more challenging.

4. Prediction: Truncated Emphatic TD

In this paper, we address the two open problems in Section 3 simultaneously by replacing the original followon trace in emphatic TD methods with truncated followon traces. Assuming $F_{-1} \doteq 0$, the original followon trace F_t in (3) can be expanded as

$$\begin{aligned} F_t &= i_t + \gamma \rho_{t-1} F_{t-1} \\ &= i_t + \gamma \rho_{t-1} i_{t-1} + \gamma^2 \rho_{t-1} \rho_{t-2} F_{t-2} \\ &= i_t + \gamma \rho_{t-1} i_{t-1} + \gamma^2 \rho_{t-1} \rho_{t-2} i_{t-2} + \gamma^3 \rho_{t-1} \rho_{t-2} \rho_{t-3} F_{t-3} \\ &= \dots \\ &= \sum_{j=0}^{t} \gamma^j \rho_{t-j:t-1} i_{t-j}, \end{aligned} \tag{11}$$

where i_t is shorthand for $i(S_t)$ and

$$\rho_{j:k} \doteq \begin{cases} \rho_j \rho_{j+1} \cdots \rho_k & j \leq k \\ 1 & j > k \end{cases}$$

is shorthand for the product of importance sampling ratios. Clearly, F_t depends on all the history from time steps 0 to t. The idea of truncated followon traces, introduced in Yu (2012, 2015, 2017), is, for a fixed length n, to compute the followon trace F_t as if F_{t-n-1} was 0. More specifically, let $F_{t,n}$ be the truncated followon traces of length n; we have

$$F_{t,n} \doteq \begin{cases} \sum_{j=0}^{n} \gamma^j \rho_{t-j:t-1} i_{t-j} & t \geq n \\ F_t & t < n \end{cases}. \tag{12}$$

For example, if $n = 2$, we then compute $F_{t,2}$ for any t as

$$F_{t,2} = i_t + \gamma \rho_{t-1} i_{t-1} + \gamma^2 \rho_{t-1} \rho_{t-2} i_{t-2}.$$

In this paper, we propose to replace F_t with $F_{t,n}$ in emphatic TD methods. Apparently, for a fixed n, the variance of $F_{t,n}$ is guaranteed to be bounded. By contrast, Sutton et al. (2016) show that the variance of F_t can be infinite. We refer to emphatic TD methods using the truncated traces as *truncated emphatic TD methods*. For example, Truncated Emphatic TD is given in Algorithm 1, where we adopt the convention that $i_t = \rho_t = 0$ for any $t < 0$.

To compute $F_{t,n}$, one needs to store $2n$ extra scalars: $\rho_{t-1}, \dots, \rho_{t-n}, i_{t-1}, \dots, i_{t-n}$. Such memory overhead is inevitable even for naive on-policy multi-step TD methods (Chapter

---

Algorithm 1: Truncated Emphatic TD

$S_0 \sim p_0(\cdot)$
$t \leftarrow 0$
while *True* **do**
  Sample $A_t \sim \mu(\cdot|S_t)$
  Execute A_t, get R_{t+1}, S_{t+1}
  $\rho_t \leftarrow \frac{\pi(A_t|S_t)}{\mu(A_t|S_t)}$
  $F_{t,n} \leftarrow 0$
  for $k = 0, \dots, n$ **do**
    $F_{t,n} \leftarrow i_{t-n+k} + \gamma \rho_{t-n+k-1} F_{t,n}$
  end
  $w_{t+1} \leftarrow w_t + \alpha_t F_{t,n} \rho_t (R_{t+1} + \gamma x_{t+1}^\top w_t - x_t^\top w_t) x_t$
  $t \leftarrow t + 1$
end

7.1 of Sutton and Barto 2018). The computation of $F_{t,n}$ can indeed be done incrementally at the cost of maintaining one more extra scalar:

$$\Delta_t \doteq \frac{\rho_t i_t}{\rho_{t-n-1} i_{t-n-1}} \Delta_{t-1},$$
$$F_{t,n} \doteq i_t + \gamma \rho_{t-1} F_{t-1,n} - \Delta_t.$$

Overall, we argue that compared with ETD(0) in Sutton et al. (2016), the additional memory and computational cost of Truncated Emphatic TD is negligible. We now analyze Truncated Emphatic TD with the following assumptions:

Assumption 4.1 *The Markov chain induced by the behavior policy μ is ergodic.*

Assumption 4.2 $\mu(a|s) > 0$ *holds for any* (s, a).

Assumption 4.3 *The feature matrix X has full column rank.*

Assumptions 4.1 and 4.3 are standard in the off-policy RL literature. Assumption 4.2 can indeed be weakened to the canonical coverage assumption $\pi(a|s) > 0 \implies \mu(a|s) > 0$ (Yu, 2015). Then for a state s, we can simply consider only actions a such that $\mu(a|s) > 0$, i.e., different states have different action spaces. All the analysis presented in this section still hold. We use Assumption 4.2 mainly to simplify presentation.

When analyzing the original ETD, we have to consider the chain $\{(F_t, S_t, A_t)\}$ evolving in the space $\mathbb{R} \times \mathcal{S} \times \mathcal{A}$ (see, e.g., Yu 2015). The space $\mathbb{R}$ is not even countable, making it hard to analyze the chain $\{(F_t, S_t, A_t)\}$ even with Assumption 4.1. With the truncated followon trace $F_{t,n}$, we only need to consider the chain $\{(S_{t-n}, A_{t-n}, \dots, S_t, A_t)\}$ which evolves in a *finite* space $(\mathcal{S} \times \mathcal{A})^n$. The ergodicity of this chain follows immediately from Assumption 4.1. Once the ergodicity is established, we can analyze the limiting update matrix under the corresponding invariant distribution.

The additional hyperparameter n in (12) defines a hard truncation. By contrast, the additional hyperparameter β in (9) defines a soft truncation. As discussed in Section 3, a desired β does not always exist since we require β to be both sufficiently large and sufficiently small. By contrast, we will show soon that a desired n always exists because we only require n to be sufficiently large. Further, to analyze ETD(λ, β) with the soft truncation, we still need to work on the chain $\{(F_{t,\beta}, S_t, A_t)\}$, whose behavior is hard to analyze. Consequently, though the asymptotic convergence of ETD(λ, β) in prediction may be established similarly to Yu (2015) for certain β, so far no finite sample analysis is available for ETD(λ, β) in prediction, much less control. Nevertheless, we believe the soft truncation and the hard truncation are two different directions for variance reduction. The soft truncation is analogous to computing the return G_t with a discount factor different from γ (see, e.g., Romoff et al. 2019); the hard truncation is analogous to computing the return G_t with a fixed horizon (see, e.g., Asis et al. 2020). It is straightforward to combine the two techniques together. For example, we can consider $F_{t,\beta,n}$ defined as

$$F_{t,\beta,n} \doteq \begin{cases} \sum_{j=0}^{n} \beta^j \rho_{t-j:t-1} i_{t-j} & t \geq n \\ F_{t,\beta} & t < n \end{cases}.$$

This combination, however, deviates from the main purpose of this paper and is saved for future work.

We now study the truncated trace $F_{t,n}$. Similar to (5), we study the limit of the conditional expectation of the truncated followon trace and define

$$m_n(s) \doteq \lim_{t \to \infty} \mathbb{E}\left[F_{t,n} | S_t = s\right].$$

When $n = \infty$, this m_∞ is referred to as *emphasis* in Zhang et al. (2020b). We, therefore, refer to m_n as *truncated emphasis* for a finite n.

Lemma 1 *Let Assumptions 4.1 and 4.2 hold. Then*

$$m_n = \sum_{j=0}^{n} \gamma^j D_\mu^{-1} (P_\pi^\top)^j D_\mu i,$$
$$m \doteq \lim_{n \to \infty} m_n = D_\mu^{-1}(I - \gamma P_\pi^\top)^{-1} D_\mu i.$$

The proof of Lemma 1 is provided in Section B.1. By definition, the weighting vector f in (4) involved in M of the ETD update (3) satisfies $f = D_\mu m$. Similarly, we define $f_n \doteq D_\mu m_n$.

Lemma 2 *Let Assumptions 4.1 and 4.2 hold. Then*

$$\|m_n - m\|_1 \leq \gamma^{n+1} \frac{d_{\mu,max}}{d_{\mu,min}} \|m\|_1,$$
$$\|f_n - f\|_\infty \leq \gamma^{n+1} \frac{d_{\mu,max}^2}{d_{\mu,min}} \|m\|_1,$$

where $d_{\mu,max} \doteq \max_s d_\mu(s)$ *and* $d_{\mu,min} \doteq \min_s d_\mu(s)$.

The proof of Lemma 2 is provided in Section B.2. The M matrix of the ETD(0) update (3) is $X^\top D_f(\gamma P_\pi - I)X$. Similarly, it can be shown that the M matrix of Truncated Emphatic TD (Algorithm 1) is $X^\top D_{f_n}(\gamma P_\pi - I)X$. Lemma 2 asserts that f_n approaches f geometrically fast. Consequently, we can expect $X^\top D_{f_n}(\gamma P_\pi - I)X$ to be n.d. if n is not too small.

Lemma 3 *Under Assumptions 4.1, 4.2, and 4.3, if*

$$\gamma^{n+1} < \frac{\lambda_{min} d_{\mu,min}}{d_{\mu,max}^2 \|\gamma P_\pi - I\| \|m\|_1}, \tag{13}$$

then $X^\top D_{f_n}(\gamma P_\pi - I)X$ is n.d., where λ_{min} is the minimum eigenvalue of

$$\frac{1}{2}\Big(D_f(I - \gamma P_\pi) + (I - \gamma P_\pi^\top)D_f\Big).$$

Sutton et al. (2016) prove that $\lambda_{min} > 0$. The proof of Lemma 3 is provided in Section B.3. Since the LHS of (13) diminishes geometrically as n increases, we argue that in practice we do not need a very large n. Recall that the motivation of using the followon trace F_t is to ensure the limiting update matrix to be n.d. Lemma 3 shows that to ensure this negative definiteness, we do not need to use all history to compute F_t. *Earlier steps contribute little to this negative definiteness due to discounting but introduce large variance due to the products of importance sampling ratios.* As suggested by (13), the desired value of n depends on the magnitude of the emphasis m, which is determined together by the behavior policy μ, the target policy π, the structure of the MDP, and the magnitude of the interest i. In general, when the magnitude of the emphasis increases, the desired truncation length also increases. In practice, we propose to treat the truncation length n as an additional hyperparameter, as estimating the desired n without access to the transition kernel p can be very challenging, which we leave for future work.

We can now show the asymptotic convergence of Truncated Emphatic TD using the standard ODE-based approach.

Assumption 4.4 *The learning rates $\{\alpha_t\}$ are positive, nonincreasing, and satisfy*

$$\sum_t \alpha_t = \infty, \sum_t \alpha_t^2 < \infty.$$

Theorem 4 *Let the assumptions and conditions of Lemma 3 hold. Let Assumption 4.4 hold. Then the iterates $\{w_t\}$ generated by Truncated Empathic TD (Algorithm 1) satisfy*

$$\lim_{t\to\infty} w_t = w_{*,n} \quad \text{a.s., where}$$
$$w_{*,n} \doteq -A_n^{-1} b_n,\ A_n \doteq X^\top D_{f_n}(\gamma P_\pi - I)X,\ b_n \doteq X^\top D_{f_n} r_\pi.$$

The proof of Theorem 4 is provided in Section B.4, which, after the negative definiteness of A_n is established with Lemma 3, follows the same routine as the convergence proof of on-policy TD(λ) in Proposition 6.4 of Bertsekas and Tsitsiklis (1996).

We now give a finite sample analysis of Projected Truncated Emphatic TD (Algorithm 2). Algorithm 2 is different from Algorithm 1 in that it adopts an additional projection

Π_R when updating the weight w_t. Here Π_R denotes the projection onto the ball of a radius R centered at the origin w.r.t. ℓ_2 norm. Introducing such a projection is common practice in finite sample analysis of TD methods (Bhandari et al., 2018; Zou et al., 2019). This projection is mainly used to control the errors introduced by Markovian samples. If i.i.d. samples are used instead, such projection can indeed be eliminated (Bhandari et al., 2018; Dalal et al., 2018).

Algorithm 2: Projected Truncated Emphatic TD

```
S_0 ~ p_0(·)
t ← 0
while True do
    Sample A_t ~ μ(·|S_t)
    Execute A_t, get R_{t+1}, S_{t+1}
    ρ_t ← π(A_t|S_t) / μ(A_t|S_t)
    F_{t,n} ← 0
    for k = 0, ..., n do
        F_{t,n} ← i_{t-n+k} + γ ρ_{t-n+k-1} F_{t,n}
    end
    w_{t+1} ← Π_R ( w_t + α_t F_{t,n} ρ_t (R_{t+1} + γ x_{t+1}^⊤ w_t − x_t^⊤ w_t) x_t )
    t ← t + 1
end
```

Theorem 5 *Let the assumptions and conditions of Lemma 3 hold. Let $R \geq \|w_{*,n}\|$. With proper learning rates $\{\alpha_t\}$, for sufficiently large t,*

$$\mathbb{E}\left[\|w_t - w_{*,n}\|^2\right] = \mathcal{O}\left(\frac{\ln^3 t}{t}\right).$$

The proof of Theorem 5 is omitted to avoid verbatim repetition since it is just a special case of a more general result in the control setting (Theorem 12). The conditions on learning rates and the constants hidden by $\mathcal{O}(\cdot)$ are also similar to those of Theorem 12. We now analyze the performance of $w_{*,n}$.

Lemma 6 *Let $\kappa \doteq \min_s \frac{d_\mu(s)i(s)}{f(s)}$. Let Assumptions 4.1, 4.2, and 4.3 hold. If*

$$\gamma^{n+1} < \frac{\kappa d_{\mu,min} \min_s i(s) d_\mu(s)}{d_{\mu,max}^2 \|I - \gamma P_\pi^\top\|_\infty \|m\|_1}, \tag{14}$$

then $\Pi_{f_n}\mathcal{T}_\pi$ is a $\sqrt{\gamma}$-contraction in $\|\cdot\|_{f_n}$ and

$$\|Xw_{*,n} - v_\pi\|_{f_n} \leq \frac{1}{\sqrt{1-\gamma}} \|\Pi_{f,n} v_\pi - v_\pi\|_{f_n}. \tag{15}$$

Here Π_{f_n} denotes the projection onto the column space of X w.r.t. the norm induced by f_n, i.e.,

$$\Pi_{f_n} v \doteq X \arg\min_w \|Xw - v\|_{f_n}^2.$$

The proof of Lemma 6 is similar to Hallak et al. (2016) and is provided in Section B.5. Again, the LHS of (14) diminishes geometrically. So in practice, n might not need to be too large. Lemma 6 characterizes the performance of the fixed points of Truncated ETD methods in prediction settings. In the following we highlight two points regarding those fixed points from different truncation length.

1. We argue that those fixed points are equally good. The $\|\Pi_{f,n} v_\pi - v_\pi\|_{f_n}$ term in (15) is the representation error resulting from the limit of the capacity of the linear function approximator. With different truncation length, we use different norm (i.e., $\|\cdot\|_{f_n}$) to measure the representation error. The multiplicative factor $\frac{1}{\sqrt{1-\gamma}}$, however, does not depend on n. In other words, as long as n is sufficiently large in the sense of (14), the exact value of n, including $n = \infty$ (i.e., no truncation), does not seem to affect the performance of the fixed point much. The intuition is straightforward. Comparing (12) and (11), it is easy to see that by using the truncation, we discard the term $\sum_{j=n+1}^{t} \gamma^j \rho_{t-j:t-1} i_{t-j}$ corresponding to earlier transitions from steps 0 to $t-n-1$. This term has a large, possibly infinite, variance because of the product of importance sampling ratios. The expectation of this term is, however, negligible because the expectation of the importance sampling ratios are well bounded (see the proof of Lemma 2) and the multiplicative factor γ^j is negligible. It is the expectation, not the variance, of the trace that determines the performance of the corresponding fixed point. Consequently, the truncation proposed in this work does not seem to yield a compromise in the performance of the fixed point. The truncation proposed in this work is more like a free variance reduction instead of a bias-variance tradeoff.

2. We argue that those fixed points are better than the fixed points of gradient TD methods minimizing d_μ-induced mean squared projected Bellman errors (e.g., GTD in Sutton et al. 2008, GTD2 and TDC in Sutton et al. 2009, Gradient Tree Backup and Gradient Retrace in Touati et al. 2018). This is because the performance of the fixed points of Truncated ETD methods can be well-bounded by the representation error, provided that the length of the truncation is sufficiently large. By contrast, the performance of the fixed points of gradient TD methods can be arbitrarily worse, no matter how small the representation error is (Kolter, 2011).

5. Control: Emphatic Approximate Value Iteration

The study of the canonical approximate value iteration (6) is essential to the study of the on-policy control algorithm SARSA (Melo et al., 2008; Zou et al., 2019). Similarly, in this section, we study approximate value iteration from an off-policy perspective, which prepares us for the off-policy control algorithm in the next section. In the rest of this paper, we write f, m, f_n, m_n, κ (defined in Lemma 6), and λ_{min} (defined in Lemma 3) as $f_{\mu,\pi}, m_{\mu,\pi}, f_{n,\mu,\pi}, m_{n,\mu,\pi}, \kappa_{\mu,\pi}, \lambda_{min,\mu,\pi}$ to explicitly acknowledge their dependence on μ and π.

The canonical approximate value iteration operator in (6) is in a sense on-policy in that the projection operator is defined w.r.t. a norm induced by the policy of the current iteration. In this section, we study approximation value iteration from an off-policy perspective, i.e., with a projection operator defined w.r.t. a different norm. Let π_w and μ_w be target and behavior policies respectively. They depend on w, the parameters used for estimating the value function, through the value function estimate $v = Xw \in \mathbb{R}^{|\mathcal{S}|}$, e.g., they can be softmax policies (cf. (7)) with different temperatures. We consider the iterates $\{v_k \doteq Xw_k\}$ generated by

$$v_{k+1} \doteq \Pi_{f_{n,\mu_{w_k},\pi_{w_k}}} \mathcal{T}_{\pi_{w_k}} v_k.$$

We call this scheme *emphatic approximate value iteration* as the projection operator is defined w.r.t. the norm induced by the (truncated) followon trace. In the rest of this section, we show that emphatic approximate value iteration adopts at least one fixed point.

With Λ_M denoting the closure of $\left\{\mu_w \mid w \in \mathbb{R}^K\right\}$ and Λ_Π denoting the closure of $\left\{\pi_w \mid w \in \mathbb{R}^K\right\}$, we make the following assumptions.

Assumption 5.1 *Both π_w and μ_w are continuous in w.*

Assumption 5.2 *For any $\mu \in \Lambda_M$, the Markov chain induced by μ is ergodic and $\mu(a|s) > 0$ holds for all (s, a).*

Assumption 5.1 is standard in analyzing approximate value iteration (De Farias and Van Roy, 2000). If π_w is not continuous in w, even the canonical approximate value iteration can fail to have a fixed point (De Farias and Van Roy, 2000). The ergodicity assumption of all the policies in the closure in Assumption 5.2 is also standard for analyzing control algorithms, in both on-policy (Marbach and Tsitsiklis, 2001) and off-policy (Zhang et al., 2021b,a) settings. One common strategy to ensure this ergodicity in closure is to mix a softmax policy with a uniformly random policy, assuming the uniformly random policy always induces an ergodic chain.

We now define two helper functions to understand how n should be selected in emphatic approximate value iteration.

$$n_1(\mu, \pi) \doteq \frac{\ln(\lambda_{min,\mu,\pi} d_{\mu,min}) - \ln\left(d^2_{\mu,max} \|\gamma P_\pi - I\| \|m_{\mu,\pi}\|_1\right)}{\ln \gamma} - 1,$$
$$n_2(\mu, \pi) \doteq \frac{\ln\left(\kappa_{\mu,\pi} d_{\mu,min} \min_s i(s) d_\mu(s)\right) - \ln\left(d^2_{\mu,max} \left\|I - \gamma P_\pi^\top\right\|_\infty \|m_{\mu,\pi}\|_1\right)}{\ln \gamma} - 1.$$

Here n_1 and n_2 correspond to the conditions of n in Lemmas 3 and 6 respectively. Assumption 5.2 ensures that n_1 and n_2 are well defined on $\Lambda_M \times \Lambda_\Pi$. The invariant distribution d_μ is continuous in μ (see, e.g., Lemma 9 of Zhang et al. 2021b), the minimum eigenvalue $\lambda_{min,\mu,\pi}$ is continuous in the elements of the matrix (see, e.g., Corollary 8.6.2 of Golub and Loan 1996) and thus is also continuous in μ and π, and both Λ_M and Λ_Π are compact. Therefore, $\sup_{\mu \in \Lambda_M, \pi \in \Lambda_\Pi} \max\{n_1(\mu, \pi), n_2(\mu, \pi)\} < \infty$ by the extreme value theorem. This allows us to select n as suggested by the following lemma.

Lemma 7 *Let Assumptions 4.3, 5.1, and 5.2 hold. If*

$$n > \sup_{\mu \in \Lambda_M, \pi \in \Lambda_\Pi} \max\{n_1(\mu, \pi), n_2(\mu, \pi)\}, \tag{16}$$

then there exists at least one w_* such that

$$Xw_* = \Pi_{f_{n,\mu_{w_*},\pi_{w_*}}} \mathcal{T}_{\pi_{w_*}} Xw_*.$$

The proof of Theorem 7 is provided in B.6, which follows the same steps of De Farias and Van Roy (2000) but generalizes their results from (on-policy) approximate value iteration to emphatic approximate value iteration.

6. Control: Truncated Emphatic Expected SARSA

We now present our control algorithm, Truncated Emphatic Expected SARSA. Unlike planning methods such as approximate value iteration, learning methods for control like SARSA usually work directly on action-value estimates. To this end, we overload notation for the ease of presentation. In particular, we overload the feature function x as $x : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^K$ to denote a state-action feature function. Correspondingly, the feature matrix X is now a matrix in $\mathbb{R}^{|\mathcal{S}\times\mathcal{A}|\times K}$ whose (s, a)-th row is $x(s, a)^\top$. The transition matrix P_π is now a matrix in $\mathbb{R}^{|\mathcal{S}\times\mathcal{A}|\times|\mathcal{S}\times\mathcal{A}|}$ to denote the state-action pair transition, i.e.,

$$P_\pi((s, a), (s', a')) \doteq p(s'|s, a)\pi(a'|s').$$

Consequently, the Bellman operator is overloaded as

$$\mathcal{T}_\pi q \doteq r + \gamma P_\pi q.$$

The stationary distribution d_μ is now in $\mathbb{R}^{|\mathcal{S}\times\mathcal{A}|}$ and denotes the invariant state-action pair distribution under the policy μ. D_μ is then a diagonal matrix in $\mathbb{R}^{|\mathcal{S}\times\mathcal{A}|\times|\mathcal{S}\times\mathcal{A}|}$. The interest function i is now from $\mathcal{S} \times \mathcal{A}$ to $(0, +\infty)$ to denote user's preference for each state-action pair. The followon trace F_t is now defined as

$$F_t \doteq i_t + \gamma\rho_t F_{t-1},$$

which is the same as the followon trace used in ELSTDQ(λ) in White (2017). Correspondingly, the truncated trace is defined as

$$F_{t,n} \doteq \begin{cases} \sum_{j=0}^{n} \gamma^j \rho_{t-j+1:t} i_{t-j} & t \geq n \\ F_t & t < n \end{cases}.$$

The truncated emphasis $m_{n,\mu,\pi}$ is now in $\mathbb{R}^{|\mathcal{S}\times\mathcal{A}|}$ and defined as

$$m_{n,\mu,\pi}(s, a) \doteq \lim_{t\to\infty} \mathbb{E}[F_{t,n}|S_t = s, A_t = a].$$

Other notation is also overloaded accordingly, e.g.,

$$m_{\mu,\pi} \doteq \lim_{n\to\infty} m_{n,\mu,\pi},\ f_{n,\mu,\pi} \doteq D_\mu m_{n,\mu,\pi},\ f_{\mu,\pi} \doteq D_\mu m_{\mu,\pi}.$$

Previous theoretical results also hold with the overloaded notation for state-action pairs. In particular, we have

Lemma 8 *Let Assumptions 4.3, 5.1, and 5.2 hold. Define*

$$n_1(\mu, \pi) \doteq \frac{\ln(\lambda_{min,\mu,\pi} d_{\mu,min}) - \ln\left(d^2_{\mu,max} \|\gamma P_\pi - I\| \|m_{\mu,\pi}\|_1\right)}{\ln \gamma} - 1,$$

$$n_2(\mu, \pi) \doteq \frac{\ln\left(\kappa_{\mu,\pi} d_{\mu,min} \min_{s,a} i(s, a) d_\mu(s, a)\right) - \ln\left(d^2_{\mu,max} \left\|I - \gamma P_\pi^\top\right\|_\infty \|m_{\mu,\pi}\|_1\right)}{\ln \gamma} - 1,$$

where $\lambda_{min,\mu,\pi}$ is the minimum eigenvalue of

$$\frac{1}{2}\left(D_{f_{\mu,\pi}}(I-\gamma P_\pi)+(I-\gamma P_\pi^\top)D_{f_{\mu,\pi}}\right),$$

$d_{\mu,min} \doteq \min_{s,a} d_\mu(s,a), d_{\mu,max} \doteq \max_{s,a} d_{\mu,max}(s,a), \kappa_{\mu,\pi} \doteq \min_{s,a} \frac{d_\mu(s,a)i(s,a)}{f(s,a)}$. *If*

$$n > \sup_{\mu\in\Lambda_M,\pi\in\Lambda_\Pi} \max\left\{n_1(\mu,\pi), n_2(\mu,\pi)\right\}$$

holds, then

(i). For any $\mu \in \Lambda_M, \pi \in \Lambda_\Pi$, $X^\top D_{f_{n,\mu,\pi}}(\gamma P_\pi - I)X$ is n.d.,

(ii). For any $\mu \in \Lambda_M, \pi \in \Lambda_\Pi$, $\Pi_{f_{n,\mu,\pi}}\mathcal{T}_\pi$ is a $\sqrt{\gamma}$ contraction in $\|\cdot\|_{f_{n,\mu,\pi}}$,

(iii). There exists at least one w_* such that

$$Xw_* = \Pi_{f_{n,\mu_{w_*},\pi_{w_*}}}\mathcal{T}_{\pi_{w_*}}Xw_*. \tag{17}$$

We use $\mathcal{W}_*$ to denote the set of all such w_*.

The proof of Lemma 8 is omitted since it is a verbatim repetition of the proofs of Lemmas 3, 6, and 7.

The iterative update scheme (17) is emphatic approximate value iteration applied to action-value estimation. To implement this scheme incrementally in a learning sense, we propose Truncated Emphatic Off-Policy Expected SARSA (Algorithm 3). When computing $F_{t,n}$, we require that the previous importance sampling ratios be recomputed with the current weight w_t. This requirement is mainly for the ease of asymptotic analysis and is eliminated in Projected Truncated Emphatic Expected SARSA, for which we provide a finite sample analysis.

We can now present our asymptotic convergence analysis of Algorithm 3. We first study the properties of the possible fixed points. We can rewrite (17) as

$$A_{w_*}w_* + b_{w_*} = 0,$$

where

$$\begin{aligned} A_w &\doteq X^\top D_{f_{n,\mu_w,\pi_w}}(\gamma P_{\pi_w} - I)X, \\ b_w &\doteq X^\top D_{f_{n,\mu_w,\pi_w}} r. \end{aligned}$$

Consequently,

$$w_* = A_{w_*}^{-1} b_{w_*}.$$

Since Λ_M and Λ_Π are compact, both π_w and μ_w are continuous in w, the RHS of the above equation is bounded from above by the extreme value theorem. Consequently, there exists a constant $R < \infty$ such that

$$\sup_{w_*\in\mathcal{W}_*} \|w_*\| \le R.$$

Algorithm 3: Truncated Emphatic Expected SARSA

$S_0 \sim p_0(\cdot)$
$A_0 \sim \mu_{w_0}(\cdot|S_0)$
$t \leftarrow 0$
while *True* **do**
  Execute A_t, get R_{t+1}, S_{t+1}
  $A_{t+1} \sim \mu_{w_t}(\cdot|S_{t+1})$
  $\rho_t \leftarrow \frac{\pi_{w_t}(A_t|S_t)}{\mu_{w_t}(A_t|S_t)}$
  $F_{t,n} \leftarrow 0$
  for $k = 0, \dots, n$ **do**
    $F_{t,n} \leftarrow i_{t-n+k} + \gamma \frac{\pi_{w_t}(A_{t-n+k}|S_{t-n+k})}{\mu_{w_t}(A_{t-n+k}|S_{t-n+k})} F_{t,n}$
  end
  $w_{t+1} \leftarrow w_t + \alpha_t F_{t,n}(R_{t+1} + \gamma \sum_a \pi_{w_t}(a|S_{t+1}) x(S_{t+1}, a)^\top w_t - x_t^\top w_t) x_t$
  $t \leftarrow t+1$
end

We then make several regularization conditions on the policies π_w and μ_w. For the analysis of on-policy SARSA (8), it is commonly assumed that the policy π_w is Lipschitz continuous in w and the Lipschitz constant is not too large (Perkins and Precup, 2002; Zou et al., 2019). This technical assumption is mainly used to ensure that a small change in the value estimate does not result in a big difference in the policy thus enforces certain smoothness of the overall learning process. Without such assumptions, even on-policy SARSA can chatter and fail to converge (Gordon, 1996, 2001). In this paper, we adopt similar assumptions in our off-policy setting.

Assumption 6.1 *Both μ_w and π_w are Lipschitz continuous in w, i.e., there exist constants L_μ and L_π such that for any $s \in \mathcal{S}, a \in \mathcal{A}$,*

$$
\begin{aligned}
|\pi_w(a|s) - \pi_{w'}(a|s)| &\leq L_\mu \big\|w - w'\big\|, \\
|\mu_w(a|s) - \mu_{w'}(a|s)| &\leq L_\mu \big\|w - w'\big\|.
\end{aligned}
$$

The Lipschitz continuity of the policies immediately implies the Lipschitz continuity of A_w and b_w governing the expected updates of Algorithm 3.

Lemma 9 *Let Assumptions 5.2 and 6.1 hold. There exist positive constants C_1, C_2, C_3, and C_4 such that for any w, w'*

$$
\begin{aligned}
\|A_w - A_{w'}\| &\leq (C_1 L_\mu + C_2 L_\pi)\big\|w - w'\big\|, \\
\|b_w - b_{w'}\| &\leq (C_3 L_\mu + C_4 L_\pi)\big\|w - w'\big\|.
\end{aligned}
$$

The proof Lemma 9 is provided in Section B.7. Under the conditions of Lemma 8, for any w, the matrix

$$
M(w) \doteq \frac{1}{2}\left(X^\top D_{f_{n,\mu_w,\pi_w}}(I - \gamma P_{\pi_w})X + X^\top (I - \gamma P_{\pi_w}^\top) D_{f_{n,\mu_w,\pi_w}} X\right)
$$

is p.d. For a symmetric positive definite matrix M, let $\lambda(M)$ denote the smallest eigenvalue of M. For any $w \in \mathbb{R}^K$, we have $\lambda(M(w)) > 0$. By the continuity of eigenvalues in the elements of the matrix, the compactness of Λ_M and Λ_Π, and the extreme value theorem, we have

$$\inf_{w \in \mathbb{R}^K} \lambda(M(w)) > 0.$$

This allows us to make the following assumptions about the Lipschitz constants L_μ and L_π, akin to Perkins and Precup (2002); Zou et al. (2019).

Assumption 6.2 *L_μ and L_π are small enough such that*

$$\lambda'_{min} \doteq \inf_{w \in \mathcal{W}} \lambda(M(w)) - ((C_1 L_\mu + C_2 L_\pi) R + C_3 L_\mu + C_4 L_\pi) > 0.$$

With these regularizations on π_w and μ_w, we can now present a high probability asymptotic convergence analysis for Algorithm 3.

Theorem 10 *Let the assumptions and conditions in Lemma 8 hold. Let Assumptions 4.4, 6.1, and 6.2 hold. Then for any compact set $\mathcal{W} \subset \mathbb{R}^K$ and any $w \in \mathcal{W}$, there exists a constant $C_\mathcal{W}$ such that for any $w_* \in \mathcal{W}_*$, the iterates $\{w_t\}$ generated by Algorithm 3 satisfy*

$$\Pr\Big(\lim_{t \to \infty} w_t = w_* \mid w_0 = w\Big) \geq 1 - C_\mathcal{W} \sum_{t=0}^{\infty} \alpha_t^2. \tag{18}$$

This immediately implies that $\mathcal{W}_*$ contains only one element (under the conditions of this theorem).

In (18), $C_\mathcal{W}$ depends on the compact set $\mathcal{W}$ from which the weight w_0 is selected. For (18) to be nontrivial, the learning rates have to be small enough, depending on the choice of initial weights. The proof of Theorem 10 is provided in Section B.8 and depends on Theorem 13 of Benveniste et al. (1990).[2]

We now analyze the convergence rate of Projected Truncated Emphatic Expected SARSA (Algorithm 4). Unlike Algorithm 3, when computing $F_{t,n}$ in Algorithm 4, we do *not* need to recompute previous importance sampling ratios. Similar to Assumption 6.2, we make the following assumption about the Lipschitz constants L_μ and L_π for analyzing Algorithm 4.

Assumption 6.3 *L_μ and L_π are not too large such that*

$$\lambda''_{min} \doteq \inf_{w_* \in \mathcal{W}_*} \lambda(M(w_*)) - ((C_1 L_\mu + C_2 L_\pi) R + C_3 L_\mu + C_4 L_\pi) > 0.$$

When defining λ'_{min} in Assumption 6.2, the infimum is taken over all possible w. When defining λ''_{min} in Assumption 6.3, the infimum is taken over only $\mathcal{W}_*$. This improvement is made possible by the introduction of the projection Π_R.

To analyze the convergence rate of Algorithm 4, it is crucial to know how fast the induced chain mixes, which is provided by the following lemma.

[2] It might be possible to obtain an almost sure convergence of Algorithm 3 like Theorem 4 by invoking Theorem 17 of Benveniste et al. (1990). Doing so requires verifying (1.9.5) of Benveniste et al. (1990). If how Melo et al. (2008) verify (1.9.5) was documented in the context of on-policy SARSA with linear function approximation, it is expected that (1.9.5) can also be similarly verified in the context of Algorithm 3.

Algorithm 4: Projected Truncated Emphatic Expected SARSA

Initialize w_0 such that $\|w_0\| \leq R$
$S_0 \sim p_0(\cdot)$
$A_0 \sim \mu_{w_0}(\cdot|S_0)$
$t \leftarrow 0$
while *True* **do**
  Execute A_t, get R_{t+1}, S_{t+1}
  $A_{t+1} \sim \mu_{w_t}(\cdot|S_{t+1})$
  $\rho_t \leftarrow \frac{\pi_{w_t}(A_t|S_t)}{\mu_{w_t}(A_t|S_t)}$
  $F_{t,n} \leftarrow 0$
  for $k = 0, \dots, n$ **do**
    $F_{t,n} \leftarrow i_{t-n+k} + \gamma \rho_{t-n+k} F_{t,n}$
  end
  $w_{t+1} \leftarrow \Pi_R \left(w_t + \alpha_t F_{t,n}(R_{t+1} + \gamma \sum_a \pi_{w_t}(a|S_{t+1}) x(S_{t+1}, a)^\top w_t - x_t^\top w_t) x_t\right)$
  $t \leftarrow t + 1$
end

Lemma 11 *(Lemma 1 of Zhang et al. (2021a)) Let Assumption 5.2 hold. Then there are constants $C_0 > 0$ and $\kappa \in (0,1)$ such that for any $w \in \mathbb{R}^K$ and $t \geq 0$, the chain $\{S_t\}_{t=0,1,\dots}$ induced by the policy μ_w satisfies*

$$\sum_{s \in \mathcal{S}} \left|\Pr(S_t = s) - \bar{d}_{\mu_w}(s)\right| \leq C_0 \kappa^t,$$

where $\bar{d}_{\mu_w}$ is the invariant state distribution induced by the policy μ_w.

Lemma 11 is usually referred to as *uniform mixing* since the mixing rate κ does not depend on w. This uniform mixing appears to be a technical assumption in Zou et al. (2019). We are now ready to present our finite sample analysis.

Theorem 12 *Let the assumptions and conditions in Lemma 8 hold. Let Assumptions 6.1 and 6.3 hold. Set the learning rate $\{\alpha_t\}$ in Algorithm 3 to*

$$\alpha_t \doteq \frac{1}{2\alpha_\lambda(t+1)}, \tag{19}$$

where $\alpha_\lambda \in (0, \lambda''_{min})$ is some constant. Then for any $w_* \in \mathcal{W}_*$, for sufficiently large t (in the sense that $t - \mathcal{O}(\ln t) > n$), the weight vector w_t generated by Algorithm 4 satisfies

$$\mathbb{E}\left[\|w_t - w_*\|^2\right] = \mathcal{O}\left(\frac{\ln^3 t}{t}\right).$$

This immediately implies that $\mathcal{W}_*$ contains only one element (under the conditions of this theorem).

The proof of Theorem 12 and the constants hidden by $\mathcal{O}(\cdot)$ are detailed in Section B.9. The proof follows the same steps as Zou et al. (2019) but generalizes the analysis of the on-policy SARSA in Zou et al. (2019) to the off-policy setting and includes backward traces, which are not included in Zou et al. (2019).

In this section, we present (Projected) Truncated Emphatic Expected SARSA as a convergent off-policy control algorithm with linear function approximation. Importantly, in Algorithms 3 and 4, the behavior policy is a function of the current action-value estimates and thus changes every time step and can be very different from the target policy. These two features are common in practice (see, e.g. Mnih et al. (2015)) but rarely appreciated in existing literature. For example, in Greedy-GQ (Maei et al., 2010; Wang and Zou, 2020), a control algorithm in the family of the gradient TD methods, the behavior policy is assumed to be fixed. In the convergent analysis of linear Q-learning (Melo et al., 2008; Lee and He, 2019), the behavior policy is assumed to be sufficiently close to the policy that linear Q-learning is expected to converge to.

7. Related Work

Besides gradient TD and emphatic TD methods, there are also other methods for addressing the deadly triad, including density-ratio-based methods (Hallak and Mannor, 2017; Liu et al., 2018; Gelada and Bellemare, 2019; Nachum et al., 2019; Zhang et al., 2020a) and target-network-based methods (Zhang et al., 2021b). Density-ratio-based methods rely on learning the density ratio $\frac{d_\pi(s)}{d_\mu(s)}$ directly via function approximation. This ratio can then be used to reweight the off-policy TD update (2) (Hallak and Mannor, 2017) if the goal is to learn the value function of the target policy or reweight rewards directly when computing the empirical average of rewards (Liu et al., 2018) if the goal is to obtain a scalar performance metric of the target policy. Density ratio learning suffers from the same critical problem as emphasis learning and gradient TD methods. Under general conditions for linear function approximation, it is hard to bound the distance between the learned density ratio and the ground truth. Consequently, the downstream updates relying on the learned density ratio rarely have rigorous performance guarantees. By contrast, the performance of the value function estimation resulting from emphatic TD methods is well bounded (Hallak et al. (2016) and Lemma 6).

Target-network-based methods rely on the use of a target network (Mnih et al., 2015) for bootstrapping. Though provably convergent, the methods of Zhang et al. (2021b) require a sufficiently large ridge regularization, introducing extra bias. By contrast, emphatic TD methods do not require any extra regularization.

Variance reduction is an active research area in RL (e.g., Du et al. 2017; Papini et al. 2018), which is usually achieved by designing proper control variates, see, e.g., Johnson and Zhang (2013). By contrast, we reduce variance by truncating the followon trace directly. This technique is specifically designed for emphatic TD methods and we leave the possible combination of truncated emphatic TD methods and standard variance reduction techniques for future work.

	$n=\infty$	$n=0$	$n=2$	$n=4$	$n=8$	$\beta=0.8$
$\pi(dashed=0\|s)$	-	-	-	-	-	-
$\pi(dashed=0.02\|s)$	-	-	-	10^4	10^{14}	-
$\pi(dashed=0.04\|s)$	10^7	-	10^1	10^1	10^9	10^9
$\pi(dashed=0.06\|s)$	-	-	10^2	10^0	10^4	10^4
$\pi(dashed=0.08\|s)$	-	-	10^{-1}	10^0	10^7	10^7
$\pi(dashed=0.1\|s)$	-	-	10^{-11}	10^0	10^2	10^4

Table 1: Average variance of curves in Figure 7. Each curve in Figure 7 consists of 100 data points. The average variance of those data points is reported in this table. Here we consider only successful configurations whose averaged prediction error at the end of training is smaller than 5. The average variance of other curves are not included and denoted as "-". Other tables in this section also follow this reporting protocol.

8. Experiments

In this section, we empirically investigate the proposed truncated emphatic TD methods, focusing on the effect of n. The implementation is made publicly available to facilitate future research.[3]

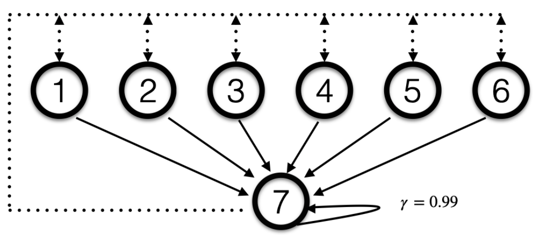

Figure 1: Baird's counterexample from Chapter 11.2 of Sutton and Barto (2018). The figure is taken from Zhang et al. (2020b). There are two actions available at each state, `dashed` and `solid`. The `solid` action always leads to state 7. The `dashed` action leads to one of states 1 - 6, with equal probability. The discount factor is $\gamma = 0.99$. The reward is always 0. The initial state is sampled uniformly from all the seven states.

We first use Baird's counterexample as the benchmark, which is illustrated in Figure 1. We consider three different settings: prediction, control with a fixed behavior policy, and control with a changing behavior policy. For the prediction setting, we consider a behavior policy $\mu(\texttt{solid}|s) = \frac{1}{7}$ and $\mu(\texttt{dashed}|s) = \frac{6}{7}$, which is the same as the behavior policy used in Sutton and Barto (2018). We consider different target policies from $\pi(\texttt{dashed}|s) = 0$ to $\pi(\texttt{dashed}|s) = 0.1$. We consider linear function approximation, where the features and the initialization of the weight vector are the same as Section D.2 of Zhang et al. (2021b). We benchmark Algorithm 1 with different selection of n. When $n = \infty$, Algorithm 1 reduces to the original ETD(0). When $n = 0$, Algorithm 1 reduces to the naive off-policy TD. We use

[3] `https://github.com/ShangtongZhang/DeepRL`

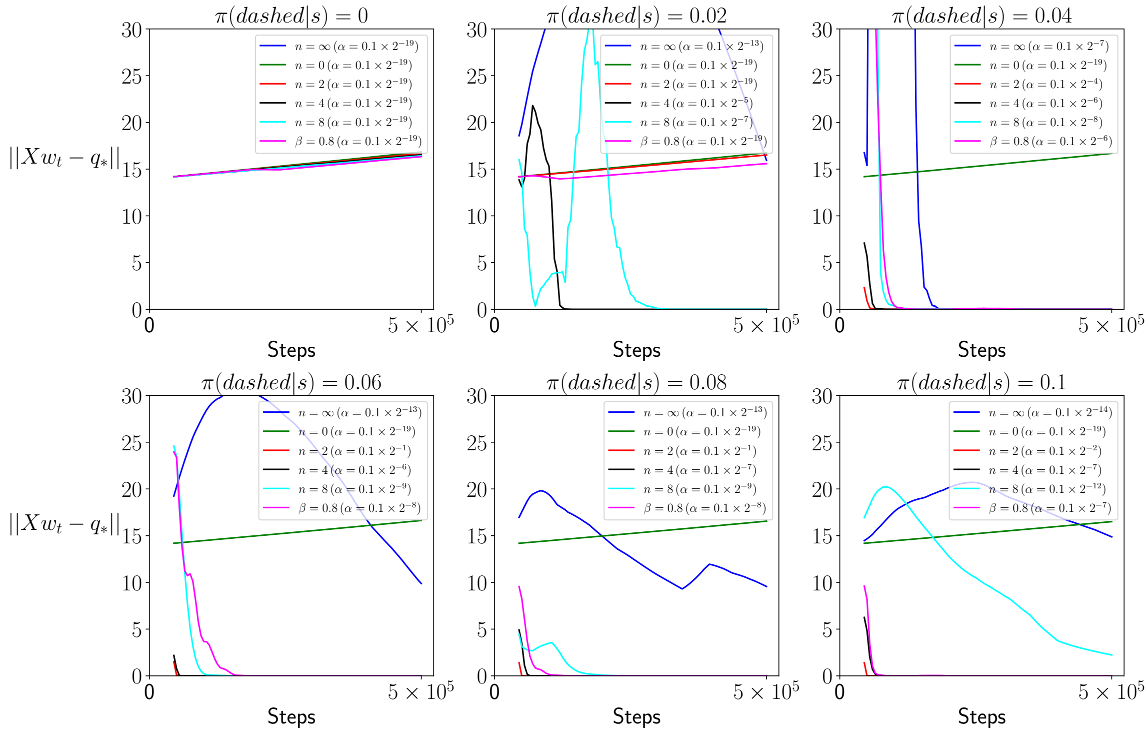

Figure 2: Truncated Emphatic TD and ETD(0, β) in the prediction setting. To improve readability, this figure contains only *one representative run* and uses a sliding window of size 10 for smoothing. A more informative but harder to read version including 30 independent runs without smoothing is provided in Figure 7 in Section C. The curves in the two figures share similar trends and all the discussion in the paper is based on the comprehensive results in Figure 7.

a fixed learning rate α, which is tuned from $\Lambda_\alpha \doteq \left\{0.1 \times 2^0, 0.1 \times 2^{-1}, \dots, 0.1 \times 2^{-19}\right\}$ for each n, with 30 independent runs. We report learning curves with the learning rate minimizing the value prediction error at the end of training. Additionally, we also benchmark ETD$(0, \beta)$, where we replace the $F_{t,n}$ in Algorithm 1 with the trace $F_{t,\beta}$ computed via (9). We tune β in $\{0.1, 0.2, 0.4, 0.8\}$. For each β, we tune the learning rate α in Λ_α as before. The interest is 1 for all states (i.e., $i(s) \equiv 1 \, \forall s$). We report the learning curves with the best β. All curves are averaged over 30 independent runs with shaded regions indicating *standard errors*, unless otherwise specified. This experimental and reporting protocol is also used in all the remaining experiments in this paper.

As shown by Figures 2 and 7 with $n = 0$, the naive off-policy TD makes no progress in this prediction setting. The curve is almost flat because the best learning rate is 0.1×2^{-19}; using any larger learning rate simply accelerates divergence. As shown by the curves with $n = \infty$, naive ETD(0) does make some progresses when $\pi(\texttt{dashed}|s) > 0$, though the final prediction errors at the end of training are usually large. By contrast, using $n = 4$ leads to quick convergence in all the tasks with $\pi(\texttt{dashed}|s) > 0$. Reducing n from 4 to 2 also works when $\pi(\texttt{dashed}|s) \geq 0.04$ and increasing n from 4 to 8 significantly increases the variance.

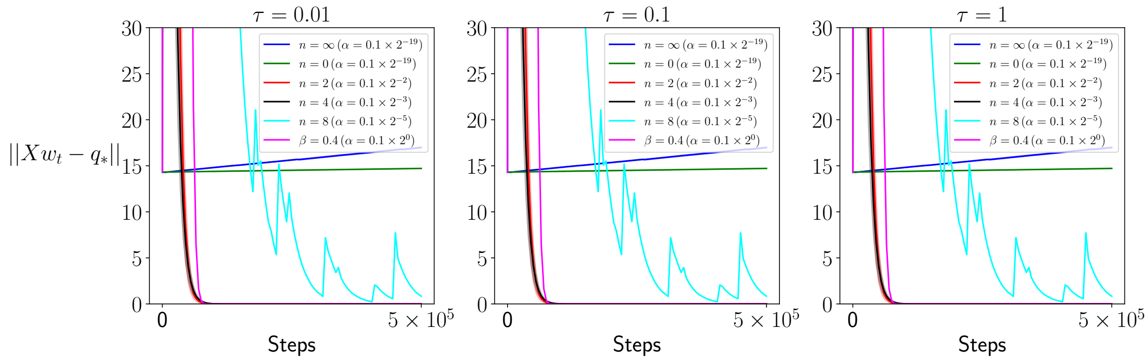

Figure 3: Truncated Emphatic Expected SARSA and its β-variant in the control setting with a fixed behavior policy. The shaded regions are invisible for some curves because their standard errors are too small.

Obviously increasing n leads to a larger variance, so in practice we want to find the smallest n. Moreover, though ETD$(0, \beta)$ converges when $\pi(\texttt{dashed}|s) \geq 0.04$, it usually exhibits larger variance than our Truncated ETD with $n = 2$ or $n = 4$ (Table 1). We conjecture that this is because the trace (9) still relies on all the history. Consider, e.g., $\pi(\texttt{dashed}|s) = 0.02$: the maximum importance ratio is $\rho_{max} = 0.98 \times 7 = 6.86$. If $\beta\rho_{max} > 1$, there is still a chance that the trace in (9) goes to infinity since it depends on all the history. However, requiring $\beta\rho_{max} < 1$ would require using a small β, which itself could also lead to instability. By contrast, with truncation, $F_{t,n}$ is always guaranteed to be bounded. The results suggest that our hard truncation also has empirical advantages over the soft truncation in Hallak et al. (2016), besides the theoretical advantages of enabling finite sample analysis for both prediction and control settings. It can be analytically computed that for all $\pi(\texttt{dashed}|s) \in \{0, 0.02, 0.04, 0.06, 0.08, 0.1\}$, the desired n as suggested by Lemma 3 is around 700. The n we use in the experiments is much smaller than the suggested one. This is because Lemma 3 has to be conservative enough to cope well with all possible MDPs. In this work, we focus on establishing the existence of such an n and giving an initial but possibly loose bound. We leave the improvement of Lemma 3 for future work. For computational experiments, we recommend to treat n as an additional hyperparameter.

When $\pi(\texttt{dashed}|s) = 0$, which is used in the original Baird's counterexample, no selection of n or β is able to make any progress. The failure of ETD(0) with this target policy is also observed by Sutton and Barto (2018). This target policy is particularly challenging because the its off-policyness is the largest in all the tested target policies, making it hard to observe progresses in computational experiments. Though truncation is not guaranteed to always reduce the variance to desired levels while maintaining convergence, our experiments in the prediction setting do suggest it is a promising approach. We leave a more in-depth investigation with this target policy for future work.

In the control setting with a fixed behavior policy, we benchmark Algorithm 4 with different selection of n, as well as its β-variant (cf. (9)). In particular, we set the radius of the ball for projection to be infinity (i.e., the projection is now an identity mapping). Consequently, when $n = \infty$, our implementation of Algorithm 4 becomes a straightforward extension of ETD(0) to the control setting. We use the same behavior policy as the

	$n=\infty$	$n=0$	$n=2$	$n=4$	$n=8$	$\beta=0.8$
$\tau=0$	-	-	10^4	10^3	10^6	10^{11}
$\tau=0.01$	-	-	10^4	10^3	10^6	10^{11}
$\tau=0.1$	-	-	10^4	10^3	10^6	10^{11}

Table 2: Average variance of curves in Figure 3. Here $n=4$ has smaller variance than $n=2$ because the former converges slightly faster.

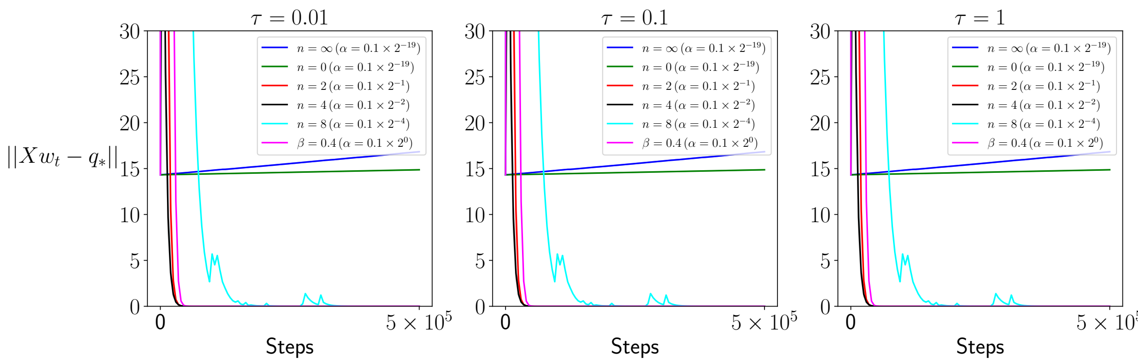

Figure 4: Truncated Emphatic Expected SARSA and its β-variant in the control setting with a changing behavior policy. The shaded regions are invisible for some curves because their standard errors are too small.

prediction setting. The target policy is a softmax policy with a temperature τ:

$$\pi(\texttt{dashed}|s) \doteq \frac{\exp\left(q(s, \texttt{dashed})/\tau\right)}{\exp\left(q(s, \texttt{dashed})/\tau\right) + \exp\left(q(s, \texttt{solid})/\tau\right)}.$$

We test three different temperatures $\tau \in \{0.01, 0.1, 1\}$. When τ approaches 0, the target policies become more and more greedy. Consequently, Algorithm 4 approaches Q-learning. As shown in Figure 3, neither the naive off-policy expected SARSA (i.e., $n=0$) nor the naive extension of ETD(0) (i.e., $n=\infty$) makes any progress in this setting. By contrast, our Truncated Empathic Expected SARSA consistently converges, with lower variance than its β-variant (Table 2).

In the control setting with a changing behavior policy, we still benchmark Algorithm 4 with a different selection of n and its β-variant. The target policy is still the softmax policy

	$n=\infty$	$n=0$	$n=2$	$n=4$	$n=8$	$\beta=0.8$
$\tau=0$	-	-	10^3	10^2	10^6	10^6
$\tau=0.01$	-	-	10^3	10^2	10^6	10^6
$\tau=0.1$	-	-	10^3	10^2	10^6	10^6

Table 3: Average variance of curves in Figure 4. Here $n=4$ has smaller variance than $n=2$ because the former converges slightly faster.

with a temperature τ. The behavior policy is now a mixture policy same as the one used in Zhang et al. (2021b). At each time step, with probability 0.9, the behavior policy is the same as the behavior policy used in the prediction setting; with probability 0.1, the behavior policy is a softmax policy with temperature 1. As shown by Figure 4 and Table 3, the results in this setting are similar to the previous setting with a fixed behavior policy but the variance with $n \in \{2, 4\}$ is reduced. This is because the behavior policy is now related to the target policy, i.e., the off-policyness is reduced.

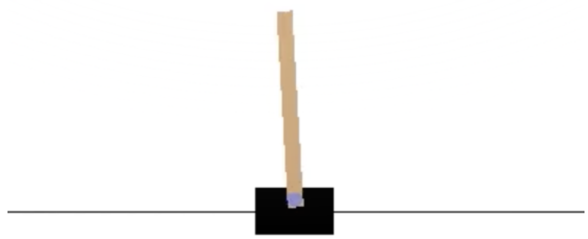

Figure 5: CartPole. At each time step, we observe the `velocity`, `acceleration`, `angular velocity`, and `angular acceleration` of the pole and move the car `left` or `right` to keep the pole balanced. The reward is +1 every time step. An episode ends if a maximum of 1000 steps is reached or the pole falls.

We further evaluate Truncated Emphatic TD methods in the CartPole domain (Figure 5), which is a classical nonsynthetic control problem. We use tile coding (Sutton, 1995) to map the four-dimensional observation (velocity, acceleration, angular velocity, angular acceleration) to a binary vector in $\mathbb{R}^{1024}$ and then apply linear function approximation. In particular, we use the tile coding software recommended in Chapter 10.1 of Sutton and Barto (2018). We benchmark Algorithm 4 and its β-variant (cf. (9)), following the same hyperparameter tuning protocol as in Baird's counterexample. We use $\gamma = 0.99$ and $i(s) = 1$. The target policy is a softmax policy with temperature $\tau = 0.01$. The behavior policy is a ϵ-softmax policy with $\epsilon = 0.95$ and $\tau = 1$. In other words, at each time step, with probability 0.95, the agent selects an action according to a uniformly random policy; with probability 0.05, the agent selects an action according to a softmax policy with temperature $\tau = 1$. We grant large randomness to the behavior policy to enlarge the off-policyness of the problem, making it more challenging. We evaluate the agent every 5×10^3 steps during the training process for 10 episodes and report the averaged undiscounted episodic return. Figure 6 (Left) investigates the effect of different truncation length. We recall that the learning rate α is tuned from Λ_α maximizing the evaluation return at the end of the training. With $n = \infty$ (i.e., no truncation), the agent barely learns anything. With $n = 0$ (i.e., naive off-policy expected SARSA without followon trace), the agent reaches a reasonable performance level but using $n = 4$ performs better. Using $n = 2$ performs better than using $n = 4$ in the middle of the training but the performance drops near the end of the training. We conjecture that this may suggest that a truncation length of 2 is not enough to stabilize the off-policy training in the tested problem. We note that being able to achieve a reasonable performance with $n = 0$ does not mean there is no stability issue with $n = 0$, since divergence to infinity and failing to learn at all is not the only consequence of instability. For example, the iterates can also chatter in a bounded region (Gordon, 1996, 2001), which might be accountable for the early plateau of the curve with $n = 0$. Increasing n improves stability and might help escape from the early plateau. Figure 6 (Right) further investigates the soft truncation using (9). We recall that β is tuned from $\{0.1, 0.2, 0.4, 0.8\}$. Using the soft truncation with $\beta = 0.2$ performs similar to using the hard truncation with $n = 4$. It can, however, be computed that the data points of the curve with $\beta = 0.2$ has an average variance around 1.8×10^4 while that of $n = 4$ is around 7×10^3. This suggests that our proposed hard truncation might be a better option for variance reduction than the existing soft truncation for the tested problem.

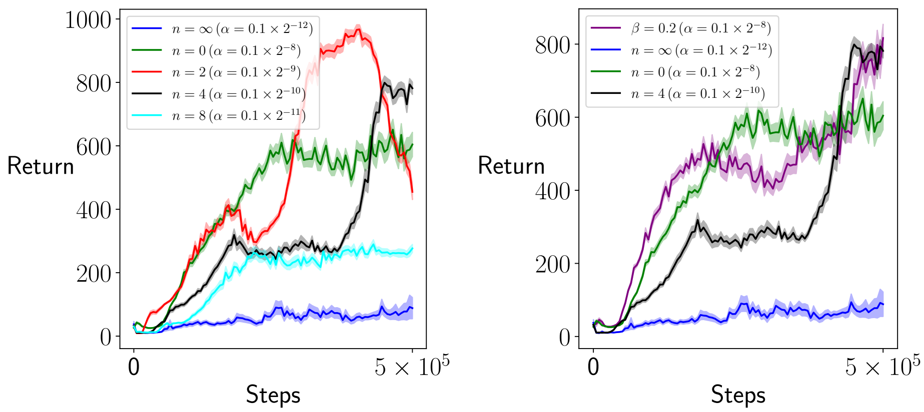

Figure 6: Truncated Emphatic Expected SARSA and its β-variant in the CartPole domain.

9. Conclusion

In this paper, we addressed the two open problems in emphatic TD methods simultaneously by using truncated followon traces. Our analysis is limited to ETD(0) but the extension to ETD(λ) is straightforward, which we leave for future work. The idea of using truncated traces as a variance reduction technique can also be applied to other trace-based off-policy RL algorithms, e.g., GTD(λ) (Maei, 2011), and other variants of followon traces, e.g., ETD(λ, β) and NETD, which we also leave for future work. In this paper, we mainly focused on value-based methods. A possibility for future work is to equip followon-trace-based actor-critic algorithms (e.g., Imani et al. 2018; Zhang et al. 2020b) with the truncated followon trace. Further, similar to the canonical approximate value iteration, bounding the performance of the fixed points of emphatic approximate value iteration also remains an open problem. In this paper, we restricted our empirical study to linear function approximation. Empirically investigating truncated emphatic TD methods with large neural networks like Jiang et al. (2021) is also a possibility for future work.

Acknowledgments

The authors thank the action editor and the anonymous reviewers for their insightful feedback. SZ is generously funded by the Engineering and Physical Sciences Research Council (EPSRC). This project has received funding from the European Research Council under the European Union's Horizon 2020 research and innovation programme (grant agreement number 637713). The experiments were made possible by a generous equipment grant from NVIDIA.

Appendix A. ODE-Based Convergent Results

A.1 Proposition 4.8 of Bertsekas and Tsitsiklis (1996)

Consider the iterates $\{w_t\}$ evolving in $\mathbb{R}^K$ defined as

$$w_{t+1} \doteq w_t + \alpha_t \left(A(O_t) w_t + b(O_t) \right),$$

where $\{O_t\}$ denote a Markov chain in a space $\mathcal{O}$, $\{\alpha_t\}$ is a sequence of learning rates, A and b are functions from $\mathcal{O}$ to $\mathbb{R}^{K \times K}$ and $\mathbb{R}^K$ respectively.

Assumption A.1 *$\{\alpha_t\}$ is a deterministic, positive, nonincreasing sequence such that*

$$\sum_t \alpha_t = \infty, \sum_t \alpha_t^2 < \infty.$$

Assumption A.2 *The chain $\{O_t\}$ adopts a unique invariant distribution denoted by $d_{\mathcal{O}}$.*

Assumption A.3 *The matrix $\bar{A} \doteq \mathbb{E}_{O_t \sim d_{\mathcal{O}}}[A(O_t)]$ is n.d..*

Assumption A.4 $\sup_{o \in \mathcal{O}} \|A(o)\| < \infty, \sup_{o \in \mathcal{O}} \|b(o)\| < \infty$

Assumption A.5 *There exist constants $C_0 > 0$ and $\kappa \in (0, 1)$ such that*

$$\left\| \mathbb{E}\left[A(O_t) \right] - \bar{A} \right\| \leq C_0 \kappa^t,$$
$$\left\| \mathbb{E}\left[b(O_t) \right] - \bar{b} \right\| \leq C_0 \kappa^t,$$

where $\bar{b} \doteq \mathbb{E}_{O_t \sim d_{\mathcal{O}}}[b(O_t)]$.

Theorem 13 *Let Assumptions A.1 - A.5 hold. Then*

$$\lim_{t \to \infty} w_t = \bar{A}^{-1} \bar{b} \quad a.s..$$

A.2 Theorem 13 of Benveniste et al. (1990)

Consider the iterates $\{w_t\}$ evolving in $\mathbb{R}^K$ defined as

$$w_{t+1} \doteq w_t + \alpha_t H(w_t, O_{t+1}), \tag{20}$$

where $\left\{O_t \in \mathbb{R}^L\right\}$ are random variables, $\{\alpha_t\}$ is a sequence of learning rates, H is a function from $\mathbb{R}^K \times \mathbb{R}^L$ to $\mathbb{R}^K$. We use $\mathcal{F}_t$ to denote the σ-field generated by $\{w_0, O_0, O_1, \ldots, O_t\}$ and make the following assumptions:

Assumption A.6 *$\{\alpha_t\}$ is a deterministic, positive, nonincreasing sequence such that*

$$\sum_t \alpha_t = \infty, \sum_t \alpha_t^2 < \infty.$$

Assumption A.7 *There exists a family $\left\{P_w \mid w \in \mathbb{R}^L\right\}$ of parameterized transition probabilities P_w on $\mathbb{R}^L$ such that for any $B \in \mathcal{B}(\mathbb{R}^K)$,*

$$\Pr(O_{t+1} \in B | \mathcal{F}_t) = P_{w_t}(O_t, B).$$

Additionally, for any function f defined on $\mathbb{R}^L$, we define $(P_w f)(o) \doteq \int f(y) P_w(o, dy)$. Here $\mathcal{B}(\cdot)$ denotes the Borel sets.

Assumption A.8 *Let D be an open subset of $\mathbb{R}^K$. For any compact subset Q of D, there exists constants C_1, q_1 (depending on Q), such that for any $w \in Q$ and any o, we have*

$$\|H(w, o)\| \leq C_1(1 + \|o\|^{q_1}).$$

Assumption A.9 *There exists a function $h : D \to \mathbb{R}^K$, and for each $w \in D$ a function $\nu_w : \mathbb{R}^L \to \mathbb{R}^K$, such that*

(i) h is locally Lipschitz continuous on D

(ii) $\nu_w(o) - (P_w \nu_w)(o) = H(w, o) - h(w)$ holds for all $w \in D, o \in \mathbb{R}^L$

(iii) for all compact subsets Q of D, there exist constants C_2, C_3, q_2, q_3 (depending on Q), such that for all $w, w' \in Q, z \in \mathbb{R}^L$,

$$\|\nu_w(o)\| < C_2(1 + \|o\|^{q_2}),$$
$$\|(P_w \nu_w)(o) - (P_{w'} \nu_{w'})(o)\| \leq C_3 \big\|w - w'\big\| (1 + \|o\|^{q_3}).$$

Assumption A.10 *For any compact subset Q of D and any $q > 0$, there exists constant C_4 (depending on Q, q) such that for all t, $o \in \mathbb{R}^L, w \in \mathbb{R}^K$,*

$$\mathbb{E}\left[\mathbb{I}(\{w_k \in Q, k \leq t\})(1 + \|O_{t+1}\|^q) \mid O_0 = o, w_0 = w\right] \leq C_4(1 + \|o\|^q),$$

where $\mathbb{I}$ is the indicator function.

Assumption A.11 *There exist a function $U \in \mathcal{C}^2(\mathbb{R}^K)$ and $w_* \in D$ such that*

(i) $U(w) \to C \leq +\infty$ if $w \to \partial D$ or $\|w\| \to \infty$

(ii) $U(w) < C$ for all $w \in D$

(iii) $U(w) \geq 0$, where the equality holds i.f.f. $w = w_*$

(iv) $\left\langle \frac{\mathrm{d}U(w)}{\mathrm{d}w}, h(w) \right\rangle \leq 0$ for all $w \in D$, where the equality holds i.f.f. $w = w_*$.

Theorem 14 *(Theorem 13 of Benveniste et al. (1990) (p. 236)) Let Assumptions A.6 - A.11 hold. For any compact $Q \subset D$, there exist constants C_0, q_0 such that for all $w \in Q, o \in \mathbb{R}^L$, the iterates $\{w_t\}$ generated by (20) satisfy*

$$\Pr\Big(\lim_{t \to \infty} w_t = w_* \mid O_0 = o, w_0 = w\Big) \geq 1 - C_0(1 + \|o\|^{q_0}) \sum_{t=0}^{\infty} \alpha_t^2.$$

A.3 Theorem 13 of Benveniste et al. (1990) with a Finite Chain

Consider the iterates $\{w_t\}$ evolving in $\mathbb{R}^K$ defined as

$$w_{t+1} \doteq w_t + \alpha_t \bar{H}(w_t, O_{t+1}), \tag{21}$$

where $\{O_t\}$ are random variables evolving in a *finite* space $\mathcal{O}$, $\{\alpha_t\}$ is a sequence of learning rates, $\bar{H}$ is a function from $\mathbb{R}^K \times \mathcal{O}$ to $\mathbb{R}^K$. Without loss of generality, let $\mathcal{O} \doteq \{1, 2, \dots, N\} \subset \mathbb{R}$. We make the following assumptions.

Assumption A.12 $\{\alpha_t\}$ *is a deterministic, positive, nonincreasing sequence such that*

$$\sum_t \alpha_t = \infty, \sum_t \alpha_t^2 < \infty.$$

Assumption A.13 *There exists a family* $\left\{\bar{P}_w \in \mathbb{R}^{N \times N} \mid w \in \mathbb{R}^K\right\}$ *of parameterized transition matrices such that the random variables* $\{O_t\}$ *evolve according to*

$$O_{t+1} \sim \bar{P}_{w_t}(O_t, \cdot)$$

Let Λ_w *be the closure of* $\left\{\bar{P}_w \mid w \in \mathbb{R}^K\right\}$*, for any* $P \in \Lambda_w$*, the Markov chain in* $\mathcal{O}$ *induced by the transition matrix* P *is ergodic. We use* d_P *to denote the invariant distribution of the chain induced by* P*. In particular,* d_w *denotes the invariant distribution of the chain induced by* $\bar{P}_w$*. We define*

$$h(w) \doteq \sum_{o \in \mathcal{O}} d_w(o) \bar{H}(w, o).$$

Assumption A.14 $\bar{P}_w$ *is Lipschitz continuous in* w*. For any compact* $Q \subset \mathbb{R}^K$ *and any* $o \in \mathcal{O}$*,* $\bar{H}(w, o)$ *is Lipschitz continuous in* w *on* Q*.*

Assumption A.15 *There exist function* $U \in \mathcal{C}^2(\mathbb{R}^K)$ *and* $w_* \in \mathbb{R}^K$ *such that*

(i) $U(w) \to \infty$ *when* $\|w\| \to \infty$

(ii) $U(w) < \infty$ *for all* $w \in \mathbb{R}^K$

(iii) $U(w) \geq 0$*, where the equality holds i.f.f.* $w = w_*$

(iv) $\left\langle \frac{\mathrm{d}U(w)}{\mathrm{d}w}, h(w) \right\rangle \leq 0$ *for all* $w \in \mathbb{R}^K$*, where the equality holds i.f.f.* $w = w_*$*.*

Corollary 15 *Under Assumptions A.12 - A.15, for any compact set* $Q \subset \mathbb{R}^K$*, there exists constants* C_0 *(depending on* Q*) such that for all* $w \in Q, o \in \mathcal{O}$*, the iterates* $\{w_t\}$ *generated by* (21) *satisfy*

$$\Pr\Big(\lim_{t \to \infty} w_t = w_* \mid O_0 = o, w_0 = w\Big) \geq 1 - C_0 \sum_{t=0}^{\infty} \alpha_t^2.$$

Proof We proceed by expressing (21) in the form of (20) and invoking Theorem 14. Let

$$H(w, o) \doteq \begin{cases} \bar{H}(w, o) & o \in \mathcal{O} \\ h(w) & o \notin \mathcal{O} \end{cases}.$$

Then (21) can be rewritten as

$$w_{t+1} \doteq w_t + \alpha_t H(w_t, O_{t+1}),$$

which has the same form as (20). Here the L in $\mathbb{R}^L$ is 1 and we consider D to be $\mathbb{R}^K$.

Assumption A.6 is identical to Assumption A.12. Assumption A.11 is implied by Assumption A.15 via considering $C = \infty$.

To verify Assumption A.7, let

$$P_w(o, B) \doteq \begin{cases} \sum_{o'} \delta_{o'}(B) \bar{P}_w(o, o') & o \in \mathcal{O} \\ \mathcal{N}(B) & o \notin \mathcal{O} \end{cases},$$

where $\delta_{o'}(B)$ is the Dirac measure, $\mathcal{N}(\cdot)$ denotes the normal distribution (we can use any well-defined distribution on $\mathbb{R}$ here). Then Assumption A.7 follows from Assumption A.13.

We now verify Assumption A.8. From Assumption A.14 and the finiteness of $\mathcal{O}$, for any compact Q, $\bar{H}(w, o)$ is bounded on Q. So $h(w)$ is also bounded on Q. Then the boundedness of $H(w, o)$ on Q follows immediately.

We now verify Assumption A.9(i). First, for any compact $Q \subset \mathbb{R}^K$, $\bar{H}(w, o)$ is Lipschitz continuous in w and bounded on Q. $d_w(o)$ is apparently bounded. By Assumption A.13, for any $P \in \Lambda_w$, the chain induced by P is ergodic. It can be easily proved (see, e.g., Lemma 9 of Zhang et al. (2021b)) that d_w is also Lipschitz continuous in w. The Lipschitz continuity of $h(w)$ on Q then follows immediately from the fact that the product of two bounded Lipschitz functions are still bounded and Lipschitz. Since we are free to choose any Q, $h(w)$ is locally Lipschitz continuous in $\mathbb{R}^K$.

We verify Assumption A.9(ii) by constructing auxiliary Markov Reward Processes (MRPs) and using standard properties of MRPs. To construct the i-th MRP ($i = 1, \dots, K$), let $H_{w,i}$ denote a vector in $\mathbb{R}^N$ whose i-th element is $H_i(w, o)$, the i-th element of $H(w, o)$. For any $w \in \mathbb{R}^K, o \in \mathcal{O}$, we define a vector $\bar{\nu}_w(o)$ in $\mathbb{R}^K$ by defining its i-th element $\bar{\nu}_{w,i}(o)$ as

$$\bar{\nu}_{w,i}(o) \doteq \mathbb{E}\left[\sum_{k=0}^{\infty} [H_{w,i}(O_k) - h_i(w)] \mid O_0 = o, O_{k+1} \sim P_w(O_k, \cdot)\right],$$

where $h_i(w)$ is the i-the element of $h(w)$. By definition, $\bar{\nu}_{w,i}$ is the bias vector of the MRP induced by $\bar{P}_w$ in $\mathcal{O}$ with the reward vector being $H_{w,i}$. Since $\bar{P}_w$ induces an ergodic chain under Assumption A.13, $\bar{\nu}_w$ is always well defined. Moreover, $h_i(w)$ is the gain of this MRP. It follows from Chapter 8.2.1 of Puterman (2014) that for any $w \in \mathbb{R}^K$ and $o \in \mathcal{O}$,

$$\bar{\nu}_{w,i}(o) = H_{w,i}(o) - h_i(w) + \sum_{o'} \bar{\nu}_{w,i}(o') \bar{P}_w(o, o') \tag{22}$$

$$\bar{\nu}_{w,i} = H_{\bar{P}_w} H_{w,i},$$

where $H_P \doteq (I - P + 1d_P^\top)^{-1}(I - 1d_P^\top)$ is the fundamental matrix of the chain induced by a transition matrix P. Define

$$\nu_w(o) \doteq \begin{cases} \bar\nu_w(o) & o \in \mathcal{O} \\ 0 & o \notin \mathcal{O} \end{cases}.$$

It is then easy to verify that for $o \in \mathcal{O}$,

$$(P_w \nu_w)(o) = \int \nu_w(y) P_w(o, dy) = \int \nu_w(y) \sum_{o'} \delta_{o'}(dy) \bar P_w(o, o') = \sum_{o'} \nu_w(o') \bar P_w(o, o').$$

For $o \notin \mathcal{O}$, $(P_w \nu_w)(o) = 0$. For $o \in \mathcal{O}$, Assumption A.9(ii) holds since it is just (22). For $o \notin \mathcal{O}$, Assumption A.9(ii) holds as well since since both LHS and RHS are 0.

We now verify Assumption A.9(iii). Since d_P is Lipschitz continuous in P for all $P \in \Lambda_w$ and Λ_w is compact, we have $\sup_{P \in \Lambda_w} \left\|(I - P + 1d_P^\top)^{-1}\right\| < \infty$ by the extreme value theorem. Using the ergodicity of the chain induced by P and

$$\left\|X^{-1} - Y^{-1}\right\| = \left\|X^{-1} Y Y^{-1} - X^{-1} X Y^{-1}\right\| \leq \left\|X^{-1}\right\| \left\|X - Y\right\| \left\|Y^{-1}\right\|,$$

it is then easy to see that $H_{\bar P_w}$ is bounded and Lipschitz continuous in w. Consequently, for any compact $Q \subset \mathbb{R}^K, w \in Q, w' \in Q$, ν_w is bounded on Q and

$$\begin{aligned} \left\|\bar\nu_{w,i} - \bar\nu_{w',i}\right\| &\leq \left\|H_{\bar P_w} - H_{\bar P_{w'}}\right\| \left\|H_{w,i}\right\| + \left\|H_{\bar P_{w'}}\right\| \left\|H_{w,i} - H_{w',i}\right\| \\ &\leq C_1 \left\|w - w'\right\| + C_2 \left\|w - w'\right\|, \end{aligned}$$

where the constant C_1 results from the Lipschitz continuity of $H_{\bar P_w}$ and the boundedness of $H(w, o)$ on Q, the constant C_2 results from the Lipschitz continuity of $H(w, o)$ on Q and the boundedness of $H_{\bar P_w}$. Since by Assumption A.14, $\bar P_w$ in Lipschitz continuous in w, the Lipschitz continuity of $P_w \nu_w$ in Q follows immediately, which completes the verification of Assumption A.9(iii).

Assumption A.10 is trivial since $\mathcal{O}$ is finite, which completes the proof. ∎

Appendix B. Proofs

B.1 Proofs of Lemmas 1

Proof Let $\tau_j \doteq (s_j, a_j, s_{j-1}, a_{j-1}, \dots s_1, a_1)$, $\Gamma_j \doteq (S_{t-j}, A_{t-j}, \dots, S_{t-1}, A_{t-1})$,

$$
\begin{aligned}
m_{t,n}(s) &\doteq \mathbb{E}[F_t^n | S_t = s] \qquad (23)\\
&= \sum_{j=0}^n \gamma^j \mathbb{E}\left[\rho_{t-j:t-1} i_{t-j} \middle| S_t = s\right] \\
&= \sum_{j=0}^n \gamma^j \sum_{\tau_j \in (\mathcal{S}\times\mathcal{A})^j} \Pr(\Gamma_j = \tau_j | S_t = s) \mathbb{E}[\rho_{t-j:t-1} i_{t-j} | \Gamma_j = \tau_j, S_t = s] \\
& \qquad \text{(Law of total expectation)} \\
&= \sum_{j=0}^n \gamma^j \sum_{\tau_j \in (\mathcal{S}\times\mathcal{A})^j} \frac{\Pr(\Gamma_j = \tau_j, S_t = s)}{\Pr(S_t = s)} \mathbb{E}[\rho_{t-j:t-1} i_{t-j} | \Gamma_j = \tau_j, S_t = s] \\
& \qquad \text{(Bayes' rule)} \\
&= \sum_{j=0}^n \gamma^j \sum_{\tau_j \in (\mathcal{S}\times\mathcal{A})^j} \frac{\Pr(\Gamma_j = \tau_j, S_t = s)}{\Pr(S_t = s)} i(s_j) \rho(s_j, a_j) \cdots \rho(s_1, a_1) \\
&= \sum_{j=0}^n \gamma^j \sum_{\tau_j \in (\mathcal{S}\times\mathcal{A})^j} \frac{\Pr(S_{t-j} = s_j) P_\pi(s_j, s_{j-1}) \cdots P_\pi(s_2, s_1) P_\pi(s_1, s)}{\Pr(S_t = s)} i(s_j) \\
&= \sum_{j=0}^n \gamma^j \sum_{s_j} \frac{\Pr(S_{t-j} = s_j) P_\pi^j(s_j, s)}{\Pr(S_t = s)} i(s_j)
\end{aligned}
$$

Assumption 4.1 implies

$$\lim_{t\to\infty} \Pr(S_t = s) = d_\mu(s).$$

Consequently,

$$
\begin{aligned}
m_n(s) &= \lim_{t\to\infty} m_{t,n}(s) \\
&= \sum_{j=0}^n \gamma^j \sum_{s_j \in \mathcal{S}} \frac{d_\mu(s_j) P_\pi^j(s_j, s)}{d_\mu(s)} i(s_j).
\end{aligned}
$$

In a matrix form,

$$
\begin{aligned}
m_n &= \sum_{j=0}^n \gamma^j D_\mu^{-1} (P_\pi^\top)^j D_\mu i, \\
m &= \lim_{n\to\infty} m_n = D_\mu^{-1} (I - \gamma P_\pi^\top)^{-1} D_\mu i.
\end{aligned}
$$

■

B.2 Proof of Lemma 2

Proof

$$\begin{aligned}
m_n - m &= \sum_{j=n+1}^{\infty} \gamma^j D_\mu^{-1} (P_\pi^\top)^j D_\mu i \\
&= D_\mu^{-1} \left(\sum_{j=n+1}^{\infty} \gamma^j (P_\pi^\top)^j \right) D_\mu i \\
&= D_\mu^{-1} \gamma^{n+1} (P_\pi^\top)^{n+1} (I - \gamma P_\pi^\top)^{-1} D_\mu i \\
&= \gamma^{n+1} D_\mu^{-1} (P_\pi^\top)^{n+1} D_\mu m,
\end{aligned}$$

implying

$$\begin{aligned}
\|m_n - m\|_1 &\leq \gamma^{n+1} \left\|D_\mu^{-1}\right\|_1 \|D_\mu\|_1 \|m\|_1 \quad (\text{Using } \left\|P_\pi^\top\right\|_1 = \|P_\pi\|_\infty = 1) \\
&= \gamma^{n+1} \frac{d_{\mu,max}}{d_{\mu,min}} \|m\|_1, \\
\|f_n - f\|_\infty &\leq \|D_\mu\|_\infty \|m_n - m\|_\infty \leq d_{\mu,max} \|m_n - m\|_1 \leq d_{\mu,max} \|m_n - m\|_\infty \\
&= \gamma^{n+1} \frac{d_{\mu,max}^2}{d_{\mu,min}} \|m\|_1
\end{aligned}$$

■

B.3 Proof of Lemma 3

Proof Let D_{f_n} be a diagonal matrix whose diagonal entry is f_n and $D_{f_n - f}$ be a diagonal matrix whose diagonal entry is $f_n - f$. We have

$$\begin{aligned}
y^\top D_{f_n} (\gamma P_\pi - I) y &= y^\top D_{f_n - f} (\gamma P_\pi - I) y + y^\top D_f (\gamma P_\pi - I) y \\
&\leq \|y\|^2 \|D_{f_n - f}\| \|\gamma P_\pi - I\| + y^\top D_f (\gamma P_\pi - I) y \\
&\leq \|y\|^2 \|f_n - f\|_\infty \|\gamma P_\pi - I\| + y^\top D_f (\gamma P_\pi - I) y
\end{aligned}$$

Similarly,

$$y^\top (\gamma P_\pi^\top - I) D_{f_n} y \leq \|y\|^2 \|f_n - f\|_\infty \left\|\gamma P_\pi^\top - I\right\| + y^\top (\gamma P_\pi^\top - I) D_f y.$$

Combining the above two inequalities together, we get

$$\begin{aligned}
&\frac{1}{2} y^\top \left(D_{f_n} (\gamma P_\pi - I) + (\gamma P_\pi^\top - I) D_{f_n} \right) y \\
\leq &\|f_n - f\|_\infty \|\gamma P_\pi - I\| \|y\|^2 + \frac{1}{2} y^\top \left(D_f (\gamma P_\pi - I) + (\gamma P_\pi^\top - I) D_f \right) y
\end{aligned}$$

(Invariance of ℓ_2 norm under transpose)

$$\le (\|f_n - f\|_\infty \|\gamma P_\pi - I\| - \lambda_{min})\|y\|^2$$

(Eigendecomposition of real symmetric matrices)

$$\le (\gamma^{n+1} \frac{d_{\mu,max}^2}{d_{\mu,min}} \|m\|_1 \|\gamma P_\pi - I\| - \lambda_{min})\|y\|^2 \quad \text{(Lemma 2)}$$

As long as the condition (13) holds, the above inequality asserts that

$$D_{f_n}(\gamma P_\pi - I) + (\gamma P_\pi^\top - I)D_{f_n}$$

is n.d., implying $D_{f_n}(\gamma P_\pi - I)$ is n.d.. This together with Assumption 4.3 completes the proof. ■

B.4 Proof of Theorem 4

Proof Let $O_t \doteq (S_{t-n}, A_{t-n}, \dots, S_t, A_t, S_{t+1})$ be a sequence of random variables generated by Algorithm 1. Let $o_t \doteq (s_{t-n}, a_{t-n}, \dots, s_t, a_t, s_{t+1})$ and define functions

$$A(o_t) \doteq \left(\sum_{j=0}^{n} \gamma^j \left(\prod_{k=t-j}^{t-1} \frac{\pi(a_k|s_k)}{\mu(a_k|s_k)}\right) i(s_{t-j})\right) \frac{\pi(a_t|s_t)}{\mu(a_t|s_t)} x(s_t) \left(\gamma x(s_{t+1})^\top - x(s_t)^\top\right),$$
$$b(o_t) \doteq \left(\sum_{j=0}^{n} \gamma^j \left(\prod_{k=t-j}^{t-1} \frac{\pi(a_k|s_k)}{\mu(a_k|s_k)}\right) i(s_{t-j})\right) \frac{\pi(a_t|s_t)}{\mu(a_t|s_t)} x(s_t) r(s_t, a_t).$$

Here o_t is just placeholder for defining $A(o_t)$ and $b(o_t)$. Then the update for $\{w_t\}$ in Algorithm 1 can be expressed as

$$w_{t+1} = w_t + \alpha_t \left(A(O_t) w_t + b(O_t)\right).$$

We now proceed to confirming its convergence via verifying Assumptions A.1 - A.5 thus invoking Theorem 13.

Assumption A.1 is identical to Assumption 4.4. Assumption A.2 follows directly from Assumption 4.1. And it is easy to see the invariant distribution of $\{O_t\}$ is

$$d_{\mathcal{O}}(o_t) = d_\mu(s_{t-n})\mu(a_{t-n}|s_{t-n})p(s_{t-n+1}|s_{t-n}, a_{t-n}) \cdots p(s_{t+1}|s_t, a_t). \tag{24}$$

Moreover,

$$\begin{aligned}
\bar{A} \doteq & \mathbb{E}_{O_t \sim d_{\mathcal{O}}}\left[A(O_t)\right] \\
= & \mathbb{E}_{O_t \sim d_{\mathcal{O}}}\left[F_{t,n}\rho_t x_t(\gamma x_{t+1}^\top - x_t^\top)\right] \\
= & \sum_{s,a,s'} d_\mu(s)\mu(a|s)p(s'|s,a)\mathbb{E}\left[F_{t,n}\rho_t x_t(\gamma x_{t+1}^\top - x_t^\top) \middle| S_t = s, A_t = a, S_{t+1} = s'\right]
\end{aligned}$$

(Law of total expectation)

$$
\begin{aligned}
&= \sum_{s,a,s'} d_\mu(s)\pi(a|s)p(s'|s,a)\mathbb{E}\left[F_{t,n}|S_t = s\right] x(s)(\gamma x(s')^\top - x(s)^\top) \\
&\hspace{20em} \text{(Conditional independence and Markov property)} \\
&= \sum_{s,a,s'} d_\mu(s)\pi(a|s)p(s'|s,a)m_n(s)x(s)(\gamma x(s')^\top - x(s)^\top) \quad \text{(Using (23) and (24))} \\
&= \sum_{s,a,s'} f_n(s)\pi(a|s)p(s'|s,a)x(s)(\gamma x(s')^\top - x(s)^\top) \quad \text{(Definition of } f_n\text{)} \\
&= X^\top D_{f_n}(\gamma P_\pi - I)X.
\end{aligned}
$$

In the above equation, we have abused the notation slightly to use O_t to denote random variables sampled from $d_\mathcal{O}$. Similarly, it can be shown that

$$
\bar{b} \doteq \mathbb{E}_{O_t \sim d_\mathcal{O}}\left[b(O_t)\right] = X^\top D_{f_n} r_\pi.
$$

Lemma 3 confirms that $\bar{A}$ is n.d., verifying Assumption A.3. Assumption A.4 is obvious since $|\mathcal{S}|, |\mathcal{A}|, n$ are all finite. Assumption A.5 follows immediately from the geometrically mixing rate of ergodic Markov chain. For example,

$$
\begin{aligned}
\left\|\mathbb{E}[A(O_t)] - \bar{A}\right\| &= \left\| \sum_{o \in \mathcal{O}} \left(\Pr(O_t = o) - d_\mathcal{O}(o)\right) A(o) \right\| \\
&\leq \max_o \|A(o)\| \sum_o |\Pr(O_t = o) - d_\mathcal{O}(o)| \\
&\leq C_0 \kappa^t,
\end{aligned}
$$

for some $C_0 > 0$ and $\kappa \in (0, 1)$. Here the last inequality is a standard result, see, e.g., Theorem 4.9 of Levin and Peres (2017).

Note this procedure cannot be used to verify the convergence of the original ETD(0), where we would need to consider $O_t = (F_t, S_t, A_t)$. Since F_t involves in $\mathbb{R}$, Assumption A.4 cannot be verified. ■

B.5 Proof of Lemma 6

Proof Since $i(s) > 0$ holds for any s and P_π is nonnegative, from Lemma 1 it is easy to see for any $n_1 > n_2$,

$$
m_{n_1}(s) > m_{n_2}(s)
$$

always holds. Then by the definition of f_n,

$$
f_{n_1}(s) > f_{n_2}(s)
$$

holds as well. In particular, for any $n \geq 1$,

$$f(s) > f_n(s) > f_0(s) = d_\mu(s) i(s) > 0.$$

For any v, we have

$$\begin{aligned}
\gamma \|P_\pi v\|_{f_n}^2 &= \gamma \sum_s f_n(s) \left(\sum_{s'} P_\pi(s, s') v(s') \right)^2 \\
&\leq \gamma \sum_s f_n(s) \sum_{s'} P_\pi(s, s') v^2(s') \quad \text{(Jensen's inequality)} \\
&= \gamma \sum_{s'} v^2(s') \sum_s f_n(s) P_\pi(s, s') \\
&= v^\top diag(\gamma P_\pi^\top f_n) v \\
&= v^\top diag\Big(f_n - (I - \gamma P_\pi^\top) f_n \Big) v \\
&= v^\top diag\Big(f_n - (I - \gamma P_\pi^\top) f + (I - \gamma P_\pi^\top)(f - f_n) \Big) v \\
&= \|v\|_{f_n}^2 - v^\top diag\Big((I - \gamma P_\pi^\top) f \Big) v + v^\top diag\Big((I - \gamma P_\pi^\top)(f - f_n) \Big) v \\
&= \|v\|_{f_n}^2 - \|v\|_{f_0}^2 + v^\top diag\Big((I - \gamma P_\pi^\top)(f - f_n) \Big) v \\
&\qquad \text{(Using } (I - \gamma P_\pi^\top) f = (I - \gamma P_\pi^\top)(I - \gamma P_\pi^\top)^{-1} D_\mu i = f_0 \text{)} \\
&\leq \|v\|_{f_n}^2 - \|v\|_{f_0}^2 + \Big\| (I - \gamma P_\pi^\top)(f - f_n) \Big\|_\infty \|v\|^2 \\
&\qquad \text{(Property of } \ell_2 \text{ norm of a diagonal matrix)} \\
&\leq \|v\|_{f_n}^2 - \|v\|_{f_0}^2 + \Big\| (I - \gamma P_\pi^\top) \Big\|_\infty \gamma^{n+1} \frac{d_{\mu,max}^2}{d_{\mu,min}} \|m\|_1 \|v\|^2 \quad \text{(Lemma 2)} \\
&\leq \|v\|_{f_n}^2 - \|v\|_{f_0}^2 + \kappa \min_s i(s) d_\mu(s) \|v\|^2 \quad \text{(Using (14))} \\
&\leq \|v\|_{f_n}^2 - \|v\|_{f_0}^2 + \kappa \|v\|_{f_n}^2 \quad \text{(Using } \min_{s'} i(s') d_\mu(s') \leq f_0(s) < f_n(s) \text{)} \\
&= (1 + \kappa) \|v\|_{f_n}^2 - \|v\|_{f_0}^2 \\
&= (1 + \kappa) \|v\|_{f_n}^2 - \sum_s v(s)^2 d_\mu(s) i(s) \\
&= (1 + \kappa) \|v\|_{f_n}^2 - \sum_s v(s)^2 f(s) \frac{d_\mu(s) i(s)}{f(s)} \\
&\leq (1 + \kappa) \|v\|_{f_n}^2 - \kappa \sum_s v(s)^2 f(s) \quad \text{(Definition of } \kappa \text{ and } f(s) > 0\text{)} \\
&\leq (1 + \kappa) \|v\|_{f_n}^2 - \kappa \sum_s v(s)^2 f_n(s) \quad \text{(Using } f(s) > f_n(s)\text{)} \\
&= \|v\|_{f_n}^2
\end{aligned}$$

Consequently,

$$\|\mathcal{T}_\pi v_1 - \mathcal{T}_\pi v_2\|_{f_n}^2 = \gamma^2 \|P_\pi(v_1 - v_2)\|_{f_n}^2 \leq \gamma \|v_1 - v_2\|_{f_n}^2,$$

implying that $\mathcal{T}_\pi$ is a $\sqrt{\gamma}$-contraction in $\|\cdot\|_{f_n}$. Since Π_{f_n} is nonexpansive in $\|\cdot\|_{f_n}$, it is easy to see that $\Pi_{f_n}\mathcal{T}_\pi$ is a $\sqrt{\gamma}$ contraction in $\|\cdot\|_{f_n}$ as well.

Further, let $D_{f_n} \doteq diag(f_n)$, it is easy to verify that

$$\Pi_{f_n} = X(X^\top D_{f_n} X)^{-1} X^\top D_{f_n}.$$

Consequently,

$$\begin{aligned} & A_n w_{*,n} = b_n \\ \iff & X^\top D_{f_n}(\gamma P_\pi - I) X w_{*,n} = X^\top D_{f_n} r_\pi \\ \iff & X^\top D_{f_n}(r_\pi + \gamma P_\pi X w_{*,n}) = X^\top D_{f_n} X w_{*,n} \\ \iff & X(X^\top D_{f_n} X)^{-1} X^\top D_{f_n}(r_\pi + \gamma P_\pi X w_{*,n}) = X(X^\top D_{f_n} X)^{-1} X^\top D_{f_n} X w_{*,n} \\ \iff & \Pi_{f_n}\mathcal{T}_\pi(X w_{*,n}) = X w_{*,n}. \end{aligned} \tag{25}$$

Then,

$$\begin{aligned} \|X w_{*,n} - v_\pi\|_{f_n}^2 &= \|X w_{*,n} - \Pi_{f_n} v_\pi\|_{f_n}^2 + \|\Pi_{f_n} v_\pi - v_\pi\|_{f_n}^2 \\ & \qquad \text{(Pythagorean theorem)} \\ &= \|\Pi_{f_n}\mathcal{T}_\pi(X w_{*,n}) - \Pi_{f_n}\mathcal{T}_\pi v_\pi\|_{f_n}^2 + \|\Pi_{f_n} v_\pi - v_\pi\|_{f_n}^2 \\ &\leq \gamma \|X w_{*,n} - v_\pi\|_{f_n}^2 + \|\Pi_{f_n} v_\pi - v_\pi\|_{f_n}^2. \end{aligned}$$

Rearranging terms completes the proof. ■

B.6 Proof of Lemma 7

Proof If (16) holds, then Lemma 6 implies that for any $u \in \Lambda_M$ and $\pi \in \Lambda_\pi$, $\Pi_{f_{n,\mu,\pi}}\mathcal{T}_\pi$ is a $\sqrt{\gamma}$-contraction in $\|\cdot\|_{f_{n,\mu,\pi}}$. We use $X w_{n,\mu,\pi}$ to denote its unique fixed point. Lemma 3 ensures that $X^\top D_{f_{n,\mu,\pi}}(I - \gamma P_\pi) X$ is p.d.. Similar to (25), it is easy to verify that

$$w_{n,\mu,\pi} = (X^\top D_{f_{n,\mu,\pi}}(I - \gamma P_\pi) X)^{-1} X^\top D_{f_{n,\mu,\pi}} r_\pi,$$

from which it is easy to see $w_{n,\mu,\pi}$ is continuous in μ and π since the invariant distribution d_μ is continuous in μ.

Similar to De Farias and Van Roy (2000), we first define several helper functions. For any policy $\mu \in \Lambda_M$, $\pi \in \Lambda_\Pi$, and $\eta > 0$, let

$$\begin{aligned} g_{\mu,\pi}(w) &\doteq X^\top D_{f_{n,\mu,\pi}}(\mathcal{T}_\pi Xw - Xw) \\ &= X^\top D_{f_{n,\mu,\pi}} X (X^\top D_{f_{n,\mu,\pi}} X)^{-1} X^\top D_{f_{n,\mu,\pi}} (\mathcal{T}_\pi Xw - Xw) \\ &= X^\top D_{f_{n,\mu,\pi}} (\Pi_{f_{n,\mu,\pi}} \mathcal{T}_\pi Xw - Xw), \\ g(w) &\doteq X^\top D_{f_{n,\mu_w,\pi_w}} (\mathcal{T}_{\pi_w} Xw - Xw) \\ &= X^\top D_{f_{n,\mu_w,\pi_w}} (\Pi_{f_{n,\mu_w,\pi_w}} \mathcal{T}_{\pi_w} Xw - Xw), \\ z^\eta_{\mu,\pi}(w) &\doteq w + \eta g_{\mu,\pi}(w), \\ z^\eta(w) &\doteq w + \eta g(w). \end{aligned}$$

We have

$$\begin{aligned} & z^\eta_{\mu,\pi}(w) = w \\ \iff & g_{\mu,\pi}(w) = 0 \\ \iff & X^\top D_{f_{n,\mu,\pi}} \mathcal{T}_\pi Xw = X^\top D_{f_{n,\mu,\pi}} Xw \\ \iff & X(X^\top D_{f_{n,\mu,\pi}} X)^{-1} X^\top D_{f_{n,\mu,\pi}} \mathcal{T}_\pi Xw = X(X^\top D_{f_{n,\mu,\pi}} X)^{-1} X^\top D_{f_{n,\mu,\pi}} Xw \\ \iff & \Pi_{f_{n,\mu,\pi}} \mathcal{T}_\pi Xw = Xw, \end{aligned}$$

i.e., Xw is a fixed point of $\Pi_{f_{n,\mu,\pi}} \mathcal{T}_\pi$ if and only if w is a fixed point of $z^\eta_{\mu,\pi}$. With the same procedure, we can also show

$$z^\eta(w) = w \iff \Pi_{f_{n,\mu_w,\pi_w}} \mathcal{T}_{\pi_w}(Xw) = Xw.$$

This suggests that to study the fixed points of emphatic approximate value iteration is to study the fixed points of z^η with any $\eta > 0$. To this end, we first study $z^\eta_{\mu,\pi}$ with the following lemma, which is analogous to Lemma 5.4 of De Farias and Van Roy (2000).

Lemma 16 *There exists an $\eta_0 > 0$ such that for all $\eta \in (0, \eta_0)$, there exists a constant $\beta_\eta \in (0,1)$ such that for all $\mu \in \Lambda_M, \pi \in \Lambda_\Pi$,*

$$\left\| z^\eta_{\mu,\pi}(w) - w_{n,\mu,\pi} \right\| \leq \beta_\eta \| w - w_{n,\mu,\pi} \|.$$

Proof By the contraction property of $\Pi_{f_{n,\mu,\pi}} \mathcal{T}_\pi$,

$$\left\| \Pi_{f_{n,\mu,\pi}} \mathcal{T}_\pi Xw - Xw_{n,\mu,\pi} \right\|_{f_{n,\mu,\pi}} \leq \sqrt{\gamma} \| Xw - Xw_{n,\mu,\pi} \|_{f_{n,\mu,\pi}}.$$

Consequently,

$$\begin{aligned} & (w - w_{n,\mu,\pi})^\top g_{\mu,\pi}(w) \\ =& (Xw - Xw_{n,\mu,\pi})^\top D_{f_{n,\mu,\pi}} (\Pi_{f_{n,\mu,\pi}} \mathcal{T}_\pi Xw - Xw) \\ =& (Xw - Xw_{n,\mu,\pi})^\top D_{f_{n,\mu,\pi}} \left(\Pi_{f_{n,\mu,\pi}} \mathcal{T}_\pi Xw - Xw_{n,\mu,\pi} + Xw_{n,\mu,\pi} - Xw \right) \\ \leq& \| Xw - Xw_{n,\mu,\pi} \|_{f_{n,\mu,\pi}} \left\| \Pi_{f_{n,\mu,\pi}} \mathcal{T}_\pi Xw - Xw_{n,\mu,\pi} \right\|_{f_{n,\mu,\pi}} - \| Xw - Xw_{n,\mu,\pi} \|^2_{f_{n,\mu,\pi}} \end{aligned}$$

$$
\begin{aligned}
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(Cauchy-Schwarz inequality)}\\
\leq& (\sqrt{\gamma}-1)\|Xw - Xw_{n,\mu,\pi}\|^2_{f_{n,\mu,\pi}} \quad \text{(Property of contraction)}\\
=& (\sqrt{\gamma}-1)(w - w_{n,\mu,\pi})^\top (X^\top D_{f_{n,\mu,\pi}} X)(w - w_{n,\mu,\pi}).
\end{aligned}
$$

Since $X^\top D_{f_{n,\mu,\pi}} X$ is symmetric and p.d., eigenvalues are continuous in the elements of the matrix, Λ_M and Λ_Π are compact, by the extreme value theorem, there exists a constant $C_1 > 0$ (the infimum over the smallest eigenvalues of all $X^\top D_{f_{n,\mu,\pi}} X$), independent of μ and π, such that for all y,

$$
y^\top X^\top D_{f_{n,\mu,\pi}} X y \geq C_1 \|y\|^2.
$$

Consequently,

$$
(w - w_{n,\mu,\pi})^\top g_{\mu,\pi}(w) \leq (\sqrt{\gamma}-1) C_1 \|w - w_{n,\mu,\pi}\|^2. \tag{26}
$$

Moreover, let x_i be the i-th column X, we have

$$
\begin{aligned}
\|g_{\mu,\pi}(w)\|^2 &= \sum_{i=1}^K \left(x_i^\top D_{f_{n,\mu,\pi}} (\Pi_{f_{n,\mu,\pi}} \mathcal{T}_\pi X w - Xw) \right)^2\\
&\leq \sum_{i=1}^K \|x_i\|^2_{f_{n,\mu,\pi}} \left\| \Pi_{f_{n,\mu,\pi}} \mathcal{T}_\pi X w - Xw \right\|^2_{f_{n,\mu,\pi}} \quad \text{(Cauchy-Schwarz inequality)}\\
&\leq \sum_{i=1}^K \|x_i\|^2_{f_{n,\mu,\pi}} \left(\left\| \Pi_{f_{n,\mu,\pi}} \mathcal{T}_\pi X w - Xw_{n,\mu,\pi} \right\|_{f_{n,\mu,\pi}} + \|Xw_{n,\mu,\pi} - Xw\|_{f_{n,\mu,\pi}} \right)^2\\
&\leq (\sqrt{\gamma}+1)^2 \sum_{i=1}^K \|x_i\|^2_{f_{n,\mu,\pi}} \|Xw_{n,\mu,\pi} - Xw\|^2_{f_{n,\mu,\pi}} \quad (\sqrt{\gamma}\text{-contraction})\\
&\leq (\sqrt{\gamma}+1)^2 \left(\sum_{i=1}^K \|x_i\|^2_{f_{n,\mu,\pi}} \right) \left\| X^\top D_{f_{n,\mu,\pi}} X \right\| \|w - w_{n,\mu,\pi}\|^2.
\end{aligned}
$$

By the extreme value theorem,

$$
\sup_{\mu \in \Lambda_\mu, \pi \in \Lambda_\pi} \left(\sum_{i=1}^K \|x_i\|^2_{f_{n,\mu,\pi}} \right) \left\| X^\top D_{f_{n,\mu,\pi}} X \right\| < \infty.
$$

Consequently, there exists a constant $C_2 > 0$, independent of μ and π, such that

$$
\|g_{\mu,\pi}(w)\|^2 \leq C_2 \|w - w_{n,\mu,\pi}\|^2. \tag{27}
$$

Combining (26) and (27) yields

$$
\begin{aligned}
\left\| z^\eta_{\mu,\pi}(w) - w_{n,\mu,\pi} \right\|^2 &= \|w + \eta g_{\mu,\pi}(w) - w_{n,\mu,\pi}\|^2\\
&= \|w - w_{n,\mu,\pi}\|^2 + 2\eta (w - w_{n,\mu,\pi})^\top g_{\mu,\pi}(w) + \eta^2 \|g_{\mu,\pi}(w)\|^2\\
&\leq (1 - 2\eta(1 - \sqrt{\gamma})\, C_1 + \eta^2 C_2) \|w - w_{n,\mu,\pi}\|^2
\end{aligned}
$$

Then for all $\eta < \eta_0 \doteq 2C_1(1-\sqrt{\gamma})/C_2$, we have $\beta_\eta \doteq \sqrt{1-2\eta(1-\sqrt{\gamma})C_1+\eta^2 C_2} < 1$. ■

We are now ready to study z^η with the previous lemma, analogously to Theorem 5.2 of De Farias and Van Roy (2000). Note $\mathcal{W} \doteq \{w_{n,\mu,\pi} \mid \mu \in \Lambda_M, \pi \in \Lambda_\Pi\}$ is a compact set by the continuity of $w_{n,\mu,\pi}$ in μ and π. Let $C_3 \doteq \sup_{w \in W} \|w\|$ and take some η in $(0, \eta_0)$, we have for any w

$$\|z^\eta(w)\| \leq \|z^\eta(w) - w_{n,\mu_w,\pi_w}\| + \|w_{n,\mu_w,\pi_w}\|$$

(w_{n,μ_w,π_w} denotes the fixed point of $\Pi_{f_{n,\mu,\pi}} \mathcal{T}_\pi$ with μ being μ_w and π being π_w.)

$$\begin{aligned}
&= \left\| z^\eta_{\mu_w,\pi_w}(w) - w_{n,\mu_w,\pi_w} \right\| + \|w_{n,\mu_w,\pi_w}\| \\
&\leq \beta_\eta \|w - w_{n,\mu_w,\pi_w}\| + C_3 \\
&\leq \beta_\eta \|w\| + (1+\beta_\eta) C_3.
\end{aligned}$$

Since $\beta_\eta < 1$, we define

$$\mathcal{W}_2 \doteq \left\{ w \in \mathbb{R}^K \mid \|w\| < \frac{1+\beta_\eta}{1-\beta_\eta} C_3 \right\}.$$

It is easy to verify that

$$w \in \mathcal{W}_2 \implies z^\eta(w) \in \mathcal{W}_2.$$

The Brouwer fixed point theorem then asserts that $z^\eta(w)$ adopts at least one fixed point in $\mathcal{W}_2$, which completes the proof. ■

B.7 Proof of Lemma 9

Proof Recall

$$\begin{aligned}
A_w &= X^\top D_{f_{n,\mu_w,\pi_w}} (\gamma P_{\pi_w} - I) X, \\
f_{n,\mu_w,\pi_w} &= \sum_{j=0}^n \gamma^j (P_{\pi_w}^\top)^j D_{\mu_w} i.
\end{aligned}$$

According to Lemma 9 of Zhang et al. (2021b), the invariant distribution d_μ is Lipschitz continuous w.r.t. μ in Λ_M under Assumption 5.2. Consequently, Assumption 6.1 implies that D_{μ_w} is Lipschitz continuous in w. It is then easy to see f_{n,μ_w,π_w} is Lipschitz continuous in w, using the fact that the product of two bounded Lipschitz continuous functions is still Lipschitz continuous. The Lipschitz continuity of A_w then follows easily, so does that of b_w. ■

B.8 Proof of Theorem 10

Proof Let $o_t \doteq (s_{t-n}, a_{t-n}, \dots, s_t, a_t, s_{t+1})$. Define

$$\delta_w(s, a, s') \doteq r(s, a) + \gamma \sum_{a'} \pi_w(a'|s') x(s', a')^\top w - x(s, a)^\top w,$$

$$\bar{H}(w, o_t) \doteq \left(\sum_{j=0}^{n} \gamma^j \left(\prod_{k=t-j+1}^{t} \frac{\pi_w(a_k|s_k)}{\mu_w(a_k|s_k)} \right) i(s_{t-j}, a_{t-j}) \right) \delta_w(s_t, a_t, s_{t+1}) x(s_t, a_t).$$

Note here o_t is just a placeholder for defining the function g. Let $O_t \doteq (S_{t-n}, A_{t-n}, \dots, S_t, A_t, S_{t+1})$ be a sequence of random variables generated by Algorithm 3. Then the update of w in Algorithm 3 can be expressed as

$$w_{t+1} = w_t + \alpha_t \bar{H}(w_t, O_t).$$

We now prove Theorem 10 by verifying Assumptions A.12 - A.15 thus invoking Corollary 15. Assumption A.12 is identical to Assumption 4.4.

Assumption A.13 is verified by the sampling procedure $A_{t+1} \sim \mu_{w_t}(\cdot|S_{t+1})$ in Algorithm 3 and Assumption 5.2. Similar to the proof of Theorem 4, it is easy to compute that the $h(w)$ of Assumption A.13 in our setting is

$$h(w) = A_w w + b_w.$$

For Assumption A.14, the Lipschitz continuity of the transition kernel is fulfilled by Assumption 6.1. By Assumption 5.2, there exists a constant $C_0 > 0$ such that $\mu_w(a|s) \geq C_0 > 0$ holds for any w, a, s. Then it is easy to see $\bar{H}(w, o_t)$ is Lipschitz continuous on any compact set $Q \subset \mathbb{R}^K$.

We now verify Assumption A.15. For any $w_* \in \mathcal{W}_*$, let

$$U(w) \doteq \frac{1}{2} \|w - w_*\|^2.$$

Then Assumption A.15 (i) - (iii) trivially holds. To verify Assumption A.15 (iv), let $\tilde{w} \doteq w - w_*$. We have

$$\begin{aligned}
&\left\langle \frac{\mathrm{d}U(w)}{\mathrm{d}w}, h(w) \right\rangle \\
=&\langle w - w_*, h(w) - h(w_*) \rangle \quad \text{(Using } h(w_*) = 0) \\
=&\langle w - w_*, A_w w + b_w - A_w w_* + A_w w_* - A_{w_*} w_* - b_{w_*} \rangle \\
=&\tilde{w}^\top A_w \tilde{w} + \tilde{w}^\top (A_w - A_{w_*}) w_* + \tilde{w}^\top (b_w - b_{w_*}) \\
\leq&\tilde{w}^\top A_w \tilde{w} + \|\tilde{w}\|^2 (C_1 L_\mu + C_2 L_\pi) R + \|\tilde{w}\|^2 (C_3 L_\mu + C_4 L_\pi) \\
=&\frac{1}{2} \tilde{w}(A_w + A_w^\top) \tilde{w} + \|\tilde{w}\|^2 (C_1 L_\mu + C_2 L_\pi) R + \|\tilde{w}\|^2 (C_3 L_\mu + C_4 L_\pi) \\
=&-\tilde{w} \left(M(w) - \left((C_1 L_\mu + C_2 L_\pi) R + (C_3 L_\mu + C_4 L_\pi) \right) I \right) \tilde{w} \\
\leq&-\lambda'_{min} \|w - w_*\|^2,
\end{aligned}$$

where the last inequality results from the positive definiteness of the matrix

$$M(w) - ((C_1 L_\mu + C_2 L_\pi) R + (C_3 L_\mu + C_4 L_\pi))\, I$$

under Assumption 6.2. Assumption A.15 (iv) then follows immediately.

With Assumptions A.12 - A.15 fulfilled, (18) follows immediately from Corollary 15. If there is a $w'_* \in \mathcal{W}_*$ and $w'_* \neq w_*$, repeating the previous procedure yields

$$\Pr\Big(\lim_{t\to\infty} w_t = w'_* \mid w_0 = w\Big) \geq 1 - C_{\mathcal{W}} \sum_{t=0}^{\infty} \alpha_t^2.$$

Using small enough $\{\alpha_t\}$ such that

$$1 - C_{\mathcal{W}} \sum_{t=0}^{\infty} \alpha_t^2 > 0.5$$

yields

$$\Pr\Big(\lim_{t\to\infty} w_t = w'_* \mid w_0 = w\Big) + \Pr\Big(\lim_{t\to\infty} w_t = w_* \mid w_0 = w\Big) > 1,$$

which is a contraction. Consequently, under the conditions of this theorem, $\mathcal{W}_*$ contains only one element. ■

B.9 Proof of Theorem 12

Proof Readers familiar with Zou et al. (2019) should find this proof straightforward. We mainly follow the framework of Zou et al. (2019) except for some additional error terms introduced by the truncated followon traces. We include the proof here mainly for completeness. We, however, remark that it is the use of the truncated followon trace and Lemma 8 that make this straightforwardness possible in our off-policy setting.

Let $o_t \doteq (s_{t-n}, a_{t-n}, \dots, s_t, a_t, s_{t+1})$. For a sequence of weight vectors $(z_{t-n}, \dots, z_t)$ in $\mathbb{R}^K$, define

$$\delta_z(s, a, s') \doteq r(s, a) + \gamma \sum_{a'} \pi_z(a'|s') x(s', a')^\top z - x(s, a)^\top z,$$

$$g(z_{t-n}, \dots, z_t, o_t) \doteq \left(\sum_{j=0}^{n} \gamma^j \left(\prod_{k=t-j+1}^{t} \frac{\pi_{z_k}(a_k|s_k)}{\mu_{z_k}(a_k|s_k)} \right) i(s_{t-j}, a_{t-j}) \right) \delta_{z_t}(s_t, a_t, s_{t+1}) x(s_t, a_t).$$

Note here both o_t and $z_{t-n}, \dots, z_t$ are just placeholders for defining the function g, and we adopt the convention that $\prod_{k=i}^{j}(\cdot) = 1$ if $j < i$. Let $O_t \doteq (S_{t-n}, A_{t-n}, \dots, S_t, A_t, S_{t+1})$ be a sequence of random variables generated by Algorithm 4. Then the update to w in Algorithm 4 can be expressed as

$$w_{t+1} = \Pi_R \left(w_t + \alpha_t g(w_{t-n}, \dots, w_t, O_t) \right).$$

For the ease of presentation, we define

$$g(z, o_t) \doteq g(z, z, \ldots, z, o_t),$$
$$\bar{g}(z) \doteq \mathbb{E}_{o_t \sim \mu_z}[g(z, o_t)]$$

as shorthand. By $o_t \sim \mu_z$, we mean

$$s_{t-n} \sim \bar{d}_{\mu_z}, a_{t-n} \sim \mu_z(\cdot|s_{t-n}), s_{t-n+1} \sim p(\cdot|s_{t-n}, a_{t-n}), \ldots, a_t \sim \mu_z(\cdot|s_t), s_{t+1} \sim p(\cdot|s_t, a_t).$$

It can be easily computed that

$$\bar{g}(z) = X^\top D_{f_{n,\mu_z,\pi_z}}(\gamma P_{\pi_z} - I)Xz + X^\top D_{f_{n,\mu_z,\pi_z}} r.$$

Consider a w_* in $\mathcal{W}_*$, we have

$$\bar{g}(w_*) \doteq A_{w_*} w_* + b_{w_*} = 0.$$

For any $\tau > 0$, we have

$$\begin{aligned}
&\|w_{t+1} - w_*\|^2 \\
\leq &\|w_t + \alpha_t g(w_{t-n}, \ldots, w_t, O_t) - w_*\|^2 \quad (\Pi_R \text{ is nonexpansive}) \\
= &\|w_t - w_*\|^2 + \alpha_t^2 \|g(w_{t-n}, \ldots, w_t, O_t)\|^2 + 2\alpha_t \langle w_t - w_*, g(w_{t-n}, \ldots, w_t, O_t) \rangle \\
= &\|w_t - w_*^2\| \\
&+ \alpha_t^2 \|g(w_{t-n}, \ldots, w_t, O_t)\|^2 && (28) \\
&+ 2\alpha_t \underbrace{\left(\langle w_t - w_*, g(w_{t-n}, \ldots, w_t, O_t) \rangle - \langle w_{t-n-\tau} - w_*, \bar{g}(w_{t-n-\tau}) \rangle\right)}_{err_t} && (29) \\
&+ 2\alpha_t \langle w_{t-n-\tau} - w_*, \bar{g}(w_{t-n-\tau}) - \bar{g}(w_*) \rangle, && (30)
\end{aligned}$$

where we adopt the convention that $w_{t-n-\tau} \equiv w_0$ if $t - n - \tau < 0$. Using Lemmas 17 and 18 to bound (28) and (30) yields

$$\mathbb{E}\left[\|w_{t+1} - w_*\|^2\right] \leq \mathbb{E}\left[\|w_t - w_*\|^2\right] + \alpha_t^2 U_g^2 + 2\alpha_t \mathbb{E}\left[err_t\right] - 2\alpha_t \alpha_\lambda \mathbb{E}\left[\|w_t - w_*\|^2\right].$$

Dividing by $2\alpha_t$ in both sides yields

$$\frac{1}{2\alpha_t}\mathbb{E}\left[\|w_{t+1} - w_*\|^2\right] \leq \frac{1}{2\alpha_t}\mathbb{E}\left[\|w_t - w_*\|^2\right] + \frac{1}{2}\alpha_t U_g^2 + \mathbb{E}\left[err_t\right] - \alpha_\lambda \mathbb{E}\left[\|w_t - w_*\|^2\right] \quad (31)$$

Using the definition of α_t in (19) yields

$$\alpha_\lambda (t+1) \mathbb{E}\left[\|w_{t+1} - w_*\|^2\right] \leq \alpha_\lambda t \mathbb{E}\left[\|w_t - w_*\|^2\right] + \frac{1}{2}\alpha_t U_g^2 + \mathbb{E}\left[err_t\right].$$

For some fixed T, let $\tau_0 \doteq \min\{\tau : C_0 \kappa^\tau < \alpha_T\}$. Using the definition of α_t in (19), it can be easily computed that

$$\tau_0 = \lceil \frac{\ln(2\alpha_\lambda (T+1) C_0)}{\ln \kappa^{-1}} \rceil = \mathcal{O}(\ln T)$$

where $\lceil \cdot \rceil$ is the ceiling function. Here we assume T is large enough such that

$$\tau_0 < T - n.$$

Telescoping (31) for $t = 0, \dots, T$ with $\tau = \tau_0$ yields

$$\begin{aligned}
&\alpha_\lambda T \mathbb{E}\left[\|w_T - w_*\|^2\right] \\
\leq &\sum_{t=0}^{T-1} \frac{1}{2} \alpha_t U_g^2 + \sum_{t=0}^{T-1} \mathbb{E}[err_t] \\
= &\sum_{t=0}^{T-1} \frac{1}{2} \frac{1}{2\alpha_\lambda (t+1)} U_g^2 + \sum_{t=0}^{\tau_0 + n} \mathbb{E}[err_t] + \sum_{t=n+\tau_0+1}^{T-1} \mathbb{E}[err_t] \\
\leq &\frac{U_g^2}{4\alpha_\lambda} \ln T + (n + \tau_0 + 1) 4 R U_g + \sum_{t=n+\tau_0+1}^{T-1} \mathbb{E}[err_t],
\end{aligned} \tag{32}$$

where the last inequality results from

$$\sum_{t=0}^{T-1} \frac{1}{t+1} \leq \ln T$$

and the first part of Lemma 19. Using the second part of Lemma 19 with $\tau = \tau_0$ to bound the last term of (32) yields

$$
\begin{aligned}
&\sum_{t=n+\tau_0+1}^{T-1} \mathbb{E}[err_t] \\
\leq &\sum_{t=n+\tau_0+1}^{T-1} \left(C_5 \sum_{j=t-n-\tau_0}^{t-1} \alpha_j + C_6 \sum_{k=t-n-\tau_0}^{t-2} \sum_{j=t-n-\tau_0}^{k} \alpha_j + C_7 C_0 \kappa^{\tau_0 - 1} \right) \\
= &\frac{1}{2\alpha_\lambda} \sum_{t=n+\tau_0+1}^{T-1} \left(C_5 \sum_{j=t-n-\tau_0}^{t-1} \frac{1}{j+1} + C_6 \sum_{k=t-n-\tau_0}^{t-2} \sum_{j=t-n-\tau_0}^{k} \frac{1}{j+1} + C_7 C_0 \kappa^{\tau_0 - 1} \right) \\
\leq &\frac{1}{2\alpha_\lambda} \sum_{t=n+\tau_0+1}^{T-1} \left(C_5 \ln \frac{t}{t-n-\tau_0} + C_6 \sum_{k=t-n-\tau_0}^{t-2} \ln \frac{k+1}{t-n-\tau_0} + C_7 C_0 \kappa^{\tau_0 - 1} \right) \\
\leq &\frac{1}{2\alpha_\lambda} \sum_{t=n+\tau_0+1}^{T-1} \left((C_5 + C_6(n+\tau_0)) \ln \frac{t}{t-n-\tau_0} + C_7 C_0 \kappa^{\tau_0 - 1} \right) \\
\leq &\frac{1}{2\alpha_\lambda} \sum_{t=n+\tau_0+1}^{T-1} \left((C_5 + C_6(n+\tau_0)) \ln \frac{t}{t-n-\tau_0} + \frac{C_7}{\kappa} \alpha_T \right) \\
= &\frac{1}{2\alpha_\lambda} \sum_{t=n+\tau_0+1}^{T-1} \left((C_5 + C_6(n+\tau_0)) \ln \frac{t}{t-n-\tau_0} + \frac{C_7}{2\alpha_\lambda \kappa} \frac{1}{T+1} \right) \\
\leq &\frac{1}{2\alpha_\lambda} (C_5 + C_6(n+\tau_0)) \ln \prod_{t=n+\tau_0+1}^{T-1} \frac{t}{t-n-\tau_0} + \frac{C_7}{2\alpha_\lambda \kappa} \\
\leq &\frac{1}{2\alpha_\lambda} (C_5 + C_6(n+\tau_0)) \ln \frac{(T-1)\cdots(T-1-n-\tau_0)}{(n+\tau_0)\cdots 1} + \frac{C_7}{2\alpha_\lambda \kappa} \\
\leq &\frac{1}{2\alpha_\lambda} (C_5 + C_6(n+\tau_0))(n+\tau_0) \ln T + \frac{C_7}{2\alpha_\lambda \kappa}.
\end{aligned}
$$

Plugging the above inequality back into (32) yields

$$
\begin{aligned}
&\mathbb{E}\left[\|w_T - w_*\|^2 \right] \\
\leq &\frac{U_g^2}{4\alpha_\lambda^2} \frac{\ln T}{T} + \frac{4RU_g}{\alpha_\lambda} \frac{(n+\tau_0+1)}{T} + \frac{1}{2\alpha_\lambda^2} (C_5 + C_6(n+\tau_0))(n+\tau_0) \frac{\ln T}{T} + \frac{C_7}{2\alpha_\lambda^2 \kappa T} \\
= &\mathcal{O}\left(\frac{\ln^3 T}{T} \right).
\end{aligned}
$$

If there is also a $w'_* \in \mathcal{W}_*$, repeating the above procedure yields

$$
\mathbb{E}\left[\|w_T - w'_*\|^2 \right] = \mathcal{O}\left(\frac{\ln^3 T}{T} \right).
$$

Consequently,

$$\left\|w_* - w'_*\right\| = \mathbb{E}\left[\left\|w_* - w'_*\right\|\right] \leq \mathbb{E}\left[\left\|w_T - w'_*\right\|\right] + \mathbb{E}\left[\left\|w_T - w_*\right\|\right] = \mathcal{O}\left(\sqrt{\frac{\ln^3 T}{T}}\right).$$

Letting T approaches infinity yields $w_* = w'_*$, i.e., $\mathcal{W}_*$ contains only one element under the condition of this theorem, which completes the proof. ■

Lemma 17 *(Bound of* (28)*) There exists a constant U_g such that*

$$\|g(w_{t-n}, \dots, w_t, O_t)\|^2 \leq U_g^2$$

Proof Due to the projection Π_R, we have $\|w_t\| \leq R$ holds for all t. By the definition of g, it is easy to compute that

$$\|g(w_{t-n}, \dots, w_t, O_t)\| \leq \underbrace{(n+1)\rho_{max}^n i_{max}(r_{max} + (1+\gamma)Rx_{max})x_{max}}_{U_g},$$

where $i_{max} \doteq \max_{s,a} i(s,a), r_{max} \doteq \max_{s,a} |r(s,a)|, x_{max} \doteq \max_{s,a} \|x(s,a)\|$,

$$\rho_{\max} \doteq \sup_{\mu \in \Lambda_M, \pi \in \Lambda_\pi, s, a} \frac{\pi(s,a)}{\mu(s,a)}.$$

Assumption 5.2 and the extreme value theorem ensures that $\rho_{max} < \infty$. ■

Lemma 18 *(Bound of* (30)*)*

$$\langle w_{t-n-\tau} - w_*, \bar{g}(w_{t-n-\tau}) - \bar{g}(w_*)\rangle \leq -\alpha_\lambda \|w_t - w_*\|^2$$

Proof Let $\tilde{w} \doteq w_{t-n-\tau} - w_*$, we have

$$\begin{aligned}
&\langle w_{t-n-\tau} - w_*, \bar{g}(w_{t-n-\tau}) - \bar{g}(w_*)\rangle \\
=&\left\langle \tilde{w}, A_{w_{t-n-\tau}} w_{t-n-\tau} + b_{w_{t-n-\tau}} - A_{w_*} w_* - b_{w_*}\right\rangle \\
=&\left\langle \tilde{w}, A_{w_{t-n-\tau}} w_{t-n-\tau} - A_{w_*} w_{t-n-\tau} + A_{w_*} w_{t-n-\tau} - A_{w_*} w_* + b_{w_{t-n-\tau}} - b_{w_*}\right\rangle \\
=&\tilde{w}^\top A_{w_*} \tilde{w} + \tilde{w}(A_{w_{t-n-\tau}} - A_{w_*}) w_{t-n-\tau} + \tilde{w}^\top (b_{w_{t-n-\tau}} - b_{w_*}) \\
\leq&\tilde{w}^\top A_{w_*} \tilde{w} + \|\tilde{w}\|^2 (C_1 L_\mu + C_2 L_\pi) R + \|\tilde{w}\|^2 (C_3 L_\mu + C_4 L_\pi) \\
\leq& - \tilde{w}^\top \left(M(w_*) - \left((C_1 L_\mu + C_2 L_\pi) R + (C_3 L_\mu + C_4 L_\pi)\right) I\right) \tilde{w} \\
\leq& - \lambda''_{min} \|\tilde{w}\|^2 \\
\leq& - \alpha_\lambda \|\tilde{w}\|^2.
\end{aligned}$$

■

Lemma 19 *(Bound of* (29)*) Let*

$$err_t \doteq \langle w_t - w_*, g(w_{t-n}, \ldots, w_t, O_t)\rangle - \langle w_{t-n-\tau} - w_*, \bar{g}(w_{t-n-\tau})\rangle.$$

Then for any t and τ,

$$\|err_t\| \leq 4RU_g.$$

If $t-n-\tau > 0$, there exist positive constants C_5, C_6, independent of t, such that

$$\mathbb{E}\left[err_t\right] \leq C_5 \sum_{j=t-n-\tau}^{t-1} \alpha_j + C_6 \sum_{k=t-n-\tau}^{t-2} \sum_{j=t-n-\tau}^{k} \alpha_j + C_7 C_0 \kappa^{\tau-1}.$$

Proof If $t-n-\tau < 0$,

$$\|err_t\| \leq \|w_t - w_*\| \|g(w_{t-n}, \ldots, w_t, O_t)\| + \|w_{t-n-\tau} - w_*\| \|\bar{g}(w_{t-n-\tau})\| \leq 4RU_g.$$

When $t-n-\tau > 0$, similar to Zou et al. (2019), we define an auxiliary Markov chain $\left\{\tilde{S}_t, \tilde{A}_t\right\}$ as

$$\left\{\tilde{S}_t, \tilde{A}_t\right\} : \cdots \underbrace{\rightarrow}_{\mu_{w_{t-n-\tau}}} S_{t-n-\tau+2} \underbrace{\rightarrow}_{\mu_{w_{t-n-\tau}}} \tilde{S}_{t-n-\tau+3} \underbrace{\rightarrow}_{\mu_{w_{t-n-\tau}}} \tilde{S}_{t-n-\tau+4} \rightarrow \cdots,$$
$$\left(\{S_t, A_t\} : \cdots \underbrace{\rightarrow}_{\mu_{w_{t-n-\tau}}} S_{t-n-\tau+2} \underbrace{\rightarrow}_{\mu_{w_{t-n-\tau+1}}} S_{t-n-\tau+3} \underbrace{\rightarrow}_{\mu_{w_{t-n-\tau+2}}} S_{t-n-\tau+4} \rightarrow \cdots\right)$$

i.e., the new chain is the same as the chain generated by Algorithm 4 (i.e., the chain (S_t, A_t)) before $S_{t-n-\tau+2}$, after which the new chain is generated by following a fixed behavior policy $\mu_{w_{t-n-\tau}}$ instead of the changing behavior policies $\mu_{w_{t-n-\tau+1}}, \mu_{w_{t-n-\tau+2}}, \cdots$ as the original chain. Let

$$\tilde{O}_t \doteq (\tilde{S}_{t-n}, \tilde{A}_{t-n}, \ldots, \tilde{S}_t, \tilde{A}_t, \tilde{S}_{t+1}),$$

we have

$$\begin{aligned}
err_t =& \langle w_t - w_*, g(w_{t-n}, \ldots, w_t, O_t)\rangle - \langle w_{t-n-\tau} - w_*, \bar{g}(w_{t-n-\tau}\rangle \\
=& \langle w_t - w_*, g(w_{t-n}, \ldots, w_t, O_t) - g(w_t, O_t)\rangle && (33)\\
&+ \langle w_t - w_*, g(w_t, O_t)\rangle - \langle w_{t-n-\tau} - w_*, g(w_{t-n-\tau}, O_t)\rangle && (34)\\
&+ \left\langle w_{t-n-\tau} - w_*, g(w_{t-n-\tau}, O_t) - g(w_{t-n-\tau}, \tilde{O}_t)\right\rangle && (35)\\
&+ \left\langle w_{t-n-\tau} - w_*, g(w_{t-n-\tau}, \tilde{O}_t) - \bar{g}(w_{t-n-\tau})\right\rangle. && (36)
\end{aligned}$$

Using Lemmas 20, 21, 22, and 24 to bound (33), (34), (35), and (36) yields

$$\begin{aligned}\mathbb{E}\left[err_t\right] \leq & 2nRL_g \sum_{j=t-n}^{t-1} \alpha_j + (2RL_g + U_g)U_g \sum_{j=t-n-\tau}^{t-1} \alpha_j \\ & + 2R|\mathcal{A}|L_\mu U_g^2 \sum_{k=t-n-\tau}^{t-2} \sum_{j=t-n-\tau}^{k} \alpha_j + 2RU_g C_0 \kappa^{\tau-1} \\ \leq & \underbrace{(2nRL_g + (2RL_g + U_g)U_g)}_{C_5} \sum_{j=t-n-\tau}^{t-1} \alpha_j \\ & + \underbrace{2R|\mathcal{A}|L_\mu U_g^2}_{C_6} \sum_{k=t-n-\tau}^{t-2} \sum_{j=t-n-\tau}^{k} \alpha_j + \underbrace{2RU_g}_{C_7} C_0 \kappa^{\tau-1}\end{aligned}$$

■

Lemma 20 *(Bound of* (33)*) There exists a positive constant L_g such that*

$$\langle w_t - w_*, g(w_{t-n}, \dots, w_t, O_t) - g(w_t, O_t)\rangle \leq 2nRL_g \sum_{j=t-n}^{t-1} \alpha_j.$$

Proof First, for any $t' > t$, we have

$$\|w_{t'} - w_t\| \leq U_g \sum_{j=t}^{t'-1} \alpha_j$$

by using triangle inequalities with $w_{t+1}, w_{t+2}, \dots, w_{t'-1}$ and Lemma 17. It is then easy to show that $g(w_{t-n}, \dots, w_t, O_t)$ is Lipschitz in its first argument:

$$\begin{aligned}&\|g(w_{t-n}, w_{t-n+1}, w_{t-n+2}, \dots, w_t, O_t) - g(w_t, w_{t-n+1}, w_{t-n+2}, \dots, w_t, O_t)\| \\ \leq & \underbrace{\frac{(n+1)(L_\mu + L_\pi)\rho_{max}^n (r_{max} + 2x_{max}R)x_{max}}{\mu_{min}^2}}_{L_g} \|w_{t-n} - w_t\| \leq L_g U_g \sum_{j=t-n}^{t-1} \alpha_j,\end{aligned}$$

where $\mu_{min} \doteq \inf_{s,a} \mu_w(a|s)$. By the extreme value theorem, Assumption 5.2 implies that $\mu_{min} > 0$. Similarly, g is also Lipschitz continuous in its second argument:

$$\|g(w_t, w_{t-n+1}, w_{t-n+2} \dots, w_t, O_t) - g(w_t, w_t, w_{t-n+2}, \dots, w_t, O_t)\| \leq L_g U_g \sum_{j=t-n+1}^{t-1} \alpha_j.$$

Repeating this procedure for the third to n-th argument $(w_{t-n+2}, \dots, w_{t-1})$ and putting them together with the triangle inequality yields

$$\|g(w_{t-n}, \dots, w_t, O_t) - g(w_t, O_t)\| \leq nL_g U_g \sum_{j=t-n}^{t-1} \alpha_j.$$

Consequently,

$$\langle w_t - w_*, g(w_{t-n}, \dots, w_t, O_t) - g(w_t, O_t)\rangle \leq 2nRL_gU_g \sum_{j=t-n}^{t-1} \alpha_j.$$

■

Lemma 21 *(Bound of* (34)*)*

$$\langle w_t - w_*, g(w_t, O_t)\rangle - \langle w_{t-n-\tau} - w_*, g(w_{t-n-\tau}, O_t)\rangle \leq (2RL_g + U_g)U_g \sum_{j=t-n-\tau}^{t-1} \alpha_j$$

Proof

$$\begin{aligned}
&\langle w_t - w_*, g(w_t, O_t)\rangle - \langle w_{t-n-\tau} - w_*, g(w_{t-n-\tau}, O_t)\rangle \\
=&\langle w_t - w_*, g(w_t, O_t) - g(w_{t-n-\tau}, O_t)\rangle - \langle w_{t-n-\tau} - w_*, g(w_{t-n-\tau}, O_t)\rangle \\
&+ \langle w_t - w_*, g(w_{t-n-\tau}, O_t)\rangle \\
=&\langle w_t - w_*, g(w_t, O_t) - g(w_{t-n-\tau}, O_t)\rangle + \langle w_t - w_{t-n-\tau}, g(w_{t-n-\tau}, O_t)\rangle \\
\leq&2RL_g\|w_t - w_{t-n-\tau}\| + U_g\|w_t - w_{t-n-\tau}\| \\
\leq&(2RL_g + U_g)U_g \sum_{j=t-n-\tau}^{t-1} \alpha_j
\end{aligned}$$

■

Lemma 22 *(Bound of* (35)*)*

$$\mathbb{E}\left[\left\langle w_{t-n-\tau} - w_*, g(w_{t-n-\tau}, O_t) - g(w_{t-n-\tau}, \tilde{O}_t)\right\rangle\right] \leq 2R|\mathcal{A}|L_\mu U_g^2 \sum_{k=t-n-\tau}^{t-2} \sum_{j=t-n-\tau}^{k} \alpha_j$$

Proof Let $\Sigma_{t-n-\tau} \doteq (w_0, w_1, \dots, w_{t-n-\tau}, S_0, A_0, \dots, S_{t-n-\tau+1}, A_{t-n-\tau+1})$. We have

$$\begin{aligned}
&\mathbb{E}\left[\left\langle w_{t-n-\tau} - w_*, g(w_{t-n-\tau}, O_t) - g(w_{t-n-\tau}, \tilde{O}_t)\right\rangle\right] \\
=&\mathbb{E}\left[\mathbb{E}\left[\left\langle w_{t-n-\tau} - w_*, g(w_{t-n-\tau}, O_t) - g(w_{t-n-\tau}, \tilde{O}_t)\right\rangle \mid \Sigma_{t-n-\tau}\right]\right] \\
&\text{(Law of total expectation)} \\
=&\mathbb{E}\left[\left\langle w_{t-n-\tau} - w_*, \mathbb{E}\left[g(w_{t-n-\tau}, O_t) - g(w_{t-n-\tau}, \tilde{O}_t) \mid \Sigma_{t-n-\tau}\right]\right\rangle\right]
\end{aligned}$$

(Conditional independence)

$$
\begin{aligned}
\leq & \mathbb{E}\left[\|w_{t-n-\tau} - w_*\| \left\| \mathbb{E}\left[g(w_{t-n-\tau}, O_t) - g(w_{t-n-\tau}, \tilde{O}_t) \mid \Sigma_{t-n-\tau}\right] \right\| \right] \\
\leq & 2R\mathbb{E}\left[\left\| \mathbb{E}\left[g(w_{t-n-\tau}, O_t) - g(w_{t-n-\tau}, \tilde{O}_t) \mid \Sigma_{t-n-\tau}\right] \right\| \right] \\
\leq & 2R|\mathcal{A}|L_\mu U_g^2 \sum_{k=t-n-\tau}^{t-2} \sum_{j=t-n-\tau}^{k} \alpha_j,
\end{aligned}
$$

where the last inequality comes from Lemma 23. ■

Lemma 23

$$
\left\| \mathbb{E}\left[g(w_{t-n-\tau}, O_t) - g(w_{t-n-\tau}, \tilde{O}_t) \mid \Sigma_{t-n-\tau}\right] \right\| \leq |\mathcal{A}|L_\mu U_g^2 \sum_{k=t-n-\tau}^{t-2} \sum_{j=t-n-\tau}^{k} \alpha_j
$$

Proof In the proof of this lemma, all expectations ($\mathbb{E}$) and probabilities (Pr) are conditioned on $\Sigma_{t-n-\tau}$. We suppress this condition in the presentation for improving readability. Given $t, n, \tau, \Sigma_{t-n-\tau}$, for any time step j such that $t - n - \tau + 1 \leq j \leq t$, we use $\mathcal{W}_j \subset \mathbb{R}^K$ to denote the set of all possible values of w_j. It is easy to see that $\mathcal{W}_j$ is always a *finite* set depending on $t, n, \tau, \Sigma_{t-n-\tau}$. This allows us to use summation instead of integral to further improve readability. We have

$$
\begin{aligned}
& \left\| \mathbb{E}\left[g(w_{t-n-\tau}, \tilde{O}_t) - g(w_{t-n-\tau}, O_t)\right] \right\| \\
= & \left\| \sum_{o_t} \left(\Pr\left(\tilde{O}_t = o_t\right) - \Pr(O_t = o_t)\right) g(w_{t-n-\tau}, o_t) \right\|
\end{aligned}
$$

(Conditional independence of O_t and $\tilde{O}_t$ given $\Sigma_{t-n-\tau}$)

$$
\leq U_g \sum_{o_t} \left| \Pr\left(\tilde{O}_t = o_t\right) - \Pr(O_t = o_t) \right|.
$$

In the rest of this proof we bound $\left| \Pr\left(\tilde{O}_t = o_t\right) - \Pr(O_t = o_t) \right|$. To start,

$$
\begin{aligned}
& \Pr(O_t = o_t) \\
= & \sum_{z_{t-1} \in \mathcal{W}_{t-1}} \Pr(w_{t-1} = z_{t-1}, A_t = a_t, S_{t+1} = s_{t+1}, S_{t-n} = s_{t-n}, \dots, S_t = s_t)
\end{aligned}
$$

(Law of total probability)

$$
= \sum_{z_{t-1}} \Pr\left(A_t = a_t, S_{t+1} = s_{t+1} \mid \substack{S_{t-n} = s_{t-n} \\ \cdots \\ S_t = s_t \\ w_{t-1} = z_{t-1}}\right) \Pr\left(w_{t-1} = z_{t-1} \mid \substack{S_{t-n} = s_{t-n} \\ \cdots \\ S_t = s_t}\right) \Pr(S_{t-n} = s_{t-n}, \dots, S_t = s_t)
$$

(Chain rule of joint distribution)

$$=\sum_{z_{t-1}} \mu_{z_{t-1}}(a_t|s_t)p(s_{t+1}|s_t,a_t) \Pr\Big(w_{t-1} = z_{t-1} \mid {}^{S_{t-n}=s_{t-n}}_{\substack{\dots \\ S_t=s_t}}\Big) \Pr(S_{t-n} = s_{t-n}, \dots, S_t = s_t).$$

Further,

$$\begin{aligned}
&\Pr\Big(\tilde{O}_t = o_t\Big) \\
=&\mu_{w_{t-n-\tau}}(a_t|s_t)p(s_{t+1}|s_t,a_t) \Pr\Big(\tilde{S}_{t-n} = s_{t-n}, \dots, \tilde{S}_t = s_t\Big) \\
&\qquad \text{(Definition of the auxiliary chain)} \\
=&\mu_{w_{t-n-\tau}}(a_t|s_t)p(s_{t+1}|s_t,a_t) \Pr\Big(\tilde{S}_{t-n} = s_{t-n}, \dots, \tilde{S}_t = s_t\Big) \sum_{z_{t-1}} \Pr\Big(w_{t-1} = z_{t-1} \mid {}^{S_{t-n}=s_{t-n}}_{\substack{\dots \\ S_t=s_t}}\Big).
\end{aligned}$$

Consequently,

$$\begin{aligned}
&\sum_{o_t} \Big|\Pr(O_t = o_t) - \Pr\Big(\tilde{O}_t = o_t\Big)\Big| \\
\leq& \sum_{s_{t-n},\dots,s_t,a_t,z_{t-1}} \Pr\Big(w_{t-1} = z_{t-1} \mid {}^{S_{t-n}=s_{t-n}}_{\substack{\dots \\ S_t=s_t}}\Big) \times \\
&\Big|\mu_{z_{t-1}}(a_t|s_t) \Pr(S_{t-n}, \dots, S_t = s_t) - \mu_{w_{t-n-\tau}}(a_t|s_t) \Pr\Big(\tilde{S}_{t-n} = s_{t-n}, \dots, \tilde{S}_t = s_t\Big)\Big| \\
\leq& \sum_{s_{t-n},\dots,s_t,a_t,z_{t-1}} \Pr\Big(w_{t-1} = z_{t-1} \mid {}^{S_{t-n}=s_{t-n}}_{\substack{\dots \\ S_t=s_t}}\Big) \times \\
&\Bigg(\Big|\mu_{z_{t-1}}(a_t|s_t) \Pr(S_{t-n}, \dots, S_t = s_t) - \mu_{w_{t-n-\tau}}(a_t|s_t) \Pr(S_{t-n} = s_{t-n}, \dots, S_t = s_t)\Big| + \\
&\Big|\mu_{w_{t-n-\tau}}(a_t|s_t) \Pr(S_{t-n}, \dots, S_t = s_t) - \mu_{w_{t-n-\tau}}(a_t|s_t) \Pr\Big(\tilde{S}_{t-n} = s_{t-n}, \dots, \tilde{S}_t = s_t\Big)\Big|\Bigg) \\
\leq& \sum_{s_{t-n},\dots,s_t,a_t,z_{t-1}} \Pr\Big(w_{t-1} = z_{t-1} \mid {}^{S_{t-n}=s_{t-n}}_{\substack{\dots \\ S_t=s_t}}\Big) \times \\
&\Bigg(\Big|\mu_{z_{t-1}}(a_t|s_t) - \mu_{w_{t-n-\tau}}(a_t|s_t)\Big| \Pr(S_{t-n} = s_{t-n}, \dots, S_t = s_t) + \\
&\mu_{w_{t-n-\tau}}(a_t|s_t)\Big|\Pr(S_{t-n}, \dots, S_t = s_t) - \Pr\Big(\tilde{S}_{t-n} = s_{t-n}, \dots, \tilde{S}_t = s_t\Big)\Big|\Bigg)
\end{aligned} \tag{37}$$

Since $z_{t-1} \in \mathcal{W}_{t-1}$, we have

$$|\mu_{z_{t-1}}(a_t|s) - \mu_{w_{t-n-\tau}}(a_t|s)| \leq L_\mu \|z_{t-1} - w_{t-n-\tau}\| \leq L_\mu U_g \sum_{j=t-n-\tau}^{t-2} \alpha_j.$$

Plugging the above inequality back to (37) yields

$$
\begin{aligned}
&\sum_{o_t} \left| \Pr(O_t = o_t) - \Pr\Big(\tilde{O}_t = o_t\Big) \right| \\
\leq &\sum_{s_{t-n}, \dots, s_t, a_t, z_{t-1}} \Pr\Big(w_{t-1} = z_{t-1} \mid {}^{S_{t-n}=s_{t-n}}_{\substack{\dots \\ S_t = s_t}}\Big) \times \\
&\Bigg(\Pr(S_{t-n} = s_{t-n}, \dots, S_t = s_t) L_\mu U_g \sum_{j=t-n-\tau}^{t-2} \alpha_j + \\
&\mu_{w_{t-n-\tau}}(a_t|s_t) \left| \Pr(S_{t-n}, \dots, S_t = s_t) - \Pr\Big(\tilde{S}_{t-n} = s_{t-n}, \dots, \tilde{S}_t = s_t\Big) \right| \Bigg) \\
= &|\mathcal{A}| L_\mu U_g \sum_{j=t-n-\tau}^{t-2} \alpha_j + \sum_{s_{t-n}, \dots, s_t} \left| \Pr(S_{t-n}, \dots, S_t = s_t) - \Pr\Big(\tilde{S}_{t-n} = s_{t-n}, \dots, \tilde{S}_t = s_t\Big) \right|.
\end{aligned}
$$

Recursively using the above inequality $n+1$ times yields

$$
\begin{aligned}
&\sum_{o_t} \left| \Pr\Big(\tilde{O}_t = o_t\Big) - \Pr(O_t = o_t) \right| \qquad (38) \\
\leq &|\mathcal{A}| L_\mu U_g \left(\sum_{j=t-n-\tau}^{t-2} \alpha_j + \dots + \sum_{j=t-n-\tau}^{t-n-2} \alpha_j \right) + \sum_{s_{t-n}} \left| \Pr(S_{t-n} = s_{t-n}) - \Pr\Big(\tilde{S}_{t-n} = s_{t-n}\Big) \right|
\end{aligned}
$$

We now bound the last term in the above equation. We have

$$
\begin{aligned}
&\Pr(S_{t-n} = s_{t-n}) \\
= &\sum_s \Pr(S_{t-n-1} = s, S_{t-n} = s_{t-n}) \\
= &\sum_s \Pr(S_{t-n-1} = s) \Pr(S_{t-n} = s_{t-n} | S_{t-n-1} = s) \\
= &\sum_{s,a} \Pr(S_{t-n-1} = s) \Pr(S_{t-n} = s_{t-n}, A_{t-n-1} = a | S_{t-n-1} = s) \\
= &\sum_{s,a} \Pr(S_{t-n-1} = s) \mathbb{E}_{w_{t-n-2}} \left[\Pr(S_{t-n} = s_{t-n}, A_{t-n-1} = a | S_{t-n-1} = s, w_{t-n-2}) \right] \\
= &\sum_{s,a} \Pr(S_{t-n-1} = s) \mathbb{E}_{w_{t-n-2}} \left[\mu_{w_{t-n-2}}(a|s) p(s_{t-n}|s,a) \right]
\end{aligned}
$$

Similarly,

$$
\Pr\Big(\tilde{S}_{t-n} = s_{t-n}\Big) = \sum_{s,a} \Pr\Big(\tilde{S}_{t-n-1} = s\Big) \mu_{w_{t-n-\tau}}(a|s) p(s_{t-n}|s,a).
$$

Consequently,

$$
\begin{aligned}
&\sum_{s_{t-n}} \left|\Pr(S_{t-n} = s_{t-n}) - \Pr\Big(\tilde{S}_{t-n} = s_{t-n}\Big)\right| \\
=&\sum_{s,a} \left|\Pr(S_{t-n-1} = s)\mathbb{E}_{w_{t-n-2}}\left[\mu_{w_{t-n-2}}(a|s)\right] - \Pr\Big(\tilde{S}_{t-n-1} = s\Big)\mu_{w_{t-n-\tau}}(a|s)\right| \\
\leq&\sum_{s,a} \left|\Pr(S_{t-n-1} = s)\mathbb{E}_{w_{t-n-2}}\left[\mu_{w_{t-n-2}}(a|s)\right] - \Pr\Big(\tilde{S}_{t-n-1} = s\Big)\mathbb{E}_{w_{t-n-2}}\left[\mu_{w_{t-n-2}}(a|s)\right]\right| + \\
&\sum_{s,a} \left|\Pr\Big(\tilde{S}_{t-n-1} = s\Big)\mathbb{E}_{w_{t-n-2}}\left[\mu_{w_{t-n-2}}(a|s)\right] - \Pr\Big(\tilde{S}_{t-n-1} = s\Big)\mu_{w_{t-n-\tau}}(a|s)\right| \\
=&\sum_{s} \left|\Pr(S_{t-n-1} = s) - \Pr\Big(\tilde{S}_{t-n-1} = s\Big)\right| + \\
&\sum_{s,a} \Pr\Big(\tilde{S}_{t-n-1} = s\Big)\left|\mathbb{E}_{w_{t-n-2}}\left[\mu_{w_{t-n-2}}(a|s)\right] - \mu_{w_{t-n-\tau}}(a|s)\right| \\
\leq&\sum_{s} \left|\Pr(S_{t-n-1} = s) - \Pr\Big(\tilde{S}_{t-n-1} = s\Big)\right| + \\
&\sum_{s,a} \Pr\Big(\tilde{S}_{t-n-1} = s\Big)\max_s\left|\mathbb{E}_{w_{t-n-2}}\left[\mu_{w_{t-n-2}}(a|s)\right] - \mu_{w_{t-n-\tau}}(a|s)\right|
\end{aligned}
$$

Since

$$
\begin{aligned}
&\left|\mathbb{E}_{w_{t-n-2}}\left[\mu_{w_{t-n-2}}(a|s)\right] - \mu_{w_{t-n-\tau}}(a|s)\right| \\
=&\left|\mathbb{E}_{w_{t-n-2}}\left[\mu_{w_{t-n-2}}(a|s) - \mu_{w_{t-n-\tau}}(a|s)\right]\right| \\
\leq&\mathbb{E}_{w_{t-n-2}}\left[\left|\mu_{w_{t-n-2}}(a|s) - \mu_{w_{t-n-\tau}}(a|s)\right|\right] \\
\leq&U_g L_\mu \sum_{j=t-n-\tau}^{t-n-3} \alpha_j,
\end{aligned}
$$

we have

$$
\begin{aligned}
&\sum_{s} \left|\Pr(S_{t-n} = s) - \Pr\Big(\tilde{S}_{t-n} = s\Big)\right| \\
\leq&\sum_{s} \left|\Pr(S_{t-n-1} = s) - \Pr\Big(\tilde{S}_{t-n-1} = s\Big)\right| + |\mathcal{A}|U_g L_\mu \sum_{j=t-n-\tau}^{t-n-3} \alpha_j.
\end{aligned}
$$

Applying the above inequality recursively yields

$$
\sum_{s} \left|\Pr(S_{t-n} = s) - \Pr\Big(\tilde{S}_{t-n} = s\Big)\right| \leq |\mathcal{A}|U_g L_\mu \left(\sum_{j=t-n-\tau}^{t-n-3} \alpha_j + \cdots + \sum_{j=t-n-\tau}^{t-n-\tau} \alpha_j\right) \tag{39}
$$

as

$$
\Pr(S_{t-n-\tau+2} = s) = \Pr\Big(\tilde{S}_{t-n-\tau+2} = s\Big)
$$

by the construction of the auxiliary chain. Plugging (39) back to (38) yields

$$\sum_{o_t}\left|\Pr\Big(\tilde{O}_t = o_t\Big) - \Pr(O_t = o_t)\right| \leq |\mathcal{A}| L_\mu U_g \sum_{k=t-n-\tau}^{t-2} \sum_{j=t-n-\tau}^{k} \alpha_j,$$

which completes the proof. ■

Lemma 24 *(Bound of* (36)*)*

$$\mathbb{E}\left[\Big\langle w_{t-n-\tau} - w_*, g(w_{t-n-\tau}, \tilde{O}_t) - \bar{g}(w_{t-n-\tau})\Big\rangle\right] \leq 2RU_g C_0 \kappa^{\tau-1}$$

Proof

$$\begin{aligned}
&\mathbb{E}\left[\Big\langle w_{t-n-\tau} - w_*, g(w_{t-n-\tau}, \tilde{O}_t) - \bar{g}(w_{t-n-\tau})\Big\rangle\right] \\
=&\mathbb{E}\left[\mathbb{E}\left[\Big\langle w_{t-n-\tau} - w_*, g(w_{t-n-\tau}, \tilde{O}_t) - \bar{g}(w_{t-n-\tau})\Big\rangle \mid \Sigma_{t-n-\tau}\right]\right] \\
=&\mathbb{E}\left[\Big\langle w_{t-n-\tau} - w_*, \mathbb{E}\left[g(w_{t-n-\tau}, \tilde{O}_t) - \bar{g}(w_{t-n-\tau}) \mid \Sigma_{t-n-\tau}\right]\Big\rangle\right] \\
\leq&\mathbb{E}\left[\|w_{t-n-\tau} - w_*\|\left\|\mathbb{E}\left[g(w_{t-n-\tau}, \tilde{O}_t) - \bar{g}(w_{t-n-\tau}) \mid \Sigma_{t-n-\tau}\right]\right\|\right] \\
\leq&2R\left\|\mathbb{E}\left[g(w_{t-n-\tau}, \tilde{O}_t) - \bar{g}(w_{t-n-\tau}) \mid \Sigma_{t-n-\tau}\right]\right\|
\end{aligned}$$

We now bound $\left\|\mathbb{E}\left[g(w_{t-n-\tau}, \tilde{O}_t) - \bar{g}(w_{t-n-\tau}) \mid \Sigma_{t-n-\tau}\right]\right\|$. In the rest of the proof, all expectations ($\mathbb{E}$) and probabilities (Pr) are conditioned on $\Sigma_{t-n-\tau}$. We suppress the condition in the presentation for improving readability. Let $\bar{O}_t \doteq (\bar{S}_{t-n}, \bar{A}_{t-n}, \dots, \bar{S}_t, \bar{A}_t, \bar{S}_{t+1})$ be a sequence of random variables such that

$$\bar{S}_{t-n} \sim \bar{d}_{\mu_{w_{t-n-\tau}}}, \bar{A}_{t-n} \sim \mu_{w_{t-n-\tau}}(\cdot|\bar{S}_{t-n}), \dots, \bar{A}_t \sim \mu_{w_{t-n-\tau}}(\cdot|\bar{S}_t), \bar{S}_{t+1} \sim p(\cdot|S_t, A_t).$$

Then

$$\begin{aligned}
&\left\|\mathbb{E}\left[g(w_{t-n-\tau}, \tilde{O}_t) - \bar{g}(w_{t-n-\tau})\right]\right\| \\
=&\left\|\mathbb{E}\left[g(w_{t-n-\tau}, \tilde{O}_t) - g(w_{t-n-\tau}, \bar{O}_t)\right]\right\| \\
=&\left\|\sum_{o_t}\left(\Pr\Big(\tilde{O}_t = o_t\Big) - \Pr\big(\bar{O}_t = o_t\big)\right) g(w_{t-n-\tau}, o_t)\right\| \\
\leq&U_g \sum_{o_t}\left|\Pr\Big(\tilde{O}_t = o_t\Big) - \Pr\big(\bar{O}_t = o_t\big)\right| \\
=&U_g \sum_{o_t}\left|\Pr\Big(\tilde{S}_{t-n} = s_{t-n}\Big) - \Pr\big(\bar{S}_{t-n} = s_{t-n}\big)\right| \mu_{w_{t-n-\tau}}(a_{t-n}|s_{t-n}) p(s_{t-n+1}|s_{t-n}, a_{t-n}) \\
&\quad \cdots \mu_{w_{t-n-\tau}}(a_t|s_t) p(s_{t+1}|s_t, a_t) \\
=&U_g \sum_{s_{t-n}}\left|\Pr\Big(\tilde{S}_{t-n} = s_{t-n}\Big) - \Pr\big(\bar{S}_{t-n} = s_{t-n}\big)\right| \\
\leq&U_g C_0 \kappa^{\tau-1} \quad \text{(Lemma 11 and the construction of the auxiliary chain)} \quad ,
\end{aligned}$$

which completes the proof. ■

Appendix C. Complementary Plots

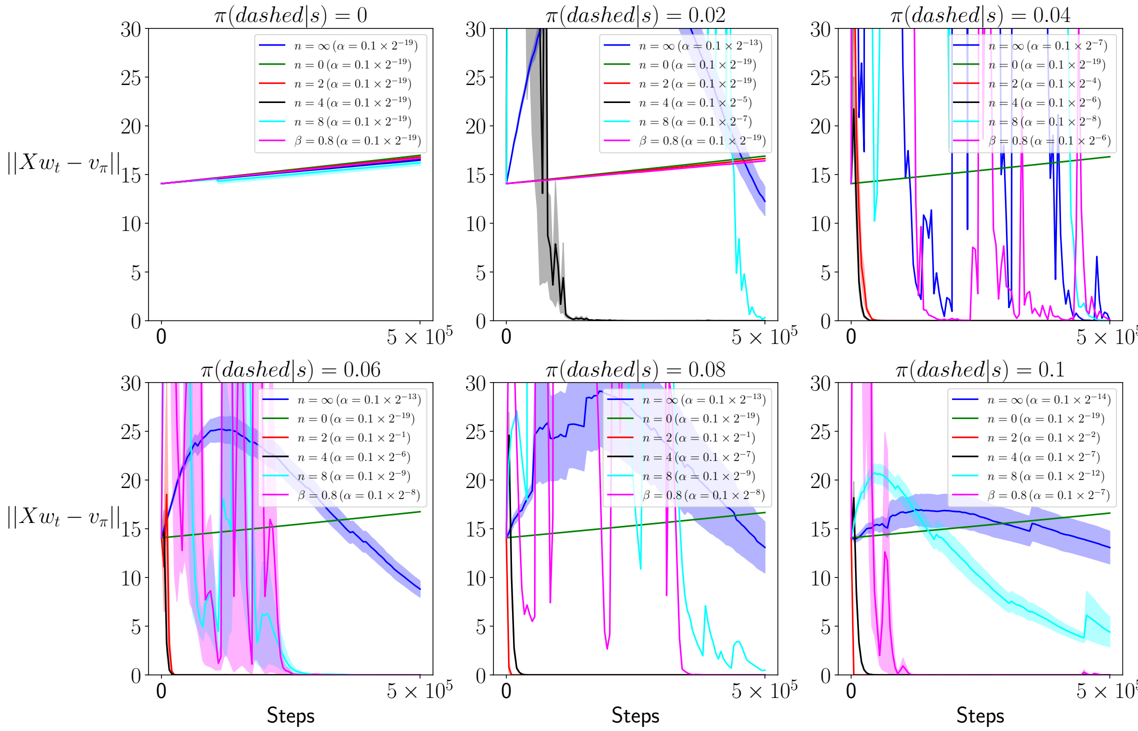

Figure 7: Truncated Emphatic TD and ETD(0, β) in the prediction setting.

References

Kristopher De Asis, Alan Chan, Silviu Pitis, Richard S. Sutton, and Daniel Graves. Fixed-horizon temporal difference methods for stable reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

Leemon C. Baird. Residual algorithms: Reinforcement learning with function approximation. In *Proceedings of the International Conference on Machine Learning*, 1995.

Marc G. Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, 2013.

Albert Benveniste, Michel Métivier, and Pierre Priouret. *Adaptive Algorithms and Stochastic Approximations.* Springer, 1990.

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

Jalaj Bhandari, Daniel Russo, and Raghav Singal. A finite time analysis of temporal difference learning with linear function approximation. In *Proceedings of the Conference on Learning Theory*, 2018.

Gal Dalal, Balázs Szörényi, Gugan Thoppe, and Shie Mannor. Finite sample analyses for td(0) with function approximation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

Daniela Pucci De Farias and Benjamin Van Roy. On the existence of fixed points for approximate value iteration and temporal-difference learning. *Journal of Optimization Theory and Applications*, 2000.

Simon S. Du, Jianshu Chen, Lihong Li, Lin Xiao, and Dengyong Zhou. Stochastic variance reduction methods for policy evaluation. In *Proceedings of the International Conference on Machine Learning*, 2017.

Gabriel Dulac-Arnold, Daniel Mankowitz, and Todd Hester. Challenges of real-world reinforcement learning. *arXiv preprint arXiv:1904.12901*, 2019.

Lasse Espeholt, Hubert Soyer, Rémi Munos, Karen Simonyan, Volodymyr Mnih, Tom Ward, Yotam Doron, Vlad Firoiu, Tim Harley, Iain Dunning, Shane Legg, and Koray Kavukcuoglu. IMPALA: scalable distributed deep-rl with importance weighted actor-learner architectures. In *Proceedings of the International Conference on Machine Learning*, 2018.

Carles Gelada and Marc G. Bellemare. Off-policy deep reinforcement learning by bootstrapping the covariate shift. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019.

Gene H. Golub and Charles F. Van Loan. *Matrix Computations, Third Edition*. Johns Hopkins University Press, 1996. ISBN 978-0-8018-5414-9.

Geoffrey J Gordon. Chattering in sarsa (lambda)-a cmu learning lab internal report. 1996.

Geoffrey J Gordon. Reinforcement learning with function approximation converges to a region. In *Advances in neural information processing systems*, 2001.

Assaf Hallak and Shie Mannor. Consistent on-line off-policy evaluation. In *Proceedings of the International Conference on Machine Learning*, 2017.

Assaf Hallak, Aviv Tamar, Rémi Munos, and Shie Mannor. Generalized emphatic temporal difference learning: Bias-variance analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2016.

Ehsan Imani, Eric Graves, and Martha White. An off-policy policy gradient theorem using emphatic weightings. In *Advances in Neural Information Processing Systems*, 2018.

Ray Jiang, Tom Zahavy, Adam White, Zhongwen Xu, Matteo Hessel, Charles Blundell, and Hado van Hasselt. Emphatic algorithms for deep reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, 2021.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*, 2013.

J. Zico Kolter. The fixed points of off-policy TD. In *Advances in Neural Information Processing Systems*, 2011.

Chandrashekar Lakshminarayanan and Csaba Szepesvári. Linear stochastic approximation: How far does constant step-size and iterate averaging go? In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2018.

Donghwan Lee and Niao He. A unified switching system perspective and ode analysis of q-learning algorithms. *arXiv preprint arXiv:1912.02270*, 2019.

David A Levin and Yuval Peres. *Markov chains and mixing times*. American Mathematical Soc., 2017.

Long Ji Lin. Self-improving reactive agents based on reinforcement learning, planning and teaching. *Machine Learning*, 1992.

Qiang Liu, Lihong Li, Ziyang Tang, and Dengyong Zhou. Breaking the curse of horizon: Infinite-horizon off-policy estimation. In *Advances in Neural Information Processing Systems*, 2018.

Hamid Reza Maei. *Gradient temporal-difference learning algorithms*. PhD thesis, University of Alberta, 2011.

Hamid Reza Maei, Csaba Szepesvári, Shalabh Bhatnagar, and Richard S. Sutton. Toward off-policy learning control with function approximation. In *Proceedings of the International Conference on Machine Learning*, 2010.

Peter Marbach and John N. Tsitsiklis. Simulation-based optimization of markov reward processes. *IEEE Transactions on Automatic Control*, 2001.

Francisco S. Melo, Sean P. Meyn, and M. Isabel Ribeiro. An analysis of reinforcement learning with function approximation. In *Proceedings of the International Conference on Machine Learning*, 2008.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin A. Riedmiller, Andreas Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 2015.

Ofir Nachum, Yinlam Chow, Bo Dai, and Lihong Li. Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections. In *Advances in Neural Information Processing Systems*, 2019.

Matteo Papini, Damiano Binaghi, Giuseppe Canonaco, Matteo Pirotta, and Marcello Restelli. Stochastic variance-reduced policy gradient. In *Proceedings of the International Conference on Machine Learning*, 2018.

Theodore J. Perkins and Doina Precup. A convergent form of approximate policy iteration. In *Advances in Neural Information Processing Systems*, 2002.

Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Joshua Romoff, Peter Henderson, Ahmed Touati, Yann Ollivier, Emma Brunskill, and Joelle Pineau. Separating value functions across time-scales. *arXiv preprint arXiv:1902.01883*, 2019.

Gavin A Rummery and Mahesan Niranjan. *On-line Q-learning using connectionist systems*. University of Cambridge, Department of Engineering Cambridge, UK, 1994.

Rayadurgam Srikant and Lei Ying. Finite-time error bounds for linear stochastic approximation andtd learning. In *Proceedings of the Conference on Learning Theory*, 2019.

Richard S. Sutton. Learning to predict by the methods of temporal differences. *Machine Learning*, 1988.

Richard S Sutton. Generalization in reinforcement learning: Successful examples using sparse coarse coding. *Advances in Neural Information Processing Systems*, 1995.

Richard S Sutton and Andrew G Barto. *Reinforcement Learning: An Introduction (2nd Edition)*. MIT press, 2018.

Richard S. Sutton, Csaba Szepesvári, and Hamid Reza Maei. A convergent o(n) temporal-difference algorithm for off-policy learning with linear function approximation. In *Advances in Neural Information Processing Systems*, 2008.

Richard S. Sutton, Hamid Reza Maei, Doina Precup, Shalabh Bhatnagar, David Silver, Csaba Szepesvári, and Eric Wiewiora. Fast gradient-descent methods for temporal-difference learning with linear function approximation. In *Proceedings of the International Conference on Machine Learning*, 2009.

Richard S. Sutton, Joseph Modayil, Michael Delp, Thomas Degris, Patrick M. Pilarski, Adam White, and Doina Precup. Horde: a scalable real-time architecture for learning knowledge from unsupervised sensorimotor interaction. In *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems*, 2011.

Richard S. Sutton, Ashique Rupam Mahmood, and Martha White. An emphatic approach to the problem of off-policy temporal-difference learning. *Journal of Machine Learning Research*, 2016.

Ahmed Touati, Pierre-Luc Bacon, Doina Precup, and Pascal Vincent. Convergent tree backup and retrace with function approximation. In *Proceedings of the International Conference on Machine Learning*, 2018.

John N. Tsitsiklis and Benjamin Van Roy. Analysis of temporal-diffference learning with function approximation. In *Advances in Neural Information Processing Systems*, 1996.

Yue Wang and Shaofeng Zou. Finite-sample analysis of greedy-gq with linear function approximation under markovian noise. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 2020.

Martha White. Unifying task specification in reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, 2017.

Huizhen Yu. Least squares temporal difference methods: An analysis under general conditions. *SIAM Journal on Control and Optimization*, 2012.

Huizhen Yu. On convergence of emphatic temporal-difference learning. In *Proceedings of the Conference on Learning Theory*, 2015.

Huizhen Yu. On convergence of some gradient-based temporal-differences algorithms for off-policy learning. *arXiv preprint arXiv:1712.09652*, 2017.

Shangtong Zhang, Bo Liu, and Shimon Whiteson. GradientDICE: Rethinking generalized offline estimation of stationary values. In *Proceedings of the International Conference on Machine Learning*, 2020a.

Shangtong Zhang, Bo Liu, Hengshuai Yao, and Shimon Whiteson. Provably convergent two-timescale off-policy actor-critic with function approximation. In *Proceedings of the International Conference on Machine Learning*, 2020b.

Shangtong Zhang, Remi Tachet des Combes, and Romain Laroche. Global optimality and finite sample analysis of softmax off-policy actor critic under state distribution mismatch. *arXiv preprint arXiv:2111.02997*, 2021a.

Shangtong Zhang, Hengshuai Yao, and Shimon Whiteson. Breaking the deadly triad with a target network. In *Proceedings of the International Conference on Machine Learning*, 2021b.

Shaofeng Zou, Tengyu Xu, and Yingbin Liang. Finite-sample analysis for SARSA with linear function approximation. In *Advances in Neural Information Processing Systems*, 2019.