The Emphatic Approach to Average-Reward Policy Evaluation

Jiamin He
University of Alberta
jiamin12@ualberta.ca

Yi Wan
University of Alberta
wan6@ualberta.ca

A. Rupam Mahmood
University of Alberta
armahmood@ualberta.ca

Abstract

Off-policy policy evaluation has been a longstanding problem in reinforcement learning. This paper looks at this problem under the average-reward formulation with function approximation. Differential temporal-difference (TD) learning has been proposed recently and has shown great potential compared to previous average-reward learning algorithms. In the tabular setting, off-policy differential TD is guaranteed to converge. However, the convergence guarantee cannot be carried through the function approximation setting. To address the instability of off-policy differential TD, we investigate the emphatic approach proposed for the discounted formulation. Specifically, we introduce average emphatic trace for average-reward off-policy learning. We further show that without any variance reduction techniques, the new trace suffers from slow learning due to high variance of importance sampling ratios. Finally, we show that differential emphatic TD(β), extended from the discounted setting, can save us from the high variance while introducing bias. Experimental results on a counterexample show that differential emphatic TD(β) performs better than an existing competitive off-policy algorithm.

1 Introduction

Off-policy learning is an important topic in reinforcement learning. Particularly, off-policy predictions are an essential component in model learning, options learning (Sutton, Precup, & Singh, 1999), and life-long learning (Sutton, Bowling, & Pilarski, 2022; White, Modayil, & Sutton, 2012). It has been extensively investigated in the discounted setting. Here, the goal is to estimate the value function of a *target policy* with off-policy data collected by a different *behavior policy*. In the online setting, Temporal-Difference (TD) learning is the most celebrated approach. It is known that off-policy TD(λ), which corrects the probability of action selection with *importance sampling ratios*, has guaranteed convergence in the tabular setting. However, moving on to the function approximation setting, TD(λ) is proven to diverge in some cases, even with linear function approximation. This well-known issue is called the *deadly traid*, which characterizes the instability of TD algorithms with function approximation and off-policy learning.

Existing efforts in addressing the deadly triad can be roughly divided into value-based and density-ratio-based methods. Unlike value-based methods which don't maintain an estimation of the density-ratio, most of density-ratio-based methods (Liu et al., 2018; R. Zhang et al., 2020; S. Zhang, Liu, & Whiteson, 2020; Mousavi et al., 2020) apply to both the discounted setting and the average-reward setting. However, it is shown that a value-based method consistently performs better than a competitive density-ratio-based method in the average-reward setting (S. Zhang, Wan, et al., 2021). In this paper, we will mainly focus on value-based methods. Existing value-based methods for the discounted setting can be categorized into three classes. Gradient TD (GTD) and emphatic TD (ETD) algorithms are two classes that are designed with a convergence guarantee. Another group of algorithms aims for faster learning instead of guaranteed convergence. Examples are Vtrace(λ) (Espeholt et al., 2018) and ABTD(ζ) (Mahmood, Yu, & Sutton, 2017).

Deep Reinforcement Learning Workshop at NeurIPS 2022.

Compared to the discounted setting, the average-reward setting is more challenging and has been less studied. The goal of average-reward policy evaluation is to estimate the *reward rate* and the *differential value function*. In this setting, there are two important variants of TD learning algorithms: average-cost TD (Tsitsiklis & Van Roy, 1999) and differential TD (Wan, Naik, & Sutton, 2021). It is previously known that off-policy differential TD will converge in the tabular setting (Wan et al., 2021), but the convergence guarantee doesn't carry over to the function approximation setting (S. Zhang, Wan, et al., 2021). For off-policy average-cost TD, it is not clear yet whether it can diverge or not in the tabular setting, but this paper will show its divergence in the function approximation setting. In addressing the instability, to the best of our knowledge, gradient TD is the first and only method extended to the average-reward setting among the three classes of off-policy remedies mentioned above for the discounted setting. In this paper, we will take the next step to investigate the emphatic approach in the average-reward setting.

Our contributions are as follows: Firstly, we illustrate the advantage of differential TD algorithms over average-cost TD algorithms in the problem of average-reward off-policy policy evaluation with function approximation. Secondly, we complete the spectrum of the emphatic algorithms for the continuing setting with the proposed average emphatic trace. However, we show that this theoretically sound trace suffers from high variance. Finally, to avoid the high variance of the new trace, we extend generalized ETD (ETD(β); Hallak et al., 2016) to the average-reward setting. We show that differential ETD(β) achieves the best performance on a counterexample in terms of the asymptotic error and step-size sensitivity compared to existing value-based algorithms, including a competitive off-policy algorithm. We also include some rudimentary results of the nonlinear variants of these algorithms on MuJoCo tasks.

2 Background

Problem formulation

We consider an infinite horizon Markov Decision Process (MDP), which is defined as a tuple $\langle \mathcal{S}, \mathcal{A}, P, r \rangle$ where $\mathcal{S}$ is the finite state space, $\mathcal{A}$ is the finite action space, $P(s'|s,a)$ is the transition function, and r is the deterministic reward function. The policy of the agent is defined as $\pi(a|s) : \mathcal{S} \times \mathcal{A} \to [0,1]$. We denote the size of the state space and the action space by $S = |\mathcal{S}|$ and $A = |\mathcal{A}|$. Assuming the target policy π induces a unichain, we consider the average-reward formulation where the agent's performance is evaluated with the *reward rate* defined as

$$r(\pi) \doteq \lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} \mathbb{E}[R_t | S_0, A_{0:t-1} \sim \pi],$$

where S_0 can be sampled from an arbitrary state distribution under our assumption. It is also useful to define a *differential value function* over states $v_\pi(s) : \mathcal{S} \to \mathbb{R}$:

$$v_\pi(s) \doteq \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \sum_{t=1}^{k} \mathbb{E}[R_t - r(\pi) | S_0 = s, A_{0:t-1} \sim \pi].$$

Particularly, we consider the setting of *online off-policy policy evaluation with linear function approximation* where the agent needs to estimate both the reward rate and the differential value function of a target policy π while interacting with the environment with a behavior policy μ. 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 an action based on the behavior policy μ and observes the next state feature ϕ_{t+1} and reward R_{t+1}. With linear function approximation, the agent approximates the reward rate with parameter $\bar{R}$ and the differential value function with $\hat{v}(s) = \boldsymbol{\phi}(s)^\top \boldsymbol{\theta}$ 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. This assumption ensures the unique existence of the stationary distribution of the behavior policy $\boldsymbol{d}_\mu$ and that of the target policy $\boldsymbol{d}_\pi$. Moreover, we define $\boldsymbol{D}_{\boldsymbol{v}} \doteq diag(\boldsymbol{v})$ 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 $\boldsymbol{D}_{\boldsymbol{v}}$ for some vector $\boldsymbol{v}$, i.e., $\|\boldsymbol{x}\|_{\boldsymbol{v}} = \sqrt{\boldsymbol{x}^\top \boldsymbol{D}_{\boldsymbol{v}} \boldsymbol{x}}$. We also define $\boldsymbol{e} = [1, 1, \cdots, 1]^\top \in \mathbb{R}^S$.

Off-policy average-cost TD and off-policy differential TD

In this section, we present two instances from the two major temporal-difference (TD) learning methods for average-reward reinforcement learning, average-cost TD (Tsitsiklis & Van Roy, 1999) and differential TD (Wan et al., 2021). For simplicity, we consider their off-policy one-step versions: off-policy average-cost TD(0) and off-policy differential TD(0). Most of the arguments should be easily extendable to their multi-step versions. For the rest of the paper, we will drop the multi-step notation, referring to them as off-policy average-cost TD and off-policy differential TD. The same goes for other algorithms in the remaining of the paper. The off-policy average-cost TD makes the following update:

$$\begin{aligned} \boldsymbol{\theta}_{t+1} &= \boldsymbol{\theta}_t + \alpha\rho_t\delta_t\boldsymbol{\phi}_t, \\ \bar{R}_{t+1} &= \bar{R}_t + \alpha\rho_t\eta(R_{t+1} - \bar{R}_t), \\ \delta_t &= R_{t+1} - \bar{R}_t + \boldsymbol{\phi}_{t+1}^\top\boldsymbol{\theta}_t - \boldsymbol{\phi}_t^\top\boldsymbol{\theta}_t, \end{aligned} \tag{1}$$

where $\rho_t \doteq \frac{\pi(A_t|S_t)}{\mu(A_t|S_t)}$ is the importance sampling ratio, $\alpha > 0$ is the primary step-size parameter, and $\eta > 0$ is a second step-size parameter to control how fast the reward rate is updated compared to the differential value function. Let $\boldsymbol{u}_t = [\bar{R}_t, \boldsymbol{\theta}_t^\top]^\top$, then we can rewrite Update (1) as follows:

$$\boldsymbol{u}_{t+1} = \boldsymbol{u}_t + \alpha\rho_t\left(\underbrace{\begin{bmatrix} \eta R_{t+1} \\ \boldsymbol{\phi}_t R_{t+1} \end{bmatrix}}_{\boldsymbol{b}_t} - \underbrace{\begin{bmatrix} \eta & \mathbf{0}^\top \\ \boldsymbol{\phi}_t & \boldsymbol{\phi}_t(\boldsymbol{\phi}_t - \boldsymbol{\phi}_{t+1})^\top \end{bmatrix}}_{\boldsymbol{A}_t}\boldsymbol{u}_t\right). \tag{2}$$

Similar to previous work (Sutton, Mahmood, & White, 2016), we only analyze the stability of the update. If the update is stable, with similar techniques from earlier work by Precup, Sutton, and Dasgupta (2001), we can prove that the update converges with probability one and bounded error. To analyze the stability of the stochastic Update (2), we will look at the corresponding expected update $\bar{\boldsymbol{u}}_{t+1} = \bar{\boldsymbol{u}}_t + \alpha(\boldsymbol{b}^{\text{ac}} - \boldsymbol{A}^{\text{ac}}\bar{\boldsymbol{u}}_t)$ where $\boldsymbol{A}^{\text{ac}}$ is defined as follows (Tsitsiklis & Van Roy, 1999):

$$\boldsymbol{A}^{\text{ac}} = \lim_{t\to\infty}\mathbb{E}_\mu[\rho_t\boldsymbol{A}_t] = \begin{bmatrix} \eta & \mathbf{0}^\top \\ \boldsymbol{\Phi}^\top\boldsymbol{d}_\mu & \boldsymbol{\Phi}^\top\boldsymbol{D}_\mu(\boldsymbol{I} - \boldsymbol{P}_\pi)\boldsymbol{\Phi} \end{bmatrix}.$$

On the other hand, off-policy differential TD performs the below update:

$$\begin{aligned} \boldsymbol{\theta}_{t+1} &= \boldsymbol{\theta}_t + \alpha\rho_t\delta_t\boldsymbol{\phi}_t, \\ \bar{R}_{t+1} &= \bar{R}_t + \alpha\rho_t\eta\delta_t, \\ \delta_t &= R_{t+1} - \bar{R}_t + \boldsymbol{\phi}_{t+1}^\top\boldsymbol{\theta}_t - \boldsymbol{\phi}_t^\top\boldsymbol{\theta}_t. \end{aligned} \tag{3}$$

Follow the same definitions and arguments for off-policy average-cost TD, we have the corresponding expected update $\bar{\boldsymbol{u}}_{t+1} = \bar{\boldsymbol{u}}_t + \alpha(\boldsymbol{b}^{\text{diff}} - \boldsymbol{A}^{\text{diff}}\bar{\boldsymbol{u}}_t)$ for off-policy differential TD where $\boldsymbol{A}^{\text{diff}}$ can be expressed as follows (S. Zhang, Wan, et al., 2021):

$$\boldsymbol{A}^{\text{diff}} = \begin{bmatrix} \eta & \boldsymbol{d}_\mu^\top(\boldsymbol{I} - \boldsymbol{P}_\pi)\boldsymbol{\Phi} \\ \boldsymbol{\Phi}^\top\boldsymbol{d}_\mu & \boldsymbol{\Phi}^\top\boldsymbol{D}_\mu(\boldsymbol{I} - \boldsymbol{P}_\pi)\boldsymbol{\Phi} \end{bmatrix}.$$

3 Instability of average-reward TD algorithms

Following Sutton et al. (2016), we analyze the $\boldsymbol{A}$ matrix in the expected update $\bar{\boldsymbol{u}}_{t+1} = \bar{\boldsymbol{u}}_t + \alpha(\boldsymbol{b} - \boldsymbol{A}\bar{\boldsymbol{u}}_t)$ of different algorithms. We call the matrix $\boldsymbol{D}_\mu(\boldsymbol{I} - \boldsymbol{P}_\pi)$ the *big key matrix* and the matrix $\boldsymbol{\Phi}^\top\boldsymbol{D}_\mu(\boldsymbol{I} - \boldsymbol{P}_\pi)\boldsymbol{\Phi}$ the *key matrix*. If the eigenvalues of an $\boldsymbol{A}$ matrix all have positive real parts, we say the $\boldsymbol{A}$ matrix is stable. If the $\boldsymbol{A}$ matrix of the expected update of an algorithm is stable, we say the algorithm is stable. The stability of an algorithm is the necessary condition for the convergence of the algorithm (Sutton et al., 2016). In average-reward on-policy setting, the big key matrix is positive semi-definite. To ensure the stability of the $\boldsymbol{A}$ matrix, it is usually assumed that $\boldsymbol{e}$ is not in the column space of $\boldsymbol{\Phi}$ (Tsitsiklis & Van Roy, 1999; S. Zhang, Wan, et al., 2021), which guarantees the positive definiteness of the key matrix. Further, since $\mu = \pi$ and $\boldsymbol{d}_\mu^\top(\boldsymbol{I} - \boldsymbol{P}_\pi)\boldsymbol{\Phi} = \boldsymbol{d}_\pi^\top(\boldsymbol{I} - \boldsymbol{P}_\pi)\boldsymbol{\Phi} = 0$, the $\boldsymbol{A}$ matrices for on-policy average-cost TD and on-policy differential TD are identical and have the

same eigenvalues all of whose real parts are positive under such an assumption. Thus, the on-policy stability for both algorithms is ensured.

In the off-policy case, the big key matrix may not be positive semi-definite. Thus, the key matrix may not be positive definite. In fact, we lose the stability guarantee of both algorithms in the off-policy setting. Next, we will present two corresponding counterexamples of off-policy average-cost TD and off-policy differential TD.

Instability of off-policy average-cost TD

We first look at off-policy average-cost TD. The two-state MDP with $c = 1$ in Figure 1 is a counterexample in which off-policy average-cost TD is unstable. In this example, the stationary distributions for both target policy and behavior policy are very straightforward: $\boldsymbol{d}_\pi = [0.4, 0.6]^\top$ and $\boldsymbol{d}_\mu = [0.6, 0.4]^\top$. The transition probability matrix induced by the target policy is $\boldsymbol{P}_\pi = \begin{bmatrix} 0.4 & 0.6 \\ 0.4 & 0.6 \end{bmatrix}$. The big key matrix is

$$\boldsymbol{D}_\mu(\boldsymbol{I} - \boldsymbol{P}_\pi) = \begin{bmatrix} 0.6 & 0 \\ 0 & 0.4 \end{bmatrix} \times \begin{bmatrix} 0.6 & -0.6 \\ -0.4 & 0.4 \end{bmatrix} = \begin{bmatrix} 0.36 & -0.36 \\ -0.16 & 0.16 \end{bmatrix}.$$

Further, the key matrix of this counterexample is

$$\begin{aligned} \boldsymbol{\Phi}^\top \boldsymbol{D}_\mu(\boldsymbol{I} - \boldsymbol{P}_\pi)\boldsymbol{\Phi} &= \begin{bmatrix} 1 & 2 \end{bmatrix} \times \begin{bmatrix} 0.36 & -0.36 \\ -0.16 & 0.16 \end{bmatrix} \times \begin{bmatrix} 1 \\ 2 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 2 \end{bmatrix} \times \begin{bmatrix} -0.36 \\ 0.16 \end{bmatrix} = -0.04. \end{aligned}$$

Now, the eigenvalues of $\boldsymbol{A}^{\text{ac}}$ consist of η and the eigenvalues of the key matrix, in this case, -0.04. Then, by definition, off-policy average-cost TD is unstable. However, off-policy differential TD is stable in this counterexample with a wide range of η. For example, when $\eta = 1$, we have $\boldsymbol{\Phi}^\top \boldsymbol{d}_\mu = 1.4$, $\boldsymbol{d}_\mu^\top(\boldsymbol{I} - \boldsymbol{P}_\pi)\boldsymbol{\Phi} = -0.2$, and $\boldsymbol{A}^{\text{diff}} = \begin{bmatrix} 1 & -0.2 \\ 1.4 & -0.04 \end{bmatrix}$. The eigenvalues of $\boldsymbol{A}^{\text{diff}}$ are $\frac{12}{25} + i\frac{\sqrt{6}}{25}$ and $\frac{12}{25} - i\frac{\sqrt{6}}{25}$ which have positive real parts. Thus, off-policy differential TD is stable in this counterexample.

Instability of off-policy differential TD

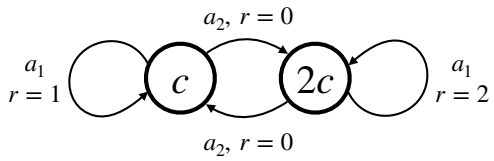
To settle the instability of off-policy differential TD, we need to modify the counterexample for off-policy average-cost TD. Notice that for off-policy average-cost TD, changing the value of c or η would only change the magnitude but not the sign of the eigenvalues of the $\boldsymbol{A}^{\text{ac}}$ matrix. Thus, average-cost TD will always be unstable regardless of any values of c and η. However, this is not the case for off-policy differential TD. When we change the value of c or η, the value and the sign of the eigenvalues of the $\boldsymbol{A}^{\text{diff}}$ matrix will change. Setting out to choose a c such that differential TD is unstable regardless of the value of η, we found that our purpose is fulfilled by choosing $c = \sqrt{175}$. In this case, the $\boldsymbol{A}^{\text{diff}}$ matrix is

$$\boldsymbol{A}^{\text{diff}} = \begin{bmatrix} \eta & \boldsymbol{d}_\mu^\top(\boldsymbol{I} - \boldsymbol{P}_\pi)\boldsymbol{\Phi} \\ \boldsymbol{\Phi}^\top \boldsymbol{d}_\mu & \boldsymbol{\Phi}^\top \boldsymbol{D}_\mu(\boldsymbol{I} - \boldsymbol{P}_\pi)\boldsymbol{\Phi} \end{bmatrix} = \begin{bmatrix} \eta & -0.2\sqrt{175} \\ 1.4\sqrt{175} & -7 \end{bmatrix},$$

whose trace is $\eta - 7$ and whose determinant is $7(7 - \eta)$. Thus, the eigenvalues of the $\boldsymbol{A}^{\text{diff}}$ matrix, λ_1 and λ_2, satisfy $\lambda_1 + \lambda_2 = \eta - 7$ and $\lambda_1\lambda_2 = 7(7 - \eta)$. When $\eta \neq 7$, at least one of the eigenvalues would be negative. When $\eta = 7$, both eigenvalues are zero. Thus, off-policy differential TD is always unstable in this counterexample.

Comparing average-cost TD and differential TD

It is known that off-policy differential TD is guaranteed to converge while off-policy average-cost TD has no guarantee at present in the tabular setting (Wan et al., 2021). Here in the function approximation setting, we have shown that both off-policy average-cost TD and off-policy differential



$$\pi(a_1 \mid c) = \pi(a_2 \mid 2c) = 0.4$$
$$\mu(a_1 \mid c) = \mu(a_2 \mid 2c) = 0.6$$

Figure 1: A two-state MDP with undetermined features. When $c = 1$, off-policy average-cost TD is unstable regardless of the choice of η while off-policy differential TD is stable with a large set of η. When $c = \sqrt{175}$, both algorithms are unstable.

TD can be unstable. Also, the degrees of their instability differ: For the MDP in Figure 1, off-policy average-cost TD is always unstable regardless of the choices of c and η, while off-policy differential TD can be stable for a large set of c and η. This demonstrates a clear advantage of off-policy differential TD over off-policy average-cost TD when the key matrix has at least one eigenvalue whose real part is negative.

For the case where the key matrix has a full set of eigenvalues whose real parts are positive, off-policy average-cost TD is guaranteed to be stable. On the other hand, we don't have a general result for off-policy differential TD yet. We conjecture that off-policy differential TD will also be stable. To shed some light on this question, we prove that when the key matrix is positive definite, off-policy differential TD is stable. The result is presented with the following proposition, whose proof can be found in the Appendix:

Proposition 1. *If* $\Delta \doteq \min_{\|\boldsymbol{\theta}\|_2=1} \boldsymbol{\theta}^\top \boldsymbol{\Phi}^\top \boldsymbol{D}_\mu (\boldsymbol{I} - \boldsymbol{P}_\pi)\boldsymbol{\Phi}\boldsymbol{\theta} > 0$*, then with sufficiently large* η*, matrix* $\boldsymbol{A}^{diff}$ *is positive definite.*

Having settled the potential benefit of differential TD in the off-policy setting, we will use it as the base average-reward algorithm in the remaining of the paper. Next, we will discuss how we can alleviate the instability issue by reweighting updates.

4 Average-reward emphatic TD learning

Addressing the instability of TD by reweighting updates

In the discounted reinforcement learning literature, there are two distinct approaches that try to address the instability by reweighting the updates using importance sampling ratios. The first approach (Precup et al., 2001) uses a full importance sampling ratios product to completely correct the state distribution to the on-policy distribution, which can be directly adopted to the average-reward setting. Adopting it to off-policy differential TD, the update with full importance-sampling (IS) correction is as follows:

$$\begin{aligned} \boldsymbol{\theta}_{t+1} &= \boldsymbol{\theta}_t + \alpha \rho_t H_t \delta_t \boldsymbol{\phi}_t, \\ \bar{R}_{t+1} &= \bar{R}_t + \alpha \rho_t H_t \eta \delta_t, \\ \delta_t &= R_{t+1} - \bar{R}_t + \boldsymbol{\phi}_{t+1}^\top \boldsymbol{\theta}_t - \boldsymbol{\phi}_t^\top \boldsymbol{\theta}_t, \\ H_{t+1} &= \rho_t H_t, H_0 = 1. \end{aligned} \tag{4}$$

We refer to temporal difference algorithms that use H_t for reweighting as *full IS TD* and the above algorithm instance as differential full IS TD. By fully correcting the state distribution, the expected update returns to the on-policy expected update where both average-cost TD and differential TD converge. Reweighting the updates with the contemplation of how likely it will arrive at the current state started from the very first time step by following the target policy instead of the behavior policy, full IS TD is often thought of as the *alternative life* view. However, this approach only works in theory but never becomes a practical algorithm due to the high-variance nature of the importance sampling ratios product.

Another relatively more practical approach is *Emphatic TD* (ETD; Sutton et al., 2016), which adopts the *excursion* view instead of the alternative life view. In the excursion view, the updates are

reweighted towards a state distribution between the stationary distribution of the behavior policy $\boldsymbol{d}_\mu$ and that of the target policy $\boldsymbol{d}_\pi$. At every time step, a new excursion is started from the current state. The excursion will last many steps until termination and be reweighted by importance sampling ratios. With a fixed discount factor as soft termination, we can think that the excursions would never terminate, but old excursions would be decayed every time step. Correcting the state distribution of updates closer to $\boldsymbol{d}_\pi$, the emphatic approach has proven to be effective in several environments with linear function approximation (Ghiassian & Sutton, 2021a, 2021b) and been successfully extended to deep reinforcement learning (Jiang et al., 2021, 2022). However, moving on the discussion to the average-reward setting, ETD from the excursion view is undefined since there is no discounting or termination here, and the expectation of the original followon trace without decaying would go to infinity. In the following section, we will introduce how to define the corresponding emphatic trace for the average-reward setting from the excursion view.

Average emphatic TD

This section introduces a novel way of correcting the state distribution. Instead of performing an unnormalized weighted average over different excursions started from previous states like ETD, we use a normalized uniform average over excursions:

$$\begin{aligned}
\boldsymbol{\theta}_{t+1} &= \boldsymbol{\theta}_t + \alpha\rho_t F_t \delta_t \boldsymbol{\phi}_t, \\
\bar{R}_{t+1} &= \bar{R}_t + \alpha\rho_t F_t \eta \delta_t, \\
\delta_t &= R_{t+1} - \bar{R}_t + \boldsymbol{\phi}_{t+1}^\top \boldsymbol{\theta}_t - \boldsymbol{\phi}_t^\top \boldsymbol{\theta}_t, \\
F_{t+1} &= \frac{t-1}{t}\rho_t F_t + \frac{1}{t}, F_0 = 1.
\end{aligned} \tag{5}$$

We refer to temporal difference algorithms that use this F_t for reweighting as *average emphatic TD* (AETD) and the above algorithm instance as differential AETD. Next, we will analyze the property of the expected update of differential AETD.

Following the derivation of ETD by Sutton et al. (2016) and the above procedure for off-policy average-cost TD, the $\boldsymbol{A}^{\text{ae}}$ matrix of expected update of differential AETD, $\bar{\boldsymbol{u}}_{t+1} = \bar{\boldsymbol{u}}_t + \alpha(\boldsymbol{b}^{\text{ae}} - \boldsymbol{A}^{\text{ae}}\bar{\boldsymbol{u}}_t)$, is

$$\boldsymbol{A}^{\text{ae}} = \begin{bmatrix} \eta & \boldsymbol{f}^\top(\boldsymbol{I} - \boldsymbol{P}_\pi)\boldsymbol{\Phi} \\ \boldsymbol{\Phi}^\top \boldsymbol{f} & \boldsymbol{\Phi}^\top \boldsymbol{D}_f (\boldsymbol{I} - \boldsymbol{P}_\pi)\boldsymbol{\Phi} \end{bmatrix},$$

where $f(s) = d_\mu(s) \lim_{t\to\infty} \mathbb{E}_\mu[F_t | S_t = s]$. Plug in F_t defined in Update (5), we have

$$\begin{aligned}
f(s) &= d_\mu(s) \lim_{t\to\infty} \mathbb{E}_\mu \left[\frac{t-1}{t}\rho_{t-1}F_{t-1} + \frac{1}{t} | S_t = s \right] \\
&= d_\mu(s) \lim_{t\to\infty} \mathbb{E}_\mu[\rho_{t-1}F_{t-1} | S_t = s] \\
&= d_\mu(s) \lim_{t\to\infty} \sum_{\bar{s},\bar{a}} \mathbb{P}(S_{t-1} = \bar{s}, A_{t-1} = \bar{a} | S_t = s) \frac{\pi(\bar{a}|\bar{s})}{\mu(\bar{a}|\bar{s})} \mathbb{E}_\mu[F_{t-1} | S_{t-1} = \bar{s}] \\
&= d_\mu(s) \lim_{t\to\infty} \sum_{\bar{s},\bar{a}} \frac{\mathbb{P}(S_{t-1} = \bar{s}, A_{t-1} = \bar{a}, S_t = s)}{\mathbb{P}(S_t = s)} \frac{\pi(\bar{a}|\bar{s})}{\mu(\bar{a}|\bar{s})} \mathbb{E}_\mu[F_{t-1} | S_{t-1} = \bar{s}] \\
&= d_\mu(s) \sum_{\bar{s},\bar{a}} \frac{d_\mu(\bar{s})\mu(\bar{a}|\bar{s})p(s|\bar{s},\bar{a})}{d_\mu(s)} \frac{\pi(\bar{a}|\bar{s})}{\mu(\bar{a}|\bar{s})} \lim_{t\to\infty} \mathbb{E}_\mu[F_{t-1} | S_{t-1} = \bar{s}] \\
&= \sum_{\bar{s},\bar{a}} \pi(\bar{a}|\bar{s}) p(s|\bar{s},\bar{a}) d_\mu(\bar{s}) \lim_{t\to\infty} \mathbb{E}_\mu[F_{t-1} | S_{t-1} = \bar{s}] \\
&= \sum_{\bar{s},\bar{a}} [\boldsymbol{P}_\pi]_{\bar{s}s} f(\bar{s}).
\end{aligned}$$

In vector form, we have $\boldsymbol{f}^\top = \boldsymbol{f}^\top \boldsymbol{P}_\pi$. Since the expectations of importance sampling ratios are one and $F_0 = 1$, the expectation of F_t will remain one. This implies that $\sum_s f(s) = 1$. Together with $\boldsymbol{f}^\top = \boldsymbol{f}^\top \boldsymbol{P}_\pi$, we can infer that $\boldsymbol{f} = \boldsymbol{d}_\pi$. Thus, AETD will correct the state distribution back to the on-policy distribution, which is the same as full IS TD. However, compared with full IS TD,

the variance of the trace is greatly tamed down by the averaging. Note that the above analysis of $\boldsymbol{f}$ applies to all AETD algorithms. Here, specifically, differential AETD has the same stability as on-policy differential TD, which is the best we can hope for theoretically. We present the stability of differential AETD with the following proposition, which is proven above:

Proposition 2. *The $\boldsymbol{A}^{ae}$ matrix is the same as on-policy differential TD:*

$$\boldsymbol{A}^{ae} = \begin{bmatrix} \eta & \mathbf{0}^\top \\ \boldsymbol{\Phi}^\top \boldsymbol{d}_\pi & \boldsymbol{\Phi}^\top \boldsymbol{D}_\pi (\boldsymbol{I} - \boldsymbol{P}_\pi)\boldsymbol{\Phi} \end{bmatrix}.$$

Thus, differential AETD and its expected update are stable.

For the convergence analysis, we leave it for future work.

AETD as an extension of generalized ETD

In this section, we adopt generalized ETD (ETD(β); Hallak et al., 2016) to the average-reward setting and describe an alternative way to look at AETD. ETD(β) is the generalization of ETD with the decaying factor of the trace, β, as a free parameter. Taking values from $(0, 1)$, β controls a bias-variance trade-off. It is reported that ETD(β) with an intermediate value of β achieves better performance than ETD (Ghiassian & Sutton, 2021a, 2021b). Adopting it to the average-reward setting, differential ETD(β) performs the following update:

$$\begin{aligned} \boldsymbol{\theta}_{t+1} &= \boldsymbol{\theta}_t + \alpha \rho_t F_t \delta_t \boldsymbol{\phi}_t, \\ \bar{R}_{t+1} &= \bar{R}_t + \alpha \rho_t F_t \eta \delta_t, \\ \delta_t &= R_{t+1} - \bar{R}_t + \boldsymbol{\phi}_{t+1}^\top \boldsymbol{\theta}_t - \boldsymbol{\phi}_t^\top \boldsymbol{\theta}_t, \\ F_{t+1} &= \beta \rho_t F_t + 1, F_0 = 1. \end{aligned} \tag{6}$$

Similar to the discounted episodic setting (Hallak et al., 2016), when $\beta = 0$, differential ETD(β) will degenerate to off-policy differential TD. However, when $\beta = 1$, the original ETD(β) is problematic because the expectation of F_t would go to infinity in the continuing case. From this point of view, AETD could be viewed as the proper extension of ETD(β) when $\beta = 1$ since it treats every time step equally and maintains a bounded expectation. Then, conversely, as a method in the spectrum of ETD(β) algorithms, AETD can also be used in the discounted setting. It is reasonable to conjecture that AETD's advantage of having less bias will come into play when the MDP has terminations, in which the variance issue is more amicable.

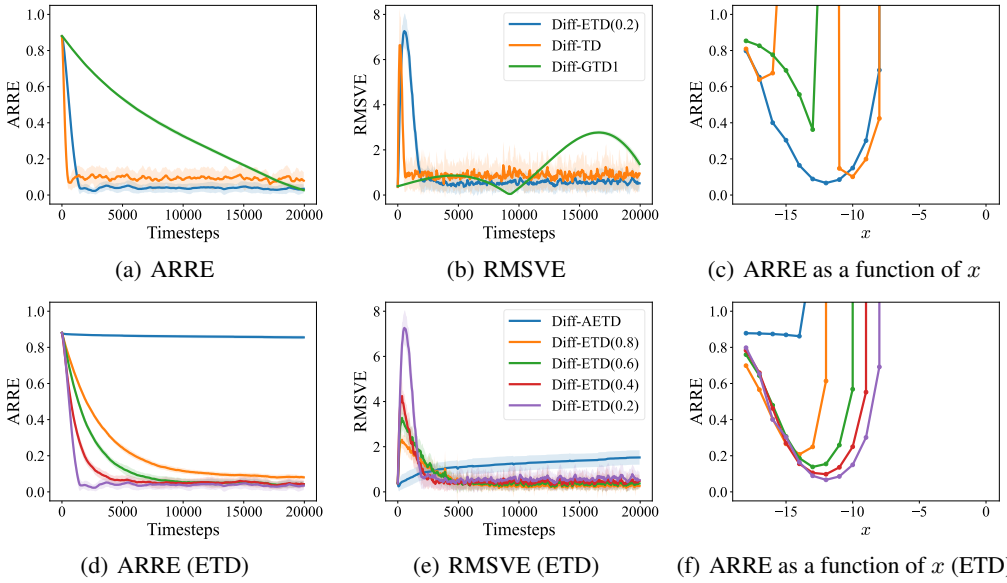
Coming back to the average-reward setting, for $0 < \beta < 1$, the stability of differential ETD(β) is not clear yet. We conjecture that for every β, there exists an average-reward MDP, in which differential ETD(β) is unstable. However, we will show that differential ETD(β) does help in alleviating the instability of off-policy differential TD in Section 5. From a practical perspective, it would be interesting to provide the conditions under which differential ETD(β) converges. We leave this for future work.

5 Experiments

This section presents the experimental results of the emphatic algorithms, differential AETD (Diff-AETD) and differential ETD(β) (Diff-ETD(β)). We compare them with the following baselines: off-policy differential TD (Diff-TD; Wan et al., 2021) and differential GTD1 (Diff-GTD1; S. Zhang, Wan, et al., 2021). For differential ETD(β), the decaying parameter β is chosen from $\{0.2, 0.4, 0.6, 0.8\}$. Note that when $\beta = 0$, differential ETD(β) is equivalent to differential TD; when $\beta = 1$, the sensible version of emphatic algorithms is differential AETD. For the evaluation metric, we use the absolute reward rate error (ARRE) $|r(\pi) - \bar{R}_t|$ to evaluate the accuracy of the reward rate estimation. When possible, we use Tsisiklis and Van Roy's variant (Tsitsiklis & Van Roy, 1999) of root-mean-squared value error (RMSVE) $\inf_{c \in \mathbb{R}} \|\hat{\boldsymbol{v}}_t - (\boldsymbol{v}_\pi + c\boldsymbol{e})\|_{\boldsymbol{d}_\mu}$ for differential value function estimation following Wan et al. (2021).

Results on the counterexample

We first use the two-state MDP in Figure 1 to demonstrate the effectiveness of the emphatic algorithms. We use constant step sizes $\alpha = 2^x$ for all algorithms where $x \in \{-18, -17, \cdots, -1, 0\}$. We use a



(a) ARRE (b) RMSVE (c) ARRE as a function of x

(d) ARRE (ETD) (e) RMSVE (ETD) (f) ARRE as a function of x (ETD)

Figure 2: Results on the two-state MDP with $c = \sqrt{175}$. See text for details.

fixed constant second step size $\eta = 1$ for all compared algorithms. For differential GTD1, we do not use ridge regularization on θ since our examples have a unique solution. We run each algorithm for $20,000$ steps. The results are averaged over 30 independent runs and reported with the best performing hyperparameters, with which the area under the learning curve of ARRE is the smallest.

Here, we report the results on the counterexample with $c = \sqrt{175}$ in Figure 2. Additional results with other value of c can be found in the Appendix. The first two columns plot ARRE and RMSVE, respectively. The third column plot the step-size sensitivity (recall that $\alpha = 2^x$). From the top row of the figure, we can see that differential ETD(β) is the best-performing algorithm compared to baselines in this environment. It achieves both the lowest ARRE and RMSVE. Also, it has a stable convergence and a nice U-shaped sensitivity curve. For both off-policy differential TD and differential GTD1, the bounded ARREs with small step sizes in Figure 2(c) may have hidden the divergence of the algorithms because of insufficient training steps. In Section 3, we have proven that off-policy differential TD is unstable in this counterexample, which means its expected update diverges. Nevertheless, off-policy differential TD itself converges but with higher errors and within a very narrow range of step sizes. Interestingly, divergence occurs with step sizes smaller than the converging step sizes. Finally, differential GTD1 doesn't converge with the chosen step sizes, which contradict its convergence guarantee. We suspect the reason is that differential GTD1 is highly sensitive to the magnitude of the features (see the convergence of differential GTD1 in Figure B.1(c) in the Appendix). The converging step sizes may be out of the scope we are searching for.

The bottom row of Figure 2 shows how different emphatic algorithms perform. We can easily notice that differential AETD can barely learn in this environment. It suffers not only from the high variance of importance sampling ratios but also from the large features in this counterexample. On the other hand, differential ETD(β) is almost immune to the large features (see also Figure B.1(d) in the Appendix). It achieves the lowest ARRE with $\beta = 0.2$ within the training steps. However, from Figure 2(d), it is reasonable to conjecture that a high value of β would achieve the lowest ARRE if more training steps are given. This can also be supported by Figure 2(e), which shows that higher values of β have slightly lower RMSVEs. From Figure 2(f), we can see that as β becomes larger, the width of the U-shaped curve becomes smaller in accordance with the observations from (Ghiassian & Sutton, 2021a, 2021b), ETD(β) suffers from high variance more with higher value of β.

Results on MuJoCo tasks with nonlinear function approximation

In this section, we present some rudimentary experiment results on four MuJoCo tasks. Following S. Zhang, Wan, et al. (2021), we first train a deterministic policy π_0 with TD3 (Fujimoto, Hoof, &

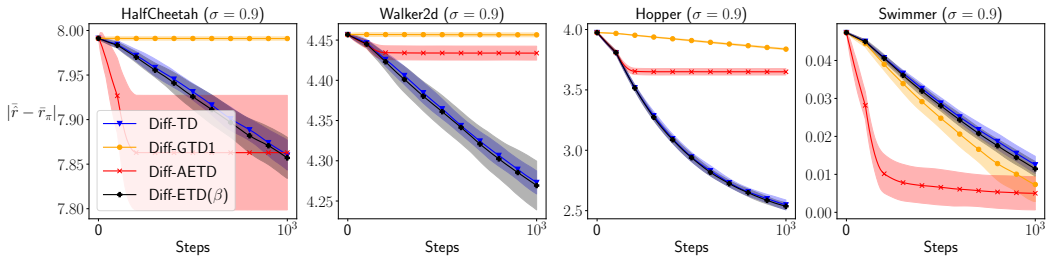


Figure 3: Results on four MuJoCo tasks. See text for details.

Meger, 2018). To avoid ill-posed importance sampling ratios, we use a stochastic target policy π by introducing Gaussian noise with zero mean and variance 0.1 to π_0. The behavior policy is composed similarly with Gaussian noise with zero mean and variance σ which takes value from $\{0.5, 0.7, 0.9\}$. For all the algorithms, the learning rate is tuned from $\{1, 0.5, 0.1, 0.05, 0.01, 0.005\}$. To stabilize training, we clip all the importance sampling ratios to a maximum value of 1. The results are averaged over 30 independent runs and reported with the best performing hyperparameters, with which the ARRE is the smallest at the end of the training following S. Zhang, Wan, et al. (2021). For other implementation details, please refer to the Appendix.

Here, we report the results with $\sigma = 0.9$ in Figure 3. The results with other values of σ are deferred to the Appendix. From the figure, we can see that differential ETD(β) performs similarly to off-policy differential TD with slightly lower errors. On the other hand, differential AETD is very unstable and performs inconsistently across different tasks compared to other algorithms. Specifically, it performs the best in Swimmer, but its learning plateaus very quickly in the other three tasks. Finally, differential GTD1 exhibits learning difficulty in all the tasks but Swimmer, in which it performs slightly better than differential ETD(β) and off-policy differential TD.

6 Conclusions and future work

In this paper, we have discussed some of our observations on the problem of average-reward off-policy policy evaluation with linear function approximation. Specifically, we find that differential TD is relatively more robust than average-cost TD in the off-policy setting. However, off-policy differential TD without any treatments still suffers from instability. We illustrate the instability of off-policy average-cost TD and off-policy differential TD with corresponding counterexamples. To address the instability, we investigate the emphatic approach for off-policy learning, which is well understood in the discounted setting. Our contributions to this subject are twofold. Firstly, we complete the spectrum of the emphatic algorithms for the continuing setting with the average emphatic trace. This completion is of two senses. In a way, average emphatic TD can be considered the extension of ETD to the average-reward setting. Alternatively, it can also be regarded as the proper form of ETD(β) for $\beta = 1$ and applicable to the discounted setting. Secondly, we validate the effectiveness of the emphatic algorithms in the average-reward scenario through an empirical study on a two-state MDP family, including the counterexamples. We demonstrate the bias-variance trade-off controlled by the decaying parameter β, mirroring the same observation in the discounted setting. More importantly, we show that differential ETD(β) is the best performing algorithm in the two-state MDPs we tested in terms of the asymptotic error and step-size sensitivity. Additionally, we extend the differential emphatic algorithms to nonlinear function approximation. Rudimentary experiment results show batch differential TD is not sensitive to reweighting updates by expected emphasis on MuJoCo robotic simulation tasks. Further investigation is needed for the nonlinear function approximation setting.

Despite the potential of the emphatic algorithms for average-reward reinforcement learning, it is not clear yet under what conditions, with what values of β, differential ETD(β) is guaranteed to converge. Further, if it converges, we would also like to know the bias of the point of convergence. We hope to answer these critical theoretical questions in the future.

References

Espeholt, L., Soyer, H., Munos, R., Simonyan, K., Mnih, V., Ward, T., ... Kavukcuoglu, K. (2018). IMPALA: Scalable distributed deep-RL with importance weighted actor-learner architectures. In *Proceedings of the 35th International Conference on Machine Learning* (Vol. 80, pp. 1407–1416). PMLR.

Fujimoto, S., Hoof, H., & Meger, D. (2018). Addressing function approximation error in actor-critic methods. In *International conference on machine learning* (pp. 1587–1596).

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

Ghiassian, S., & Sutton, R. S. (2021b). An empirical comparison of off-policy prediction learning algorithms on the collision task. *arXiv preprint arXiv:2106.00922*.

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

Jiang, R., Zahavy, T., Xu, Z., White, A., Hessel, M., Blundell, C., & Van Hasselt, H. (2021). Emphatic algorithms for deep reinforcement learning. In *International Conference on Machine Learning* (pp. 5023–5033).

Jiang, R., Zhang, S., Chelu, V., White, A., & van Hasselt, H. (2022). Learning expected emphatic traces for deep RL. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 36, pp. 7015–7023).

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

Mahmood, A. R., Yu, H., & Sutton, R. S. (2017). Multi-step off-policy learning without importance sampling ratios. *arXiv preprint arXiv:1702.03006*.

Mousavi, A., Li, L., Liu, Q., & Zhou, D. (2020). Black-box off-policy estimation for infinite-horizon reinforcement learning. *arXiv preprint arXiv:2003.11126*.

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

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

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

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

Tsitsiklis, J. N., & Van Roy, B. (1999). Average cost temporal-difference learning. *Automatica*, *35*(11), 1799–1808.

Wan, Y., Naik, A., & Sutton, R. S. (2021). Learning and planning in average-reward markov decision processes. In *International Conference on Machine Learning* (pp. 10653–10662).

White, A., Modayil, J., & Sutton, R. S. (2012). Scaling life-long off-policy learning. In *2012 ieee international conference on development and learning and epigenetic robotics (icdl)* (pp. 1–6).

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

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

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

Zhang, S., Wan, Y., Sutton, R. S., & Whiteson, S. (2021). Average-reward off-policy policy evaluation with function approximation. In *International Conference on Machine Learning* (pp. 12578–12588).

Zhang, S., Zhang, Z., & Maguluri, S. T. (2021). Finite sample analysis of average-reward td learning and q-learning. *Advances in Neural Information Processing Systems*, *34*, 1230–1242.

A Proof of Proposition 1

The proof is inspired by the proof of Lemma 2 in the work of S. Zhang, Zhang, and Maguluri (2021) which provides finite sample analysis for on-policy average-cost TD. Recall that $\boldsymbol{u} = [\bar{R}, \boldsymbol{\theta}^\top]^\top$ and

$$\boldsymbol{A}^{\text{diff}} = \begin{bmatrix} \eta & \boldsymbol{d}_\mu^\top(\boldsymbol{I} - \boldsymbol{P}_\pi)\boldsymbol{\Phi} \\ \boldsymbol{\Phi}^\top \boldsymbol{d}_\mu & \boldsymbol{\Phi}^\top \boldsymbol{D}_\mu(\boldsymbol{I} - \boldsymbol{P}_\pi)\boldsymbol{\Phi} \end{bmatrix},$$

we can rewrite the minimization problem $\min_{\|\boldsymbol{u}\|_2=1} \boldsymbol{u}^\top \boldsymbol{A}\boldsymbol{u}$ as

$$\min_{\sqrt{\bar{R}^2+\|\boldsymbol{\theta}\|_2^2}=1} \eta\bar{R}^2 + \bar{R}\boldsymbol{\theta}^\top\boldsymbol{\Phi}^\top\boldsymbol{d}_\mu + \bar{R}\boldsymbol{d}_\mu^\top(\boldsymbol{I} - \boldsymbol{P}_\pi)\boldsymbol{\Phi}\boldsymbol{\theta} + \boldsymbol{\theta}^\top\boldsymbol{\Phi}^\top\boldsymbol{D}_\mu(\boldsymbol{I} - \boldsymbol{P}_\pi)\boldsymbol{\Phi}\boldsymbol{\theta}.$$

Now for any $\bar{R} \in \mathbb{R}$ and $\boldsymbol{\theta} \in \mathbb{R}^d$, we have

$$\begin{aligned}
\left|\bar{R}\boldsymbol{\theta}^\top\boldsymbol{\Phi}^\top\boldsymbol{d}_\mu + \bar{R}\boldsymbol{d}_\mu^\top(\boldsymbol{I} - \boldsymbol{P}_\pi)\boldsymbol{\Phi}\boldsymbol{\theta}\right| &= \left|\bar{R}\boldsymbol{d}_\mu^\top(2\boldsymbol{I} - \boldsymbol{P}_\pi)\boldsymbol{\Phi}\boldsymbol{\theta}\right| \\
&\le \left|\bar{R}\right| \left|\boldsymbol{d}_\mu^\top(2\boldsymbol{I} - \boldsymbol{P}_\pi)\boldsymbol{\Phi}\boldsymbol{\theta}\right| \\
&\le \left|\bar{R}\right| \left(2\left|\boldsymbol{d}_\mu^\top\boldsymbol{\Phi}\boldsymbol{\theta}\right| + \left|\boldsymbol{d}_\mu^\top\boldsymbol{P}_\pi\boldsymbol{\Phi}\boldsymbol{\theta}\right|\right) \\
&\le \left|\bar{R}\right| \left(2\|\boldsymbol{d}_\mu\|_1\|\boldsymbol{\Phi}\boldsymbol{\theta}\|_\infty + \|\boldsymbol{P}_\pi^\top\boldsymbol{d}_\mu\|_1\|\boldsymbol{\Phi}\boldsymbol{\theta}\|_\infty\right) \\
&= 3\left|\bar{R}\right| \|\boldsymbol{\Phi}\boldsymbol{\theta}\|_\infty \\
&\le 3\left|\bar{R}\right| \max_{s\in\mathcal{S}} \|\boldsymbol{\phi}(s)\|_2\|\boldsymbol{\theta}\|_2 \\
&\le 3B\left|\bar{R}\right| \|\boldsymbol{\theta}\|_2.
\end{aligned}$$

The last step is due to we assume a finite state space which implies there exist $B \ge 0$ such that $\max_{s\in\mathcal{S}} \|\boldsymbol{\phi}(s)\|_2 \le B$. Next, by the definition of Δ, for any $\boldsymbol{\theta} \in \mathbb{R}^d$, we have

$$\boldsymbol{\theta}^\top\boldsymbol{\Phi}^\top\boldsymbol{D}_\mu(\boldsymbol{I} - \boldsymbol{P}_\pi)\boldsymbol{\Phi}\boldsymbol{\theta} \ge \Delta\|\boldsymbol{\theta}\|_2^2.$$

Then we have

$$\begin{aligned}
&\min_{\sqrt{\bar{R}^2+\|\boldsymbol{\theta}\|_2^2}=1} \eta\bar{R}^2 + \bar{R}\boldsymbol{\theta}^\top\boldsymbol{\Phi}^\top\boldsymbol{d}_\mu + \bar{R}\boldsymbol{d}_\mu^\top(\boldsymbol{I} - \boldsymbol{P}_\pi)\boldsymbol{\Phi}\boldsymbol{\theta} + \boldsymbol{\theta}^\top\boldsymbol{\Phi}^\top\boldsymbol{D}_\mu(\boldsymbol{I} - \boldsymbol{P}_\pi)\boldsymbol{\Phi}\boldsymbol{\theta} \\
&\ge \min_{\sqrt{\bar{R}^2+\|\boldsymbol{\theta}\|_2^2}=1} \eta\bar{R}^2 - 3B\left|\bar{R}\right| \|\boldsymbol{\theta}\|_2 + \Delta\|\boldsymbol{\theta}\|_2^2 \\
&= \min_{\bar{R}\in[-1,1]} \eta\left|\bar{R}\right|^2 - 3B\left|\bar{R}\right|\sqrt{1-\left|\bar{R}\right|^2} + \Delta\left(1-\left|\bar{R}\right|^2\right) \\
&= \min_{x\in[0,1]} \eta x - 3B\sqrt{x\,(1-x)} + \Delta\,(1-x) \\
&= \Delta + \min_{x\in[0,1]} (\eta-\Delta)\,x - 3B\sqrt{x\,(1-x)}.
\end{aligned}$$

When $\eta \ge \Delta + 3B\sqrt{\left(\frac{3B}{\Delta}\right)^2 - 1}$, we have

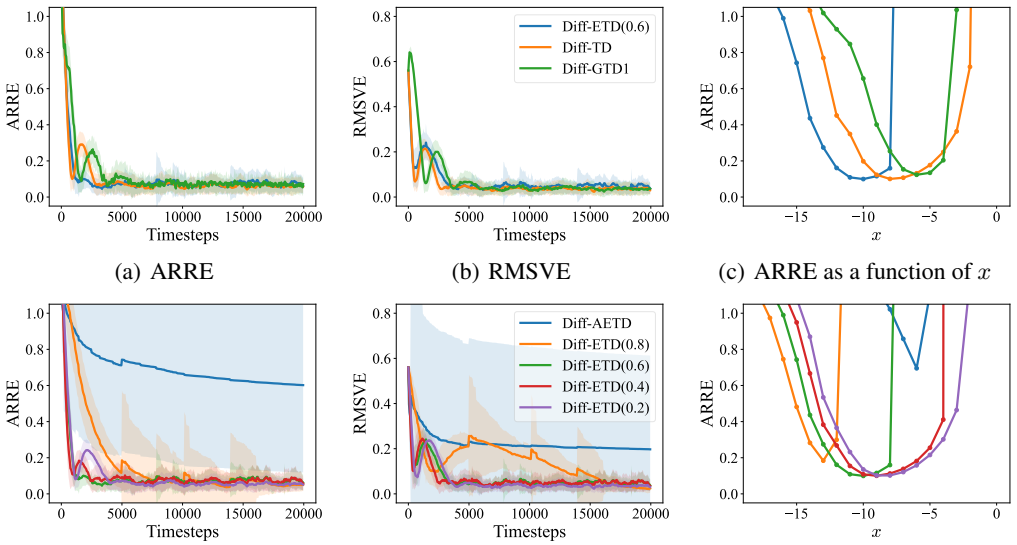
$$\min_{x\in[0,1]} (\eta-\Delta)\,x - 3B\sqrt{x\,(1-x)} \ge -\frac{\Delta}{2}.$$

Consequently, $\min_{\|\boldsymbol{u}\|_2=1} \boldsymbol{u}^\top\boldsymbol{A}\boldsymbol{u} \ge \frac{\Delta}{2} > 0$. □

B Additional experiment results and details

Results on the two-state MDP with small features

To draw a complete picture of how differential emphatic algorithms perform, we also report results on the two-state MDP with $c = 1$. To make the task more challenging, we increase the gap



(a) ARRE (b) RMSVE (c) ARRE as a function of x

(d) ARRE (ETD) (e) RMSVE (ETD) (f) ARRE as a function of x (ETD)

Figure B.1: Results on the two-state MDP with $c = 1$.

between the behavior and target policies. In the experiment, we set $\pi(a_1|c) = \pi(a_2|2c) = 0.2$ and $\mu(a_1|c) = \mu(a_2|2c) = 0.8$.

The results are shown in Figure B.1. From the top row of the figure, we can see that similar performance is achieved by the two baseline methods and differential ETD(β), among which differential GTD1 has a slightly higher ARRE. Suffering from high variance, differential AETD learns much more slowly than these methods. Nevertheless, its performance in this small-feature setting is much better than in the previous large-feature environment. In fact, it is much better than differential full IS TD, which is not shown in the results but barely learns. As for step-size sensitivity, from Figure B.1(c), we can see that off-policy differential TD is the least sensitive and differential GTD1 is the most sensitive. Although differential ETD(β) with the selected value of β is more sensitive to step sizes compared to off-policy differential TD, we can see that from Figure B.1(f), with $\beta = 0.2$, differential ETD(β) actually achieves almost the same performance as off-policy differential TD.

Implementation details for experiments on MuJoCo tasks

There are two challenges when applying differential AETD and differential ETD(β) in complex tasks like MuJoCo tasks that require nonlinear function approximation. Firstly, although having a lower variance than full IS TD, AETD and ETD(β) still suffer from the high variance issue. This is exacerbated when neural networks are applied. Secondly, calculating the emphatic traces requires trajectory-based data instead of transition-based data, which is more challenging to maintain when a data buffer is needed for training neural networks. To address these two issues, we adopt the strategy to learn an *expected emphasis* (S. Zhang, Liu, Yao, & Whiteson, 2020; Jiang et al., 2022) in the nonlinear setting. Specifically, we estimate the following expected emphasis:

$$\tilde{f}(s) = \lim_{t\to\infty} \mathbb{E}_\mu[F_t|S_t = s], \tag{7}$$

where F_t is the emphasis defined in update (5) or update (6). We use single-step bootstrap target to learn the expected emphasis. Specifically, for every transition pair $(S_t, A_t, R_{t+1}, S_{t+1})$, $\tilde{f}(S_{t+1})$ is updated towards $\rho_t \tilde{f}(S_t)$ for AETD and $\beta\rho_t \tilde{f}(S_t) + 1$ for ETD(β).

Following S. Zhang, Wan, et al. (2021), we use batch updates to stabilize the training of neural networks. To this end, we collect data with the behavior policy for 10^6 steps. The implementations of off-policy differential TD and differential GTD1 are extended from S. Zhang, Wan, et al.. For differential AETD and differential ETD(β), we use neural networks to parameterize $\boldsymbol{v}$ and $\tilde{\boldsymbol{f}}$. Similar to off-policy differential TD, we use a target network for $\tilde{\boldsymbol{f}}$ which is updated every 100 steps.

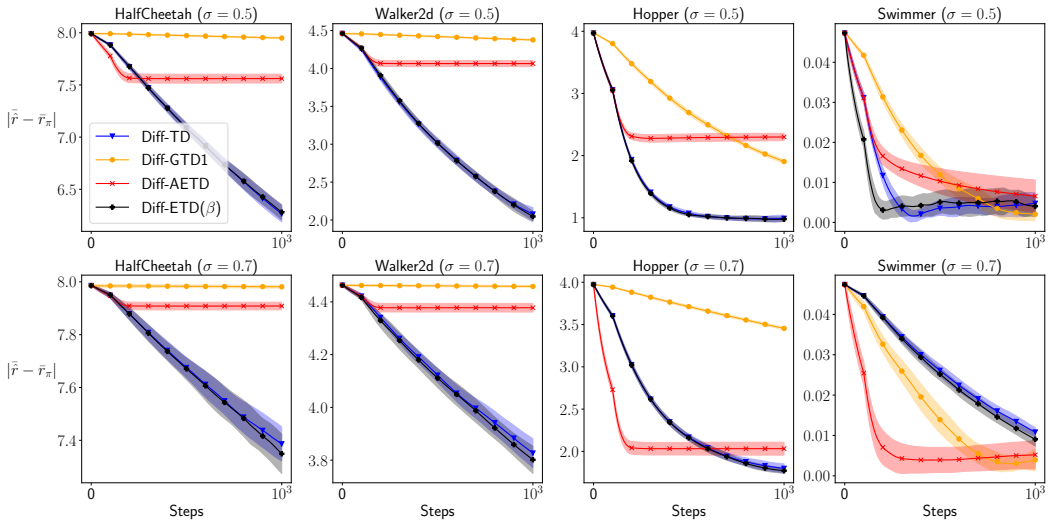


Figure B.2: Results on four MuJoCo tasks.

More results on MuJoCo tasks

The performance of different algorithms in MuJoCo tasks with $\sigma = 0.5$ and $\sigma = 0.7$ is shown in Figure B.2. The conclusion of these results are similar to that of Figure 3.