Provably Convergent Two-Timescale Off-Policy Actor-Critic with Function Approximation

Shangtong Zhang [1] **Bo Liu** [2] **Hengshuai Yao** [3] **Shimon Whiteson** [1]

Abstract

We present the first provably convergent two-timescale off-policy actor-critic algorithm (COF-PAC) with function approximation. Key to COF-PAC is the introduction of a new critic, the *emphasis critic*, which is trained via Gradient Emphasis Learning (GEM), a novel combination of the key ideas of Gradient Temporal Difference Learning and Emphatic Temporal Difference Learning. With the help of the emphasis critic and the canonical value function critic, we show convergence for COF-PAC, where the critics are linear, and the actor can be nonlinear.

1. Introduction

The policy gradient theorem and the corresponding actor-critic algorithm (Sutton et al., 2000; Konda, 2002) have recently enjoyed great success in various domains, e.g., defeating the top human player in Go (Silver et al., 2016), achieving human-level control in Atari games (Mnih et al., 2016). However, the canonical actor-critic algorithm is on-policy and hence suffers from significant data inefficiency (e.g., see Mnih et al. (2016)). To address this issue, Degris et al. (2012) propose the Off-Policy Actor-Critic (Off-PAC) algorithm. Off-PAC has been extended in various ways, e.g., off-policy Deterministic Policy Gradient (DPG, Silver et al. 2014), Deep Deterministic Policy Gradient (DDPG, Lillicrap et al. 2015), Actor Critic with Experience Replay (ACER, Wang et al. 2016), off-policy Expected Policy Gradient (EPG, Ciosek & Whiteson 2017), TD3 (Fujimoto et al., 2018), and IMPALA (Espeholt et al., 2018). Off-PAC and its extensions have enjoyed great empirical success as the canonical on-policy actor-critic algorithm. There is, however, a theoretical gap between the canonical on-policy actor-critic and Off-PAC. Namely, on-policy actor-critic has a two-timescale convergent analysis under function approximation (Konda, 2002), but Off-PAC is convergent only in the tabular setting (Degris et al., 2012). While there have been several attempts to close this gap (Imani et al., 2018; Maei, 2018; Zhang et al., 2019; Liu et al., 2019), none of them is convergent under function approximation without imposing strong assumptions (e.g., assuming the critic converges).

In this paper, we close this long-standing theoretical gap via the Convergent Off-Policy Actor-Critic (COF-PAC) algorithm, the first provably convergent two-timescale off-policy actor-critic algorithm. COF-PAC builds on Actor-Critic with Emphatic weightings (ACE, Imani et al. 2018), which reweights Off-PAC updates with *emphasis* through the *followon trace* (Sutton et al., 2016). The emphasis accounts for off-policy learning by adjusting the state distribution, and the followon trace approximates the emphasis (see Sutton et al. 2016).[4] Intuitively, estimating the emphasis of a state using the followon trace is similar to estimating the value of a state using a single Monte Carlo return. Thus it is not surprising that the followon trace can have unbounded variance (Sutton et al., 2016) and large emphasis approximation error, complicating the convergence analysis of ACE.

Instead of using the followon trace, we propose a novel stochastic approximation algorithm, Gradient Emphasis Learning (GEM), to approximate the emphasis in COF-PAC, inspired by Gradient TD methods (GTD, Sutton et al. 2009b;a), Emphatic TD methods (ETD, Sutton et al. 2016), and reversed TD methods (Wang et al., 2007; 2008; Hallak & Mannor, 2017; Gelada & Bellemare, 2019). We prove the almost sure convergence of GEM, as well as other GTD-style algorithms, with linear function approximation under a slowly changing target policy. With the help of GEM, we prove the convergence of COF-PAC, where the policy parameterization can be nonlinear, and the convergence level is the same as the on-policy actor-critic (Konda, 2002).

[1]University of Oxford [2]Auburn University [3]Huawei Technologies. Correspondence to: Shangtong Zhang <shangtong.zhang@cs.ox.ac.uk>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

[4]We use emphasis to denote the limit of the expectation of the followon trace, which is slightly different from Sutton et al. (2016) and is clearly defined in the next section.

2. Background

We use $||x||_\Xi \doteq \sqrt{x^\top \Xi x}$ to denote the norm induced by a positive definite matrix Ξ, which induces the matrix norm $||A||_\Xi \doteq \sup_{||x||_\Xi=1} ||Ax||_\Xi$. To simplify notation, we write $||\cdot||$ for $||\cdot||_I$, where I is the identity matrix. All vectors are column vectors. We use "0" to denote an all-zero vector and an all-zero matrix when the dimension can be easily deduced from the context, and similarly for "1". When it does not confuse, we use vectors and functions interchangeably. Proofs are in the appendix.

We consider an infinite-horizon Markov Decision Process (MDP) with a finite state space $\mathcal{S}$ with $|\mathcal{S}|$ states, a finite action space $\mathcal{A}$ with $|\mathcal{A}|$ actions, a transition kernel $p : \mathcal{S}\times\mathcal{S}\times\mathcal{A} \to [0,1]$, a reward function $r : \mathcal{S}\times\mathcal{A}\times\mathcal{S} \to \mathbb{R}$, and a discount factor $\gamma \in [0,1)$. At time step t, an agent at a state S_t takes an action A_t according to $\mu(\cdot|S_t)$, where $\mu : \mathcal{A}\times\mathcal{S} \to [0,1]$ is a *fixed behavior policy*. The agent then proceeds to a new state S_{t+1} according to $p(\cdot|S_t, A_t)$ and gets a reward $R_{t+1} \doteq r(S_t, A_t, S_{t+1})$. In the off-policy setting, the agent is interested in a *target policy* π. We use $G_t \doteq \sum_{k=1}^{\infty} \gamma^{k-1} R_{t+k}$ to denote the return at time step t when following π instead of μ. Consequently, we define the state value function v_π and the state action value function q_π as $v_\pi(s) \doteq \mathbb{E}_\pi[G_t|S_t = s]$ and $q_\pi(s,a) \doteq \mathbb{E}_\pi[G_t|S_t = s, A_t = a]$. We use $\rho(s,a) \doteq \frac{\pi(a|s)}{\mu(a|s)}$ to denote the importance sampling ratio and define $\rho_t \doteq \rho(S_t, A_t)$ (Assumption 1 below ensures ρ is well-defined). We sometimes write ρ as ρ_π to emphasize its dependence on π.

Policy Evaluation: We consider linear function approximation for policy evaluation. Let $x : \mathcal{S} \to \mathbb{R}^{K_1}$ be the state feature function, and $\tilde{x} : \mathcal{S}\times\mathcal{A} \to \mathbb{R}^{K_2}$ be the state-action feature function. We use $X \in \mathbb{R}^{|\mathcal{S}|\times K_1}$ and $\tilde{X} \in \mathbb{R}^{N_{sa}\times K_2} (N_{sa} \doteq |\mathcal{S}|\times|\mathcal{A}|)$ to denote feature matrices, where each row of X is $x(s)$ and each row of $\tilde{X}$ is $\tilde{x}(s,a)$. We use as shorthand that $x_t \doteq x(S_t), \tilde{x}_t \doteq \tilde{x}(S_t, A_t)$. Let $d_\mu \in \mathbb{R}^{|\mathcal{S}|}$ be the stationary distribution of μ; we define $\tilde{d}_\mu \in \mathbb{R}^{N_{sa}}$ where $\tilde{d}_\mu(s,a) \doteq d_\mu(s)\mu(a|s)$. We define $D \doteq diag(d_\mu) \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$ and $\tilde{D} \doteq diag(\tilde{d}_\mu) \in \mathbb{R}^{N_{sa}\times N_{sa}}$. Assumption 1 below ensures d_μ exists and D is invertible, as well as $\tilde{D}$. Let $P_\pi \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$ be the state transition matrix and $\tilde{P}_\pi \in \mathbb{R}^{N_{sa}\times N_{sa}}$ be the state-action transition matrix, i.e., $P_\pi(s,s') \doteq \sum_a \pi(a|s)p(s'|s,a), \tilde{P}_\pi((s,a),(s',a')) \doteq p(s'|s,a)\pi(a'|s')$. We use $v \doteq X\nu, q \doteq \tilde{X}u$ to denote estimates for v_π, q_π respectively, where ν, u are learnable parameters.

We first consider GTD methods. For a vector $v \in \mathbb{R}^{|\mathcal{S}|}$, we define a projection $\Pi v \doteq Xy^*, y^* \doteq \arg\min_y ||Xy - v||_D^2$. We have $\Pi = X(X^\top DX)^{-1}X^\top D$ (Assumption 2 below ensures the existence of $(X^\top DX)^{-1}$). Similarly, for a vector $q \in \mathbb{R}^{N_{sa}}$, we define a projection $\tilde{\Pi} \doteq \tilde{X}(\tilde{X}^\top\tilde{D}\tilde{X})^{-1}\tilde{X}^\top\tilde{D}$. The value function v_π is the unique fixed point of the Bellman operator $\mathcal{T} : \mathcal{T}v \doteq r_\pi + \gamma P_\pi v$ where $r_\pi(s) \doteq \sum_{a,s'} \pi(a|s)p(s'|s,a)r(s,a,s')$. Similarly, q_π is the unique fixed point for the operator $\tilde{\mathcal{T}} : (\tilde{\mathcal{T}}q)(s,a) \doteq \tilde{r} + \gamma\tilde{P}_\pi q$, where $\tilde{r} \in \mathbb{R}^{N_{sa}}$ and $\tilde{r}(s,a) \doteq \sum_{s'} p(s'|s,a)r(s,a,s')$. GTD2 (Sutton et al., 2009a) learns the estimate v for v_π, by minimizing $||\Pi\mathcal{T}v - v||_D^2$. GQ(0) (Maei, 2011) learns the estimate q for q_π by minimizing $||\tilde{\Pi}\tilde{\mathcal{T}}q - q||_{\tilde{D}}^2$. Besides GTD methods, ETD methods are also used for off-policy policy evaluation. ETD(0) updates ν as

$$M_t \doteq i(S_t) + \gamma\rho_{t-1}M_{t-1}, \tag{1}$$

$$\nu_{t+1} \doteq \nu_t + \alpha M_t\rho_t(R_{t+1} + \gamma x_{t+1}^\top\nu_t - x_t^\top\nu_t)x_t^\top, \tag{2}$$

where α is a step size, M_t is the followon trace, and $i : \mathcal{S} \to [0,\infty)$ is the interest function reflecting the user's preference for different states (Sutton et al., 2016). The interest is usually set to 1 for all states (Sutton et al., 2016), meaning all states are equally important. But if we are interested in learning an optimal policy for only a subset of states (e.g., the initial states), we can set the interest to 1 for those states and to 0 otherwise (White, 2017). The interest function can also be regarded as a generalization of the initiation set in the option framework (White, 2017). We refer the reader to White (2017) for more usage of the interest function.

Control: Off-policy actor-critic methods (Degris et al., 2012; Imani et al., 2018) aim to maximize the excursion objective $J(\pi) \doteq \sum_s d_\mu(s)i(s)v_\pi(s)$ by adapting the target policy π. We assume π is parameterized by $\theta \in \mathbb{R}^K$, and use θ, π, π_θ interchangeably in the rest of this paper when it does not confuse. All gradients are taken w.r.t. θ unless otherwise specified. According to the off-policy policy gradient theorem (Imani et al., 2018), the policy gradient is $\nabla J(\theta) = \sum_s \bar{m}(s)\sum_a q_\pi(s,a)\nabla\pi(a|s)$, where $\bar{m} \doteq (I - \gamma P_\pi^\top)^{-1}Di \in \mathbb{R}^{|\mathcal{S}|}$. We rewrite $\bar{m}$ as $DD^{-1}(I - \gamma P_\pi^\top)^{-1}Di$ and define

$$m_\pi \doteq D^{-1}(I - \gamma P_\pi^\top)^{-1}Di.$$

We therefore have $\bar{m} = Dm_\pi$, yielding

$$\nabla J(\theta) = \mathbb{E}_{s\sim d_\mu, a\sim\mu(\cdot|s)}[m_\pi(s)\psi_\theta(s,a)q_\pi(s,a)],$$

where $\psi_\theta(s,a) \doteq \rho_\theta(s,a)\nabla\log\pi(a|s) \in \mathbb{R}^K$. We refer to m_π as the *emphasis* in the rest of this paper. To compute $\nabla J(\theta)$, we need m_π and q_π, to which we typically do not have access. Degris et al. (2012) ignore the emphasis m_π and update θ as $\theta_{t+1} \leftarrow \theta_t + \alpha\rho_t q_\pi(S_t, A_t)\nabla\log\pi(A_t|S_t)$ in Off-PAC, which is theoretically justified only in the tabular setting.[5] Imani et al. (2018) approximate $m_\pi(S_t)$ with the followon trace M_t, yielding the ACE update

[5] See Errata in Degris et al. (2012)

$\theta_{t+1} \leftarrow \theta_t + \alpha M_t \rho_t q_\pi(S_t, A_t) \nabla \log \pi(A_t|S_t)$. Assuming $\lim_{t\to\infty} \mathbb{E}_\mu[M_t|S_t = s]$ exists and π is fixed, Sutton et al. (2016) show that $\lim_{t\to\infty} \mathbb{E}_\mu[M_t|S_t = s] = m_\pi(s)$. The existence of this limit is later established in Lemma 1 in Zhang et al. (2019).

2.1. Assumptions

Assumption 1. *The Markov chain induced by the behavior policy μ is ergodic, and $\forall(s,a)$, $\mu(a|s) > 0$.*

Assumption 2. *The matrices $C \doteq X^\top D X, \tilde{C} \doteq \tilde{X}^\top \tilde{D} \tilde{X}$ are nonsingular.*

Assumption 3. *There exists a constant $C_0 < \infty$ such that $\forall(s,a,\theta,\bar{\theta})$,*

$$||\psi_\theta(s,a)|| \leq C_0, ||\nabla\psi_\theta(s,a)|| \leq C_0$$
$$|\pi_\theta(a|s) - \pi_{\bar{\theta}}(a|s)| \leq C_0||\theta - \bar{\theta}||,$$
$$||\psi_\theta(s,a) - \psi_{\bar{\theta}}(s,a)|| \leq C_0||\theta - \bar{\theta}||.$$

Remark 1. *The nonsingularity in Assumption 2 is commonly assumed in GTD methods (Sutton et al., 2009b;a; Maei, 2011) and can be satisfied by using linearly independent features. Assumption 3 contains common assumptions for policy parameterization akin to those of Sutton et al. (2000); Konda (2002).*

Lemma 1. *Under Assumptions (1, 3), there exists a constant $C_1 < \infty$ such that $\forall(\theta,\bar{\theta})$*

$$||\nabla J(\theta)|| \leq C_1,\ ||\nabla J(\theta) - \nabla J(\bar{\theta})|| \leq C_1||\theta - \bar{\theta}||,$$
$$||\tfrac{\partial^2 J(\theta)}{\partial\theta_i \partial\theta_j}|| \leq C_1.$$

Lemma 2. *Under Assumption 1, $||P_\pi||_D = ||D^{-1}P_\pi^\top D||_D$*

3. Gradient Emphasis Learning

Motivation: The followon trace M_t has two main disadvantages. First, when we use M_t to approximate $m_\pi(S_t)$ (e.g., in ACE), the approximation error tends to be large. M_t is a random variable and although its conditional expectation $\mathbb{E}_\mu[M_t|S_t = s]$ converges to $m_\pi(S_t)$ under a fixed target policy π, M_t itself can have unbounded variance (Sutton et al., 2016), indicating the approximation error $|M_t - m_\pi(S_t)|$ can be unbounded. Moreover, in our actor-critic setting, where π keeps changing, it is not clear whether this convergence holds or not. Theoretically, this large approximation error may preclude a convergent analysis for ACE. Empirically, this large variance makes ETD hard to use in practice. For example, as pointed out in Sutton & Barto (2018), "it is nigh impossible to get consistent results in computational experiments" (for ETD) in Baird's counterexample (Baird, 1995), a common off-policy learning benchmark.

Second, it is hard to query the emphasis $m_\pi(s)$ for a given state s using the followon trace. As M_t is only a scalar, it is almost memoryless. To obtain an emphasis estimation for a given state using the followon trace, we have to simulate a trajectory long enough to go into the mixing stage and visit that particular state, which is typically difficult in offline training. This lack of memory is also a cause of the large approximation error.

In this paper, we propose a novel stochastic approximation algorithm, Gradient Emphasis Learning (GEM), to learn m_π using function approximation. GEM can track the true emphasis m_π under a changing target policy π.

Algorithm Design: We consider linear function approximation, and our estimate for m_π is $m \doteq Xw$, where $w \in \mathbb{R}^{K_1}$ is the learnable parameters. For a vector $y \in \mathbb{R}^{|\mathcal{S}|}$, we define an operator $\hat{\mathcal{T}}$ as $\hat{\mathcal{T}}y \doteq i + \gamma D^{-1} P_\pi^\top D y$.

Proposition 1. *$\hat{\mathcal{T}}$ is a contraction mapping w.r.t. some weighted maximum norm and m_π is its unique fixed point.*

The proof involves arguments from Bertsekas & Tsitsiklis (1989), where the choice of the weighted maximum norm depends on $\gamma D^{-1} P_\pi^\top D$. Our operator $\hat{\mathcal{T}}$ is a generalization of the discounted COP-TD operator $Y_{\hat{\gamma}}$ (Gelada & Bellemare, 2019), where $Y_{\hat{\gamma}} y \doteq (1-\hat{\gamma})1 + \hat{\gamma} D^{-1} P_\pi^\top D y$ and $\hat{\gamma}$ is a scalar similar to γ. They show that $Y_{\hat{\gamma}}$ is contractive only when $\hat{\gamma}$ is small enough. Here our Proposition 1 proves contraction for any $\gamma < 1$. Although $\hat{\mathcal{T}}$ and $Y_{\hat{\gamma}}$ are similar, they are designed for different purposes. Namely, $Y_{\hat{\gamma}}$ is designed to learn a density ratio, while $\hat{\mathcal{T}}$ is designed to learn the emphasis. Emphasis generalizes density ratio in that users are free to choose the interest i in $\hat{\mathcal{T}}$.

Given Proposition 1, it is tempting to compose a semi-gradient update rule for updating w analogously to discounted COP-TD, where the incremental update for w_t is $(i(S_{t+1}) + \gamma\rho_t x_t^\top w_t - x_{t+1}^\top w_t) x_{t+1}$. This semi-gradient update, however, can diverge for the same reason as the divergence of off-policy linear TD: the key matrix $D(I - \gamma P_\pi)$ is not guaranteed to be negative semi-definite (see Sutton et al. (2016)). Motivated by GTD methods, we seek an approximate solution m that satisfies $m = \Pi\hat{\mathcal{T}}m$ via minimizing a projected objective $||\Pi\bar{\delta}_w||_D^2$, where $\bar{\delta}_w \doteq \hat{\mathcal{T}}(Xw) - Xw$. For reasons that will soon be clear, we also include ridge regularization, yielding the objective

$$J^{m_\pi}(w) \doteq \tfrac{1}{2}||\Pi\bar{\delta}_w||_D^2 + \tfrac{1}{2}\eta||w||^2, \tag{3}$$

where $\eta > 0$ is the weight of the ridge term. We can now compute $\nabla_w J^{m_\pi}(w)$ following a similar routine as Sutton et al. (2009a). When sampling $\nabla_w J^{m_\pi}(w)$, we use another set of parameters $\kappa \in \mathbb{R}^{K_1}$ to address the double sampling issue as proposed by Sutton et al. (2009a). See Sutton et al. (2009a) for details of the derivation. This derivation, however, provides only an intuition behind GEM and has little to do with the actual convergence proof for two reasons. First, in an actor-critic setting, π keeps changing,

as does J^{m_π}. Second, we consider sequential Markovian data $\{S_0, A_0, S_1, \dots\}$. The proof in Sutton et al. (2009a) assumes i.i.d. data, i.e., each state S_t is sampled from d_μ independently. Compared with the i.i.d. assumption, the Markovian assumption is more practical in RL problems. We now present the GEM algorithm, which updates κ and w recursively as

GEM:

$$\begin{aligned}\bar{\delta}_t &\leftarrow i(S_{t+1}) + \gamma\rho_t x_t^\top w_t - x_{t+1}^\top w_t,\\ \kappa_{t+1} &\leftarrow \kappa_t + \alpha_t(\bar{\delta}_t - x_{t+1}^\top \kappa_t)x_{t+1},\\ w_{t+1} &\leftarrow w_t + \alpha_t\big((x_{t+1} - \gamma\rho_t x_t)x_{t+1}^\top \kappa_t - \eta w_t\big),\end{aligned} \tag{4}$$

where $\eta > 0$ is a constant, α_t is a deterministic sequence satisfying the Robbins-Monro condition (Robbins & Monro, 1951), i.e., $\{\alpha_t\}$ is non-increasing positive and $\sum_t \alpha_t = \infty, \sum_t \alpha_t^2 < \infty$. Similar to Sutton et al. (2009a), we define $d_t^\top \doteq [\kappa_t^\top, w_t^\top]$ and rewrite the GEM update as

$$d_{t+1} = d_t + \alpha_t(h(Y_t) - G_{\theta_t}(Y_t)d_t),$$

where $Y_t \doteq (S_t, A_t, S_{t+1})$. With $y \doteq (s, a, s')$, we define

$$\begin{aligned}A_\theta(y) &\doteq x(s')(x(s') - \gamma\rho_\theta(s,a)x(s))^\top,\\ C(y) &\doteq x(s')x(s')^\top,\\ G_\theta(y) &\doteq \begin{bmatrix} C(y) & A_\theta(y) \\ -A_\theta(y)^\top & \eta I \end{bmatrix}, h(y) \doteq \begin{bmatrix} i(s')x(s') \\ 0 \end{bmatrix}.\end{aligned}$$

Let $d_\mathcal{Y}(y) \doteq d_\mu(s)\mu(a|s)p(s'|s,a)$, the limiting behavior of GEM is then governed by

$$\begin{aligned}A(\theta) &\doteq \mathbb{E}_{d_\mathcal{Y}}[A_\theta(y)] = X^\top(I - \gamma P_\theta^\top)DX,\\ \bar{G}(\theta) &\doteq \mathbb{E}_{d_\mathcal{Y}}[G_\theta(y)] = \begin{bmatrix} C & A(\theta) \\ -A(\theta)^\top & \eta\mathbf{I} \end{bmatrix},\\ \bar{h} &\doteq \mathbb{E}_{d_\mathcal{Y}}[h(y)] = \begin{bmatrix} X^\top Di \\ 0 \end{bmatrix}.\end{aligned} \tag{5}$$

Readers familiar with GTD2 (Sutton et al., 2009a) may find that the $\bar{G}(\theta)$ in Eq (5) is different from its counterpart in GTD2 in that the bottom right block of $\bar{G}(\theta)$ is ηI while that block in GTD2 is 0. This ηI results from the ridge regularization in the objective $J^{m_\pi}(w)$, and this block has to be strictly positive definite[6] to ensure the positive definiteness of $\bar{G}(\theta)$. In general, any regularization in the form of $||w_t||_\Xi^2$ is sufficient. We assume the ridge to simplify notation.

As we consider an actor-critic setting where the policy θ is changing every step, we pose the following condition on the changing rate of θ:

[6]In this paper, by positive definiteness for an asymmetric square matrix X, we mean there exists a constant $\epsilon > 0$ such that $\forall y, y^\top Xy \geq \epsilon||y||^2$.

Condition 1. *(Assumption 3.1(3) in Konda (2002)) The random sequence $\{\theta_t\}$ satisfies $||\theta_{t+1} - \theta_t|| \leq \beta_t H_t$, where $\{H_t\}$ is some nonnegative process with bounded moments and $\{\beta_t\}$ is a nonincreasing deterministic sequence satisfying the Robbins-Monro condition such that $\sum_t(\frac{\beta_t}{\alpha_t})^d < \infty$ for some $d > 0$.*

When we consider a policy evaluation setting where θ is fixed, this condition is satisfied automatically. We show later that this condition is also satisfied in COF-PAC. We now characterize the asymptotic behavior of GEM.

Theorem 1. *(Convergence of GEM) Under Assumptions (1, 2) and Condition 1, the iterate $\{d_t\}$ generated by (4) satisfies $\sup_t ||d_t|| < \infty$ and $\lim_{t\to\infty} ||\bar{G}(\theta_t)d_t - \bar{h}|| = 0$ almost surely.*

Lemma 3. *Under Assumptions (1,2), when $\eta > 0$, $\bar{G}(\theta)$ is nonsingular and $\sup_\theta ||\bar{G}(\theta)^{-1}|| < \infty$.*

By simple block matrix inversion, Theorem 1 implies

$$\begin{aligned}&\lim_{t\to\infty} ||w_{\theta_t}^*(\eta) - w_t|| = 0, \text{where}\\ &w_\theta^*(\eta) \doteq (A(\theta)^\top C^{-1}A(\theta) + \eta I)^{-1}A(\theta)^\top C^{-1}X^\top Di.\end{aligned}$$

Konda (2002) provides a general theorem for stochastic approximation algorithms to track a slowly changing linear system. To prove Theorem 1, we verify that GEM indeed satisfies all the assumptions (listed in the appendix) in Konda's theorem. Particularly, that theorem requires $\bar{G}(\theta)$ to be strictly positive definite, which is impossible if $\eta = 0$. This motivates the introduction of the ridge regularization in J^{m_π} defined in Eq. (3). *Namely, the ridge regularization is essential in the convergence of GEM under a slowly changing target policy.* Introducing regularization in the GTD objective is not new. Mahadevan et al. (2014) introduce the proximal GTD learning framework to integrate GTD algorithms with first-order optimization-based regularization via saddle-point formulations and proximal operators. Yu (2017) introduces a general regularization term for improving robustness. Du et al. (2017) introduce ridge regularization to improve the convexity of the objective. However, their analysis is conducted with the saddle-point formulation of the GTD objective (Liu et al., 2015; Macua et al., 2015) and requires a fixed target policy, which is impractical in our control setting. We are the first to establish the tracking ability of GTD-style algorithms under a slowly changing target policy by introducing ridge regularization, which ensures the driving term $\bar{G}(\theta)$ is strictly positive definite. Without this ridge regularization, we are not aware of any existing work establishing this tracking ability. Note our arguments do not apply when $\eta = 0$ and π is changing, which is an open problem. However, if $\eta = 0$ and π is fixed, we can use arguments from Yu (2017) to prove convergence. In this scenario, assuming $A(\theta)$ is nonsingular, $\{w_t\}$ converges to $w_\theta^*(0) = A(\theta)^{-1}X^\top Di$ and we have

Proposition 2. $\Pi\hat{\mathcal{T}}(Xw_\theta^*(0)) = Xw_\theta^*(0)$.

Similarly, we introduce ridge regularization in the q-value analogue of GTD2, which we call GQ2. GQ2 updates u recursively as

GQ2:

$$\begin{aligned}
\delta_t &\leftarrow R_{t+1} + \gamma\rho_{t+1}\tilde{x}_{t+1}^\top u_t - \tilde{x}_t^\top u_t,\\
\tilde{\kappa}_{t+1} &\leftarrow \tilde{\kappa}_t + \alpha_t(\delta_t - \tilde{x}_t^\top \tilde{\kappa}_t)\tilde{x}_t,\\
u_{t+1} &\leftarrow u_t + \alpha_t\big((\tilde{x}_t - \gamma\rho_{t+1}\tilde{x}_{t+1})\tilde{x}_t^\top \tilde{\kappa}_t - \eta u_t\big). \quad (6)
\end{aligned}$$

Similarly, we define $\tilde{d}_t^\top \doteq [\tilde{\kappa}_t^\top, u_t^\top]$,

$$\tilde{A}(\theta) \doteq \tilde{X}^\top \tilde{D}(I - \gamma\tilde{P}_\theta)\tilde{X},$$
$$\tilde{G}(\theta) \doteq \begin{bmatrix} \tilde{C} & \tilde{A}(\theta) \\ -\tilde{A}(\theta)^\top & \eta I \end{bmatrix}, \tilde{h} \doteq \begin{bmatrix} \tilde{X}^\top \tilde{D}\tilde{r} \\ 0 \end{bmatrix}.$$

Theorem 2. *(Convergence of GQ2) Under Assumptions (1, 2) and Condition 1, the iterate $\{\tilde{d}_t\}$ generated by (6) satisfies $\sup_t ||\tilde{d}_t|| < \infty$ and $\lim_{t\to\infty} ||\tilde{G}(\theta_t)\tilde{d}_t - \tilde{h}|| = 0$ almost surely.*

Similarly, we have

$$\lim_{t\to\infty} ||u_{\theta_t}^*(\eta) - u_t|| = 0, \text{where}$$
$$u_\theta^*(\eta) \doteq (\tilde{A}(\theta)^\top \tilde{C}^{-1}\tilde{A}(\theta) + \eta I)^{-1}\tilde{A}(\theta)^\top \tilde{C}^{-1}\tilde{X}^\top \tilde{D}\tilde{r}.$$

Comparing the update rules of GEM and GQ2, it now becomes clear that GEM is "reversed" GQ2. In particular, the $A(\theta)$ in GEM is the "transpose" of the $\tilde{A}(\theta)$ in GQ2. Such reversed TD methods have been explored by Hallak & Mannor (2017); Gelada & Bellemare (2019), both of which rely on the operator $D^{-1}P_\pi^\top D$ introduced by Hallak & Mannor (2017). Previous methods implement this operator under the semi-gradient paradigm (Sutton, 1988). By contrast, GEM is a full gradient. The techniques in GEM can be applied immediately to the discounted COP-TD (Gelada & Bellemare, 2019) to improve its convergence from a small enough $\hat{\gamma}$ to any $\hat{\gamma} < 1$. Applying GEM-style update to COP-TD (Hallak & Mannor, 2017) is still an open problem as COP-TD involves a nonlinear projection, whose gradient is hard to compute.

4. Convergent Off-Policy Actor-Critic

To estimate $\nabla J(\theta)$, we use GEM and GQ2 to estimate m_π and q_π respectively, yielding COF-PAC (Algorithm 1). In COF-PAC, we require both $\{\alpha_t\}$ and $\{\beta_t\}$ to be deterministic and nonincreasing and satisfy the Robbins-Monro condition. Furthermore, there exists some $d > 0$ such that $\sum_t (\frac{\beta_t}{\alpha_t})^d < \infty$. These are common stepsize conditions in two-timescale algorithms (see Borkar (2009)). Like Konda (2002), we also use adaptive stepsizes $\Gamma_1 : \mathbb{R}^{K_1} \to \mathbb{R}$ and $\Gamma_2 : \mathbb{R}^{K_2} \to \mathbb{R}$ to ensure θ changes slowly enough. We now pose the same condition on $\Gamma_i (i = 1, 2)$ as Konda (2002). There exist constants C_1, C_2 satisfying $0 < C_1 < C_2 < \infty$ such that for any vector $d, \bar{d}$, the following properties hold: $||d||\Gamma_i(d) \in [C_1, C_2]$, $|\Gamma_i(d) - \Gamma_i(\bar{d})| \leq \frac{C_2||d-\bar{d}||}{1+||d||+||\bar{d}||}$. Konda (2002) provides an example for Γ_i. Let $C_0 > 0$ be some constant, then we define Γ_i as $\Gamma_i(d) = \mathbb{I}_{||d||<C_0} + \frac{1+C_0}{1+||d||}\mathbb{I}_{||d||\geq C_0}$, where $\mathbb{I}$ is the indicator function. It is easy to verify that the above conditions on stepsizes $(\alpha_t, \beta_t, \Gamma_1, \Gamma_2)$, together with Assumptions (1,3), ensure that $\Gamma_1(w_t)\Gamma_2(u_t)\Delta_t$ is bounded. Condition 1 on the policy changing rate, therefore, indeed holds. Consequently, Theorems 1 and 2 hold when the target policy π is updated according to COF-PAC.

Algorithm 1 COF-PAC

Ensure: $\eta > 0$
Initialize $w_0, \kappa_0, u_0, \tilde{\kappa}_0, \theta_0$
$t \leftarrow 0$
Get S_0, A_0
while True **do**
  Execute A_t, get R_{t+1}, S_{t+1}
  Sample $A_{t+1} \sim \mu(\cdot|S_{t+1})$
  $\bar{\delta}_t \leftarrow i(S_{t+1}) + \gamma\rho_t x_t^\top w_t - x_{t+1}^\top w_t$
  $\kappa_{t+1} \leftarrow \kappa_t + \alpha_t(\bar{\delta}_t - x_{t+1}^\top \kappa_t)x_{t+1}$
  $w_{t+1} \leftarrow w_t + \alpha_t\big((x_{t+1} - \gamma\rho_t x_t)x_{t+1}^\top \kappa_t - \eta w_t\big)$
  $\delta_t \leftarrow R_{t+1} + \gamma\rho_{t+1}\tilde{x}_{t+1}^\top u_t - \tilde{x}_t^\top u_t$
  $\tilde{\kappa}_{t+1} \leftarrow \tilde{\kappa}_t + \alpha_t(\delta_t - \tilde{x}_t^\top \tilde{\kappa}_t)\tilde{x}_t$
  $u_{t+1} \leftarrow u_t + \alpha_t\big((\tilde{x}_t - \gamma\rho_{t+1}\tilde{x}_{t+1})\tilde{x}_t^\top \tilde{\kappa}_t - \eta u_t\big)$
  $\Delta_t \leftarrow \rho_t(w_t^\top x_t)(u_t^\top \tilde{x}_t)\nabla \log \pi_\theta(A_t|S_t)$
  $\theta_{t+1} \leftarrow \theta_t + \beta_t\Gamma_1(w_t)\Gamma_2(u_t)\Delta_t$
  $t \leftarrow t + 1$
end while

We now characterize the asymptotic behavior of COF-PAC. The limiting policy update in COF-PAC is

$$\hat{g}(\theta) \doteq \sum_s d_\mu(s)\big(x(s)^\top w_\theta^*(\eta)\big)\sum_a \mu(a|s)\psi_\theta(s,a)\big(\tilde{x}(s,a)^\top u_\theta^*(\eta)\big).$$

The bias introduced by the estimates m and q is $b(\theta) \doteq \nabla J(\theta) - \hat{g}(\theta)$, which determines the asymptotic behavior of COF-PAC:

Theorem 3. *(Convergence of COF-PAC) Under Assumptions (1-3), the iterate $\{\theta_t\}$ generated by COF-PAC (Algorithm 1) satisfies $\liminf_t \Big(||\nabla J(\theta_t)|| - ||b(\theta_t)||\Big) \leq 0$, almost surely, i.e., $\{\theta_t\}$ visits any neighborhood of the set $\{\theta : ||\nabla J(\theta)|| \leq ||b(\theta)||\}$ infinitely many times.*

The proof is inspired by Konda (2002). According to Theorem 3, COF-PAC reaches the same convergence level as the canonical on-policy actor-critic (Konda, 2002). Together

with the fact that $\nabla J(\theta)$ is Lipschitz continuous and β_t is diminishing, it is easy to see θ_t will eventually remain in the neighborhood $\{\theta : ||\nabla J(\theta)|| \leq ||b(\theta)||\}$ in Theorem 3 for arbitrarily long time. When π_θ is close to μ in the sense of the following Assumption 4(a), we can provide an explicit bound for the bias $b(\theta)$. However, failing to satisfy Assumption 4 does not necessarily imply the bias is large. The bound here is indeed loose and is mainly to provide an intuition for the source of the bias.

Assumption 4. *(a) The following two matrices are positive semidefinite :*

$$F_\theta \doteq \begin{bmatrix} C & X^\top P_\theta^\top D X \\ X^\top D P_\theta X & C \end{bmatrix},$$
$$\tilde{F}_\theta \doteq \begin{bmatrix} \tilde{C} & \tilde{X}^\top \tilde{D} \tilde{P}_\theta \tilde{X} \\ \tilde{X}^\top \tilde{P}_\theta^\top \tilde{D} \tilde{X} & \tilde{C} \end{bmatrix}.$$

(b) $\inf_\theta |\det(A(\theta))| > 0, \inf_\theta |\det(\tilde{A}(\theta))| > 0$.
(c) The Markov chain induced by π_θ is ergodic.

Remark 2. *Part (a) is from Kolter (2011), which ensures π_θ is not too far away from μ. The non-singularity of $A(\theta)$ and $\tilde{A}(\theta)$ for each fixed θ is commonly assumed (Sutton et al., 2009a;b; Maei, 2011). In part (b), we make a slightly stronger assumption that their determinants do not approach 0 during the optimization of θ.*

Proposition 3. *Under Assumptions (1-4), let d_θ be the stationary distribution under π_θ and define $\tilde{d}_\theta(s,a) \doteq d_\theta(s)\pi_\theta(a|s), D_\theta \doteq diag(d_\theta), \tilde{D}_\theta \doteq diag(\tilde{d}_\theta)$, we have*

$$||b(\theta)||_D \leq C_0\eta + C_1 \frac{1+\gamma\kappa(D^{-\frac{1}{2}} D_\theta^{\frac{1}{2}})}{1-\gamma} ||m_{\pi_\theta} - \Pi m_{\pi_\theta}||_D + C_2 \frac{1+\gamma\kappa(\tilde{D}^{-\frac{1}{2}} \tilde{D}_\theta^{\frac{1}{2}})}{1-\gamma} ||q_{\pi_\theta} - \tilde{\Pi} q_{\pi_\theta}||_{\tilde{D}},$$

where $\kappa(\cdot)$ is the condition number of a matrix w.r.t. ℓ_2 norm and C_0, C_1, C_2 are some positive constants.

The bias $b(\theta)$ comes from the bias of both the q_π estimate and the m_π estimate. The bound of the q_π estimate follows directly from Kolter (2011). The proof from Kolter (2011), however, can not be applied to analyze the m_π estimate until Lemma 2 is established.

Compatible Features: One possible approach to eliminate the bias $b(\theta)$ is to consider *compatible features* as in the canonical on-policy actor-critic (Sutton et al., 2000; Konda, 2002). Let Ψ be a subspace and $< \cdot, \cdot >_\Psi$ be an inner product, which induces a norm $||\cdot||^\Psi$. We define a projection Π_Ψ as $\Pi_\Psi y \doteq \arg\min_{\bar{y} \in \Psi} ||\bar{y} - y||^\Psi$. For any vector y and a vector $\bar{y} \in \Psi$, we have $< y - \Pi_\Psi y, \bar{y} >_\Psi = 0$ by Pythagoras. Based on this equality, Konda (2002) designs compatible features for an on-policy actor-critic. Inspired by Konda (2002), we now design compatible features for COF-PAC.

Let $\hat{m}_\theta, \hat{q}_\theta$ be estimates for m_π, q_π. With slight abuse of notations, we define

$$\hat{g}(\theta) \doteq \sum_s d_\mu(s) \hat{m}_\theta(s) \sum_a \mu(a|s) \psi_\theta(s,a) \hat{q}_\theta(s,a),$$

which is the limiting policy update. The bias $\nabla J(\theta) - \hat{g}(\theta)$ can then be decomposed as $b_1(\theta) + b_2(\theta)$, where

$$b_1(\theta) \doteq \sum_s d_\mu(s)(m_\pi(s) - \hat{m}_\theta(s))\phi_1^\theta(s),$$
$$\phi_1^\theta(s) \doteq \sum_a \mu(a|s)\psi_\theta(s,a)\hat{q}_\theta(s,a),$$
$$b_2(\theta) \doteq \sum_{s,a} d_{\mu,m}(s,a)\phi_2^\theta(s,a)(q_\pi(s,a) - \hat{q}_\theta(s,a)),$$
$$d_{\mu,m}(s,a) \doteq d_\mu(s)m_\pi(s)\mu(a|s),\ \phi_2^\theta(s,a) \doteq \psi_\theta(s,a).$$

For an $i \in [1, \ldots, K]$, we consider $\phi_{1,i}^\theta \in \mathbb{R}^{|\mathcal{S}|}$, where $\phi_{1,i}^\theta(s)$ is the i-th element of $\phi_1^\theta(s) \in \mathbb{R}^K$. Let Ψ_1 denote the subspace in $\mathbb{R}^{|\mathcal{S}|}$ spanned by $\{\phi_{1,i}^\theta\}_{i=1,\ldots,K}$. We define an inner product $< y_1, y_2 >_{\Psi_1} \doteq \sum_s d_\mu(s) y_1(s) y_2(s)$. Then we can write $b_{1,i}(\theta)$, the i-the element of $b_1(\theta)$, as

$$b_{1,i}(\theta) = < m_\pi - \hat{m}_\theta, \phi_{1,i}^\theta >_{\Psi_1}.$$

If our estimate $\hat{m}_\theta$ satisfies $\hat{m}_\theta = \Pi_{\Psi_1} m_\pi$, we have $b_1(\theta) = 0$. This motivates learning the estimate $\hat{m}_\theta$ via minimizing $J_{\Psi_1} \doteq ||\Pi_{\Psi_1} m_\pi - \hat{m}_\theta||^{\Psi_1}$. One possibility is to consider linear function approximation for $\hat{m}_\theta$ and use $\{\phi_{1,i}^\theta\}$ as features. Similarly, we consider the subspace Ψ_2 in $\mathbb{R}^{N_{sa}}$ spanned by $\{\phi_{2,i}^\theta\}$ and define the inner product according to $d_{\mu,m}$. We then aim to learn $\hat{q}_\theta$ via minimizing $J_{\Psi_2} \doteq ||\Pi_{\Psi_2} q_\pi - \hat{q}_\theta||^{\Psi_2}$. Again, we can consider linear function approximation with features $\{\phi_{2,i}^\theta\}$. In general, any feature, whose feature space contains Ψ_1 or Ψ_2, are compatible features. Due to the change of θ, compatible features usually change every time step (Konda, 2002). Note if we consider a state value critic instead of a state-action value critic, the computation of compatible features will involve the transition kernel p, to which we do not have access.

In the on-policy setting, Monte Carlo or TD(1) can be used to train a critic with compatible features (Sutton et al., 2000; Konda, 2002). In the off-policy setting, one could consider a GEM analogue of GTD(λ) (Yu, 2017) with $\lambda = 1$ to minimize J_{Ψ_1}. To minimize J_{Ψ_2}, one could consider a q-value analogue of ETD(λ) (Yu, 2015) with $\lambda = 1$. We leave the convergent analysis for those analogues under a changing target policy for future work.

5. Experiments

We design experiments to answer the following questions: (a) Can GEM approximate the emphasis as promised? (b) Can the GEM-learned emphasis boost performance compared with the followon trace? All curves are averaged over 30 independent runs. Shadowed regions indicate one standard derivation. All the implementations are made publicly available for future research.[7]

[7] https://github.com/ShangtongZhang/DeepRL

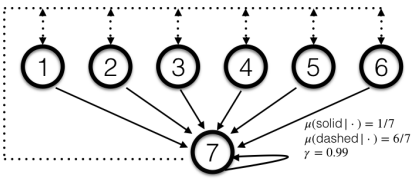

Figure 1. A variant of Baird's counterexample. This figure is adapted from Sutton & Barto (2018). The `solid` action always leads to the state 7 and a reward 0, and the `dashed` action leads to states 1 - 6 with equal probability and a reward $+1$.

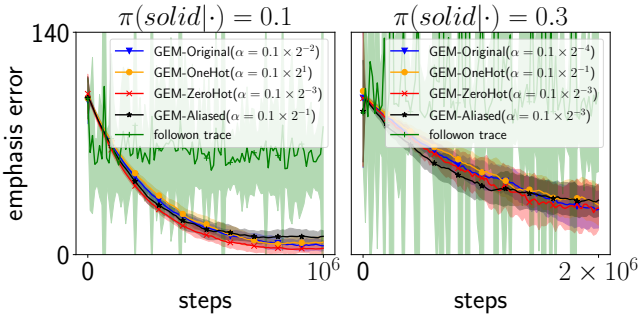

Figure 2. Averaged emphasis approximation error in last 1000 steps for the followon trace and GEM with different features. Learning rates used are bracketed.

Approximating Emphasis: We consider variants of Baird's counterexample (Baird, 1995; Sutton & Barto, 2018) as shown in Figure 1. There are two actions and the behavior policy μ always chooses the `dashed` action with probability $\frac{6}{7}$. The initial state is chosen from all the states with equal probability, and the interest i is 1 for all states. We consider four different sets of features: original features, one-hot features, zero-hot features, and aliased features. Original features are the features used by Sutton & Barto (2018), where the feature for each state lies in $\mathbb{R}^8$ (details in the appendix). This set of features is uncommon as in practice the number of states is usually much larger than the number of features. One-hot features use one-hot encoding, where each feature lies in $\mathbb{R}^7$, which indeed degenerates to a tabular setting. Zero-hot features are the complements of one-hot features, e.g., the feature of the state 1 is $[0, 1, 1, 1, 1, 1, 1]^\top \in \mathbb{R}^7$. The quantities of interest, e.g., m_π and v_π, can be expressed accurately under all the three sets of features. In the fourth set of features, we consider state aliasing. Namely, we still consider the original features, but now the feature of the state 7 is modified to be identical as the feature of the state 6. The last two dimensions of features then become identical for all states, and therefore we removed them, resulting in features lying in $\mathbb{R}^6$. Now the quantities of interest may not lie in the feature space.

In this section, we compare the accuracy of approximating the emphasis m_π with GEM (Eq (4)) and the followon trace (Eq (1)). We report the emphasis approximation error in Figure 2. At time step t, the emphasis approximation error is computed as $|M_t - m_\pi(S_t)|$ and $|w_t^\top x(S_t) - m_\pi(S_t)|$ for the followon trace and GEM respectively, where m_π is computed analytically, $M_{-1} = 0$, and w_0 is drawn from a unit normal distribution. For GEM, we consider a fixed learning rate α and tune it from $\{0.1 \times 2^1, \ldots, 0.1 \times 2^{-6}\}$. We consider two target policies: $\pi(\texttt{solid}|\cdot) = 0.1$ and $\pi(\texttt{solid}|\cdot) = 0.3$.

As shown in Figure 2, the GEM approximation enjoys lower variance than the followon trace approximation and has a lower approximation error under all four sets of features. Interestingly, when the original features are used, the C matrix is indeed singular, which violates Assumption 2. However, the algorithm does not diverge. This may suggest that the Assumption 2 can be relaxed in practice.

Policy Evaluation: The followon trace M_t is originally used in ETD to reweight updates (Eq (1) and Eq (2)). We compare ETD(0) with GEM-ETD(0), which updates ν as

$$\nu_{t+1} \leftarrow \nu_t + \alpha_2 \hat{M}_t \rho_t (R_{t+1} + \gamma x_{t+1}^\top \nu_t - x_t^\top \nu_t) x_t^\top,$$

where $\hat{M}_t \doteq w_t^\top x_t$ and w_t is updated according to GEM (Eq (4)) with a fixed learning rate α_1. If we assume m_π lies in the column space of X, a convergent analysis of GEM-ETD(0) is straightforward.

We consider a target policy $\pi(\texttt{solid}|\cdot) = 0.05$. We report the Root Mean Squared Value Error (*RMSVE*) at each time step during training in Figure 3. RMSVE is computed as $||v - v_\pi||_D$, where v_π is computed analytically. For ETD(0), we tune the learning rate α from $\{0.1 \times 2^0, \ldots, 0.1 \times 2^{-19}\}$. For GEM-ETD(0), we set $\alpha_1 = 0.025$ and tune α_2 in the same range as α. For both algorithms, we report the results with learning rates that minimized the area under the curve (AUC) in the solid lines in Figure 3. In our policy evaluation experiments, GEM-ETD(0) has a clear win over ETD(0) under all four sets of features. Note the AUC-minimizing learning rate for ETD(0) is usually several orders smaller than that of GEM-ETD(0), which explains why ETD(0) curves tend to have smaller variance than GEM-ETD(0) curves. When we decrease the learning rate of GEM-ETD(0) (as indicated by the red dashed lines in Figure 3), the variance of GEM-ETD(0) can be reduced, and the AUC is still smaller than that of ETD(0).

GEM-ETD is indeed a way to trade off bias and variance. If the states are heavily aliased, the GEM emphasis estimation may be heavily biased, as will GEM-ETD. We do not claim that GEM-ETD is always better than ETD. For example, when we set the target policy to $\pi(\texttt{solid}|\cdot) = 1$, there was no observable progress for both GEM-ETD(0) and ETD(0)

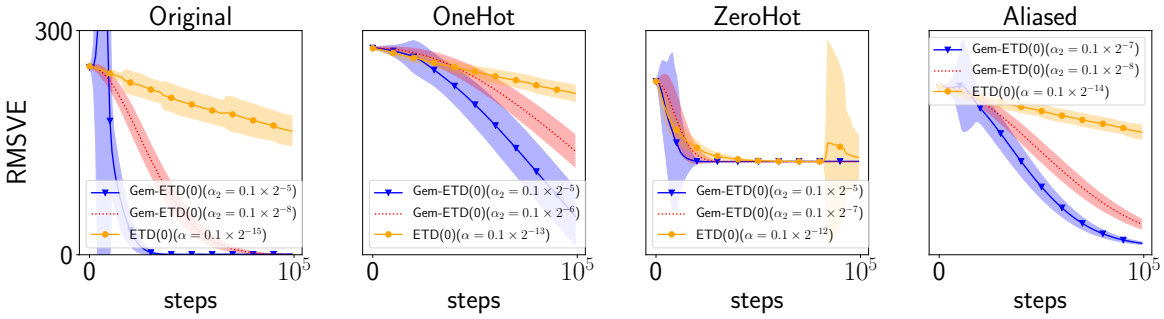

Figure 3. Averaged RMSVE in recent 1000 steps for GEM-ETD(0) and ETD(0) with four different sets of features.

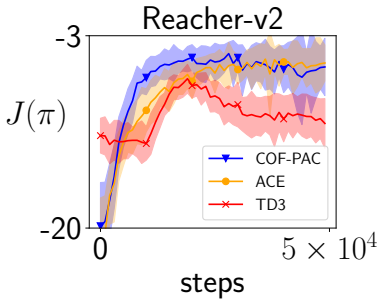

Figure 4. Comparison between ACE, COF-PAC and TD3 with a uniformly random behavior policy.

with reasonable computation resources.[8] When it comes to the bias-variance trade-off, the optimal choice is usually task-dependent. Our empirical results suggest GEM-ETD is a promising approach for this trade-off. ETD(0) is a special case of ETD(λ, β) (Hallak et al., 2016), where λ and β are used for bias-variance trade-off. Similarly, we can have GEM-ETD(λ, β) by introducing λ and β to our GEM operator $\hat{\mathcal{T}}$ analogously to ETD(λ, β). A comparison between ETD(λ, β) and GEM-ETD(λ, β) is a possibility for future work.

Control: We benchmarked COF-PAC, ACE and TD3 (Fujimoto et al., 2018) in `Reacher-v2` from OpenAI Gym (Brockman et al., 2016). Our implementation is based on Zhang et al. (2019), and we inherited their hyperparameters. Like Gelada & Bellemare (2019); Zhang et al. (2019), we consider *uniformly random* behavior policy. Neural networks are used to parameterize π, v_π, m_π. A semi-gradient version of GEM is used to train m_π inspired by the success of semi-gradient methods in large scale RL (Mnih et al., 2015). Details are provided in the appendix. We trained both algorithms for 5×10^4 steps and evaluate $J(\pi)$ every 10^3 steps. According to Figure 4, COF-PAC solves the task faster than ACE and is more stable than TD3 in this tested domain.

[8]This target policy is problematic for GEM-ETD(0) mainly because the magnitude of $\bar{\delta}_t$ in Eq (4) varies dramatically across different states, which makes the supervised learning of κ hard.

6. Related Work

Our work relies on results from Konda (2002) with two fundamental differences: (1) The work by Konda (2002) focuses on only the on-policy setting and cannot be naturally extended to the off-policy counterpart, while we work on the off-policy setting by incorporating state-of-the-art techniques such as GTD, emphatic learning, and reversed TD. (2) The learning architecture has substantial differences in that Konda (2002) considers one TD critic while we consider two GTD-style critics. As the structures of TD algorithms and GTD-style algorithms are dramatically different, applying Konda's arguments for a TD critic to our two GTD-style critics is not straightforward.

Maei (2018) proposes the Gradient Actor-Critic algorithm under a different objective, $\sum_s d_\mu(s)v(s)$, for off-policy learning with function approximation. This objective differs from the excursion objective in that it replaces the true value function v_π with an estimate v. Consequently, the optimal policy under this objective depends on the features used to approximate the value function, and this approximation of the excursion objective can be arbitrarily poor. Maei (2018) tries to show the convergence of a GTD critic under a slowly changing target policy with results from Konda (2002). In this paper, we show that GTD has to be regularized before the results from Konda (2002) can take over. Furthermore, the policy gradient estimator Maei (2018) proposes is also based on the followon trace. That estimator tracks the true gradient only in a limiting sense under a fixed π (Maei, 2018, Theorem 2) and has potentially unbounded variance, similar to how M_t tracks $m_\pi(S_t)$. It is unclear if that policy gradient estimator can track the true policy gradient under a changing π or not. To address this issue, we instead use function approximation to learn the emphasis directly.

Off-PAC has inspired the invention of many other off-policy actor-critic algorithms (e.g., off-policy DPG, DDPG, ACER,

off-policy EPG, TD3, IMPALA), all of which, like Off-PAC, ignore the emphasis and thus are not theoretically justified under function approximation. Another line of policy-based off-policy algorithms involves reward shaping via policy entropy. In particular, SBEED (Dai et al., 2017) is an off-policy actor-critic algorithm with a finite-sample analysis on the statistical error. The convergence analysis of SBEED (Theorem 5 in Dai et al. (2017)) is conducted within a bi-level optimization framework, assuming the exact solution of the inner optimization problem can be obtained. With function approximation, requiring the exact solution is usually impractical due to representation error. Even if the exact solution is representable, computing it explicitly is still expensive (c.f. solving a least-squares regression problem with a large feature matrix exactly). By contrast, our work adopts a two-timescale perspective, where we do not need to obtain the exact minimizer of $J^{m_\pi}(w)$ every step. Other works in this line of research include Nachum et al. (2017a); O'Donoghue et al. (2016); Schulman et al. (2017); Nachum et al. (2017b); Haarnoja et al. (2017; 2018), which are mainly developed in a tabular setting and do not have a convergence analysis under function approximation.

Liu et al. (2019) propose to reweight the Off-PAC update via the density ratio between π and μ. This density ratio can be learned by either Liu et al. (2018) as Liu et al. (2019) did or (discounted) COP-TD, Nachum et al. (2019a), Uehara & Jiang (2019), Zhang et al. (2020a), and GradientDICE (Zhang et al., 2020b). The convergence of those density ratio learning algorithms under a slowly changing target policy is, however, unclear. For GradientDICE with linear function approximation, it is possible to employ our arguments for proving Theorem 1 to prove its convergence under a slowly changing target policy and thus give a convergent analysis for this reweighted Off-PAC in a two-timescale form. We leave this for future work. Zhang et al. (2019) propose a new objective based on the density ratio from Gelada & Bellemare (2019), yielding Generalized Off-Policy Actor-Critic (Geoff-PAC), whose convergence is also unclear.

In concurrent work (AlgaeDICE), Nachum et al. (2019b) propose a new objective for off-policy actor-critic and reformulate the policy optimization problem into a minimax problem. Primal-dual algorithms can then take over. Nachum et al. (2019b) show the primal variable works similarly to an actor, and the dual variable works similarly to a critic. It is possible to provide a two-timescale convergent analysis for AlgaeDICE when the dual variable is linear and the primal variable is nonlinear using arguments from this paper, which we also leave for future work.

Although the actor runs at a slower timescale and the critic runs at a faster timescale, we remark that applying the two-timescale convergent arguments from Chapter 6.1 of Borkar (2009) to the actor-critic setting with function approximation is in general hard. Due to function approximation, the equilibrium of the ordinary differential equation (ODE) associated with the critic usually depends on features. Consequently, the ODE associated with the actor depends on the features as well, making it hard to analyze. To eliminate the influence of the representation error resulting from function approximation, compatible features and eligibility trace with $\lambda = 1$ seem necessary, both of which, however, do not fit well into arguments from Borkar (2009).

7. Conclusion

We have presented the first provably convergent two-timescale off-policy actor-critic with function approximation via introducing the emphasis critic and establishing the tracking ability of GTD-style algorithms under a slowly changing target policy. Future work can extend COF-PAC with non-linear critics via considering projection onto the tangent plane as Maei (2011). Conducting a finite sample analysis of COF-PAC with techniques in Zou et al. (2019); Xu et al. (2020b;a) is also a possibility for future work.

Acknowledgments

SZ is generously funded by the Engineering and Physical Sciences Research Council (EPSRC). This project has received funding from the European Research Council under the European Union's Horizon 2020 research and innovation programme (grant agreement number 637713). The experiments were made possible by a generous equipment grant from NVIDIA. BLs research is funded by the National Science Foundation (NSF) under grant NSF IIS1910794, Amazon Research Award, and Adobe gift fund.

References

Baird, L. Residual algorithms: Reinforcement learning with function approximation. *Machine Learning*, 1995.

Bertsekas, D. P. and Tsitsiklis, J. N. *Parallel and distributed computation: numerical methods*. Prentice hall Englewood Cliffs, NJ, 1989.

Borkar, V. S. *Stochastic approximation: a dynamical systems viewpoint*. Springer, 2009.

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., and Zaremba, W. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

Ciosek, K. and Whiteson, S. Expected policy gradients. *arXiv preprint arXiv:1706.05374*, 2017.

Dai, B., Shaw, A., Li, L., Xiao, L., He, N., Liu, Z., Chen, J., and Song, L. Sbeed: Convergent reinforcement learning

with nonlinear function approximation. *arXiv preprint arXiv:1712.10285*, 2017.

Degris, T., White, M., and Sutton, R. S. Off-policy actor-critic. *arXiv preprint arXiv:1205.4839*, 2012.

Du, S. S., Chen, J., Li, L., Xiao, L., and Zhou, D. Stochastic variance reduction methods for policy evaluation. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.

Espeholt, L., Soyer, H., Munos, R., Simonyan, K., Mnih, V., Ward, T., Doron, Y., Firoiu, V., Harley, T., Dunning, I., et al. Impala: Scalable distributed deep-rl with importance weighted actor-learner architectures. *arXiv preprint arXiv:1802.01561*, 2018.

Fujimoto, S., Van Hoof, H., and Meger, D. Addressing function approximation error in actor-critic methods. *arXiv preprint arXiv:1802.09477*, 2018.

Gelada, C. and Bellemare, M. G. Off-policy deep reinforcement learning by bootstrapping the covariate shift. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*, 2019.

Haarnoja, T., Tang, H., Abbeel, P., and Levine, S. Reinforcement learning with deep energy-based policies. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1352–1361. JMLR. org, 2017.

Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. *arXiv preprint arXiv:1801.01290*, 2018.

Hallak, A. and Mannor, S. Consistent on-line off-policy evaluation. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.

Hallak, A., Tamar, A., Munos, R., and Mannor, S. Generalized emphatic temporal difference learning: Bias-variance analysis. In *Proceedins of 30th AAAI Conference on Artificial Intelligence*, 2016.

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

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

Konda, V. R. *Actor-critic algorithms*. PhD thesis, Massachusetts Institute of Technology, 2002.

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

Liu, B., Liu, J., Ghavamzadeh, M., Mahadevan, S., and Petrik, M. Finite-sample analysis of proximal gradient td algorithms. In *UAI*, 2015.

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

Liu, Y., Swaminathan, A., Agarwal, A., and Brunskill, E. Off-policy policy gradient with state distribution correction. *arXiv preprint arXiv:1904.08473*, 2019.

Macua, S. V., Chen, J., Zazo, S., and Sayed, A. H. Distributed policy evaluation under multiple behavior strategies. *IEEE Transactions on Automatic Control*, 2015.

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

Maei, H. R. Convergent actor-critic algorithms under off-policy training and function approximation. *arXiv preprint arXiv:1802.07842*, 2018.

Mahadevan, S., Liu, B., Thomas, P., Dabney, W., Giguere, S., Jacek, N., Gemp, I., and Liu, J. Proximal reinforcement learning: A new theory of sequential decision making in primal-dual spaces. *arXiv preprint arXiv:1405.6757*, 2014.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. Human-level control through deep reinforcement learning. *Nature*, 2015.

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., and Kavukcuoglu, K. Asynchronous methods for deep reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning*, 2016.

Nachum, O., Norouzi, M., Xu, K., and Schuurmans, D. Bridging the gap between value and policy based reinforcement learning. In *Advances in Neural Information Processing Systems*, 2017a.

Nachum, O., Norouzi, M., Xu, K., and Schuurmans, D. Trust-pcl: An off-policy trust region method for continuous control. *arXiv preprint arXiv:1707.01891*, 2017b.

Nachum, O., Chow, Y., Dai, B., and Li, L. Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections. *arXiv preprint arXiv:1906.04733*, 2019a.

Nachum, O., Dai, B., Kostrikov, I., Chow, Y., Li, L., and Schuurmans, D. Algaedice: Policy gradient from arbitrary experience. *arXiv preprint arXiv:1912.02074*, 2019b.

O'Donoghue, B., Munos, R., Kavukcuoglu, K., and Mnih, V. Combining policy gradient and q-learning. *arXiv preprint arXiv:1611.01626*, 2016.

Robbins, H. and Monro, S. A stochastic approximation method. *The Annals of Mathematical Statistics*, 1951.

Schulman, J., Chen, X., and Abbeel, P. Equivalence between policy gradients and soft q-learning. *arXiv preprint arXiv:1704.06440*, 2017.

Silver, D., Lever, G., Heess, N., Degris, T., Wierstra, D., and Riedmiller, M. Deterministic policy gradient algorithms. In *Proceedings of the 31st International Conference on Machine Learning*, 2014.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 2016.

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

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

Sutton, R. S., McAllester, D. A., Singh, S. P., and Mansour, Y. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems*, 2000.

Sutton, R. S., Maei, H. R., Precup, D., Bhatnagar, S., Silver, D., Szepesvári, C., and Wiewiora, E. Fast gradient-descent methods for temporal-difference learning with linear function approximation. In *Proceedings of the 26th International Conference on Machine Learning*, 2009a.

Sutton, R. S., Maei, H. R., and Szepesvári, C. A convergent $o(n)$ temporal-difference algorithm for off-policy learning with linear function approximation. In *Advances in Neural Information Processing Systems*, 2009b.

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*, 2016.

Uehara, M. and Jiang, N. Minimax weight and q-function learning for off-policy evaluation. *arXiv preprint arXiv:1910.12809*, 2019.

Wang, T., Bowling, M., and Schuurmans, D. Dual representations for dynamic programming and reinforcement learning. In *2007 IEEE International Symposium on Approximate Dynamic Programming and Reinforcement Learning*, 2007.

Wang, T., Bowling, M., Schuurmans, D., and Lizotte, D. J. Stable dual dynamic programming. In *Advances in neural information processing systems*, 2008.

Wang, Z., Bapst, V., Heess, N., Mnih, V., Munos, R., Kavukcuoglu, K., and de Freitas, N. Sample efficient actor-critic with experience replay. *arXiv preprint arXiv:1611.01224*, 2016.

White, M. Unifying task specification in reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.

Xu, T., Wang, Z., and Liang, Y. Improving sample complexity bounds for actor-critic algorithms. *arXiv preprint arXiv:2004.12956*, 2020a.

Xu, T., Wang, Z., and Liang, Y. Non-asymptotic convergence analysis of two time-scale (natural) actor-critic algorithms. *arXiv preprint arXiv:2005.03557*, 2020b.

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

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

Zhang, R., Dai, B., Li, L., and Schuurmans, D. Gendice: Generalized offline estimation of stationary values. In *International Conference on Learning Representations*, 2020a.

Zhang, S., Boehmer, W., and Whiteson, S. Generalized off-policy actor-critic. In *Advances in Neural Information Processing Systems*, 2019.

Zhang, S., Liu, B., and Whiteson, S. Gradientdice: Rethinking generalized offline estimation of stationary values. In *Proceedings of the 37th International Conference on Machine Learning*, 2020b.

Zou, S., Xu, T., and Liang, Y. Finite-sample analysis for sarsa with linear function approximation. In *Advances in Neural Information Processing Systems*, 2019.