Doubly Robust Off-Policy Actor-Critic: Convergence and Optimality

Tengyu Xu [1] **Zhuoran Yang** [2] **Zhaoran Wang** [3] **Yingbin Liang** [1]

Abstract

Designing off-policy reinforcement learning algorithms is typically a very challenging task, because a desirable iteration update often involves an expectation over an on-policy distribution. Prior off-policy actor-critic (AC) algorithms have introduced a new critic that uses the density ratio for adjusting the distribution mismatch in order to stabilize the convergence, but at the cost of potentially introducing high biases due to the estimation errors of both the density ratio and value function. In this paper, we develop a doubly robust off-policy AC (DR-Off-PAC) for discounted MDP, which can take advantage of learned nuisance functions to reduce estimation errors. Moreover, DR-Off-PAC adopts a single timescale structure, in which both actor and critics are updated simultaneously with constant stepsize, and is thus more sample efficient than prior algorithms that adopt either two timescale or nested-loop structure. We study the finite-time convergence rate and characterize the sample complexity for DR-Off-PAC to attain an ϵ-accurate optimal policy. We also show that the overall convergence of DR-Off-PAC is doubly robust to the approximation errors that depend only on the expressive power of approximation functions. To the best of our knowledge, our study establishes the first overall sample complexity analysis for a single time-scale off-policy AC algorithm.

1. Introduction

In reinforcement learning (RL) (Sutton & Barto, 2018), policy gradient and its variant actor-critic (AC) algorithms have achieved enormous success in various domains such as game playing (Mnih et al., 2016), Go (Silver et al., 2016), robotic (Haarnoja et al., 2018), etc. However, these successes usually rely on the access to *on-policy* samples, i.e., samples collected online from on-policy visitation (or stationary) distribution. However, in many real-world applications, online sampling during a learning process is costly and unsafe (Gottesman et al., 2019). This necessitates the use of *off-policy* methods, which use dataset sampled from a *behavior* distribution. Since the policy gradient is expressed in the form of the on-policy expectation, it is challenging to estimate the policy gradient with off-policy samples. A common approach to implement actor-critic algorithms in the off-policy setting is to simply ignore the distribution mismatch between on- and off-policy distributions (Degris et al., 2012; Silver et al., 2014; Lillicrap et al., 2016; Fujimoto et al., 2018; Wang et al., 2016; Houthooft et al., 2018; Meuleau et al., 2001) but it has been demonstrated that such distribution mismatch can often result in divergence and poor empirical performance (Liu et al., 2019).

Several attempts have been made to correct the distribution mismatch in off-policy actor-critic's update by introducing a reweighting factor in policy update (Imani et al., 2018; Zhang et al., 2019b; Liu et al., 2019; Zhang et al., 2019c; Maei, 2018), but so far only COF-PAC (Zhang et al., 2019c) and OPPOSD (Liu et al., 2019) have been theoretically shown to converge without making strong assumptions about the estimation quality. Specifically, COF-PAC reweights the policy update with emphatic weighting approximated by a linear function, and OPPOSD reweights the policy with a density correction ratio learned by a method proposed in (Liu et al., 2018). Although both COF-PAC and OPPOSD show much promise by stabilizing the convergence, the convergence results in (Zhang et al., 2019c) and (Liu et al., 2019) indicate that both algorithms may suffer from a **large bias error** induced by estimations of both reweighting factor and value function.

The doubly robust method arises as a popular technique to reduce such a **bias error**, in which the bias vanishes as long as some (but not necessarily the full set of) estimations are accurate. Such an approach has been mainly applied to the off-policy *evaluation* problem (Tang et al., 2019; Jiang & Li, 2016; Dudík et al., 2011; 2014), and the development of such a method for solving the policy *optimization* problem is rather limited. (Huang & Jiang, 2020) derives a doubly robust policy gradient for finite-horizon Markov Decision

[1]Department of Electrical and Computer Engineering, The Ohio State University [2]Departments of Industrial Engineering & Management Sciences, Northwestern University [3] Department of Operations Research and Financial Engineering, Princeton University. Correspondence to: Tengyu Xu <xu.3260@osu.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

Process (MDP) and only for the on-policy setting. (Kallus & Uehara, 2020) proposed a doubly robust policy gradient estimator for the off-policy setting, but only for infinite-horizon averaged MDP, which does not extend easily to discounted MDP. Moreover, model-free implementation of such doubly robust policy gradient estimators typically requires the estimation of several nuisance functions via samples, but previous works proposed only methods for critic to estimate those nuisances in the finite-horizon setting, which cannot extend efficiently to the infinite-horizon setting.

Thus, our first goal is to propose a novel doubly robust policy gradient estimator for infinite-horizon discounted MDP, and further design efficient model-free critics to estimate nuisance functions so that such an estimator can be effectively incorporated to yield a doubly robust off-policy actor-critic algorithm.

On the theory side, previous work has established only the **doubly robust estimation**, i.e., the policy gradient estimator is doubly robust (Huang & Jiang, 2020; Kallus & Uehara, 2020). However, it is very unclear that by incorporating such a doubly robust estimator into an actor-critic algorithm, whether the overall convergence of the algorithm remains doubly robust, i.e., enjoys **doubly robust optimality gap**. Several reasons may eliminate such a nice property. For example, the alternating update between actor and critic does not allow critics' each estimation to be sufficiently accurate, so that doubly robust estimation may not hold at each round of iteration. Furthermore, the optimality gap of the overall convergence of an algorithm depends on interaction between critics' estimation error and actor's update error as well as other sampling variance errors, so that the double robust *estimation* does not necessarily yield the doubly robust *optimality gap*.

Thus, our second goal is to establish a finite-time convergence guarantee for our proposed algorithm, and show that the optimality gap of the overall convergence of our algorithm remains doubly robust.

1.1. Main Contributions

Doubly Robust Estimator: We propose a new method to derive a doubly robust policy gradient estimator for an infinite-horizon discounted MDP. Comparing with the previously proposed estimators that adjust only the distribution mismatch (Liu et al., 2019; Imani et al., 2018; Zhang et al., 2019b;c), our new estimator significantly reduces the bias error when two of the four nuisances in our estimator are accurate (and is hence doubly robust). We further propose a new recursive method for critics to estimate the nuisances in the infinite-horizon off-policy setting. Based on our proposed new estimator and nuisance estimation methods, we develop a model-free doubly robust off-policy actor-critic (DR-Off-PAC) algorithm.

Doubly Robust Optimality Gap: We provide the finite-time convergence analysis for our DR-Off-PAC algorithm with single timescale updates. We show that DR-Off-PAC is guaranteed to converge to the optimal policy, and the optimality gap of the overall convergence is also doubly robust to the approximation errors. This result is somewhat surprising, because the doubly robust policy gradient update suffers from both non-vanishing optimization error and approximation error at each iteration, whereas the double robustness of the optimality gap is independent of the optimization error. This also indicates that we can improve the optimality gap of DR-Off-PAC by adopting a powerful function class to estimate certain nuisance functions.

Our work is the first that characterizes the doubly robust optimality gap for the overall convergence of off-policy actor-critic algorithms, for which we develop new tools for analyzing actor-critic and critic-critic error interactions.

1.2. Related Work

The first off-policy actor-critic algorithm is proposed in (Degris et al., 2012) as Off-PAC, and has inspired the invention of many other off-policy actor-critic algorithms such as off-policy DPG (Silver et al., 2014), DDPG (Lillicrap et al., 2016), TD3 (Fujimoto et al., 2018), ACER (Wang et al., 2016), and off-policy EPG (Houthooft et al., 2018), etc, all of which have the distribution mismatch between the sampling distribution and visitation (or stationary) distribution of updated policy, and hence are not provably convergent under function approximation settings.

In one line of studies, off-policy design adopts reward shaping via entropy regularization and optimizes over a different objective function that does not require the knowledge of behaviour sampling (Haarnoja et al., 2018; O'Donoghue et al., 2016; Dai et al., 2018; Nachum et al., 2017; 2018; Schulman et al., 2017; Haarnoja et al., 2017; Tosatto et al., 2020). Although the issue of distribution mismatch is avoided for this type of algorithms, they do not have convergence guarantee in general settings. The distribution mismatch issue is also avoided in a gradient based algorithm AlgaeDICE (Nachum et al., 2019), in which the original problem is reformulated into a minimax problem. However, since non-convex minimax objective is in general difficult to optimize, the convergence of AlgaeDICE is not clear.

In another line of works, efforts have been made to address the issue of distribution mismatch in Off-PAC. (Imani et al., 2018) developed actor-critic with emphatic weighting (ACE), in which the convergence of Off-PAC is ameliorated by using emphatic weighting (Sutton et al., 2016). Inspired by ACE and the density ratio in (Gelada & Bellemare, 2019), (Zhang et al., 2019b) proposed Geoff-PAC to optimize a new objective. Based on Geoff-PAC, (Lyu et al., 2020) further applied the variance reduction technique in

(Cutkosky & Orabona, 2019) to develop a new algorithm VOMPS/ACE-STORM. However, since the policy gradient estimator with emphatic weighting is only unbiased in asymptotic sense and emphatic weighting usually suffers from unbounded variance, the convergence of ACE, Geoff-PAC and VOMPS/ACE-STORM are in general not clear. So far, only limited off-policy actor-critic algorithms have been shown to have guaranteed convergence. (Zhang et al., 2019c) proposed a provably convergent two timescale off-policy actor-critic via learning the emphatic weights with linear features, and (Liu et al., 2019) proposed to reweight the off-PAC update via learning the density ratio with the approach in (Liu et al., 2018). However, both convergence results in (Liu et al., 2019) and (Zhang et al., 2019c) suffer from bias errors of function approximation, and the two timescale update and the double-loop structure adopted in (Zhang et al., 2019c) and (Liu et al., 2019), respectively, can cause significant sample inefficiency. Recently, (Kallus & Uehara, 2020) proposed an off-policy gradient method with doubly robust policy gradient estimator. However, they also adopted an inefficient double-loop structure and the overall convergence of the algorithm with such an estimator was not shown to have the doubly robust property. In contrast to previous works, we develop a new doubly robust off-policy actor-critic that provably converges with the overall convergence also being doubly robust to the function approximation errors. Our algorithm adopts a single-timescale update scheme, and is thus more sample efficient than the previous methods (Liu et al., 2019; Zhang et al., 2019c; Kallus & Uehara, 2020).

2. Background and Problem Formulation

In this section, we introduce the background of Markov Decision Process (MDP) and problem formulation. We consider an infinite-horizon MDP described by $(\mathcal{S}, \mathcal{A}, \mathsf{P}, r, \mu_0, \gamma)$, where $\mathcal{S}$ denotes the set of states, $\mathcal{A}$ denotes the set of actions, and $\mathsf{P}(s'|s,a)$ denotes the transition probability from state $s \in \mathcal{S}$ to state s' with action $a \in \mathcal{A}$. Note that $|\mathcal{S}|$ and $|\mathcal{A}|$ can be infinite such that $\mathsf{P}(s'|s,a)$ is then a Markov kernel. Let $r(s,a,s')$ be the reward that an agent receives if the agent takes an action a at state s and the system transits to state s'. Moreover, we denote μ_0 as the distribution of the initial state $s_0 \in \mathcal{S}$ and $\gamma \in (0,1)$ as the discount factor. Let $\pi(a|s)$ be the policy which is the probability of taking action a given current state s. Then, for a given policy π, we define the state value function as $V_\pi(s) = \mathbb{E}[\gamma^t r(s_t, a_t, s_{t+1})|s_0 = s, \pi]$ and the state-action value function as $Q_\pi(s,a) = \mathbb{E}[\gamma^t r(s_t, a_t, s_{t+1})|s_0 = s, a_0 = a, \pi]$. Note that $V_\pi(s) = \mathbb{E}_\pi[Q_\pi(s,a)|s]$ and $Q_\pi(s,a)$ satisfies the following Bellman equation:

$$Q_\pi(s,a) = R(s,a) + \gamma \mathcal{P}_\pi Q_\pi(s,a), \tag{1}$$

where $R(s,a) = \mathbb{E}[r(s,a,s')|s,a]$ and

$$\mathcal{P}_\pi Q_\pi(s,a) \coloneqq \mathbb{E}_{s' \sim \mathsf{P}(\cdot|s,a), a' \sim \pi(\cdot|s')}[Q_\pi(s', a')].$$

We further define the expected total reward function as $J(\pi) = (1-\gamma)\mathbb{E}[\gamma^t r(s_t, a_t, s_{t+1})|s_0 \sim \mu_0, \pi] = \mathbb{E}_{\mu_0}[V_\pi(s)] = \mathbb{E}_{\nu_\pi}[r(s,a,s')]$, where $\nu_\pi(s,a) = (1-\gamma)\sum_{t=0}^{\infty} \gamma^t \mathsf{P}(s_t = s, a_t = a|s_0 \sim \mu_0, \pi)$ is the visitation distribution. The visitation distribution satisfies the following "inverse" Bellman equation:

$$\nu_\pi(s', a') = \pi(a'|s')[(1-\gamma)\mu_0(s') + \gamma \int_{(s,a)} \mathsf{P}(s'|s,a)\nu_\pi(s,a) ds da]. \tag{2}$$

In policy optimization, the agent's goal is to find an optimal policy π^* that maximizes $J(\pi)$, i.e., $\pi^* = \operatorname{argmax}_\pi J(\pi)$. We consider the setting in which policy π is parametrized by $w \in \mathbb{R}^d$. Then, the policy optimization is to solve the problem $\max_w J(\pi_w)$. In the sequel we write $J(\pi_w) := J(w)$ for notational simplicity. A popular approach to solve such a maximization problem is the policy gradient method, in which we update the policy in the gradient ascent direction as $w_{t+1} = w_t + \alpha \nabla_w J(w_t)$. A popular form of $\nabla_w J(w)$ is derived by (Sutton et al., 2000) as

$$\nabla_w J(w) = \mathbb{E}_{\nu_{\pi_w}}[Q_{\pi_w}(s,a) \nabla_w \log \pi_w(a|s)]. \tag{3}$$

In the on-policy setting, many works adopt the policy gradient formulation in eq. (3) to estimate $\nabla_w J(w)$, which requires sampling from the visitation distribution ν_{π_w} and Monte Carlo rollout from policy π_w to estimate the value function $Q_{\pi_w}(s,a)$ (Zhang et al., 2019a; Xiong et al., 2020).

In this paper we focus on policy optimization in the behavior-agnostic off-policy setting. Specifically, we are given access to samples from a fixed distribution $\{(s_i, a_i, r_i, s_i')\} \sim \mathcal{D}_d$, where the state-action pair (s_i, a_i) is sampled from an ***unknown*** distribution $d(\cdot) : \mathcal{S} \times \mathcal{A} \to [0,1]$, the successor state s_i' is sampled from $\mathsf{P}(\cdot|s_i, a_i)$ and r_i is the received reward. We also have access to samples generated from the initial distribution, i.e., $s_{0,i} \sim \mu_0$. In the behavior-agnostic off-policy setting, it is difficult to estimate $\nabla_w J(w)$ directly with the form in eq. (3), as neither ν_{π_w} nor Monte Carlo rollout sampling is accessible. Thus, our goal is to develop an *efficient* algorithm to estimate $\nabla_w J(w)$ with off-policy samples from $\mathcal{D}_d$, and furthermore, establish the convergence guarantee for our proposed algorithm.

3. DR-Off-PAC: Algorithm and Convergence

In this section, we first develop a new doubly robust policy gradient estimator and then design a new doubly robust off-policy actor-critic algorithm.

3.1. Doubly Robust Policy Gradient Estimator

In this subsection, we construct a new doubly robust policy gradient estimator for an infinite-horizon discounted MDP. We first denote the density ratio as $\rho_{\pi_w} = \nu_{\pi_w}(s,a)/d(s,a)$, and denote the derivative of Q_{π_w} and ρ_{π_w} as $d^q_{\pi_w}$ and $d^\rho_{\pi_w}$, respectively.

Previous constructions (Kallus & Uehara, 2020) for such an estimator directly combine the policy gradient with a number of error terms under various filtrations to guarantee the double robustness. Such a method does not appear to extend easily to the discounted MDP. Specifically, the method in (Kallus & Uehara, 2020) considers finite-horizon MDP with $\gamma = 1$, and further extends their result to infinite-horizon average-reward MDP. Their extension relies on the fact that the objective function $J(w)$ in average-reward MDP is independent of the initial distribution μ_0. In contrast, $J(w)$ in discounted-reward MDP depends on μ_0. Any direct extension necessarily results in a bias due to the lack of the initial distribution, which is unknown a priori, and hence loses the doubly robust property.

To derive a doubly robust gradient estimator in the discounted MDP setting, we first consider a bias reduced estimator of the objective $J(w)$ with off-policy sample (s,a,r,s') and s_0, and then take the derivative of such an estimator to obtain a doubly robust policy gradient estimator. The idea behind this derivation is that as long as the objective estimator has small bias, the gradient of such an estimator can also have small bias. More detailed discussion can be referred to the supplement material.

Given sample $s_0 \sim \mu_0(\cdot)$ and $(s,a,r,s') \sim \mathcal{D}_d$ and estimators $\hat{Q}_{\pi_w}$, $\hat{\rho}_{\pi_w}$, $\hat{d}^q_{\pi_w}$ and $\hat{d}^\rho_{\pi_w}$, our constructed doubly robust policy gradient error is given as follows.

$$
\begin{aligned}
&G_{\text{DR}}(w) \\
&= (1-\gamma)\Big(\hat{Q}_{\pi_w}(s_0,a_0)\nabla_w \log \pi_w(a_0|s_0) + \hat{d}^q_{\pi_w}(s_0,a_0)\Big) \\
&+ \hat{d}^\rho_{\pi_w}(s,a)\left(r(s,a,s') - \hat{Q}_{\pi_w}(s,a) + \gamma \hat{Q}_{\pi_w}(s',a')\right) \\
&+ \hat{\rho}_{\pi_w}(s,a)\Big[-\hat{d}^q_{\pi_w}(s,a) \\
&+ \gamma\Big(\hat{Q}_{\pi_w}(s',a')\nabla_w \log \pi_w(a'|s') + \hat{d}^q_{\pi_w}(s',a')\Big)\Big], \qquad (4)
\end{aligned}
$$

where $a_0 \sim \pi_w(\cdot|s_0)$ and $a' \sim \pi_w(\cdot|s')$. The following theorem establishes that our proposed estimator G_{DR} satisfies the doubly robust property.

Theorem 1. *The bias error of estimator $G_{DR}(w)$ in eq. (4) satisfies*

$$
\begin{aligned}
&\mathbb{E}[G_{DR}(w)] - \nabla_w J(w) \\
&= -\mathbb{E}[\varepsilon_\rho(s,a)\varepsilon_{d^q}(s,a)] - \mathbb{E}[\varepsilon_{d^\rho}(s,a)\varepsilon_q(s,a)] \\
&\quad + \gamma\mathbb{E}[\varepsilon_\rho(s,a)\varepsilon_q(s',a')\nabla_w \log(a'|s')] \\
&\quad + \gamma\mathbb{E}[\varepsilon_\rho(s,a)\varepsilon_{d^q}(s',a')] + \gamma\mathbb{E}[\varepsilon_\rho(s,a)\varepsilon_q(s',a')],
\end{aligned}
$$

where the estimation errors are defined as

$$
\begin{aligned}
&\varepsilon_\rho = \rho_{\pi_w} - \hat{\rho}_{\pi_w}, \quad \varepsilon_q = Q_{\pi_w} - \hat{Q}_{\pi_w}, \\
&\varepsilon_{d^\rho} = d^\rho_{\pi_w} - \hat{d}^\rho_{\pi_w}, \quad \varepsilon_{d^q} = d^q_{\pi_w} - \hat{d}^q_{\pi_w}.
\end{aligned}
$$

Theorem 1 shows that the estimation error of $G_{\text{DR}}(w)$ takes a **multiplicative** form of pairs of individual estimation errors rather than the summation over all errors. Such a structure thus exhibits a **three-way doubly robust** property. Namely, as long as **one** of the three pairs $(\hat{\rho}_{\pi_w}, \hat{d}^\rho_{\pi_w})$, $(\hat{Q}_{\pi_w}, \hat{d}^q_{\pi_w})$, $(\hat{\rho}_{\pi_w}, \hat{Q}_{\pi_w})$ are accurately estimated, our estimator $G_{\text{DR}}(w)$ is unbiased, i.e., $\mathbb{E}[G_{\text{DR}}(w)] - \nabla_w J(w) = 0$. There is no need for all of the individual errors to be small.

3.2. Estimation of Nuisance Functions

In order to incorporate the doubly robust estimator eq. (4) into an actor-critic algorithm, we develop critics to respectively construct efficient estimators $\hat{Q}_{\pi_w}$, $\hat{\rho}_{\pi_w}$, $\hat{d}^q_{\pi_w}$, $\hat{d}^\rho_{\pi_w}$ in $G_{\text{DR}}(w)$ in the linear function approximation setting.

Critic I: Value function $\hat{Q}_{\pi_w}$ and density ratio $\hat{\rho}_{\pi_w}$. In the off-policy evaluation problem, (Yang et al., 2020) shows that the objective function $J(w)$ can be expressed by the following primal linear programming (LP):

$$
\begin{aligned}
&\min_{Q_{\pi_w}} \quad (1-\gamma)\mathbb{E}_{\mu_0 \pi_w}[Q_{\pi_w}(s,a)] \\
&\text{s.t.,} \quad Q_{\pi_w}(s,a) = R(s,a) + \gamma \mathcal{P}_\pi Q_{\pi_w}(s,a),
\end{aligned}
$$

with the corresponding dual LP given by

$$
\begin{aligned}
&\max_{\nu_{\pi_w}} \quad \mathbb{E}_{\nu_{\pi_w}}[R(s,a)] \\
&\text{s.t.,} \quad \nu_\pi(s',a') = (1-\gamma)\mu_0(s')\pi(a'|s') + \gamma \mathcal{P}^*_\pi \nu_\pi(s,a).
\end{aligned}
$$

Then, the value function $Q_{\pi_w}(s,a)$ and the distribution correction ratio $\rho_{\pi_w}(s,a)$ can be learned by solving the following regularized Lagrangian:

$$
\begin{aligned}
&\min_{\hat{\rho}_{\pi_w} \geq 0} \max_{\hat{Q}_{\pi_w}, \eta} L(\hat{\rho}_{\pi_w}, \hat{Q}_{\pi_w}, \eta) \\
&:= (1-\gamma)\mathbb{E}_{\mu_0}[\hat{Q}_{\pi_w}(s,a)] + \mathbb{E}_{\mathcal{D}_d}[\rho_{\pi_w}(s,a)(r(s,a,s') \\
&\quad + \gamma \hat{Q}_{\pi_w}(s',a') - \hat{Q}_{\pi_w}(s,a))] - \frac{1}{2}\mathbb{E}_{\mathcal{D}_d}[\hat{Q}_{\pi_w}(s,a)^2] \\
&\quad + \mathbb{E}_{\mathcal{D}_d}[\eta \hat{\rho}_{\pi_w}(s,a) - \eta] - 0.5\eta^2. \qquad (5)
\end{aligned}
$$

We construct $\hat{\rho}_{\pi_w}$ and $\hat{Q}_{\pi_w}$ with linearly independent feature $\phi(s,a) \in \mathbb{R}^{d_1}$: $\hat{\rho}_{\pi_w}(s,a) = \phi(s,a)^\top \theta_\rho$ and $\hat{Q}_{\pi_w}(s,a) = \phi(s,a)^\top \theta_q$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$. In such a case, $L(\rho_{\pi_w}, Q_{\pi_w}, \eta)$ is strongly-concave in both θ_q and η, and convex in θ_ρ. We denote the global optimum of $L(\theta_\rho, \theta_q, \eta)$ as $\theta^*_{\rho,w}$, $\theta^*_{q,w}$ and η^*_w. The errors of approximating Q_{π_w} and ρ_{π_w} with estimators $\hat{Q}_{\pi_w}(s,a,\theta^*_{q,w}) = \phi(s,a)^\top \theta^*_{q,w}$ and $\hat{\rho}_{\pi_w}(s,a,\theta^*_{q,w}) = \phi(s,a)^\top \theta^*_{\rho,w}$, respectively, are defined as

$$
\epsilon_q = \max\Big\{ \max_w \sqrt{\mathbb{E}_{\mathcal{D}}[(\hat{Q}_{\pi_w}(s,a,\theta^*_{q,w}) - Q_{\pi_w}(s,a))^2]},
$$

$$\max_w \sqrt{\mathbb{E}_{\mathcal{D}\cdot\pi_w}[(\hat{Q}_{\pi_w}(s',a',\theta^*_{q,w}) - Q_{\pi_w}(s',a'))^2]}\Big\},$$
$$\epsilon_\rho = \max_w \sqrt{\mathbb{E}_{\mathcal{D}}[(\hat{\rho}_{\pi_w}(s,a,\theta^*_{\rho,w}) - \rho_{\pi_w}(s,a))^2]}.$$

To solve the minimax optimization problem in eq. (5), we adopt stochastic gradient descent-ascent method with mini-batch samples $\mathcal{B}_t = \{(s_i, a_i, r_i, s_i')\}_{i=1\cdots N} \sim \mathcal{D}_d$, $a_i' \sim \pi_{w_t}(\cdot|s_i')$ and $\mathcal{B}_{t,0} = \{(s_{0,i})\}_{i=1\cdots N} \sim \mu_0$, $a_{0,i} \sim \pi_{w_t}(\cdot|s_{0,i}')$, which update parameters recursively as follows

$$\begin{aligned}
\delta_{t,i} &= (1-\gamma)\phi_{0,i} + \gamma\phi_i^\top\theta_{\rho,t}\phi_i' - \phi_i^\top\theta_{\rho,t}\phi_i \\
\eta_{t+1} &= \theta_{\rho,t} + \beta_1\frac{1}{N}\sum_{i\in\mathcal{B}_t}(\phi_i^\top\theta_{\rho,t} - 1 - \eta_t) \\
\theta_{q,t+1} &= \Gamma_{R_q}\Big[\theta_{q,t} + \beta_1\frac{1}{N}\sum_{i\in\mathcal{B}_t,\mathcal{B}_{t,0}}(\delta_{t,i} - \phi_i^\top\theta_{q,t}\phi_i)\Big] \\
\theta_{\rho,t+1} &= \Gamma_{R_\rho}\Big[\theta_{\rho,t} - \beta_1\frac{1}{N}\sum_{i\in\mathcal{B}_t}(r_i\phi_i + \gamma\phi_i'^\top\theta_{q,t}\phi_i \\
&\qquad - \phi_i^\top\theta_{q,t}\phi_i + \eta_t\phi_i)\Big],
\end{aligned} \tag{6}$$

where Γ_R indicates the projection onto a ball with radius R. Such a projection operator stabilizes the algorithm (Konda & Tsitsiklis, 2000; Bhatnagar et al., 2009). Note that the iteration in eq. (6) is similar to but difference from the GradientDICE update in (Zhang et al., 2020b), as GradientDICE can learn only the density ratio ρ_{π_w}, while our approach in eq. (6) can learn both the value function Q_{π_w} and the density ratio ρ_{π_w}.

Critic II: Derivative of value function $\hat{d}^q_{\pi_w}$. Taking derivative on both sides of eq. (1) yields

$$\begin{aligned}
d^q_{\pi_w}(s,a) &= \gamma\mathbb{E}[d^q_{\pi_w}(s',a')|s,a] \\
&\quad + \gamma\mathbb{E}[Q_{\pi_w}(s',a')\nabla_w\log\pi_w(a'|s')|s,a],
\end{aligned} \tag{7}$$

We observe that eq. (7) takes a form analogous to the Bellman equation in eq. (1), and thus suggests a recursive approach to estimate $d^q_{\pi_w}$, similarly to temporal difference (TD) learning. Specifically, suppose we estimate $d^q_{\pi_w}$ with a feature matrix $x(s,a) \in \mathbb{R}^{d_3\times d}$, i.e., $\hat{d}^q_{\pi_w}(s,a) = x(s,a)^\top\theta_{d_q}$ for all $(s,a) \in \mathcal{S}\times\mathcal{A}$. Replace $Q_{\pi_w}(s,a)$ with its estimator $\hat{Q}_{\pi_w}(s,a) = \phi(s,a)^\top\theta_q$ in eq. (7). The temporal difference error is then given as

$$\begin{aligned}
\delta_{d_q}(s,a,\theta_q) &= \gamma x(s',a')^\top\theta_{d_q} \\
&\quad + \gamma\phi(s',a')^\top\theta_q\nabla_w\log\pi_w(a'|s') - x(s,a)^\top\theta_{d_q}
\end{aligned}$$

and θ_{d^q} can be updated with the TD-like semi-gradient

$$\theta_{d^q,t+1} = \theta_{d^q,t} + \beta_2 x(s,a)\delta_{d^q}(s,a,\theta_{d^q,t}). \tag{8}$$

However, in the off-policy setting, the iteration in eq. (8) may not converge due to the off-policy sampling. To solve such an issue, we borrow the idea from gradient TD (GTD) and formulate the following strongly convex objective

$$\begin{aligned}
&H(\theta_{d_q},\theta_q) \\
&= \mathbb{E}[x(s,a)\delta_{d_q}(s,a,\theta_q)]^\top\mathbb{E}[x(s,a)\delta_{d_q}(s,a,\theta_q)].
\end{aligned}$$

We denote the global optimum of $H(\theta_{d_q},\theta^*_{q,w})$ as $\theta^*_{d_q,w}$, i.e., $H(\theta^*_{d_q,w},\theta^*_{q,w}) = 0$. The approximation error of estimating $d^q_{\pi_w}$ with estimator $\hat{d}^q_{\pi_w}(s',a',\theta^*_{d_q,w}) = x(s,a)^\top\theta^*_{d_q,w}$ is defined as

$$\epsilon_{d_q} = \max\Big\{\max_w\sqrt{\mathbb{E}_{\mathcal{D}}\left[\left\|\hat{d}^q_{\pi_w}(s,a,\theta^*_{d_q,w}) - d^q_{\pi_w}(s,a)\right\|_2^2\right]},$$
$$\max_w\sqrt{\mathbb{E}_{\mathcal{D}_d\cdot\pi_w}\left[\left\|\hat{d}^q_{\pi_w}(s',a',\theta^*_{d_q,w}) - d^q_{\pi_w}(s',a')\right\|_2^2\right]}\Big\}.$$

Similarly to GTD, we introduce an auxiliary variable w_{d^q} to avoid the issue of double sampling when using gradient based approach to minimize $H(\theta_{d_q},\theta_q)$. With mini-batch samples $\mathcal{B}_t = \{(s_i,a_i,s_i')\}_{i=1\cdots N} \sim \mathcal{D}_d$, we have the following update for θ_{d_q}.

$$\begin{aligned}
\theta_{d_q,t+1} &= \theta_{d_q,t} + \beta_3\frac{1}{N}\sum_{i\in\mathcal{B}_t}(x_i - \gamma x_i')x_i^\top w_{d_q,t}, \\
w_{d_q,t+1} &= w_{d_q,t} + \beta_3\frac{1}{N}\sum_{i\in\mathcal{B}_t}(x_i\delta_{d_q,i}(\theta_{q,t}) - w_{d_q,t}).
\end{aligned} \tag{9}$$

Critic III: Derivative of density ratio $\hat{d}^\rho_{\pi_w}$. We denote $\psi_{\pi_w}(s,a) := \nabla_w\log(\nu_{\pi_w}(s,a))$, and construct an estimator for $d^\rho_{\pi_w}$ as $\hat{d}^\rho_{\pi_w}(s,a) = \hat{\rho}_{\pi_w}(s,a)\hat{\psi}_{\pi_w}(s,a)$, where $\hat{\rho}_{\pi_w}$ and $\hat{\psi}_{\pi_w}$ are approximation of ρ_{π_w} and ψ_{π_w}, respectively. Note that eq. (2) can be rewritten in the following alternative form

$$\nu_{\pi_w}(\tilde{s}',a') = \int\pi_w(a'|\tilde{s}')\tilde{\mathsf{P}}(\tilde{s}'|s,a)\nu_{\pi_w}(s,a)dsda, \tag{10}$$

where $\tilde{\mathsf{P}}(\cdot|s,a) = (1-\gamma)\mu_0 + \gamma\mathsf{P}(\cdot|s,a)$. Taking derivative on both sides of eq. (10) and using $\nabla g(w) = g(w)\nabla\log g(w)$, we obtain

$$\begin{aligned}
&\nu_{\pi_w}(\tilde{s}',a')\psi_{\pi_w}(\tilde{s}',a') \\
&= \nabla_w\log(\pi_w(a'|\tilde{s}'))\cdot\Big[\pi_w(a'|\tilde{s}') \\
&\quad \int_{s,a}\tilde{\mathsf{P}}(\tilde{s}'|s,a)\nu_{\pi_w}(s,a)dsda\Big] \\
&\quad + \textstyle\int_{s,a}\Big[\pi_w(a'|\tilde{s}')\tilde{\mathsf{P}}(\tilde{s}'|s,a)\nu_{\pi_w}(s,a)\Big]\psi_{\pi_w}(s,a)dsda \\
&= \nabla_w\log(\pi_w(a'|\tilde{s}'))\cdot\nu_{\pi_w}(\tilde{s}',a') \\
&\quad + \textstyle\int_{s,a}\Big[\pi_w(a'|\tilde{s}')\tilde{\mathsf{P}}(\tilde{s}'|s,a)\nu_{\pi_w}(s,a)\Big]\psi_{\pi_w}(s,a)dsda \\
&= \nabla_w\log(\pi_w(a'|\tilde{s}'))\cdot\nu_{\pi_w}(\tilde{s}',a') \\
&\quad + \textstyle\int_{s,a}\nu_{\pi_w}(\tilde{s}',a')P(s,a|\tilde{s}',a')\,\psi_{\pi_w}(s,a)dsda
\end{aligned}$$

where the second equality follows because $\pi_w(a'|\tilde{s}')\int_{s,a}\tilde{\mathsf{P}}(\tilde{s}'|s,a)\nu_{\pi_w}(s,a)dsda = \nu_{\pi_w}(\tilde{s}',a')$, and the third equality follows because if $(s,a)\sim\nu_{\pi_w}(\cdot)$, then $(\tilde{s}',a')\sim\nu_{\pi_w}(\cdot)$, and Bayes' theorem implies that $\frac{\pi_w(a'|\tilde{s}')\tilde{\mathsf{P}}(\tilde{s}'|s,a)\nu_{\pi_w}(s,a)}{\nu_{\pi_w}(\tilde{s}',a')} = P(s,a|\tilde{s}',a')$. Then, dividing both sides by $\nu_{\pi_w}(\tilde{s}',a')$ yields

$$\psi_{\pi_w}(\tilde{s}',a') = \nabla_w \log(\pi_w(a'|\tilde{s}')) + \int_{s,a} P(s,a|\tilde{s}',a')\psi_{\pi_w}(s,a)dsda. \tag{11}$$

With linear function approximation, we estimate $\psi_{\pi_w}(s,a)$ with feature matrix $\varphi(s,a)\in\mathbb{R}^{d_2\times d}$ i.e., $\hat{\psi}_{\pi_w}(s,a) = \varphi(s,a)^\top\theta_\psi$ for all $(s,a)\in\mathcal{S}\times\mathcal{A}$. The temporal difference error is given as

$$\delta_\psi(\tilde{s}',a') = \nabla_w \log\pi_w(a'|\tilde{s}') + \varphi(s,a)^\top\theta_\psi - \varphi(\tilde{s}',a')^\top\theta_\psi, \tag{12}$$

Note that in eq. (12), we require $\tilde{s}'\sim\tilde{\mathsf{P}}(\cdot|s,a)$. To obtain a sample triple $(s,a,\tilde{s}')$ from such a "hybrid" transition kernel, for a given sample (s,a,s'), we take a Bernoulli choice between s' and $s_0\sim\mu_0$ with probability γ and $1-\gamma$, respectively, to obtain a state $\tilde{s}'$ that satisfies the requirement. Then, similarly to how we obtain the estimator $\hat{d}^q_{\pi_w}$, we adopt the method in GTD to formulate the following objective

$$F(\theta_\psi) = \mathbb{E}[\varphi(s',a')\delta_\psi(\tilde{s}',a')]^\top\mathbb{E}[\varphi(s',a')\delta_\rho(\tilde{s}',a')]. \tag{13}$$

We denote the global optimum of $F(\theta_\psi)$ as $\theta^*_{\psi,w}$, i.e., $F(\theta^*_{\psi,w}) = 0$, and define the approximation error of estimating $d^\rho_{\pi_w}$ with estimator $\hat{d}^\rho_{\pi_w}(s,a,\theta^*_{\rho,w},\theta^*_{\psi,w}) = \phi(s,a)^\top\theta^*_{\rho,w}\varphi(s,a)^\top\theta^*_{\psi,w}$ as

$$\epsilon_{d_\rho} = \max_w \sqrt{\mathbb{E}_{\mathcal{D}}\left[\left\|\hat{d}^\rho_{\pi_w}(s,a,\theta^*_{\rho,w},\theta^*_{\psi,w}) - d^\rho_{\pi_w}(s,a)\right\|_2^2\right]}.$$

Given mini-batch samples $\mathcal{B}_t = \{(s_i,a_i,s_i')\}_{i=1\cdots N}\sim\mathcal{D}_d$, $a_i'\sim\pi_{w_t}(\cdot|s_i')$ and $\mathcal{B}_{t,0} = \{(s_{0,i},a_{0,i})\}_{i=1\cdots N}\sim\mu_0$, we have the following update for θ_ψ:

$$\theta_{\psi,t+1} = \theta_{\psi,t} + \beta_2\frac{1}{N}\sum_{i\in\tilde{\mathcal{B}}_t}(\varphi_i' - \varphi_i)\varphi_i'^\top w_{\psi,t},$$
$$w_{\psi,t+1} = w_{\psi,t} + \beta_2\frac{1}{N}\sum_{i\in\tilde{\mathcal{B}}_t}(\varphi_i'\delta_{\psi,i} - w_{\psi,t}), \tag{14}$$

where $w_{\psi,t}$ is the auxiliary variable that we introduce to avoid the double sampling issue.

DR-Off-PAC Estimator. Given parameters $\theta_{\rho,t}$, $\theta_{q,t}$, $\theta_{\psi,t}$ and $\theta_{d_q,t}$, the doubly robust policy gradient can be obtained as follows

$$\begin{aligned} &G^i_{\text{DR}}(w_t) \\ &= (1-\gamma)\left(\phi_{0,i}^\top\theta_{q,t}\nabla_w\log\pi_w(s_{0,i},a_{0,i}) + x_{0,i}^\top\theta_{d_q,t}\right) \\ &+ \psi_i^\top\theta_{\psi,t}(r(s_i,a_i,s_i') - \phi_i^\top\theta_{q,t} + \gamma\mathbb{E}_{\pi_{w_t}}[\phi_i'^\top\theta_{q,t}]) \\ &+ \phi_i^\top\theta_{\rho,t}(-x_i^\top\theta_{d_q,t} \\ &\qquad + \gamma\phi_i^\top\theta_{q,t}\nabla_w\log\pi_w(s_{t,i},a_{t,i}) + x_i^\top\theta_{d_q,t}). \end{aligned} \tag{15}$$

Algorithm 1 DR-Off-PAC

Initialize: Policy parameter w_0, and estimator parameters $\theta_{q,0}$, $\theta_{\rho,0}$, $\theta_{d_q,0}$ and $\theta_{\psi,0}$.
for $t=0,\cdots,T-1$ **do**
  Obtain mini-batch samples $\mathcal{B}_t\sim\mathcal{D}_d$ and $\mathcal{B}_{t,0}\sim\mu_0$
  Critic I: Update density ratio and value function estimation via eq. (6): $\theta_{q,t},\theta_{\rho,t}\to\theta_{q,t+1},\theta_{\rho,t+1}$
  Critic II: Update derivative of value function estimation via eq. (9): $\theta_{d_q,t}\to\theta_{d_q,t+1}$
  Critic III: Update derivative of density ratio estimation via eq. (14): $\theta_{\psi,t}\to\theta_{\psi,t+1}$
  Actor: Update policy parameter via eq. (15)
  $w_{t+1} = w_t + \alpha\frac{1}{N}\sum_i G^i_{\text{DR}}(w_t)$
end for
Output: $w_{\hat{T}}$ with $\hat{T}$ chosen uniformly in $\{0,\cdots,T-1\}$

DR-Off-PAC Algorithm. We now propose a doubly robust off-policy actor-critic (DR-Off-PAC) algorithm as detailed in Algorithm 1. The stepsizes β_1, β_2, β_3, and α are set to be $\Theta(1)$ to yield a single-timescale update, i.e., all parameters are updated equally fast. At each iteration, critics I, II, and III perform one-step update respectively for parameters θ_q, θ_ρ, θ_ψ, and θ_{d_q}, and then actor performs one-step policy update based on all critics' return. Note that Algorithm 1 is inherently a tri-level optimization process, as the update of w depends on θ_ρ, θ_q, θ_ψ, and θ_{d_q}, in which the update of θ_{d_q} depends on θ_q. Thus the interactions between actor and critics and between critic and critic are more complicated than previous actor-critic algorithms that solve bilevel problems (Konda & Tsitsiklis, 2000; Bhatnagar, 2010; Xu et al., 2020b). Due to the single timescale scheme that Algorithm 1 adopts, actor's update is based on inexact estimations of critics, which can significantly affect the overall convergence of the algorithm. Interestingly, as we will show in the next section, Algorithm 1 is guaranteed to converge to the optimal policy, and at the same time attains doubly robust optimality gap with respect to approximation errors.

4. Convergence Analysis of DR-Off-PAC

In this section, we establish the local and global convergence rate for DR-Off-PAC in the single-timescale update setting.

4.1. Local Convergence

We first state a few standard technical assumptions, which have also been adopted in previous studies (Xu et al., 2020b; 2019; Zhang et al., 2020a;b)

Assumption 1. *For any $(s,a) \in \mathcal{S} \times \mathcal{A}$ and $w \in \mathbb{R}^d$, there exists a constant $C_d > 0$ such that $\rho_{\pi_w}(s,a) > C_d$.*

Assumption 2. *For any $(s,a) \in \mathcal{S} \times \mathcal{A}$, there exist positive constants C_ϕ, C_φ, C_ψ, and C_x such that the following hold: (1) $\|\phi(s,a)\|_2 \leq C_\phi$; (2) $\|\varphi(s,a)\|_2 \leq C_\varphi$; (3) $\|\psi(s,a)\|_2 \leq C_\psi$; (4) $\|x(s,a)\|_2 \leq C_x$.*

Assumption 3. *The matrices $A = \mathbb{E}_{\mathcal{D}_d \cdot \pi_w}[(\phi - \gamma\phi')\phi^\top]$, $B = \mathbb{E}_{\tilde{\mathcal{D}}_d \cdot \pi_w}[(\varphi - \varphi')\varphi'^\top]$ and $C = \mathbb{E}_{\mathcal{D}_d \cdot \pi_w}[(\gamma x' - x)x^\top]$ are nonsingular.*

Assumption 4. *For any $w, w' \in \mathbb{R}^d$ and any $(s,a) \in \mathcal{S} \times \mathcal{A}$, there exist positive constants C_{sc}, L_{sc}, and L_π such that the following hold: (1) $\|\nabla_w \log \pi_w(a|s)\|_2 \leq C_{sc}$; (2) $\|\nabla_w \log \pi_w(a|s) - \nabla_w \log \pi_{w'}(a|s)\|_2 \leq L_{sc} \|w - w'\|_2$; (3) $\|\pi_w(\cdot|s) - \pi_{w'}(\cdot|s)\|_{TV} \leq L_\pi \|w - w'\|_2$, where $\|\cdot\|_{TV}$ denotes the total-variation norm.*

The following theorem characterizes the convergence rate of Algorithm 1, as well as its doubly robust optimality gap.

Theorem 2 (Local convergence)**.** *Consider the DR-Off-PAC in Algorithm 1. Suppose Assumption 1 - 4 hold. Let the stepsize $\alpha, \beta_1, \beta_2, \beta_3 = \Theta(1)$. We have*

$$\mathbb{E}[\left\|\nabla_w J(w_{\hat{T}})\right\|_2] \leq \Theta\left(\frac{1}{\sqrt{T}}\right) + \Theta\left(\frac{1}{\sqrt{N}}\right) + \Theta(\epsilon_\rho \epsilon_{d_q} + \epsilon_{d_\rho} \epsilon_q + \epsilon_\rho \epsilon_q).$$

Theorem 2 shows that Algorithm 1 is guaranteed to converge to a first-order stationary point (i.e., locally optimal policy). In particular, the optimality gap (i.e., the overall convergence error) scales as $(\epsilon_\rho \epsilon_{d_q} + \epsilon_{d_\rho} \epsilon_q + \epsilon_\rho \epsilon_q)$. Thus, the optimality gap of Algorithm 1 is **3-way doubly robust** with respect to the function approximation errors, i.e., the optimality gap is small as long as one of the three pairs $(\epsilon_\rho, \epsilon_q)$, $(\epsilon_\rho, \epsilon_{d_\rho})$, $(\epsilon_q, \epsilon_{d_q})$ is small.

There are two key differences between the doubly robust properties characterized in Theorem 2 and Theorem 1. (a) At the high level, Theorem 1 characterizes the doubly robust property only for the policy gradient estimator, and such a property has been characterized in the previous work for other estimators. In contrast, Theorem 2 characterizes the doubly robust property for the optimality gap of the overall convergence of an algorithm, which has not been characterized in any of the previous studies. (b) At the more technical level, the estimation error ε defined in Theorem 1 captures both the optimization error ϵ_{opt} determined by how well we solve the nuisances estimation problem, and the approximation error ϵ_{approx} determined by the representation power of approximation function classes. Thus, Theorem 1 shows that $G_{\text{DR}}(w)$ is doubly robust to the per-iteration estimation errors that depend on both the optimization process and the approximation function class. As a comparison, Theorem 2 indicates that the optimality gap of DR-Off-PAC is doubly robust only to approximation errors determined by the approximation function class, which implies that the doubly robust property of the overall convergence of DR-Off-PAC is not affected by the optimization process.

Now in order to attain an optimization target accuracy ϵ (besides the doubly robust optimality gap), we let $T = \Theta(1/\epsilon^2)$ and $B = \Theta(1/\epsilon^2)$. Then Theorem 2 indicates that Algorithm 1 converges to an ϵ-accurate stationary point with the total sample complexity $NT = \Theta(1/\epsilon^4)$. This result outperforms the best known sample complexity of on-policy actor-critic algorithm by an factor of $\mathcal{O}(\log(1/\epsilon))$ in (Xu et al., 2020b). Such an improvement is mainly due to the single-loop structure that we adopt in Algorithm 1, in which critics inherit the most recently output from the last iteration as actor updates in order to be more sample efficient. But critic in the nested-loop algorithm in (Xu et al., 2020b) always restarts from an random initialization after each actor's update, which yields more sample cost.

4.2. Global Convergence

In this subsection, we establish the global convergence guarantee for DR-Off-PAC in Algorithm 1. We first make the following standard assumption on the Fisher information matrix induced by the policy class π_w.

Assumption 5. *For all $w \in \mathbb{R}^d$, the Fisher information matrix induced by policy π_w and initial state distribution μ_0 satisfies*

$$F(w) = \mathbb{E}_{\nu_{\pi_w}}[\nabla_w \log \pi_w(a|s) \nabla_w \log \pi_w(a|s)^\top] \succeq \lambda_F \cdot I_d,$$

for some constant $\lambda_F > 0$.

Assumption 5 essentially states that $F(w)$ is well-conditioned. This assumption can be satisfied by some commonly used policy classes. More detailed justification of such an assumption can be referred to Appendix B.2 in (Liu et al., 2020).

We further define the following *compatible function approximation* error as

$$\epsilon_{compat} = \max_{w \in \mathbb{R}^d} \sqrt{\mathbb{E}_{\nu_{\pi^*}}[(A_{\pi_w}(s,a) - (1-\gamma)\chi_{\pi_w}^{*\top} \nabla_w \log \pi_w(a|s))^2]},$$

where $A_{\pi_w}(s,a) = Q_{\pi_w}(s,a) - V_{\pi_w}(s)$ is the advantage function and $\chi_{\pi_w}^{*\top} = F(w)^{-1} \nabla_w J(w)$. Such an error ϵ_{compat} captures the approximating error of the advantage function by the score function. It measures the capacity of the policy class π_w, and takes small or zero values if the expressive power of the policy class is large (Wang et al., 2019; Agarwal et al., 2019).

The following theorem establishes the global convergence guarantee for Algorithm 1.

Theorem 3 (Global convergence)**.** *Consider the DR-Off-PAC update in Algorithm 1. Suppose Assumption 1, 2, 3 and 5 hold. For the same parameter setting as in Theorem 2, we have*

$$J(\pi^*) - J(w_{\hat{T}}) \leq \frac{\epsilon_{compat}}{1-\gamma} + \Theta\left(\frac{1}{\sqrt{T}}\right) + \Theta\left(\frac{1}{\sqrt{N}}\right) + \Theta(\epsilon_\rho \epsilon_{d_q} + \epsilon_{d_\rho} \epsilon_q + \epsilon_\rho \epsilon_q)$$

Theorem 3 shows that Algorithm 1 is guaranteed to converge to the global optimum at a sublinear rate, and the optimality gap is bounded by $\Theta(\epsilon_{compat}) + \Theta(\epsilon_\rho \epsilon_{d_q} + \epsilon_{d_\rho} \epsilon_q + \epsilon_\rho \epsilon_q)$. Note that the error term $\Theta(\epsilon_{compat})$ is introduced by the parametrization of policy and thus exists even for exact policy gradient algorithm (Liu et al., 2020; Wang et al., 2019). The global convergence of DR-Off-PAC in Theorem 3 also enjoys doubly robust optimality gap as in Theorem 2. By letting $T = \Theta(1/\epsilon^2)$ and $N = \Theta(1/\epsilon^2)$, Algorithm 1 converges to an ϵ-level global optimum (besides the approximation errors) with a total sample complexity $NT = \Theta(1/\epsilon^4)$. This result matches the global convergence rate of single-loop actor-critic in (Xu et al., 2020c; Fu et al., 2020).

5. Experiments

We conduct empirical experiments to answer the following two questions: (a) does the overall convergence of DR-Off-PAC doubly robust to function approximation errors as Theorem 2 & 3 indicate? (2) how does DR-Off-PAC compare with other off-policy methods?

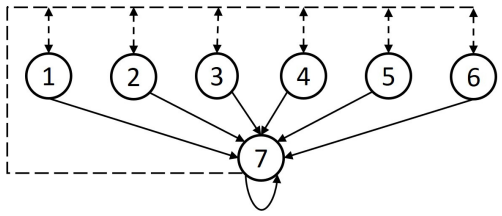

Figure 1. A variant of Baird's counterexample.

We consider a variant of Baird's counterexample (Baird, 1995; Sutton & Barto, 2018) as shown in Figure 1. There are two actions represented by solid line and dash line, respectively. The *solid* action always leads to state 7 and a reward 0, and the *dash* action leads to states 1-6 with equal probability and a reward $+1$. The initial distribution μ_0 chooses all states s with equal probability $\frac{1}{7}$ and the behavior distribution chooses all state-action pairs (s, a) with equal probability $\frac{1}{14}$. We consider two types of one-hot features for estimating the nuisances: complete feature (CFT) and incomplete feature (INCFT), where CFT for each (s, a) lies in $\mathbb{R}^{14}$ and INCFT for each (s, a) lies in $\mathbb{R}^d$ with $(d < 14)$. Note that CFT has large enough expressive power so that the approximation error is zero, while INCFT does not have enough expressive power, and thus introduces non-vanishing approximation errors. In our experiments, we consider fixed learning rates 0.1, 0.5, 0.1, 0.05, 0.01 for updating w, θ_q, θ_ψ, θ_{dq}, and $\theta_{d\rho}$, respectively, and we set the mini-batch size as $N = 5$. All curves are averaged over 20 independent runs.

Doubly Robust Optimality Gap: We first investigate how the function approximation error affects the optimality gap of the overall convergence of DR-Off-PAC. In this experiment, we set the dimension of INCFTs as 0, which results in trivial critics that always provide constant estimations. We consider the following four feature settings for critics

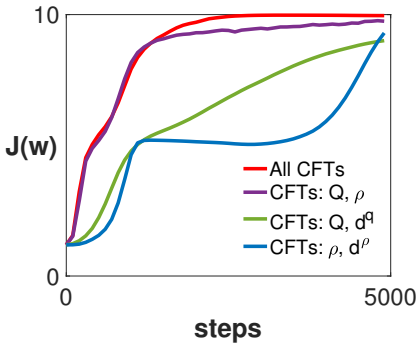

Figure 2. DR-Off-PAC under difference feature settings.

to estimate the nuisance functions (Q, ρ, d^q, d^ρ): **(I)** all nuisances with CFTs. **(II)** (Q, ρ) with CFTs and (d^ρ, d^q) with INCFTs; **(III)** (Q, d^q) with CFTs and (ρ, d^ρ) with INCFTs; **(IV)** (ρ, d^ρ) with CFTs and (Q, d^q) with INCFTs. The results are provided in Figure 2. We can see that DR-Off-PAC with all nuisances estimated by CFTs (red line) enjoys the fastest convergence speed and smallest optimality gap, and DR-Off-PAC with only two nuisances estimated with CFTs can still converge to the same optimal policy as the red line, validating the doubly robust optimality gap in the overall convergence characterized by Theorem 2 and Theorem 3.

Comparison to AC-DC: As we have mentioned before, previous provably convergent off-policy actor-critic algorithms introduce an additional critic to correct the distribution mismatch (Liu et al., 2019; Zhang et al., 2019c). Such a strategy can be viewed as a special case of DR-Off-PAC when both θ_{dq} and θ_ψ equal zero. Here we call such a type of algorithms as actor-critic with distribution correction (AC-DC). In this experiment, we set the dimension of INCFTs as 4 and compare the convergence of DR-Off-AC and AC-DC in the settings considered in our previous experiment. The learning curves of DR-Off-PAC and AC-DC are reported in Figure 3. We can see that the overall convergence of DR-

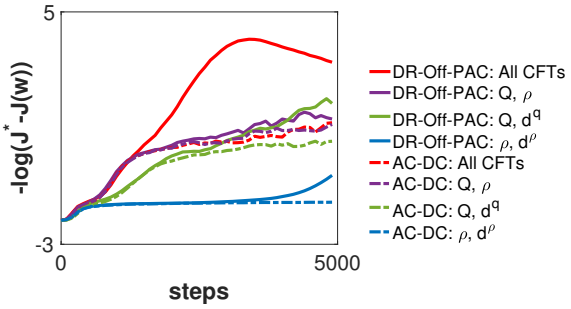

Figure 3. Comparison between DR-Off-PAC and AC-DC.

Off-PAC (each solid line) outperforms that of AC-DC (dash line with the same color) for all feature settings (where each color corresponds to one feature setting). Specifically, In **(III)** or **(IV)**, when either Q or ρ is estimated with incomplete features, the performance of AC-DC is significantly impeded by the approximation error and thus has lower accuracy, whereas DR-Off-PAC has better convergence performance by mitigating the effect of such approximation errors via the doubly robust property. Interestingly, even in the settings where both Q and ρ are estimated with complete features (S_1 and S_2) so that AC-DC is expected to achieve zero optimality gap, our DR-Off-PAC still converges faster and more accurately than AC-DC, demonstrating that DR-Off-PAC can improve the convergence of AC-DC even when both ρ and Q are estimated with a complete approximation function class.

6. Conclusion

In this paper, we first develop a new doubly robust policy gradient estimator for an infinite-horizon discounted MDP, and propose new methods to estimate the nuisances in the off-policy setting. Based on such an estimator, we propose a doubly robust off-policy algorithm called DR-Off-PAC for solving the policy optimization problem. We further study the finite-time convergence of DR-Off-PAC under the single timescale update setting. We show that DR-Off-PAC provably converges to the optimal policy, with the optimality gap being doubly robust to approximation errors that depend only on the expressive power of function classes. For future work, it is interesting to incorporate variance reduction technique (Xu et al., 2020a; Cutkosky & Orabona, 2019) to DR-Off-PAC to improve its convergence performance.

Acknowledgements

The work of T. Xu and Y. Liang was supported in part by the U.S. National Science Foundation under the grants CCF-1761506 and CCF-1900145. Z. Wang acknowledges National Science Foundation (Awards 2048075, 2008827, 2015568, 1934931), Simons Institute (Theory of Reinforcement Learning), Amazon, J.P. Morgan, and Two Sigma for their supports. Z. Yang acknowledges Simons Institute (Theory of Reinforcement Learning).

References

Agarwal, A., Kakade, S. M., Lee, J. D., and Mahajan, G. Optimality and approximation with policy gradient methods in Markov decision processes. *arXiv preprint arXiv:1908.00261*, 2019.

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

Bhatnagar, S. An actor–critic algorithm with function approximation for discounted cost constrained Markov decision processes. *Systems & Control Letters*, 59(12): 760–766, 2010.

Bhatnagar, S., Sutton, R. S., Ghavamzadeh, M., and Lee, M. Natural actor–critic algorithms. *Automatica*, 45(11): 2471–2482, 2009.

Cutkosky, A. and Orabona, F. Momentum-based variance reduction in non-convex SGD. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pp. 15236–15245, 2019.

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. In *Proc. International Conference on Machine Learning (ICML)*, pp. 1125–1134, 2018.

Degris, T., White, M., and Sutton, R. S. Off-policy actor-critic. In *Proc. International Conference on Machine Learning (ICML)*, pp. 179–186, 2012.

Dudík, M., Langford, J., and Li, L. Doubly robust policy evaluation and learning. In *Proc. International Conference on Machine Learning (ICML)*, pp. 1097–1104, 2011.

Dudík, M., Erhan, D., Langford, J., Li, L., et al. Doubly robust policy evaluation and optimization. *Statistical Science*, 29(4):485–511, 2014.

Fu, Z., Yang, Z., and Wang, Z. Single-timescale actor-critic provably finds globally optimal policy. *arXiv preprint arXiv:2008.00483*, 2020.

Fujimoto, S., Hoof, H., and Meger, D. Addressing function approximation error in actor-critic methods. In *Proc. International Conference on Machine Learning (ICML)*, pp. 1587–1596, 2018.

Gelada, C. and Bellemare, M. G. Off-policy deep reinforcement learning by bootstrapping the covariate shift. In

Proc. AAAI Conference on Artificial Intelligence (AAAI), volume 33, pp. 3647–3655, 2019.

Gottesman, O., Johansson, F., Komorowski, M., Faisal, A., Sontag, D., Doshi-Velez, F., and Celi, L. A. Guidelines for reinforcement learning in healthcare. *Nature medicine*, 25(1):16–18, 2019.

Haarnoja, T., Tang, H., Abbeel, P., and Levine, S. Reinforcement learning with deep energy-based policies. In *Proc. International Conference on Machine Learning (ICML)*, pp. 1352–1361, 2017.

Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. Soft actor-critic: off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *Proc. International Conference on Machine Learning (ICML)*, pp. 1861–1870, 2018.

Houthooft, R., Chen, Y., Isola, P., Stadie, B., Wolski, F., Ho, O. J., and Abbeel, P. Evolved policy gradients. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pp. 5400–5409, 2018.

Huang, J. and Jiang, N. From importance sampling to doubly robust policy gradient. In *Proc. International Conference on Machine Learning (ICML)*, pp. 4434–4443, 2020.

Imani, E., Graves, E., and White, M. An off-policy policy gradient theorem using emphatic weightings. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pp. 96–106, 2018.

Jiang, N. and Li, L. Doubly robust off-policy value evaluation for reinforcement learning. In *Proc. International Conference on Machine Learning (ICML)*, pp. 652–661, 2016.

Kallus, N. and Uehara, M. Statistically efficient off-policy policy gradients. *arXiv preprint arXiv:2002.04014*, 2020.

Konda, V. R. and Tsitsiklis, J. N. Actor-critic algorithms. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pp. 1008–1014, 2000.

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. In *Proc. International Conference on Learning Representations (ICLR)*, 2016.

Liu, Q., Li, L., Tang, Z., and Zhou, D. Breaking the curse of horizon: Infinite-horizon off-policy estimation. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pp. 5356–5366, 2018.

Liu, Y., Swaminathan, A., Agarwal, A., and Brunskill, E. Off-policy policy gradient with state distribution correction. In *Proc. International Conference on Learning Representations (ICLR)*, 2019.

Liu, Y., Zhang, K., Basar, T., and Yin, W. An improved analysis of (variance-reduced) policy gradient and natural policy gradient methods. *In Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 33, 2020.

Lyu, D., Qi, Q., Ghavamzadeh, M., Yao, H., Yang, T., and Liu, B. Variance-reduced off-policy memory-efficient policy search. *arXiv preprint arXiv:2009.06548*, 2020.

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.

Meuleau, N., Peshkin, L., and Kim, K.-E. Exploration in gradient-based reinforcement learning. 2001.

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 *Proc. International Conference on Machine Learning (ICML)*, pp. 1928–1937, 2016.

Morimura, T., Uchibe, E., Yoshimoto, J., Peters, J., and Doya, K. Derivatives of logarithmic stationary distributions for policy gradient reinforcement learning. *Neural computation*, 22(2):342–376, 2010.

Nachum, O., Norouzi, M., Xu, K., and Schuurmans, D. Bridging the gap between value and policy based reinforcement learning. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pp. 2775–2785, 2017.

Nachum, O., Norouzi, M., Xu, K., and Schuurmans, D. Trust-PCL: An off-policy trust region method for continuous control. In *Proc. International Conference on Learning Representations (ICLR)*, 2018.

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

O'Donoghue, B., Munos, R., Kavukcuoglu, K., and Mnih, V. Combining policy gradient and q-learning. In *Proc. International Conference on Learning Representations (ICLR)*, 2016.

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 *Proc. International Conference on Machine Learning (ICML)*, pp. 387–395, 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*, 529(7587):484–489, 2016.

Sutton, R. S. and Barto, A. G. *Reinforcement learning: An introduction*. 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 *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pp. 1057–1063, 2000.

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

Tang, Z., Feng, Y., Li, L., Zhou, D., and Liu, Q. Doubly robust bias reduction in infinite horizon off-policy estimation. *arXiv preprint arXiv:1910.07186*, 2019.

Tosatto, S., Carvalho, J., Abdulsamad, H., and Peters, J. A nonparametric off-policy policy gradient. In *Proc. International Conference on Artificial Intelligence and Statistics*, pp. 167–177. PMLR, 2020.

Wang, L., Cai, Q., Yang, Z., and Wang, Z. Neural policy gradient methods: Global optimality and rates of convergence. *arXiv preprint arXiv:1909.01150*, 2019.

Wang, Z., Bapst, V., Heess, N., Mnih, V., Munos, R., Kavukcuoglu, K., and de Freitas, N. Sample efficient actor-critic with experience replay. In *Proc. International Conference on Learning Representations (ICLR)*, 2016.

Xiong, H., Xu, T., Liang, Y., and Zhang, W. Non-asymptotic convergence of Adam-type reinforcement learning algorithms under Markovian sampling. *arXiv preprint arXiv:2002.06286*, 2020.

Xu, P. T., Wang, Z., Zhou, Y., and Liang, Y. Reanalysis of variance reduced temporal difference learning. In *Proc. International Conference on Learning Representations (ICLR)*, 2020a.

Xu, T., Zou, S., and Liang, Y. Two time-scale off-policy TD learning: Non-asymptotic analysis over markovian samples. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pp. 10633–10643, 2019.

Xu, T., Wang, Z., and Liang, Y. Improving sample complexity bounds for (natural) actor-critic algorithms. *In Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 33, 2020b.

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

Yang, M., Nachum, O., Dai, B., Li, L., and Schuurmans, D. Off-policy evaluation via the regularized lagrangian. *arXiv preprint arXiv:2007.03438*, 2020.

Zhang, K., Koppel, A., Zhu, H., and Başar, T. Global convergence of policy gradient methods to (almost) locally optimal policies. *arXiv preprint arXiv:1906.08383*, 2019a.

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

Zhang, S., Boehmer, W., and Whiteson, S. Generalized off-policy actor-critic. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pp. 2001–2011, 2019b.

Zhang, S., Liu, B., Yao, H., and Whiteson, S. Provably convergent off-policy actor-critic with function approximation. *arXiv preprint arXiv:1911.04384*, 2019c.

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