Doubly Robust Off-Policy Actor-Critic Algorithms in Reinforcement Learning

Anonymous authors
Paper under double-blind review

Abstract

We study off-policy actor-critic algorithms with doubly robust value function estimators. Doubly robust policy evaluation methods have previously been introduced for RL (Jiang and Li, 2016), followed by extending as a variance reduction technique in policy gradients (). In this work, we propose and analyse the significance of doubly robust policy evaluation in off-policy actor-critic algorithms. todo from here. We study the problem of off-policy critic evaluation in several variants of value-based off-policy actor-critic algorithms, where we require an off-policy critic evaluation step. Despite enormous success of off-policy policy gradients on control tasks, existing general methods suffer from high variance and instability. This is partly because the policy improvement step depends on gradient of the high variance estimate of the value function. In this work, we present a new way of off-policy policy evaluation in actor-critic, based on the doubly robust estimators Jiang and Li (2016). We extend the doubly robust estimator from off-policy policy evaluation (OPE) to off-policy actor-critic algorithms, where we estimate the rewards with a reward function approximator. Additionally, we analyse convergence of our proposed off-policy actor-critic algorithm and show that under certain conditions and assumptions, under off-policy data, the resulting algorithm converges to the optimal policy at a sublinear rate. Following this, we further investigate the bias-variance of the resulting off-policy policy gradient estimator as in doubly robust policy evaluation. Furthermore, in cases where the rewards are stochastic or noisy which can lead to high variance, doubly robust critic estimation can improve performance under corrupted reward signals, indicating its usefulness for robust and safe reinforcement learning.

1 Introduction

Policy gradient based methods are widely popular in deep reinforcement learning (RL) for solving continuous control tasks Schulman *et al.* (2015). Several variants of off-policy value gradient based methods have been proposed recently Haarnoja *et al.* (2018); Lillicrap *et al.* (2015) with the goal to solve complex manipulation while being sample efficient due to the ability to re-use off-policy data. Existing literature in RL on off-policy evaluation has a long history Precup (2000), Jiang and Li (2016) where the goal is to estimate the value of a policy using data sampled from another behaviour policy. Off-policy methods generally suffer from high variance due to importance sampling corrections, although several approaches have introduced bias by learning a performance model to reduce variance. Additionally, in several control and robotic tasks, the reward function may be corrupted or noisy, e.g rewards from sensors. Stochastic rewards may often make the off-policy learning process more difficult, especially for learning complex manipulation behaviours.

In this work, we extend the existing doubly robust estimators for off-policy evaluation (OPE) to the control setting, in off-policy actor-critic algorithms. In off-policy value-based policy gradient algorithms such as DDPG and SAC, the critic evaluates the performance of a policy and the policy is improved based on the critic estimate. Often existing value-based policy gradient algorithms, however, suffer from high variance and instability particularly in continuous control tasks Henderson *et al.* (2018). This problem can be further exacerbated in practical sensory-motor driven robotic applications where the reward functions are often noisy and corrupted. As such, existing off-policy policy gradient algorithms would be quite unreliable for use in the real world. We propose doubly robust estimation for critic evaluation towards the goal of reducing variance in the critic estimates, often better stability and safe improvements in performance.

We aim to merge the gap between off-policy evaluation (OPE) estimators with guarantees towards unbiasedness and low variance, and estimators used in off-policy actor-critic based methods. Our goal is to achieve low variance regression based critic evaluation while keeping the critic estimator unbiased. We propose a novel approach towards extending doubly robust estimators, based on a combination of direct model-based approach and model-free IS based estimators in the off-policy actor-critic setting for deep RL. We achieve this by proposing a reward function estimator, to estimate the reward function of the MDP, based on which we can estimate the DR estimator relying on the predicted MDP rewards.

Our **key contributions** are as follows: (i) We extend the doubly robust estimators to the off-policy actor-critic setting in RL, for low variance critic evaluation in policy gradient. (ii) We present a novel formulation for the model-based part of the reward estimator to derive the doubly robust (DR) estimator and use it for minimizing the mean squared error in critic evaluation. (iii) Our proposed doubly robust (DR) estimators for actor-critic algorithms can also be interpreted as a novel formulation as an action-dependent control variate, while keeping the policy gradient estimator unbiased but lowering variance of the gradient estimates. (iv) We find that extending the DR estimator for off-policy actor-critic can be significantly useful for reducing variance of the critic evaluation, since in most off-policy value gradient based approaches, the estimate of the critic plays a key role in performance. Our proposed extension is evaluated on a wide range of benchmark continuous control tasks, for both growing batch and fixed batch off-policy deep RL settings. Doubly robust estimators in actor-critic can significantly reduce the variance in critic evaluation under stochastic rewards, leading to robust and safe algorithms for practical applications. We evaluate our proposed algorithm on noisy reward versions of existing control benchmark tasks, and find that using DR estimation can significantly reduce variance and improve performance.

2 Preliminaries and Background

Lewis: Explicitly point out the dataset in this section In policy gradient methods, the aim is to learn a parameterized policy $\pi_\theta(a|s)$ to maximize the discounted sum of cumulative returns along the sampled trajectories, given by $J(\pi_\theta) = \mathbb{E}_{\pi_\theta}[\sum_{i=t+1}^{\infty} \gamma^i r(s_i, a_i)]$. Based on the policy gradient theorem Sutton *et al.* (2000), we can improve the policy parameters θ using the policy gradient, which can be computed with Monte-Carlo estimation $\nabla_\theta J(\theta) = \mathbb{E}\pi_\theta[\nabla_\theta \log \pi_\theta(a|s) Q^{\pi_\theta}(s,a)]$, where, the $\mathbb{E}$ above uses samples under the current policy π. Often the policy gradient estimator can suffer from high variance, and hence an advantage function or a state dependent baseline. Instead of on-policy gradient estimators, which can be sample-inefficient in practice, due to their inability to re-use data from past experiences, often an off-policy gradient estimate is preferred, based on the deterministic policy gradient (DPG) theorem Silver *et al.* (2014). For continuous control tasks, the DDPG algorithm Lillicrap *et al.* (2015) is often used due to their ease ability to learn from experience replay buffer. The off-policy policy gradient estimator is given by $\nabla_\theta J(\theta) = \mathbb{E}\mu[\nabla_\theta Q^w(s, \pi_\theta(s)]$, where Q^w is a critic estimate and π_θ is a deterministic policy which outputs continuous actions, to allow directly finding the gradient of value function.

Doubly Robust Off-Policy Evaluation : In off-policy evaluation, given a fixed batch or historical data generated by some behaviour or unknown policies $\beta(a,s)$, often the goal is to produce an estimate of the value function $V^\pi(s)$ such that the estimator has low mean squared error (MSE) between the true value function V^{π_e} and the estimated $V^\pi(s)$. Doubly robust estimation is an idea extended from statistics to produce regression estimates, lowering the MSE, in the case of missing or incomplete data. The idea of DR estimators extended from statistics were initially proposed for the contextual bandits setting where often the assumption was that the estimated reward function is given $\widehat{R}(s,a)$, to define the DR estimator for contextual bandits $V_{DR} = \widehat{V}(s) + \rho\big[r - \widehat{R}(s,a)\big]$, where $\widehat{R}(s,a)$ is the estimated reward, ρ being the IS corrections for the mismatch in the action distributions. From there, DR estimators were extended to the off-policy evaluation in RL setting Jiang and Li (2016),Thomas and Brunskill (2016a) to reduce the variance of off-policy evaluation, while keeping the regression based estimators unbiased. Jiang and Li (2016) argued that instead of using importance sampling corrections, which is unbiased, but can have high variance, it is better to use DR estimators in off-policy evaluation tasks. The key step in DR estimator is to use the following unbiased estimator $V_{DR}(s) = \widehat{V}(s) + \rho[r(s,a) + \gamma V_{DR}(s') - \widehat{Q}(s,a)]$, where we replace V^{π} with V^{π}_{DR} to denote a DR estimation of the off-policy evaluation. A key requirement in DR estimators is to use an approximation to the MDP model since the $\widehat{V}$ requires the rewards from an approximation of the MDP. In other words, $\widehat{R}$ used to compute $\widehat{V}$ is the model's prediction of the reward. Given the samples from past data and an approximate model of the MDP, the goal of DR estimators is to produce a low variance regression mean equated error estimate $\text{MSE}(V_{DR}, V^{\pi})$.

3 Approach

In this work, we extend the existing value-based off-policy policy gradient algorithms such as DDPG Lillicrap *et al.* (2015) and Soft Actor-Critic (SAC) Haarnoja *et al.* (2018) with a doubly robust (DR) estimation of the critic Q^{DR}_{ϕ}. We propose to use DR based critic estimation in the critic evaluation step, to reduce the variance of the critic estimate in value-based policy gradient algorithms. Since value-based gradient algorithms relies on directly finding the gradient of the action-value function, we hypothesize that reducing the variance of critic estimation can significantly improve the performance and lead to better stability in these algorithms.

3.1 Policy Gradient with Doubly Robust Estimator

In this section, we derive the doubly robust estimator for policy gradient algorithms. The key idea of this estimator comes from observing doubly robust estimation in OPE which provides an unbiased but low variance estimation of the value function. This depends on the estimated reward function $\widehat{R}(s,a)$, where accurate estimation of the MDP rewards can reduce the variance of the value function. In the next section, we will discuss our approach to estimate the MDP rewards $\widehat{R}(s,a)$ for a practical algorithm. By using the unbiased DR estimator as a control variate Thomas and Brunskill (2016b), for the off-policy actor-critic algorithm, we can achieve a low variance unbiased estimator of the critic, which helps to improve the stability of existing off-policy policy gradient algorithms. The policy gradient update is given by $\nabla_{\theta} J(\theta) = \mathbb{E}_{s \sim d_{\beta}} \Big[\nabla_{\theta} Q^{DR}_{\phi}(s, \pi_{\theta}(s)) \Big]$, where the critic and policies are separately parameterized with ϕ and θ, and we update the policy parameters with stochastic gradient optimization. Considering algorithms such as DDPG, here we denote the policy improvement phase with deterministic policies $\pi_{\theta}(s)$, which has been extended to stochastic Gaussian policies with a reparameterization trick in algorithms such as Soft Actor-Critic (SAC) Haarnoja *et al.* (2018). The key step in our algorithm is that, we replace the critic as in DDPG with a doubly robust estimation of the critic denoted by $Q^{DR}_{\phi}(s, \pi_{\theta}(s))$, such that the critic now minimizes the mean squared regression loss with the following TD error :

$$Q^{DR}_{\phi}(s,a) = \widehat{Q}(s, \pi_{\theta}(s)) + \Big[r(s,a) + \gamma Q^{DR}_{\phi}(s', \pi_{\theta}(s')) - \widehat{V}(s) \Big] \tag{3.1}$$

$$Q^{DR}_{\phi}(s_t, a_t) = \widehat{Q}(s_t, a_t) + \rho_t(\theta) \cdot \Big(r_t + \gamma Q^{DR}_{\phi}(s_{t+1}, a_{t+1}) - \widehat{V}(s_t) \Big), \tag{3.2}$$

where $\rho_t(\theta) = \pi_{\theta}(a_t|s_t)/\pi_0(a_t|s_t)$ is the importance ratio. where $\widehat{Q}(s, \pi_{\theta}(s))$ is following the DR estimate for action-value functions $\widehat{Q}$ instead of the value function $\widehat{V}$. Equation 3.1 shows that the critic update for the policy gradient now requires a separate estimation action-value $\widehat{Q}$ and value function $\widehat{V}$ as well, based on the predicted MDP rewards Jiang and Li (2016), ie, model-based estimation of the reward function, for the doubly robust estimation of the critic. In the next section, we discuss our approach for approximation of the rewards $\widehat{R}(s,a)$.

3.2 Reward Function Approximator

Since the doubly robust estimation of the critic requires an approximate MDP model, we estimate the true rewards of the MDP $R(s,a)$ with an approximate reward $\widehat{R}(s,a)$ by using a separately parameterized function approximator with parameters ψ_R, denoted by $\widehat{R} = f_{\psi_R}(s,a)$. The MDP rewards are estimated based on samples from the replay buffer, and we use a similar approach as in Romoff *et al.* (2018) where we train the reward function estimator based on the following regression loss

$$L(\psi_R) = \mathbb{E}_{s,a,r \sim Buffer}[(\widehat{R}(s,a) - R(s,a))^2] \tag{3.3}$$

We then use this reward for further estimating the approximated action-value function $\widehat{Q}(s,a)$ and value function $\widehat{V}(s)$ that are required for the DR estimation. Note that, to estimate this, which is a form of the advantage function or a control variate, we typically use a separate function approximator for the control variate estimate. We use another separately parameterized network with parameters ψ_{QV} which outputs both the approximated $\widehat{Q}$ and $\widehat{V}$ and trained with the samples from the replay buffer

$$\mathcal{L}_{\psi_{QV}} = \mathbb{E}_{s,a,r,s' \sim Buffer}[(\widehat{R}(s,a) + \gamma \widehat{Q}(s', \pi_\theta(s')) - \widehat{Q}(s,a))^2] \tag{3.4}$$

where we use the approximated reward $\widehat{R}(s,a)$ in the TD error for minimizing the loss $\mathcal{L}(\psi_{QV})$. Based on minimizing the losses in equation 3.3 and 3.4, we therefore get an approximation of $\widehat{R}$, $\widehat{Q}$ and $\widehat{V}$ that are required for doubly robust critic estimation.

3.3 Algorithm

Our entire algorithm requires actor and critic updates, additional estimates of$\widehat{R}$, $\widehat{Q}$ and $\widehat{V}$, thereby, we are effectively introducing a three time-scale algorithm. Therefore, in our algorithm, to ensure convergence, we use the large learning rate $\alpha_{\widehat{Q}}$ for the approximated $\widehat{Q}$, followed by a larger learning rate for the critic estimate compared to the actor as in actor-critic algorithms. In our algorithm, as detailed below, we therefore require the MDP rewards $\widehat{R}(s,a)$ to be predicted first, based on which we can train the function approximator with a TD error based on $\widehat{R}(s,a)$ to compute $\widehat{Q}(s,a)$ and $\widehat{V}(s)$. We then use the estimates from the model for off-policy DR evaluation of the critic Q_ϕ^{DR} such as to get unbiased but low variance estimates of the critic. Following this, we can then use *any* off-policy value-based policy gradient algorithm including DDPG, SAC or TD3.

Requirement of Independence: Following Jiang and Li (2016), note that, we use a different set of samples from the replay buffer to estimate $\widehat{R}(s,a)$, $\widehat{Q}(s,a)$ and $\widehat{V}(s)$ than the samples used for actor and critic updates. Although this is not a strict requirement that the samples used for estimating $\pi_\theta(a,s)$ and $\widehat{R}(s,a)$ be independent of each other, we find that using separate random samples for the model-based estimates and the model-free updates typically works better in practice.

3.4 Stochastic and Noisy Rewards

We further consider the case of using stochastic and noisy rewards, where conventional algorithms often fail due to high variance estimates of the gradient. We consider the setting where in addition to the MDP rewards $R(s,a)$, we add a Gaussian noise $\mathcal{N}(\mu,\sigma)$ to the true rewards, to make a corrupted version of the rewards. This is similar to the Corrupted Reward MDP (CRMDP) Everitt *et al.* (2017), similar to the stochastic reward control experiments considered in Romoff *et al.* (2018). This is similar to many practical robotic tasks where the reward function may often be noisy due to noise in the sensory data. We further examine the significance of using the DR estimator in cases with noisy rewards, given by $\widetilde{r}(s,a) = r(s,a) + \mathcal{N}(\mu,\sigma)$ and examine the significance of DR estimator to reduce the variance of noisy critic estimates. For our experiments, we add noise to examine performance in presence of stochastic rewards.

4 Convergence Analysis

In this section, we establish the global convergence of the doubly robust off-policy actor-critic algorithm presented in Algorithm 1. To do this, we build from the Soft Actor-Critic Haarnoja *et al.* (2018) off-policy gradient algorithm, and establish convergence guarantees for the policy improvement and policy evaluation steps, drawing inspiration similar to Liu *et al.* (2019). We present a fully off-policy policy optimization algorithm that can handle off-policy data in both the policy improvement and the policy evaluation step. We show for the first time that an off-policy policy optimization algorithm can still achieve global convergence at a sublinear rate, with parameterized neural network policies, even when using data from a different behaviour policy. Our key contribution is in showing a convergent off-policy policy optimization algorithm, which has been lacking in the literature for a long time. We show that both policy evaluation and policy improvement steps can fully handle off-policy data.

4.1 Reward Estimation Error in Doubly Robust Evaluation

Since our doubly robust estimator requires estimating the MDP rewards $\widehat{R}$, we first characterize how misspecification of the MDP rewards, and the error induced by it propagate along the optimization steps. To do this, we considered estimating the MDP rewards with a neural network approximator. We consider a D layer feedforward network with ReLU activation functions. We now analyse the error introduced by the reward function approximator $f_{\psi_R}(s,a)$ and establish the reward function error propagation error. Under policy π, let us denote the cumulative return with $R(s,a)$ and $\widehat{r}(s,a)$. Our goal is to estimate the rewards $\widehat{R}(s,a)$ parameterized by ψ. We can therefore quantify the reward gap term under the policy π given the two reward functions, to get the per step reward error :

$$\begin{aligned}\text{Gap}(\pi) &= \max_{\psi} \mathbb{E}_{(s,a)\sim d_\pi(s,a)}\Bigg[J(\pi,r) - J(\pi,r_\psi)\Bigg] \\ &\equiv \sum_{k=0}^{T-1} \mathbb{E}_{(s,a)\sim d_\pi(s,a)}\Big[L(\theta_k, r_{\psi'}) - J(\theta_k, r_\psi)\Big]\end{aligned} \tag{4.1}$$

We can then bound the optimality gap by upper bounding the difference of the expected cumulative reward under reward function r_ψ. We consider upper bounding the reward error term $L(\theta, r_{\psi'}) - L(\theta, r_\psi)$, by first characterizing the approximate convexity of $L(\theta,\psi)$ w.r.t ψ which can be given as

Lemma 4.1. *Assuming state and state-action density functions are upper bounded by an absolute constant $c \geq 0$, and under standard assumptions of concentrability coefficients in related literature, it holds for any $\psi \in S_{B_\psi}$ that*

$$\begin{aligned}\mathbb{E}_{\text{init}}\big[L(\theta_k,\psi') - L(\theta_k,\psi_k)\big] = \mathbb{E}_{\text{init}}\big[\nabla_\psi L(\theta_k,\psi_k)^\top(\psi' - \psi_k)\big] + \\ \mathcal{O}\big((1-\gamma)^{-1}\cdot B_\psi^{3/2}\cdot m^{-1/4}\big).\end{aligned} \tag{4.2}$$

The above characterizes the approximate convexity of the reward estimation problem $L(\theta,\psi)$ w.r.t to the reward function parameters ψ. For detailed proofs, see Chen *et al.* (2020).

Our analysis for characterizing the reward function error follows similarly, as in imitation learning frameworks and analysing the overall convergence of standard generative adversarial imitation learning Ho and Ermon (2016), Chen *et al.* (2020). Using this above, we can further upper bound the term $L(\theta_k,\psi') - L(\theta_k,\psi)$, adopting from Chen *et al.* (2020)

Lemma 4.2. *We can provide an upper bound characterizing the error based on the reward estimation as follows :*

$$\begin{aligned}\eta\cdot\big\{L(\psi_k,\psi') - L(\psi_k,\psi_k)\big\} \leq ||\psi_k - \psi_k^{'}||_2^2 - \\ ||\psi_{k+1} - \psi'||_2^2 - ||\psi_{k+1} - \psi_k||_2^2 + \eta\cdot\Delta_k\end{aligned} \tag{4.3}$$

where following similar analysis as in Chen *et al.* (2020), we have :

$$\begin{aligned}\mathbb{E}_{\text{init}}\Big[|\Delta_k|\Big] = \eta\cdot\big((2+\lambda\cdot L_\psi)^2 + \sigma^2/N\big) + \\ 2B_\psi\cdot(\sigma^2/N)^{1/2} + \mathcal{O}\big((1-\gamma)^{-1}B_\psi^{3/2}\cdot m^{-1/4}\big).\end{aligned} \tag{4.4}$$

Overall, we can characterize the per-step error in the reward function estmation as follows, which upper bounds the error in the reward estimation term required for our doubly robust estimator.

$$\begin{aligned}L(\theta,\psi') - L(\theta,\psi) \leq \eta^{-1}\cdot\Big[||\psi_k - \psi'||_2^2 - ||\psi_{k+1} - \psi'||_2^2 - \\ ||\psi_{k+1} - \psi_k||_2^2\Big] + \Delta_k\end{aligned} \tag{4.5}$$

4.2 Convergent Policy Improvement with Off-Policy Data

To establish global convergence for the policy improvement step, we propose a variant of the SAC algorithm similar in spirit to that of TRPO but one that can fully utilize off-policy samples from an experience replay buffer required for policy improvement step in algorithm 1. Note that unlike recent works analysing global convergence of policy gradient algorithms , , in the on-policy case, there is little to no work analysing convergence of off-policy policy gradient algorithms such as DDPG or SAC. This is beyond the scope of our work too. However, we propose a variant of SAC, by re-formulating it as a constrained optimization problem similar to TRPO, and analyse theoretical convergence properties of the resulting update.

We consider energy based policies of the form $\pi_{\theta_k} \propto \exp\{\tau_k^{-1} f_{\theta_k}\}$, such that given an estimator Q_{ω_k} of $Q_{\pi_{\theta_k}}$. In SAC, the policy parameters are learned by directly minimizing the expected KL-divergence term

$J(\pi_\theta) = \mathbb{E}_{s\sim\mathcal{D}}\Big[\mathrm{KL}\Big(\pi_\theta(\cdot \mid s)||\frac{\exp\left(Q_\phi(s,\cdot)\right)}{Z_\phi(s)}\Big)\Big]$. In general, this objective is solved by applying the reparameterization trick to obtain an unbiased DDPG style gradient estimator for stochastic policies. Alternately, the improved policy can also be obtained analytically, assuming $Q_\phi(s, a)$ is estimated, by solving an entropy regularized value function $\Omega(Q_\phi(s, \cdot))$, where Ω denotes entropy regularization, or a KL regularized value function $\mathrm{KL}(\pi_\theta||\pi_0)$ where π_0 is a uniform policy. Following this observation, we propose a KL regularized variant of SAC, where for energy based policies, we can perform policy gradient by solving an equivalent mean squared error term, one that can fully handle off-policy data. Given an entropy regularized value function as in SAC Haarnoja *et al.* (2018), we can write down the policy gradient objective similar to TRPO with a KL constraint, but the key difference being that the objective is computed under behaviour policy state-action samples $d_\mu(s, a)$.

$$\begin{aligned}\mathcal{L}(\theta) = \mathbb{E}_{d_{\mu_k}(s,a)}\big[< Q_{\phi_k}(s,\cdot), \pi_\theta(\cdot\,|\,s) > -\\ \beta_k \cdot \mathrm{KL}(\pi_\theta(\cdot\,|\,s)\,\|\,\pi_{\theta_k}(\cdot\,|\,s))\big],\end{aligned} \tag{4.6}$$

We consider energy based policies of the form $\pi_{\theta_k} \propto \exp\{\tau_k^{-1} f_{\theta_k}\}$, such that given an estimator Q_{ω_k} of $Q_{\pi_{\theta_k}}$, the update $\mathcal{L}(\theta)$ gives

$$\begin{aligned}\widehat{\pi}_{k+1} \leftarrow \underset{\pi}{\mathrm{argmax}}\{\mathbb{E}_{\pi_{\theta_k}}[Q_{\phi_k}(s,\cdot), \pi(\cdot\,|\,s) -\\ \beta_k \cdot \mathrm{KL}(\pi(\cdot\,|\,s)\,\|\,\pi_{\theta_k}(\cdot\,|\,s))]\}\end{aligned} \tag{4.7}$$

such that the analytical solution of this, as in SAC by directly solving the regularized value function gives the closed form solution (as in Liu *et al.* (2019)) given by $\widehat{\pi}_{k+1} \propto \exp\{\beta_k^{-1} Q_{\phi_k} + \tau_k^{-1} f_{\theta_k}\}$.

This means that we can update the policy, as the policy improvement step, similar to following the policy gradient theorem, in a state-wise manner. In other words, instead of explicitly following the policy gradient theorem to find a gradient direction to improve the policy, we have essentially defined a closed form solution to the policy improvement step (replacing the policy gradient step). Therfore, instead of maximizing the cumulative return objective $J(\pi_\theta)$ following the policy gradient theorem, we can instead write down a MSE subproblem for estimating the policy parameters, as below

$$\begin{aligned}\theta_{k+1} \leftarrow \underset{\theta}{\mathrm{argmin}}\,\mathbb{E}_{\mu_k}\big[\big(f_\theta(s,a) - \tau_{k+1}\cdot(\beta_k^{-1} Q_{\phi_k}(s,a) +\\ \tau_k^{-1} f_{\theta_k}(s,a))\big)^2\big],\end{aligned} \tag{4.8}$$

Again, note that the above problem can be solved using *off-policy* data collected under a fixed random behaviour policy μ_k, since the solutuon to the policy improvement step holds state-wise, allowing for efficient use of off-policy data (e.g stored in an experience replay or batch data). By re-formulating the SAC policy gradient update as in our doubly robust actor-critic algorithm to be equivalent to minimizing a MSE term, we can therefore establish global convergence and optimality of the resulting policy improvement step by adopting the convergence guarantees from Liu *et al.* (2019), as stated below :

Lemma 4.3 (Convergence of Off-Policy SAC update, adopted from Liu *et al.* (2019))**.** For the policy sequence $\{\pi_{\theta_k}\}_{k=1}^K$ attained by closed form solution of policy improvement

$$\min_{0\leq k\leq K}\{\mathcal{L}(\pi^*) - \mathcal{L}(\pi_{\theta_k})\} \leq \frac{\beta^2 \log|\mathcal{A}| + M + \beta^2 \sum_{k=0}^{K-1}(\varepsilon_k + \varepsilon_k')}{(1-\gamma)\beta\cdot\sqrt{K}}.$$

Here $\varepsilon_k = \tau_{k+1}^{-1}\epsilon_{k+1}\cdot\phi_k^* + \beta_k^{-1}\epsilon_k'\cdot\psi_k^*$ and $\varepsilon_k' = |\mathcal{A}|\cdot\tau_{k+1}^{-2}\epsilon_{k+1}^2$, where

$$\epsilon_{k+1} = \mathcal{O}(R_f^2 T^{-1/2} + R_f^{5/2} m_f^{-1/4} + R_f^3 m_f^{-1/2})$$

$$\epsilon_k' = \mathcal{O}(R_Q^2 T^{-1/2} + R_Q^{5/2} m_Q^{-1/4} + R_Q^3 m_Q^{-1/2}).$$

Also, we have $M = 2\mathbb{E}_{\nu^*}[\max_{a\in\mathcal{A}}(Q_{\omega_0}(s,a))^2] + 2R_f^2$.

For more details, see Theorem 4.9 in Liu *et al.* (2019). The key to our approach involves showing a policy improvement bound for the off-policy policy optimization step using off-policy data as in the doubly robust actor-critic algorithm presented in Algorithm 1.

4.3 Convergence of Doubly Robust Policy Evaluation

We now discuss convergence analysis in the policy evaluation step with the doubly robust off-policy estimator $Q^{\pi}_{DR}(s,a)$. We can characterize the error from solving the subproblem of policy evaluation. We make similar assumptions as in Assumptions 4.1, 4.2 and 4.4 in Liu *et al.* (2019), and show that the expected Bellman error for the policy evaluation step satisfies

Lemma 4.4 (Policy Evaluation Error with Doubly Robust Estimator adopted from Liu *et al.* (2019))**.** We set $T \geq 64/(1-\gamma)^2$ and the stepsize to be $\eta = T^{-1/2}$. Within the k-th iteration of policy evaluation algorithm, the output $Q_{\overline{\omega}}$ of the policy evaluation algorithm satisfies

$$\begin{aligned} &\mathbb{E}_{\text{init},\sigma_k}[(Q_{\overline{\phi},DR}(s,a) - Q^{\pi_{\theta_k}}(s,a))^2] = \\ &\mathcal{O}(R_Q^2 T^{-1/2} + R_Q^{5/2} m_Q^{-1/4} + R_Q^3 m_Q^{-1/2}). \end{aligned} \tag{4.9}$$

The key difference with existing policy evaluation error from Liu *et al.* (2019) is that due our reward function approximation, we would incur additional errors in the overall value function approximation, due to misspecification of the reward function, as discussed in the previous section

Proof. See detailed proof in § **??**. □

For more details and complete derivation for the policy evaluation error, see Liu *et al.* (2019) .We establish that since the policy evaluation step involves minimizing the MSBE w.r.t to the parameters ϕ, the policy evaluation error is similar to that esablished in Liu *et al.* (2019) since the MDP parameters $\widehat{R}$ and ψ_{QV} does not depend on the parameters ϕ. The above theorem therefore establishes that the policy improvement TRPO step under off-policy samples collected from $\mu(a,s)$ can still converge to the globally optimal solution. This further establishes the global convergence of the maximization step of $\mathcal{L}(\theta,\phi)$ in our overall minimax alternative optimization framework. The above theorem for the error in the doubly robust policy evaluation characterizes the same error as in the infinite dimensional mirror descent corresponding to neural TRPO/PPO with completely off-policy data. Therefore, as in Liu *et al.* (2019), this error will also decay to zero at the rate of $\frac{1}{\sqrt{T}}$ when the width $m_f = m_Q$ is sufficiently large, and T is the number of TD and SGD iterations. The detailed proof for the policy evaluation error with doubly robust estimation of $Q_{\pi,DR}(s,a)$ is as similar as that in Liu *et al.* (2019).

5 Bias Variance Analysis of Doubly Robust Actor-Critic

In this section we perform a bias variance analysis of the gradient estimator where the action-value function $Q_(\phi)^{DR}$ is estimated using doubly robust estimator. We show at first that the gradient estimate of the policy gradient is biased and then proceed towards analysing the variance of this estimator. Let τ be a given trajectory. The MC return for a given trajectory is given by $G(\tau)$. Since we are doing temporal difference, we done the return for a given trajectory as $\widetilde{G}(\tau)$. Where where, $\widetilde{G}_t = \widehat{Q}(s_t, \pi_\theta(s_t) + [\widehat{R}_t(s_t,a_t) + \gamma Q_{\phi}^{DR}(s_{t+1},a_{t+1}) - \widehat{V}(s_t)]$. The bias of the gradient estimator is computed as follows, which further re-arranging would give the bias as in equation 5.2

$$\begin{aligned} \mathbb{E}_{\tau\sim\pi_\theta}\Bigg[\Big(\sum_{t=1}^{T} \widehat{R}_t(s_t,a_t)\nabla_\theta \log \pi_\theta(a_t \mid s_t)\Big) \\ + \Big(\sum_{t=1}^{T}\big[\widehat{Q}(s_t,\pi_\theta(s_t)) + \gamma Q_{\phi}^{DR}(s_{t+1},a_{t+1}) \\ - \widehat{V}(s_t)\big]\nabla_\theta \log \pi_\theta(a_t \mid s_t)\Big) - \sum_{t=1}^{T} R_t(s_t,a_t)\nabla_\theta \log \pi_\theta(a_t \mid s_t)\Bigg] \end{aligned} \tag{5.1}$$

$$\begin{aligned} \texttt{bias} = \mathbb{E}_{\tau\sim\pi_\theta}\Bigg[\sum_{t=1}^{T}\Big(\widehat{R}_(s_t,a_t) - R_t(s_t,a_t)\Big)\nabla_\theta \log \pi_\theta(a_t \mid s_t)\Bigg] \\ + \mathbb{E}_{\tau\sim\pi_\theta}\Bigg[\Big(\sum_{t=1}^{T}\big[\widehat{Q}(s_t,\pi_\theta(s_t)) + \gamma Q_{\phi}^{DR}(s_{t+1},a_{t+1}) \\ - \widehat{V(s_t)}\big]\nabla_\theta \log \pi_\theta(a_t \mid s_t)\Big)\Bigg] \end{aligned} \tag{5.2}$$

We have a biased estimator where the bias is given by equation (5.2). The first in equation (5.2) also includes the bias due to the estimation of the reward function given by $\mathbb{E}[f_{\psi_R}(s_t, a_t)] - R_t(s_t, a_t)$. The variance of the gradient of the estimator is computed using the following,

$$
\begin{aligned}
&\mathrm{Var}\Big[\sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(a_t \mid s_t)\widetilde{G}_t\Big] \\
&\approx \sum_{t=1}^{T} \mathbb{E}_{\tau\sim\pi_\theta}\Big[\big(\nabla_\theta \log \pi_\theta(a_t \mid s_t)\widetilde{G}_t\big)^2\Big] \\
&\quad - \Big[\mathbb{E}_{\tau\sim\pi_\theta}\Big[\nabla_\theta \log \pi_\theta(a_t \mid s_t)\widetilde{G}_t\Big]\Big]^2
\end{aligned} \tag{5.3}
$$

where it is an approximation since we are computing the variance of the sum by sum of the variances which is not true in general. Under this assumption, we can write Var$(X) = \mathbb{E}[X^2] - [E(X)]^2$. The first term of equation (5.3) can be factored assuming independence of the values inside the expectation and substituting the value of $\widetilde{G}_t$ we have the following expression.

$$
\begin{aligned}
\approx \sum_{t=1}^{T} \mathbb{E}_{\tau\sim\pi_\theta}\Big[\big(\nabla_\theta \log \pi_\theta(a_t \mid s_t)\big)^2\Big]\mathbb{E}_{\tau\sim\pi_\theta}\Big[\big(\widehat{Q}(s_t, \pi_\theta(s_t) \\
+ [\widehat{R}_t(s_t, a_t) + \gamma Q_\phi^{DR}(s_{t+1}, a_{t+1}) - \widehat{V}(s_t)]\big)^2\Big]
\end{aligned}
$$

Additionally, since we are estimating the reward $\widehat{R}$, the variance of the reward function estimator is further added to the variance of the policy gradient estimation. In case where the reward function was not estimated, the variance of the gradient estimate will further reduce. Furthermore, the second term of equation (5.3), does not go to zero, since our estimator is biased, leading to further reduction in variance of the estimator. Existing literature Jiang and Li (2016) shows that the variance of value function estimation can be significantly reduced using a doubly robust estimator. In this section, we show that using a doubly robust critic evaluation, we can reduce the variance of both the policy gradient and critic estimator, which would explain our observed improvements in performance.

6 Experimental Results

In all our experiments, we compare with existing value-based off-policy gradient algorithms including DDPG, SAC and TD3, and highlight the significance of reducing variance of the critic estimate with DR estimator, in terms of performance improvement. We evaluate the performance of different actor critic algorithms with a doubly robust critic estimator on several continuous control Mujoco tasks Todorov *et al.* (2012). Experiments are evaluated on the Half-Cheetah-v1, Walker-2d-v1, Hopper-v1, environments, and evaluated an average over 5 runs with random seeds.

Figure 1 shows that using the DR estimator, we can obtain improvements in the performance over standard baselines. In case of SAC and TD3 algorithms, for Hopper-v1, we obtain an equivalent performance to that of the baselines. For experiments with the HalfCheetah-v1, environemnts, we observe that the baselines for SAC is higher than our DR estimator. These results show the case where the per step rewards are exactly observed by the agent (with out any noise) and that the agent directly use these rewards in the reward estimator. We also observe that for most of the mujoco environments, the variance of the DR estimator is lower compared to the baselines. Figure 2 shows the performance comparison using DR estimator where reward prediction is done in the presence of a Gaussian noise added on top of true rewards. Since the reward estimator that we use performs best if the rewards are noisy we observe that the DR estimator thus obtained significantly outperforms in the majority of the mujoco tasks. We also run our DR estimator with different noise levels added to the rewards. In Figure 2 shows the performance plots where the variance of the added Gaussian noise is $\sigma = 0.5$. We also observe that performance with DR estimator achieves lower variance compared to the baselines.

7 Related Work

Off policy evaluation in Markov decision processes is the task of evaluating a expected return of one policy with data generated by a different *behavior policy*. In Hanna *et al.* (2019), the authors propose a regression importance sampling method where the behavior policy is estimated using the same set of data which are used for calculating the importance sampling estimate. Such an estimate of the behavior policy leads to a lower mean squared error for off policy evaluation compared with the true behavior policy. Methods related to regression importance sampling had been studied for Monte Carlo methods in Henmi *et al.* (2007). Doubly

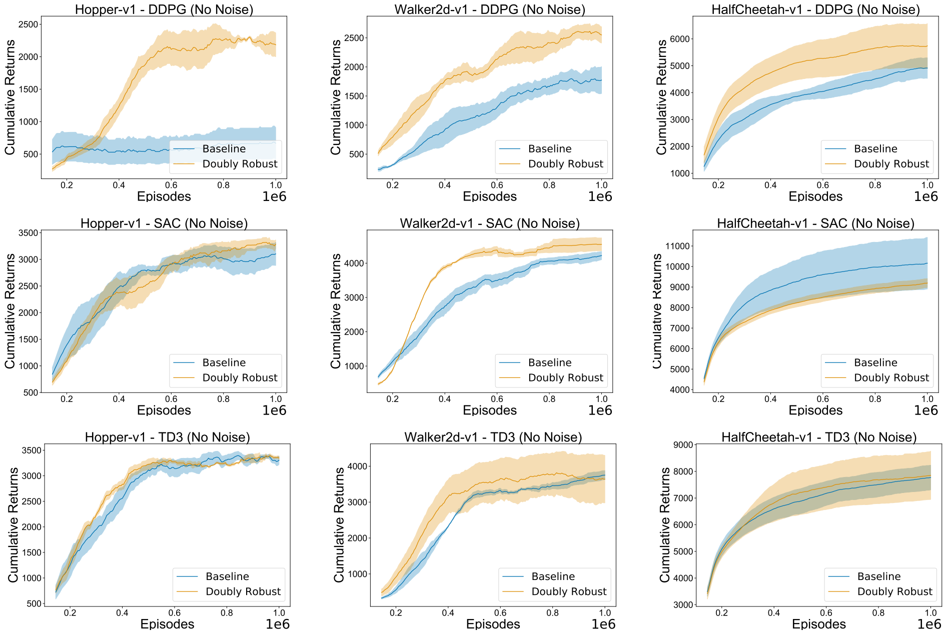

Figure 1: Performance comparison of DDPG,SAC and TD3 with DR estimator using rewards with no noise

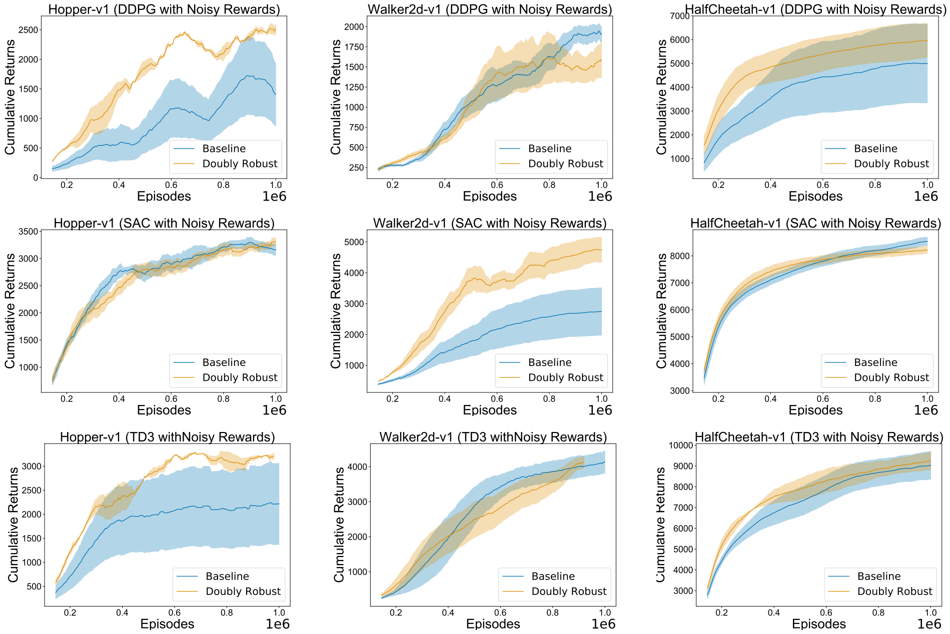

Figure 2: Performance comparison of DDPG, SAC and TD3 with DR estimator and baseline using noisy rewards where Gaussian noise σ added to the per-step rewards

Robust estimators for off policy value evaluation had been studied in Jiang and Li (2016), where the authors used Doubly Robust estimates for Bandits as a control variate for variance reduction. An extension of this work was proposed in Thomas and Brunskill (2016a), which uses doubly robust estimator and proposes a way to mix between model based estimates and importance sampling based estimates to predict the performance of a policy with historical data where the data was generated using a behavior policy. Since Doubly Robust estimators, require a reward predictor, we have adapted the reward estimator method in Romoff *et al.* (2018). Other data driven approaches in reward estimation has been discussed in Fu *et al.* (2018); Hadfield-Menell *et al.* (2017). A more extensive view of the Doubly Robust estimator has been proposed by Farajtabar *et al.* (2018)

where the authors present the formulation for learning DR model in RL and the model parameters are learned by minimizing the variance of the DR estimators.

8 Discussion and Conclusion

Our proposed algorithm uses doubly robust estimators, for providing unbiased and low variance critic evaluation. This is particularly useful since majority of popular off-policy methods for control tasks rely on direct value based policy gradient estimates where having useful estimates of the value function plays an important role. We find that since the DR estimator plays the role of control variates to reduce variance of the critic estimate, it has a significant effect in terms of improving performance and lowering variance of existing popular off-policy gradient algorithms. We achieve DR estimation for the model-free setting by using a separate reward function estimator to predict the MDP rewards, since DR estimators use a combination of model-fre and model-based approaches. Our approach of estimating the rewards with a separate function approximator plays a further important role in settings where the reward function is stochastic and corrupted. We find that existing policy gradient algorithms can perform poorly in stochastic reward environments, due to high variance in the critic estimates. In such cases, DR estimators can be quite useful for further reducing the variance. Our algorithm plays an important step towards robust and safe RL methods, which is a crucial step for extending current advances of deep RL algorithms for real world applications.

References

Minshuo Chen, Yizhou Wang, Tianyi Liu, Zhuoran Yang, Xingguo Li, Zhaoran Wang, and Tuo Zhao. On computation and generalization of generative adversarial imitation learning. *CoRR*, abs/2001.02792, 2020.

Tom Everitt, Victoria Krakovna, Laurent Orseau, and Shane Legg. Reinforcement learning with a corrupted reward channel. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 4705–4713, 2017.

Mehrdad Farajtabar, Yinlam Chow, and Mohammad Ghavamzadeh. More robust doubly robust off-policy evaluation. *arXiv preprint arXiv:1802.03493*, 2018.

Justin Fu, Avi Singh, Dibya Ghosh, Larry Yang, and Sergey Levine. Variational inverse control with events: A general framework for data-driven reward definition. In *Advances in Neural Information Processing Systems*, pages 8538–8547, 2018.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 1856–1865, 2018.

Dylan Hadfield-Menell, Smitha Milli, Pieter Abbeel, Stuart J Russell, and Anca Dragan. Inverse reward design. In *Advances in neural information processing systems*, pages 6765–6774, 2017.

Josiah Hanna, Scott Niekum, and Peter Stone. Importance sampling policy evaluation with an estimated behavior policy. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, pages 2605–2613, 2019.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 3207–3214, 2018.

Masayuki Henmi, Ryo Yoshida, and Shinto Eguchi. Importance sampling via the estimated sampler. *Biometrika*, 94(4):985–991, 2007.

Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 4565–4573, 2016.

Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, pages 652–661, 2016.

Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

Boyi Liu, Qi Cai, Zhuoran Yang, and Zhaoran Wang. Neural trust region/proximal policy optimization attains globally optimal policy. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 10564–10575, 2019.

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

Joshua Romoff, Peter Henderson, Alexandre Piché, Vincent François-Lavet, and Joelle Pineau. Reward estimation for variance reduction in deep reinforcement learning. In *2nd Annual Conference on Robot Learning, CoRL 2018, Zürich, Switzerland, 29-31 October 2018, Proceedings*, pages 674–699, 2018.

John Schulman, Sergey Levine, Pieter Abbeel, Michael I. Jordan, and Philipp Moritz. Trust region policy optimization. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pages 1889–1897, 2015.

David Silver, Guy Lever, Nicolas Heess, Thomas Degris, Daan Wierstra, and Martin Riedmiller. Deterministic policy gradient algorithms. In *International Conference on Machine Learning*, 2014.

Richard S Sutton, David A McAllester, Satinder P Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in neural information processing systems*, 2000.

Philip S. Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, pages 2139–2148, 2016.

Philip S. Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, pages 2139–2148, 2016.

Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5026–5033. IEEE, 2012.

A Variance Analysis

With the current estimator

$$Q_{\text{DR}}^{H+1-k} = \widehat{V}(s_k) + \frac{\pi(a_k|s_k)}{\pi_0(a_k|s_k)}\Big(r_k + \gamma Q_{\text{DR}}^{H-k} - \widehat{Q}(s_k, a_k)\Big). \tag{A.1}$$

We first show that the estimator Q_{DR} is unbiased. Recall that π_0 denotes the behavior policy collecting off-policy data, and π denotes the target policy to be evaluated. We have for any $k \in [H]$

$$\begin{aligned}\mathbb{E}_k^{\pi_0}[Q_{\text{DR}}^{H+1-k}] &= \mathbb{E}_k^{\pi_0}\big[\widehat{Q}(s_k, a_k)\big] + \mathbb{E}_k^{\pi_0}\rho_k\big[r_k + \gamma Q_{\text{DR}}^{H-k} - \widehat{V}(s_k)\big] \\ &= \mathbb{E}_k^{\pi_0}\big[\widehat{Q}(s_k, a_k)\big] + \mathbb{E}_k^{\pi}\big[r_k + \gamma Q_{\text{DR}}^{H-k} - \widehat{V}(s_k)\big] \\ &= \mathbb{E}_k^{\pi_0}\big[\widehat{Q}(s_k, a_k)\big] - \widehat{V}(s_k) + \mathbb{E}_k^{\pi}\big[r_k + \gamma Q_{\text{DR}}^{H-k}\big] \end{aligned} \tag{A.2}$$

We use the shorthand $\mathbb{E}_k = \mathbb{E}[\cdot \mid s_1, a_1, \cdots, s_{k-1}, a_{k-1}]$ for conditional expectations, and $\text{Var}_k[\cdot]$ for variances similarly. Likewise, we use $\mathbb{E}_\mu[\cdot] = \mathbb{E}_{\mu,P}[\cdot]$ for the expectation w.r.t. the randomness of policy μ and the transition function P. For notational simplicity, we write $a_k \sim \pi(\cdot|s_k)$ as $a_k \sim \pi$ for any $k \geq 1$ and a stochastic policy π. We also write $Q(s_k, a_k)$ for $Q^{\pi,H+1-k}(s_k, a_k)$. We denote by σ_0 the stationary distribution of the behavior policy μ.

Theorem A.1. Let $\delta(s, a) = \widehat{Q}(s_k, a_k) - Q(s_k, a_k)$. We characterize the variance of the doubly robust estimator Q_{DR}^{H+1-k} for $Q^{\pi,H+1-k}$ recursively as below,

$$\begin{aligned}\text{Var}_k[Q_{\text{DR}}^{H+1-k}] = \mathbb{E}_k\Big[\text{Var}_k\big[\rho_k\delta(s_k, a_k) \mid s_k\big]\Big] + \mathbb{E}_k\Big[\text{Var}_{k+1}\big[\rho_k^2 r_k\big]\Big] + \gamma^2\mathbb{E}_k\Big[\text{Var}_{k+1}\big[\rho_k^2 Q_{\text{DR}}^{H-k}\big]\Big] \\ + \text{Var}_k\big[V(s_k)\big]\end{aligned} \tag{A.3}$$

for $k \in [H]$. Here the boundary value of the variance is $\text{Var}[Q_{\text{DR}}^0 \mid s_H, a_H] = 0$.

Proof. We use the shorthand $\mathbb{E}_k = \mathbb{E}[\cdot \mid s_1, a_1, \cdots, s_{k-1}, a_{k-1}]$ for conditional expectations, and $\text{Var}_k[\cdot]$ for variances similarly. For notational simplicity, we write $a_k \sim \pi(\cdot|s_k)$ as $a_k \sim \pi$ for any $k \geq 1$ and a stochastic policy π. We also write $Q(s_k, a_k)$ for $Q^{\pi,H+1-k}(s_k, a_k)$. By definition we have

$$\begin{aligned}\text{Var}_k[Q_{\text{DR}}^{H+1-k}] &= \mathbb{E}_k\Big[\big(Q_{\text{DR}}^{H+1-k}\big)^2\Big] - \Big(\mathbb{E}_{k+1}\mathbb{E}_{s_k}\mathbb{E}_{a_k\sim\pi}\big[Q(s_k, a_k)\big]\Big)^2 \\ &\overset{(a)}{=} \mathbb{E}_k\Big[\Big(\widehat{V}(s_k) + \rho_k\big(r_t + \gamma Q_{\text{DR}}^{H-k} - \widehat{Q}(s_k, a_k)\big)\Big)^2 - V(s_k)^2\Big] + \text{Var}_k\big[V(s_k)\big] \\ &= \mathbb{E}_k\Big[\Big(-\rho_k\delta(s_k, a_k) + \widehat{V}(s_k) + \rho_k\big(r_t + \gamma Q_{\text{DR}}^{H-k} - Q(s_k, a_k)\big)\Big)^2 - V(s_k)^2\Big] \\ &\quad + \text{Var}_k\big[V(s_k)\big] \\ &\overset{(b)}{=} \mathbb{E}_k\Big[\Big(-\rho_k\delta(s_k, a_k) + \widehat{V}(s_k) + \rho_k\big(r_k - R(s_k, a_k)\big) + \rho_k\gamma\big(Q_{\text{DR}}^{H-k} - \mathbb{E}_{k+1}\big[V(s_{k+1})\big]\big)\Big)^2 \\ &\quad - Q(s_k, a_k)^2\Big] + \text{Var}_k\big[V(s_k)\big] \\ &\overset{(c)}{=} \mathbb{E}_k\Big[\mathbb{E}_k\Big[\big(-\rho_k\delta(s_k, a_k) + \widehat{V}(s_k)\big)^2 - V(s_k)^2\Big|s_k\Big]\Big] + \mathbb{E}_k\Big[\mathbb{E}_{k+1}\Big[\rho_k^2\big(r_k - R(s_k, a_k)\big)^2\Big]\Big] \\ &\quad + \mathbb{E}_k\Big[\mathbb{E}_{k+1}\Big[\rho_k^2\gamma^2\Big(Q_{\text{DR}}^{H-k} - \mathbb{E}_{k+1}\big[V(s_{k+1})\big]\Big)^2\Big]\Big] + \text{Var}_k\big[V(s_k)\big] \\ &\overset{(d)}{=} \mathbb{E}_k\Big[\text{Var}_k\big[-\rho_k\delta(s_k, a_k) + \widehat{V}(s_k) \mid s_k\big]\Big] + \mathbb{E}_k\Big[\text{Var}_{k+1}\big[\rho_k^2 r_k\big]\Big] \\ &\quad + \gamma^2\mathbb{E}_k\Big[\text{Var}_k\big[\rho_k^2 Q_{\text{DR}}^{H-k} \mid s_k, a_k\big]\Big] + \text{Var}_k\big[V(s_k)\big] \\ &= \mathbb{E}_k\Big[\text{Var}_k\big[\rho_k\delta(s_k, a_k) \mid s_k\big]\Big] + \mathbb{E}_k\Big[\text{Var}_{k+1}\big[\rho_k^2 r_k\big]\Big] + \gamma^2\mathbb{E}_k\Big[\text{Var}_{k+1}\big[\rho_k^2 Q_{\text{DR}}^{H-k}\big]\Big] \\ &\quad + \text{Var}_k\big[V(s_k)\big],\end{aligned} \tag{A.4}$$

where (a) and (d) hold since $V(s_k) = \mathbb{E}_{a\sim\pi}[Q(s_k, a)]$, (c) follows from the fact that conditioned on $\{s_k, a_k\}$, $Q_{\text{DR}}^{H-k} - \mathbb{E}_{k+1}\big[V(s_{k+1})\big]$ and $r_k - R(s_k, a_k)$ are independent with zero means. In addition, (b) holds by the definition of action-value function Q. Hence, we conclude the proof. □

B MSE OF THE DOUBLY ROBUST ESTIMATOR FOR THE ACTION-VALUE FUNCTION

Let $\rho(t) = \pi_{\theta_t}$ We calculate the mean squared error (MSE) of our doubly robust estimator $Q^{(t)}_{\text{DR}}$ at the t-th iteration of policy improvement as follows,

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{(s,a)\sim(\sigma_0,\mu)}\left[\mathbb{E}_\mu\left[\left(Q^{(t)}_{\text{DR}}(s,a) - Q^{\pi_{\theta_t}}(s,a)\right)^2 \,\middle|\, s,a\right]\right] \\
&= \mathop{\mathbb{E}}_{(s,a)\sim(\sigma_0,\mu)}\left[\text{Var}_\mu\left[Q^{(t)}_{\text{DR}}(s,a) \,\middle|\, s,a\right]\right] + \mathop{\mathbb{E}}_{(s,a)\sim(\sigma_0,\mu)}\left[\left(\mathbb{E}_\mu\left[Q^{(t)}_{\text{DR}}(s,a)\right] - Q^{\pi_{\theta_t}}(s,a)\right)^2 \,\middle|\, s,a\right] \\
&= \text{Var}_\mu\left[Q^{(t)}_{\text{DR}}(s,a)\right] + \mathop{\mathbb{E}}_{(s,a)\sim(\sigma_0,\mu)}\left[\left(\mathbb{E}_\mu\left[Q^{(t)}_{\text{DR}}(s,a)\right]\right)^2 \,\middle|\, s,a\right] - 2\mathbb{E}_\mu\left[Q^{(t)}_{\text{DR}}(s,a)\right]\mathbb{E}_\mu\left[Q^{\pi_{\theta_t}}(s,a)\right] \\
&\quad + \mathbb{E}_\mu\left[\left(Q^{\pi_{\theta_t}}(s,a)\right)^2\right] \\
&= \text{Var}_\mu\left[Q^{(t)}_{\text{DR}}(s,a)\right] + \mathop{\mathbb{E}}_{(s,a)\sim(\sigma_0,\mu)}\left[\left(\mathbb{E}_{\pi_{\theta_t}}\left[Q^{\pi_{\theta_t}}(s,a)\right]\right)^2 \,\middle|\, s,a\right] - 2\mathbb{E}_\mu\left[Q^{(t)}_{\text{DR}}(s,a)\right]\mathbb{E}_\mu\left[Q^{\pi_{\theta_t}}(s,a)\right] \\
&\quad + \mathbb{E}_\mu\left[\left(Q^{\pi_{\theta_t}}(s,a)\right)^2\right],
\end{aligned}
\tag{B.1}
$$

where $\text{Var}_\mu\left[Q^{(t)}_{\text{DR}}(s,a)\right] = \text{Var}_1[Q^H_{\text{DR}}]$ is characterized in (A.4).

Here unbiasedness is defined as

$$
\mathbb{E}_{(s_1,a_1)\sim(\mu,\pi_0)}[\mathbb{E}_{s_2,a_2,\ldots}[Q^H_{\text{DR}}(s_1,a_1)]] = \mathbb{E}_{(s,a)\sim(\mu,\pi)}[Q(s,a)] \tag{B.2}
$$

or

$$
\mathbb{E}_{(s_1,a_1)\sim(\mu,\pi_0)}[\mathbb{E}_{s_2,a_2,\ldots}[Q^H_{\text{DR}}(s_1,a_1)]] = \mathbb{E}_{(s,a)\sim(\mu,\pi_0)}[Q(s,a)] \tag{B.3}
$$

C ERROR PROPAGATION OF DROPAC

Let π^* be the optimal policy with ν^* being its stationary state distribution and σ^* being its stationary state-action distribution. Recall that, as defined in (**??**), $\widehat{\pi}_{k+1}$ is the ideal improved policy based on Q_{ω_k}, which is an estimator of the exact action-value function $Q^{\pi_{\theta_k}}$. Accordingly, we define the ideal improved policy based on $Q^{\pi_{\theta_k}}$ as

$$
\pi_{k+1} = \mathop{\text{argmax}}_{\pi}\left\{\mathbb{E}_{\nu_k}\left[Q^{\pi_{\theta_k}}(s,\cdot), \pi(\cdot,s) - \beta_k \cdot \text{KL}(\pi(\cdot\,|\,s)\,\|\,\pi_{\theta_k}(\cdot\,|\,s))\right]\right\}. \tag{C.1}
$$

By the same proof of Proposition **??**, we have $\pi_{k+1} \propto \exp\{\beta_k^{-1}Q^{\pi_{\theta_k}} + \tau_k^{-1}f_{\theta_k}\}$, which is also an energy-based policy.

We define the following quantities related to density ratios between policies or stationary distributions,

$$
\phi_k^* = \mathbb{E}_{\widetilde{\sigma}_k}[|\,\mathrm{d}\pi^*/\,\mathrm{d}\pi_0 - \,\mathrm{d}\pi_{\theta_k}/\,\mathrm{d}\pi_0|^2]^{1/2}, \quad \psi_k^* = \mathbb{E}_{\sigma_k}[|\,\mathrm{d}\sigma^*/\,\mathrm{d}\sigma_k - \,\mathrm{d}\nu^*/\,\mathrm{d}\nu_k|^2]^{1/2}, \tag{C.2}
$$

where $\mathrm{d}\pi^*/\,\mathrm{d}\pi_0$, $\mathrm{d}\pi_{\theta_k}/\,\mathrm{d}\pi_0$, $\mathrm{d}\sigma^*/\,\mathrm{d}\sigma_k$, and $\mathrm{d}\nu^*/\,\mathrm{d}\nu_k$ are the Radon-Nikodym derivatives. A closely related quantity known as the concentrability coefficient is commonly used in the literature (**?????**). In comparison, as our analysis is based on stationary distributions, our definitions of ϕ_k^* and ψ_k^* are simpler in that they do not require unrolling the state-action sequence. Then we have the following lemma that quantifies how the errors of policy improvement and policy evaluation propagate into the infinite-dimensional policy space.

Lemma C.1 (Error Propagation)**.** Suppose that the policy improvement error in Line **??** of Algorithm **??** satisfies

$$
\mathbb{E}_{\widetilde{\sigma}_k}\left[\left(f_{\theta_{k+1}}(s,a) - \tau_{k+1}\cdot(\beta_k^{-1}Q_{\omega_k}(s,a) - \tau_k^{-1}f_{\theta_k}(s,a))\right)^2\right] \leq \Delta_{k+1}, \tag{C.3}
$$

and the policy evaluation error in Line **??** of Algorithm **??** satisfies

$$
\mathbb{E}_{\sigma_k}\left[\left(Q^{(t)}_{\text{DR}}(s,a) - Q^{\pi_{\theta_t}}(s,a)\right)^2\right] \leq \Delta'_k. \tag{C.4}
$$

For π_{k+1} defined in (C.1) and $\pi_{\theta_{k+1}}$ obtained in Line **??** of Algorithm **??**, we have

$$
\left|\mathbb{E}_{\nu^*}\left[\log(\pi_{\theta_{k+1}}(\cdot\,|\,s)/\pi_{k+1}(\cdot\,|\,s)), \pi^*(\cdot\,|\,s) - \pi_{\theta_k}(\cdot\,|\,s)\right]\right| \leq \varepsilon_k, \tag{C.5}
$$

where $\varepsilon_k = \tau_{k+1}^{-1}\Delta_{k+1}\cdot\phi_{k+1}^* + \beta_k^{-1}\Delta'_k\cdot\psi_k^*$.

Proof of Lemma C.1. We first have

$$\pi_{k+1}(a \,|\, s) = \exp\{\beta_k^{-1} Q^{\pi_{\theta_k}}(s,a) + \tau_k^{-1} f_{\theta_k}(s,a)\}/Z_{k+1}(s),$$

and

$$\pi_{\theta_{k+1}}(a \,|\, s) = \exp\{\tau_{k+1}^{-1} f_{\theta_{k+1}}(s,a)\}/Z_{\theta_{k+1}}(s).$$

Here $Z_{k+1}(s), Z_{\theta_{k+1}}(s) \in \mathbb{R}$ are normalization factors, which are defined as

$$\begin{aligned} Z_{k+1}(s) &= \sum_{a' \in \mathcal{A}} \exp\{\beta_k^{-1} Q^{\pi_{\theta_k}}(s,a') + \tau_k^{-1} f_{\theta_k}(s,a')\}, \\ Z_{\theta_{k+1}}(s) &= \sum_{a' \in \mathcal{A}} \exp\{\tau_{k+1}^{-1} f_{\theta_{k+1}}(s,a')\}, \end{aligned} \tag{C.6}$$

respectively. Thus, we reformulate the inner product in (C.5) as

$$\begin{aligned} &\log \pi_{\theta_{k+1}}(\cdot \,|\, s) - \log \pi_{k+1}(\cdot \,|\, s), \pi^*(\cdot \,|\, s) - \pi_{\theta_k}(\cdot \,|\, s) \\ &\quad = \tau_{k+1}^{-1} f_{\theta_{k+1}}(s,\cdot) - (\beta_k^{-1} Q^{\pi_{\theta_k}}(s,\cdot) + \tau_k^{-1} f_{\theta_k}(s,\cdot)), \pi^*(\cdot \,|\, s) - \pi_{\theta_k}(\cdot \,|\, s), \end{aligned} \tag{C.7}$$

where we use the fact that

$$\begin{aligned} &\log Z_{k+1}(s) - \log Z_{\theta_{k+1}}(s), \pi^*(\cdot \,|\, s) - \pi_{\theta_k}(\cdot \,|\, s) \\ &\quad = (\log Z_{k+1}(s) - \log Z_{\theta_{k+1}}(s)) \sum_{a' \in \mathcal{A}} (\pi^*(a' \,|\, s) - \pi_{\theta_k}(a' \,|\, s)) = 0. \end{aligned}$$

Thus, it remains to upper bound the right-hand side of (C.7). We first decompose it to two terms, namely the error from learning the Q-function and the error from fitting the improved policy, that is,

$$\begin{aligned} &\tau_{k+1}^{-1} f_{\theta_{k+1}}(s,\cdot) - (\beta_k^{-1} Q^{\pi_{\theta_k}}(s,\cdot) + \tau_k^{-1} f_{\theta_k}(s,\cdot)), \pi^*(\cdot \,|\, s) - \pi_{\theta_k}(\cdot \,|\, s) \\ &\quad = \underbrace{\tau_{k+1}^{-1} f_{\theta_{k+1}}(s,\cdot) - (\beta_k^{-1} Q_{\omega_k}(s,\cdot) + \tau_k^{-1} f_{\theta_k}(s,\cdot)), \pi^*(\cdot \,|\, s) - \pi_{\theta_k}(\cdot \,|\, s)}_{\text{(i)}} \\ &\qquad + \underbrace{\beta_k^{-1} Q_{\omega_k}(s,\cdot) - \beta_k^{-1} Q^{\pi_{\theta_k}}(s,\cdot), \pi^*(\cdot \,|\, s) - \pi_{\theta_k}(\cdot \,|\, s)}_{\text{(ii)}}. \end{aligned} \tag{C.8}$$

Upper Bounding (i): We have

$$\begin{aligned} &\tau_{k+1}^{-1} f_{\theta_{k+1}}(s,\cdot) - (\beta_k^{-1} Q_{\omega_k}(s,\cdot) + \tau_k^{-1} f_{\theta_k}(s,\cdot)), \pi^*(\cdot \,|\, s) - \pi_{\theta_k}(\cdot \,|\, s) \qquad \text{(C.9)} \\ &\quad = \tau_{k+1}^{-1} f_{\theta_{k+1}}(s,\cdot) - (\beta_k^{-1} Q_{\omega_k}(s,\cdot) + \tau_k^{-1} f_{\theta_k}(s,\cdot)), \pi_0(\cdot \,|\, s) \cdot \left(\frac{\pi^*(\cdot \,|\, s)}{\pi_0(\cdot \,|\, s)} - \frac{\pi_{\theta_k}(\cdot \,|\, s)}{\pi_0(\cdot \,|\, s)}\right). \end{aligned}$$

Taking expectation with respect to $s \sim \nu^*$ on the both sides of (C.9) and using the Cauchy-Schwarz inequality, we obatin

$$\begin{aligned} &\big|\mathbb{E}_{\nu^*}\big[\tau_{k+1}^{-1} f_{\theta_{k+1}}(s,\cdot) - (\beta_k^{-1} Q_{\omega_k}(s,\cdot) + \tau_k^{-1} f_{\theta_k}(s,\cdot)), \pi^*(\cdot \,|\, s) - \pi_{\theta_k}(\cdot \,|\, s)\big]\big| \\ &= \left|\int_{\mathcal{S}} \tau_{k+1}^{-1} f_{\theta_{k+1}}(s,\cdot) - (\beta_k^{-1} Q_{\omega_k}(s,\cdot) + \tau_k^{-1} f_{\theta_k}(s,\cdot)), \pi_0(\cdot \,|\, s) \cdot \left(\frac{\pi^*(\cdot \,|\, s)}{\pi_0(\cdot \,|\, s)} - \frac{\pi_{\theta_k}(\cdot \,|\, s)}{\pi_0(\cdot \,|\, s)}\right) \cdot \nu^*(s)\,\mathrm{d}s\right| \\ &= \left|\int_{\mathcal{S} \times \mathcal{A}} \big(\tau_{k+1}^{-1} f_{\theta_{k+1}}(s,a) - (\beta_k^{-1} Q_{\omega_k}(s,a) + \tau_k^{-1} f_{\theta_k}(s,a))\big) \cdot \left(\frac{\pi^*(a \,|\, s)}{\pi_0(a \,|\, s)} - \frac{\pi_{\theta_k}(a \,|\, s)}{\pi_0(a \,|\, s)}\right) \mathrm{d}\widetilde{\sigma}_k(s,a)\right| \\ &\le \mathbb{E}_{\widetilde{\sigma}_k}\big[\big(\tau_{k+1}^{-1} f_{\theta_{k+1}}(s,a) - (\beta_k^{-1} Q_{\omega_k}(s,a) + \tau_k^{-1} f_{\theta_k}(s,a))\big)^2\big]^{1/2} \cdot \mathbb{E}_{\widetilde{\sigma}_k}\left[\left|\frac{\mathrm{d}\pi^*}{\mathrm{d}\pi_0} - \frac{\mathrm{d}\pi_{\theta_k}}{\mathrm{d}\pi_0}\right|^2\right]^{1/2} \\ &\le \tau_{k+1}^{-1} \Delta_{k+1} \cdot \phi_k^*, \end{aligned} \tag{C.10}$$

where in the last inequality we use the error bound in (C.3) and the definition of ϕ_k^* in (C.2).

Upper Bounding (ii): By the Cauchy-Schwartz inequality, we have

$$\begin{aligned} &|\mathbb{E}_{\nu^*}[\beta_k^{-1} Q_{\omega_k}(s,\cdot) - \beta_k^{-1} Q^{\pi_{\theta_k}}(s,\cdot), \pi^*(\cdot \,|\, s) - \pi_{\theta_k}(\cdot \,|\, s)]| \\ &\quad = \left|\int_{\mathcal{S} \times \mathcal{A}} (\beta_k^{-1} Q_{\omega_k}(s,a) - \beta_k^{-1} Q^{\pi_{\theta_k}}(s,a)) \cdot \left(\frac{\pi^*(a \,|\, s)}{\pi_{\theta_k}(a \,|\, s)} - \frac{\pi_{\theta_k}(a \,|\, s)}{\pi_{\theta_k}(a \,|\, s)}\right) \cdot \frac{\nu^*(s)}{\nu_k(s)}\,\mathrm{d}\sigma_k(s,a)\right| \\ &\quad \le \mathbb{E}_{\sigma_k}[(\beta_k^{-1} Q_{\omega_k}(s,a) - \beta_k^{-1} Q^{\pi_{\theta_k}}(s,a))^2]^{1/2} \cdot \mathbb{E}_{\sigma_k}\left[\left|\frac{\mathrm{d}\sigma^*}{\mathrm{d}\sigma_k} - \frac{\mathrm{d}\nu^*}{\mathrm{d}\nu_k}\right|^2\right]^{1/2} \\ &\quad \le \beta_k^{-1} \Delta_k' \cdot \psi_k^*, \end{aligned} \tag{C.11}$$

where in the last inequality we use the error bound in (C.4) and the definition of ψ_k^* in (C.2). Finally, combining (C.7), (C.8), (C.10), and (C.11), we have

$$
\begin{aligned}
&|\mathbb{E}_{\nu^*}[\log \pi_{\theta_{k+1}}(\cdot \,|\, s) - \log \pi_{k+1}(\cdot \,|\, s), \pi^*(\cdot \,|\, s) - \pi_{\theta_k}(\cdot \,|\, s)]| \\
&\quad \leq \tau_{k+1}^{-1}\Delta_{k+1} \cdot \phi_k^* + \beta_k^{-1}\Delta'_k \cdot \psi_k^*,
\end{aligned}
$$

which concludes the proof of Lemma C.1. □

The following lemma characterizes the difference between $f_{\theta_{k+1}}$ and f_{θ_k}.

Lemma C.2 (Stepwise Energy Difference)**.** Under the same conditions of Lemma C.1, we have

$$
\mathbb{E}_{\nu^*}[\|\tau_{k+1}^{-1} f_{\theta_{k+1}}(s, \cdot) - \tau_k^{-1} f_{\theta_k}(s, \cdot)\|_\infty^2] \leq 2\Delta'_k + 2\beta_k^{-2}M,
$$

where $\varepsilon'_k = |\mathcal{A}| \cdot \tau_{k+1}^{-2}\Delta_{k+1}^2$ and $M = 2\mathbb{E}_{\nu^*}[\max_{a\in\mathcal{A}}(Q_{\omega_0}(s,a))^2] + 2R_f^2$.

Proof. See Appendix **??** for a detailed proof. □

Intuitively, the bounded difference between $f_{\theta_{k+1}}$ and $f_{\theta_{k+1}}$ quantified in Lemma C.2 is due to the KL-regularization in (**??**), which keeps the updated policy $\pi_{\theta_{k+1}}$ from being too far away from the current policy π_{θ_k}.

The differences characterized in Lemmas C.1 and C.2 play key roles in establishing the global convergence of neural DROPAC.

Proof of Lemma C.2. By the triangle inequality, we have

$$
\begin{aligned}
&\|\tau_{k+1}^{-1} f_{\theta_{k+1}}(s, \cdot) - \tau_k^{-1} f_{\theta_k}(s, \cdot)\|_\infty^2 \\
&\quad \leq 2\|\tau_{k+1}^{-1} f_{\theta_{k+1}}(s, \cdot) - \tau_k^{-1} f_{\theta_k}(s, \cdot) - \beta_k^{-1} Q_{\omega_k}(s, \cdot)\|_\infty^2 + 2\|\beta_k^{-1} Q_{\omega_k}(s, \cdot)\|_\infty^2. \qquad \text{(C.12)}
\end{aligned}
$$

For the first term on the right-hand side of (C.12), we have

$$
\mathbb{E}_{\nu^*}[\|\tau_{k+1}^{-1} f_{\theta_{k+1}}(s, \cdot) - \tau_k^{-1} f_{\theta_k}(s, \cdot) - \beta_k^{-1} Q_{\omega_k}(s, \cdot)\|_\infty^2] \leq |\mathcal{A}| \cdot \tau_{k+1}^{-2}\Delta_{k+1}^2. \qquad \text{(C.13)}
$$

For the second term on the right-hand side of (C.12), we have

$$
\mathbb{E}_{\nu^*}[\|\beta_k^{-1} Q_{\omega_k}(s, \cdot)\|_\infty^2] \leq \beta_k^{-2} \cdot \mathbb{E}_{\nu^*}\Big[\max_{a\in\mathcal{A}} 2(Q_{\omega_0}(s,a))^2 + 2R_f^2\Big] = \beta_k^{-2}M, \qquad \text{(C.14)}
$$

where we use the 1-Lipschitz continuity of Q_ω in ω and the constraint $\|\omega_k - \omega_0\|_2 \leq R_\omega$. Then, taking expectation with respect to $s \sim \nu^*$ on the both sides of (C.12) and plugging in (C.13) and (C.14), we finish the proof of Lemma C.2. □

D GLOBAL CONVERGENCE OF DROPAC

We track the progress of neural PPO in Algorithm **??** using the expected total reward

$$
\mathcal{L}(\pi) = \mathbb{E}_{\nu^*}[V^\pi(s)] = \mathbb{E}_{\nu^*}[Q^\pi(s, \cdot), \pi(\cdot \,|\, s)], \qquad \text{(D.1)}
$$

where ν^* is the stationary state distribution of the optimal policy π^*. The following theorem characterizes the global convergence of $\mathcal{L}(\pi_{\theta_k})$ towards $\mathcal{L}(\pi^*)$. Recall that T_f and T_Q are the numbers of SGD and TD iterations in Lines **??** and **??** of Algorithm **??**, while ϕ_k^* and ψ_k^* are defined in (C.2).

Theorem D.1 (Global Rate of Convergence of Neural PPO)**.** Suppose that Assumptions **??**, **??**, and **??** hold. For the policy sequence $\{\pi_{\theta_k}\}_{k=1}^K$ attained by neural PPO in Algorithm **??**, we have

$$
\min_{0\leq k\leq K}\big\{\mathcal{L}(\pi^*) - \mathcal{L}(\pi_{\theta_k})\big\} \leq \frac{\beta^2\log|\mathcal{A}| + M + \beta^2\sum_{k=0}^{K-1}(\varepsilon_k + \varepsilon'_k)}{(1-\gamma)\beta\cdot\sqrt{K}}.
$$

Here $\varepsilon_k = \tau_{k+1}^{-1}\Delta_{k+1} \cdot \phi_k^* + \beta_k^{-1}\Delta'_k \cdot \psi_k^*$ and $\varepsilon'_k = |\mathcal{A}| \cdot \tau_{k+1}^{-2}\Delta_{k+1}^2$, where

$$
\Delta_{k+1} = \mathcal{O}(R_f^2 T^{-1/2} + R_f^{5/2} m_f^{-1/4} + R_f^3 m_f^{-1/2}), \;\; \epsilon'_k = \mathcal{O}(R_Q^2 T^{-1/2} + R_Q^{5/2} m_Q^{-1/4} + R_Q^3 m_Q^{-1/2}).
$$

Also, we have $M = 2\mathbb{E}_{\nu^*}[\max_{a\in\mathcal{A}}(Q_{\omega_0}(s,a))^2] + 2R_f^2$.

Proof. See Section **??** for a detailed proof of Theorem D.1. □

We first present the performance difference lemma of **?**. Recall that the expected total reward $\mathcal{L}(\pi)$ is defined in (D.1) and ν^* is the stationary state distribution of the optimal policy π^*.

Lemma D.2 (Performance Difference)**.** For $\mathcal{L}(\pi)$ defined in (D.1), we have

$$\mathcal{L}(\pi) - \mathcal{L}(\pi^*) = (1-\gamma)^{-1} \cdot \mathbb{E}_{\nu^*}[Q^\pi(s,\cdot), \pi(\cdot \,|\, s) - \pi^*(\cdot \,|\, s)].$$

Proof. See Appendix **??** for a detailed proof. □

Proof of Lemma D.2. The proof follows that of Lemma 6.1 in **?**. By the definition of $V^\pi(s)$ in (**??**), we have

$$\begin{aligned}
\mathbb{E}_{\nu^*}[V^{\pi^*}(s)] &= \sum_{t=0}^{\infty} \gamma^t \cdot \mathbb{E}_{a_t \sim \pi^*(\cdot \,|\, s_t), s_t \sim (\mathcal{P}^{\pi^*})^t \nu^*}\bigl[(1-\gamma)\cdot r(s_t,a_t)\bigr] \qquad \text{(D.2)}\\
&= \sum_{t=0}^{\infty} \gamma^t \cdot \mathbb{E}_{a_t \sim \pi^*(\cdot \,|\, s_t), s_t \sim (\mathcal{P}^{\pi^*})^t \nu^*}\bigl[(1-\gamma)\cdot r(s_t,a_t) + V^\pi(s_t) - V^\pi(s_t)\bigr]\\
&= \sum_{t=0}^{\infty} \gamma^t \cdot \mathbb{E}_{s_{t+1}\sim \mathcal{P}(\cdot \,|\, s_t,a_t), a_t \sim \pi^*(\cdot \,|\, s_t), s_t \sim (\mathcal{P}^{\pi^*})^t \nu^*}\bigl[(1-\gamma)\cdot r(s_t,a_t) + \gamma \cdot V^\pi(s_{t+1}) - V^\pi(s_t)\bigr]\\
&\quad + \mathbb{E}_{\nu^*}[V^\pi(s)],
\end{aligned}$$

where the third inequality is obtained by taking $\mathbb{E}_{\nu^*}[V^\pi(s_0)] = \mathbb{E}_{\nu^*}[V^\pi(s)]$ out and, correspondingly, delaying $V^\pi(s_t)$ by one time step to $V^\pi(s_{t+1})$ in each term of the summation. Note that for the advantage function, by definition of the action-value function, we have

$$A^\pi(s,a) = Q^\pi(s,a) - V^\pi(s) = (1-\gamma)\cdot r(s,a) + \gamma \cdot \mathbb{E}_{s' \sim \mathcal{P}(\cdot \,|\, s,a)}[V^\pi(s')] - V^\pi(s),$$

which together with (D.2) implies

$$\begin{aligned}
\mathbb{E}_{\nu^*}[V^{\pi^*}(s)] &= \sum_{t=0}^{\infty} \gamma^t \cdot \mathbb{E}_{a_t \sim \pi^*(\cdot \,|\, s_t), s_t \sim (\mathcal{P}^{\pi^*})^t \nu^*}[A^\pi(s_t,a_t)] + \mathbb{E}_{\nu^*}[V^\pi(s)]\\
&= (1-\gamma)^{-1} \cdot \mathbb{E}_{\sigma^*}[A^\pi(s,a)] + \mathbb{E}_{\nu^*}[V^\pi(s)]. \qquad \text{(D.3)}
\end{aligned}$$

Here the second equality follows from $(\mathcal{P}^{\pi^*})^t \nu^* = \nu^*$ for any $t \geq 0$ and $\sigma^* = \pi^* \nu^*$. Finally, note that for any given $s \in \mathcal{S}$,

$$\begin{aligned}
\mathbb{E}_{\pi^*}[A^\pi(s,a)] = \mathbb{E}_{\pi^*}[Q^\pi(s,a) - V^\pi(s)] &= Q^\pi(s,\cdot), \pi^*(\cdot \,|\, s) - Q^\pi(s,\cdot), \pi(\cdot \,|\, s)\\
&= Q^\pi(s,\cdot), \pi^*(\cdot \,|\, s) - \pi(\cdot \,|\, s). \qquad \text{(D.4)}
\end{aligned}$$

Plugging (D.4) into (D.3) and recalling the definition of $\mathcal{L}(\pi)$ in (D.1), we finish the proof of Lemma D.2. □

Since the optimal policy π^* maximizes the value function $V^\pi(s)$ with respect to π for any $s \in \mathcal{S}$, we have $\mathcal{L}(\pi^*) = \mathbb{E}_{\nu^*}[V^{\pi^*}(s)] \geq \mathbb{E}_{\nu^*}[V^\pi(s)] = \mathcal{L}(\pi)$ for any π. As a result, we have

$$\mathbb{E}_{\nu^*}[Q^\pi(s,\cdot), \pi(\cdot \,|\, s) - \pi^*(\cdot \,|\, s)] \leq 0, \quad \text{for any } \pi. \qquad \text{(D.5)}$$

Under the variational inequality framework (**?**), (D.5) corresponds to the monotonicity of the mapping Q^π evaluated at π^* and any π. Note that the classical notion of monotonicity requires the evaluation at any pair π' and π, while we restrict π' to π^* in (D.5). Hence, we refer to (D.5) as one-point monotonicity. In the context of nonconvex optimization, the mapping Q^π can be viewed as the gradient of $\mathcal{L}(\pi)$ at π, which lives in the dual space, while π lives in the primal space. Another condition related to (D.5) in nonconvex optimization is known as dissipativity (**?**).

The following lemma establishes the one-step descent of the KL-divergence in the infinite-dimensional policy space, which follows from the analysis of mirror descent (**??**) as well as the fact that given any ν_k, the subproblem of policy improvement in (C.1) can be solved for each $s \in \mathcal{S}$ individually.

Lemma D.3 (One-Step Descent)**.** For the ideal improved policy π_{k+1} defined in (C.1) and the current policy π_{θ_k}, we have that, for any $s \in \mathcal{S}$,

$$\begin{aligned}
&\mathrm{KL}(\pi^*(\cdot \,|\, s) \,\|\, \pi_{\theta_{k+1}}(\cdot \,|\, s)) - \mathrm{KL}(\pi^*(\cdot \,|\, s) \,\|\, \pi_{\theta_k}(\cdot \,|\, s))\\
&\quad \leq \log(\pi_{\theta_{k+1}}(\cdot \,|\, s)/\pi_{k+1}(\cdot \,|\, s)), \pi_{\theta_k}(\cdot \,|\, s) - \pi^*(\cdot \,|\, s) - \beta_k^{-1} \cdot Q^{\pi_{\theta_k}}(s,\cdot), \pi^*(\cdot \,|\, s) - \pi_{\theta_k}(\cdot \,|\, s)\\
&\qquad - 1/2 \cdot \|\pi_{\theta_{k+1}}(\cdot \,|\, s) - \pi_{\theta_k}(\cdot \,|\, s)\|_1^2 - \tau_{k+1}^{-1} f_{\theta_{k+1}}(s,\cdot) - \tau_k^{-1} f_{\theta_k}(s,\cdot), \pi_{\theta_k}(\cdot \,|\, s) - \pi_{\theta_{k+1}}(\cdot \,|\, s).
\end{aligned}$$

Proof. See Appendix **??** for a detailed proof. □

Proof of Lemma D.3. First, we have

$$\begin{aligned}
&\mathrm{KL}(\pi^*(\cdot\,|\,s)\,\|\,\pi_{\theta_k}(\cdot\,|\,s)) - \mathrm{KL}(\pi^*(\cdot\,|\,s)\,\|\,\pi_{\theta_{k+1}}(\cdot\,|\,s))\\
&\quad= \log(\pi_{\theta_{k+1}}(\cdot\,|\,s)/\pi_{\theta_k}(\cdot\,|\,s)), \pi^*(\cdot\,|\,s)\\
&\quad= \log(\pi_{\theta_{k+1}}(\cdot\,|\,s)/\pi_{\theta_k}(\cdot\,|\,s)), \pi^*(\cdot\,|\,s) - \pi_{\theta_{k+1}}(\cdot\,|\,s) + \mathrm{KL}(\pi_{\theta_{k+1}}(\cdot\,|\,s)\,\|\,\pi_{\theta_k}(\cdot\,|\,s))\\
&\quad= \log(\pi_{\theta_{k+1}}(\cdot\,|\,s)/\pi_{\theta_k}(\cdot\,|\,s)) - \beta_k^{-1}Q^{\pi_{\theta_k}}(s,\cdot), \pi^*(\cdot\,|\,s) - \pi_{\theta_k}(\cdot\,|\,s)\\
&\qquad\quad + \beta_k^{-1}\cdot Q^{\pi_{\theta_k}}(s,\cdot), \pi^*(\cdot\,|\,s) - \pi_{\theta_k}(\cdot\,|\,s) + \mathrm{KL}(\pi_{\theta_{k+1}}(\cdot\,|\,s)\,\|\,\pi_{\theta_k}(\cdot\,|\,s))\\
&\qquad\quad + \log(\pi_{\theta_{k+1}}(\cdot\,|\,s)/\pi_{\theta_k}(\cdot\,|\,s)), \pi_{\theta_k}(\cdot\,|\,s) - \pi_{\theta_{k+1}}(\cdot\,|\,s).
\end{aligned} \tag{D.6}$$

Recall that $\pi_{k+1} \propto \exp\{\tau_k^{-1}f_{\theta_k} + \beta_k^{-1}Q^{\pi_{\theta_k}}\}$ and $Z_{k+1}(s)$ and $Z_{\theta_k}(s)$ are defined in (C.6). Also recall that we have $\log Z_{\theta_k}(s), \pi(\cdot\,|\,s) - \pi'(\cdot\,|\,s) = \log Z_k(s), \pi(\cdot\,|\,s) - \pi'(\cdot\,|\,s) = 0$ for all k, π, and π', which implies that, on the right-hand-side of (D.6),

$$\begin{aligned}
&\log\pi_{\theta_k}(\cdot\,|\,s) + \beta_k^{-1}Q^{\pi_{\theta_k}}(s,\cdot), \pi^*(\cdot\,|\,s) - \pi_{\theta_k}(\cdot\,|\,s)\\
&\quad= \tau_k^{-1}f_{\theta_k}(s,\cdot) + \beta_k^{-1}Q^{\pi_{\theta_k}}(s,\cdot), \pi^*(\cdot\,|\,s) - \pi_{\theta_k}(\cdot\,|\,s) - \log Z_{\theta_k}(s), \pi^*(\cdot\,|\,s) - \pi_{\theta_k}(\cdot\,|\,s)\\
&\quad= \tau_k^{-1}f_{\theta_k}(s,\cdot) + \beta_k^{-1}Q^{\pi_{\theta_k}}(s,\cdot), \pi^*(\cdot\,|\,s) - \pi_{\theta_k}(\cdot\,|\,s) - \log Z_{k+1}(s), \pi^*(\cdot\,|\,s) - \pi_{\theta_k}(\cdot\,|\,s)\\
&\quad= \log\pi_{k+1}(\cdot\,|\,s), \pi^*(\cdot\,|\,s) - \pi_{\theta_k}(\cdot\,|\,s),
\end{aligned} \tag{D.7}$$

and

$$\begin{aligned}
&\log(\pi_{\theta_{k+1}}(\cdot\,|\,s)/\pi_{\theta_k}(\cdot\,|\,s)), \pi_{\theta_k}(\cdot\,|\,s) - \pi_{\theta_{k+1}}(\cdot\,|\,s)\\
&\quad= \tau_{k+1}^{-1}f_{\theta_{k+1}}(s,\cdot) - \tau_k^{-1}f_{\theta_k}(s,\cdot), \pi_{\theta_k}(\cdot\,|\,s) - \pi_{\theta_{k+1}}(\cdot\,|\,s)\\
&\qquad - \log Z_{\theta_{k+1}}(s), \pi_{\theta_k}(\cdot\,|\,s) - \pi_{\theta_{k+1}}(\cdot\,|\,s) + \log Z_{\theta_k}(s), \pi_{\theta_k}(\cdot\,|\,s) - \pi_{\theta_{k+1}}(\cdot\,|\,s)\\
&\quad= \tau_{k+1}^{-1}f_{\theta_{k+1}}(s,\cdot) - \tau_k^{-1}f_{\theta_k}(s,\cdot), \pi_{\theta_k}(\cdot\,|\,s) - \pi_{\theta_{k+1}}(\cdot\,|\,s).
\end{aligned} \tag{D.8}$$

Plugging (D.7) and (D.8) into (D.6), we obtain

$$\begin{aligned}
&\mathrm{KL}(\pi^*(\cdot\,|\,s)\,\|\,\pi_{\theta_k}(\cdot\,|\,s)) - \mathrm{KL}(\pi^*(\cdot\,|\,s)\,\|\,\pi_{\theta_{k+1}}(\cdot\,|\,s)) \qquad\qquad (\mathrm{D.9})\\
&\quad= \log(\pi_{\theta_{k+1}}(\cdot\,|\,s)/\pi_{k+1}(\cdot\,|\,s)), \pi^*(\cdot\,|\,s) - \pi_{\theta_k}(\cdot\,|\,s) + \beta_k^{-1}\cdot Q^{\pi_{\theta_k}}(s,\cdot), \pi^*(\cdot\,|\,s) - \pi_{\theta_k}(\cdot\,|\,s)\\
&\qquad + \tau_{k+1}^{-1}f_{\theta_{k+1}}(s,\cdot) - \tau_k^{-1}f_{\theta_k}(s,\cdot), \pi_{\theta_k}(\cdot\,|\,s) - \pi_{\theta_{k+1}}(\cdot\,|\,s) + \mathrm{KL}(\pi_{\theta_{k+1}}(\cdot\,|\,s)\,\|\,\pi_{\theta_k}(\cdot\,|\,s))\\
&\quad\geq \log(\pi_{\theta_{k+1}}(\cdot\,|\,s)/\pi_{k+1}(\cdot\,|\,s)), \pi^*(\cdot\,|\,s) - \pi_{\theta_k}(\cdot\,|\,s) + \beta_k^{-1}\cdot Q^{\pi_{\theta_k}}(s,\cdot), \pi^*(\cdot\,|\,s) - \pi_{\theta_k}(\cdot\,|\,s)\\
&\qquad + \tau_{k+1}^{-1}f_{\theta_{k+1}}(s,\cdot) - \tau_k^{-1}f_{\theta_k}(s,\cdot), \pi_{\theta_k}(\cdot\,|\,s) - \pi_{\theta_{k+1}}(\cdot\,|\,s) + 1/2\cdot\|\pi_{\theta_{k+1}}(\cdot\,|\,s) - \pi_{\theta_k}(\cdot\,|\,s)\|_1^2,
\end{aligned}$$

where in the last inequality we use the Pinsker's inequality. Rearranging the terms in (D.9), we finish the proof of Lemma D.3. □

Based on Lemmas D.2 and D.3, we prove Theorem D.1 by casting neural PPO as infinite-dimensional mirror descent with primal and dual errors, whose impact is characterized in Lemma C.1. In particular, we employ the ℓ_1-ℓ_∞ pair of primal-dual norms.

Proof of Theorem D.1. Taking expectation with respect to $s \sim \nu^*$ and invoking Lemmas C.1 and D.3, we have

$$\begin{aligned}
&\mathbb{E}_{\nu^*}[\mathrm{KL}(\pi^*(\cdot\,|\,s)\,\|\,\pi_{\theta_{k+1}}(\cdot\,|\,s))] - \mathbb{E}_{\nu^*}[\mathrm{KL}(\pi^*(\cdot\,|\,s)\,\|\,\pi_{\theta_k}(\cdot\,|\,s))]\\
&\quad\leq \varepsilon_k - \beta_k^{-1}\cdot\mathbb{E}_{\nu^*}[Q^{\pi_{\theta_k}}(s,\cdot), \pi^*(\cdot\,|\,s) - \pi_{\theta_k}(\cdot\,|\,s)] - 1/2\cdot\mathbb{E}_{\nu^*}[\|\pi_{\theta_{k+1}}(\cdot\,|\,s) - \pi_{\theta_k}(\cdot\,|\,s)\|_1^2]\\
&\qquad - \mathbb{E}_{\nu^*}[\tau_{k+1}^{-1}f_{\theta_{k+1}}(s,\cdot) - \tau_k^{-1}f_{\theta_k}(s,\cdot), \pi_{\theta_k}(\cdot\,|\,s) - \pi_{\theta_{k+1}}(\cdot\,|\,s)].
\end{aligned}$$

By Lemma D.2 and the Hölder's inequality, we further have

$$\begin{aligned}
&\mathbb{E}_{\nu^*}[\mathrm{KL}(\pi^*(\cdot\,|\,s)\,\|\,\pi_{\theta_{k+1}}(\cdot\,|\,s))] - \mathbb{E}_{\nu^*}[\mathrm{KL}(\pi^*(\cdot\,|\,s)\,\|\,\pi_{\theta_k}(\cdot\,|\,s))]\\
&\quad\leq \varepsilon_k - (1-\gamma)\beta_k^{-1}\cdot(\mathcal{L}(\pi^*) - \mathcal{L}(\pi_{\theta_k})) - 1/2\cdot\mathbb{E}_{\nu^*}[\|\pi_{\theta_{k+1}}(\cdot\,|\,s) - \pi_{\theta_k}(\cdot\,|\,s)\|_1^2]\\
&\qquad + \mathbb{E}_{\nu^*}\big[\|\tau_{k+1}^{-1}f_{\theta_{k+1}}(s,\cdot) - \tau_k^{-1}f_{\theta_k}(s,\cdot)\|_\infty\cdot\|\pi_{\theta_k}(\cdot\,|\,s) - \pi_{\theta_{k+1}}(\cdot\,|\,s)\|_1\big]\\
&\quad\leq \varepsilon_k - (1-\gamma)\beta_k^{-1}\cdot(\mathcal{L}(\pi^*) - \mathcal{L}(\pi_{\theta_k})) + 1/2\cdot\mathbb{E}_{\nu^*}[\|\tau_{k+1}^{-1}f_{\theta_{k+1}}(s,\cdot) - \tau_k^{-1}f_{\theta_k}(s,\cdot)\|_\infty^2]\\
&\quad\leq \varepsilon_k - (1-\gamma)\beta_k^{-1}\cdot(\mathcal{L}(\pi^*) - \mathcal{L}(\pi_{\theta_k})) + (\varepsilon_k' + \beta_k^{-2}M),
\end{aligned} \tag{D.10}$$

where in the second inequality we use $2xy - y^2 \leq x^2$ and in the last inequality we use Lemma C.2. Rearranging the terms in (D.10), we have

$$\begin{aligned}
&(1-\gamma)\beta_k^{-1}\cdot(\mathcal{L}(\pi^*) - \mathcal{L}(\pi_{\theta_k})) \qquad\qquad (\mathrm{D.11})\\
&\quad\leq \mathbb{E}_{\nu^*}[\mathrm{KL}(\pi^*(\cdot\,|\,s)\,\|\,\pi_{\theta_{k+1}}(\cdot\,|\,s))] - \mathbb{E}_{\nu^*}[\mathrm{KL}(\pi^*(\cdot\,|\,s)\,\|\,\pi_{\theta_k}(\cdot\,|\,s))] + \beta_k^{-2}M + \varepsilon_k + \varepsilon_k'.
\end{aligned}$$

Telescoping (D.11) for $k+1 \in [K]$, we obtain

$$\begin{aligned}
&\sum_{k=0}^{K-1} (1-\gamma)\beta_k^{-1} \cdot \big(\mathcal{L}(\pi_{\theta_k}) - \mathcal{L}(\pi^*)\big) \\
&\quad \le \mathbb{E}_{\nu^*}\big[\mathrm{KL}\big(\pi^*(\cdot \,|\, s) \,\|\, \pi_{\theta_K}(\cdot \,|\, s)\big)\big] - \mathbb{E}_{\nu^*}\big[\mathrm{KL}\big(\pi^*(\cdot \,|\, s) \,\|\, \pi_{\theta_0}(\cdot \,|\, s)\big)\big] \\
&\qquad + M\sum_{k=0}^{K-1} \beta_k^{-2} + \sum_{k=0}^{K-1} (\varepsilon_k + \varepsilon_k').
\end{aligned}$$

Note that we have (i) $\sum_{k=0}^{K-1} \beta_k^{-1} \cdot \big(\mathcal{L}(\pi^*) - \mathcal{L}(\pi_{\theta_k})\big) \ge \big(\sum_{k=0}^{K-1} \beta_k^{-1}\big) \cdot \min_{0 \le k \le K}\{\mathcal{L}(\pi^*) - \mathcal{L}(\pi_{\theta_k})\}$, (ii) $\mathbb{E}_{\nu^*}[\mathrm{KL}(\pi^*(\cdot \,|\, s) \,\|\, \pi_{\theta_0}(\cdot \,|\, s))] \le \log|\mathcal{A}|$ due to the uniform initialization of policy, and that (iii) the KL-divergence is nonnegative. Hence, we have

$$\min_{0 \le k \le K} \big\{\mathcal{L}(\pi^*) - \mathcal{L}(\pi_{\theta_k})\big\} \le \frac{\log|\mathcal{A}| + M\sum_{k=0}^{K-1} \beta_k^{-2} + \sum_{k=0}^{K-1} (\varepsilon_k + \varepsilon_k')}{(1-\gamma)\sum_{k=0}^{K-1} \beta_k^{-1}}. \tag{D.12}$$

Setting the penalty parameter $\beta_k = \beta\sqrt{K}$, we have $\sum_{k=0}^{K-1} \beta_k^{-1} = \beta^{-1}\sqrt{K}$ and $\sum_{k=0}^{K-1} \beta_k^{-2} = \beta^{-2}$, which together with (D.12) concludes the proof of Theorem D.1. □