Consistent Emphatic Temporal-Difference Learning

Abstract

Off-policy policy evaluation has been a critical and challenging problem in reinforcement learning, and Temporal-Difference (TD) learning is one of the most important approaches for addressing it. Notably, Full Importance-Sampling TD is the only existing off-policy TD method that is guaranteed to find the on-policy TD fixed point in the linear function approximation setting but, unfortunately, has a high variance and is scarcely practical. This notorious high variance issue motivates the introduction of Emphatic TD, which tames down the variance but has a biased fixed point. Inspired by these two methods, we propose a new consistent algorithm with a transient bias, which strikes a balance between bias and variance. Further, we unify the new algorithm with several existing algorithms and obtain a new family of consistent algorithms called *Consistent Emphatic TD* (CETD(λ, β, ν)), which can control a smooth bias-variance trade-off by varying the speed at which the transient bias fades. Through theoretical analysis and experiments on a didactic example, we validate the consistency of CETD(λ, β, ν). Moreover, we show that CETD(λ, β, ν) converges faster to the lowest error in a complex task with a high variance.

1 INTRODUCTION

Off-policy learning is a critical area in reinforcement learning (RL). Particularly, off-policy policy evaluation (OPPE), also known as off-policy prediction, is an essential component in model learning, options learning (Sutton et al., 1999), and life-long learning (Sutton et al., 2022; White et al., 2012). The goal of OPPE is to estimate the value function of a *target policy* with data collected by a different *behavior policy*. We refer to the data collected by the target policy as *on*-policy data and the data collected by the behavior policy *off*-policy data. In this paper, we consider the problem of OPPE with linear function approximation.

In the online RL setting, where the algorithm makes incremental updates and does not store data samples, Temporal-Difference (TD) learning is a ubiquitous family of algorithms, and On-policy TD(λ) is an essential approach to on-policy prediction (Sutton, 1988). In OPPE, there have been substantial efforts in obtaining the *on-policy fixed point*, to which On-policy TD(λ) converges with on-policy data (Precup et al., 2001; Hallak & Mannor, 2017; Liu et al., 2018; Nachum et al., 2019; Zhang et al., 2020a,b). There is a good reason for targeting the on-policy fixed point: It can produce a good approximation of the target policy's value function (Tsitsiklis & Van Roy, 1996). We say such an off-policy TD algorithm is *consistent* if it has the same fixed point as the on-policy fixed point, and we resort to consistent off-policy TD algorithms as our approach.

To the best of our knowledge, Full Importance-Sampling TD (Full IS TD(λ), Precup et al., 2001) is the only consistent off-policy TD algorithm without making any significant assumption, while others require additional assumptions on the structure of the problem, especially the feature representation (Hallak & Mannor, 2017; Liu et al., 2018; Nachum et al., 2019; Zhang et al., 2020a,b). To obtain the on-policy fixed point, Full IS TD(λ) reweights the TD update with a full importance-sampling-ratio (IS-ratio) product, which is the multiplication of the IS ratios at every time step. However, Full IS TD(λ) barely works in practice due to the high variance of the full IS-ratio product.

Following Full IS TD(λ)'s idea of calibrating the fixed point of TD by reweighting the TD update, Emphatic TD (ETD(λ), Sutton et al., 2016) uses an alternative emphatic weighting, which is a geometrically weighted sum of IS-ratio products accumulated from every time step. While the IS-ratio products with fewer terms do mitigate the variance issue, however, they also induce a persistent bias, deviating from the on-policy TD fixed point. Further, to obtain a smooth bias-

Submitted to the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023). **To be used for reviewing only**.

variance trade-off, Hallak et al. (2016) proposed ETD(λ, β), which unifies Off-policy TD(λ) and ETD(λ) with a tunable parameter β. Yet, ETD(λ, β) would lose the stability guarantee when β is smaller than an instance-dependent condition number that is difficult to determine.

In this paper, we propose a new consistent algorithm inspired by Full IS TD(λ) and ETD(λ), which strikes a better balance between bias and variance. The new algorithm renovates the idea of ETD(λ), introducing a transient bias to achieve a lower variance than Full IS TD(λ) while retaining consistency as the bias fades away over time. To make the new algorithm more practical, we introduce extra hyperparameters to control a smooth bias-variance trade-off by unifying the new algorithm with existing algorithms. The resulting new family of consistent algorithms called *Consistent Emphatic TD* (CETD(λ, β, ν)) has a more general stability guarantee than ETD(λ, β), the same fixed point as On-policy TD(λ), and much better performance than Full IS TD(λ), the only consistent TD method previously. Through experiments on a didactic example, we verify the consistency of CETD(λ, β, ν). Experiment results on a more complex task with high variance also show CETD(λ, β, ν)'s faster convergence to the lowest error compared to the competitors mentioned above. To the best of our knowledge, CETD(λ, β, ν) is the *first practical consistent* algorithm for off-policy TD learning with linear function approximation.

2 BACKGROUND

We consider an infinite horizon Markov Decision Process (MDP), which is defined as a tuple $\langle \mathcal{S}, \mathcal{A}, P, r, \gamma \rangle$ where $\mathcal{S}$ is the finite state space, $\mathcal{A}$ is the finite action space, $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition function, $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function, and $\gamma < 1$ is the discount factor. The policy of the agent is defined as $\pi : \mathcal{S} \to \Delta(\mathcal{A})$. We denote the size of the state space and the action space by $S \doteq |\mathcal{S}|$ and $A \doteq |\mathcal{A}|$. The discounted value function is defined as

$$v_\pi(s) \doteq \mathbb{E}_\pi \left[\sum_{t=1}^{\infty} \gamma^{t-1} r(S_t, A_t) | S_0 = s \right].$$

Particularly, we consider the setting of *online off-policy policy evaluation with linear function approximation* where the agent needs to estimate the value function of a target policy π while interacting with the environment with a behavior policy μ without storing samples. We assume the MDP is parameterized by the feature function $\boldsymbol{\phi} : \mathcal{S} \to \mathbb{R}^d$ or equivalently the feature matrix $\boldsymbol{\Phi} \in \mathbb{R}^{S \times d}$, where d is the dimension of the feature. At each time step t, instead of observing S_t, the agent observes the feature of the state $\boldsymbol{\phi}_t \doteq \boldsymbol{\phi}(S_t)$. Then the agent selects action A_t based on the behavior policy μ and observes the next state feature $\boldsymbol{\phi}_{t+1}$ and reward $R_{t+1} = r(S_t, A_t)$. The importance-sampling ratio at time step t is defined as $\rho_t = \frac{\pi(A_t|S_t)}{\mu(A_t|S_t)}$. With linear function approximation, the agent approximates the value function with $\hat{v}(s; \boldsymbol{\theta}) = \boldsymbol{\phi}(s)^\top \boldsymbol{\theta}$ or in matrix form, $\hat{\boldsymbol{v}} = \boldsymbol{\Phi}\boldsymbol{\theta}$, where $\boldsymbol{\theta} \in \mathbb{R}^d$ is a parameter vector. We assume the Markov chains induced by the behavior policy μ and the target policy π are irreducible, which ensures the unique existence of corresponding stationary distributions $d_\mu \in \Delta(\mathcal{S})$ and $d_\pi \in \Delta(\mathcal{S})$. Similar to identifying $\hat{v}$ as $\hat{\boldsymbol{v}}$, we also identify d_μ as $\boldsymbol{d}_\mu$ and d_π as $\boldsymbol{d}_\pi$. Moreover, we define $\boldsymbol{D_v} \doteq diag(\boldsymbol{v})$ for some vector $\boldsymbol{v}$. Specifically, we use $\boldsymbol{D}_\pi$ for $\boldsymbol{D}_{\boldsymbol{d}_\pi}$ and $\boldsymbol{D}_\mu$ for $\boldsymbol{D}_{\boldsymbol{d}_\mu}$. We use $\| \cdot \|_{\boldsymbol{v}}$ to denote the vector norm induced by $\boldsymbol{D_v}$ for some vector $\boldsymbol{v}$, i.e., $\|\boldsymbol{x}\|_{\boldsymbol{v}} = \sqrt{\boldsymbol{x}^\top \boldsymbol{D_v} \boldsymbol{x}}$.

Stability and Consistency We will define the stability of a stochastic algorithm with an update of the following form:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \alpha(\boldsymbol{b}_t - \boldsymbol{A}_t \boldsymbol{\theta}_t),$$

where α is the step size parameter, $\{\boldsymbol{\theta}_t\}_{t=0}^\infty$ is the sequence of weight vectors generated by the algorithm, and $\{(\boldsymbol{A}_t, \boldsymbol{b}_t)\}_{t=0}^\infty$ is a sequence of random matrices and vectors that depend on the problem and the algorithm. Define $\boldsymbol{A} \doteq \lim_{t\to\infty} \mathbb{E}_\mu[\boldsymbol{A}_t]$ and $\boldsymbol{b} \doteq \lim_{t\to\infty} \mathbb{E}_\mu[\boldsymbol{b}_t]$. Using $\boldsymbol{A}$ and $\boldsymbol{b}$, we can form a deterministic algorithm:

$$\bar{\boldsymbol{\theta}}_{t+1} = \bar{\boldsymbol{\theta}}_t + \alpha(\boldsymbol{b} - \boldsymbol{A}\bar{\boldsymbol{\theta}}_t),$$

which we call the *expected update* of the stochastic algorithm. We use the definition of the stability of a stochastic algorithm from Sutton et al. (2016): A stochastic algorithm and its expected update are *stable* if the expected update converges to a unique fixed point under any initialization. As discussed in Sutton et al. (2016), the stability of a stochastic algorithm is essential to its convergence: If a stochastic algorithm is stable, then its parameter vector may converge with probability one with a proper step-size scheduling. Besides, if the stochastic algorithm converges, then it would converge to the fixed point of its expected update, $\bar{\boldsymbol{\theta}} = \boldsymbol{A}^{-1}\boldsymbol{b}$. Next, we define the consistency of an off-policy TD algorithm with hyperparameter λ as follows: such an algorithm is considered *consistent* if its expected update converges to the same fixed point as On-policy TD(λ). Further, a consistent off-policy TD algorithm will also be stable, as On-policy TD(λ) is stable.

In our pursuit of consistent off-policy TD algorithms, we will review a line of work that has made progress toward this goal by reweighting the TD update.

Off-Policy TD(λ) Off-policy TD(λ) (Precup, 2000) is the earliest effort in this line of work. In the one-step case, Off-policy TD(0) makes the following update:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \alpha \rho_t \delta_t \boldsymbol{\phi}_t, \tag{1}$$

$$\delta_t = R_{t+1} + \gamma \boldsymbol{\phi}_{t+1}^\top \boldsymbol{\theta}_t - \boldsymbol{\phi}_t^\top \boldsymbol{\theta}_t, \tag{2}$$

where α is the step size parameter. Compared to On-policy TD(0), Off-policy TD(0) uses the IS ratio ρ_t to correct the

probability of selecting action A_t at time step t, which allows Off-policy TD(0) to converge to the true value function v_π if the feature representation is tabular. This convergence guarantee is true for any $\lambda \in [0, 1]$ in the tabular case. However, Off-policy TD(λ) is shown to diverge in various counterexamples with general linear features (Baird, 1995; Sutton & Barto, 2018). This divergence issue is due to the distribution of Update (1) could deviate too much from the on-policy distribution, which is well explained in Sutton et al. (2016). What is worse, the *off-policy fixed point*, the fixed point that Off-policy TD(0) converges to if it does converge, could have an unbounded error (Kolter, 2011).

Full IS TD(λ) To address the divergence of Off-policy TD(λ), Precup et al. (2001) introduced the idea of correcting the distribution of Update (1) by reweighting it. The algorithm they proposed, Full IS TD(λ), reweights the update with a *full* IS-ratio product, the product of all the IS ratios up to the current time step. Before we bring in the update of Full IS TD(0), we introduce a general description of one-step TD algorithms with an unspecified trace F_t:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \alpha \rho_t F_t \delta_t \boldsymbol{\phi}_t, \tag{3}$$

where δ_t is the TD error defined in Eq. (2). For Off-policy TD(0), $F_t = 1$. In the case of Full IS TD(0), $F_t = \rho_{t-1}\rho_{t-2}\cdots\rho_0$ or $F_t = \rho_{t-1}F_{t-1}$ with $F_0 = 1$, which always corrects the distribution back to the on-policy distribution completely (Precup et al., 2001). In general, for Full IS TD(λ), $\mathbb{E}_\mu[F_t|S_t = s]$ is equal to $\frac{\mathbb{P}_\pi(S_t=s)}{\mathbb{P}_\mu(S_t=s)}$, which converges to the density ratio $\frac{d_\pi(s)}{d_\mu(s)}$. Consequently, the expected update of Full IS TD(λ) is the same as On-policy TD(λ), implying its consistency. Since F_t given $S_t = s$ is an unbiased estimator of $\frac{\mathbb{P}_\pi(S_t=s)}{\mathbb{P}_\mu(S_t=s)}$, Full IS TD($\lambda$) is considered an *unbiased* off-policy TD algorithm. However, Full IS TD(λ) is scarcely practical due to variance issues, which motivates the next idea in this line.

ETD(λ) Instead of using the full IS-ratio product to weight the update, Sutton et al. (2016) proposed to use the emphatic weighting:

$$F_t = \gamma \rho_{t-1} F_{t-1} + 1, \text{with } F_0 = 1, \tag{4}$$

which is termed the *followon trace*, a geometrically weighted sum of IS-ratio products accumulated from different time steps. By introducing *incomplete* IS-ratio products into the weighting, the proposed algorithm, ETD(λ), reduces the variance and remains stable. However, these incomplete IS-ratio products also introduce persistent bias into F_t, causing ETD(λ) to converge to a biased fixed point.

3 TOWARDS PRACTICAL CONSISTENT TD LEARNING

AVERAGE FOLLOWON TRACE

As discussed in the last section, the unbiased and the only consistent method, Full IS TD(λ), is not practical due to the high variance issue. On the other hand, ETD(λ), an effective remedy to the variance issue, is biased and deviates from our objective of finding the on-policy fixed point. Then, *could we find a good trade-off point between Full IS TD(λ) and ETD(λ)?* Specifically, could we find a method that is consistent and has a milder variance compared to Full IS TD(λ)? The answer is, fortunately, yes. Inspired by the idea of using incomplete IS-ratio products to reduce the variance, we propose to use the below average emphatic weighting:

$$F_t = \frac{t}{t+1}\rho_{t-1}F_{t-1} + \frac{1}{t+1}, \text{with } F_0 = 1, \tag{5}$$

which we term the *average followon trace*. If we expand the average followon trace, we can see that it is a uniform average of the IS-ratio products. The motivation for using an average weighting instead of the geometrically decayed weighting is to gradually reduce bias by de-emphasizing the new IS-ratio product at each time step. In fact, the bias would vanish in the limit, and the resulting one-step algorithm defined by Update (3) and the average followon trace is consistent. We call this algorithm *Average Emphatic TD(0)* (AETD(0)) and present its consistency in Theorem 3.1, of which the formal statement and proof are deferred to the appendix.

Theorem 3.1 (Informal)**.** *Under some mild assumptions, AETD(0) is stable and consistent.*

Now, the idea of using a uniformly weighted sum of IS-ratio products to reweight the TD update is not completely new. Hallak et al. (2016) generalized ETD(λ) to include a tunable decay parameter β to control the decay rate of the followon trace (Eq. (4)). The resulting algorithm, ETD(λ, β), uses the following emphatic weighting:

$$F_t = \beta \rho_{t-1} F_{t-1} + 1, \text{with } F_0 = 1. \tag{6}$$

When choosing $\beta = 1$, the trace would equally weight each IS-ratio product with weight 1. However, equally weighting the products would be problematic in their case because the expectation of F_t would converge to infinity in the limit. Nevertheless, ETD(λ, β) presents an interesting strategy to achieve a smooth bias-variance trade-off. By unifying existing methods and introducing a controllable parameter, β, ETD(λ, β) achieves a smooth bias-variance trade-off: When $\beta = 0$, ETD(λ, β) becomes Off-policy TD(λ), which has a lower variance but a larger bias; When $\beta = \gamma$, ETD(λ, β) becomes ETD(λ), which has a lower bias but a larger variance. Then, *could we adopt the same strategy to attain a smooth bias-variance trade-off for AETD(0)?*

Table 1: The coefficients of different IS-ratio products in F_t.

IS-ratio Product	Off-policy TD(λ)	Scaled ETD(λ, β)	Full IS TD(λ)	AETD(λ)	CETD(λ, β, ν)	ETD(λ, β)
1	1	$1-\beta$	0	$1/(t+1)$	$f(t)$	1
ρ_{t-1}	0	$\beta(1-\beta)$	0	$1/(t+1)$	$(1-f(t))f(t-1)$	β
$\rho_{t-1}\rho_{t-2}$	0	$\beta^2(1-\beta)$	0	$1/(t+1)$	$\Pi_{k=t-1}^{t}(1-f(k))f(t-2)$	β^2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\Pi_{k=2}^{t}\rho_{k-1}$	0	$\beta^{t-1}(1-\beta)$	0	$1/(t+1)$	$\Pi_{k=2}^{t}(1-f(k))f(1)$	β^{t-1}
$\Pi_{k=1}^{t}\rho_{k-1}$	0	β^t	1	$1/(t+1)$	$\Pi_{k=1}^{t}(1-f(k))$	β^t

A SMOOTH BIAS-VARIANCE TRADE-OFF

In this section, we try to answer the question asked above. In particular, we seek a way to control the speed at which the bias fades and retain the consistency of AETD(0). To attain this goal, we consider the unification of AETD(0) with both Off-policy TD(0) and Full IS TD(0), the one with the least variance but the greatest bias and the one with the least bias but the greatest variance.

We start by unifying AETD(0) and Full IS TD(0). The traces that the two methods use are $F_t = \frac{t}{t+1}\rho_{t-1}F_{t-1} + \frac{1}{t+1}$ and $F_t = \rho_{t-1}F_{t-1}$. To unify these two traces, we introduce a tunable parameter $\beta' \in [0,1]$ to the average followon trace: $F_t^{(1)} = (1-\beta'(t+1)^{-1})\rho_{t-1}F_{t-1}^{(1)} + \beta'(t+1)^{-1}$ with $F_0^{(1)} = 1$. Then, when $\beta' = 0$, $F_t^{(1)}$ becomes $F_t^{(1)} = \rho_{t-1}F_{t-1}^{(1)}$, which corresponds to the trace of Full IS TD(0); When $\beta' = 1$, $F_t^{(1)}$ becomes the average followon trace.

Similarly, we can unify AETD(0) and Off-policy TD(0) with another tunable parameter ν and a new trace: $F_t^{(2)} = (1-(t+1)^{-\nu})\rho_{t-1}F_{t-1}^{(2)} + (t+1)^{-\nu}$ with $F_0^{(2)} = 1$. When $\nu = 0$, $F_t^{(2)}$ becomes constant 1, which corresponds to the trace of Off-policy TD(0); When $\nu = 1$, $F_t^{(2)}$ becomes the average followon trace.

Then we further unify the two traces, $F_t^{(1)}$ and $F_t^{(2)}$, leading us to a third trace with two parameters β' and ν: $F_t^{(3)} = (1-\beta'(t+1)^{-\nu})\rho_{t-1}F_{t-1}^{(3)} + \beta'(t+1)^{-\nu}$ with $F_0^{(3)} = 1$. It seems that we have finished the work of unification. However, we found that when $\nu = 0$, the trace becomes $F_t = (1-\beta')\rho_{t-1}F_{t-1} + \beta'$, which is also a geometrically weighted sum of IS-ratio products as in ETD(λ, β). To further obtain the same decay rate in the resulting trace and the followon trace (Eq. 6), we replace β' with $1-\beta$ in $F_t^{(3)}$, and name the resulting trace *general followon trace*:

$$F_t = (1-f(t))\rho_{t-1}F_{t-1} + f(t), \text{with } F_0 = 1, \quad (7)$$

where $f(t) \doteq (1-\beta)(t+1)^{-\nu}$ with $\beta \in [0,1]$ and $\nu \in [0,1]$. Note that when we choose $\nu = 0$, the resulting trace would be $F_t = \beta\rho_{t-1}F_{t-1} + (1-\beta)$ with $F_0 = 1$, which we call the scaled followon trace. The resulting one-step algorithm is subsequently called Scaled ETD(0, β). Although the scaled followon trace has the same decay rate as the original followon trace (Eq. (4)), it is downscaled by $1-\beta$ (see Table 1). This discrepancy, however, is not a qualitative difference because the constant factor $1-\beta$ could be absorbed in the step size hyperparameter (see Update (3)). In addition, if we take the downscaling view, we will notice that the full IS-ratio product term is not downscaled to $\beta^t(1-\beta)$. However, this minor difference would not prevent Scaled ETD(0, β) from sharing the same theory and empirical performance as ETD(0, β) (with the step size scaled). Thus, Scaled ETD(0, β) could be viewed as a slight variant of ETD(0, β).

Having settled the relationship between Scaled ETD(0, β) and ETD(0, β), we are now ready to name the algorithm that unifies AETD(0), Off-policy TD(0), Full IS TD(0), and Scaled ETD(0, β). We call the resulting one-step algorithm *General Emphatic TD(0)* (GETD(0)), which is defined by Update (3) and the general followon trace.

So far, we have only introduced the one-step form of AETD and GETD. By applying the same idea of uniform averaging and the same strategy of unification for the multi-step bootstrapping case, we can obtain the multi-step versions of them. Here, we present the unified algorithm with multi-step bootstrapping called *General Emphatic TD* (GETD(λ, β, ν)), which makes the following update[1]:

$$\begin{aligned}
\boldsymbol{\theta}_{t+1} &= \boldsymbol{\theta}_t + \alpha\delta_t\boldsymbol{z}_t,\\
\delta_t &= R_{t+1} + \gamma\boldsymbol{\phi}_{t+1}^\top\boldsymbol{\theta}_t - \boldsymbol{\phi}_t^\top\boldsymbol{\theta}_t,\\
\boldsymbol{z}_t &= \rho_t(\gamma\lambda\boldsymbol{z}_{t-1} + M_t\boldsymbol{\phi}_t), \text{with } \boldsymbol{z}_{-1} = \mathbf{0},\\
M_t &= (1-g(t))\,F_t + g(t),\\
F_t &= (1-f(t))\,\rho_{t-1}F_{t-1} + f(t), \text{with } F_0 = 1,
\end{aligned} \quad (8)$$

where $g(t)$ and $f(t)$ are defined as follows:

$$g(t) \doteq \lambda\left(\frac{1-\beta}{t+1}\right)^{\nu} \text{ and } f(t) \doteq \frac{1-\beta}{(t+1)^{\nu}}. \quad (9)$$

Similar to the one-step bootstrapping case, GETD(λ, β, ν) subsumes AETD(λ) , Off-policy TD(λ), Full IS TD(λ), Scaled ETD(λ), and Scaled ETD(λ, β). A list of the updates of all these algorithms is included in the appendix.

[1]For simplicity, we have not included general state-dependent interest, discounting, and bootstrapping functions as in previous work (Sutton et al., 2016). However, GETD(λ, β, ν) can be easily extended to those cases.

CONSISTENT EMPHATIC TD

In this section, we examine the product of the unification. The question here is, *does the introduced tunable decay parameters β and ν bring us a smooth bias-variance trade-off while retaining the consistency?* Fortunately, the answer is, again, yes. We name this new class of consistent algorithms with tunable decay parameters as *Consistent Emphatic TD* (CETD(λ, β, ν)). Specifically, CETD(λ, β, ν) is defined by Update (8) with $\beta \in [0, 1)$ and $\nu \in (0, 1]$, excluding existing methods in GETD(λ, β, ν). We present the consistency of CETD(λ, β, ν) in Theorem 3.2, of which the formal statement and the proof are deferred to the appendix.

Theorem 3.2 (Informal)**.** *Under some mild assumptions, for any $\beta \in [0, 1)$ and $\nu \in (0, 1]$, CETD(λ, β, ν) is stable and consistent.*

Remark 3.3. CETD(λ, β, ν) is stable for any values of $\beta \in [0, 1)$ and $\nu \in (0, 1]$. This is significantly stronger than ETD(λ, β) (Hallak et al., 2016). In their case, ETD(λ, β) is stable only with $\beta > \beta_0$ where $\beta_0 \leq \gamma$ is an instance-dependent condition number.

Remark 3.4. CETD(λ, β, ν) is consistent for any values of $\beta \in [0, 1)$ and $\nu \in (0, 1]$. This is, again, significantly stronger than ETD(λ, β). For any $\beta \in [0, 1)$, ETD(λ, β) has a persistent bias. In particular, the bias would increase as the value of β decrease. At the extreme end when $\beta = 0$, ETD(λ, β) becomes Off-policy TD(λ), which could have unbounded bias (Kolter, 2011). On the other hand, CETD(λ, β, ν)'s bias is transient for any $\beta \in [0, 1)$ and $\nu \in (0, 1]$.

Having settled the consistency of CETD(λ, β, ν), we now discuss the bias-variance trade-off we obtained. Figure 1 plots the landscape of GETD(λ, β, ν), which clearly illustrates the relationship between CETD(λ, β, ν) and other algorithms. Starting from AETD(λ), intuitively, as the value of ν decreases, the algorithm would get closer to Off-policy TD(λ) with the variance decreased but the bias increased; Meanwhile, as the value of β increases, the algorithm would get closer to Full IS TD(λ) with the bias decreased but the variance increased. More generally, it would hold for CETD(λ, β, ν) that increasing β or ν would reduce the bias and increase the variance, and vice versa.

To better analyze the bias-variance trade-off that β and ν control, we study CETD(λ, β, ν)'s three instances, which cover a diagonal line and two edges of CETD(λ, β, ν) (see Figure 1). The first instance is CETD1(λ, β), which corresponds to a diagonal line of CETD(λ, β, ν). In this diagonal line, the value of ν is always the same as the value of β. This line has the special property that it connects Off-policy TD(λ) and Full IS TD(λ). The update of CETD1(λ, β) is the same as Update (8) but with $g(t)$ and $f(t)$ specified as the following:

$$g(t) \doteq \lambda \left(\frac{1-\beta}{t+1}\right)^{\beta} \text{ and } f(t) \doteq \frac{1-\beta}{(t+1)^{\beta}}. \tag{10}$$

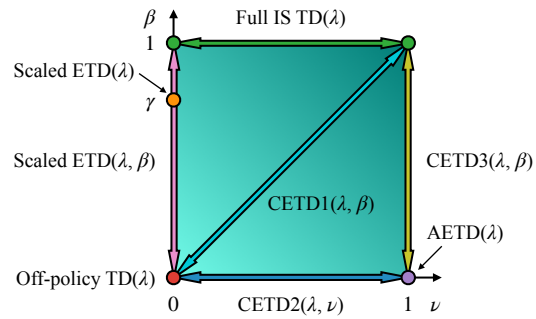


Figure 1: The landscape of GETD(λ, β, ν). The square excluding the top and left edges represents CETD(λ, β, ν). The darkness of the color at each point inside the square represents the magnitude of F_t's variance.

The second instance is CETD2(λ, ν), which corresponds to the bottom edge of CETD(λ, β, ν). In this line, β is always 0. Towards the left endpoint, CETD2(λ, ν) approaches Off-policy TD(λ); At the right endpoint, CETD2(λ, ν) becomes AETD(λ). The update of CETD2(λ, ν) is the same as Update (8) but with $g(t)$ and $f(t)$ specified as the following:

$$g(t) \doteq \lambda(t+1)^{-\nu} \text{ and } f(t) \doteq (t+1)^{-\nu}. \tag{11}$$

The third instance is CETD3(λ, β), which corresponds to the right edge of CETD(λ, β, ν). In this edge, ν is always 1. At the bottom endpoint, CETD3(λ, β) becomes AETD(λ). Towards the top endpoint, CETD3(λ, β) approaches Full IS TD(λ). The update of CETD3(λ, β) is the same as Update (8) but with $g(t)$ and $f(t)$ specified as the following:

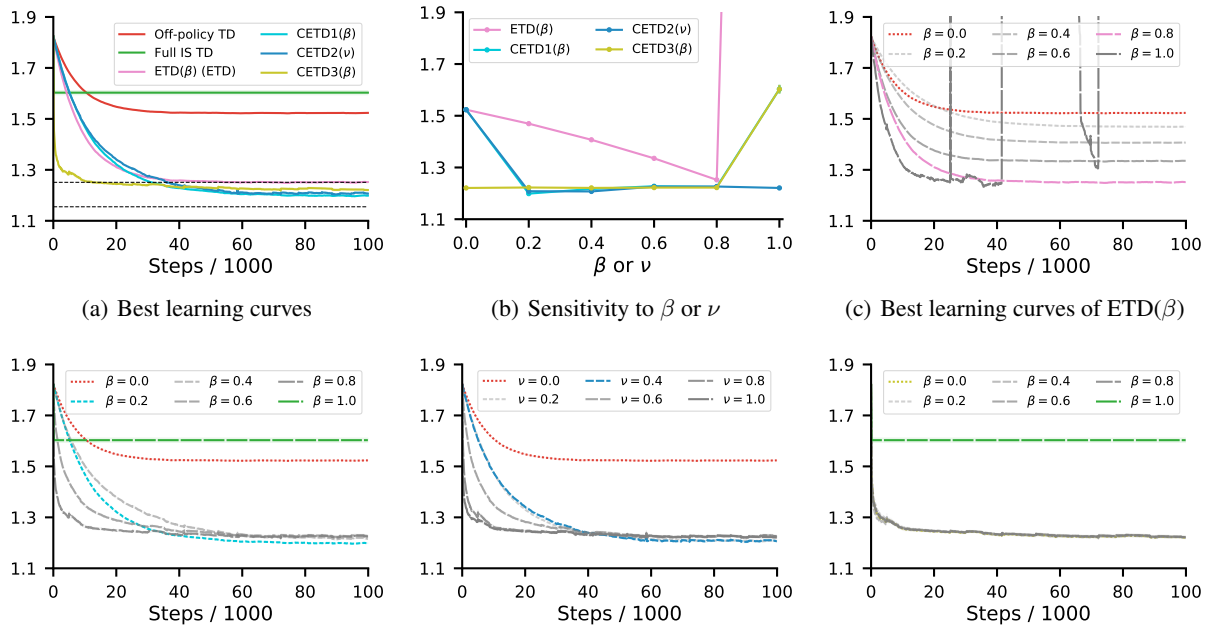
$$g(t) \doteq \frac{\lambda(1-\beta)}{t+1} \text{ and } f(t) \doteq \frac{(1-\beta)}{t+1}. \tag{12}$$

4 EXPERIMENTS

In this section, we present experiments that demonstrate CETD(λ, β, ν)'s empirical practicality in the one-step case. For simplicity, we drop the λ argument in all the algorithms. For example, we will use CETD(β, ν) for CETD(0, β, ν). The results for the multi-step case are similar to the one-step case, which are deferred to the appendix. To evaluate the quality of the learned $\boldsymbol{\theta}$, we use the root-mean-square-value error for our metric:

$$\overline{\text{RMSVE}}(\boldsymbol{\theta}) = \|\hat{\boldsymbol{v}}_{\boldsymbol{\theta}} - \boldsymbol{v}_{\pi}\|_{\boldsymbol{d}_{\pi}},$$

For all the experiments, we use constant step sizes $\alpha = 2^x$ for all algorithms where $x \in \{-18, -17, \cdots, -1, 0\}$. For *tunable algorithms* with an adjustable decay parameter (ETD(β), CETD1(β), CETD2(ν), and CETD3(β)), the decay parameter (β or ν) is chosen from $\{0.0, 0.2, 0.4, 0.6, 0.8, 1.0\}$. Notice that ETD($\beta$), CETD1($\beta$), and CETD2($\nu$) with $\beta = 0.0$ or $\nu = 0.0$ are the same as Off-policy TD; CETD1(β) and CETD3(β)



(a) Best learning curves (b) Sensitivity to β or ν (c) Best learning curves of ETD(β)

(d) Best learning curves of CETD1(β) (e) Best learning curves of CETD2(ν) (f) Best learning curves of CETD1(β)

Figure 2: Performance of different algorithms on the Two-state task. The y-axis shows $\overline{\text{RMSVE}}$. The dash lines from top to bottom in Figure (a) show $\overline{\text{RMSVE}}(\theta^*_{\text{ETD}}) \approx 1.251$ and $\overline{\text{RMSVE}}(\theta^*_{\text{On}}) \approx 1.155$, respectively.

with $\beta = 1.0$ are the same as Full IS TD; ETD(β) with $\beta = 1.0$ is an unsound method with a followon trace whose expectation would blow up to infinity in the limit. All the results are reported with the best-performing step size, with which the final error is the smallest. The final error is calculated by averaging the errors in the last 1% of the training steps. Compared to the area under the learning curve (AUC), the final error is favored because it is a better reflection of how the algorithm performs asymptotically.

CONSISTENCY OF CETD(β, ν)

Two-State Didactic Task To verify the consistency of CETD(β, ν), we designed a didactic task with two states (see Figure 3). In this task, the target policy π will go to the left state from any state with a probability of 0.6, while the probability for the behavior policy is 0.4. The discount factor γ is 0.8. The on-policy fixed point for this task is $\theta^*_{\text{On}} \approx 0.8257$, which induces an error of $\overline{\text{RMSVE}}(\theta^*_{\text{On}}) \approx 1.155$. The off-policy fixed point is $\theta^*_{\text{Off}} \approx 0.3061$, which induces an error of $\overline{\text{RMSVE}}(\theta^*_{\text{Off}}) \approx 1.523$. The fixed point of ETD (ETD(β) with $\beta = 0.9$) is $\theta^*_{\text{ETD}} \approx 0.7392$, which induces an error of $\overline{\text{RMSVE}}(\theta^*_{\text{ETD}}) \approx 1.251$. Thus, consistent algorithms have a theoretical advantage in this task because their fixed point is the on-policy fixed point, which has the lowest $\overline{\text{RMSVE}}$. In this task, we run each algorithm for $100,000$ steps and present the experiment results in Figure 2, which are averaged over 100 independent runs.

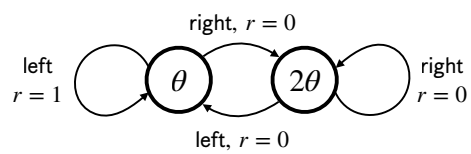
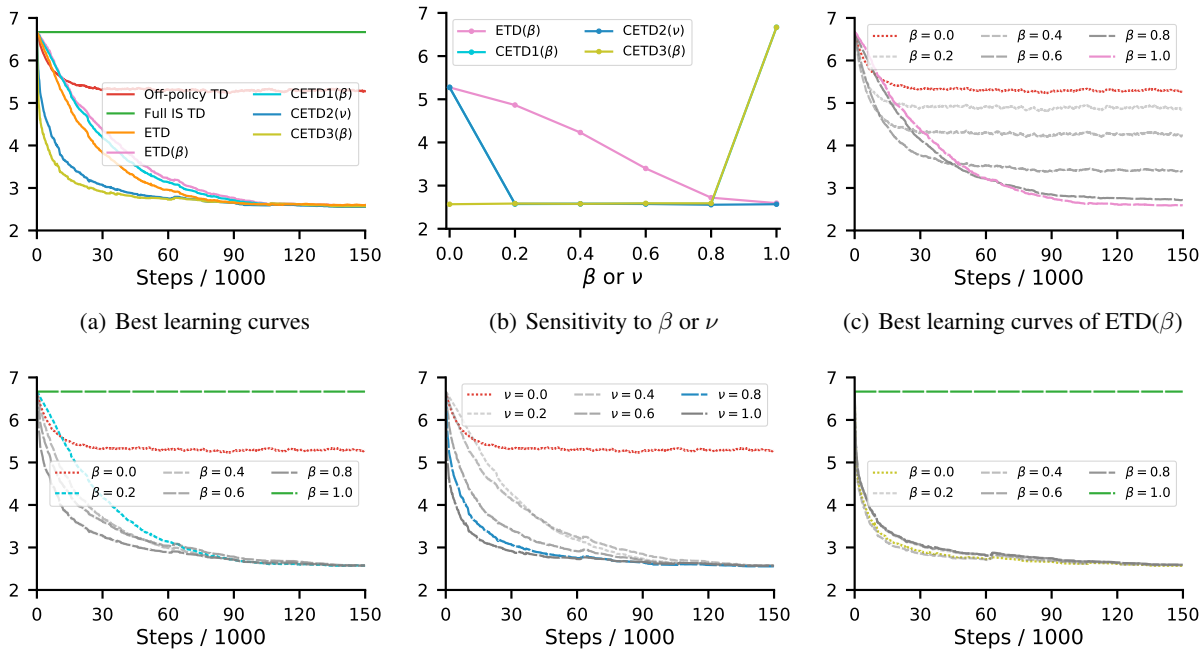


Figure 3: The Two-state Didactic task. The values of the two states are approximated by θ and 2θ, respectively.

From Figure 2(a), we can see that all CETD instances achieve the lowest error between $\overline{\text{RMSVE}}(\theta^*_{\text{ETD}}) \approx 1.251$ and $\overline{\text{RMSVE}}(\theta^*_{\text{On}}) \approx 1.155$ (the dash lines). Since the transient bias has not completely faded away within the given time steps, the error is still slightly larger than the theoretical optimal error of 1.155. Nevertheless, this is the best performance among existing algorithms and a significant improvement over the only existing consistent algorithm, Full IS TD, which does not learn due to the high variance issue. On the other hand, ETD (ETD(β) with $\beta = \gamma = 0.8$) is the second-tier algorithm in this task, achieving its theoretical optimal error of 1.251. For Off-policy TD, it also converges to its fixed point, which induces a significantly larger error of $\overline{\text{RMSVE}}(\theta^*_{\text{Off}}) \approx 1.523$.

Figure 2(c)-2(f) plot the learning curves of tunable algorithms with fixed values of the decay parameter. From Figure 2(c), it is evident that ETD(β) converges to solutions with large biases for most values of β. For $\beta = 1$, its error suddenly explodes after some steps, demonstrating the unsoundness of ETD(β) with $\beta = 1$ in the continuing case. For



(a) Best learning curves (b) Sensitivity to β or ν (c) Best learning curves of ETD(β)

(d) Best learning curves of CETD1(β) (e) Best learning curves of CETD2(ν) (f) Best learning curves of CETD3(β)

Figure 4: Performance of different algorithms on the Rooms task. The y-axis shows $\overline{\text{RMSVE}}$.

CETD instances (see Figure 2(d)-2(f)), with the decay parameter in interval $[0.2, 0.8]$, they all converge smoothly to the lowest error in accordance with the theory that CETD(β, ν) are consistent for any $\beta \in [0, 1)$ and $c \in (0, 1]$.

Figure 2(b) summarizes these results. We can conclude that all CETD instances consistently enjoy low error compared to existing algorithms, which implies CETD(β, ν)'s consistency across all its hyperparameter choices.

PRACTICALITY OF CETD(β, ν)

Rooms Task To further test the performance of CETD(β, ν) in more complex tasks with higher variances, we extended the episodic Rooms task proposed by Ghiassian & Sutton (2021) to the continuing setting. The task specification largely follows from Ghiassian & Sutton (2021), and the modifications can be found in the appendix. Compared to the Two-state task, the Rooms task has more states and complex feature representation. Moreover, the differences between the target policies and the behavior policy are larger, inducing a much larger variance. We run each algorithm for $150,000$ steps and 30 runs. To provide a stable result and better illustrate the advantage of CETD(β, ν), we present Interquartile Mean (IQM) results in Figure 4, which are more robust and statistically efficient compared to the mean or median results (Agarwal et al., 2021). Additionally, we present the mean results of all runs in the appendix to provide a comprehensive performance profile.

From Figure 4(a), we can see that ETD, ETD(β), and all CETD instances achieve similar final errors. Among them, CETD2(ν) and CETD3(β) learn the fastest. Same as in the Two-state task, Off-policy TD converges very fast to a solution with a significant bias, while Full IS TD does not learn at all.

Figure 4(c)-4(f) plot the learning curves of tunable algorithms with fixed values of the decay parameter. From Figure 4(c), we can see that as the value of β increases, the bias of the solution ETD(β) found becomes smaller, and the learning also becomes slower. For CETD1(β) and CETD2(ν) (see Figure 4(d) and 4(e)), they learn faster with larger values of the decay parameter. On the other hand, CETD3(β) is not very sensitive to the value of β (see Figure 4(f)).

Figure 4(b) summarizes the above results. We can see that even in the high variance setting, CETD instances are still better: They converge faster to the lowest error and are less sensitive to the decay parameter compared to ETD(β).

THE BIAS-VARIANCE TRADE-OFF

Having presented the advantages of the CETD instances in a didactic example and a more complex task with high variance, we now focus on analyzing the bias-variance trade-off that β and ν control to provide more insights on the properties of CETD(β, ν). Specifically, we look at the bias and variance of the general followon trace F_t in Eq. (7) for different algorithms. Ideally, the random variable F_t given

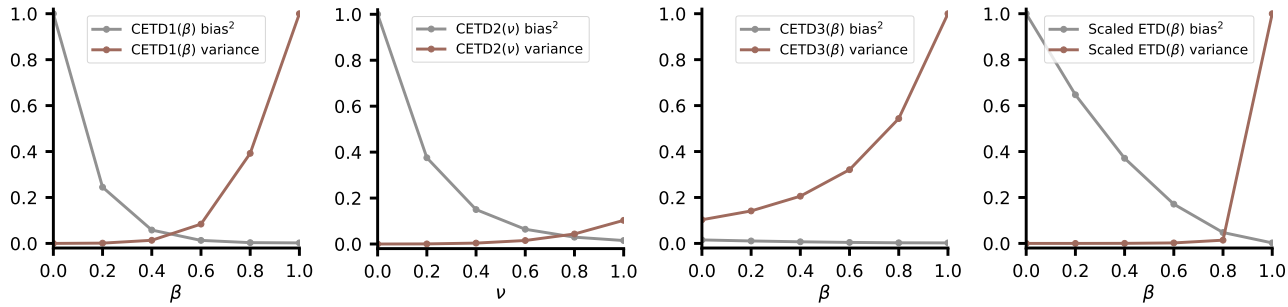


Figure 5: The bias-variance trade-off of different algorithms. The y-axis shows the normalized bias and variance of F_t.

$S_t = s$ should have an expectation of $\frac{\mathbb{P}_\pi(S_t=s)}{\mathbb{P}_\mu(S_t=s)}$, which converges to the density ratio in the limit and corrects the distribution of the updates back to the on-policy distribution. Full IS TD is such an algorithm with zero bias. However, the variance of Full IS TD's F_t is unbearably high. For the CETD instances, they have a non-zero bias within finite time steps, but the bias converges to zero asymptotically. The benefit of having some transient bias is that the CETD instances could enjoy a lower variance. On the other hand, Scaled ETD(β) has a much lower variance but a persistent bias. We performed experiments on the Two-state task. In order to obtain an accurate estimation, we sample $100,000$ trajectories of length 30 to calculate the bias and variance of F_t. The results are presented in Figure 5. In addition, we provide more details and discussions on the bias-variance analysis in the appendix.

From Figure 5, we can see that, for all algorithms, as the decay parameter (β or ν) becomes larger, the bias would decrease, and the variance would increase. However, the magnitude and the speed of the changes differ. For CETD1(β), the bias and variance curves are quite symmetric. On the leftmost point with $\beta = 0$, CETD1(β) becomes Off-policy TD which has the lowest variance but the largest bias; On the right-most point with $\beta = 1$, CETD1(β) becomes Full IS TD which has the lowest bias but the largest variance; The points in the middle achieve a good bias-variance trade-off. Note that Scaled ETD(β) also connects Off-policy TD and Full IS TD. However, it is not a consistent method and could only become less biased as β approaches 1. The result for Scaled ETD(β) also indicates the bias would only decrease to a relatively low level when β is quite large (in this case, at least 0.8), not to mention that its bias is persistent. In contrast the bias of the CETD instances would fade away as more time steps are given. Besides, CETD2(ν) and CETD3(β) together form a polygonal line connecting Off-policy TD and Full IS TD. The bias and the variance curves of the two algorithms combined form a similar shape to that of CETD1(β) but much wider. As a result, these two algorithms are less sensitive to the decay parameter but also put themselves at risk of not achieving the best trade-off. Generally, CETD2(ν) would hold the best trade-off point in tasks with high variance.

5 CONCLUSIONS AND DISCUSSION

In this paper, we have introduced a new consistent off-policy algorithm. To attain a smooth bias-variance trade-off, we unify the new algorithm with some existing algorithms (Precup, 2000; Precup et al., 2001; Sutton et al., 2016; Hallak et al., 2016). The resulting unified algorithm contains a new family of consistent algorithms, Consistent Emphatic TD (CETD(λ, β, ν)), which has several desired theoretical and empirical properties. First, different from ETD(λ, β), CETD(λ, β, ν) is guaranteed to be stable regardless of the values of its hyperparameters. Second, while ETD(λ, β) has a biased fixed point, CETD(λ, β, ν) has the same fixed point as On-policy TD(λ). Third, the bias-variance trade-off that its hyperparameters control makes CETD(λ, β, ν) practical, providing an effective remedy to Full IS TD(λ), the only consistent method previously. To the best of our knowledge, CETD(λ, β, ν) is the *first practical consistent* algorithm for off-policy TD learning with linear function approximation. By constraining the value of CETD(λ, β, ν)'s hyperparameters, we obtain its three instances with the same number of hyperparameters as ETD(λ, β). Experiment results on a didactic example and a more complex task with high variance show a competitive performance of the instances, validating the consistency and practicality of CETD(λ, β, ν).

Despite having the ability to control a smooth bias-variance trade-off, CETD(λ, β, ν) still suffers from a high variance issue to some degree (see the appendix). This issue roots in all importance-sampling-based methods including ETD(λ, β) and Full IS TD(λ). Potential treatments include periodically restarting or truncating the followon trace (Guan et al., 2021; Zhang & Whiteson, 2022) and learning an expected followon trace (Zhang et al., 2020c; Jiang et al., 2022). Investigating these and new approaches to further reduce the variance is one direction to be explored. Another unanswered question is the convergence of CETD(λ, β, ν). Similar to ETD(λ) (Sutton et al., 2016), we have provided the stability guarantee of CETD(λ, β, ν), which is an important necessary condition of its convergence. Similar to proving the convergence of ETD(λ) (Yu, 2016), significant technical challenges may present in proving the convergence of CETD(λ, β, ν). Thus, we leave it for future work.

References

Agarwal, R., Schwarzer, M., Castro, P. S., Courville, A. C., and Bellemare, M. Deep reinforcement learning at the edge of the statistical precipice. *Advances in neural information processing systems*, 34:29304–29320, 2021.

Baird, L. Residual algorithms: Reinforcement learning with function approximation. In *Machine Learning Proceedings 1995*, pp. 30–37. Elsevier, 1995.

Ghiassian, S. and Sutton, R. S. An empirical comparison of off-policy prediction learning algorithms in the four rooms environment. *arXiv preprint arXiv:2109.05110*, 2021.

Guan, Z., Xu, T., and Liang, Y. PER-ETD: A polynomially efficient emphatic temporal difference learning method. *arXiv preprint arXiv:2110.06906*, 2021.

Hallak, A. and Mannor, S. Consistent on-line off-policy evaluation. In *International Conference on Machine Learning*, pp. 1372–1383. PMLR, 2017.

Hallak, A., Tamar, A., Munos, R., and Mannor, S. Generalized emphatic temporal difference learning: Bias-variance analysis. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

Jiang, R., Zhang, S., Chelu, V., White, A., and van Hasselt, H. Learning expected emphatic traces for deep RL. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.

Kolter, J. The fixed points of off-policy TD. *Advances in Neural Information Processing Systems*, 24, 2011.

Liu, Q., Li, L., Tang, Z., and Zhou, D. Breaking the curse of horizon: Infinite-horizon off-policy estimation. *Advances in Neural Information Processing Systems*, 31, 2018.

Nachum, O., Chow, Y., Dai, B., and Li, L. Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections. *Advances in Neural Information Processing Systems*, 32, 2019.

Precup, D. Eligibility traces for off-policy policy evaluation. *Computer Science Department Faculty Publication Series*, pp. 80, 2000.

Precup, D., Sutton, R. S., and Dasgupta, S. Off-policy temporal-difference learning with function approximation. In *ICML*, pp. 417–424, 2001.

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

Sutton, R. S. and Barto, A. G. *Reinforcement learning: An introduction*. MIT press, 2018.

Sutton, R. S., Precup, D., and Singh, S. Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence*, 112(1-2): 181–211, 1999.

Sutton, R. S., Mahmood, A. R., and White, M. An emphatic approach to the problem of off-policy temporal-difference learning. *The Journal of Machine Learning Research*, 17 (1):2603–2631, 2016.

Sutton, R. S., Bowling, M. H., and Pilarski, P. M. The Alberta Plan for AI research. *arXiv preprint arXiv:2208.11173*, 2022.

Tsitsiklis, J. and Van Roy, B. Analysis of temporal-diffference learning with function approximation. *Advances in neural information processing systems*, 9, 1996.

White, A., Modayil, J., and Sutton, R. S. Scaling life-long off-policy learning. In *2012 IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL)*, pp. 1–6. IEEE, 2012.

Yu, H. Weak convergence properties of constrained emphatic temporal-difference learning with constant and slowly diminishing stepsize. *The Journal of Machine Learning Research*, 17(1):7745–7802, 2016.

Zhang, R., Dai, B., Li, L., and Schuurmans, D. GenDICE: Generalized offline estimation of stationary values. *arXiv preprint arXiv:2002.09072*, 2020a.

Zhang, S. and Whiteson, S. Truncated emphatic temporal difference methods for prediction and control. *Journal of Machine Learning Research*, 23(153):1–59, 2022.

Zhang, S., Liu, B., and Whiteson, S. GradientDICE: Rethinking generalized offline estimation of stationary values. In *International Conference on Machine Learning*, pp. 11194–11203. PMLR, 2020b.

Zhang, S., Liu, B., Yao, H., and Whiteson, S. Provably convergent two-timescale off-policy actor-critic with function approximation. In *International Conference on Machine Learning*, pp. 11204–11213. PMLR, 2020c.