CUP: A CONSERVATIVE UPDATE POLICY ALGORITHM FOR SAFE REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Safe reinforcement learning (RL) is still very challenging since it requires the agent to consider both return maximization and safe exploration. In this paper, we propose CUP, a **C**onservative **U**pdate **P**olicy algorithm with a theoretical safety guarantee. The derivation of CUP is based on surrogate functions with respect to our new proposed bounds. Although using bounds as surrogate functions to design safe RL algorithms have appeared in some existing works, we develop it at least three aspects: **(i)** We provide a rigorous theoretical analysis to extend the bounds with respect to generalized advantage estimator (GAE). GAE significantly reduces variance while maintains a tolerable level of bias, which is an efficient step for us to design CUP; **(ii)** The proposed bounds are more compact than existing works, i.e., using the proposed bounds as surrogate functions are better local approximations to the objective and constraints. **(iii)** The bound of worst-case safe constraint violation of CUP is more compact than the existing safe RL algorithms, which explains why CUP is so good in practice. Finally, extensive experiments show the effectiveness of CUP where the agent satisfies safe constraints.

1 INTRODUCTION

Reinforcement learning (RL) (Sutton & Barto, 1998) has achieved significant successes in many fields (Mnih et al., 2015; Silver et al., 2017; OpenAI, 2019; Afsar et al., 2021), robotics (Deisenroth et al., 2013), playing Go (Silver et al., 2016; 2017), Starcraft (Vinyals et al., 2019), Dota (OpenAI, 2019), and recommendation system (Afsar et al., 2021). However, most RL algorithms improve the performance under the assumption that an agent is free to explore any behaviors. In real-world applications, only considering return maximization is not enough, and we also need to consider safe behaviors. For example, a robot agent should avoid playing actions that irrevocably harm its hardware, and a recommender system should avoid presenting offending items to users. Thus, it is crucial to consider *safe exploration* for RL, which is usually formulated as constrained Markov decision processes (CMDP) (Altman, 1999).

It is challenging to solve CMDP since traditional approaches (e.g., Q-learning (Watkins, 1989) & policy gradient (Williams, 1992)) usually violate the safe exploration constraints, which is undesirable for safe RL. Recently, Achiam et al. (2017); Yang et al. (2020); Bharadhwaj et al. (2021) suggest to use some surrogate functions to replace the objective and constraints. However, their implementations involve some convex approximations to the non-convex objective and safe constraints, which leads to many error sources and troubles. Concretely, Achiam et al. (2017); Yang et al. (2020); Bharadhwaj et al. (2021) approximate the non-convex objective (or constraints) with first-order or second Taylor expansion, but their implementations still lack a theory to show the error difference between the original objective (or constraints) and its convex approximations. Besides, their approaches involve the inverse of a high-dimension Fisher information matrix, which causes their algorithms to require a costly computation for each update when solving high-dimensional RL problems.

Our Main Work. To address above problems, we propose the *conservative update policy* (CUP) algorithm with a theoretical safety guarantee. We derive the CUP bases on some new proposed surrogate functions with respect to objective and constraints and provide a practical implementation of CUP that does not depend on any convex approximation to adapt high-dimensional safe RL.

Concretely, in Section 3, Theorem 1 shows generalized difference bounds between two arbitrary policies for the objective and constraints. Those bounds provide principled approximations to the

objective and constraints, which are theoretical foundations for us to use those bounds as surrogate functions to replace objective and constraints to design algorithms.

Although using difference bound to replace objective or constraints has appeared in some existing works (e.g., (Kakade & Langford, 2002; Schulman et al., 2015; Achiam et al., 2017)), Theorem 1 improves their bounds at least two aspects: **(i)** Firstly, our rigorous theoretical analysis extends the bound with respect to generalized advantage estimator (GAE) (Schulman et al., 2016). GAE significantly reduces variance while maintains a tolerable level of bias, which is one of the critical steps for us to design efficient algorithms in the later section. Although Zhang et al. (2020); Kang et al. (2021) have applied GAE to solve safe RL problems, their approaches are empirical and lack a theoretical analysis with respect to GAE. Thus, our result provides a theory to illustrate the effectiveness of the work (Zhang et al., 2020; Kang et al., 2021). **(ii)** Our new bounds refine classic difference bounds. For example, our bounds are more compact than Achiam et al. (2017), i,e., using our new bounds as surrogate functions are better local approximations to the objective and constraints. Besides, the surrogate functions with respect to our new bounds are more accessible to be estimated from the samples than the approaches appears in (Kakade & Langford, 2002; Schulman et al., 2015)), for more discussions, please see Remark 1.

In Section 4, we provide the necessary details of the proposed CUP. The CUP contains two steps: it performs a policy improvement at first, then it projects the policy back onto the safe region to reconcile the constraint violation. Theorem 2 shows a lower bound on policy improvement and an upper bound on constraint violation for CUP at each update. Notably, the result in Theorem 2 shows the bound of CUP is more compact than state-of-the-art safe RL algorithms: CPO (Achiam et al., 2017, Proposition 1-2), PCPO (Yang et al., 2020, Theorem 1) and FOCOPS (Zhang et al., 2020), which provides a partial explanation for why CUP is so good in practice. For more discussions, please refer to Remark 2. Finally, we provide a practical implementation of sample-based CUP. Such an implementation allows us to use deep neural networks to train a model. Mainly, CUP does not depend on any convex approximation for objective and constraints, and it optimizes the objective according to the first-order optimizer. Extensive high-dimensional experiments on continuous control tasks show the effectiveness of CUP where the agent satisfies safe constraints.

2 PRELIMINARIES

Reinforcement learning (RL) (Sutton & Barto, 1998) is often formulated as a *Markov decision process* (MDP) (Puterman, 2014) that is a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathbb{P}, r, \rho_0, \gamma)$. Here $\mathcal{S}$ is state space, $\mathcal{A}$ is action space. $\mathbb{P}(s^{'}|s,a)$ is probability of state transition from s to $s^{'}$ after playing a. $r(\cdot): \mathcal{S} \times \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, and $r(s'|s,a)$ denotes the reward that the agent observes when state transition from s to $s^{'}$ after it plays a. $\rho_0(\cdot): \mathcal{S} \to [0,1]$ is the initial state distribution and $\gamma \in (0,1)$.

A stationary parameterized policy $\pi_{\boldsymbol{\theta}}$ is a probability distribution defined on $\mathcal{S} \times \mathcal{A}$, $\pi_{\boldsymbol{\theta}}(a|s)$ denotes the probability of playing a in state s. We use $\Pi_{\boldsymbol{\theta}}$ to denote the set of all stationary policies, where $\Pi_{\boldsymbol{\theta}} = \{\pi_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \mathbb{R}^p\}$, and $\boldsymbol{\theta}$ is a parameter needed to be learned. Let $\mathbf{P}_{\pi_{\boldsymbol{\theta}}} \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$ be a state transition probability matrix, and their components are: $\mathbf{P}_{\pi_{\boldsymbol{\theta}}}[s,s'] = \sum_{a\in\mathcal{A}} \pi_{\boldsymbol{\theta}}(a|s)\mathbb{P}(s'|s,a) =: \mathbb{P}_{\pi_{\boldsymbol{\theta}}}(s^{'}|s)$, which denotes one-step state transformation probability from s to $s^{'}$ by executing $\pi_{\boldsymbol{\theta}}$. Let $\tau = \{s_t, a_t, r_{t+1}\}_{t\geq 0} \sim \pi_{\boldsymbol{\theta}}$ be a trajectory generated by $\pi_{\boldsymbol{\theta}}$, where $s_0 \sim \rho_0(\cdot)$, $a_t \sim \pi_{\boldsymbol{\theta}}(\cdot|s_t)$, $s_{t+1} \sim \mathbb{P}(\cdot|s_t, a_t)$, and $r_{t+1} = r(s_{t+1}|s_t, a_t)$. We use $\mathbb{P}_{\pi_{\boldsymbol{\theta}}}(s_t = s^{'}|s)$ to denote the probability of visiting the state $s^{'}$ after t time steps from the state s by executing $\pi_{\boldsymbol{\theta}}$. Due to the Markov property in MDP, $\mathbb{P}_{\pi_{\boldsymbol{\theta}}}(s_t = s^{'}|s)$ is $(s, s^{'})$-th component of the matrix $\mathbf{P}^t_{\pi_{\boldsymbol{\theta}}}$, i.e., $\mathbb{P}_{\pi_{\boldsymbol{\theta}}}(s_t = s^{'}|s) = \mathbf{P}^t_{\pi_{\boldsymbol{\theta}}}[s, s^{'}]$. Finally, let $d^{s_0}_{\pi_{\boldsymbol{\theta}}}(s) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \mathbb{P}_{\pi_{\boldsymbol{\theta}}}(s_t = s|s_0)$ be the stationary state distribution of the Markov chain (starting at s_0) induced by policy $\pi_{\boldsymbol{\theta}}$. We define $d^{\rho_0}_{\pi_{\boldsymbol{\theta}}}(s) = \mathbb{E}_{s_0\sim\rho_0(\cdot)}[d^{s_0}_{\pi_{\boldsymbol{\theta}}}(s)]$ as the discounted state visitation distribution on initial distribution $\rho_0(\cdot)$.

The *state value function* of $\pi_{\boldsymbol{\theta}}$ is defined as $V_{\pi_{\boldsymbol{\theta}}}(s) = \mathbb{E}_{\pi_{\boldsymbol{\theta}}}[\sum_{t=0}^{\infty}\gamma^t r_{t+1}|s_0 = s]$, where $\mathbb{E}_{\pi_{\boldsymbol{\theta}}}[\cdot|\cdot]$ denotes a conditional expectation on actions which are selected by $\pi_{\boldsymbol{\theta}}$. Its *state-action value function* is $Q_{\pi_{\boldsymbol{\theta}}}(s,a) = \mathbb{E}_{\pi_{\boldsymbol{\theta}}}[\sum_{t=0}^{\infty}\gamma^t r_{t+1}|s_0 = s, a_0 = a]$, and advantage function is $A_{\pi_{\boldsymbol{\theta}}}(s,a) = Q_{\pi_{\boldsymbol{\theta}}}(s,a) - V_{\pi_{\boldsymbol{\theta}}}(s)$. The goal of reinforcement learning is to maximize $J(\pi_{\boldsymbol{\theta}})$:

$$J(\pi_{\boldsymbol{\theta}}) = \mathbb{E}_{s\sim d^{\rho_0}_{\pi_{\boldsymbol{\theta}}}(\cdot)}[V_{\pi_{\boldsymbol{\theta}}}(s)]. \tag{1}$$

2.1 POLICY GRADIENT AND GENERALIZED ADVANTAGE ESTIMATOR (GAE)

Policy gradient (Williams, 1992; Sutton et al., 2000) is widely used to solve policy optimization, which maximizes the expected total reward by repeatedly estimating the gradient $g = \nabla J(\pi_{\boldsymbol{\theta}})$. Schulman et al. (2016) summarize several different related expressions for the policy gradient:

$$g = \nabla J(\pi_{\boldsymbol{\theta}}) = \mathbb{E}\left[\sum_{t=0}^{\infty} \Psi_t \nabla \log \pi_{\boldsymbol{\theta}}(a_t|s_t)\right], \tag{2}$$

where Ψ_t can be total discounted reward of the trajectory, value function, advantage function or temporal difference (TD) error. As stated by Schulman et al. (2016), the choice $\Psi_t = A(s_t, a_t)$ yields almost the lowest possible variance, which is consistent with the theoretical analysis (Greensmith et al., 2004; Wu et al., 2018). Furthermore, Schulman et al. (2016) propose generalized advantage estimator (GAE) $\hat{A}_t^{\text{GAE}(\gamma,\lambda)}(s_t, a_t)$ to replace Ψ_t: for any $\lambda \in [0, 1]$,

$$\hat{A}_t^{\text{GAE}(\gamma,\lambda)}(s_t, a_t) = \sum_{\ell=0}^{\infty} (\gamma\lambda)^{\ell} \delta_{t+\ell}^{V}, \tag{3}$$

where $\delta_t^V = r_{t+1} + \gamma V(s_{t+1}) - V(s_t)$ is TD error, and $V(\cdot)$ is an estimator of value function. GAE is an efficient technique for data efficiency and reliable performance of reinforcement learning.

2.2 SAFE REINFORCEMENT LEARNING

Safe RL (Ray et al., 2019) is often formulated as a constrained MDP (CMDP) $\mathcal{M} \cup \mathcal{C}$ (Altman, 1999), which is a standard MDP $\mathcal{M}$ augmented with an additional constraint set $\mathcal{C}$. The set $\mathcal{C} = \{(c_i, b_i)\}_{i=1}^m$, where c_i are cost functions: $c_i : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, and limits are b_i, $i = 1, \cdot, m$. The *cost-return* is defined as: $J^{c_i}(\pi_{\boldsymbol{\theta}}) = \mathbb{E}_{\pi_{\boldsymbol{\theta}}}\left[\sum_{t=0}^{\infty} \gamma^t c_i(s_t, a_t)\right]$, then we define the feasible policy set $\Pi_{\mathcal{C}}$ as:

$$\Pi_{\mathcal{C}} = \cap_{i=1}^m \left\{\pi_{\boldsymbol{\theta}} \in \Pi_{\boldsymbol{\theta}} \text{ and } J^{c_i}(\pi_{\boldsymbol{\theta}}) \leq b_i\right\}.$$

The goal of CMDP is to search the optimal policy $\pi_\star$ such that

$$\pi_\star = \arg\max_{\pi_{\boldsymbol{\theta}} \in \Pi_{\mathcal{C}}} J(\pi_{\boldsymbol{\theta}}). \tag{4}$$

Furthermore, we define value functions, action-value functions, and advantage functions for the auxiliary costs in analogy to $V_{\pi_{\boldsymbol{\theta}}}, Q_{\pi_{\boldsymbol{\theta}}}$, and $A_{\pi_{\boldsymbol{\theta}}}$, with c_i replacing r respectively, we denote them as $V_{\pi_{\boldsymbol{\theta}}}^{c_i}, Q_{\pi_{\boldsymbol{\theta}}}^{c_i}$, and $A_{\pi_{\boldsymbol{\theta}}}^{c_i}$. For example, $V_{\pi_{\boldsymbol{\theta}}}^{c_i}(s) = \mathbb{E}_{\pi_{\boldsymbol{\theta}}}\left[\sum_{t=0}^{\infty} \gamma^t c_i(s_t, a_t)|s_0 = s\right]$. Without loss of generality, we will restrict our discussion to the case of one constraint with a cost function c and upper bound b. Finally, we extend the GAE w.r.t. auxiliary cost function c:

$$\hat{A}_{C,t}^{\text{GAE}(\gamma,\lambda)}(s_t, a_t) = \sum_{\ell=0}^{\infty} (\gamma\lambda)^{\ell} \delta_{t+\ell}^{C}, \tag{5}$$

where $\delta_t^C = r_{t+1} + \gamma C(s_{t+1}) - C(s_t)$ is TD error, and $C(\cdot)$ is an estimator of cost function c.

3 GENERALIZED POLICY PERFORMANCE DIFFERENCE BOUNDS

In this section, we show some generalized policy optimization performance bounds for $J(\pi_{\boldsymbol{\theta}})$ and $J^c(\pi_{\boldsymbol{\theta}})$. The proposed bounds provide some new certain surrogate functions w.r.t. the objective and cost function, which are theoretical foundations for us to design efficient algorithms to improve policy performance and satisfy constraints. Additionally, those bounds refine or extend some existing works (e.g., (Kakade & Langford, 2002; Schulman et al., 2015; Achiam et al., 2017)) to GAE case that significantly reduces variance while maintains a tolerable level of bias, which is one of the key steps for us to propose efficient algorithms in the later section.

Before we present our new bounds, let us revisit a classic result about policy performance difference from (Kakade & Langford, 2002), i.e., the next Eq.(6),

$$J(\pi_{\boldsymbol{\theta}}) - J(\pi_{\boldsymbol{\theta}'}) = (1-\gamma)^{-1} \mathbb{E}_{s \sim d_{\pi_{\boldsymbol{\theta}}}^{\rho_0}(\cdot)} \mathbb{E}_{a \sim \pi_{\boldsymbol{\theta}}(\cdot|s)}[A_{\pi_{\boldsymbol{\theta}'}}(s, a)]. \tag{6}$$

Eq.(6) shows a difference between two arbitrary policies $\pi_{\boldsymbol{\theta}}$ and $\pi_{\boldsymbol{\theta}'}$ with different parameters $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$. However, as stated by Zanger et al. (2021), Eq.(6) is very intractable for sampling-based policy optimization since it requires the data comes from a fixed policy $\pi_{\boldsymbol{\theta}}$. In this section, our new bound will refine the result (6). For more discussions about the difference between our new bound and Eq.(6), please refer to Remark 1.

3.1 SOME ADDITIONAL NOTATIONS

We use a bold lowercase letter to denote a vector, e.g., $\mathbf{a} = (a_1, a_2, \cdots, a_n)$, and its i-th element $\mathbf{a}[i] =: a_i$. Let $\varphi(\cdot) : \mathcal{S} \to \mathbb{R}$ be a function defined on $\mathcal{S}$, $\delta_t^{\varphi} = r(s_{t+1}|s_t, a_t) + \gamma\varphi(s_{t+1}) - \varphi(s_t)$ is TD error w.r.t. $\varphi(\cdot)$. For two arbitrary policies $\pi_{\boldsymbol{\theta}}$ and $\pi_{\boldsymbol{\theta}'}$, we denote $\delta^{\varphi}_{\pi_{\boldsymbol{\theta}},t}(s)$ as the expectation of TD error, and define $\Delta_t^{\varphi}(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}'}, s)$ as the difference between $\delta^{\varphi}_{\pi_{\boldsymbol{\theta}},t}(s)$ and $\delta^{\varphi}_{\pi_{\boldsymbol{\theta}'},t}(s)$: $\forall s \in \mathcal{S}$,

$$\delta^{\varphi}_{\pi_{\boldsymbol{\theta}},t}(s) = \mathop{\mathbb{E}}_{\substack{s_t\sim\mathbb{P}_{\pi_{\boldsymbol{\theta}}}(\cdot|s)\\ a_t\sim\pi_{\boldsymbol{\theta}}(\cdot|s_t)\\ s_{t+1}\sim\mathbb{P}(\cdot|s_t,a_t)}} \left[\delta_t^{\varphi}\right], \Delta_t^{\varphi}(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}'}, s) = \mathop{\mathbb{E}}_{\substack{s_t\sim\mathbb{P}_{\pi_{\boldsymbol{\theta}'}}(\cdot|s)\\ a_t\sim\pi_{\boldsymbol{\theta}'}(\cdot|s_t)\\ s_{t+1}\sim\mathbb{P}(\cdot|s_t,a_t)}} \left[\left(\frac{\pi_{\boldsymbol{\theta}}(a_t|s_t)}{\pi_{\boldsymbol{\theta}'}(a_t|s_t)} - 1\right)\delta_t^{\varphi}\right].$$

Furthermore, we introduce two vectors $\boldsymbol{\delta}^{\varphi}_{\pi_{\boldsymbol{\theta}},t}, \boldsymbol{\Delta}_t^{\varphi}(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}'}) \in \mathbb{R}^{|\mathcal{S}|}$, and their components are:

$$\boldsymbol{\delta}^{\varphi}_{\pi_{\boldsymbol{\theta}},t}[s] = \delta^{\varphi}_{\pi_{\boldsymbol{\theta}},t}(s), \quad \boldsymbol{\Delta}_t^{\varphi}(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}'})[s] = \Delta_t^{\varphi}(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}'}, s). \tag{7}$$

Let matrix $\mathbf{P}^{(\lambda)}_{\pi_{\boldsymbol{\theta}}} = (1-\gamma\lambda)\sum_{t=0}^{\infty}(\gamma\lambda)^t\mathbf{P}^{t+1}_{\pi_{\boldsymbol{\theta}}}$, where $\lambda \in [0, 1]$. It is similar to the normalized discounted distribution $d^{\rho_0}_{\pi_{\boldsymbol{\theta}}}(s)$, we extend it to λ-version and denote it as $d^{\lambda}_{\pi_{\boldsymbol{\theta}}}(s)$:

$$d^{\lambda}_{\pi_{\boldsymbol{\theta}}}(s) = \mathbb{E}_{s_0\sim\rho_0(\cdot)}\left[(1-\tilde{\gamma})\sum_{t=0}^{\infty}\tilde{\gamma}^t\mathbb{P}^{(\lambda)}_{\pi_{\boldsymbol{\theta}}}(s_t = s|s_0)\right],$$

where $\tilde{\gamma} = \frac{\gamma(1-\lambda)}{1-\gamma\lambda}$, the probability $\mathbb{P}^{(\lambda)}_{\pi_{\boldsymbol{\theta}}}(s_t = s|s_0)$ is the (s_0, s)-th component of the matrix product $\left(\mathbf{P}^{(\lambda)}_{\pi_{\boldsymbol{\theta}}}\right)^t$. Finally, we introduce a vector $\mathbf{d}^{\lambda}_{\pi_{\boldsymbol{\theta}}} \in \mathbb{R}^{|\mathcal{S}|}$, and its components are: $\mathbf{d}^{\lambda}_{\pi_{\boldsymbol{\theta}}}[s] = d^{\lambda}_{\pi_{\boldsymbol{\theta}}}(s)$.

3.2 MAIN RESULTS

Theorem 1 (Generalized Policy Performance Difference)**.** *For any function $\varphi(\cdot) : \mathcal{S} \to \mathbb{R}$, for two arbitrary policies $\pi_{\boldsymbol{\theta}}$ and $\pi_{\boldsymbol{\theta}'}$, for any $p, q \in [1, \infty)$ such that $\frac{1}{p} + \frac{1}{q} = 1$, we define two error terms:*

$$\epsilon^{\varphi,(\lambda)}_{p,q,t}(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}'}) =: \|\mathbf{d}^{\lambda}_{\pi_{\boldsymbol{\theta}}} - \mathbf{d}^{\lambda}_{\pi_{\boldsymbol{\theta}'}}\|_p \|\boldsymbol{\delta}^{\varphi}_{\pi_{\boldsymbol{\theta}},t}\|_q, \tag{8}$$

$$L^{\varphi,\pm}_{p,q}(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}'}) =: \frac{1}{1-\tilde{\gamma}}\sum_{t=0}^{\infty}\gamma^t\lambda^t\mathbb{E}_{s\sim d^{\lambda}_{\pi_{\boldsymbol{\theta}'}}(\cdot)}\left[\Delta_t^{\varphi}(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}'}, s) \pm \epsilon^{\varphi,(\lambda)}_{p,q,t}(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}'})\right]. \tag{9}$$

Then, the following bound w.r.t. policy performance difference $J(\pi_{\boldsymbol{\theta}}) - J(\pi_{\boldsymbol{\theta}'})$ holds:

$$L^{\varphi,-}_{p,q,}(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}'}) \le J(\pi_{\boldsymbol{\theta}}) - J(\pi_{\boldsymbol{\theta}'}) \le L^{\varphi,+}_{p,q,}(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}'}). \tag{10}$$

We provide its proof in Appendix E. The bound (10) is *tight*, i.e., if $\pi_{\boldsymbol{\theta}} = \pi_{\boldsymbol{\theta}'}$, all the three terms in Eq.(10) are zero identically. From Eq.(9), we know the performance difference bound $L^{\varphi,\pm}_{p,q}(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}'})$ (10) can be interpreted by two distinct difference parts: **(i)** the first difference part, i.e., the expectation $\Delta_t^{\varphi}(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}'}, s)$, which is determined by the difference between TD errors of $\pi_{\boldsymbol{\theta}}$ and $\pi_{\boldsymbol{\theta}'}$; **(ii)** the second difference part, i.e., the discounted distribution difference $\epsilon^{\varphi,(\lambda)}_{p,q,t}(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}'})$, which is determined by the gap between the normalized discounted distribution of $\pi_{\boldsymbol{\theta}}$ and $\pi_{\boldsymbol{\theta}'}$. Thus, the difference of both TD errors and discounted distribution determine the policy difference $J(\pi_{\boldsymbol{\theta}}) - J(\pi_{\boldsymbol{\theta}'})$.

The different choices of p and q lead Eq.(10) to be different bounds. If $p = 1, q = \infty$, we denote $\epsilon^{\varphi}_{\pi_{\boldsymbol{\theta}},t} =: \|\boldsymbol{\delta}^{\varphi}_{\pi_{\boldsymbol{\theta}},t}\|_q = \max_{s_t\in\mathcal{S}}\mathbb{E}_{a_t\sim\pi_{\boldsymbol{\theta}}(\cdot|s_t), s_{t+1}\sim\mathbb{P}(\cdot|s_t,a_t)}[|\delta_t^{\varphi}|]$, then, according to Lemma 2 (see Appendix E.3), when $p = 1, q = \infty$, then error $\epsilon^{\varphi,(\lambda)}_{p,q,t}(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}'})$ is reduced to:

$$\epsilon^{\varphi,(\lambda)}_{p,q,t}(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}'})\big|_{p=1,q=\infty} \le \frac{1}{1-\tilde{\gamma}}\cdot\frac{\gamma(1-\lambda)\epsilon^{\varphi}_{\pi_{\boldsymbol{\theta}},t}}{|1-2\gamma\lambda|\mathcal{S}||\mathcal{A}||}\mathbb{E}_{s\sim d^{\lambda}_{\pi_{\boldsymbol{\theta}'}}(\cdot)}\left[2D_{\text{TV}}(\pi_{\boldsymbol{\theta}'}, \pi_{\boldsymbol{\theta}})[s]\right],$$

where $D_{\text{TV}}(\pi_{\boldsymbol{\theta}'}, \pi_{\boldsymbol{\theta}})[s]$ is the total variational divergence between action distributions at state s, i.e.,

$$2D_{\text{TV}}(\pi_{\boldsymbol{\theta}'}, \pi_{\boldsymbol{\theta}})[s] = \sum_{a\in\mathcal{A}}\left|\pi_{\boldsymbol{\theta}'}(a|s) - \pi_{\boldsymbol{\theta}}(a|s)\right|.$$

Finally, let $\varphi = V_{\pi_{\boldsymbol{\theta}'}}$, the left side of (10) in Theorem 1 implies a lower bound of performance difference, which illustrates the worse case of approximation error, we present it in Proposition 1.

Proposition 1 (Worse case approximation error). *For any two policies $\pi_{\boldsymbol{\theta}}$ and $\pi_{\boldsymbol{\theta}'}$, let $\epsilon^{V}_{\pi_{\boldsymbol{\theta}}}(\pi_{\boldsymbol{\theta}'}) =: \sup_{t\in\mathbb{N}^+}\{\epsilon^{\varphi}_{\pi_{\boldsymbol{\theta}},t} : \varphi = V_{\pi_{\boldsymbol{\theta}'}}\}$, then the following bound holds*

$$J(\pi_{\boldsymbol{\theta}}) - J(\pi_{\boldsymbol{\theta}'}) \tag{11}$$
$$\geq \frac{1}{1-\tilde{\gamma}}\mathbb{E}_{s\sim d^{\lambda}_{\pi_{\boldsymbol{\theta}'}}(\cdot),a\sim\pi_{\boldsymbol{\theta}}(\cdot|s)}\left[A^{\mathrm{GAE}(\gamma,\lambda)}_{\pi_{\boldsymbol{\theta}'}}(s,a) - \frac{2\gamma(1-\lambda)\epsilon^{V}_{\pi_{\boldsymbol{\theta}}}(\pi_{\boldsymbol{\theta}'})}{(1-\gamma\lambda)\left|1-2\gamma\lambda|\mathcal{S}||\mathcal{A}|\right|}D_{\mathrm{TV}}(\pi_{\boldsymbol{\theta}'},\pi_{\boldsymbol{\theta}})[s]\right].$$

If $\lambda \to 0$, then the distribution $d^{\lambda}_{\pi_{\boldsymbol{\theta}'}}(\cdot)$ is reduced to $d^{\rho_0}_{\pi_{\boldsymbol{\theta}'}}(\cdot)$ and the bound (11) is reduced to

$$J(\pi_{\boldsymbol{\theta}}) - J(\pi_{\boldsymbol{\theta}'}) \geq \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\rho_0}_{\pi_{\boldsymbol{\theta}'}}(\cdot),a\sim\pi_{\boldsymbol{\theta}}(\cdot|s)}\left[A_{\pi_{\boldsymbol{\theta}'}}(s,a) - 2\gamma\epsilon^{V}_{\pi_{\boldsymbol{\theta}}}(\pi_{\boldsymbol{\theta}'})D_{\mathrm{TV}}(\pi_{\boldsymbol{\theta}'},\pi_{\boldsymbol{\theta}})[s]\right]. \tag{12}$$

Let us review (Achiam et al., 2017, Corollary 1), which shows

$$J(\pi_{\boldsymbol{\theta}}) - J(\pi_{\boldsymbol{\theta}'}) \geq \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\rho_0}_{\pi_{\boldsymbol{\theta}'}}(\cdot),a\sim\pi_{\boldsymbol{\theta}}(\cdot|s)}\left[A_{\pi_{\boldsymbol{\theta}'}}(s,a) - 2\frac{\gamma\epsilon^{V}_{\pi_{\boldsymbol{\theta}}}(\pi_{\boldsymbol{\theta}'})}{1-\gamma}D_{\mathrm{TV}}(\pi_{\boldsymbol{\theta}'},\pi_{\boldsymbol{\theta}})[s]\right]. \tag{13}$$

Comparing (12) to (13), our new bound (12) is slightly tighter than the bound shown by (Achiam et al., 2017). Concretely, our result improves the bound (13) by a factor $\frac{1}{1-\gamma}$. Since the refined bound (11) contains GAE technique that significantly reduces variance while maintains a tolerable level of bias (Schulman et al., 2016), which implies using the bound (11) as a surrogate function could improve performance potentially.

Remark 1 (Comparison with (Kakade & Langford, 2002)). *The result (11) develops the classic performance difference (6) at least three aspects. Firstly, the bound (11) extends from the advantage $A_{\pi_{\boldsymbol{\theta}'}}$ (6) to GAE function $A^{\mathrm{GAE}(\gamma,\lambda)}_{\pi_{\boldsymbol{\theta}'}}$. Secondly, the following term in Eq.(11):*

$$(1-\tilde{\gamma})^{-1}\mathbb{E}_{s\sim d^{\lambda}_{\pi_{\boldsymbol{\theta}'}}(\cdot),a\sim\pi_{\boldsymbol{\theta}}(\cdot|s)}\left[A^{\mathrm{GAE}(\gamma,\lambda)}_{\pi_{\boldsymbol{\theta}'}}(s,a)\right] \tag{14}$$

is an approximation for the difference $J(\pi_{\boldsymbol{\theta}}) - J(\pi_{\boldsymbol{\theta}'})$, while Eq.(6) shows an identity for difference $J(\pi_{\boldsymbol{\theta}}) - J(\pi_{\boldsymbol{\theta}'})$. Comparison to Eq.(6), the proposed Eq.(14) uses the state distribution $d^{\lambda}_{\pi_{\boldsymbol{\theta}'}}(\cdot)$ instead of $d^{\rho_0}_{\pi_{\boldsymbol{\theta}}}(\cdot)$, which is known the first oder expansion with respect to the policy $\pi_{\boldsymbol{\theta}}$ around the neighborhood around $\pi_{\boldsymbol{\theta}'}$(Kakade & Langford, 2002; Achiam et al., 2017). Finally, although Eq.(6) provides an identity for $J(\pi_{\boldsymbol{\theta}}) - J(\pi_{\boldsymbol{\theta}'})$, it never shows an error bound of the first oder expansion for the performance difference $J(\pi_{\boldsymbol{\theta}}) - J(\pi_{\boldsymbol{\theta}'})$, and the proposed bound (11) makes up for such a weakness. Such a bound (11) can be viewed as the worse-case approximation error, which provides a fresh surrogate function for us to design algorithms in the later section.

Let $\varphi = V^{c}_{\pi_{\boldsymbol{\theta}'}}$, Theorem 1 implies an upper bound of cost function as presented in the next Proposition 2, we will use it to make guarantee for safe policy optimization.

Proposition 2. *For any two policies $\pi_{\boldsymbol{\theta}}$ and $\pi_{\boldsymbol{\theta}'}$, let $\epsilon^{C}_{\pi_{\boldsymbol{\theta}}}(\pi_{\boldsymbol{\theta}'}) =: \sup_{t\in\mathbb{N}^+}\{\epsilon^{\varphi}_{\pi_{\boldsymbol{\theta}},t} : \varphi = V^{c}_{\pi_{\boldsymbol{\theta}'}}\}$, then*

$$J^{c}(\pi_{\boldsymbol{\theta}}) - J^{c}(\pi_{\boldsymbol{\theta}'}) \tag{15}$$
$$\leq \frac{1}{1-\tilde{\gamma}}\mathbb{E}_{s\sim d^{\lambda}_{\pi_{\boldsymbol{\theta}'}}(\cdot),a\sim\pi_{\boldsymbol{\theta}}(\cdot|s)}\left[A^{\mathrm{GAE}(\gamma,\lambda)}_{\pi_{\boldsymbol{\theta}'},C}(s,a) + \frac{2\gamma(1-\lambda)\epsilon^{C}_{\pi_{\boldsymbol{\theta}}}(\pi_{\boldsymbol{\theta}'})}{(1-\gamma\lambda)\left|1-2\gamma\lambda|\mathcal{S}||\mathcal{A}|\right|}D_{\mathrm{TV}}(\pi_{\boldsymbol{\theta}'},\pi_{\boldsymbol{\theta}})[s]\right],$$

where we calculate $A^{\mathrm{GAE}(\gamma,\lambda)}_{\pi_{\boldsymbol{\theta}'},C}(s,a)$ according to the data sampled from $\pi_{\boldsymbol{\theta}'}$ and (5).

All above bound results (11) and (15) can be extended for a total variational divergence to KL-divergence between policies, which are desirable for policy optimization. We obtain

$$\mathbb{E}_{s\sim d^{\lambda}_{\pi_{\boldsymbol{\theta}'}}(\cdot)}\left[D_{\mathrm{TV}}(\pi_{\boldsymbol{\theta}'},\pi_{\boldsymbol{\theta}})[s]\right] \leq \mathbb{E}_{s\sim d^{\lambda}_{\pi_{\boldsymbol{\theta}'}}(\cdot)}\left[\sqrt{\frac{1}{2}\mathrm{KL}(\pi_{\boldsymbol{\theta}'},\pi_{\boldsymbol{\theta}})[s]}\right] \leq \sqrt{\frac{1}{2}\mathbb{E}_{s\sim d^{\lambda}_{\pi_{\boldsymbol{\theta}'}}(\cdot)}\left[\mathrm{KL}(\pi_{\boldsymbol{\theta}'},\pi_{\boldsymbol{\theta}})[s]\right]}, \tag{16}$$

where $\mathrm{KL}(\cdot,\cdot)$ is KL-divergence, and $\mathrm{KL}(\pi_{\boldsymbol{\theta}'},\pi_{\boldsymbol{\theta}})[s] = \mathrm{KL}(\pi_{\boldsymbol{\theta}'}(\cdot|s),\pi_{\boldsymbol{\theta}}(\cdot|s))$; the first inequality follows Pinsker's inequality (Csiszár & Körner, 2011) and the second inequality follows Jensen's inequality. According to (16), we obtain the next Proposition 3.

Proposition 3. *All the bounds in (11) and (15) hold if we make the following substitution:*

$$\mathbb{E}_{s\sim d^{\lambda}_{\pi_{\boldsymbol{\theta}'}}(\cdot)}\left[D_{\mathrm{TV}}(\pi_{\boldsymbol{\theta}'},\pi_{\boldsymbol{\theta}})[s]\right] \leftarrow \sqrt{\frac{1}{2}\mathbb{E}_{s\sim d^{\lambda}_{\pi_{\boldsymbol{\theta}'}}(\cdot)}\left[\mathrm{KL}(\pi_{\boldsymbol{\theta}'},\pi_{\boldsymbol{\theta}})[s]\right]}.$$

4 METHODOLOGY: A CONSERVATIVE UPDATE POLICY (CUP)

According to the bounds in Proposition 1-3, we develop new surrogate functions to replace the objective and constraints. Inspired by two recent works (Yang et al., 2020; Zhang et al., 2020), we propose the CUP (conservative update policy) algorithm that is a two-step approach contains *performance improvement* and *projection*. Theorem 2 proves the proposed CUP guarantees the policy improvement and safe constraints.

Step 1: Performance Improvement. According to Proposition 1 and Proposition 3, for an appropriate coefficient α_k, we update policy as follows,

$$\pi_{\boldsymbol{\theta}_{k+\frac{1}{2}}} = \arg\max_{\pi_{\boldsymbol{\theta}} \in \Pi_{\boldsymbol{\theta}}} \left\{ \mathbb{E}_{s\sim d^{\lambda}_{\pi_{\boldsymbol{\theta}_k}}(\cdot), a\sim\pi_{\boldsymbol{\theta}}(\cdot|s)} \left[A^{\mathrm{GAE}(\gamma,\lambda)}_{\pi_{\boldsymbol{\theta}_k}}(s,a) \right] - \alpha_k \sqrt{\mathbb{E}_{s\sim d^{\lambda}_{\pi_{\boldsymbol{\theta}_k}}(\cdot)} \left[\mathrm{KL}(\pi_{\boldsymbol{\theta}_k}, \pi_{\boldsymbol{\theta}})[s] \right]} \right\}. \tag{17}$$

This step is a typical minimization-maximization (MM) algorithm (Hunter & Lange, 2004), it includes return maximization and minimization the distance between old policy and new policy.

Step 2: Projection. According to Proposition 2 and Proposition 3, for an appropriate coefficient β_k, we project the policy $\pi_{\boldsymbol{\theta}_{k+\frac{1}{2}}}$ onto the safe constraint set. Concretely, we use a measure $D(\cdot,\cdot)$ (e.g., KL divergence or ℓ_2-norm) to minimize distance between $\pi_{\boldsymbol{\theta}_{k+\frac{1}{2}}}$ and $\pi_{\boldsymbol{\theta}}$, and require the new policy satisfies the safe constraint:

$$\pi_{\boldsymbol{\theta}_{k+1}} = \arg\min_{\pi_{\boldsymbol{\theta}} \in \Pi_{\boldsymbol{\theta}}} D\left(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}_{k+\frac{1}{2}}}\right), \tag{18}$$

$$\text{s.t. } J^c(\pi_{\boldsymbol{\theta}_k}) + \frac{1}{1-\tilde{\gamma}} \mathbb{E}_{s\sim d^{\lambda}_{\pi_{\boldsymbol{\theta}_k}}(\cdot), a\sim\pi_{\boldsymbol{\theta}}(\cdot|s)} \left[A^{\mathrm{GAE}(\gamma,\lambda)}_{\pi_{\boldsymbol{\theta}_k}, C}(s,a) \right] + \beta_k \sqrt{\mathbb{E}_{s\sim d^{\lambda}_{\pi_{\boldsymbol{\theta}_k}}(\cdot)} \left[\mathrm{KL}(\pi_{\boldsymbol{\theta}_k}, \pi_{\boldsymbol{\theta}})[s] \right]} \leq b.$$

Until now, the particular choice of surrogate function is heuristically motivated, we show the policy and safe constraint guarantee of the proposed CUP in Theorem 2, and its proof shown in Appendix F.

Theorem 2. *Let* $\delta_k = \mathbb{E}_{s\sim d^{\lambda}_{\pi_{\boldsymbol{\theta}_k}}(\cdot)} \left[\mathrm{KL}\left(\pi_{\boldsymbol{\theta}_k}, \pi_{\boldsymbol{\theta}_{k+\frac{1}{2}}}\right)[s] \right]$*, if* $\pi_{\boldsymbol{\theta}_k}$ *and* $\pi_{\boldsymbol{\theta}_{k+1}}$ *are related to (17)-(18), then the lower bound on policy improvement, and upper bound on constraint violation are*

$$J(\pi_{\boldsymbol{\theta}_{k+1}}) - J(\pi_{\boldsymbol{\theta}_k}) \geq -\frac{\gamma(1-\lambda)\alpha_k\sqrt{2\delta_k}\epsilon^V_{\pi_{\boldsymbol{\theta}}}(\pi_{\boldsymbol{\theta}'})}{(1-\gamma)\left|1-2\gamma\lambda|\mathcal{S}||\mathcal{A}|\right|}, J^c(\pi_{\boldsymbol{\theta}_{k+1}}) \leq b + \frac{\gamma(1-\lambda)\beta_k\sqrt{2\delta_k}\epsilon^C_{\pi_{\boldsymbol{\theta}}}(\pi_{\boldsymbol{\theta}'})}{(1-\gamma)\left|1-2\gamma\lambda|\mathcal{S}||\mathcal{A}|\right|}.$$

Remark 2. *Let* $\lambda \to 0$*, according to Theorem 2, the performance and cost constraint of CUP satisfies*

$$J(\pi_{\boldsymbol{\theta}_{k+1}}) - J(\pi_{\boldsymbol{\theta}_k}) \geq -\frac{\gamma\alpha_k\sqrt{2\delta_k}\epsilon^V_{\pi_{\boldsymbol{\theta}}}(\pi_{\boldsymbol{\theta}'})}{(1-\gamma)}, \; J^c(\pi_{\boldsymbol{\theta}_{k+1}}) \leq b + \frac{\gamma\beta_k\sqrt{2\delta_k}\epsilon^C_{\pi_{\boldsymbol{\theta}}}(\pi_{\boldsymbol{\theta}'})}{(1-\gamma)}. \tag{19}$$

The bounds of CUP in (19) achieves at $\mathcal{O}(\frac{\alpha_k\gamma}{1-\gamma})$ *or* $\mathcal{O}(\frac{\beta_k\gamma}{1-\gamma})$*, which is more compact than the bounds of CPO (Achiam et al., 2017, Proposition 1-2), PCPO (Yang et al., 2020, Theorem 1) and FOCOPS (Zhang et al., 2020) where their bounds achieve at* $\mathcal{O}\left(\frac{\gamma}{(1-\gamma)^2}\right)$.

Practical Implementation Now, we present our sample-based implementation for CUP (17)-(18). Our main idea is to estimate the objective and constraints in (17)-(18) with samples collected by current policy $\pi_{\boldsymbol{\theta}_k}$, then solving its optimization problem via first-order optimizer. Due to the limitation of space, we present pseudo-code of CUP in Algorithm 1 (see Appendix B).

For each $\{(s_t, a_t, r_{t+1}, c_{t+1})\}_{t=1}^T \sim \pi_{\boldsymbol{\theta}_k}$, firstly, we update performance improvement step as:

$$\boldsymbol{\theta}_{k+\frac{1}{2}} = \arg\max_{\boldsymbol{\theta}} \left\{ \frac{1}{T}\sum_{t=1}^{N} \frac{\pi_{\boldsymbol{\theta}}(a_t|s_t)}{\pi_{\boldsymbol{\theta}_k}(a_t|s_t)} \hat{A}_t - \alpha_k \sqrt{\hat{D}_{\mathrm{KL}}(\pi_{\boldsymbol{\theta}_k}, \pi_{\boldsymbol{\theta}})} \right\}, \tag{20}$$

where $\hat{A}_t$ is a estimator of $A^{\mathrm{GAE}(\gamma,\lambda)}_{\pi_{\boldsymbol{\theta}_k}}(s,a)$, $\hat{D}_{\mathrm{KL}}(\pi_{\boldsymbol{\theta}_k}, \pi_{\boldsymbol{\theta}}) = \frac{1}{T}\sum_{t=1}^{N} \mathrm{KL}(\pi_{\boldsymbol{\theta}_k}(\cdot|s_t), \pi_{\boldsymbol{\theta}}(\cdot|s_t))$.

Then we update projection step by replacing the distance function D by KL-divergence, and we use a soft constraint instead of the hard constraint (18),

$$\boldsymbol{\theta}_{k+1} = \arg\min_{\boldsymbol{\theta}} \left\{ \frac{1}{T}\sum_{t=1}^{T} \mathrm{KL}\left(\pi_{\boldsymbol{\theta}_{k+\frac{1}{2}}}(a_t|s_t), \pi_{\boldsymbol{\theta}}(a_t|s_t)\right) + \beta_k \mathcal{L}_c \right\}, \tag{21}$$

where $\mathcal{L}_c = \hat{J}^C + \frac{1}{1-\bar{\gamma}} \cdot \frac{1}{T} \sum_{t=1}^{T} \frac{\pi_{\boldsymbol{\theta}}(a_t|s_t)}{\pi_{\boldsymbol{\theta}_k}(a_t|s_t)} \hat{A}_t^C + \alpha_k \sqrt{\frac{1}{T} \sum_{t=1}^{T} \mathrm{KL}(\pi_{\text{old}}(a_t|s_t), \pi_{\boldsymbol{\theta}}(a_t|s_t))} - b$, $\hat{J}^C$ and $\hat{A}_t^C$ are estimators for cost-return and cost-advantage correspondingly.

5 Related Work

This section reviews some typical ways to solve safe reinforcement learning: local policy search, Lagrangian approach, and constrained policy optimization (CPO). We provide more comparisons and discussion in Appendix A and Table 3.

Local Policy Search and Lagrangian Approach. A direct way to solve CMDP (4) is to apply *local policy search* (Peters & Schaal, 2008; Pirotta et al., 2013) over the policy space $\Pi_{\mathcal{C}}$, i.e.,

$$\pi_{\boldsymbol{\theta}_{k+1}} = \arg \max_{\pi_{\boldsymbol{\theta}} \in \Pi_{\boldsymbol{\theta}}} J(\pi_{\boldsymbol{\theta}}), \text{ s.t. } J^c(\pi_{\boldsymbol{\theta}}) \leq b, \text{ and } D(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}_k}) < \delta, \tag{22}$$

where δ is a positive scalar, $D(\cdot, \cdot)$ is some distance measure. For practice, the local policy search (22) is challenging to implement because it requires evaluation of the constraint function c to determine whether a proposed point π is feasible (Zhang et al., 2020). Besides, when updating policy according to samples, local policy search (22) requires off-policy evaluation (Achiam et al., 2017), which is very challenging for high-dimension control problem (Duan et al., 2016; Yang et al., 2018; 2021a). Thus, local policy search (22) looks simple, but it is impractical for high-dimension policy optimization.

The standard way to solve CMDP (4) is Lagrangian approach (Chow et al., 2017; Xu et al., 2021) that is also known as primal-dual policy optimization:

$$(\pi_\star, \lambda_\star) = \arg \min_{\lambda \geq 0} \max_{\pi_{\boldsymbol{\theta}} \in \Pi_{\boldsymbol{\theta}}} \left\{ J(\pi_{\boldsymbol{\theta}}) - \lambda (J^c(\pi_{\boldsymbol{\theta}}) - b) \right\}. \tag{23}$$

Although extensive canonical algorithms are proposed to solve problem (23), e.g., (Liang et al., 2018; Tessler et al., 2019; Paternain et al., 2019; Le et al., 2019; Russel et al., 2020; Xu et al., 2020; Satija et al., 2020; Chen et al., 2021), the policy updated by Lagrangian approach may be infeasible w.r.t. CMDP (4). This is hazardous in reinforcement learning when one needs to execute the intermediate policy (which may be unsafe) during training (Chow et al., 2018).

Constrained Policy Optimization (CPO). Recently, CPO (Achiam et al., 2017) suggests to replace the cost constraint with a surrogate cost function which evaluates the constraint $J^c(\pi_{\boldsymbol{\theta}})$ according to the samples collected from the current policy $\pi_{\boldsymbol{\theta}_k}$:

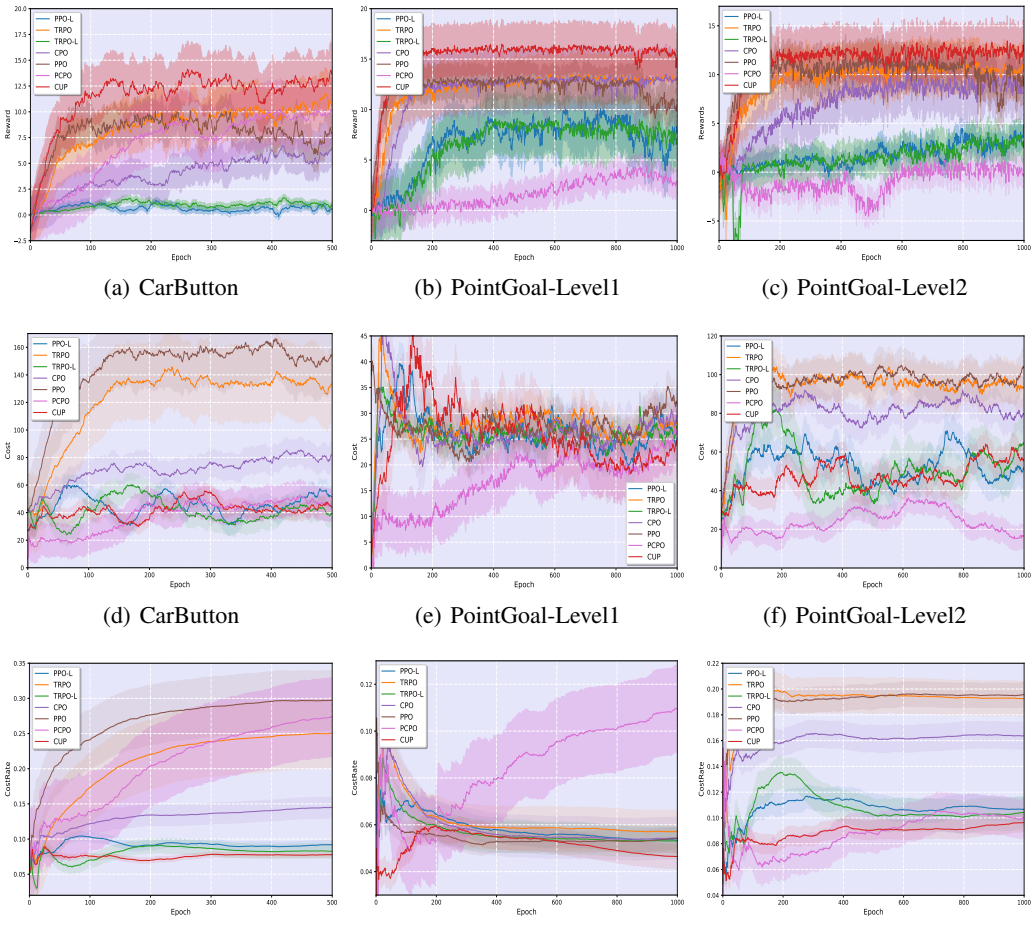
$$\pi_{\boldsymbol{\theta}_{k+1}} = \arg \max_{\pi_{\boldsymbol{\theta}} \in \Pi_{\boldsymbol{\theta}}} \quad \mathbb{E}_{s \sim d^{\rho_0}_{\pi_{\boldsymbol{\theta}_k}}(\cdot), a \sim \pi_{\boldsymbol{\theta}}(\cdot|s)} \left[A_{\pi_{\boldsymbol{\theta}_k}}(s, a) \right] \tag{24}$$

$$\text{s.t. } J^c(\pi_{\boldsymbol{\theta}_k}) + \frac{1}{1-\gamma} \mathbb{E}_{s \sim d^{\rho_0}_{\pi_{\boldsymbol{\theta}_k}}(\cdot), a \sim \pi_{\boldsymbol{\theta}}(\cdot|s)} \left[A^c_{\pi_{\boldsymbol{\theta}_k}}(s, a) \right] \leq b, \tag{25}$$

$$\bar{D}_{\mathrm{KL}}(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}_k}) = \mathbb{E}_{s \sim d^{\rho_0}_{\pi_{\boldsymbol{\theta}_k}}(\cdot)} [\mathrm{KL}(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}_k})[s]] \leq \delta. \tag{26}$$

Existing recent works (e.g., (Achiam et al., 2017; Vuong et al., 2019; Yang et al., 2020; Han et al., 2020; Bisi et al., 2020; Bharadhwaj et al., 2021)) try to find some convex approximations to replace the term $A_{\pi_{\boldsymbol{\theta}_k}}(s, a)$ and $\bar{D}_{\mathrm{KL}}(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}_k})$ Eq.(24)-(26). Such first-order and second-order approximations turn a non-convex problem (24)-(26) to be a convex problem, it seems to make a simple solution, but this approach results in many error sources and troubles in practice. Firstly, it still lacks a theory analysis to show the difference between the non-convex problem (24)-(26) and its convex approximation. Policy optimization is a typical non-convex problem (Yang et al., 2021b); its convex approximation may introduce some error for its original issue. Secondly, CPO updates parameters according to conjugate gradient (Süli & Mayers, 2003), and its solution involves the inverse Fisher information matrix, which requires expensive computation for each update. Later, Yang et al. (2020) propose projected-based constrained policy optimization (PCPO) that also uses second-order approximation, which also results in an expensive computation.

Instead of using a convex approximation for the objective function, the proposed CUP algorithm improves CPO and PCPO at least two aspects. Firstly, the CUP directly optimizes the surrogate objective function via the first-order method, and it does not depend on any convex approximation. Thus, the CUP effectively avoids the expensive computation for the inverse Fisher information matrix. Secondly, CUP extends the surrogate objective function to GAE. Although Zhang et al. (2020) has used the GAE technique in experiments, to the best of our knowledge, it still lacks a rigorous theoretical analysis involved GAE before we propose CUP.



(a) CarButton (b) PointGoal-Level1 (c) PointGoal-Level2

(d) CarButton (e) PointGoal-Level1 (f) PointGoal-Level2

(g) CarButton (h) PointGoal-Level1 (i) PointGoal-Level2

Figure 1: Learning curves for reward, cost, and cost rate on Gym ships with three pre-made robot.

6 EXPERIMENTS

In this section, we show the effectiveness of CUP on three different sets of experiments (including seven tasks): (i) robots with speed limit (Zhang et al., 2020); (ii) circle task (Achiam et al., 2017); (iii) robot options and desiderata (Ray et al., 2019).

For task (i), we train different robotic agents to move along a straight line or a two-dimensional plane, but the robot's speed is constrained for safety purposes. For task (ii), the agent is rewarded for running in a wide circle but is constrained to stay within a safe region smaller than the radius of the target circle. Task (iii) is safety Gym ships with three pre-made robots that we use in the benchmark environments from (Ray et al., 2019). All of those details are provided in Appendix G.1.

Baseline algorithms. We compare CUP to CPO (Achiam et al., 2017), PCPO (Yang et al., 2020), FOCOPS (Zhang et al., 2020) in the task (i) and task (ii). To make a more comprehensive comparison, we compare CUP with the unconstrained algorithms TRPO (Schulman et al., 2015) and PPO (Schulman et al., 2017), and compare CUP with two additional safe RL algorithms TRPO-Lagrangian and PPO-Lagrangian, that combine the Lagrangian approach with TRPO and PPO.

Robots with Speed Limit and circle task. Table 1 shows that both CUP and FOCOPS consistently enforce approximate constraint satisfaction while CUP has a higher performance than FOCOPS. CUP outperforms CPO and PCPO significantly for both reward and cost. Those observations suggest the projection step of CUP helps the agent to learn the safe constraints. We notice PCPO also has a projection step, CUP performs better than PCPO due to CUP learning the objective and constraints

Table 1: Bootstrap mean with 100 bootstrap samples of reward/cost return after training on robot with speed limit environments. Cost thresholds are in brackets under the environment names.

Environment		FOCOPS	CPO	PCPO	**CUP**
Walker2d-v3	Reward	1798.1 ± 0.3	1076.9 ± 9.8	1039.5 ± 5.2	$\mathbf{2964.3 \pm 10.8}$
(82)	Cost	82.3 ± 0.03	107.82 ± 1.16	100.25 ± 0.67	73.9 ± 0.09
Hopper-v3	Reward	1869.3 ± 2.8	1056.0 ± 5.0	1071.1 ± 4.6	$\mathbf{2409.8 \pm 5.6}$
(82)	Cost	83.1 ± 0.1	90.0 ± 8.2	74.8 ± 8.2	80.0 ± 0.1
AntCircle-v0	Reward	1206.3 ± 159.1	423.3 ± 12.6	342.9 ± 5.5	$\mathbf{1879.7 \pm 79.4}$
(50)	Cost	44.1 ± 4.2	51.3 ± 1.5	51.2 ± 2.4	49.3 ± 2.0
HumanoidCircle-v0	Reward	963.0 ± 40.0	329.5 ± 1.7	244.5 ± 7.5	$\mathbf{1029 \pm 49.0}$
(50)	Cost	50.6 ± 1.9	46.0 ± 0.4	47.1 ± 1.3	48.4 ± 2.8

Table 2: Normalized metrics from the conclusion of training averaged over various slates of environments and three random seeds per environment.

SGPoint	$\bar{J}_r$	$\bar{M}_c$	$\bar{\rho}_c$	SGCar	$\bar{J}_r$	$\bar{M}_c$	$\bar{\rho}_c$	SGDoggo	$\bar{J}_r$	$\bar{M}_c$	$\bar{\rho}_c$
PPO	1.0	1.0	1.0		1.0	1.0	1.0		1.0	1.0	1.0
PPO-L	0.552	0.553	0.638		0.299	0.241	0.237		0.0	0.028	0.288
TRPO	1.077	0.906	0.991		1.153	0.899	0.874		0.704	1.492	1.108
TRPO-L	0.726	0.672	0.628		0.302	**0.182**	0.226		0.061	0.016	0.283
CPO	0.957	0.794	0.869		0.801	0.406	0.501		0.560	1.071	0.891
PCPO	0.226	**0.321**	0.830		1.066	0.234	0.993		0.890	0.843	0.528
CUP	**1.303**	0.507	**0.452**		**1.472**	0.191	**0.206**		**1.096**	**0.007**	**0.214**

under a non-convex function while PCPO uses its convex approximation, which is one motivation for us to propose CUP.

Safety Gym Ships with Three Pre-made Robots. We compare the algorithms on three environments SGPoint, SGCar, and SGDoggo, which are all six Point/Car/Doggo robot environments with constraints in Safety Gym. Thus, it is necessary to introduce the rule for comparing the aggregate performance of algorithms across many environments by (Ray et al., 2019), where it assigns each environment $\mathcal{E}$ a set of characteristic metrics, $\bar{J}_r^{\mathcal{E}}$, $\bar{J}_c^{\mathcal{E}}$, $\bar{\rho}_c^{\mathcal{E}}$ and compute normalized return $\bar{J}_r$, normalized constraint violation $\bar{M}_c$, and normalized cost rate $\bar{\rho}_c$:

$$\bar{J}_r = \frac{J(\pi_{\boldsymbol{\theta}})}{\bar{J}_r^{\mathcal{E}}},\ \bar{M}_c = \frac{\max\{0, J_c(\pi_{\boldsymbol{\theta}}) - d\}}{\max\{10^{-6}, \bar{J}_c^{\mathcal{E}} - d\}},\ \bar{\rho}_c = \frac{\bar{\rho}_c}{\bar{\rho}_c^{\mathcal{E}}}.$$

Figure 1 shows the results from benchmarking unconstrained and constrained RL algorithms on all Point level 1 and 2 environments, and the approximately constraint-satisfying training run (CostRate curves). All the cost are shown in $[25, 30]$ In Table 2, we show the normalized metrics from the conclusion of training averaged over various slates of environments and three random seeds per environment. All experiments were run with three random seeds. Results show that cost and rewards trade off happens in SGPoint and SGCar, while CUP achieves the best performance in those two environments. In the SGDoggo environment, CUP achieves the best performance and constraint satisfaction over all the baseline algorithms.

7 Conclusion

In this paper, we propose the CUP algorithm with a theoretical safety guarantee. We derive the CUP bases on some new proposed surrogate functions w.r.t. objective and constraints and the practical implementation of CUP does not depend on any convex approximation. Extensive experiments on continuous control tasks show the effectiveness of CUP where the agent satisfies safe constraints.

REFERENCES

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *Proceedings of International Conference on Machine Learning (ICML)*, volume 70, pp. 22–31, 2017.

M Mehdi Afsar, Trafford Crump, and Behrouz Far. Reinforcement learning based recommender systems: A survey. *arXiv preprint arXiv:2101.06286*, 2021.

Eitan Altman. *Constrained Markov decision processes*. CRC Press, 1999.

Richard Bellman. A markovian decision process. *Journal of mathematics and mechanics*, 6(5): 679–684, 1957.

Homanga Bharadhwaj, Aviral Kumar, Nicholas Rhinehart, Sergey Levine, Florian Shkurti, and Animesh Garg. Conservative safety critics for exploration. In *International Conference on Learning Representations (ICLR)*, 2021.

Lorenzo Bisi, Luca Sabbioni, Edoardo Vittori, Matteo Papini, and Marcello Restelli. Risk-averse trust region optimization for reward-volatility reduction. In Christian Bessiere (ed.), *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pp. 4583–4589, 2020.

Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

Yi Chen, Jing Dong, and Zhaoran Wang. A primal-dual approach to constrained markov decision processes. *arXiv preprint arXiv:2101.10895*, 2021.

Yinlam Chow, Mohammad Ghavamzadeh, Lucas Janson, and Marco Pavone. Risk-constrained reinforcement learning with percentile risk criteria. *The Journal of Machine Learning Research*, 18(1):6070–6120, 2017.

Yinlam Chow, Ofir Nachum, Edgar Duenez-Guzman, and Mohammad Ghavamzadeh. A lyapunov-based approach to safe reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Imre Csiszár and János Körner. *Information theory: coding theorems for discrete memoryless systems*. Cambridge University Press, 2011.

Gal Dalal, Krishnamurthy Dvijotham, Matej Vecerik, Todd Hester, Cosmin Paduraru, and Yuval Tassa. Safe exploration in continuous action spaces. *arXiv preprint arXiv:1801.08757*, 2018.

Marc Peter Deisenroth, Gerhard Neumann, and Jan Peters. A survey on policy search for robotics. *Foundations and Trends® in Machine Learning*, 2013.

Yan Duan, Xi Chen, Rein Houthooft, John Schulman, and Pieter Abbeel. Benchmarking deep reinforcement learning for continuous control. In *International Conference on Machine Learning (ICML)*, pp. 1329–1338, 2016.

Evan Greensmith, Peter L Bartlett, and Jonathan Baxter. Variance reduction techniques for gradient estimates in reinforcement learning. *Journal of Machine Learning Research (JMLR)*, 5(Nov): 1471–1530, 2004.

Minghao Han, Lixian Tian, Yuanand Zhang, Jun Wang, and Wei Pan. Reinforcement learning control of constrained dynamic systems with uniformly ultimate boundedness stability guarantee. *arXiv preprint arXiv:2011.06882*, 2020.

David R Hunter and Kenneth Lange. A tutorial on mm algorithms. *The American Statistician*, 58(1): 30–37, 2004.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of International Conference on Machine Learning (ICML)*, volume 2, pp. 267–274, 2002.

Bingyi Kang, Shie Mannor, and Jiashi Feng. Learning safe policies with cost-sensitive advantage estimation, 2021. https://openreview.net/forum?id=uVnhiRaW3J.

Hoang Le, Cameron Voloshin, and Yisong Yue. Batch policy learning under constraints. In *International Conference on Machine Learning (ICML)*, pp. 3703–3712, 2019.

Qingkai Liang, Fanyu Que, and Eytan Modiano. Accelerated primal-dual policy optimization for safe reinforcement learning. *arXiv preprint arXiv:1802.06480*, 2018.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529, 2015.

OpenAI. Openai five defeats dota 2 world champions, 2019. https://openai.com/blog/openai-five-defeats-dota-2-world-champions/.

Santiago Paternain, Luiz FO Chamon, Miguel Calvo-Fullana, and Alejandro Ribeiro. Constrained reinforcement learning has zero duality gap. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Jan. Peters and Stefan. Schaal. Reinforcement learning of motor skills with policy gradients. *Neural Netw*, 21(4):682–697, 2008.

M. Pirotta, M. Restelli, A. Pecorino, and D. Calandriello. Safe policy iteration. In *International Conference on Machine Learning (ICML)*, pp. 307–315, 2013.

Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Alex Ray, Joshua Achiam, and Dario Amodei. Benchmarking Safe Exploration in Deep Reinforcement Learning. 2019.

Reazul Hasan Russel, Mouhacine Benosman, and Jeroen Van Baar. Robust constrained-mdps: Soft-constrained robust policy optimization under model uncertainty. *arXiv preprint arXiv:2010.04870*, 2020.

Harsh Satija, Philip Amortila, and Joelle Pineau. Constrained markov decision processes via backward value functions. In *International Conference on Machine Learning (ICML)*, pp. 8502–8511, 2020.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International Conference on Machine Learning (ICML)*, pp. 1889–1897, 2015.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. *International Conference on Learning Representations (ICLR)*, 2016.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484, 2016.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *Nature*, 550(7676):354, 2017.

Endre Süli and David F Mayers. *An introduction to numerical analysis*. Cambridge university press, 2003.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 1998.

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 (NeurIPS)*, pp. 1057–1063, 2000.

Chen Tessler, Daniel J Mankowitz, and Shie Mannor. Reward constrained policy optimization. *International Conference on Learning Representation (ICLR)*, 2019.

Oriol Vinyals, Igor Babuschkin, Junyoung Chung, Michael Mathieu, Max Jaderberg, Wojciech M Czarnecki, Andrew Dudzik, Aja Huang, Petko Georgiev, Richard Powell, et al. Alphastar: Mastering the real-time strategy game starcraft ii. *DeepMind blog*, 2, 2019.

Quan Vuong, Yiming Zhang, and Keith W Ross. Supervised policy update for deep reinforcement learning. In *International Conference on Learning Representation (ICLR)*, 2019.

Christopher John Cornish Hellaby Watkins. Learning from delayed rewards. 1989.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.

Cathy Wu, Aravind Rajeswaran, Yan Duan, Vikash Kumar, Alexandre M Bayen, Sham Kakade, Igor Mordatch, and Pieter Abbeel. Variance reduction for policy gradient with action-dependent factorized baselines. *International Conference on Learning Representation (ICLR)*, 2018.

Tengyu Xu, Yingbin Liang, and Guanghui Lan. A primal approach to constrained policy optimization: Global optimality and finite-time analysis. *arXiv preprint arXiv:2011.05869*, 2020.

Tengyu Xu, Yingbin Liang, and Guanghui Lan. A primal approach to constrained policy optimization: Global optimality and finite-time analysis. *International Conference on Machine Learning (ICML)*, 2021.

Long Yang, Minhao Shi, Qian Zheng, Wenjia Meng, and Gang Pan. A unified approach for multi-step temporal-difference learning with eligibility traces in reinforcement learning. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pp. 2984–2990, 2018.

Long Yang, Gang Zheng, Yu Zhang, Qian Zheng, Pengfei Li, and Gang Pan. On convergence of gradient expected sarsa (λ). In *AAAI*, 2021a.

Long Yang, Qian Zheng, and Gang Pan. Sample complexity of policy gradient finding second-order stationary points. In *AAAI*, 2021b.

Tsung-Yen Yang, Justinian Rosca, Karthik Narasimhan, and Peter J Ramadge. Projection-based constrained policy optimization. In *International Conference on Learning Representation (ICLR)*, 2020.

Moritz A Zanger, Karam Daaboul, and J Marius Zöllner. Safe continuous control with constrained model-based policy optimization. *arXiv preprint arXiv:2104.06922*, 2021.

Yiming Zhang, Quan Vuong, and Keith Ross. First order constrained optimization in policy space. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

A ADDITIONAL DISCUSSION ABOUT RELATED WORK

This section reviews three typical safe reinforcement learning algorithms: CPO (Achiam et al., 2017), PCPO (Yang et al., 2020) and FOCOPS (Zhang et al., 2020). Those algorithms also use new surrogate functions to replace the objective and constraints, which resembles the proposed CUP algorithm. The goal is to present the contribution of our work.

A.1 CPO (ACHIAM ET AL., 2017)

For a given policy $\pi_{\boldsymbol{\theta}_k}$, CPO updates new policy $\pi_{\boldsymbol{\theta}_{k+1}}$ as follows:

$$\pi_{\boldsymbol{\theta}_{k+1}} = \arg\max_{\pi_{\boldsymbol{\theta}} \in \Pi_{\boldsymbol{\theta}}} \quad \mathbb{E}_{s\sim d^{\rho_0}_{\pi_{\boldsymbol{\theta}_k}}(\cdot), a\sim\pi_{\boldsymbol{\theta}}(\cdot|s)} \left[A_{\pi_{\boldsymbol{\theta}_k}}(s,a)\right] \tag{27}$$

$$\text{s.t. } J^c(\pi_{\boldsymbol{\theta}_k}) + \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\rho_0}_{\pi_{\boldsymbol{\theta}_k}}(\cdot), a\sim\pi_{\boldsymbol{\theta}}(\cdot|s)} \left[A^c_{\pi_{\boldsymbol{\theta}_k}}(s,a)\right] \leq b, \tag{28}$$

$$\bar{D}_{\mathrm{KL}}(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}_k}) = \mathbb{E}_{s\sim d^{\rho_0}_{\pi_{\boldsymbol{\theta}_k}}(\cdot)}[\mathrm{KL}(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}_k})[s]] \leq \delta. \tag{29}$$

It is impractical to solve the problem (24) directly due to the computational cost. (Achiam et al., 2017) suggest to find some convex approximations to replace the term $A_{\pi_{\boldsymbol{\theta}_k}}(s,a)$ and $\bar{D}_{\mathrm{KL}}(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}_k})$ Eq.(24)-(26).

Concretely, according to (6), Achiam et al. (2017) suggest to use first-order Taylor expansion of $J(\pi_{\boldsymbol{\theta}})$ to replace the objective (24) as follows,

$$\frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\rho_0}_{\pi_{\boldsymbol{\theta}_k}}(\cdot), a\sim\pi_{\boldsymbol{\theta}_k}(\cdot|s)} \left[\frac{\pi_{\boldsymbol{\theta}}(a|s)}{\pi_{\boldsymbol{\theta}_k}(a|s)} A_{\pi_{\boldsymbol{\theta}_k}}(s,a)\right] = J(\pi_{\boldsymbol{\theta}}) - J(\pi_{\boldsymbol{\theta}_k}) \approx (\boldsymbol{\theta} - \boldsymbol{\theta}_k)^\top \nabla_{\boldsymbol{\theta}} J(\pi_{\boldsymbol{\theta}}).$$

Similarly, Achiam et al. (2017) use the following approximations to turn the constrained policy optimization (24)-(26) to be a convex problem,

$$\frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\rho_0}_{\pi_{\boldsymbol{\theta}_k}}(\cdot), a\sim\pi_{\boldsymbol{\theta}_k}(\cdot|s)} \left[\frac{\pi_{\boldsymbol{\theta}}(a|s)}{\pi_{\boldsymbol{\theta}_k}(a|s)} A^c_{\pi_{\boldsymbol{\theta}_k}}(s,a)\right] \approx (\boldsymbol{\theta} - \boldsymbol{\theta}_k)^\top \nabla_{\boldsymbol{\theta}} J^c(\pi_{\boldsymbol{\theta}}), \tag{30}$$

$$\bar{D}_{\mathrm{KL}}(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}_k}) \approx (\boldsymbol{\theta} - \boldsymbol{\theta}_k)^\top \mathbf{H} (\boldsymbol{\theta} - \boldsymbol{\theta}_k), \tag{31}$$

where $\mathbf{H}$ is Hessian matrix of $\bar{D}_{\mathrm{KL}}(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}_k})$, i.e.,

$$\mathbf{H}[i,j] =: \frac{\partial^2}{\partial \boldsymbol{\theta}_i \partial \boldsymbol{\theta}_j} \mathbb{E}_{s\sim d^{\rho_0}_{\pi_{\boldsymbol{\theta}_k}}(\cdot)} \left[\mathrm{KL}(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}_k})[s]\right],$$

Eq.(31) is the second-oder approximation of (26).

Let $\lambda_\star, \nu_\star$ is the dual solution of the following problem

$$\lambda_\star, \nu_\star = \arg\max_{\lambda\geq 0, \nu \geq 0} \left\{ \frac{-1}{2\lambda}\left(\mathbf{g}^\top \mathbf{H}^{-1} \mathbf{g} - 2\nu r + s v^2\right) + \nu c - \frac{\lambda\delta}{2} \right\};$$

where $\mathbf{g} = \nabla_{\boldsymbol{\theta}} \mathbb{E}_{s\sim d^{\rho_0}_{\pi_{\boldsymbol{\theta}_k}}(\cdot), a\sim\pi_{\boldsymbol{\theta}}(\cdot|s)} \left[A_{\pi_{\boldsymbol{\theta}_k}}(s,a)\right]$, $\mathbf{a} = \nabla_{\boldsymbol{\theta}} \mathbb{E}_{s\sim d^{\rho_0}_{\pi_{\boldsymbol{\theta}_k}}(\cdot), a\sim\pi_{\boldsymbol{\theta}}(\cdot|s)} \left[A^c_{\pi_{\boldsymbol{\theta}_k}}(s,a)\right]$, $r = \mathbf{g}^\top \mathbf{H} \mathbf{a}$, $s = \mathbf{a}^\top \mathbf{H}^{-1} \mathbf{a}$, and $c = J^c(\pi_{\boldsymbol{\theta}_k}) - b$.

Finally, CPO updates parameters according to conjugate gradient as follows: if approximation to CPO is feasible:

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k + \frac{1}{\lambda_\star} \mathbf{H}^{-1}(\mathbf{g} - \nu_\star \mathbf{a}),$$

else,

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \sqrt{\frac{2\delta}{\mathbf{a}^\top \mathbf{H}^{-1} \mathbf{a}}} \mathbf{H}^{-1} \mathbf{a}.$$

A.2 PCPO (YANG ET AL., 2020)

Projection-Based Constrained Policy Optimization (PCPO) is an iterative method for optimizing policies in a two-step process: the first step performs a local reward improvement update, while the second step reconciles any constraint violation by projecting the policy back onto the constraint set.

Reward Improvement:

$$\pi_{\boldsymbol{\theta}_{k+\frac{1}{2}}} = \arg\max_{\pi_{\boldsymbol{\theta}} \in \Pi_{\boldsymbol{\theta}}} \mathbb{E}_{s\sim d^{\rho_0}_{\pi_{\boldsymbol{\theta}_k}}(\cdot), a\sim\pi_{\boldsymbol{\theta}}(\cdot|s)} \left[A_{\pi_{\boldsymbol{\theta}_k}}(s,a)\right],$$
$$\text{s.t.} \bar{D}_{\text{KL}}(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}_k}) = \mathbb{E}_{s\sim d^{\rho_0}_{\pi_{\boldsymbol{\theta}_k}}(\cdot)}[\text{KL}(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}_k})[s]] \leq \delta;$$

Projection:

$$\pi_{\boldsymbol{\theta}_{k+1}} = \arg\min_{\pi_{\boldsymbol{\theta}} \in \Pi_{\boldsymbol{\theta}}} D\left(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}_{k+\frac{1}{2}}}\right),$$
$$\text{s.t. } J^c(\pi_{\boldsymbol{\theta}_k}) + \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\rho_0}_{\pi_{\boldsymbol{\theta}_k}}(\cdot), a\sim\pi_{\boldsymbol{\theta}}(\cdot|s)} \left[A^c_{\pi_{\boldsymbol{\theta}_k}}(s,a)\right] \leq b.$$

Then, Yang et al. (2020) follows CPO (Achiam et al., 2017) uses convex approximation to original problem, and calculate the update rule as follows,

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \sqrt{\frac{2\delta}{\mathbf{g}^\top \mathbf{H}^{-1}\mathbf{g}}}\mathbf{H}^{-1}\mathbf{g} - \max\left(0, \frac{\sqrt{\frac{2\delta}{\mathbf{g}^\top \mathbf{H}^{-1}\mathbf{g}}}\mathbf{a}^\top\mathbf{H}^{-1}\mathbf{g} + c}{\mathbf{a}^\top\mathbf{L}^{-1}\mathbf{a}}\right)\mathbf{L}^{-1}\mathbf{a},$$

where $\mathbf{L} = \mathbf{I}$ if D is ℓ_2-norm, and $\mathbf{L} = \mathbf{H}$ if D is KL-divergence.

A.3 FOCOPS (ZHANG ET AL., 2020)

Zhang et al. (2020) propose the First Order Constrained Optimization in Policy Space (FOCOPS) that is a two-step approach. We present it as follows.

Step1: Finding the optimal update policy. Firstly, for a given policy $\pi_{\boldsymbol{\theta}k}$, we find an optimal update policy $\pi^\star$ by solving the optimization problem (27)-(29) in the non-parameterized policy space.

$$\pi^\star = \arg\max_{\pi\in\Pi} \quad \mathbb{E}_{s\sim d^{\rho_0}_{\pi_{\boldsymbol{\theta}_k}}(\cdot), a\sim\pi(\cdot|s)} \left[A_{\pi_{\boldsymbol{\theta}_k}}(s,a)\right] \tag{32}$$
$$\text{s.t. } J^c(\pi_{\boldsymbol{\theta}_k}) + \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\rho_0}_{\pi_{\boldsymbol{\theta}_k}}(\cdot), a\sim\pi(\cdot|s)} \left[A^c_{\pi_{\boldsymbol{\theta}_k}}(s,a)\right] \leq b, \tag{33}$$
$$\bar{D}_{\text{KL}}(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}_k}) = \mathbb{E}_{s\sim d^{\rho_0}_{\pi_{\boldsymbol{\theta}_k}}(\cdot)}[\text{KL}(\pi, \pi_{\boldsymbol{\theta}_k})[s]] \leq \delta. \tag{34}$$

If $\pi_{\boldsymbol{\theta}_k}$ is feasible, then the optimal policy for (32)-(34) takes the following form:

$$\pi^\star(a|s) = \frac{\pi_{\boldsymbol{\theta}_k}(a|s)}{Z_{\lambda,\nu}(s)} \exp\left(\frac{1}{\lambda}\left(A_{\pi_{\boldsymbol{\theta}_k}}(s,a) - \nu A^c_{\pi_{\boldsymbol{\theta}_k}}(s,a)\right)\right), \tag{35}$$

where $Z_{\lambda,\nu}(s)$ is the partition function which ensures (35) is a valid probability distribution, λ and ν are solutions to the optimization problem:

$$\min_{\lambda,\nu\geq 0} \lambda\nu + \nu\tilde{b} + \lambda\mathbb{E}_{s\sim d^{\rho_0}_{\pi_{\boldsymbol{\theta}_k}}(\cdot), a\sim\pi^\star(\cdot|s)}\left[Z_{\lambda,\nu}(s)\right],$$

the term $\tilde{b} = (1-\gamma)(b - J^c(\pi_{\boldsymbol{\theta}_k}))$.

Step 2: Projection Then, we project the policy found in the previous step back into the parameterized policy space $\Pi_{\boldsymbol{\theta}}$ by solving for the closest policy $\pi_{\boldsymbol{\theta}} \in \Pi_{\boldsymbol{\theta}}$ to $\pi^\star$ in order to obtain $\pi_{\boldsymbol{\theta}_{k+1}}$:

$$\boldsymbol{\theta}_{k+1} = \arg\min_{\boldsymbol{\theta}} \mathbb{E}_{s\sim d^{\rho_0}_{\pi_{\boldsymbol{\theta}_k}}(\cdot)}[\text{KL}(\pi_{\boldsymbol{\theta}}, \pi^\star)[s]].$$

A.4 COMPARISON TO CUP

Comparing to CPO and PCPO, the implementation of CUP does not depend on any convex approximations. CPO learns its objective with the deep neural network via the first-order method (see Appendix B).

Concretely, CPO and PCPO approximate the non-convex objective (or constraints) with first-order or second Taylor expansion, but their implementations still lack a theory to show the error difference between the original objective (or constraints) and its convex approximations. Additionally, their approaches involve the inverse of a high-dimension Fisher information matrix, which causes their algorithms to require a costly computation for each update when solving high-dimensional RL problems. While the proposed CUP does not depend on any convex approximations, it learns the policy via first-order optimization approaches. Thus, CUP does not involve the inverse of a high-dimension Fisher information matrix, which implies CUP requires less memory than CPO and PCPO.

Although FOCOPS is also a non-convex implementation, it heavily depends on the current best-satisfied policy. It is known that the current best policy may not be the optimal policy, and FOCOPS requires to project this policy back into the parametric policy space, which implies FOCOPS reduce the chances for an agent to explore the environment since it may lose in a locally optimal solution. While the proposed CUP does not depend on the current optimal policy, in fact, CUP requires the agent to learn the policy according to (17), the numerical solution is not the current optimal policy, which helps CUP to explore the environment.

Table 3: Comparison of some safe reinforcement algorithms.

Algorithm	Optimization problem	Implementation	Remark
CPO (Achiam et al., 2017)	$\pi_{\boldsymbol{\theta}_{k+1}} = \arg\max_{\pi_{\boldsymbol{\theta}} \in \Pi_{\boldsymbol{\theta}}} \mathbb{E}_{s\sim d^{\rho_0}_{\pi_{\boldsymbol{\theta}_k}}(\cdot), a\sim\pi_{\boldsymbol{\theta}}(\cdot\|s)} \left[A_{\pi_{\boldsymbol{\theta}_k}}(s,a)\right]$, s.t. $J^c(\pi_{\boldsymbol{\theta}_k}) + \mathbb{E}_{s\sim d^{\rho_0}_{\pi_{\boldsymbol{\theta}_k}}(\cdot), a\sim\pi_{\boldsymbol{\theta}}(\cdot\|s)} \left[A^c_{\pi_{\boldsymbol{\theta}_k}}(s,a)\right] \leq b$, $\bar{D}_{\mathrm{KL}}(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}_k}) = \mathbb{E}_{s\sim d^{\rho_0}_{\pi_{\boldsymbol{\theta}_k}}(\cdot)}[\mathrm{KL}(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}_k})[s]] \leq \delta$.	$\boldsymbol{\theta}_{k+1} = \arg\max_{\boldsymbol{\theta}} \mathbf{g}^\top(\boldsymbol{\theta} - \boldsymbol{\theta}_k)$, s.t. $c + \mathbf{b}^\top(\boldsymbol{\theta} - \boldsymbol{\theta}_k) \leq 0$, $\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_k)^\top \mathbf{H}(\boldsymbol{\theta} - \boldsymbol{\theta}_k) \leq \delta$.	Convex Implementation
PCPO (Yang et al., 2020)	Reward Improvement: $\pi_{\boldsymbol{\theta}_{k+\frac{1}{2}}} = \arg\max_{\pi_{\boldsymbol{\theta}} \in \Pi_{\boldsymbol{\theta}}} \mathbb{E}_{s\sim d^{\rho_0}_{\pi_{\boldsymbol{\theta}_k}}(\cdot), a\sim\pi_{\boldsymbol{\theta}}(\cdot\|s)} \left[A_{\pi_{\boldsymbol{\theta}_k}}(s,a)\right]$, s.t. $\bar{D}_{\mathrm{KL}}(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}_k}) = \mathbb{E}_{s\sim d^{\rho_0}_{\pi_{\boldsymbol{\theta}_k}}(\cdot)}[\mathrm{KL}(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}_k})[s]] \leq \delta$; Projection: $\pi_{\boldsymbol{\theta}_{k+1}} = \arg\min_{\pi_{\boldsymbol{\theta}} \in \Pi_{\boldsymbol{\theta}}} D\left(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}_{k+\frac{1}{2}}}\right)$, s.t. $J^c(\pi_{\boldsymbol{\theta}_k}) + \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\rho_0}_{\pi_{\boldsymbol{\theta}_k}}(\cdot), a\sim\pi_{\boldsymbol{\theta}}(\cdot\|s)} \left[A^c_{\pi_{\boldsymbol{\theta}_k}}(s,a)\right] \leq b$.	Reward Improvement: $\boldsymbol{\theta}_{k+\frac{1}{2}} = \arg\max_{\boldsymbol{\theta}} \mathbf{g}^\top(\boldsymbol{\theta} - \boldsymbol{\theta}_k)$, s.t. $\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_k)^\top \mathbf{H}(\boldsymbol{\theta} - \boldsymbol{\theta}_k) \leq \delta$; Projection: $\pi_{\boldsymbol{\theta}_{k+1}} = \arg\min_{\boldsymbol{\theta}} \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_k)^\top \mathbf{L}(\boldsymbol{\theta} - \boldsymbol{\theta}_k)$, s.t. $c + \mathbf{b}^\top(\boldsymbol{\theta} - \boldsymbol{\theta}_k) \leq 0$.	Convex Implementation
FOCOPS (Zhang et al., 2020)	Optimal update policy: $\pi^\star = \arg\max_{\pi\in\Pi} \mathbb{E}_{s\sim d^{\rho_0}_{\pi_{\boldsymbol{\theta}_k}}(\cdot), a\sim\pi(\cdot\|s)} \left[A_{\pi_{\boldsymbol{\theta}_k}}(s,a)\right]$, s.t. $J^c(\pi_{\boldsymbol{\theta}_k}) + \mathbb{E}_{s\sim d^{\rho_0}_{\pi_{\boldsymbol{\theta}_k}}(\cdot), a\sim\pi(\cdot\|s)} \left[A^c_{\pi_{\boldsymbol{\theta}_k}}(s,a)\right] \leq b$, $\bar{D}_{\mathrm{KL}}(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}_k}) = \mathbb{E}_{s\sim d^{\rho_0}_{\pi_{\boldsymbol{\theta}_k}}(\cdot)}[\mathrm{KL}(\pi, \pi_{\boldsymbol{\theta}_k})[s]] \leq \delta$; Projection: $\pi_{\boldsymbol{\theta}_{k+1}} = \arg\min_{\pi_{\boldsymbol{\theta}} \in \Pi_{\boldsymbol{\theta}}} \mathbb{E}_{s\sim d^{\rho_0}_{\pi_{\boldsymbol{\theta}_k}}(\cdot)}[\mathrm{KL}(\pi_{\boldsymbol{\theta}}, \pi^\star)[s]]$.	Optimal update policy: $\pi^\star(a\|s) = \frac{\pi_{\boldsymbol{\theta}_k}(a\|s)}{Z_{\lambda,\nu}(s)} \exp\left(\frac{1}{\lambda}\left(A_{\pi_{\boldsymbol{\theta}_k}}(s,a) - \nu A^c_{\pi_{\boldsymbol{\theta}_k}}(s,a)\right)\right)$; Projection: $\boldsymbol{\theta}_{k+1} = \arg\min_{\boldsymbol{\theta}} \mathbb{E}_{s\sim d^{\rho_0}_{\pi_{\boldsymbol{\theta}_k}}(\cdot)}[\mathrm{KL}(\pi_{\boldsymbol{\theta}}, \pi^\star)[s]]$.	Non-Convex Implementation
CUP (Our Work)	Policy Improvement: $\pi_{\boldsymbol{\theta}_{k+\frac{1}{2}}} = \arg\max_{\pi_{\boldsymbol{\theta}} \in \Pi_{\boldsymbol{\theta}}} \left\{\mathbb{E}_{s\sim d^{\lambda}_{\pi_{\boldsymbol{\theta}_k}}(\cdot), a\sim\pi_{\boldsymbol{\theta}}(\cdot\|s)} \left[A^{\mathrm{GAE}(\gamma,\lambda)}_{\pi_{\boldsymbol{\theta}_k}}(s,a)\right] - \alpha_k \sqrt{\mathbb{E}_{s\sim d^{\lambda}_{\pi_{\boldsymbol{\theta}_k}}(\cdot)}[\mathrm{KL}(\pi_{\boldsymbol{\theta}_k}, \pi_{\boldsymbol{\theta}})[s]]}\right\}$; Projection: $\pi_{\boldsymbol{\theta}_{k+1}} = \arg\min_{\pi_{\boldsymbol{\theta}} \in \Pi_{\boldsymbol{\theta}}} D\left(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}_{k+\frac{1}{2}}}\right)$, s.t. $J^c(\pi_{\boldsymbol{\theta}_k}) + \frac{1}{1-\tilde{\gamma}}\mathbb{E}_{s\sim d^{\lambda}_{\pi_{\boldsymbol{\theta}_k}}(\cdot), a\sim\pi_{\boldsymbol{\theta}}(\cdot\|s)} \left[A^{\mathrm{GAE}(\gamma,\lambda)}_{\pi_{\boldsymbol{\theta}_k}, C}(s,a)\right] + \beta_k \sqrt{\mathbb{E}_{s\sim d^{\lambda}_{\pi_{\boldsymbol{\theta}_k}}(\cdot)}[\mathrm{KL}(\pi_{\boldsymbol{\theta}_k}, \pi_{\boldsymbol{\theta}})[s]]} \leq b$.	Policy Improvement: $\boldsymbol{\theta}_{k+\frac{1}{2}} = \arg\max_{\boldsymbol{\theta}} \left\{\frac{1}{T}\sum_{t=1}^{T} \frac{\pi_{\boldsymbol{\theta}}(a_t\|s_t)}{\pi_{\boldsymbol{\theta}_k}(a_t\|s_t)} \hat{A}_t - \alpha\sqrt{\frac{1}{T}\sum_{t=1}^{T} \mathrm{KL}(\pi_{\boldsymbol{\theta}_k}(\cdot\|s_t), \pi_{\boldsymbol{\theta}}(\cdot\|s_t))}\right\}$; Projection: $\boldsymbol{\theta}_{k+1} = \arg\min_{\boldsymbol{\theta}} \frac{1}{T}\sum_{t=1}^{T} \left\{\mathrm{KL}\left(\pi_{\boldsymbol{\theta}_{k+\frac{1}{2}}}(\cdot\|s_t), \pi_{\boldsymbol{\theta}}(\cdot\|s_t)\right) + \nu_k \frac{1-\gamma\lambda}{1-\gamma} \frac{\pi_{\boldsymbol{\theta}}(a_t\|s_t)}{\pi_{\boldsymbol{\theta}_k}(a_t\|s_t)} \hat{A}^C_t\right\}$.	Non-Convex Implementation

B CONSERVATIVE POLICY UPDATE (CPU) ALGORITHM

Algorithm 1 Conservative Policy Update (CPU)

Initialize: policy network parameters $\boldsymbol{\theta}_0$; value network parameter $\boldsymbol{\omega}_0$; cost value function parameter $\boldsymbol{\nu}_0$, step-size ν_0;
Hyper-parameters: trajectory horizon T; discount rate γ; episode number M, N, mini-batch size B, positive constant α, η;
for $k = 0, 1, 2, \ldots$ **do**
Collect batch data of M episodes of horizon T in $\cup_{i=1}^{M} \cup_{t=0}^{T} \{(s_{i,t}, a_{i,t}, r_{i,t+1}, c_{i,t+1})\}$ according to current policy $\pi_{\boldsymbol{\theta}_k}$;
Estimate c-return by discount averaging on each episode: $\hat{J}_i^C = \sum_{t=0}^{T} \gamma^t c_{i,t+1}$;
Compute TD errors $\cup_{i=1}^{M} \cup_{t=0}^{T} \{\delta_{i,t}\}$, cost TD errors $\cup_{i=1}^{M} \cup_{t=0}^{T} \{\delta_{i,t}^C\}$:

$$\delta_{i,t} = r_{i,t} + \gamma V_{\boldsymbol{\omega}_k}(s_{i,t}) - V_{\boldsymbol{\omega}_k}(s_{i,t-1}),\ \delta_{i,t}^C = c_{i,t} + \gamma V_{\boldsymbol{\nu}_k}^C(s_{i,t}) - V_{\boldsymbol{\nu}_k}^C(s_{i,t-1});$$

Compute GAE: $\cup_{i=1}^{M} \cup_{t=0}^{T} \{\hat{A}_{i,t}, \hat{A}_{i,t}^C\}$: $\hat{A}_{i,t} = \sum_{j=t}^{T} (\gamma\lambda)^{j-t} \delta_{i,j}$, $\hat{A}_{i,t}^C = \sum_{j=t}^{T} (\gamma\lambda)^{j-t} \delta_{i,j}^C$;
Compute target function for value function and cost value function as follows,

$$V_{i,t}^{\text{target}} = \hat{A}_{i,t} + V_{\boldsymbol{\omega}_k}(s_{i,t}),\ \ V_{i,t}^{\text{target},C} = \hat{A}_{i,t}^C + V_{\boldsymbol{\nu}_k}^C(s_{i,t});$$

Store data: $\mathcal{D}_k = \cup_{i=1}^{M} \cup_{t=0}^{T} \left\{(a_{i,t}, s_{i,t}, \hat{A}_{i,t}, \hat{A}_{i,t}^C, V_{i,t}^{\text{target}}, V_{i,t}^{\text{target},C})\right\}$;

$\pi_{\text{old}} \leftarrow \pi_{\boldsymbol{\theta}_k}$; Policy Improvement
for $i = 0, 1, 2, \ldots, M$ **do**

$$\boldsymbol{\theta}_{k+\frac{1}{2}} = \arg\max_{\boldsymbol{\theta}} \left\{ \frac{1}{T} \sum_{t=1}^{T} \frac{\pi_{\boldsymbol{\theta}}(a_{i,t}|s_{i,t})}{\pi_{\text{old}}(a_{i,t}|s_{i,t})} \hat{A}_{i,t} - \alpha \sqrt{\frac{1}{T} \sum_{t=1}^{T} \text{KL}(\pi_{\text{old}}(\cdot|s_{i,t}), \pi_{\boldsymbol{\theta}}(\cdot|s_{i,t}))} \right\};$$

end for
$\pi_{\text{old}} \leftarrow \pi_{\boldsymbol{\theta}_{k+\frac{1}{2}}}$; Projection
$\nu_{k+1} = (\nu_k + \eta(\hat{J}_i^C - b))_+$;
for $i = 0, 1, 2, \ldots, M$ **do**

$$\boldsymbol{\theta}_{k+1} = \arg\min_{\boldsymbol{\theta}} \frac{1}{T} \sum_{t=1}^{T} \left\{ \text{KL}(\pi_{\boldsymbol{\theta}_{\text{old}}}(\cdot|s_{i,t}), \pi_{\boldsymbol{\theta}}(\cdot|s_{i,t})) + \nu_k \frac{1-\gamma\lambda}{1-\gamma} \frac{\pi_{\boldsymbol{\theta}}(a_{i,t}|s_{i,t})}{\pi_{\boldsymbol{\theta}_k}(a_{i,t}|s_{i,t})} \hat{A}_{i,t}^C \right\};$$

end for
for each mini-batch $\{(a_j, s_j, \hat{A}_j, \hat{A}_j^C, V_j^{\text{target}}, V_j^{\text{target},C})\}$ of size B from $\mathcal{D}_k$ **do**

$$\boldsymbol{\omega}_{k+1} = \arg\min_{\boldsymbol{\omega}} \sum_{j=1}^{B} \left(V_{\boldsymbol{\omega}}(s_j) - V_j^{\text{target}}\right)^2, \boldsymbol{\nu}_{k+1} = \arg\min_{\boldsymbol{\nu}} \sum_{j=1}^{B} \left(V_{\boldsymbol{\nu}}^c(s_j) - V_j^{\text{target},C}\right)^2;$$

end for
end for

B.1 PRACTICAL IMPLEMENTATION OF CUP

In this section, we present the practical implementation of CUP.

Step 1: Policy Improvement

For the first step,

$$\pi_{\boldsymbol{\theta}_{k+\frac{1}{2}}} = \arg\max_{\pi_{\boldsymbol{\theta}} \in \Pi_{\boldsymbol{\theta}}} \left\{ \mathbb{E}_{s\sim d^{\lambda}_{\pi_{\boldsymbol{\theta}_k}}(\cdot), a\sim\pi_{\boldsymbol{\theta}}(\cdot|s)} \left[A^{\mathrm{GAE}(\gamma,\lambda)}_{\pi_{\boldsymbol{\theta}_k}}(s,a) \right] - \alpha_k \sqrt{\mathbb{E}_{s\sim d^{\lambda}_{\pi_{\boldsymbol{\theta}_k}}(\cdot)} \left[\mathrm{KL}(\pi_{\boldsymbol{\theta}_k}, \pi_{\boldsymbol{\theta}})[s] \right]} \right\},$$

according to

$$\mathbb{E}_{s\sim d^{\lambda}_{\pi_{\boldsymbol{\theta}_k}}(\cdot), a\sim\pi_{\boldsymbol{\theta}}(\cdot|s)} \left[A^{\mathrm{GAE}(\gamma,\lambda)}_{\pi_{\boldsymbol{\theta}_k}}(s,a) \right] = \mathbb{E}_{s\sim d^{\lambda}_{\pi_{\boldsymbol{\theta}_k}}(\cdot), a\sim\pi_{\boldsymbol{\theta}_k}(\cdot|s)} \left[\frac{\pi_{\boldsymbol{\theta}}(a|s)}{\pi_{\boldsymbol{\theta}_k}(a|s)} A^{\mathrm{GAE}(\gamma,\lambda)}_{\pi_{\boldsymbol{\theta}_k}}(s,a) \right], \quad (36)$$

for each data sampled from $\cup_{i=1}^{M} \cup_{t=0}^{T} \{(s_{i,t}, a_{i,t}, r_{i,t+1}, c_{i,t+1})\}$ according to current policy $\pi_{\boldsymbol{\theta}_k}$, we learn the parameter $\boldsymbol{\theta}_{k+\frac{1}{2}}$ as follows,

$$\boldsymbol{\theta}_{k+\frac{1}{2}} = \arg\max_{\boldsymbol{\theta}} \left\{ \frac{1}{T} \sum_{t=1}^{T} \frac{\pi_{\boldsymbol{\theta}}(a_{i,t}|s_{i,t})}{\pi_{\boldsymbol{\theta}_k}(a_{i,t}|s_{i,t})} \hat{A}_{i,t} - \alpha \sqrt{\frac{1}{T} \sum_{t=1}^{T} \mathrm{KL}(\pi_{\boldsymbol{\theta}_k}(\cdot|s_{i,t}), \pi_{\boldsymbol{\theta}}(\cdot|s_{i,t}))} \right\},$$

which can be solved via the first order optimizer.

Step 2: Projection

Now, consider the second step:

$$\pi_{\boldsymbol{\theta}_{k+1}} = \arg\min_{\pi_{\boldsymbol{\theta}} \in \Pi_{\boldsymbol{\theta}}} D\left(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}_{k+\frac{1}{2}}}\right),$$

$$\text{s.t. } J^c(\pi_{\boldsymbol{\theta}_k}) + \frac{1}{1-\tilde{\gamma}} \mathbb{E}_{s\sim d^{\lambda}_{\pi_{\boldsymbol{\theta}_k}}(\cdot), a\sim\pi_{\boldsymbol{\theta}}(\cdot|s)} \left[A^{\mathrm{GAE}(\gamma,\lambda)}_{\pi_{\boldsymbol{\theta}_k}, C}(s,a) \right] + \beta_k \sqrt{\mathbb{E}_{s\sim d^{\lambda}_{\pi_{\boldsymbol{\theta}_k}}(\cdot)} \left[\mathrm{KL}(\pi_{\boldsymbol{\theta}_k}, \pi_{\boldsymbol{\theta}})[s] \right]} \leq b.$$

We turn the projection step as the following unconstrained problem:

$$\min_{\boldsymbol{\theta}, \nu \geq 0} \left\{ D\left(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}_{k+\frac{1}{2}}}\right) + \nu \left(J^c(\pi_{\boldsymbol{\theta}_k}) + \frac{1}{1-\tilde{\gamma}} \mathbb{E}_{s\sim d^{\lambda}_{\pi_{\boldsymbol{\theta}_k}}(\cdot), a\sim\pi_{\boldsymbol{\theta}}(\cdot|s)} \left[A^{\mathrm{GAE}(\gamma,\lambda)}_{\pi_{\boldsymbol{\theta}_k}, C}(s,a) \right] \right. \right.$$
$$\left. \left. + \beta_k \sqrt{\mathbb{E}_{s\sim d^{\lambda}_{\pi_{\boldsymbol{\theta}_k}}(\cdot)} \left[\mathrm{KL}(\pi_{\boldsymbol{\theta}_k}, \pi_{\boldsymbol{\theta}})[s] \right]} - b \right) \right\}. \quad (37)$$

In our implementation, we use KL-divergence as the distance measure $D(\cdot, \cdot)$, then

$$D\left(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}_{k+\frac{1}{2}}}\right) = \mathbb{E}_{s\sim d^{\lambda}_{\pi_{\boldsymbol{\theta}_k}}(\cdot)} \left[\mathrm{KL}\left(\pi_{\boldsymbol{\theta}_{k+\frac{1}{2}}}, \pi_{\boldsymbol{\theta}}\right)[s] \right]. \quad (38)$$

To simplify the problem, we ignore the term $\beta_k \sqrt{\mathbb{E}_{s\sim d^{\lambda}_{\pi_{\boldsymbol{\theta}_k}}(\cdot)} \left[\mathrm{KL}(\pi_{\boldsymbol{\theta}_k}, \pi_{\boldsymbol{\theta}})[s] \right]}$ due to the following two aspects: (i) firstly, β_k is adapted to the term $\frac{\gamma(1-\lambda)\sqrt{2\delta_k}\epsilon^{C}_{\pi_{\boldsymbol{\theta}_{k+1}}}(\pi_{\boldsymbol{\theta}_k})}{(1-\gamma\lambda)|1-2\gamma\lambda|\mathcal{S}||\mathcal{A}||}$, and for the high-dimensional state space or continuous action space, then β_k is very small; (ii) secondly, if D is a KL-divergence measure, then the direction of the policy optimization $D\left(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}_{k+\frac{1}{2}}}\right)$ (38) is proportional to $\beta_k \sqrt{\mathbb{E}_{s\sim d^{\lambda}_{\pi_{\boldsymbol{\theta}_k}}(\cdot)} \left[\mathrm{KL}(\pi_{\boldsymbol{\theta}_k}, \pi_{\boldsymbol{\theta}})[s] \right]}$, thus, in practice, we can only optimize the distance $D\left(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}_{k+\frac{1}{2}}}\right)$. Above discussions implies that instead of (37), we can consider the problem

$$\min_{\boldsymbol{\theta}, \nu \geq 0} \mathcal{L}(\boldsymbol{\theta}, \nu),$$

where

$$\mathcal{L}(\boldsymbol{\theta}, \nu) = D\left(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}_{k+\frac{1}{2}}}\right) + \nu \left(J^c(\pi_{\boldsymbol{\theta}_k}) + \frac{1}{1-\tilde{\gamma}} \mathbb{E}_{s\sim d^{\lambda}_{\pi_{\boldsymbol{\theta}_k}}(\cdot), a\sim\pi_{\boldsymbol{\theta}}(\cdot|s)} \left[A^{\mathrm{GAE}(\gamma,\lambda)}_{\pi_{\boldsymbol{\theta}_k}, C}(s,a) \right] - b \right).$$

Then, according to gradient decent method, we have

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta \frac{\mathcal{L}(\boldsymbol{\theta}, \nu)}{\partial \boldsymbol{\theta}}, \quad \nu \leftarrow \nu - \eta \frac{\mathcal{L}(\boldsymbol{\theta}, \nu)}{\partial \nu}. \quad (39)$$

Particularly,

$$\frac{\mathcal{L}(\boldsymbol{\theta},\nu)}{\partial\nu} = J^c(\pi_{\boldsymbol{\theta}_k}) + \frac{1}{1-\tilde{\gamma}}\mathbb{E}_{s\sim d^{\lambda}_{\pi_{\boldsymbol{\theta}_k}}(\cdot),a\sim\pi_{\boldsymbol{\theta}}(\cdot|s)}\left[A^{\mathrm{GAE}(\gamma,\lambda)}_{\pi_{\boldsymbol{\theta}_k},C}(s,a)\right] - b, \tag{40}$$

where the term $\mathbb{E}_{s\sim d^{\lambda}_{\pi_{\boldsymbol{\theta}_k}}(\cdot),a\sim\pi_{\boldsymbol{\theta}}(\cdot|s)}\left[A^{\mathrm{GAE}(\gamma,\lambda)}_{\pi_{\boldsymbol{\theta}_k},C}(s,a)\right]$ can be estimated following the idea as (36). But recall (17) is a MM-iteration, i.e., we require to minimize $\mathbb{E}_{s\sim d^{\lambda}_{\pi_{\boldsymbol{\theta}_k}}(\cdot)}\mathrm{KL}\left(\pi_{\boldsymbol{\theta}},\pi_{\boldsymbol{\theta}_k}\right)[s]$, which implies $\pi_{\boldsymbol{\theta}}$ is close to $\pi_{\boldsymbol{\theta}_k}$. Thus it is reasonable $\mathbb{E}_{s\sim d^{\lambda}_{\pi_{\boldsymbol{\theta}_k}}(\cdot),a\sim\pi_{\boldsymbol{\theta}}(\cdot|s)}\left[A^{\mathrm{GAE}(\gamma,\lambda)}_{\pi_{\boldsymbol{\theta}_k},C}(s,a)\right] \approx 0$, thus, in practice, we update ν following a simple way

$$\nu \leftarrow (\nu - \eta(J^c(\pi_{\boldsymbol{\theta}_k}) - b))_+,$$

where $(\cdot)_+$ denote the positive part, i.e., if $x \leq 0$, $(x)_+ = 0$, else $(x)_+ = x$.

Finally, according to (39), for each data sampled from $\cup_{i=1}^{M}\cup_{t=0}^{T}\{(s_{i,t},a_{i,t},r_{i,t+1},c_{i,t+1})\}$ according to current policy $\pi_{\boldsymbol{\theta}_k}$, we learn the parameter $\boldsymbol{\theta}_{k+1}$ as follows,

$$\boldsymbol{\theta}_{k+1} = \arg\min_{\boldsymbol{\theta}} \frac{1}{T}\sum_{t=1}^{T}\left\{\mathrm{KL}\left(\pi_{\boldsymbol{\theta}_{k+\frac{1}{2}}}(\cdot|s_{i,t}),\pi_{\boldsymbol{\theta}}(\cdot|s_{i,t})\right) + \nu_k\frac{1-\gamma\lambda}{1-\gamma}\frac{\pi_{\boldsymbol{\theta}}(a_{i,t}|s_{i,t})}{\pi_{\boldsymbol{\theta}_k}(a_{i,t}|s_{i,t})}\hat{A}^{C}_{i,t}\right\},$$

which can be solved via the first-order optimizer.

C NOTATIONS

C.1 MATRIX INDEX

In this paper, we use a bold capital letter to denote matrix, e.g., $\mathbf{A} = (a_{i,j}) \in \mathbb{R}^{m \times n}$, and its (i, j)-th element denoted as

$$\mathbf{A}[i, j] =: a_{i,j},$$

where $1 \leq i \leq m, 1 \leq j \leq n$. Similarly, a bold lowercase letter denotes a vector, e.g., $\mathbf{a} = (a_1, a_2, \cdots, a_n) \in \mathbb{R}^n$, and its i-th element denoted as

$$\mathbf{a}[i] =: a_i,$$

where $1 \leq i \leq n$.

C.2 KEY NOTATIONS OF REINFORCEMENT LEARNING

For convenience of reference, we list key notations that have be used in this paper.

C.2.1 VALUE FUNCTION AND DYNAMIC SYSTEM OF MDP.

$\mathbf{r}_{\pi_{\boldsymbol{\theta}}},\ R_{\pi_{\boldsymbol{\theta}}}(s),$	:	$\mathbf{r}_{\pi_{\boldsymbol{\theta}}} \in \mathbb{R}^{\lvert\mathcal{S}\rvert}$ is the expected vector reward according to $\pi_{\boldsymbol{\theta}}$, i.e., their components are: $\mathbf{r}_{\pi_{\boldsymbol{\theta}}}[s] = \sum_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}} \pi_{\boldsymbol{\theta}}(a\lvert s) r(s'\lvert s, a) =: R_{\pi_{\boldsymbol{\theta}}}(s),\ s \in \mathcal{S}$.
$\mathbf{v}_{\pi_{\boldsymbol{\theta}}},\ V_{\pi_{\boldsymbol{\theta}}}(s),$	:	$\mathbf{v}_{\pi_{\boldsymbol{\theta}}} \in \mathbb{R}^{\lvert\mathcal{S}\rvert}$ is the vector that stores all the state value functions, and its components are: $\mathbf{v}_{\pi_{\boldsymbol{\theta}}}[s] = V_{\pi_{\boldsymbol{\theta}}}(s),\ s \in \mathcal{S}$.
$\rho(\cdot), \boldsymbol{\rho}$	:	$\rho(s)$: the initial state distribution of state s; $\boldsymbol{\rho} \in \mathbb{R}^{\lvert\mathcal{S}\rvert}$, and $\boldsymbol{\rho}[s] = \rho(s)$.
$\mathbf{P}_{\pi_{\boldsymbol{\theta}}}$	:	Single-step state transition matrix by executing $\pi_{\boldsymbol{\theta}}$.
$\mathbb{P}_{\pi_{\boldsymbol{\theta}}}(s^{'}\lvert s)$	:	Single-step state transition probability from s to $s^{'}$ by executing $\pi_{\boldsymbol{\theta}}$, and it is the $(s, s^{'})$-th component of the matrix $\mathbf{P}_{\pi_{\boldsymbol{\theta}}}$, i.e., $\mathbf{P}_{\pi_{\boldsymbol{\theta}}}[s, s^{'}] = \mathbb{P}_{\pi_{\boldsymbol{\theta}}}(s^{'}\lvert s)$.
$\mathbb{P}_{\pi_{\boldsymbol{\theta}}}(s_t = s^{'}\lvert s)$	:	The probability of visiting the state $s^{'}$ after t time steps from the state s by executing $\pi_{\boldsymbol{\theta}}$, and it is the $(s, s^{'})$-th component of the matrix $\mathbf{P}_{\pi_{\boldsymbol{\theta}}}$, i.e., $\mathbf{P}^t_{\pi_{\boldsymbol{\theta}}}[s, s^{'}] = \mathbb{P}_{\pi_{\boldsymbol{\theta}}}(s_t = s^{'}\lvert s)$.
$d^{s_0}_{\pi_{\boldsymbol{\theta}}}(s),\ d^{\rho_0}_{\pi_{\boldsymbol{\theta}}}(s)$	:	The normalized discounted distribution of the future state s encountered starting at s_0 by executing $\pi_{\boldsymbol{\theta}}$: $d^{s_0}_{\pi_{\boldsymbol{\theta}}}(s) =: (1-\gamma)\sum_{t=0}^{\infty} \gamma^t \mathbb{P}_{\pi_{\boldsymbol{\theta}}}(s_t = s\lvert s_0)$. Since $s_0 \sim \rho(\cdot)$, we define $d^{\rho_0}_{\pi_{\boldsymbol{\theta}}}(s) =: \mathbb{E}_{s_0 \sim \rho(\cdot)}[d^{s_0}_{\pi_{\boldsymbol{\theta}}}(s)]$.
$\mathbf{d}^{\rho_0}_{\pi_{\boldsymbol{\theta}}}$	:	It stores all the normalized discounted state distributions $d^{\rho_0}_{\pi_{\boldsymbol{\theta}}}(s), \in \mathcal{S}$, i.e., $\mathbf{d}^{\rho_0}_{\pi_{\boldsymbol{\theta}}} \in \mathbb{R}^{\lvert\mathcal{S}\rvert}$, and its components are: $\mathbf{d}^{\rho_0}_{\pi_{\boldsymbol{\theta}}}[s] = d^{\rho_0}_{\pi_{\boldsymbol{\theta}}}(s)$.

C.2.2 EXTEND THEM TO λ-VERSION.

$\mathbf{P}^{(\lambda)}_{\pi_{\boldsymbol{\theta}}}$	:	$\mathbf{P}^{(\lambda)}_{\pi_{\boldsymbol{\theta}}} = (1-\gamma\lambda)\sum_{t=0}^{\infty}(\gamma\lambda)^t \mathbf{P}^{t+1}_{\pi_{\boldsymbol{\theta}}}$.
$\mathbb{P}^{(\lambda)}_{\pi_{\boldsymbol{\theta}}}(s^{'}\lvert s)$	:	$\mathbb{P}^{(\lambda)}_{\pi_{\boldsymbol{\theta}}}(s^{'}\lvert s) =: \mathbf{P}^{(\lambda)}_{\pi_{\boldsymbol{\theta}}}[s, s^{'}] = (1-\gamma\lambda)\sum_{t=0}^{\infty}(\gamma\lambda)^t \mathbb{P}_{\pi_{\boldsymbol{\theta}}}(s_{t+1} = s^{'}\lvert s)$.
$\mathbf{r}^{(\lambda)}_{\pi_{\boldsymbol{\theta}}},\ R^{(\lambda)}_{\pi_{\boldsymbol{\theta}}}(s)$	:	$\mathbf{r}^{(\lambda)}_{\pi_{\boldsymbol{\theta}}} = \sum_{t=0}^{\infty}(\gamma\lambda\mathbf{P}_{\pi_{\boldsymbol{\theta}}})^t \mathbf{r}_{\pi_{\boldsymbol{\theta}}};\ R^{(\lambda)}_{\pi_{\boldsymbol{\theta}}}(s) =: \mathbf{r}^{(\lambda)}_{\pi_{\boldsymbol{\theta}}}[s]$.
$\tilde{\gamma}$	:	$\tilde{\gamma} = \dfrac{\gamma(1-\lambda)}{1-\gamma\lambda}$.
$d^{s_0,\lambda}_{\pi_{\boldsymbol{\theta}}}(s)$	:	$d^{s_0,\lambda}_{\pi_{\boldsymbol{\theta}}}(s) = (1-\tilde{\gamma})\sum_{t=0}^{\infty}\tilde{\gamma}^t \mathbb{P}^{(\lambda)}_{\pi_{\boldsymbol{\theta}}}(s_t = s\lvert s_0)$.
$d^{\lambda}_{\pi_{\boldsymbol{\theta}}}(s),\ \mathbf{d}^{\lambda}_{\pi_{\boldsymbol{\theta}}}$	:	$d^{\lambda}_{\pi_{\boldsymbol{\theta}}}(s) = \mathbb{E}_{s_0 \sim \rho_0(\cdot)}\left[d^{s_0,\lambda}_{\pi_{\boldsymbol{\theta}}}(s)\right],\ \mathbf{d}^{\lambda}_{\pi_{\boldsymbol{\theta}}}[s] = d^{\lambda}_{\pi_{\boldsymbol{\theta}}}(s)$.

C.2.3 TD ERROR W.R.T. ANY FUNCTION $\varphi(\cdot)$.

δ^{φ}_t	:	$\delta^{\varphi}_t = r(s_{t+1}\lvert s_t, a_t) + \gamma\varphi(s_{t+1}) - \varphi(s_t)$.
$\delta^{\varphi}_{\pi_{\boldsymbol{\theta}},t}(s)$	:	$\delta^{\varphi}_{\pi_{\boldsymbol{\theta}},t}(s) = \mathbb{E}_{s_t \sim \mathbb{P}_{\pi_{\boldsymbol{\theta}}}(\cdot\lvert s), a_t \sim \pi_{\boldsymbol{\theta}}(\cdot\lvert s_t), s_{t+1} \sim \mathbb{P}(\cdot\lvert s_t, a_t)}\left[\delta^{\varphi}_t\right]$.
$\boldsymbol{\delta}^{\varphi}_{\pi_{\boldsymbol{\theta}},t}$	:	$\boldsymbol{\delta}^{\varphi}_{\pi_{\boldsymbol{\theta}},t}[s] = \delta^{\varphi}_{\pi_{\boldsymbol{\theta}},t}(s)$.
$\Delta^{\varphi}_t(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}'}, s)$	:	$\mathbb{E}_{s_t \sim \mathbb{P}_{\pi_{\boldsymbol{\theta}'}}(\cdot\lvert s), a_t \sim \pi_{\boldsymbol{\theta}'}(\cdot\lvert s_t), s_{t+1} \sim \mathbb{P}(\cdot\lvert s_t, a_t)}\left[\left(\dfrac{\pi_{\boldsymbol{\theta}}(a_t\lvert s_t)}{\pi_{\boldsymbol{\theta}'}(a_t\lvert s_t)} - 1\right)\delta^{\varphi}_t\right]$.
$\boldsymbol{\Delta}^{\varphi}_t(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}'})$	:	$\boldsymbol{\Delta}^{\varphi}_t(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}'})[s] = \Delta^{\varphi}_t(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}'}, s)$.

D PRELIMINARIES

In this section, we introduce some new notations about state distribution, policy optimization and λ-returns.

D.1 STATE DISTRIBUTION

We use $\mathbf{P}_{\pi_{\boldsymbol{\theta}}} \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$ to denote the state transition matrix by executing $\pi_{\boldsymbol{\theta}}$, and their components are:

$$\mathbf{P}_{\pi_{\boldsymbol{\theta}}}[s, s'] = \sum_{a\in\mathcal{A}} \pi_{\boldsymbol{\theta}}(a|s)\mathbb{P}(s'|s, a) =: \mathbb{P}_{\pi_{\boldsymbol{\theta}}}(s^{'}|s), \;\; s, s^{'} \in \mathcal{S},$$

which denotes one-step state transformation probability from s to $s^{'}$.

We use $\mathbb{P}_{\pi_{\boldsymbol{\theta}}}(s_t = s|s_0)$ to denote the probability of visiting s after t time steps from the initial state s_0 by executing $\pi_{\boldsymbol{\theta}}$. Particularly, we notice if $t = 0$, $s_t \neq s_0$, then $\mathbb{P}_{\pi_{\boldsymbol{\theta}}}(s_t = s|s_0) = 0$, i.e.,

$$\mathbb{P}_{\pi_{\boldsymbol{\theta}}}(s_t = s|s_0) = 0, \;\; t = 0 \text{ and } s \neq s_0. \tag{41}$$

Then for any initial state $s_0 \sim \rho(\cdot)$, the following holds,

$$\mathbb{P}_{\pi_{\boldsymbol{\theta}}}(s_t = s|s_0) = \sum_{s^{'}\in\mathcal{S}} \mathbb{P}_{\pi_{\boldsymbol{\theta}}}(s_t = s|s_{t-1} = s^{'})\mathbb{P}_{\pi_{\boldsymbol{\theta}}}(s_{t-1} = s^{'}|s_0). \tag{42}$$

Recall $d^{s_0}_{\pi_{\boldsymbol{\theta}}}(s)$ denotes the normalized discounted distribution of the future state s encountered starting at s_0 by executing $\pi_{\boldsymbol{\theta}}$,

$$d^{s_0}_{\pi_{\boldsymbol{\theta}}}(s) = (1-\gamma)\sum_{t=0}^{\infty} \gamma^t \mathbb{P}_{\pi_{\boldsymbol{\theta}}}(s_t = s|s_0).$$

Furthermore, since $s_0 \sim \rho_0(\cdot)$, we define

$$d^{\rho_0}_{\pi_{\boldsymbol{\theta}}}(s) = \mathbb{E}_{s_0\sim\rho_0(\cdot)}[d^{s_0}_{\pi_{\boldsymbol{\theta}}}(s)] = \int_{s_0\in\mathcal{S}} \rho_0(s_0) d^{s_0}_{\pi_{\boldsymbol{\theta}}}(s)\mathrm{d}s_0$$

as the discounted state visitation distribution over the initial distribution $\rho_0(\cdot)$. We use $\mathbf{d}^{\rho_0}_{\pi_{\boldsymbol{\theta}}} \in \mathbb{R}^{|\mathcal{S}|}$ to store all the normalized discounted state distributions, and its components are:

$$\mathbf{d}^{\rho_0}_{\pi_{\boldsymbol{\theta}}}[s] = d^{\rho_0}_{\pi_{\boldsymbol{\theta}}}(s), \;\; s \in \mathcal{S}.$$

We use $\boldsymbol{\rho}_0 \in \mathbb{R}^{|\mathcal{S}|}$ to denote initial state distribution vector, and their components are:

$$\boldsymbol{\rho}_0[s] = \rho_0(s), \;\; s \in \mathcal{S}.$$

Then, we rewrite $\mathbf{d}^{\rho_0}_{\pi_{\boldsymbol{\theta}}}$ as the following matrix version,

$$\mathbf{d}^{\rho_0}_{\pi_{\boldsymbol{\theta}}} = (1-\gamma)\sum_{t=0}^{\infty} (\gamma \mathbf{P}_{\pi_{\boldsymbol{\theta}}})^t \boldsymbol{\rho}_0 = (1-\gamma)(\mathbf{I} - \gamma\mathbf{P}_{\pi_{\boldsymbol{\theta}}})^{-1}\boldsymbol{\rho}_0. \tag{43}$$

D.2 OBJECTIVE OF MDP

Recall $\tau = \{s_t, a_t, r_{t+1}\}_{t\geq 0} \sim \pi_{\boldsymbol{\theta}}$, according to τ, we define the expected return $J(\pi_{\boldsymbol{\theta}}|s_0)$ as follows,

$$J(\pi_{\boldsymbol{\theta}}|s_0) = \mathbb{E}_{\tau\sim\pi_{\boldsymbol{\theta}}}[R(\tau)] = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{s_0}_{\pi_{\boldsymbol{\theta}}}(\cdot), a\sim\pi_{\boldsymbol{\theta}}(\cdot|s), s^{'}\sim\mathbb{P}(\cdot|s,a)}\left[r(s^{'}|s, a)\right], \tag{44}$$

where $R(\tau) = \sum_{t\geq 0}\gamma^t r_{t+1}$, and the notation $J(\pi_{\boldsymbol{\theta}}|s_0)$ is "conditional" on s_0 is to emphasize the trajectory τ starting from s_0.

Since $s_0 \sim \rho_0(\cdot)$, we define the objective of MDP as follows,

$$J(\pi_{\boldsymbol{\theta}}) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\rho_0}_{\pi_{\boldsymbol{\theta}}}(\cdot), a\sim\pi_{\boldsymbol{\theta}}(\cdot|s), s^{'}\sim\mathbb{P}(\cdot|s,a)}\left[r(s^{'}|s, a)\right]. \tag{45}$$

The goal of reinforcement learning is to solve the following optimization problem:

$$\boldsymbol{\theta}_\star = \arg\max_{\boldsymbol{\theta}\in\mathbb{R}^p} J(\pi_{\boldsymbol{\theta}}). \tag{46}$$

D.3 BELLMAN OPERATORN

Let $\mathcal{B}_{\pi_{\boldsymbol{\theta}}}$ be the *Bellman operator*:

$$\mathcal{B}_{\pi_{\boldsymbol{\theta}}} : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}, \quad v \mapsto \mathbf{r}_{\pi_{\boldsymbol{\theta}}} + \gamma \mathbf{P}_{\pi_{\boldsymbol{\theta}}} v, \tag{47}$$

where $\mathbf{r}_{\pi_{\boldsymbol{\theta}}} \in \mathbb{R}^{|\mathcal{S}|}$ is the expected reward according to $\pi_{\boldsymbol{\theta}}$, i.e., their components are:

$$\mathbf{r}_{\pi_{\boldsymbol{\theta}}}[s] = \sum_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}} \pi_{\boldsymbol{\theta}}(a|s) r(s'|s,a) =: R_{\pi_{\boldsymbol{\theta}}}(s), \;\; s \in \mathcal{S}.$$

Let $\mathbf{v}_{\pi_{\boldsymbol{\theta}}} \in \mathbb{R}^{|\mathcal{S}|}$ be a vector that stores all the state value functions, and its components are:

$$\mathbf{v}_{\pi_{\boldsymbol{\theta}}}[s] = V_{\pi_{\boldsymbol{\theta}}}(s), \;\; s \in \mathcal{S}.$$

Then, according to Bellman operator (47), we rewrite Bellman equation (Bellman, 1957) as the following matrix version:

$$\mathcal{B}_{\pi_{\boldsymbol{\theta}}} \mathbf{v}_{\pi_{\boldsymbol{\theta}}} = \mathbf{v}_{\pi_{\boldsymbol{\theta}}}. \tag{48}$$

Furthermore, we define λ-*Bellman operator* $\mathcal{B}^{\lambda}_{\pi_{\boldsymbol{\theta}}}$ as follows,

$$\mathcal{B}^{\lambda}_{\pi_{\boldsymbol{\theta}}} = (1-\lambda) \sum_{t=0}^{\infty} \lambda^t (\mathcal{B}_{\pi_{\boldsymbol{\theta}}})^{t+1},$$

which implies

$$\mathcal{B}^{\lambda}_{\pi_{\boldsymbol{\theta}}} : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}, \quad v \mapsto \mathbf{r}^{(\lambda)}_{\pi_{\boldsymbol{\theta}}} + \tilde{\gamma} \mathbf{P}^{(\lambda)}_{\pi_{\boldsymbol{\theta}}} v, \tag{49}$$

where

$$\mathbf{P}^{(\lambda)}_{\pi_{\boldsymbol{\theta}}} = (1-\gamma\lambda) \sum_{t=0}^{\infty} (\gamma\lambda)^t \mathbf{P}^{t+1}_{\pi_{\boldsymbol{\theta}}}, \;\; \mathbf{r}^{(\lambda)}_{\pi_{\boldsymbol{\theta}}} = \sum_{t=0}^{\infty} (\gamma\lambda \mathbf{P}_{\pi_{\boldsymbol{\theta}}})^t \mathbf{r}_{\pi_{\boldsymbol{\theta}}}, \;\; \tilde{\gamma} = \frac{\gamma(1-\lambda)}{1-\gamma\lambda}. \tag{50}$$

Let

$$\mathbb{P}^{(\lambda)}_{\pi_{\boldsymbol{\theta}}}(s^{'}|s) = \mathbf{P}^{(\lambda)}_{\pi_{\boldsymbol{\theta}}}[s, s^{'}] =: (1-\gamma\lambda) \sum_{t=0}^{\infty} (\gamma\lambda)^t \left(\mathbf{P}^{t+1}_{\pi_{\boldsymbol{\theta}}}[s, s^{'}] \right), \tag{51}$$

where $\mathbf{P}^{t+1}_{\pi_{\boldsymbol{\theta}}}[s, s^{'}]$ is the $(s, s^{'})$-th component of matrix $\mathbf{P}^{t+1}_{\pi_{\boldsymbol{\theta}}}$, which is the probability of visiting $s^{'}$ after $t+1$ time steps from the state s by executing $\pi_{\boldsymbol{\theta}}$, i.e.,

$$\mathbf{P}^{t+1}_{\pi_{\boldsymbol{\theta}}}[s, s^{'}] = \mathbb{P}_{\pi_{\boldsymbol{\theta}}}(s_{t+1} = s^{'}|s). \tag{52}$$

Thus, we rewrite $\mathbb{P}^{(\lambda)}_{\pi_{\boldsymbol{\theta}}}(s^{'}|s)$ (51) as follows

$$\mathbb{P}^{(\lambda)}_{\pi_{\boldsymbol{\theta}}}(s^{'}|s) = (1-\gamma\lambda) \sum_{t=0}^{\infty} (\gamma\lambda)^t \mathbb{P}_{\pi_{\boldsymbol{\theta}}}(s_{t+1} = s^{'}|s), \;\; s \in \mathcal{S}. \tag{53}$$

D.4 λ-RETURN

Furthermore, recall the following visitation sequence $\tau = \{s_t, a_t, r_{t+1}\}_{t \geq 0}$ induced by $\pi_{\boldsymbol{\theta}}$, it is similar to the probability $\mathbb{P}_{\pi_{\boldsymbol{\theta}}}(s_t = s^{'}|s_0)$, we introduce $\mathbb{P}^{(\lambda)}_{\pi_{\boldsymbol{\theta}}}(s_t = s^{'}|s_0)$ as the probability of transition from state s to state $s^{'}$ after t time steps under the dynamic transformation matrix $\mathbf{P}^{(\lambda)}_{\pi_{\boldsymbol{\theta}}}$. Then, the following equity holds

$$\mathbb{P}^{(\lambda)}_{\pi_{\boldsymbol{\theta}}}(s_t = s|s_0) = \sum_{s' \in \mathcal{S}} \mathbb{P}^{(\lambda)}_{\pi_{\boldsymbol{\theta}}}(s_t = s|s_{t-1} = s^{'}) \mathbb{P}^{(\lambda)}_{\pi_{\boldsymbol{\theta}}}(s_{t-1} = s^{'}|s_0). \tag{54}$$

Similarly, let

$$R_{\pi_{\boldsymbol{\theta}}}^{(\lambda)}(s) =: \mathbf{r}_{\pi_{\boldsymbol{\theta}}}^{(\lambda)}[s] = \sum_{t=0}^{\infty}(\gamma\lambda\mathbf{P}_{\pi_{\boldsymbol{\theta}}})^t\mathbf{r}_{\pi_{\boldsymbol{\theta}}}[s] = \sum_{t=0}^{\infty}(\gamma\lambda)^t\left(\sum_{s^{'}\in\mathcal{S}}\mathbb{P}_{\pi_{\boldsymbol{\theta}}}(s_t = s^{'}|s)R_{\pi_{\boldsymbol{\theta}}}(s^{'})\right)$$
$$= \sum_{t=0}^{\infty}\sum_{s^{'}\in\mathcal{S}}(\gamma\lambda)^t\mathbb{P}_{\pi_{\boldsymbol{\theta}}}(s_t = s^{'}|s)R_{\pi_{\boldsymbol{\theta}}}(s^{'}). \tag{55}$$

It is similar to normalized discounted distribution $d_{\pi_{\boldsymbol{\theta}}}^{\rho_0}(s)$, we introduce λ-return version of discounted state distribution $d_{\pi_{\boldsymbol{\theta}}}^{\lambda}(s)$ as follows: $\forall s \in \mathcal{S}$,

$$d_{\pi_{\boldsymbol{\theta}}}^{s_0,\lambda}(s) = (1-\tilde{\gamma})\sum_{t=0}^{\infty}\tilde{\gamma}^t\mathbb{P}_{\pi_{\boldsymbol{\theta}}}^{(\lambda)}(s_t = s|s_0), \tag{56}$$
$$d_{\pi_{\boldsymbol{\theta}}}^{\lambda}(s) = \mathbb{E}_{s_0\sim\rho_0(\cdot)}\left[d_{\pi_{\boldsymbol{\theta}}}^{s_0,\lambda}(s)\right], \tag{57}$$
$$\mathbf{d}_{\pi_{\boldsymbol{\theta}}}^{\lambda}[s] = d_{\pi_{\boldsymbol{\theta}}}^{\lambda}(s), \tag{58}$$

where $\mathbb{P}_{\pi_{\boldsymbol{\theta}}}^{(\lambda)}(s_t = s|s_0)$ is the (s_0, s)-th component of the matrix $\left(\mathbf{P}_{\pi_{\boldsymbol{\theta}}}^{(\lambda)}\right)^t$, i.e.,

$$\mathbb{P}_{\pi_{\boldsymbol{\theta}}}^{(\lambda)}(s_t = s|s_0) =: \left(\mathbf{P}_{\pi_{\boldsymbol{\theta}}}^{(\lambda)}\right)^t[s_0, s].$$

Similarly, $\mathbb{P}_{\pi_{\boldsymbol{\theta}}}^{(\lambda)}(s_t = s^{'}|s)$ is the $(s, s^{'})$-th component of the matrix $\left(\mathbf{P}_{\pi_{\boldsymbol{\theta}}}^{(\lambda)}\right)^t$, i.e.,

$$\mathbb{P}_{\pi_{\boldsymbol{\theta}}}^{(\lambda)}(s_t = s^{'}|s) =: \left(\mathbf{P}_{\pi_{\boldsymbol{\theta}}}^{(\lambda)}\right)^t[s, s^{'}].$$

Finally, we rewrite $\mathbf{d}_{\pi_{\boldsymbol{\theta}}}^{\rho_0,\lambda}$ as the following matrix version,

$$\mathbf{d}_{\pi_{\boldsymbol{\theta}}}^{\lambda} = (1-\tilde{\gamma})\sum_{t=0}^{\infty}\left(\gamma\mathbf{P}_{\pi_{\boldsymbol{\theta}}}^{(\lambda)}\right)^t\boldsymbol{\rho}_0 = (1-\tilde{\gamma})\left(\mathbf{I}-\tilde{\gamma}\mathbf{P}_{\pi_{\boldsymbol{\theta}}}^{(\lambda)}\right)^{-1}\boldsymbol{\rho}_0. \tag{59}$$

Remark 3 (λ-Return Version of Bellman Equation)**.** *According to Bellman equation (48), $\mathbf{v}_{\pi_{\boldsymbol{\theta}}}$ is fixed point of λ-operator $\mathcal{B}_{\pi_{\boldsymbol{\theta}}}^{\lambda}$, i.e.,*

$$\mathbf{v}_{\pi_{\boldsymbol{\theta}}} = \mathbf{r}_{\pi_{\boldsymbol{\theta}}}^{(\lambda)} + \tilde{\gamma}\mathbf{P}_{\pi_{\boldsymbol{\theta}}}^{(\lambda)}\mathbf{v}_{\pi_{\boldsymbol{\theta}}}. \tag{60}$$

Recall $\tau = \{s_t, a_t, r_{t+1}\}_{t\geq 0} \sim \pi_{\boldsymbol{\theta}}$, according to (60), the value function of initial state s_0 is

$$V_{\pi_{\boldsymbol{\theta}}}(s_0) = \mathbf{v}_{\pi_{\boldsymbol{\theta}}}[s_0] = \mathbf{r}_{\pi_{\boldsymbol{\theta}}}^{(\lambda)}[s_0] + \tilde{\gamma}\mathbf{P}_{\pi_{\boldsymbol{\theta}}}^{(\lambda)}\mathbf{v}_{\pi_{\boldsymbol{\theta}}}[s_0]$$
$$= R_{\pi_{\boldsymbol{\theta}}}^{(\lambda)}(s_0) + \tilde{\gamma}\sum_{s^{'}\in\mathcal{S}}\mathbb{P}_{\pi_{\boldsymbol{\theta}}}^{(\lambda)}(s_1 = s^{'}|s_0)V_{\pi_{\boldsymbol{\theta}}}(s^{'}). \tag{61}$$

We unroll the expression of (61) repeatedly, then we have

$$
\begin{aligned}
&V_{\pi_{\boldsymbol{\theta}}}(s_0) \\
=&R^{(\lambda)}_{\pi_{\boldsymbol{\theta}}}(s_0) + \tilde{\gamma} \sum_{s^{'} \in \mathcal{S}} \mathbb{P}^{(\lambda)}_{\pi_{\boldsymbol{\theta}}}(s_1 = s^{'}|s_0) \underbrace{\left(R^{(\lambda)}_{\pi_{\boldsymbol{\theta}}}(s^{'}) + \tilde{\gamma} \sum_{s^{''} \in \mathcal{S}} \mathbb{P}^{(\lambda)}_{\pi_{\boldsymbol{\theta}}}(s_2 = s^{''}|s_1 = s^{'}) V_{\pi_{\boldsymbol{\theta}}}(s^{''}) \right)}_{= V_{\pi_{\boldsymbol{\theta}}}(s^{'})} \\
=&R^{(\lambda)}_{\pi_{\boldsymbol{\theta}}}(s_0) + \tilde{\gamma} \sum_{s^{'} \in \mathcal{S}} \mathbb{P}^{(\lambda)}_{\pi_{\boldsymbol{\theta}}}(s_1 = s^{'}|s_0) R^{(\lambda)}_{\pi_{\boldsymbol{\theta}}}(s^{'}) \\
&\quad + \tilde{\gamma}^2 \sum_{s^{''} \in \mathcal{S}} \underbrace{\left(\sum_{s^{'} \in \mathcal{S}} \mathbb{P}^{(\lambda)}_{\pi_{\boldsymbol{\theta}}}(s_1 = s^{'}|s_0) \mathbb{P}^{(\lambda)}_{\pi_{\boldsymbol{\theta}}}(s_2 = s^{''}|s_1 = s^{'}) \right)}_{\overset{(54)}{=} : \mathbb{P}^{(\lambda)}_{\pi_{\boldsymbol{\theta}}}(s_2 = s^{''}|s_0)} V_{\pi_{\boldsymbol{\theta}}}(s^{''}) \\
=&R^{(\lambda)}_{\pi_{\boldsymbol{\theta}}}(s_0) + \tilde{\gamma} \sum_{s \in \mathcal{S}} \mathbb{P}^{(\lambda)}_{\pi_{\boldsymbol{\theta}}}(s_1 = s|s_0) R^{(\lambda)}_{\pi_{\boldsymbol{\theta}}}(s) + \tilde{\gamma}^2 \sum_{s \in \mathcal{S}} \mathbb{P}^{(\lambda)}_{\pi_{\boldsymbol{\theta}}}(s_2 = s|s_0) V_{\pi_{\boldsymbol{\theta}}}(s) \\
=&R^{(\lambda)}_{\pi_{\boldsymbol{\theta}}}(s_0) + \tilde{\gamma} \sum_{s \in \mathcal{S}} \mathbb{P}^{(\lambda)}_{\pi_{\boldsymbol{\theta}}}(s_1 = s|s_0) R^{(\lambda)}_{\pi_{\boldsymbol{\theta}}}(s) \\
&\quad + \tilde{\gamma}^2 \sum_{s \in \mathcal{S}} \mathbb{P}^{(\lambda)}_{\pi_{\boldsymbol{\theta}}}(s_2 = s|s_0) \left(R^{(\lambda)}_{\pi_{\boldsymbol{\theta}}}(s) + \tilde{\gamma} \sum_{s^{'} \in \mathcal{S}} \mathbb{P}^{(\lambda)}_{\pi_{\boldsymbol{\theta}}}(s_3 = s^{'}|s_2 = s) V_{\pi_{\boldsymbol{\theta}}}(s^{'}) \right) \\
=&R^{(\lambda)}_{\pi_{\boldsymbol{\theta}}}(s_0) + \tilde{\gamma} \sum_{s \in \mathcal{S}} \mathbb{P}^{(\lambda)}_{\pi_{\boldsymbol{\theta}}}(s_1 = s|s_0) R^{(\lambda)}_{\pi_{\boldsymbol{\theta}}}(s) + \tilde{\gamma}^2 \sum_{s \in \mathcal{S}} \mathbb{P}^{(\lambda)}_{\pi_{\boldsymbol{\theta}}}(s_2 = s|s_0) R^{(\lambda)}_{\pi_{\boldsymbol{\theta}}}(s) \\
&\quad + \tilde{\gamma}^3 \sum_{s^{'} \in \mathcal{S}} \underbrace{\left(\sum_{s \in \mathcal{S}} \mathbb{P}^{(\lambda)}_{\pi_{\boldsymbol{\theta}}}(s_2 = s|s_0) \mathbb{P}^{(\lambda)}_{\pi_{\boldsymbol{\theta}}}(s_3 = s^{'}|s_2 = s) \right)}_{= \mathbb{P}^{(\lambda)}_{\pi_{\boldsymbol{\theta}}}(s_3 = s^{'}|s_0)} V_{\pi_{\boldsymbol{\theta}}}(s^{'}) \\
=&R^{(\lambda)}(s_0) + \tilde{\gamma} \sum_{s \in \mathcal{S}} \mathbb{P}^{(\lambda)}_{\pi_{\boldsymbol{\theta}}}(s_1 = s|s_0) R^{(\lambda)}_{\pi_{\boldsymbol{\theta}}}(s) + \tilde{\gamma}^2 \sum_{s \in \mathcal{S}} \mathbb{P}^{(\lambda)}_{\pi_{\boldsymbol{\theta}}}(s_2 = s|s_0) R^{(\lambda)}_{\pi_{\boldsymbol{\theta}}}(s) \\
&\quad + \tilde{\gamma}^3 \sum_{s \in \mathcal{S}} \mathbb{P}^{(\lambda)}_{\pi_{\boldsymbol{\theta}}}(s_3 = s|s_0) V_{\pi_{\boldsymbol{\theta}}}(s) \\
=&\cdots \\
=&\sum_{s \in \mathcal{S}} \sum_{t=0}^{\infty} \tilde{\gamma}^t \mathbb{P}^{(\lambda)}_{\pi_{\boldsymbol{\theta}}}(s_t = s|s_0) R^{(\lambda)}_{\pi_{\boldsymbol{\theta}}}(s) \overset{(56)}{=} \frac{1}{1-\tilde{\gamma}} \sum_{s \in \mathcal{S}} d^{s_0,\lambda}_{\pi_{\boldsymbol{\theta}}}(s) R^{(\lambda)}_{\pi_{\boldsymbol{\theta}}}(s).
\end{aligned} \tag{62}
$$

According to (44) and (62), we have

$$
\begin{aligned}
J(\pi_{\boldsymbol{\theta}}) =& \sum_{s_0 \in \mathcal{S}} \rho_0(s_0) V_{\pi_{\boldsymbol{\theta}}}(s_0) \overset{(62)}{=} \frac{1}{1-\tilde{\gamma}} \sum_{s_0 \in \mathcal{S}} \rho_0(s_0) \sum_{s \in \mathcal{S}} d^{s_0,\lambda}_{\pi_{\boldsymbol{\theta}}}(s) R^{(\lambda)}_{\pi_{\boldsymbol{\theta}}}(s) \\
=& \frac{1}{1-\tilde{\gamma}} \sum_{s \in \mathcal{S}} \underbrace{\left(\sum_{s_0 \in \mathcal{S}} \rho_0(s_0) d^{s_0,\lambda}_{\pi_{\boldsymbol{\theta}}}(s) \right)}_{= d^{\lambda}_{\pi_{\boldsymbol{\theta}}}(s)} R^{(\lambda)}_{\pi_{\boldsymbol{\theta}}}(s) \\
=& \frac{1}{1-\tilde{\gamma}} \sum_{s \in \mathcal{S}} d^{\lambda}_{\pi_{\boldsymbol{\theta}}}(s) R^{(\lambda)}_{\pi_{\boldsymbol{\theta}}}(s) = \frac{1}{1-\tilde{\gamma}} \mathbb{E}_{s \sim d^{\lambda}_{\pi_{\boldsymbol{\theta}}}(\cdot)} \left[R^{(\lambda)}_{\pi_{\boldsymbol{\theta}}}(s) \right].
\end{aligned} \tag{63}
$$

Finally, we summarize above results in the following Lemma 1.

Lemma 1. *The objective $J(\pi_{\boldsymbol{\theta}})$ (45) can be rewritten as the following version:*

$$J(\pi_{\boldsymbol{\theta}}) = \frac{1}{1-\tilde{\gamma}} \sum_{s\in\mathcal{S}} d^{\lambda}_{\pi_{\boldsymbol{\theta}}}(s) R^{(\lambda)}_{\pi_{\boldsymbol{\theta}}}(s) = \frac{1}{1-\tilde{\gamma}} \mathbb{E}_{s\sim d^{\lambda}_{\pi_{\boldsymbol{\theta}}}(\cdot)} \left[R^{(\lambda)}_{\pi_{\boldsymbol{\theta}}}(s) \right].$$

E PROOF OF THEOREM 1

We need the following Proposition 4 to prove Theorem 1, which illustrates an identity for the objective function of policy optimization.

Proposition 4. *For any function $\varphi(\cdot) : \mathcal{S} \to \mathbb{R}$, for any policy $\pi_{\boldsymbol{\theta}}$, for any trajectory satisfies $\tau = \{s_t, a_t, r_{t+1}\}_{t\geq 0} \sim \pi_{\boldsymbol{\theta}}$, let*

$$\begin{aligned} \delta^{\varphi}_t &= r(s_{t+1}|s_t, a_t) + \gamma\varphi(s_{t+1}) - \varphi(s_t), \\ \delta^{\varphi}_{\pi_{\boldsymbol{\theta}},t}(s) &= \mathbb{E}_{s_t\sim\mathbb{P}_{\pi_{\boldsymbol{\theta}}}(\cdot|s), a_t\sim\pi_{\boldsymbol{\theta}}(\cdot|s_t), s_{t+1}\sim\mathbb{P}(\cdot|s_t,a_t)} \left[\delta^{\varphi}_t\right], \end{aligned}$$

then, the objective $J(\pi_{\boldsymbol{\theta}})$ (63) can be rewritten as the following version:

$$\begin{aligned} J(\pi_{\boldsymbol{\theta}}) =& \mathbb{E}_{s_0\sim\rho_0(\cdot)}[\varphi(s_0)] + \frac{1}{1-\tilde{\gamma}} \sum_{s\in\mathcal{S}} d^{\lambda}_{\pi_{\boldsymbol{\theta}}}(s) \left(\sum_{t=0}^{\infty} \gamma^t \lambda^t \delta^{\varphi}_{\pi_{\boldsymbol{\theta}},t}(s) \right) \\ =& \mathbb{E}_{s_0\sim\rho_0(\cdot)}[\varphi(s_0)] + \frac{1}{1-\tilde{\gamma}} \mathbb{E}_{s\sim d^{\lambda}_{\pi_{\boldsymbol{\theta}}}(\cdot)} \left[\sum_{t=0}^{\infty} \gamma^t \lambda^t \delta^{\varphi}_{\pi_{\boldsymbol{\theta}},t}(s) \right]. \end{aligned} \tag{64}$$

We present the proof of of Proposition 4 at the end of this section, see Section E.2.

We introduce a vector $\boldsymbol{\delta}^{\varphi}_{\pi_{\boldsymbol{\theta}},t} \in \mathbb{R}^{|\mathcal{S}|}$ and its components are: for any $s \in \mathcal{S}$

$$\boldsymbol{\delta}^{\varphi}_{\pi_{\boldsymbol{\theta}},t}[s] = \delta^{\varphi}_{\pi_{\boldsymbol{\theta}},t}(s). \tag{65}$$

Then, we rewrite the objective as the following vector version

$$J(\pi_{\boldsymbol{\theta}}) = \mathbb{E}_{s_0\sim\rho_0(\cdot)}[\varphi(s_0)] + \frac{1}{1-\tilde{\gamma}} \sum_{t=0}^{\infty} \gamma^t \lambda^t \langle \mathbf{d}^{\lambda}_{\pi_{\boldsymbol{\theta}}}, \boldsymbol{\delta}^{\varphi}_{\pi_{\boldsymbol{\theta}},t} \rangle, \tag{66}$$

where $\langle \cdot, \cdot \rangle$ denotes inner production between two vectors.

E.1 PROOF OF THEOREM 1

Theorem 1 (Generalized Policy Performance Difference) *For any function $\varphi(\cdot) : \mathcal{S} \to \mathbb{R}$, for two arbitrary policy $\pi_{\boldsymbol{\theta}}$ and $\pi_{\boldsymbol{\theta}'}$, for any $p, q \in [1, \infty)$ such that $\frac{1}{p} + \frac{1}{q} = 1$, The following bound holds:*

$$\frac{1}{1-\tilde{\gamma}} \sum_{t=0}^{\infty} \gamma^t \lambda^t M^{\varphi,-}_{p,q,t}(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}'}) \leq J(\pi_{\boldsymbol{\theta}}) - J(\pi_{\boldsymbol{\theta}'}) \leq \frac{1}{1-\tilde{\gamma}} \sum_{t=0}^{\infty} \gamma^t \lambda^t M^{\varphi,+}_{p,q,t}(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}'}), \tag{67}$$

where the terms $M^{\varphi,-}_{p,q,t}$ and $M^{\varphi,+}_{p,q,t}$ are defined in (83)-(84).

Proof. (of Theorem 1)

We consider two arbitrary policies $\pi_{\boldsymbol{\theta}}$ and $\pi_{\boldsymbol{\theta}'}$ with different parameters $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$, let

$$D^{\varphi,(\lambda)}_t(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}'}) =: \langle \mathbf{d}^{\lambda}_{\pi_{\boldsymbol{\theta}}}, \boldsymbol{\delta}^{\varphi}_{\pi_{\boldsymbol{\theta}},t} \rangle - \langle \mathbf{d}^{\lambda}_{\pi_{\boldsymbol{\theta}'}}, \boldsymbol{\delta}^{\varphi}_{\pi_{\boldsymbol{\theta}'},t} \rangle. \tag{68}$$

According to (66), we obtain performance difference as follows,

$$\begin{aligned} J(\pi_{\boldsymbol{\theta}}) - J(\pi_{\boldsymbol{\theta}'}) =& \frac{1}{1-\tilde{\gamma}} \sum_{t=0}^{\infty} \gamma^t \lambda^t \left(\langle \mathbf{d}^{\lambda}_{\pi_{\boldsymbol{\theta}}}, \boldsymbol{\delta}^{\varphi}_{\pi_{\boldsymbol{\theta}},t} \rangle - \langle \mathbf{d}^{\lambda}_{\pi_{\boldsymbol{\theta}'}}, \boldsymbol{\delta}^{\varphi}_{\pi_{\boldsymbol{\theta}'},t} \rangle \right) \\ =& \frac{1}{1-\tilde{\gamma}} \sum_{t=0}^{\infty} \gamma^t \lambda^t D^{\varphi,(\lambda)}_t(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}'}), \end{aligned} \tag{69}$$

which requires us to consider the boundedness of the difference $D_t^{\varphi,(\lambda)}(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}'})$ (68) .

Step 1: Bound the term $D_t^{\varphi,(\lambda)}(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}'})$ (68).

We rewrite the first term of (68) as follows,

$$\langle \mathbf{d}^{\lambda}_{\pi_{\boldsymbol{\theta}}}, \boldsymbol{\delta}^{\varphi}_{\pi_{\boldsymbol{\theta}},t} \rangle = \langle \mathbf{d}^{\lambda}_{\pi_{\boldsymbol{\theta}'}}, \boldsymbol{\delta}^{\varphi}_{\pi_{\boldsymbol{\theta}},t} \rangle + \langle \mathbf{d}^{\lambda}_{\pi_{\boldsymbol{\theta}}} - \mathbf{d}^{\lambda}_{\pi_{\boldsymbol{\theta}'}}, \boldsymbol{\delta}^{\varphi}_{\pi_{\boldsymbol{\theta}},t} \rangle, \tag{70}$$

which is bounded by applying Hölder's inequality to the term $\langle \mathbf{d}^{\lambda}_{\pi_{\boldsymbol{\theta}}} - \mathbf{d}^{\lambda}_{\pi_{\boldsymbol{\theta}'}}, \boldsymbol{\delta}^{\varphi}_{\pi_{\boldsymbol{\theta}},t} \rangle$, we rewrite (70) as follows,

$$\begin{aligned} &\langle \mathbf{d}^{\lambda}_{\pi_{\boldsymbol{\theta}'}}, \boldsymbol{\delta}^{\varphi}_{\pi_{\boldsymbol{\theta}},t} \rangle - \|\mathbf{d}^{\lambda}_{\pi_{\boldsymbol{\theta}}} - \mathbf{d}^{\lambda}_{\pi_{\boldsymbol{\theta}'}}\|_p \|\boldsymbol{\delta}^{\varphi}_{\pi_{\boldsymbol{\theta}},t}\|_q \\ \leq &\langle \mathbf{d}^{\lambda}_{\pi_{\boldsymbol{\theta}}}, \boldsymbol{\delta}^{\varphi}_{\pi_{\boldsymbol{\theta}},t} \rangle \leq \langle \mathbf{d}^{\lambda}_{\pi_{\boldsymbol{\theta}'}}, \boldsymbol{\delta}^{\varphi}_{\pi_{\boldsymbol{\theta}},t} \rangle + \|\mathbf{d}^{\lambda}_{\pi_{\boldsymbol{\theta}}} - \mathbf{d}^{\lambda}_{\pi_{\boldsymbol{\theta}'}}\|_p \|\boldsymbol{\delta}^{\varphi}_{\pi_{\boldsymbol{\theta}},t}\|_q, \end{aligned} \tag{71}$$

where $p, q \in [1, \infty)$ and $\frac{1}{p} + \frac{1}{q} = 1$. Let

$$\epsilon^{\varphi,(\lambda)}_{p,q,t}(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}'}) =: \|\mathbf{d}^{\lambda}_{\pi_{\boldsymbol{\theta}}} - \mathbf{d}^{\lambda}_{\pi_{\boldsymbol{\theta}'}}\|_p \|\boldsymbol{\delta}^{\varphi}_{\pi_{\boldsymbol{\theta}},t}\|_q,$$

then we rewrite Eq.(71) as follows,

$$\langle \mathbf{d}^{\lambda}_{\pi_{\boldsymbol{\theta}'}}, \boldsymbol{\delta}^{\varphi}_{\pi_{\boldsymbol{\theta}},t} \rangle - \epsilon^{\varphi,(\lambda)}_{p,q,t}(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}'}) \leq \langle \mathbf{d}^{\lambda}_{\pi_{\boldsymbol{\theta}}}, \boldsymbol{\delta}^{\varphi}_{\pi_{\boldsymbol{\theta}},t} \rangle \leq \langle \mathbf{d}^{\lambda}_{\pi_{\boldsymbol{\theta}'}}, \boldsymbol{\delta}^{\varphi}_{\pi_{\boldsymbol{\theta}},t} \rangle + \epsilon^{\varphi,(\lambda)}_{p,q,t}(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}'}). \tag{72}$$

Let

$$M^{\varphi}_t(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}'}) =: \underbrace{\langle \mathbf{d}^{\lambda}_{\pi_{\boldsymbol{\theta}'}}, \boldsymbol{\delta}^{\varphi}_{\pi_{\boldsymbol{\theta}},t} \rangle}_{\text{Term-I}} - \underbrace{\langle \mathbf{d}^{\lambda}_{\pi_{\boldsymbol{\theta}'}}, \boldsymbol{\delta}^{\varphi}_{\pi_{\boldsymbol{\theta}'},t} \rangle}_{\text{Term-II}}, \tag{73}$$

combining the (68) and (72), we achieve the boundedness of $D^{\varphi}_t(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}'})$ as follows

$$M^{\varphi}_t(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}'}) - \epsilon^{\varphi,(\lambda)}_{p,q,t}(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}'}) \leq D^{\varphi}_t(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}'}) \leq M^{\varphi}_t(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}'}) + \epsilon^{\varphi,(\lambda)}_{p,q,t}(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}'}). \tag{74}$$

Step 2: Analyze the term $M^{\varphi}_t(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}'})$ (73).

To analyze (74) further, we need to consider the first term appears in $M^{\varphi}_t(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}'})$ (73):

$$\begin{aligned} \text{Term-I (73)} =&\langle \mathbf{d}^{\lambda}_{\pi_{\boldsymbol{\theta}'}}, \boldsymbol{\delta}^{\varphi}_{\pi_{\boldsymbol{\theta}},t} \rangle \\ =&\sum_{s \in \mathcal{S}} d^{\lambda}_{\pi_{\boldsymbol{\theta}'}}(s) \delta^{\varphi}_{\pi_{\boldsymbol{\theta}},t}(s) = \mathbb{E}_{s \sim d^{\lambda}_{\pi_{\boldsymbol{\theta}'}}(\cdot)} \left[\delta^{\varphi}_{\pi_{\boldsymbol{\theta}},t}(s) \right] \qquad (75) \\ \overset{(65)}{=}&\mathbb{E}_{s \sim d^{\lambda}_{\pi_{\boldsymbol{\theta}'}}(\cdot)} \left[\mathbb{E}_{s_t \sim \mathbb{P}_{\pi_{\boldsymbol{\theta}}}(\cdot|s)} [\delta^{\varphi}_{\pi_{\boldsymbol{\theta}}}(s_t)] \right]. \qquad (76) \end{aligned}$$

We notice the following relationship

$$\delta^{\varphi}_{\pi_{\boldsymbol{\theta}},t}(s) = \mathop{\mathbb{E}}_{\substack{s_t \sim \mathbb{P}_{\pi_{\boldsymbol{\theta}}}(\cdot|s) \\ a_t \sim \pi_{\boldsymbol{\theta}}(\cdot|s_t) \\ s_{t+1} \sim \mathbb{P}(\cdot|s_t, a_t)}} [\delta^{\varphi}_t] = \mathop{\mathbb{E}}_{\substack{s_t \sim \mathbb{P}_{\pi_{\boldsymbol{\theta}'}}(\cdot|s) \\ a_t \sim \pi_{\boldsymbol{\theta}'}(\cdot|s_t) \\ s_{t+1} \sim \mathbb{P}(\cdot|s_t, a_t)}} \left[\frac{\pi_{\boldsymbol{\theta}}(a_t|s_t)}{\pi_{\boldsymbol{\theta}'}(a_t|s_t)} \delta^{\varphi}_t \right], \tag{77}$$

which holds since we use importance sampling: for any distribution $p(\cdot)$ and $q(\cdot)$, for any random variable function $f(\cdot)$,

$$\mathbb{E}_{x \sim p(x)}[f(x)] = \mathbb{E}_{x \sim q(x)} \left[\frac{p(x)}{q(x)} f(x) \right].$$

According to (75), (77), we rewrite the term $\langle \mathbf{d}^{\lambda}_{\pi_{\boldsymbol{\theta}'}}, \boldsymbol{\delta}^{\varphi}_{\pi_{\boldsymbol{\theta}},t} \rangle$ in Eq.(73) as follows,

$$\text{Term-I (73)} = \langle \mathbf{d}^{\lambda}_{\pi_{\boldsymbol{\theta}'}}, \boldsymbol{\delta}^{\varphi}_{\pi_{\boldsymbol{\theta}},t} \rangle = \sum_{s \in \mathcal{S}} d^{\lambda}_{\pi_{\boldsymbol{\theta}'}}(s) \left(\mathop{\mathbb{E}}_{\substack{s_t \sim \mathbb{P}_{\pi_{\boldsymbol{\theta}'}}(\cdot|s) \\ a_t \sim \pi_{\boldsymbol{\theta}'}(\cdot|s_t) \\ s_{t+1} \sim \mathbb{P}(\cdot|s_t, a_t)}} \left[\frac{\pi_{\boldsymbol{\theta}}(a_t|s_t)}{\pi_{\boldsymbol{\theta}'}(a_t|s_t)} \delta^{\varphi}_t \right] \right). \tag{78}$$

Now, we consider the second term appears in $M_t^{\varphi}(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}'})$ (73):

$$\begin{aligned}\text{Term-II (73)} &= \langle \mathbf{d}_{\pi_{\boldsymbol{\theta}'}}^{\lambda}, \boldsymbol{\delta}_{\pi_{\boldsymbol{\theta}'},t}^{\varphi} \rangle \\ &= \sum_{s\in\mathcal{S}} d_{\pi_{\boldsymbol{\theta}'}}^{\lambda}(s)\delta_{\pi_{\boldsymbol{\theta}'},t}^{\varphi}(s) = \sum_{s\in\mathcal{S}} d_{\pi_{\boldsymbol{\theta}'}}^{\lambda}(s) \begin{pmatrix} \mathbb{E} \\ {\scriptstyle s_t\sim\mathbb{P}_{\pi_{\boldsymbol{\theta}'}}(\cdot|s)} \\ {\scriptstyle a_t\sim\pi_{\boldsymbol{\theta}'}(\cdot|s_t)} \\ {\scriptstyle s_{t+1}\sim\mathbb{P}(\cdot|s_t,a_t)} \end{pmatrix} [\delta_t^{\varphi}]. \end{aligned} \tag{79}$$

Finally, take the results (78) and (79) to (73), we obtain the difference between $\langle \mathbf{d}_{\pi_{\boldsymbol{\theta}'}}^{\lambda}, \boldsymbol{\delta}_{\pi_{\boldsymbol{\theta}},t}^{\varphi} \rangle$ and $\langle \mathbf{d}_{\pi_{\boldsymbol{\theta}'}}^{\lambda}, \boldsymbol{\delta}_{\pi_{\boldsymbol{\theta}'},t}^{\varphi} \rangle$, i.e., we achieve a identity for $M_t^{\varphi}(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}'})$ (73) as follows,

$$\begin{aligned} M_t^{\varphi}(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}'}) &\overset{(73)}{=} \langle \mathbf{d}_{\pi_{\boldsymbol{\theta}'}}^{\lambda}, \boldsymbol{\delta}_{\pi_{\boldsymbol{\theta}},t}^{\varphi} \rangle - \langle \mathbf{d}_{\pi_{\boldsymbol{\theta}'}}^{\lambda}, \boldsymbol{\delta}_{\pi_{\boldsymbol{\theta}'},t}^{\varphi} \rangle \\ &\overset{(78),(79)}{=} \sum_{s\in\mathcal{S}} d_{\pi_{\boldsymbol{\theta}'}}^{\lambda}(s) \begin{pmatrix} \mathbb{E} \\ {\scriptstyle s_t\sim\mathbb{P}_{\pi_{\boldsymbol{\theta}'}}(\cdot|s)} \\ {\scriptstyle a_t\sim\pi_{\boldsymbol{\theta}'}(\cdot|s_t)} \\ {\scriptstyle s_{t+1}\sim\mathbb{P}(\cdot|s_t,a_t)} \end{pmatrix} \left[\left(\frac{\pi_{\boldsymbol{\theta}}(a_t|s_t)}{\pi_{\boldsymbol{\theta}'}(a_t|s_t)} - 1 \right) \delta_t^{\varphi} \right]. \end{aligned} \tag{80}$$

To simplify expression, we introduce a notation as follows,

$$\Delta_t^{\varphi}(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}'}, s) =: \underset{\substack{s_t\sim\mathbb{P}_{\pi_{\boldsymbol{\theta}'}}(\cdot|s) \\ a_t\sim\pi_{\boldsymbol{\theta}'}(\cdot|s_t) \\ s_{t+1}\sim\mathbb{P}(\cdot|s_t,a_t)}}{\mathbb{E}} \left[\left(\frac{\pi_{\boldsymbol{\theta}}(a_t|s_t)}{\pi_{\boldsymbol{\theta}'}(a_t|s_t)} - 1 \right) \delta_t^{\varphi} \right], \tag{81}$$

and we use a vector $\boldsymbol{\Delta}_t^{\varphi}(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}'}) \in \mathbb{R}^{|\mathcal{S}|}$ to store all the values $\{\Delta_t^{\varphi}(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}'}, s)\}_{s\in\mathcal{S}}$:

$$\boldsymbol{\Delta}_t^{\varphi}(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}'})[s] = \Delta_t^{\varphi}(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}'}, s).$$

Then we rewrite $\langle \mathbf{d}_{\pi_{\boldsymbol{\theta}'}}^{\lambda}, \boldsymbol{\delta}_{\pi_{\boldsymbol{\theta}},t}^{\varphi} \rangle - \langle \mathbf{d}_{\pi_{\boldsymbol{\theta}'}}^{\lambda}, \boldsymbol{\delta}_{\pi_{\boldsymbol{\theta}'},t}^{\varphi} \rangle$ (80) as follows,

$$\begin{aligned} M_t^{\varphi}(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}'}) =& \langle \mathbf{d}_{\pi_{\boldsymbol{\theta}'}}^{\lambda}, \boldsymbol{\delta}_{\pi_{\boldsymbol{\theta}},t}^{\varphi} \rangle - \langle \mathbf{d}_{\pi_{\boldsymbol{\theta}'}}^{\lambda}, \boldsymbol{\delta}_{\pi_{\boldsymbol{\theta}'},t}^{\varphi} \rangle \\ \overset{(80)}{=}& \sum_{s\in\mathcal{S}} d_{\pi_{\boldsymbol{\theta}'}}^{\lambda}(s)\Delta_t^{\varphi}(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}'}, s) = \langle \mathbf{d}_{\pi_{\boldsymbol{\theta}'}}^{\lambda}, \boldsymbol{\Delta}_t^{\varphi}(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}'}) \rangle. \end{aligned}$$

Step 3: Bound on $J(\pi_{\boldsymbol{\theta}}) - J(\pi_{\boldsymbol{\theta}'})$.

Recall (74), taking above result in it, we obtain

$$\langle \mathbf{d}_{\pi_{\boldsymbol{\theta}'}}^{\lambda}, \boldsymbol{\Delta}_t^{\varphi}(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}'}) \rangle - \epsilon_{p,q,t}^{\varphi,(\lambda)}(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}'}) \leq D_t^{\varphi}(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}'}) \leq \langle \mathbf{d}_{\pi_{\boldsymbol{\theta}'}}^{\lambda}, \boldsymbol{\Delta}_t^{\varphi}(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}'}) \rangle + \epsilon_{p,q,t}^{\varphi,(\lambda)}(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}'}). \tag{82}$$

Finally, let

$$M_{p,q,t}^{\varphi,-}(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}'}) = \langle \mathbf{d}_{\pi_{\boldsymbol{\theta}'}}^{\lambda}, \boldsymbol{\Delta}_t^{\varphi}(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}'}) \rangle - \epsilon_{p,q,t}^{\varphi,(\lambda)}(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}'}) \tag{83}$$

$$= \sum_{s\in\mathcal{S}} d_{\pi_{\boldsymbol{\theta}'}}^{\lambda}(s) \begin{pmatrix} \mathbb{E} \\ {\scriptstyle s_t\sim\mathbb{P}_{\pi_{\boldsymbol{\theta}'}}(\cdot|s)} \\ {\scriptstyle a_t\sim\pi_{\boldsymbol{\theta}'}(\cdot|s_t)} \\ {\scriptstyle s_{t+1}\sim\mathbb{P}(\cdot|s_t,a_t)} \end{pmatrix} \left[\left(\frac{\pi_{\boldsymbol{\theta}}(a_t|s_t)}{\pi_{\boldsymbol{\theta}'}(a_t|s_t)} - 1 \right) \delta_t^{\varphi} \right] - \|\mathbf{d}_{\pi_{\boldsymbol{\theta}}}^{\lambda} - \mathbf{d}_{\pi_{\boldsymbol{\theta}'}}^{\lambda}\|_p \|\boldsymbol{\delta}_{\pi_{\boldsymbol{\theta}},t}^{\varphi}\|_q$$

$$= \mathbb{E}_{s\sim d_{\pi_{\boldsymbol{\theta}'}}^{\lambda}(\cdot)} \left[\begin{pmatrix} \mathbb{E} \\ {\scriptstyle s_t\sim\mathbb{P}_{\pi_{\boldsymbol{\theta}'}}(\cdot|s)} \\ {\scriptstyle a_t\sim\pi_{\boldsymbol{\theta}'}(\cdot|s_t)} \\ {\scriptstyle s_{t+1}\sim\mathbb{P}(\cdot|s_t,a_t)} \end{pmatrix} \left[\left(\frac{\pi_{\boldsymbol{\theta}}(a_t|s_t)}{\pi_{\boldsymbol{\theta}'}(a_t|s_t)} - 1 \right) \delta_t^{\varphi} \right] \right] - \|\mathbf{d}_{\pi_{\boldsymbol{\theta}}}^{\lambda} - \mathbf{d}_{\pi_{\boldsymbol{\theta}'}}^{\lambda}\|_p \|\boldsymbol{\delta}_{\pi_{\boldsymbol{\theta}},t}^{\varphi}\|_q.$$

and

$$M^{\varphi,+}_{p,q,t}(\pi_{\boldsymbol{\theta}},\pi_{\boldsymbol{\theta}'}) = \langle \mathbf{d}^{\lambda}_{\pi_{\boldsymbol{\theta}'}}, \boldsymbol{\Delta}^{\varphi}_{t}(\pi_{\boldsymbol{\theta}},\pi_{\boldsymbol{\theta}'})\rangle + \epsilon^{\varphi,(\lambda)}_{p,q,t}(\pi_{\boldsymbol{\theta}},\pi_{\boldsymbol{\theta}'}) \tag{84}$$

$$= \sum_{s\in\mathcal{S}} d^{\lambda}_{\pi_{\boldsymbol{\theta}'}}(s) \left(\mathop{\mathbb{E}}_{\substack{s_t\sim\mathbb{P}_{\pi_{\boldsymbol{\theta}'}}(\cdot|s)\\ a_t\sim\pi_{\boldsymbol{\theta}'}(\cdot|s_t)\\ s_{t+1}\sim\mathbb{P}(\cdot|s_t,a_t)}} \left[\left(\frac{\pi_{\boldsymbol{\theta}}(a_t|s_t)}{\pi_{\boldsymbol{\theta}'}(a_t|s_t)} - 1 \right) \delta^{\varphi}_{t} \right] \right) + \|\mathbf{d}^{\lambda}_{\pi_{\boldsymbol{\theta}}} - \mathbf{d}^{\lambda}_{\pi_{\boldsymbol{\theta}'}}\|_p \|\boldsymbol{\delta}^{\varphi}_{\pi_{\boldsymbol{\theta}},t}\|_q$$

$$= \mathbb{E}_{s\sim d^{\lambda}_{\pi_{\boldsymbol{\theta}'}}(\cdot)} \left[\mathop{\mathbb{E}}_{\substack{s_t\sim\mathbb{P}_{\pi_{\boldsymbol{\theta}'}}(\cdot|s)\\ a_t\sim\pi_{\boldsymbol{\theta}'}(\cdot|s_t)\\ s_{t+1}\sim\mathbb{P}(\cdot|s_t,a_t)}} \left[\left(\frac{\pi_{\boldsymbol{\theta}}(a_t|s_t)}{\pi_{\boldsymbol{\theta}'}(a_t|s_t)} - 1 \right) \delta^{\varphi}_{t} \right] \right] + \|\mathbf{d}^{\lambda}_{\pi_{\boldsymbol{\theta}}} - \mathbf{d}^{\lambda}_{\pi_{\boldsymbol{\theta}'}}\|_p \|\boldsymbol{\delta}^{\varphi}_{\pi_{\boldsymbol{\theta}},t}\|_q.$$

According to (69) and (82), we achieve the boundedness of performance difference between two arbitrary policies $\pi_{\boldsymbol{\theta}}$ and $\pi_{\boldsymbol{\theta}'}$:

$$\underbrace{\frac{1}{1-\tilde{\gamma}} \sum_{t=0}^{\infty} \gamma^t \lambda^t M^{\varphi,-}_{p,q,t}(\pi_{\boldsymbol{\theta}},\pi_{\boldsymbol{\theta}'})}_{=:L^{\varphi,-}_{p,q,}} \leq J(\pi_{\boldsymbol{\theta}}) - J(\pi_{\boldsymbol{\theta}'}) \leq \underbrace{\frac{1}{1-\tilde{\gamma}} \sum_{t=0}^{\infty} \gamma^t \lambda^t M^{\varphi,+}_{p,q,t}(\pi_{\boldsymbol{\theta}},\pi_{\boldsymbol{\theta}'})}_{=:L^{\varphi,+}_{p,q,}}. \tag{85}$$

□

E.2 Proof of Proposition 4

Proof. (of Proposition 4).

Step 1: Rewrite the objective $J(\pi_\theta)$ in Eq.(63).

We rewrite the discounted distribution $\mathbf{d}^{\lambda}_{\pi_{\boldsymbol{\theta}}}$ (59) as follows,

$$\boldsymbol{\rho}_0 - \frac{1}{1-\tilde{\gamma}} \mathbf{d}^{\lambda}_{\pi_{\boldsymbol{\theta}}} + \frac{\tilde{\gamma}}{1-\tilde{\gamma}} \mathbf{P}^{(\lambda)}_{\pi_{\boldsymbol{\theta}}} \mathbf{d}^{\lambda}_{\pi_{\boldsymbol{\theta}}} = \mathbf{0}. \tag{86}$$

Let $\varphi(\cdot)$ be a real number function defined on the state space $\mathcal{S}$, i.e., $\varphi : \mathcal{S} \to \mathbb{R}$. Then we define a vector function $\boldsymbol{\phi}(\cdot) \in \mathbb{R}^{|\mathcal{S}|}$ to collect all the values $\{\varphi(s)\}_{s\in\mathcal{S}}$, and its components are

$$\boldsymbol{\phi}[s] = \varphi(s), \quad s \in \mathcal{S}.$$

Now, we take the inner product between the vector $\boldsymbol{\phi}$ and (86), we have

$$\begin{aligned} 0 &= \langle \boldsymbol{\rho}_0 - \frac{1}{1-\tilde{\gamma}} \mathbf{d}^{\lambda}_{\pi_{\boldsymbol{\theta}}} + \frac{\tilde{\gamma}}{1-\tilde{\gamma}} \mathbf{P}^{(\lambda)}_{\pi_{\boldsymbol{\theta}}} \mathbf{d}^{\lambda}_{\pi_{\boldsymbol{\theta}}}, \boldsymbol{\phi} \rangle \\ &= \langle \boldsymbol{\rho}_0, \boldsymbol{\phi} \rangle - \frac{1}{1-\tilde{\gamma}} \langle \mathbf{d}^{\lambda}_{\pi_{\boldsymbol{\theta}}}, \boldsymbol{\phi} \rangle + \frac{\tilde{\gamma}}{1-\tilde{\gamma}} \langle \mathbf{P}^{(\lambda)}_{\pi_{\boldsymbol{\theta}}} \mathbf{d}^{\lambda}_{\pi_{\boldsymbol{\theta}}}, \boldsymbol{\phi} \rangle. \end{aligned} \tag{87}$$

We express the first term $\langle \boldsymbol{\rho}_0, \boldsymbol{\phi} \rangle$ of (87) as follows,

$$\langle \boldsymbol{\rho}_0, \boldsymbol{\phi} \rangle = \sum_{s\in\mathcal{S}} \rho_0(s)\varphi(s) = \mathbb{E}_{s\sim\rho_0(\cdot)}[\varphi(s)]. \tag{88}$$

We express the second term $\langle \mathbf{d}^{\lambda}_{\pi_{\boldsymbol{\theta}}}, \boldsymbol{\phi} \rangle$ of (87) as follows,

$$-\frac{1}{1-\tilde{\gamma}} \langle \mathbf{d}^{\lambda}_{\pi_{\boldsymbol{\theta}}}, \boldsymbol{\phi} \rangle = -\frac{1}{1-\tilde{\gamma}} \sum_{s\in\mathcal{S}} d^{\lambda}_{\pi_{\boldsymbol{\theta}}}(s)\varphi(s) = -\frac{1}{1-\tilde{\gamma}} \mathbb{E}_{s\sim d^{\lambda}_{\pi_{\boldsymbol{\theta}}}(\cdot)}[\varphi(s)]. \tag{89}$$

We express the third term $\langle \tilde{\gamma}\mathbf{P}_{\pi_{\boldsymbol{\theta}}}^{(\lambda)}\mathbf{d}_{\pi_{\boldsymbol{\theta}}}^{\lambda}, \boldsymbol{\phi}\rangle$ of (87) as follows,

$$\begin{aligned}\frac{\tilde{\gamma}}{1-\tilde{\gamma}}\langle \mathbf{P}_{\pi_{\boldsymbol{\theta}}}^{(\lambda)}\mathbf{d}_{\pi_{\boldsymbol{\theta}}}^{\lambda}, \boldsymbol{\phi}\rangle =& \frac{\tilde{\gamma}}{1-\tilde{\gamma}}\sum_{s^{'}\in\mathcal{S}}\left(\mathbf{P}_{\pi_{\boldsymbol{\theta}}}^{(\lambda)}\mathbf{d}_{\pi_{\boldsymbol{\theta}}}^{\lambda}\right)[s^{'}]\varphi(s^{'}) \\ =& \frac{\tilde{\gamma}}{1-\tilde{\gamma}}\sum_{s^{'}\in\mathcal{S}}\left(\sum_{s\in\mathcal{S}}\mathbb{P}_{\pi_{\boldsymbol{\theta}}}^{(\lambda)}(s^{'}|s)d_{\pi_{\boldsymbol{\theta}}}^{\lambda}(s)\right)\varphi(s^{'}). \end{aligned} \tag{90}$$

According to Lemma 1, put the results (63) and (87) together, we have

$$\begin{aligned} & J(\pi_{\boldsymbol{\theta}}) \\ \overset{(63),(87)}{=}& \frac{1}{1-\tilde{\gamma}}\sum_{s\in\mathcal{S}}d_{\pi_{\boldsymbol{\theta}}}^{\lambda}(s)R_{\pi_{\boldsymbol{\theta}}}^{(\lambda)}(s) + \left\langle \boldsymbol{\rho}_0 - \frac{1}{1-\tilde{\gamma}}\mathbf{d}_{\pi_{\boldsymbol{\theta}}}^{\lambda} + \frac{\tilde{\gamma}}{1-\tilde{\gamma}}\mathbf{P}_{\pi_{\boldsymbol{\theta}}}^{(\lambda)}\mathbf{d}_{\pi_{\boldsymbol{\theta}}}^{\lambda}, \boldsymbol{\phi}\right\rangle \\ =& \mathbb{E}_{s_0\sim\rho_0(\cdot)}[\varphi(s_0)] + \frac{1}{1-\tilde{\gamma}}\sum_{s\in\mathcal{S}}d_{\pi_{\boldsymbol{\theta}}}^{\lambda}(s)\left(R_{\pi_{\boldsymbol{\theta}}}^{(\lambda)}(s) + \tilde{\gamma}\sum_{s^{'}\in\mathcal{S}}\mathbb{P}_{\pi_{\boldsymbol{\theta}}}^{(\lambda)}(s^{'}|s)\varphi(s^{'}) - \varphi(s)\right), \end{aligned} \tag{91}$$

where the last equation holds since we unfold (87) according to (88)-(90).

Step 2: Rewrite the term $\left(R_{\pi_{\boldsymbol{\theta}}}^{(\lambda)}(s) + \tilde{\gamma}\sum_{s^{'}\in\mathcal{S}}\mathbb{P}_{\pi_{\boldsymbol{\theta}}}^{(\lambda)}(s^{'}|s)\varphi(s^{'}) - \varphi(s)\right)$ in Eq.(91).

Then, we unfold the second term of (91) as follows,

$$R_{\pi_{\boldsymbol{\theta}}}^{(\lambda)}(s) + \tilde{\gamma}\sum_{s^{'}\in\mathcal{S}}\mathbb{P}_{\pi_{\boldsymbol{\theta}}}^{(\lambda)}(s^{'}|s)\varphi(s^{'}) - \varphi(s) \tag{92}$$

$$\overset{(53),(55)}{=}\sum_{t=0}^{\infty}(\gamma\lambda\mathbf{P}_{\pi_{\boldsymbol{\theta}}})^t\mathbf{r}_{\pi_{\boldsymbol{\theta}}}[s] + \tilde{\gamma}(1-\gamma\lambda)\sum_{s^{'}\in\mathcal{S}}\sum_{t=0}^{\infty}(\gamma\lambda)^t\left(\mathbf{P}_{\pi_{\boldsymbol{\theta}}}^{t+1}[s,s^{'}]\right)\varphi(s^{'}) - \varphi(s)$$

$$\overset{(50)}{=}\sum_{t=0}^{\infty}(\gamma\lambda\mathbf{P}_{\pi_{\boldsymbol{\theta}}})^t\mathbf{r}_{\pi_{\boldsymbol{\theta}}}[s] + \gamma(1-\lambda)\sum_{s^{'}\in\mathcal{S}}\sum_{t=0}^{\infty}(\gamma\lambda)^t\mathbb{P}_{\pi_{\boldsymbol{\theta}}}(s_{t+1}=s^{'}|s)\varphi(s^{'}) - \varphi(s). \tag{93}$$

Recall the terms $\mathbf{P}_{\pi_{\boldsymbol{\theta}}}^{(\lambda)}$, $\mathbf{r}_{\pi_{\boldsymbol{\theta}}}^{(\lambda)}[s]$ defined in (50)-(55),

$$R_{\pi_{\boldsymbol{\theta}}}^{(\lambda)}(s) + \gamma(1-\lambda)\sum_{s^{'}\in\mathcal{S}}\mathbb{P}_{\pi_{\boldsymbol{\theta}}}^{(\lambda)}(s^{'}|s)\varphi(s^{'}) - \varphi(s) \tag{94}$$

We consider the first term $R_{\pi_{\boldsymbol{\theta}}}^{(\lambda)}(s)$ of (92) as follows,

$$R_{\pi_{\boldsymbol{\theta}}}^{(\lambda)}(s) \overset{(50)=(55)}{=} \mathbf{r}_{\pi_{\boldsymbol{\theta}}}^{(\lambda)}[s] = \sum_{t=0}^{\infty}(\gamma\lambda)^t\mathbf{P}_{\pi_{\boldsymbol{\theta}}}^{t}\mathbf{r}_{\pi_{\boldsymbol{\theta}}}[s] = \sum_{t=0}^{\infty}\sum_{s_t\in\mathcal{S}}(\gamma\lambda)^t\mathbb{P}_{\pi_{\boldsymbol{\theta}}}(s_t|s)R_{\pi_{\boldsymbol{\theta}}}(s_t). \tag{95}$$

We consider the second term $\tilde{\gamma}\sum_{s\in\mathcal{S}}\mathbb{P}_{\pi_{\boldsymbol{\theta}}}^{(\lambda)}(s^{'}|s)\varphi(s)-\varphi(s)$ of (92) as follows,

$$
\begin{aligned}
&\tilde{\gamma}\sum_{s^{'}\in\mathcal{S}}\mathbb{P}_{\pi_{\boldsymbol{\theta}}}^{(\lambda)}(s^{'}|s)\varphi(s^{'})-\varphi(s)\\
\overset{(53)}{=}&\tilde{\gamma}(1-\gamma\lambda)\sum_{s^{'}\in\mathcal{S}}\sum_{t=0}^{\infty}(\gamma\lambda)^{t}\mathbb{P}_{\pi_{\boldsymbol{\theta}}}(s_{t+1}=s^{'}|s)\varphi(s^{'})-\varphi(s) &(96)\\
\overset{(50)}{=}&\gamma(1-\lambda)\sum_{s^{'}\in\mathcal{S}}\sum_{t=0}^{\infty}(\gamma\lambda)^{t}\mathbb{P}_{\pi_{\boldsymbol{\theta}}}(s_{t+1}=s^{'}|s)\varphi(s^{'})-\varphi(s) &(97)\\
=&\gamma\sum_{s^{'}\in\mathcal{S}}\sum_{t=0}^{\infty}(\gamma\lambda)^{t}\mathbb{P}_{\pi_{\boldsymbol{\theta}}}(s_{t+1}=s^{'}|s)\varphi(s^{'})-\sum_{s^{'}\in\mathcal{S}}\underbrace{\left(\sum_{t=0}^{\infty}(\gamma\lambda)^{t+1}\mathbb{P}_{\pi_{\boldsymbol{\theta}}}(s_{t+1}=s^{'}|s)\varphi(s^{'})\right)}_{=\sum_{t=1}^{\infty}(\gamma\lambda)^{t}\mathbb{P}_{\pi_{\boldsymbol{\theta}}}(s_{t}=s^{'}|s)\varphi(s^{'})}-\varphi(s)\\
=&\gamma\sum_{s^{'}\in\mathcal{S}}\sum_{t=0}^{\infty}(\gamma\lambda)^{t}\mathbb{P}_{\pi_{\boldsymbol{\theta}}}(s_{t+1}=s^{'}|s)\varphi(s^{'})-\underbrace{\left(\sum_{s^{'}\in\mathcal{S}}\sum_{t=1}^{\infty}(\gamma\lambda)^{t}\mathbb{P}_{\pi_{\boldsymbol{\theta}}}(s_{t}=s^{'}|s)\varphi(s^{'})+\varphi(s)\right)}_{=\sum_{s^{'}\in\mathcal{S}}\sum_{t=0}^{\infty}(\gamma\lambda)^{t}\mathbb{P}_{\pi_{\boldsymbol{\theta}}}(s_{t}=s^{'}|s)\varphi(s^{'})} &(98)\\
=&\gamma\sum_{s^{'}\in\mathcal{S}}\sum_{t=0}^{\infty}(\gamma\lambda)^{t}\mathbb{P}_{\pi_{\boldsymbol{\theta}}}(s_{t+1}=s^{'}|s)\varphi(s^{'})-\sum_{s_t\in\mathcal{S}}\sum_{t=0}^{\infty}(\gamma\lambda)^{t}\mathbb{P}_{\pi_{\boldsymbol{\theta}}}(s_{t}|s)\varphi(s), &(99)
\end{aligned}
$$

where the equation from Eq.(98) to Eq.(99) holds since: according to (41), we use the following identity

$$
\sum_{s^{'}\in\mathcal{S}}\mathbb{P}_{\pi_{\boldsymbol{\theta}}}(s_0=s^{'}|s)\varphi(s^{'})=\varphi(s).
$$

Furthermore, take the result (95) and (99) to (94), we have

$$
\begin{aligned}
&R_{\pi_{\boldsymbol{\theta}}}^{(\lambda)}(s)+\tilde{\gamma}\sum_{s^{'}\in\mathcal{S}}\mathbb{P}_{\pi_{\boldsymbol{\theta}}}^{(\lambda)}(s^{'}|s)\varphi(s^{'})-\varphi(s)\\
=&\sum_{t=0}^{\infty}(\gamma\lambda)^{t}\Bigg(\sum_{s_t\in\mathcal{S}}\mathbb{P}_{\pi_{\boldsymbol{\theta}}}(s_t|s)R_{\pi_{\boldsymbol{\theta}}}(s_t)+\gamma\sum_{s^{'}\in\mathcal{S}}\underbrace{\mathbb{P}_{\pi_{\boldsymbol{\theta}}}(s_{t+1}=s^{'}|s)\varphi(s^{'})}_{\overset{(42)}{=}\sum_{s_t\in\mathcal{S}}\mathbb{P}_{\pi_{\boldsymbol{\theta}}}(s_{t+1}=s^{'}|s_t)\mathbb{P}_{\pi_{\boldsymbol{\theta}}}(s_t|s)\varphi(s^{'})}\\
&\qquad\qquad-\sum_{s_t\in\mathcal{S}}\mathbb{P}_{\pi_{\boldsymbol{\theta}}}(s_t|s)\varphi(s_t)\Bigg) \qquad (100)\\
=&\sum_{t=0}^{\infty}(\gamma\lambda)^{t}\Bigg(\sum_{s_t\in\mathcal{S}}\mathbb{P}_{\pi_{\boldsymbol{\theta}}}(s_t|s)R_{\pi_{\boldsymbol{\theta}}}(s_t)+\gamma\sum_{s_t\in\mathcal{S}}\mathbb{P}_{\pi_{\boldsymbol{\theta}}}(s_t|s)\sum_{s_{t+1}\in\mathcal{S}}\mathbb{P}_{\pi_{\boldsymbol{\theta}}}(s_{t+1}|s_t)\varphi(s_{t+1})\\
&\qquad\qquad-\sum_{s_t\in\mathcal{S}}\mathbb{P}_{\pi_{\boldsymbol{\theta}}}(s_t|s)\varphi(s_t)\Bigg) \qquad (101)\\
=&\sum_{t=0}^{\infty}(\gamma\lambda)^{t}\sum_{s_t\in\mathcal{S}}\mathbb{P}_{\pi_{\boldsymbol{\theta}}}(s_t|s)\Bigg(\underbrace{\sum_{a_t\in\mathcal{A}}\pi_{\boldsymbol{\theta}}(a_t|s_t)\sum_{s_{t+1}\in\mathcal{S}}\mathbb{P}(s_{t+1}|s_t,a_t)r(s_{t+1}|s_t,a_t)}_{=R_{\pi_{\boldsymbol{\theta}}}(s_t)}\\
&\qquad\qquad+\gamma\underbrace{\sum_{a_t\in\mathcal{A}}\pi_{\boldsymbol{\theta}}(a_t|s_t)\sum_{s_{t+1}\in\mathcal{S}}\mathbb{P}(s_{t+1}|s_t,a_t)}_{=\mathbb{P}_{\pi_{\boldsymbol{\theta}}}(s_{t+1}|s_t)}\varphi(s_{t+1})-\varphi(s_t)\Bigg)\\
=&\sum_{t=0}^{\infty}(\gamma\lambda)^{t}\sum_{s_t\in\mathcal{S}}\mathbb{P}_{\pi_{\boldsymbol{\theta}}}(s_t|s)\sum_{a_t\in\mathcal{A}}\pi_{\boldsymbol{\theta}}(a_t|s_t)\sum_{s_{t+1}\in\mathcal{S}}\mathbb{P}(s_{t+1}|s_t,a_t)\left(r(s_{t+1}|s_t,a_t)+\gamma\varphi(s_{t+1})-\varphi(s_t)\right) \qquad (102)\\
=&\sum_{t=0}^{\infty}(\gamma\lambda)^{t}\mathbb{E}_{s_t\sim\mathbb{P}_{\pi_{\boldsymbol{\theta}}}(\cdot|s),a_t\sim\pi_{\boldsymbol{\theta}}(\cdot|s_t),s_{t+1}\sim\mathbb{P}(\cdot|s_t,a_t)}\left[r(s_{t+1}|s_t,a_t)+\gamma\varphi(s_{t+1})-\varphi(s_t)\right], \qquad (103)
\end{aligned}
$$

the equation from Eq.(99) to Eq.(100) holds since:

$$\mathbb{P}_{\pi_{\boldsymbol{\theta}}}(s_{t+1}|s)\overset{(42)}{=}\sum_{s_t\in\mathcal{S}}\mathbb{P}_{\pi_{\boldsymbol{\theta}}}(s_{t+1}|s_t)\mathbb{P}_{\pi_{\boldsymbol{\theta}}}(s_t|s);$$

the equation from Eq.(100) to Eq.(101) holds since we use the Markov property of the definition of MDP: for each time $t\in\mathbb{N}$,

$$\mathbb{P}_{\pi_{\boldsymbol{\theta}}}(s_{t+1}=s^{'}|s_t=s)=\mathbb{P}_{\pi_{\boldsymbol{\theta}}}(s^{'}|s);$$

the equation (102) the following identity:

$$\sum_{a_t\in\mathcal{A}}\pi_{\boldsymbol{\theta}}(a_t|s_t)=1,\qquad\sum_{s_{t+1}\in\mathcal{S}}\mathbb{P}(s_{t+1}|s_t,a_t)=1,$$

then

$$\varphi(s_t)=\sum_{a_t\in\mathcal{A}}\pi_{\boldsymbol{\theta}}(a_t|s_t)\sum_{s_{t+1}\in\mathcal{S}}\mathbb{P}(s_{t+1}|s_t,a)\varphi(s_t).$$

Step 3: Put all the result together.

Finally, let

$$\delta_t^{\varphi} = r(s_{t+1}|s_t, a_t) + \gamma\varphi(s_{t+1}) - \varphi(s_t),$$
$$\delta_{\pi_{\boldsymbol{\theta}},t}^{\varphi}(s) = \mathbb{E}_{s_t\sim\mathbb{P}_{\pi_{\boldsymbol{\theta}}}(\cdot|s),a_t\sim\pi_{\boldsymbol{\theta}}(\cdot|s_t),s_{t+1}\sim\mathbb{P}(\cdot|s_t,a_t)}\left[\delta_t^{\varphi}\right],$$

combining the results (91) and (103), we have

$$\begin{aligned} J(\pi_{\boldsymbol{\theta}}) =& \mathbb{E}_{s_0\sim\rho_0(\cdot)}[\varphi(s_0)] + \frac{1}{1-\tilde{\gamma}}\sum_{s\in\mathcal{S}} d_{\pi_{\boldsymbol{\theta}}}^{\lambda}(s)\left(\sum_{t=0}^{\infty}\gamma^t\lambda^t\delta_{\pi_{\boldsymbol{\theta}},t}^{\varphi}(s)\right) \\ =& \mathbb{E}_{s_0\sim\rho_0(\cdot)}[\varphi(s_0)] + \frac{1}{1-\tilde{\gamma}}\mathbb{E}_{s\sim d_{\pi_{\boldsymbol{\theta}}}^{\lambda}(\cdot)}\left[\sum_{t=0}^{\infty}\gamma^t\lambda^t\delta_{\pi_{\boldsymbol{\theta}},t}^{\varphi}(s)\right]. \end{aligned} \tag{104}$$

This concludes the proof of Proposition 4. □

E.3 LEMMA 2

Lemma 2. *Let $\|\mathbf{\Pi}_{\pi_{\boldsymbol{\theta}'}} - \mathbf{\Pi}_{\pi_{\boldsymbol{\theta}}}\|_{1,1}$ denote as the $L_{1,1}$-norm for the difference between two policy space $\{\pi_{\boldsymbol{\theta}}(a|s)\}_{(s,a)\in\mathcal{S}\times\mathcal{A}}$, $\{\pi_{\boldsymbol{\theta}'}(a|s)\}_{(s,a)\in\mathcal{S}\times\mathcal{A}}$, i.e.,*

$$\|\mathbf{\Pi}_{\pi_{\boldsymbol{\theta}'}} - \mathbf{\Pi}_{\pi_{\boldsymbol{\theta}}}\|_{1,1} =: \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\left|\pi_{\boldsymbol{\theta}'}(a|s) - \pi_{\boldsymbol{\theta}}(a|s)\right|. \tag{105}$$

The divergence between discounted future state visitation distributions, $\|\mathbf{d}_{\pi_{\boldsymbol{\theta}'}}^{\lambda} - \mathbf{d}_{\pi_{\boldsymbol{\theta}}}^{\lambda}\|_1$, is bounded as follows,

$$\|\mathbf{d}_{\pi_{\boldsymbol{\theta}'}}^{\lambda} - \mathbf{d}_{\pi_{\boldsymbol{\theta}}}^{\lambda}\|_1 \leq \frac{(1-\gamma\lambda)^2}{(1-\gamma)\left(1-\gamma\lambda\|\mathbf{\Pi}_{\pi_{\boldsymbol{\theta}'}} - \mathbf{\Pi}_{\pi_{\boldsymbol{\theta}}}\|_{1,1}\right)}\mathbb{E}_{s\sim d_{\pi_{\boldsymbol{\theta}}}^{\lambda}(\cdot)}\Big[2D_{\mathrm{TV}}(\pi_{\boldsymbol{\theta}'},\pi_{\boldsymbol{\theta}})[s]\Big]$$

and

$$\|\mathbf{d}_{\pi_{\boldsymbol{\theta}'}}^{\lambda} - \mathbf{d}_{\pi_{\boldsymbol{\theta}}}^{\lambda}\|_1 \leq \frac{(1-\gamma\lambda)^2}{(1-\gamma)\left(1-\gamma\lambda\|\mathbf{\Pi}_{\pi_{\boldsymbol{\theta}'}} - \mathbf{\Pi}_{\pi_{\boldsymbol{\theta}}}\|_{1,1}\right)}\mathbb{E}_{s\sim d_{\pi_{\boldsymbol{\theta}'}}^{\lambda}(\cdot)}\Big[2D_{\mathrm{TV}}(\pi_{\boldsymbol{\theta}'},\pi_{\boldsymbol{\theta}})[s]\Big],$$

where

$$2D_{\mathrm{TV}}(\pi_{\boldsymbol{\theta}'},\pi_{\boldsymbol{\theta}})[s] =: \sum_{a\in\mathcal{A}}\left|\pi_{\boldsymbol{\theta}'}(a|s) - \pi_{\boldsymbol{\theta}}(a|s)\right|.$$

Furthermore, we achieve the boundedness of $\|\mathbf{d}_{\pi_{\boldsymbol{\theta}'}}^{\lambda} - \mathbf{d}_{\pi_{\boldsymbol{\theta}}}^{\lambda}\|_1$ as follows,

$$\|\mathbf{d}_{\pi_{\boldsymbol{\theta}'}}^{\lambda} - \mathbf{d}_{\pi_{\boldsymbol{\theta}}}^{\lambda}\|_1 \leq \frac{1}{1-\tilde{\gamma}}\cdot\frac{1-\gamma\lambda}{|1-2\gamma\lambda|\mathcal{S}||\mathcal{A}||}\mathbb{E}_{s\sim d_{\pi_{\boldsymbol{\theta}}}^{\lambda}(\cdot)}\Big[2D_{\mathrm{TV}}(\pi_{\boldsymbol{\theta}'},\pi_{\boldsymbol{\theta}})[s]\Big],$$

$$\|\mathbf{d}_{\pi_{\boldsymbol{\theta}'}}^{\lambda} - \mathbf{d}_{\pi_{\boldsymbol{\theta}}}^{\lambda}\|_1 \leq \frac{1}{1-\tilde{\gamma}}\cdot\frac{1-\gamma\lambda}{|1-2\gamma\lambda|\mathcal{S}||\mathcal{A}||}\mathbb{E}_{s\sim d_{\pi_{\boldsymbol{\theta}'}}^{\lambda}(\cdot)}\Big[2D_{\mathrm{TV}}(\pi_{\boldsymbol{\theta}'},\pi_{\boldsymbol{\theta}})[s]\Big].$$

Proof. Recall Eq.(59), let

$$\mathbf{G}_{\pi_{\boldsymbol{\theta}}} = \left(\mathbf{I} - \tilde{\gamma}\mathbf{P}_{\pi_{\boldsymbol{\theta}}}^{(\lambda)}\right)^{-1},\ \mathbf{G}_{\pi_{\boldsymbol{\theta}'}} = \left(\mathbf{I} - \tilde{\gamma}\mathbf{P}_{\pi_{\boldsymbol{\theta}'}}^{(\lambda)}\right)^{-1},\ \mathbf{D} = \mathbf{P}_{\pi_{\boldsymbol{\theta}'}}^{(\lambda)} - \mathbf{P}_{\pi_{\boldsymbol{\theta}}}^{(\lambda)}. \tag{106}$$

Then, the following holds

$$\mathbf{G}_{\pi_{\boldsymbol{\theta}}}^{-1} - \mathbf{G}_{\pi_{\boldsymbol{\theta}'}}^{-1} = \left(\mathbf{I} - \tilde{\gamma}\mathbf{P}_{\pi_{\boldsymbol{\theta}}}^{(\lambda)}\right) - \left(\mathbf{I} - \tilde{\gamma}\mathbf{P}_{\pi_{\boldsymbol{\theta}'}}^{(\lambda)}\right) = \tilde{\gamma}\mathbf{D}. \tag{107}$$

Furthermore, by left-multiplying by $\mathbf{G}_{\pi_{\boldsymbol{\theta}}}$ and right-multiplying by $\mathbf{G}_{\pi_{\boldsymbol{\theta}'}}$, we achieve

$$\mathbf{G}_{\pi_{\boldsymbol{\theta}'}} - \mathbf{G}_{\pi_{\boldsymbol{\theta}}} = \tilde{\gamma}\mathbf{G}_{\pi_{\boldsymbol{\theta}'}}\mathbf{D}\mathbf{G}_{\pi_{\boldsymbol{\theta}}}. \tag{108}$$

Grouping all the results from (106)-(108), recall (59),

$$\mathbf{d}_{\pi_{\boldsymbol{\theta}}}^{\lambda} = (1-\tilde{\gamma})\sum_{t=0}^{\infty}\left(\gamma \mathbf{P}_{\pi_{\boldsymbol{\theta}}}^{(\lambda)}\right)^{t}\boldsymbol{\rho}_0 = (1-\tilde{\gamma})\left(\mathbf{I}-\tilde{\gamma}\mathbf{P}_{\pi_{\boldsymbol{\theta}}}^{(\lambda)}\right)^{-1}\boldsymbol{\rho}_0 = (1-\tilde{\gamma})\mathbf{G}_{\pi_{\boldsymbol{\theta}}}\boldsymbol{\rho}_0, \tag{109}$$

then we have

$$\begin{aligned}\mathbf{d}_{\pi_{\boldsymbol{\theta}'}}^{\lambda} - \mathbf{d}_{\pi_{\boldsymbol{\theta}}}^{\lambda} =& (1-\tilde{\gamma})\left(\mathbf{G}_{\pi_{\boldsymbol{\theta}'}} - \mathbf{G}_{\pi_{\boldsymbol{\theta}}}\right)\boldsymbol{\rho}_0 \\ \overset{(108)}{=}& (1-\tilde{\gamma})\tilde{\gamma}\mathbf{G}_{\pi_{\boldsymbol{\theta}'}}\mathbf{D}\mathbf{G}_{\pi_{\boldsymbol{\theta}}}\boldsymbol{\rho}_0 \\ \overset{(109)}{=}& \tilde{\gamma}\mathbf{G}_{\pi_{\boldsymbol{\theta}'}}\mathbf{D}\mathbf{d}_{\pi_{\boldsymbol{\theta}}}^{\lambda}.\end{aligned} \tag{110}$$

Applying (110), we have

$$\|\mathbf{d}_{\pi_{\boldsymbol{\theta}'}}^{\lambda} - \mathbf{d}_{\pi_{\boldsymbol{\theta}}}^{\lambda}\|_1 \overset{(110)}{\leq} \tilde{\gamma}\|\mathbf{G}_{\pi_{\boldsymbol{\theta}'}}\|_1\|\mathbf{D}\mathbf{d}_{\pi_{\boldsymbol{\theta}}}^{\lambda}\|_1. \tag{111}$$

Firstly, we bound the term $\|\mathbf{G}_{\pi_{\boldsymbol{\theta}'}}\|_1$ as follows,

$$\|\mathbf{G}_{\pi_{\boldsymbol{\theta}'}}\|_1 = \left\|\left(\mathbf{I}-\tilde{\gamma}\mathbf{P}_{\pi_{\boldsymbol{\theta}'}}^{(\lambda)}\right)^{-1}\right\|_1 \leq \sum_{t=0}^{\infty}\tilde{\gamma}^t\left\|\mathbf{P}_{\pi_{\boldsymbol{\theta}'}}^{(\lambda)}\right\|_1 = \frac{1}{1-\tilde{\gamma}} = \frac{1-\gamma\lambda}{1-\gamma}. \tag{112}$$

Now, we analyze the second term as follows,

$$\begin{aligned}&\|\mathbf{D}\mathbf{d}_{\pi_{\boldsymbol{\theta}}}^{\lambda}\|_1 \\ =& \sum_{s'\in\mathcal{S}}\left|\sum_{s\in\mathcal{S}}\mathbf{D}(s'|s)d_{\pi_{\boldsymbol{\theta}}}^{\lambda}(s)\right| \\ \overset{(53)}{=}& \sum_{s'\in\mathcal{S}}\left|\sum_{s\in\mathcal{S}}\left(\mathbb{P}_{\pi_{\boldsymbol{\theta}'}}^{(\lambda)}(s'|s) - \mathbb{P}_{\pi_{\boldsymbol{\theta}}}^{(\lambda)}(s'|s)\right)\right| d_{\pi_{\boldsymbol{\theta}}}^{\lambda}(s) \\ \overset{(53)}{=}& \sum_{s'\in\mathcal{S}}\left|(1-\gamma\lambda)\sum_{t=0}^{\infty}(\gamma\lambda)^t\sum_{s\in\mathcal{S}}\left(\mathbb{P}_{\pi_{\boldsymbol{\theta}'}}(s_{t+1}=s'|s) - \mathbb{P}_{\pi_{\boldsymbol{\theta}}}(s_{t+1}=s'|s)\right)\right| d_{\pi_{\boldsymbol{\theta}}}^{\lambda}(s) \\ \leq& \sum_{s\in\mathcal{S}}\left((1-\gamma\lambda)\sum_{t=0}^{\infty}(\gamma\lambda)^t\sum_{s'\in\mathcal{S}}\left|\mathbb{P}_{\pi_{\boldsymbol{\theta}'}}(s_{t+1}=s'|s) - \mathbb{P}_{\pi_{\boldsymbol{\theta}}}(s_{t+1}=s'|s)\right|\right) d_{\pi_{\boldsymbol{\theta}}}^{\lambda}(s)\end{aligned} \tag{113}$$

.

Before we provide a further analyze (113), we need to bound $|\mathbb{P}_{\pi_{\boldsymbol{\theta}'}}(s_{t+1}=s'|s) - \mathbb{P}_{\pi_{\boldsymbol{\theta}}}(s_{t+1}=s'|s)|$. Let $s_0 = s$, then

$$\begin{aligned}\mathbb{P}_{\pi_{\boldsymbol{\theta}}}(s_{t+1}=s'|s) \overset{42}{=}& \sum_{s_1\in\mathcal{S}}\mathbb{P}_{\pi_{\boldsymbol{\theta}}}(s_{t+1}=s'|s_1)\mathbb{P}_{\pi_{\boldsymbol{\theta}}}(s_1|s_0) \\ =& \sum_{s_1\in\mathcal{S}}\sum_{s_2\in\mathcal{S}}\mathbb{P}_{\pi_{\boldsymbol{\theta}}}(s_{t+1}=s'|s_2)\mathbb{P}_{\pi_{\boldsymbol{\theta}}}(s_2|s_1)\mathbb{P}_{\pi_{\boldsymbol{\theta}}}(s_1|s_0) \\ =& \cdots \\ =& \sum_{s_1\in\mathcal{S}}\sum_{s_2\in\mathcal{S}}\cdots\sum_{s_t\in\mathcal{S}}\left(\prod_{i=1}^{t+1}\mathbb{P}_{\pi_{\boldsymbol{\theta}}}(s_i|s_{i-1})\right) \\ =& \sum_{s_1\in\mathcal{S}}\sum_{s_2\in\mathcal{S}}\cdots\sum_{s_t\in\mathcal{S}}\left(\prod_{i=1}^{t+1}\left(\sum_{a_i\in\mathcal{A}}\mathbb{P}(s_i|s_{i-1},a_i)\pi_{\boldsymbol{\theta}}(a_i|s_{i-1})\right)\right).\end{aligned} \tag{114}$$

Similarly, we have

$$\mathbb{P}_{\pi_{\boldsymbol{\theta}'}}(s_{t+1}=s'|s) = \sum_{s_1\in\mathcal{S}}\sum_{s_2\in\mathcal{S}}\cdots\sum_{s_t\in\mathcal{S}}\left(\prod_{i=1}^{t+1}\left(\sum_{a_i\in\mathcal{A}}\mathbb{P}(s_i|s_{i-1},a_i)\pi_{\boldsymbol{\theta}'}(a_i|s_{i-1})\right)\right). \tag{115}$$

Then, according to the results (114)-(115), let $s_0 = s$, the following holds

$$
\begin{aligned}
&\sum_{s^{'}\in\mathcal{S}} |\mathbb{P}_{\pi_{\boldsymbol{\theta}'}}(s_{t+1}=s^{'}|s) - \mathbb{P}_{\pi_{\boldsymbol{\theta}}}(s_{t+1}=s^{'}|s)| \\
=&\sum_{s^{'}\in\mathcal{S}} \left| \sum_{s_1\in\mathcal{S}}\sum_{s_2\in\mathcal{S}}\cdots\sum_{s_t\in\mathcal{S}} \left(\prod_{i=1}^{t+1} \left(\sum_{a_i\in\mathcal{A}} \mathbb{P}(s_i|s_{i-1},a_i)\left(\pi_{\boldsymbol{\theta}'}(a_i|s_{i-1}) - \pi_{\boldsymbol{\theta}}(a_i|s_{i-1})\right)\right)\right)\right| \\
\leq&\sum_{s_1\in\mathcal{S}}\sum_{s_2\in\mathcal{S}}\cdots\sum_{s_t\in\mathcal{S}} \left(\prod_{i=1}^{t+1} \sum_{a_i\in\mathcal{A}} |\pi_{\boldsymbol{\theta}'}(a_i|s_{i-1}) - \pi_{\boldsymbol{\theta}}(a_i|s_{i-1})|\right) \\
=&\sum_{s_1\in\mathcal{S}}\sum_{s_2\in\mathcal{S}}\cdots\sum_{s_t\in\mathcal{S}} \left(\prod_{i=2}^{t+1} \sum_{a_i\in\mathcal{A}} |\pi_{\boldsymbol{\theta}'}(a_i|s_{i-1}) - \pi_{\boldsymbol{\theta}}(a_i|s_{i-1})|\right) \cdot \left(\sum_{a_1\in\mathcal{A}} |\pi_{\boldsymbol{\theta}'}(a_1|s_0) - \pi_{\boldsymbol{\theta}}(a_1|s_0)|\right) \\
=&\prod_{i=2}^{t+1} \left(\sum_{s_{i-1}\in\mathcal{S}}\sum_{a_i\in\mathcal{A}} |\pi_{\boldsymbol{\theta}'}(a_i|s_{i-1}) - \pi_{\boldsymbol{\theta}}(a_i|s_{i-1})|\right) \cdot \left(\sum_{a_1\in\mathcal{A}} |\pi_{\boldsymbol{\theta}'}(a_1|s_0) - \pi_{\boldsymbol{\theta}}(a_1|s_0)|\right) \\
=&\left(\underbrace{\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} |\pi_{\boldsymbol{\theta}'}(a|s) - \pi_{\boldsymbol{\theta}}(a|s)|}_{=:\left\|\boldsymbol{\Pi}_{\pi_{\boldsymbol{\theta}'}} - \boldsymbol{\Pi}_{\pi_{\boldsymbol{\theta}}}\right\|_{1,1}} \right)^t \cdot \left(\sum_{a\in\mathcal{A}} |\pi_{\boldsymbol{\theta}'}(a|s) - \pi_{\boldsymbol{\theta}}(a|s)|\right).
\end{aligned} \tag{116}
$$

Taking the result (116) to (113), we have

$$
\begin{aligned}
\|\mathbf{D}\mathbf{d}^{\lambda}_{\pi_{\boldsymbol{\theta}}}\|_1 \leq& (1-\gamma\lambda)\sum_{t=0}^{\infty}(\gamma\lambda)^t \left\|\boldsymbol{\Pi}_{\pi_{\boldsymbol{\theta}'}} - \boldsymbol{\Pi}_{\pi_{\boldsymbol{\theta}}}\right\|_{1,1}^{t} \sum_{s\in\mathcal{S}} \underbrace{\sum_{a\in\mathcal{A}} |\pi_{\boldsymbol{\theta}'}(a|s) - \pi_{\boldsymbol{\theta}}(a|s)|}_{=:2D_{\mathrm{TV}}(\pi_{\boldsymbol{\theta}'},\pi_{\boldsymbol{\theta}})[s]} d^{\lambda}_{\pi_{\boldsymbol{\theta}}}(s) \\
=&(1-\gamma\lambda)\sum_{t=0}^{\infty}(\gamma\lambda)^t \|\boldsymbol{\Pi}_{\pi_{\boldsymbol{\theta}'}} - \boldsymbol{\Pi}_{\pi_{\boldsymbol{\theta}}}\|_{1,1}^{t} \mathbb{E}_{s\sim d^{\lambda}_{\pi_{\boldsymbol{\theta}}}(\cdot)}\Big[2D_{\mathrm{TV}}(\pi_{\boldsymbol{\theta}'},\pi_{\boldsymbol{\theta}})[s]\Big] \\
=&\frac{1-\gamma\lambda}{1-\gamma\lambda\|\boldsymbol{\Pi}_{\pi_{\boldsymbol{\theta}'}} - \boldsymbol{\Pi}_{\pi_{\boldsymbol{\theta}}}\|_{1,1}} \mathbb{E}_{s\sim d^{\lambda}_{\pi_{\boldsymbol{\theta}}}(\cdot)}\Big[2D_{\mathrm{TV}}(\pi_{\boldsymbol{\theta}'},\pi_{\boldsymbol{\theta}})[s]\Big].
\end{aligned} \tag{117}
$$

Finally, according to (111), (112) and (117), we have

$$
\|\mathbf{d}^{\lambda}_{\pi_{\boldsymbol{\theta}'}} - \mathbf{d}^{\lambda}_{\pi_{\boldsymbol{\theta}}}\|_1 \leq \frac{\tilde{\gamma}}{1-\tilde{\gamma}} \cdot \frac{\gamma(1-\lambda)}{1-\gamma\lambda\|\boldsymbol{\Pi}_{\pi_{\boldsymbol{\theta}'}} - \boldsymbol{\Pi}_{\pi_{\boldsymbol{\theta}}}\|_{1,1}} \mathbb{E}_{s\sim d^{\lambda}_{\pi_{\boldsymbol{\theta}}}(\cdot)}\Big[2D_{\mathrm{TV}}(\pi_{\boldsymbol{\theta}'},\pi_{\boldsymbol{\theta}})[s]\Big]. \tag{118}
$$

□

Recall (105), we have

$$
\|\boldsymbol{\Pi}_{\pi_{\boldsymbol{\theta}'}} - \boldsymbol{\Pi}_{\pi_{\boldsymbol{\theta}}}\|_{1,1} = \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} |\pi_{\boldsymbol{\theta}'}(a|s) - \pi_{\boldsymbol{\theta}}(a|s)| \leq 2|\mathcal{S}||\mathcal{A}|. \tag{119}
$$

Then, we achieve the boundedness of $\|\boldsymbol{\Pi}_{\pi_{\boldsymbol{\theta}'}} - \boldsymbol{\Pi}_{\pi_{\boldsymbol{\theta}}}\|_{1,1}$ as follows,

$$
\|\mathbf{d}^{\lambda}_{\pi_{\boldsymbol{\theta}'}} - \mathbf{d}^{\lambda}_{\pi_{\boldsymbol{\theta}}}\|_1 \leq \frac{\tilde{\gamma}}{1-\tilde{\gamma}} \cdot \frac{1-\gamma\lambda}{|1-2\gamma\lambda|\mathcal{S}||\mathcal{A}||} \mathbb{E}_{s\sim d^{\lambda}_{\pi_{\boldsymbol{\theta}}}(\cdot)}\Big[2D_{\mathrm{TV}}(\pi_{\boldsymbol{\theta}'},\pi_{\boldsymbol{\theta}})[s]\Big]. \tag{120}
$$

F PROOF OF THEOREM 2

Theorem 2 *Let* $\delta_k = \mathbb{E}_{s\sim d^{\lambda}_{\pi_{\boldsymbol{\theta}_k}}(\cdot)}\left[\mathrm{KL}\left(\pi_{\boldsymbol{\theta}_k}, \pi_{\boldsymbol{\theta}_{k+\frac{1}{2}}}\right)[s]\right]$*, if* $\pi_{\boldsymbol{\theta}_k}$ *and* $\pi_{\boldsymbol{\theta}_{k+1}}$ *are related to (17)-(18), then the lower bound on policy improvement, and upper bound on constraint violation are*

$$J(\pi_{\boldsymbol{\theta}_{k+1}}) - J(\pi_{\boldsymbol{\theta}_k}) \geq -\frac{\gamma(1-\lambda)\alpha_k\sqrt{2\delta_k}\epsilon^{V}_{\pi_{\boldsymbol{\theta}}}(\pi_{\boldsymbol{\theta}'})}{(1-\gamma)\left|1-2\gamma\lambda|\mathcal{S}||\mathcal{A}||\right.}, J^c(\pi_{\boldsymbol{\theta}_{k+1}}) \leq b + \frac{\gamma(1-\lambda)\beta_k\sqrt{2\delta_k}\epsilon^{C}_{\pi_{\boldsymbol{\theta}}}(\pi_{\boldsymbol{\theta}'})}{(1-\gamma)\left|1-2\gamma\lambda|\mathcal{S}||\mathcal{A}||\right.}.$$

Proof. (of Theorem 2)

According to Bregman divergence, if policy $\pi_{\boldsymbol{\theta}_k}$ is feasible, policy $\pi_{\boldsymbol{\theta}_{k+1}}$ is generated according to (18), then the following

$$\mathrm{KL}\left(\pi_{\boldsymbol{\theta}_k}, \pi_{\boldsymbol{\theta}_{k+\frac{1}{2}}}\right) \geq \mathrm{KL}\left(\pi_{\boldsymbol{\theta}_k}, \pi_{\boldsymbol{\theta}_{k+1}}\right) + \mathrm{KL}\left(\pi_{\boldsymbol{\theta}_{k+1}}, \pi_{\boldsymbol{\theta}_{k+\frac{1}{2}}}\right)$$

implies

$$\delta_k = \mathbb{E}_{s\sim d^{\lambda}_{\pi_{\boldsymbol{\theta}_k}}(\cdot)}\left[\mathrm{KL}\left(\pi_{\boldsymbol{\theta}_k}, \pi_{\boldsymbol{\theta}_{k+\frac{1}{2}}}\right)[s]\right] \geq \mathbb{E}_{s\sim d^{\lambda}_{\pi_{\boldsymbol{\theta}_k}}(\cdot)}\left[\mathrm{KL}\left(\pi_{\boldsymbol{\theta}_{k+1}}, \pi_{\boldsymbol{\theta}_k}\right)[s]\right].$$

According to the asymptotically symmetry of KL divergence if we update the policy within a local region, then, we have

$$\delta_k \geq \mathbb{E}_{s\sim d^{\lambda}_{\pi_{\boldsymbol{\theta}_k}}(\cdot)}\left[\mathrm{KL}\left(\pi_{\boldsymbol{\theta}_{k+\frac{1}{2}}}, \pi_{\boldsymbol{\theta}_k}\right)[s]\right] \geq \mathbb{E}_{s\sim d^{\lambda}_{\pi_{\boldsymbol{\theta}_k}}(\cdot)}\left[\mathrm{KL}\left(\pi_{\boldsymbol{\theta}_{k+1}}, \pi_{\boldsymbol{\theta}_k}\right)[s]\right].$$

Furthermore, according to Proposition 1 and Proposition 3, we have

$$\begin{aligned}
&J(\pi_{\boldsymbol{\theta}_{k+1}}) - J(\pi_{\boldsymbol{\theta}_k}) \\
\geq&\frac{1}{1-\tilde{\gamma}}\mathbb{E}_{s\sim d^{\lambda}_{\pi_{\boldsymbol{\theta}_k}}(\cdot), a\sim\pi_{\boldsymbol{\theta}_{k+1}}(\cdot|s)}\left[A^{\mathrm{GAE}(\gamma,\lambda)}_{\pi_{\boldsymbol{\theta}_k}}(s,a) - \frac{2\gamma(1-\lambda)\epsilon^{V}_{\pi_{\boldsymbol{\theta}_{k+1}}}(\pi_{\boldsymbol{\theta}_k})}{(1-\gamma\lambda)\left|1-2\gamma\lambda|\mathcal{S}||\mathcal{A}||\right.}D_{\mathrm{TV}}(\pi_{\boldsymbol{\theta}_k}, \pi_{\boldsymbol{\theta}_{k+1}})[s]\right] \\
\geq&\frac{1}{1-\tilde{\gamma}}\mathbb{E}_{s\sim d^{\lambda}_{\pi_{\boldsymbol{\theta}_k}}(\cdot), a\sim\pi_{\boldsymbol{\theta}_{k+1}}(\cdot|s)}\left[-\frac{2\gamma(1-\lambda)\alpha_k\epsilon^{V}_{\pi_{\boldsymbol{\theta}_{k+1}}}(\pi_{\boldsymbol{\theta}_k})}{(1-\gamma\lambda)\left|1-2\gamma\lambda|\mathcal{S}||\mathcal{A}||\right.}\sqrt{\frac{1}{2}\mathbb{E}_{s\sim d^{\lambda}_{\pi_{\boldsymbol{\theta}_k}}(\cdot)}\left[\mathrm{KL}(\pi_{\boldsymbol{\theta}_k}, \pi_{\boldsymbol{\theta}_{k+1}})[s]\right]}\right] \\
\geq&\frac{1}{1-\tilde{\gamma}}\mathbb{E}_{s\sim d^{\lambda}_{\pi_{\boldsymbol{\theta}_k}}(\cdot), a\sim\pi_{\boldsymbol{\theta}_{k+1}}(\cdot|s)}\left[-\frac{\gamma(1-\lambda)\alpha_k\sqrt{2\delta_k}\epsilon^{V}_{\pi_{\boldsymbol{\theta}_{k+1}}}(\pi_{\boldsymbol{\theta}_k})}{(1-\gamma\lambda)\left|1-2\gamma\lambda|\mathcal{S}||\mathcal{A}||\right.}\right].
\end{aligned}$$

Similarly, according to Proposition 1 and Proposition 2, and since policy $\pi_{\boldsymbol{\theta}_{k+1}}$ satisfies

$$J^c(\pi_{\boldsymbol{\theta}_k}) + \frac{1}{1-\tilde{\gamma}}\mathbb{E}_{s\sim d^{\lambda}_{\pi_{\boldsymbol{\theta}_k}}(\cdot), a\sim\pi_{\boldsymbol{\theta}_{k+1}}(\cdot|s)}\left[A^{\mathrm{GAE}(\gamma,\lambda)}_{\pi_{\boldsymbol{\theta}_k},C}(s,a)\right] + \beta_k\sqrt{\mathbb{E}_{s\sim d^{\lambda}_{\pi_{\boldsymbol{\theta}_k}}(\cdot)}\left[\mathrm{KL}(\pi_{\boldsymbol{\theta}_k}, \pi_{\boldsymbol{\theta}_{k+1}})[s]\right]} \leq b, \tag{121}$$

and

$$\begin{aligned}
&J^c(\pi_{\boldsymbol{\theta}_{k+1}}) - J^c(\pi_{\boldsymbol{\theta}_k}) \qquad (122)\\
\leq&\frac{1}{1-\tilde{\gamma}}\mathbb{E}_{s\sim d^{\lambda}_{\pi_{\boldsymbol{\theta}_k}}(\cdot), a\sim\pi_{\boldsymbol{\theta}_{k+1}}(\cdot|s)}\left[A^{\mathrm{GAE}(\gamma,\lambda)}_{\pi_{\boldsymbol{\theta}_k},C}(s,a) + \frac{2\gamma(1-\lambda)\beta_k\epsilon^{C}_{\pi_{\boldsymbol{\theta}_{k+1}}}(\pi_{\boldsymbol{\theta}_k})}{(1-\gamma\lambda)\left|1-2\gamma\lambda|\mathcal{S}||\mathcal{A}||\right.}D_{\mathrm{TV}}(\pi_{\boldsymbol{\theta}_k}, \pi_{\boldsymbol{\theta}_{k+1}})[s]\right].
\end{aligned}$$

Combining (121)- (123), we have

$$\begin{aligned}
&J^c(\pi_{\boldsymbol{\theta}_{k+1}}) - J^c(\pi_{\boldsymbol{\theta}_k}) \qquad (123)\\
\leq&b + \frac{1}{1-\tilde{\gamma}}\mathbb{E}_{s\sim d^{\lambda}_{\pi_{\boldsymbol{\theta}_k}}(\cdot), a\sim\pi_{\boldsymbol{\theta}_{k+1}}(\cdot|s)}\left[\frac{2\gamma(1-\lambda)\beta_k\epsilon^{C}_{\pi_{\boldsymbol{\theta}_{k+1}}}(\pi_{\boldsymbol{\theta}_k})}{(1-\gamma\lambda)\left|1-2\gamma\lambda|\mathcal{S}||\mathcal{A}||\right.}\sqrt{\frac{1}{2}\mathbb{E}_{s\sim d^{\lambda}_{\pi_{\boldsymbol{\theta}_k}}(\cdot)}\left[\mathrm{KL}(\pi_{\boldsymbol{\theta}_k}, \pi_{\boldsymbol{\theta}_{k+1}})[s]\right]}\right] \\
\leq&b + \frac{1}{1-\tilde{\gamma}}\mathbb{E}_{s\sim d^{\lambda}_{\pi_{\boldsymbol{\theta}_k}}(\cdot), a\sim\pi_{\boldsymbol{\theta}_{k+1}}(\cdot|s)}\left[\frac{\gamma(1-\lambda)\beta_k\sqrt{2\delta_k}\epsilon^{C}_{\pi_{\boldsymbol{\theta}_{k+1}}}(\pi_{\boldsymbol{\theta}_k})}{(1-\gamma\lambda)\left|1-2\gamma\lambda|\mathcal{S}||\mathcal{A}||\right.}\right]. \qquad (124)
\end{aligned}$$

□

G EXPERIMENTS

The Python code for our implementation of CUP is provided along with this submission in the supplementary material.

G.1 ENVIRONMENT

G.1.1 ENVIRONMENT 1: ROBOTS WITH SPEED LIMIT.

We consider two tasks from MuJoCo (Brockman et al., 2016): Walker2d-v3 and Hopper-v3, where the setting of cost follows (Zhang et al., 2020). For agents move on a two-dimensional plane, the cost is calculated as follows,

$$C(s,a) = \sqrt{v_x^2 + v_y^2};$$

for agents move along a straight line, the cost is calculated as

$$C(s,a) = |v_x|,$$

where v_x, v_y are the velocities of the agent in the x and y directions respectively.

G.1.2 ENVIRONMENT 2: CIRCLE.

The Circle Environment follows (Achiam et al., 2017), and we use open-source implementation of the circle environments from `https://github.com/ymzhang01/mujoco-circle`. According to Zhang et al. (2020), those experiments were implemented in OpenAI Gym (Brockman et al., 2016) while the circle tasks in Achiam et al. (2017) were implemented in rllab (Duan et al., 2016). We also excluded the Point agent from the original experiments since it is not a valid agent in OpenAI Gym. The first two dimensions in the state space are the (x, y) coordinates of the center mass of the agent, hence the state space for both agents has two extra dimensions compared to the standard Ant-v0 and Humanoid-v0 environments from OpenAI Gym.

Now, we present some necessary details of this environment taken from (Zhang et al., 2020).

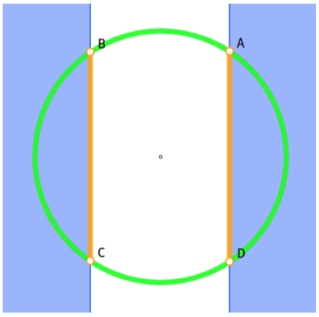


Figure 2: In the Circle task, reward is maximized by moving along the green circle. The agent is not allowed to enter the blue regions, so its optimal constrained path follows the line segments AD and BC (figure and caption taken from (Achiam et al., 2017; Zhang et al., 2020)).

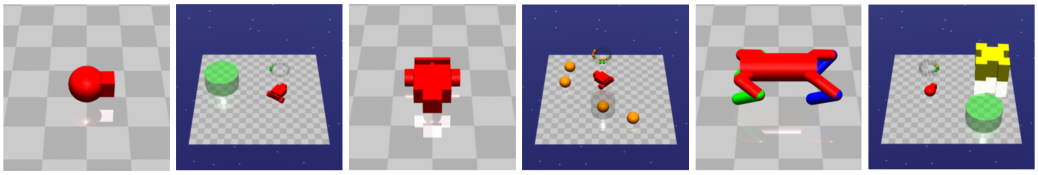
In the circle tasks, the goal is for an agent to move along the circumference of a circle while remaining within a safety region smaller than the radius of the circle. The exact geometry of the task is shown in Figure 2. The reward and cost functions are defined as:

$$R(s) = \frac{-yv_x + xv_y}{1 + |\sqrt{x^2 + y^2} - r|}, \quad C(s) = \mathbb{I}(|x| > x_{\lim}),$$

where x, y are the positions of the agent on the plane, v_x, v_y are the velocities of the agent along the x and y directions, r is the radius of the circle, and $x_{\lim}$ specifies the range of the safety region. The radius is set to $r = 10$ for both Ant and Humanoid while $x_{\lim}$ is set to 3 and 2.5 for Ant and Humanoid respectively. Note that these settings are identical to those of the circle task in Achiam et al. (2017); Zhang et al. (2020).

G.1.3 ENVIRONMENT 3: SAFETY GYM SHIPS WITH THREE PRE-MADE ROBOTS.

In Safety Gym environments, the agent perceives the world through a robot's sensors and interacts with the world through its actuators. In our paper, we consider three environment: Point, Car, Dog from (Ray et al., 2019). In this section, the presentation of those environments are taken from (Ray et al., 2019), for more details, please refer to (Ray et al., 2019, Page 8–10).



(a) Point (b) Goal (c) Car (d) Button (e) Doggo (f) Push

Figure 3: Fig (a), (c), (e) show the pre-made robots in Safety Gym. These robots are used in the benchmark environments. Fig (b), (d), (f) show the tasks for our environments. From left to right: Goal, Button, Push. In "Goal," the objective is to move the robot inside the green goal area. In "Button," the objective is to press the highlighted button (visually indicated with a faint gray cylinder). In "Push," the objective is to push the yellow box inside of the green goal area (figure and caption taken from (Ray et al., 2019)).

Point: (Figure 3 (a)). A simple robot constrained to the 2D-plane, with one actuator for turning and another for moving forward/backwards. This factored control scheme makes the robot particularly easy to control for navigation. Point has a small square in front that makes it both easier to visually determine the robot's direction, and helps the point push a box element that appears in one of our tasks.

Car: (Figure 3 (c)). Car is a slightly more complex robot that has two independently-driven parallel wheels and a free rolling rear wheel. Car is not fixed to the 2D-plane, but mostly resides in it. For this robot, both turning and moving forward/backward require coordinating both of the actuators. It is similar in design to simple robots used in education.

Doggo: (Figure 3 (e)). Doggo is a quadrupedal robot with bilateral symmetry. Each of the four legs has two controls at the hip, for azimuth and elevation relative to the torso, and one in the knee, controlling angle. It is designed such that a uniform random policy should keep the robot from falling over and generate some travel.

All actions for all robots are continuous, and linearly scaled to $[-1, +1]$, which is common for 3D robot-based RL environments and (anecdotally) improves learning with neural nets. Modulo scaling, the action parameterization is based on a mix of hand-tuning and MuJoCo actuator defaults, and we caution that it is not clear if these choices are optimal. Some safe exploration techniques are action-layer interventions, like projecting to the closest predicted safe action (Dalal et al., 2018), and these methods can be sensitive to action parameterization. As a result, action parameterization may merit more careful consideration than is usually given. Future work on action space design might be to find action parameterizations that respect physical measures we care about—for example, an action space where a fixed distance corresponds to a fixed amount of energy.

The Safety Gym environment-builder currently supports three main tasks: Goal, Button, and Push (depicted in Fig. 2). Tasks in Safety Gym are mutually exclusive, and an individual environment can only make use of a single task. Reward functions are configurable, allowing rewards to be either sparse (rewards only obtained on task completion) or dense (rewards have helpful, hand-crafted shaping terms). Task details follow:

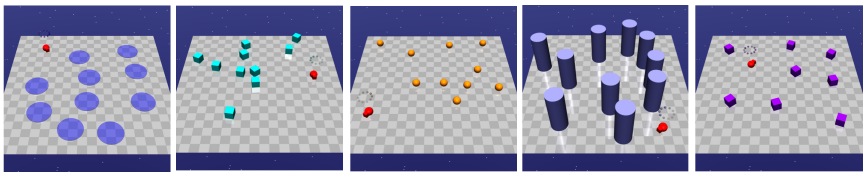
Goal: (Figure 3 (b)). Move the robot to a series of goal positions. When a goal is achieved, the goal location is randomly reset to someplace new, while keeping the rest of the layout the same. The sparse reward component is attained on achieving a goal position (robot enters the goal circle). The dense reward component gives a bonus for moving towards the goal.

Button: (Figure 3 (d)). Press a series of goal buttons. Several immobile "buttons" are scattered throughout the environment, and the agent should navigate to and press (contact) the currently-

highlighted button, which is the goal button. After the agent presses the correct button, the environment will select and highlight a new goal button, keeping everything else fixed. The sparse reward component is attained on pressing the current goal button. The dense reward component gives a bonus for moving towards the current goal button.

Push: (Figure 3 (f)). Move a box to a series of goal positions. Like the goal task, a new goal location is drawn each time a goal is achieved. The sparse reward component is attained when the yellow box enters the goal circle. The dense reward component consists of two parts: one for getting the agent closer to the box, and another for getting the box closer to the goal.

Constraint Options and Desiderata



(a) Hazards (b) Vases (c) Buttons (d) Pillars (e) Gremlins

Figure 4: Constraint elements used in our environments (figure and caption taken from (Ray et al., 2019)).

The constraint elements themselves are:

Hazards: (Figure 4 (a)). Dangerous areas to avoid. These are circles on the ground that are non-physical, and the agent is penalized for entering them.

Vases: (Figure 4 (b)). Objects to avoid. These are small blocks that represent fragile objects. The agent is penalized for touching or moving them.

Pillars: (Figure 4 (c)). Immobile obstacles. These are rigid barriers in the environment, which the agent should not touch.

Buttons: (Figure 4 (d)). Incorrect goals. When using the "buttons" goal, pressing an incorrect button is penalized.

Gremlins: (Figure 4 (e)). Moving objects. These are simple moving objects that the agent must avoid contacting. Since they are moving quickly, the agent must stay out of the way of their path of travel.

Although all constraint elements represent things for the agent to avoid, they pose different challenges for the agent by virtue of having different dynamics. To illustrate the contrast: hazards provide no physical obstacle, vases are moveable obstacles, pillars are immovable obstacles, buttons can sometimes be perceived as goals, and gremlins are actively-moving obstacles. Like reward functions in Safety Gym, cost functions are configurable in various ways; see the code for details. By default, cost functions are simple indicators for whether an unsafe interaction has occured ($c_t = 1$ if the agent has done the unsafe thing, otherwise $c_t = 0$).

Finally, SGPoint, SGCar, and SGDoggo, which are all six Point/Car/Doggo robot environments with constraints in Safety Gym, and Ray et al. (2019) have provided an implementation for those environments.

G.2 Details of Experiments

In all of those experiments, we use a two-layer feedforward neural network with a tanh activation for both policy and value networks. Experiment-specific parameters are as follows:

Parameter	Walker2d	Hopper	HumanoidCircle	AntCircle
No. of hidden layers	2	2	2	2
No. of hidden nodes	64	64	64	64
Batch size	2048	2048	50000	50000
Minibatch size	64	64	1000	1000
Rollout length	1000	1000	1000	1000
GAE parameter (cost)	0.95	0.95	0.995	0.995
GAE parameter (reward)	0.95	0.95	0.995	0.995
discounter for cost	0.99	0.99	0.995	0.995
discounter for reward	0.99	0.99	0.995	0.995
learning rate for policy	3×10^{-4}	3×10^{-4}	3×10^{-4}	3×10^{-4}
learning rate for value and reward function	3×10^{-4}	3×10^{-4}	3×10^{-4}	3×10^{-4}