Efficient Safe Meta-Reinforcement Learning: Provable Near-Optimality and Anytime Safety

Siyuan Xu & Minghui Zhu
School of Electrical Engineering and Computer Science
The Pennsylvania State University
University Park, PA 16801
{spx5032, muz16}@psu.edu

Abstract

This paper studies the problem of safe meta-reinforcement learning (safe meta-RL), where an agent efficiently adapts to unseen tasks while satisfying safety constraints at all times during adaptation. We propose a framework consisting of two complementary modules: safe policy adaptation and safe meta-policy training. The first module introduces a novel one-step safe policy adaptation method that admits a closed-form solution, ensuring monotonic improvement, constraint satisfaction at every step, and high computational efficiency. The second module develops a Hessian-free meta-training algorithm that incorporates safety constraints on the meta-policy and leverages the analytical form of the adapted policy to enable scalable optimization. Together, these modules yield three key advantages over existing safe meta-RL methods: (i) superior optimality, (ii) anytime safety guarantee, and (iii) high computational efficiency. Beyond existing safe meta-RL analyses, we prove the anytime safety guarantee of policy adaptation and provide a lower bound of the expected total reward of the adapted policies compared with the optimal policies, which shows that the adapted policies are nearly optimal. Empirically, our algorithm achieves superior optimality, strict safety compliance, and substantial computational gains—up to 70% faster training and 50% faster testing—across diverse locomotion and navigation benchmarks.

1 Introduction

Reinforcement learning (RL) [47] has achieved significant successes in various domains, from video games [37, 46, 26] to robotics [28, 27, 35, 36]. The RL problem is formulated as a Markov decision process (MDP) and aims to maximize the expected total reward. Safe RL [58, 55, 15, 57] addresses additional safety requirements, such as collision avoidance for robots [51, 21] and operation restrictions in financial management [1]. Typically, the safe RL problem is formulated as a constrained MDP (CMDP) [4], which aims to maximize the expected total reward while ensuring that the expected safety costs are below given thresholds. As noted in [15], the goals of reward maximization and constraint enforcement are not completely aligned, aggravating the challenge of the inherent trade-off between exploration and exploitation.

Meta-reinforcement learning (meta-RL) [5] aims to extract common knowledge from multiple existing RL tasks, accelerating the learning process and increasing the data efficiency of RL algorithms. Safe meta-RL [24, 55, 59] integrates safe RL and meta-RL and inherits the benefits of both. On the other hand, existing safe meta-RL methods face three new challenges: optimality, computational efficiency, and anytime safety. Meta-CRPO [24] considers an online safe meta-RL problem. In each round, it computes the task-specific policy by CRPO [55] and updates the meta-policy that has the minimal average distance to the task-specific policies of all previous tasks. However, the

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

Table 1: Comparison with existing safe meta-RL methods

Methods	Theoretical results			Experimental results	
	Safety		Bounded optimality gap	Efficiency	Optimality
	Constraint violation	Target policy			
[24]	Positive	Safety for final policy	✓	Low	Low
[12]	Positive	Safety for adapted policy	×	Low	Medium
Ours	Zero	Anytime safety	✓	High	High

meta-training does not optimize the performance of the task-specific policy adaptation, and the policies adapted from the learned meta-policy may be sub-optimal for new tasks. Meta-CPO [12] optimizes the policies adapted from the meta-policy by constraint policy optimization (CPO) [2]. Nevertheless, its computational complexity is high in both the meta-training and meta-test stages. Specifically, during the meta-training, meta-CPO solves a constrained bilevel optimization problem [52] where the constraints are present at both the upper and lower levels. The computation involved, particularly the inverse of the Hessian, is computationally expensive. During the meta-test, each policy adaptation step solves a nonconvex constrained optimization problem.

In applications of (safe) meta-RL [38, 6], during the meta-test, the agent collects the rewards/costs of state-action pairs by exploring a new, unknown CMDP and optimizes the policy based on the collected data. Therefore, it is important to guarantee anytime safety, i.e., the safety constraints must be satisfied for every policy used for the exploration. However, the anytime safety is overlooked in all existing safe meta-RL algorithms [24, 12]. Specifically, during the meta-test, they start with the meta-policy and repeatedly adapt the most recent policy into a new one by the policy adaptation algorithm, which generates a sequence of policies. Except for the final policy in the sequence, each policy, including the initial meta-policy, is used to explore the environment and collect data. Meta-CRPO [24] only quantifies the safety constraint violation of the final convergent policy in the sequence, neglecting that of intermediate policies for data collection. Meta-CPO [12] applies the CPO [2] as the policy adaptation algorithm, which can quantify the safety constraint violation of policies that have undergone at least one adaptation step. However, the safety of the meta-policy is ignored. Moreover, both meta-CRPO and meta-CPO provide positive upper bounds of the constraint violation, which do not guarantee zero violation of the safety constraints.

1.1 Main contribution

This paper develops a safe meta-RL framework consisting of two modules: safe policy adaptation and safe meta-policy training. Specifically, we introduce a novel safe policy adaptation method, which guarantees monotonic improvement, ensures safety, and provides a closed-form solution for a single safe policy adaptation step. For the meta-policy training, we impose safety constraints on the meta-policy, derive the meta-gradient, simplify its computation by leveraging the closed-form expression of the adapted policy, and develop a Hessian-free meta-training algorithm.

The proposed algorithms offer three key advantages over existing safe meta-RL methods. (i) **Superior optimality.** Our safe meta-policy training algorithm maximizes the expected accumulated reward of the policies adapted from the meta-policy, and then improves the optimality of meta-CRPO [24] and naive transfers from meta-RL, which do not consider the task-specific safe policy adaptation in the meta-training. (ii) **Anytime safety guarantee** during the meta-test. With the imposed safety constraint on the meta-policy, the safe meta-policy training produces a safe initial meta-policy. Moreover, as mentioned in (b), the safe policy adaptation guarantees safety for each step when the initial policy is safe. By integrating these two modules, anytime safety is achieved. (iii) **High computational efficiency** in both the meta-test and meta-training stages. As mentioned in (c), we derive the close-formed solution for the policy adaptation. It makes the meta-test much more efficient than those in meta-CRPO [24] and meta-CPO [12], which solve constrained optimization problems. In the meta-training, the close-formed solution of the policy adaptation is used to derive a Hessian-free meta-gradient and reduces the computation complexity of the proposed algorithm to approach that in the single-level optimization, making it more efficient than meta-CPO [12] and many meta-RL algorithms [16, 30] with the bi-level optimization steps and the computation of Hessian and Hessian inverse. We conduct experiments on seven scenarios including navigation tasks with collision avoidance and locomotion tasks to verify these advantages of the proposed algorithms.

Another major contribution of the paper is that it is the first to derive a comprehensive theoretical analysis regarding near optimality and anytime safety guarantees for safe meta-RL. First, we establish the theoretical basis of the algorithm design that guarantees anytime safety, i.e., zero constraint violation for any policy during the policy adaptation. Second, we derive a lower bound of the expected accumulated reward of the adapted policies compared to that of the task-specific optimal policies, which shows the near optimality of the proposed safe meta-RL framework. Finally, we demonstrate a trade-off between the optimality bound and constraint violation when the allowable constraint violation varies, which enables the algorithm to be adjusted to prioritize either safety or optimality.

Table 1 compares both the theoretical and experimental results between this paper and previous works [24, 12]. First, we study anytime safety and provide a zero constraint violation guarantee. In previous works, they only provided positive upper bounds for the constraint violation, and the upper bounds only work for the final policy [24] or the adapted policies [12]. Second, although [24] provides an upper bound of the optimality gap, the experimental optimality is the worst. On the other hand, [12] does not provide an optimality bound. In contrast, our method exhibits high optimality and provides a near-optimality guarantee, outperforming existing approaches in terms of both experimental and theoretical outcomes. Third, our method is more efficient than the existing methods [24, 12].

Related works. Due to the space limit, we include a section of related works in Appendix A.

1.2 Notations

Denote the l_2 norm of vectors and the spectral norm (2-norm) of matrices by $\|\cdot\|$. Denote the Kullback–Leibler divergence (KL-divergence) of probability distributions p and q defined on the same sample space $\mathcal{X}$ by $D_{\mathrm{KL}}(p\|q) \triangleq \int_{\mathcal{X}} \ln\left(\frac{p(dx)}{q(dx)}\right) p(dx)$.

2 Problem Statement

CMDP. A CMDP $\mathcal{M} \triangleq \{\mathcal{S}, \mathcal{A}, \gamma, \rho, P, r, \{c_i\}_{i=1}^p, \{d_i\}_{i=1}^p\}$ is defined by the state space $\mathcal{S}$, the action space $\mathcal{A}$, the discount factor γ, the initial state distribution ρ over $\mathcal{S}$, the transition probability $P(s'|s,a) : \mathcal{S}\times\mathcal{A}\times\mathcal{S} \to [0,1]$, the reward function $r : \mathcal{S}\times\mathcal{A}\times\mathcal{S} \to [0, r^{max}]$, p cost functions where the i-th cost function is defined as $c_i : \mathcal{S}\times\mathcal{A}\times\mathcal{S} \to [0, c_i^{max}]$ for $i = 1, \cdots, p$, and the constant d_i, which is the limit of constraint i. The state space $\mathcal{S}$ could be either a discrete space or a bounded continuous space. The action space $\mathcal{A}$ could be either discrete or continuous.

Policy. A stochastic policy $\pi : \mathcal{S} \to \mathbb{P}(\mathcal{A})$ is a mapping from states to probability distributions over action. When $\mathcal{A}$ is discrete, $\pi(a|s)$ denotes the probability of choosing action a in state s; when $\mathcal{A}$ is continuous, $\pi(a|s)$ denotes the probability density. Denote the policy space as Π. In addition, a softmax policy parameterized by $\theta \in \mathbb{R}^n$ is denoted as π_θ, where $\pi_\theta(a|s) \triangleq \frac{\exp(f_\theta(s,a))}{\int_{\mathcal{A}}\exp(f_\theta(s,a'))da'}$, $\forall(s,a) \in \mathcal{S}\times\mathcal{A}$, for continuous action space $\mathcal{A}$, or $\pi_\theta(a|s) \triangleq \frac{\exp(f_\theta(s,a))}{\sum_{a'\in\mathcal{A}}\exp(f_\theta(s,a'))}$, for discrete action space $\mathcal{A}$, and $f_\theta : \mathcal{S}\times\mathcal{A} \to \mathbb{R}$ is a continuous function for any θ.

Safe RL. For a policy π, the value function is defined as $V^\pi(s) \triangleq \mathbb{E}[\sum_{t=0}^\infty \gamma^t\, r(s_t, a_t, s_{t+1})|s_0 = s, \pi]$. The action-value function is defined as $Q^\pi(s,a) \triangleq \mathbb{E}[\sum_{t=0}^\infty \gamma^t r(s_t, a_t,\, s_{t+1})|s_0 = s, a_0 = a, \pi]$. The advantage function is defined as $A^\pi(s,a) \triangleq Q^\pi(s,a) - V^\pi(s)$. The accumulated reward function is $J(\pi) \triangleq \mathbb{E}_{s\sim\rho}[V^\pi(s)]$. Similarly, for each $i = 1, \cdots, p$, we define $V_{c_i}^\pi(s) \triangleq \mathbb{E}[\sum_{t=0}^\infty \gamma^t c_i(s_t, a_t, s_{t+1})|s_0 = s, \pi]$, $Q_{c_i}^\pi(s,a) \triangleq \mathbb{E}[\sum_{t=0}^\infty \gamma^t c_i(s_t, a_t, s_{t+1})|s_0 = s, a_0 = a, \pi]$, $A_{c_i}^\pi(s,a) \triangleq Q_{c_i}^\pi(s,a) - V_{c_i}^\pi(s)$, and $J_{c_i}(\pi) \triangleq \mathbb{E}_{s\sim\rho}\left[V_{c_i}^\pi(s)\right]$. The discounted state visitation distribution of π is defined as $\nu^\pi(s) \triangleq (1-\gamma)\mathbb{E}_{s_0\sim\rho}[\sum_{t=0}^\infty \gamma^t\, \mathbb{P}(s_t = s|\pi)]$. The safe RL problem is to maximize the accumulated reward function while the accumulated cost functions satisfy the constraints, i.e., solving the problem $\max_{\pi\in\Pi} J(\pi)$ s.t. $J_{c_i,\tau}(\pi) \le d_i,\ \forall i = 1, \cdots, p$.

Safe meta-RL with anytime safety. Safe meta-RL targets multiple safe RL tasks. Consider a space of safe RL tasks Γ, where each task $\tau \in \Gamma$ is modeled by a CMDP $\mathcal{M}_\tau \triangleq \{\mathcal{S}, \mathcal{A}, \gamma, \rho_\tau, P_\tau, r_\tau, \{c_{i,\tau}\}_{i=1}^p, \{d_{i,\tau}\}_{i=1}^p\}$. Following the notions in the above subsections, the notations $\rho_\tau, P_\tau, r_\tau, c_{i,\tau}, d_{i,\tau}$, as well as V_τ^π, $V_{c_i,\tau}^\pi$, Q_τ^π, $Q_{c_i,\tau}^\pi$, A_τ^π, $A_{c_i,\tau}^\pi$, J_τ, $J_{c_i,\tau}$, and ν_τ^π are defined for task τ. Consider a set of safe RL tasks in Γ following a probability distribution $\mathbb{P}(\Gamma)$. Safe meta-RL aims to learn the

meta-prior from $\mathbb{P}(\Gamma)$ which can be used to train a policy for an unseen task $\tau_{new} \sim \mathbb{P}(\Gamma)$ by a small number of new environment explorations with anytime safety. In specific, during the meta-training, tasks can be sampled from $\mathbb{P}(\Gamma)$, i.e., $\{\tau_j\}_{j=1}^T \sim \mathbb{P}(\Gamma)$ and the tasks' CMDPs $\{\mathcal{M}_{\tau_j}\}_{j=1}^T$ can be explored. During the meta-test, a new task τ_{new} is given, and the agent explores the CMDP $\mathcal{M}_{\tau_{new}}$ and produces the task-specific policy. Note that we consider the meta-training to be an offline stage, e.g. done in simulated environments, the safety constraints may be violated. In contrast, the policies are deployed to practical environments during the meta-test. Any policy used to explore $\mathcal{M}_{\tau_{new}}$ or used to execute the task τ_{new} should satisfy the safety constraints.

3 Safe Meta-RL Framework

The proposed safe meta-RL framework aims to learn a meta-policy π_ϕ such that it can adapt to new tasks with anytime safety guarantee. The framework includes two modules: safe policy adaptation, by which the task-specific policy π^τ for task τ is adapted from the meta-policy π_ϕ, and safe meta-training, which identifies the meta-policy π_ϕ.

Considering that the amount of data collection is limited, we expect that the task-specific policy π^τ is adapted from the meta-policy π_ϕ by a few safe policy adaptation steps and can achieve good performance and guarantee safety on the new tasks. To achieve this goal, we design the one-step safe policy adaptation in Section 3.1, which achieves significant policy improvement, guarantees safety, and holds high computational efficiency. In Section 3.2, the meta-policy training is to optimize the task-specific policy, which is adapted from the meta-policy π_ϕ by one-step safe policy adaptation.

3.1 One-step safe policy adaptation

Since data collection is limited when a new task is revealed, performing numerous policy adaptation steps to solve the original RL problem becomes impractical, as each step requires collecting a batch of data using the corresponding policy. Accordingly, we define **one-step policy adaptation as the policy adaptation that only needs to collect the data by a single policy** and derive the method for one-step safe policy adaptation in the remainder of this section.

We derive the optimization problem to achieve a one-step safe policy adaptation from the meta-policy. For task τ, the policy π^τ is adapted from the meta-policy π_ϕ by the safe policy adaptation $\mathcal{A}^s$:

$$\pi^\tau = \mathcal{A}^s(\pi_\phi, \Lambda, \Delta, \tau) \triangleq \underset{\pi\in\Pi}{\operatorname{argmax}}\, \mathbb{E}_{s\sim\nu_\tau^{\pi_\phi}, a\sim\pi(\cdot|s)}\left[A_\tau^{\pi_\phi}(s,a)\right] - \lambda\, \mathbb{E}_{s\sim\nu_\tau^{\pi_\phi}}\left[D_{KL}\left(\pi(\cdot|s)\|\pi_\phi(\cdot|s)\right)\right],$$
$$\text{s.t. } J_{c_i,\tau}(\pi_\phi) + \mathbb{E}_{\substack{s\sim\nu_\tau^{\pi_\phi}\\ a\sim\pi(\cdot|s)}}\left[\frac{A_{c_i,\tau}^{\pi_\phi}(s,a)}{1-\gamma}\right] + \lambda_{c_i}\, \mathbb{E}_{s\sim\nu_\tau^{\pi_\phi}}\left[D_{KL}\left(\pi(\cdot|s)\|\pi_\phi(\cdot|s)\right)\right] \le d_{i,\tau} + \delta_{c_i}, \quad (1)$$

where $i = 1, \cdots, p$, $\Lambda \triangleq \{\lambda, \lambda_{c_1}, \cdots, \lambda_{c_p}\}$ and $\Delta \triangleq \{\delta_{c_1}, \cdots, \delta_{c_p}\}$ are the hyper-parameters of $\mathcal{A}^s$. The safe policy adaptation $\mathcal{A}^s$ in problem (1) is inspired by the derivation of CPO [2], where both problem (1) and CPO aim to approximate the original safe RL problem. Specifically, the objective and constraint functions of problem (1) serve as upper bounds of the true objective and constraint functions $J_\tau(\pi)$ and $J_{c_i,\tau}(\pi)$ of the safe RL problem. More details about the upper bounds will be discussed in Lemma 1 of Section 5.1. More importantly, problem (1) only needs to collect state-action data points and evaluate $A_\tau^{\pi_\phi}$ for a single policy π_ϕ, which keeps the same requirement of data collection as one-step of gradient ascent in MAML [16]. Therefore, $\mathcal{A}^s$ is the one-step safe policy adaptation. On the other hand, considering a single gradient ascent in MAML is usually insufficient to identify a policy with good performance and safety, $\mathcal{A}^s$ is to completely solve (1).

The existence of the solution, the safety, and the monotonic improvement are guaranteed for $\mathcal{A}^s$. Specifically, when setting $\Delta = 0$, given that the meta-policy π_ϕ is safe for task τ, i.e., $J_{c_i,\tau}(\pi_\phi) \le d_{i,\tau}, \forall i = 1, \cdots, p$, for an appropriate hyper-parameter Λ, we have following properties: (i) the feasibility set of problem (1) is not empty; (ii) π^τ is safe for task τ, i.e., $J_{c_i,\tau}(\pi^\tau) \le d_{i,\tau}, \forall i = 1, \cdots, p$; (iii) the performance of π^τ is better than the meta-policy π_ϕ, i.e., $J_\tau(\pi^\tau) \ge J_\tau(\pi_\phi)$. The complete statements and proofs of property (i) are shown in Proposition 1 of Section 4.1; properties (ii) and (iii) under selected hyper-parameter Λ are shown in Section 5. Moreover, when the requirement of the constraint satisfaction is not strict, setting $\delta_{c_i} = 0$ for all i in problem (1) may overly restrict the policy update step. To enhance the algorithm's flexibility, we set $0 \le \delta_{c_i} \le \delta_{max}$ as an allowable constraint violation in problem (1).

As mentioned in the above properties (ii) and (iii), both CPO and problem (1) can achieve policy improvement and safety guarantee. However, the computational complexity of directly solving CPO or the constrained optimization problem (1) is high. CPO [2] and meta-CPO [12] solve an approximate problem to mitigate the issue, but the computational complexity is still high, meanwhile the safety constraint violation cannot be avoided in theory and also usually appears in practice. In contrast, the safe policy adaptation in (1) is designed to have the closed-form solution under certain Lagrangian multipliers, and then can be efficiently solved, which will be discussed in Section 4.1.

Note that problem (1), for the first time, simultaneously offers two key advantages: (a) constraint satisfaction guarantee for a single policy optimization step (policy optimization using data collected on a single policy), which enables anytime safety in each policy adaptation step during the meta-test, and (b) the closed-form solution, which significantly reduces the computational complexity of the meta-policy training. The details of the two benefits to the safe meta-RL problem will be discussed in Sections 4.1 and 5. Consequently, it is particularly well-suited for the safe meta-RL problem formulation. As the existing safe policy optimization algorithms, such as primal-dual-based algorithm in RCPO [**?**], PPO-Lagrangian [43], and CRPO [55] used by meta-CRPO, do not hold any of these two benefits, and therefore (1) cannot be replaced by these algorithms. Moreover, although some prior works [61, 34] also derive closed-form solutions of safe policy optimization, safety cannot be guaranteed in each step. Instead, safety is only guaranteed for the final convergent policy, where the trust region size ϵ is reduced to 0.

3.2 Safe meta-policy training

We obtain the optimal meta-policy π_{ϕ^*} by solving the following optimization problem:

$$\max_{\phi} \mathbb{E}_{\tau\sim\mathbb{P}(\Gamma)}[J_\tau(\mathcal{A}^s(\pi_\phi,\Lambda,\Delta,\tau))],\ \text{s.t. } J_{c_i,\tau}\left(\pi_\phi\right) \leq d_{i,\tau}+\delta_{c_i}, \forall i=1,\cdots,p \text{ and } \forall \tau\in\Gamma. \quad (2)$$

Here, $\mathbb{E}_{\tau\sim\mathbb{P}(\Gamma)}[J_\tau(\mathcal{A}^s(\pi_\phi,\Lambda,\Delta,\tau))]$ is the meta-objective function and is defined by the expected accumulated reward after the parameter is adapted by the policy adaptation, which evaluates the optimality of the meta-policy π_ϕ. We choose the constraints $J_{c_i,\tau}(\pi_\phi) \leq d_{i,\tau}+\delta_{c_i}, \forall i=1,\cdots,p$ for any task τ (similar to problem (1), we set δ_{c_i} as the allowable error). There are two reasons to set the constraints. First, as shown in Proposition 1, $J_{c_i,\tau}(\pi_\phi) \leq d_{i,\tau}+\delta_{c_i}, \forall i=1,\cdots,p$ is a sufficient condition for that the safe policy adaptation algorithm $\mathcal{A}^s(\pi_\phi,\Lambda,\Delta,\tau)$ has a solution, and further assure the safe meta-policy training (2) is well-defined. Second, the exploration of the CMDP by the meta-policy π_ϕ should be safe for each task τ to guarantee the initial policy of the policy adaptation is safe. As mentioned in Section 3.1, $\mathcal{A}^s(\pi_\phi,\Lambda,\Delta,\tau)$ is guaranteed to be safe for task τ when π_ϕ is safe, and iterative policy adaptation using $\mathcal{A}^s$ is guaranteed to be safe. Therefore, the anytime safety of the policy adaptation is guaranteed. Its formal statement is shown in Section 5.

4 Algorithm

This section introduces the efficient algorithmic solutions to solve problems (1) and (2), respectively.

4.1 Closed-form solution for safe policy adaptation

Based on the design of problem (1), we can derive its closed-form solution under certain Lagrangian multipliers, and then solve the Lagrangian multipliers to obtain the overall solution. We first derive the closed-form solution of problem (1) and show its existence in the following proposition.

Proposition 1. *Suppose that the softmax policy π_ϕ satisfies $J_{c_i,\tau}(\pi_\phi) \leq d_{i,\tau}+\delta_{c_i}, \forall i=1,\cdots,p$, the solution π^τ of the optimization problem* (1) *exists. Under certain mild constraint qualifications, there exists Lagrangian multipliers $\{u^*_{c_i,\tau}\}_{i=1}^p$ with $0 \leq u^*_{c_i,\tau} < \infty$, such that $\pi^\tau(\cdot|s) \propto \exp(f_\phi(s,\cdot) + \eta^{-1}(A^{\pi_\phi}_\tau(s,\cdot) - \sum_{i=1}^p u^*_{c_i,\tau} A^{\pi_\phi}_{c_i,\tau}(s,\cdot)))$, for any $s\in\mathcal{S}$, where $\eta \triangleq \lambda + (1-\gamma)\sum_{i=1}^p u^*_{c_i,\tau}\lambda_{c_i}$.*

The complete statement of Proposition 1 that includes the sufficient condition for the existence of $\{u^*_{c_i,\tau}\}_{i=1}^p$, as well as the proof of the proposition are shown in Appendix F.2.1. Proposition 1 shows that, when the meta-policy π_ϕ is softmax, the closed-form solution of the policy adaptation (1) is also softmax. The approximate function f_ϕ for the meta-policy π_ϕ is adapted to $f_\phi + \eta^{-1}(A^{\pi_\phi}_\tau - \sum_{i=1}^p u^*_{c_i,\tau} A^{\pi_\phi}_{c_i,\tau})$ of π^τ. With this computation, the approximate function of π^τ can be directly

obtained, which is much simpler than solving problem (1). More importantly, it can significantly reduce the computational complexity of the meta-gradient, which will be discussed in Section 4.2.

In addition, the closed-form solution in Proposition 1 implies the safe policy adaptation (1) can be reduced to the policy adaptation for an unconstrained MDP under the penalized reward function. Specifically, when we define a comprehensive reward function $\bar{r}_\tau \triangleq r_\tau - \sum_{i=1}^p u^*_{c_i,\tau} c_{i,\tau}$, then the term $A^{\pi_\phi}_\tau - \sum_{i=1}^p u^*_{c_i,\tau} A^{\pi_\phi}_{c_i,\tau}$ is the advantage function of π_ϕ for $\bar{r}_\tau$. This implies that problem (1) is equivalent to an unconstrained policy optimization problem, where the reward r_τ is penalized by the negative costs $-c_{i,\tau}$ and the weights of the cost penalty are given by the Lagrangian multiplier $u^*_{c_i,\tau}$.

Proposition 2. *Suppose that the assumption in Proposition 1 holds. Let π^u ($u \triangleq [u_1, \cdots, u_p]$) be the policy with $\pi^u(\cdot|s) \propto \exp(f_\phi(s,\cdot) + (\lambda + (1-\gamma)\sum_{i=1}^p u_i \lambda_{c_i})^{-1}(A^{\pi_\phi}_\tau(s,\cdot) - \sum_{i=1}^p u_i A^{\pi_\phi}_{c_i,\tau}(s,\cdot)))$. Then, the Lagrangian multipliers $\{u^*_{c_i,\tau}\}_{i=1}^p$ in Proposition 1 is the solution of the dual problem of* (1)*, i.e.,*

$$\min_{u \in \mathbb{R}^p_{\geq 0}} \mathbb{E}_{\substack{s\sim\nu^{\pi_\phi}_\tau \\ a\sim\pi^u}}[(A^{\pi_\phi}_\tau - \textstyle\sum_{i=1}^p u_i A^{\pi_\phi}_{c_i,\tau})(s,a) - D_{KL}\left(\pi^u(\cdot|s) \| \pi_\phi(\cdot|s)\right)] + \textstyle\sum_{i=1}^p u_i d'_{i,\tau}, \tag{3}$$

where $\eta^u \triangleq \lambda + (1-\gamma)\sum_{i=1}^p u_i \lambda_{c_i}$ and $d'_{i,\tau} \triangleq (1-\gamma)(d_{i,\tau} + \delta_{c_i} - J_{c_i,\tau}(\pi_\phi))$.

Proposition 2 shows the derivation of the Lagrangian multiplier $u^*_{c_i,\tau}$. Its proof is shown in Appendix F.2.2. With $u^*_{c_i,\tau}$. Note that problem (3) is the dual problem of (1), which is always convex. As a result, we can apply convex optimization approaches [9] to solve problem (3), and then the solution of safe policy adaptation (1) can be obtained immediately by Proposition 1. We provide an optional algorithm for solving problem (3) and its computational complexity analysis in Appendix E.1.

Algorithm 1 Safe meta-policy training algorithm

Require: Initial meta-policy π_{ϕ_0}; allowable constraint violation δ_{c_i} defined in Problems (1) and (2).

for $n = 0, \cdots, N-1$ **do**
  Sample a task τ with the CMDP $\mathcal{M}_\tau$ from the task distribution $\mathbb{P}(\Gamma)$
  Evaluate $J_{c_i,\tau}(\pi_{\phi_n})$, $A^{\pi_{\phi_n}}_\tau(\cdot,\cdot)$ and $A^{\pi_{\phi_n}}_{c_i,\tau}(\cdot,\cdot)$ by sampling data using the meta-policy π_{ϕ_n} on task τ
  if $J_{c_i,\tau}(\pi_{\phi_n}) \leq d_{i,\tau} + \delta_{c_i}, \forall i = 1, \cdots, p$ **then**
    Solve the task-specific policy π^τ and the Lagrangian multipliers $u^*_{c_i,\tau}(\pi_{\phi_n})$ with meta-policy π_{ϕ_n}
    Evaluate $Q^{\pi^\tau}_\tau(\cdot,\cdot)$ by sampling data using the task-specific policy π^τ on task τ
    Compute the meta-gradient $\nabla_\phi J_\tau(\pi^\tau)$ by (4)
    Take a step of TRPO [44] with using $\nabla_\phi J_\tau(\pi^\tau)$ towards maximize $J_\tau(\pi^\tau)$ to obtain ϕ_{n+1}
  else
    Choose any $i_n \in \{1, \cdots, p\}$ such that $J_{C_{i_n}}(\pi_{\phi_n}) > d_{i_n,\tau} + \delta_{c_{i_n}}$
    Compute the policy gradient $\nabla_\phi J_{C_{i_n},\tau}(\pi_{\phi_n}) \propto \mathbb{E}_{s\sim\nu^{\pi_{\phi_n}}_\tau, a\sim\pi_{\phi_n}(\cdot|s)}[\nabla_\phi f_{\phi_n}(s,a) A^{\pi_{\phi_n}}_{C_{i_n},\tau}(s,a)]$.
    Take a step of TRPO with using $\nabla_\phi J_{C_{i_n},\tau}(\pi_{\phi_n})$ towards minimize $J_{C_{i_n},\tau}(\pi_\phi)$ to obtain ϕ_{n+1}
  end if
end for
Return π_{ϕ_N}

4.2 Safe meta-policy training algorithm

To solve the optimization problem (2) for meta-training, we first consider the computation of the meta-gradient, i.e., $\nabla_\phi \mathbb{E}_{\tau\sim\mathbb{P}(\Gamma)}[J_\tau(\mathcal{A}^s(\pi_\phi, \Lambda, \Delta, \tau))]$. The following proposition provides the computation of $\nabla_\phi J_\tau(\mathcal{A}^s(\pi_\phi, \Lambda, \Delta, \tau))$. Notice that the Lagrangian multipliers $\{u^*_{c_i,\tau}\}_{i=1}^p$ in Propositions 1 and 2 are solved by problem (3), and thus depend on the meta-policy π_ϕ. We denote the solved Lagrangian multipliers with π_ϕ as $u^*_{c_i,\tau}(\pi_\phi)$ in the following sections.

Proposition 3. *Suppose the assumption in Proposition 1 holds. Let $\pi^\tau = \mathcal{A}^s(\pi_\phi, \Lambda, \Delta, \tau)$. Under certain conditions, we have that $\nabla_\phi J_\tau(\pi^\tau)$ exists and $\nabla_\phi J_\tau(\pi^\tau) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim\nu^{\pi^\tau}_\tau, a\sim\pi^\tau(\cdot|s)}[(\nabla_\phi \eta(\pi_\phi)^{-1} \bar{Q}^{\pi_\phi}_\tau(s,a) + \eta(\pi_\phi)^{-1} \nabla_\phi \bar{Q}^{\pi_\phi}_\tau(s,a) + \nabla_\phi f_\phi(s,a)) Q^{\pi^\tau}_\tau(s,a)]$, where $\eta(\pi_\phi) \triangleq \lambda + (1-\gamma)\sum_{i=1}^p u^*_{c_i,\tau}(\pi_\phi)\lambda_{c_i}$, and $\bar{Q}^{\pi_\phi}_\tau \triangleq Q^{\pi_\phi}_\tau - \sum_{i=1}^p u^*_{c_i,\tau}(\pi_\phi) Q^{\pi_\phi}_{c_i,\tau}$.*

We show the computations of $\nabla_\phi Q^{\pi_\phi}_\tau(\cdot)$, $\nabla_\phi Q^{\pi_\phi}_{c_i,\tau}(\cdot)$ and $\nabla_\phi u^*_{c_i,\tau}(\pi_\phi)$ in Appendices F.3.2 and F.3.3. The complete statement of Proposition 3 that includes the sufficient condition of the existence of $\nabla_\phi J_\tau(\pi^\tau)$, as well as the proof of the proposition are shown in Appendix F.3.1. In Proposition

3, the gradient $\nabla_\phi u^*_{c_i,\tau}(\pi_\phi)$ w.r.t ϕ, is the gradient of the solved Lagrangian multipliers, i.e. the optimal solution of problem (3). We apply the implicit gradient theorem for constrained optimization in [17, 52] to show the existence and the computation of $\nabla_\phi u^*_{c_i,\tau}(\pi_\phi)$, which is shown in Appendix F.3.3. We further simplify the computation of the meta-gradient as

$$\nabla_\phi J_\tau(\pi^\tau) \approx \mathbb{E}_{s\sim\nu^{\pi^\tau}_\tau, a\sim\pi^\tau(\cdot|s)}[(\nabla_\phi f_\phi(s,a) + \eta(\pi_\phi)^{-1}\tilde{\nabla}_\phi \bar{Q}^{\pi_\phi}_\tau(s,a))Q^{\pi^\tau}_\tau(s,a)], \tag{4}$$

where $\eta(\pi_\phi) \triangleq \lambda + (1-\gamma)\sum_{i=1}^p u^*_{c_i,\tau}(\pi_\phi)\lambda_{c_i}$ and $\tilde{\nabla}_\phi \bar{Q}^{\pi_\phi}_\tau = \nabla_\phi Q^{\pi_\phi}_\tau - \sum_{i=1}^p u^*_{c_i,\tau}(\pi_\phi)\nabla_\phi Q^{\pi_\phi}_{c_i,\tau}$. In (4), we take $\nabla_\phi u^*_{c_i,\tau}(\pi_\phi) = 0$ in Proposition 3 approximately. On one hand, the computation complexity of $\nabla_\phi u^*_{c_i,\tau}(\pi_\phi)$ is high, as shown in Appendix F.3.3. On the other hand, under this approximation, we only omit the small change of the Lagrangian multiplier $u^*_{c_i,\tau}(\pi_\phi)$ around the meta-policy π_ϕ, i.e., we keep the penalty to constraint violation but treat the weight of the penalty to constraint violation unchanged over a small neighbor of π_ϕ. Therefore, the omitted term is a higher-order term with a smaller impact on the meta-gradient. Note that, the meta-gradients in many meta-learning approaches include the Hessian computation, such as supervised meta-learning approaches, like MAML and iMAML [16, 42, 53], meta-RL [16, 30] and safe meta-RL approach meta-CPO [12]. In contrast, thanks to the closed-form solution (shown in Proposition 1) of the policy adaptation problem (1), the meta-gradient in (4) does not include the computations of Hessian and inverse of Hessian w.r.t. ϕ, which holds a comparable computational complexity as the policy gradient, and therefore is more computationally efficient than the above meta-learning approaches.

The safe meta-policy training algorithm aims to solve the optimization problem in (2) and is stated in Algorithm 1. To handle the constraint imposed on the meta-policy π_ϕ in problem (2), we use the idea similar to CRPO [55]. Specifically, we first check the constraint violation in line 4. If the constraints are not violated, we maximize the meta-objective; otherwise, we minimize the constraint functions. Under this procedure, we always have $J_{c_i,\tau}(\pi_{\phi_n}) \leq d_{i,\tau} + \delta_{c_i}, \forall i = 1, \cdots, p$ when computing the task-specific policy $\pi^\tau = \mathcal{A}^s(\pi_{\phi_n}, \Lambda, \Delta, \tau)$, and therefore the solution of π^τ always exists, according to Proposition 1. To stabilize the training, we use the TRPO for the policy update in lines 8 and 12, which only needs the gradient information.

5 Theoretical Results

In this section, we introduce the theoretical results of the safe meta-RL framework. Note that problem (2) is a constrained bilevel optimization problem, and the convergence and optimality analysis of solving the problem and obtaining π_{ϕ^*} are widely studied in [52, 7, 31]. So we analyze the performance of the solved meta-policy π_{ϕ^*} in our theoretical results. In particular, we introduce the necessary assumptions and notations, derive the performance guarantee for safe policy adaptation $\mathcal{A}^s$ in Section 5.1, and then derive the optimality and safety guarantee of the safe meta-RL framework in Section 5.2. We introduce an assumption and several notations used in the theoretical results.

Assumption 1. *The feasible set of problem* (2) *is not empty and bounded.*

Assumption 1 supposes problem (2) is well defined and its optimal meta-parameter ϕ^* exists. Since the reward $r_\tau \leq r^{max}$ and $c_{i,\tau} \leq c_i^{max}$, then $|A^\pi_\tau(s,a)| \leq r^{max}/(1-\gamma)$ and $|A^\pi_{c_i,\tau}(s,a)| \leq c_i^{max}/(1-\gamma)$ are upper bounded. We denote $A^{max} \triangleq \max_{\tau\in\Gamma,\pi\in\Pi}|A^\pi_\tau(s,a)|$ and $A^{max}_{c_i} \triangleq \max_{\tau\in\Gamma,\pi\in\Pi}|A^\pi_{c_i,\tau}(s,a)|$ for each $i = 1, \cdots, p$.

5.1 Monotonic improvement and anytime safety for policy adaptation

We first introduce a key lemma and show its proof in Appendix F.4.1. Here, we define $D_{TV}(\pi(\cdot|s)||\pi'(\cdot|s)) \triangleq \frac{1}{2}\sum_{a\in\mathcal{A}}|\pi(a|s) - \pi'(a|s)|$.

Lemma 1. *For any task τ, and any policies π and $\pi' \in \Pi$ with $\max_{s\in\mathcal{S}} D_{TV}(\pi||\pi') \leq \alpha$ $\mathbb{E}_{s\sim\nu^\pi_\tau}[D_{TV}(\pi(\cdot|s)\,||\pi'(\cdot|s))]\}$, we have*

$$J_\tau(\pi') \leq J_\tau(\pi) + \mathbb{E}_{s\sim\nu^\pi_\tau, a\sim\pi'(\cdot|s)}\left[\frac{A^\pi_\tau(s,a)}{1-\gamma}\right] + \frac{2\gamma\alpha A^{max}}{(1-\gamma)^2}\mathbb{E}_{s\sim\nu^\pi_\tau}\left[D_{KL}(\pi'(\cdot|s)||\pi(\cdot|s))\right]$$

$$J_\tau(\pi') \geq J_\tau(\pi) + \mathbb{E}_{s\sim\nu^\pi_\tau, a\sim\pi'(\cdot|s)}\left[\frac{A^\pi_\tau(s,a)}{1-\gamma}\right] - \frac{2\gamma\alpha A^{max}}{(1-\gamma)^2}\mathbb{E}_{s\sim\nu^\pi_\tau}\left[D_{KL}(\pi'(\cdot|s)||\pi(\cdot|s))\right].$$

The inequalities also holds when A^{π}_{τ}, $A^{\pi'}_{\tau}$, A^{max} and J_{τ} are replaced by $A^{\pi}_{c_i,\tau}$, $A^{\pi'}_{c_i,\tau}$, $A^{max}_{c_i}$, and $J_{c_i,\tau}$, for all $i = 1, \cdots, p$.

The right-hand sides of the inequalities in Lemma 1 are the objective function and constraint functions of problem (1). The first inequality in Lemma 1 (applied to $J_{c_i,\tau}(\pi')$) shows that the constraint function of problem (1) is the upper bound of the accumulated cost $J_{c_i,\tau}$. Therefore, the constraint functions in problem (1) limit the upper bound of $J_{c_i,\tau}(\pi')$ to be below the constraint limit, which also applies to $J_{c_i,\tau}(\pi')$ itself. The second inequality in Lemma 1 (applied to J_{τ}) shows that the objective function of problem (1) is the lower bound of $J_{\tau}(\pi')$. Then, $\mathcal{A}^s$ in problem (1) is to maximize the lower bound of $J_{\tau}(\pi')$, which guarantees monotonic improvement. We formalize the results in Proposition 4 and show the proofs in Appendix F.4.2.

Proposition 4. *Suppose π_{ϕ} satisfies that ϕ is bounded and $J_{c_i,\tau}(\pi_{\phi}) \leq d_{i,\tau} + \delta_{c_i}, \forall i = 1, \cdots, p$. There exists a constant α such that, when $\pi^{\tau} = \mathcal{A}^s(\pi_{\phi}, \Lambda, \Delta, \tau)$ with $\lambda \geq \frac{2\gamma\alpha A^{max}}{1-\gamma}$ and $\lambda_{c_i} \geq \frac{2\gamma\alpha A^{max}_{c_i}}{(1-\gamma)^2}$ for each $i = 1, \cdots, p$, then $J_{c_i,\tau}(\pi^{\tau}) \leq d_{i,\tau} + \delta_{c_i}$ for each i, and $J_{\tau}(\pi^{\tau}) \geq J_{\tau}(\pi_{\phi})$.*

With this proposition, we can derive the properties of monotonic improvement and anytime safety guarantee for the policy adaptation, which is stated in Corollary 1. As Assumption 1 assumes the feasible set of problem (2) is bounded, we fix the constant α in Proposition 4 for the bounded set.

Corollary 1. *Suppose that Assumptions 1 holds. Let $\lambda \geq \frac{2\gamma\alpha A^{max}}{1-\gamma}$ and $\lambda_{c_i} \geq \frac{2\gamma\alpha A^{max}_{c_i}}{(1-\gamma)^2}$ for each i. Let $\pi^{\tau}_{[k+1]} = \mathcal{A}^s(\pi^{\tau}_{[k]}, \Lambda, \Delta, \tau)$ with $\delta_{c_i} = 0$ for $k \in \mathbb{N}$, where $\pi^{\tau}_{[0]} = \pi_{\phi^*}$ being the solution of problem* (2)*. Then, for all $k \in \mathbb{N}$, $J_{c_i,\tau}(\pi^{\tau}_{[k]}) \leq d_{i,\tau}$ for each i and $J_{\tau}(\pi^{\tau}_{[k+1]}) \geq J_{\tau}(\pi^{\tau}_{[k]})$.*

When a new task $\tau \in \Gamma$ is given, we start from the meta-policy π_{ϕ^*}, iteratively implement $\mathcal{A}^s$, and generate a policy sequence $\{\pi^{\tau}_{[k]}\}^N_{k=0}$. As indicated in Corollary 1, the constraints are satisfied for each policy in the policy sequence, which shows the anytime safety of the policy adaptations.

5.2 Near-optimality and safety guarantee for one-step policy adaptation

In Section 5.1, we show the policy is always monotonically improved from π_{ϕ^*} and satisfies the safety constraints during policy adaptation. On the other hand, π_{ϕ^*} is learned from the task distribution $\mathbb{P}(\Gamma)$, which should be a good initial policy for the task sampled from $\mathbb{P}(\Gamma)$. In this section, we consider the policy that is obtained from using only one step of policy adaptation from π_{ϕ^*} and compare its optimality with the task-specific optimal policy to verify the near-optimality of the proposed safe meta-RL framework. We start by introducing several definitions.

Definitions. Define the optimal policy π^{τ}_* for task τ as $\pi^{\tau}_* \triangleq \text{argmax}_{\pi\in\Pi} J_{\tau}(\pi)$ s.t. $J_{c_i,\tau}(\pi) \leq d_{i,\tau}$. Define the ϵ-conservatively optimal policy $\pi^{\tau}_{*,[\epsilon]}$, which is optimal for τ under conservative safety constraints, i.e., $\pi^{\tau}_{*,[\epsilon]} \triangleq \text{argmax}_{\pi\in\Pi} J_{\tau}(\pi)$ s.t. $J_{c_i,\tau}(\pi) \leq d_{i,\tau} - \epsilon$, where the conservative constant $\epsilon \geq 0$, and $\pi^{\tau}_* = \pi^{\tau}_{*,[0]}$. We define the variance of a task distribution $\mathbb{P}(\Gamma)$ as $\mathcal{V}ar(\mathbb{P}(\Gamma)) \triangleq \min_{\phi} \mathbb{E}_{\tau\sim\mathbb{P}(\Gamma)}\mathbb{E}_{s\sim\nu^{\pi_{\phi}}_{\tau}}[D_{KL}(\pi^{\tau}_*(\cdot|s)||\pi_{\phi}(\cdot|s))]$, which the minimal mean square of the distances among the optimal task-specific policies π^{τ}_*, and the minimal point is denoted as $\hat{\phi}$. Similarly, the task variance under the conservative safety constraints is defined as $\mathcal{V}ar^{\epsilon}(\mathbb{P}(\Gamma)) \triangleq \min_{\phi} \mathbb{E}_{\tau\sim\mathbb{P}(\Gamma)}$ $\mathbb{E}_{s\sim\nu^{\pi_{\phi}}_{\tau}}[D_{KL}(\pi^{\tau}_{*,[\epsilon]}(\cdot|s)||\pi_{\phi}(\cdot|s))]$, and the minimal point is denoted as $\hat{\phi}^{[\epsilon]}$. The radius of $\mathbb{P}(\Gamma)$ is defined as $R(\mathbb{P}(\Gamma)) \triangleq \max_{\tau\in\Gamma,\epsilon\in E} \mathbb{E}_{s\sim\nu^{\pi_{\hat{\phi}^{[\epsilon]}}}_{\tau}}[D_{KL}(\pi^{\tau}_{*,[\epsilon]}(\cdot|s) \,||\pi_{\hat{\phi}^{[\epsilon]}}(\cdot|s))]$, where the set E is defined by $E \triangleq \{\epsilon \geq 0 : \pi^{\tau}_{*,[\epsilon]}$ exists for all $\tau \in \Gamma\}$. Note that the task variance $\mathcal{V}ar^{[\epsilon]}$ and the radius R is the inherent property of $\mathbb{P}(\Gamma)$, which measures the similarity of tasks sampled from $\mathbb{P}(\Gamma)$. For example, if the reward function r and cost c_i among tasks are similar, optimal policies $\pi^{\tau}_{*,[\epsilon]}$ are close, then $\mathcal{V}ar^{[\epsilon]}$ and R are close to 0. With the definitions, the near-optimality and safety guarantee of the safe meta-RL is shown in Theorem 1.

Theorem 1. *Suppose that Assumptions 1 holds. Let $\lambda = \frac{2\gamma\alpha A^{max}}{1-\gamma}$, $\lambda_{c_i} = \frac{2\gamma\alpha A^{max}_{c_i}}{(1-\gamma)^2}$ and $\delta_{c_i} = \frac{4\gamma\alpha A^{max}_{c_i}}{(1-\gamma)^2} R(\mathbb{P}(\Gamma)) - \epsilon$ for all $i = 1, \cdots, p$, where ϵ is chosen from $\left[0, \frac{4\gamma\alpha A^{max}_{c_i}}{(1-\gamma)^2} R(\mathbb{P}(\Gamma))\right]$. Let ϕ^* be*

the solution of problem (2)*. The solution of* $\mathcal{A}^s(\pi_{\phi^*}, \Lambda, \Delta, \tau)$ *exists, and we have*

$$\mathbb{E}_{\tau\sim\mathbb{P}(\Gamma)}[J_\tau(\mathcal{A}^s(\pi_{\phi^*}, \Lambda, \Delta, \tau))] \geq \mathbb{E}_{\tau\sim\mathbb{P}(\Gamma)}[J_\tau(\pi^\tau_{*,[\epsilon]})] - \frac{4\gamma\alpha A^{max}}{(1-\gamma)^2}\mathcal{V}ar^\epsilon(\mathbb{P}(\Gamma)), \tag{5}$$

$$J_{c_i,\tau}(\mathcal{A}^s(\pi_{\phi^*}, \Lambda, \Delta, \tau)) - d_{i,\tau} \leq \frac{4\gamma\alpha A^{max}_{c_i}}{(1-\gamma)^2}R(\mathbb{P}(\Gamma)) - \epsilon, \textit{ for any } \tau \in \Gamma. \tag{6}$$

Theorem 1 is proven in Appendix F.4.3. The theorem derives (i) the lower bound of the expected accumulated reward of the policy π^τ adapted by one time of $\mathcal{A}^s$ from the meta-parameter π_{ϕ^*} with the comparison to the task-specific (conservatively) optimal policy $\pi^\tau_{*,[\epsilon]}$. It also derives (ii) the upper bound of the constraint violation for each task τ. We further discuss the tightness of the derived bounds in Appendix G. Next, we explore two specific cases of δ_{c_i} to illustrate Theorem 1.

Case 1 (Safety guaranteed)**.** *When* $\delta_{c_i} = 0$*, the safe constraint is strictly satisfied, i.e.,* $J_{c_i,\tau}(\pi^\tau) - d_{i,\tau} \leq 0$ *for any* τ*, but the optimality comparator* $J_\tau(\pi^\tau_{*,[\epsilon]})$ *with* $\epsilon = \frac{4\gamma\alpha A^{max}_{c_i}}{(1-\gamma)^2}R(\mathbb{P}(\Gamma))$ *in* (5) *is conservatively optimal (*ϵ*-conservatively optimal).*

Case 2 (Near-optimality)**.** *When* $\delta_{c_i} = \frac{4\gamma\alpha A^{max}_{c_i}}{(1-\gamma)^2}R(\mathbb{P}(\Gamma))$*, the optimality comparator* $J_\tau(\pi^\tau_{*,[0]}) = J_\tau(\pi^\tau_*)$ *in* (5) *is all-task optimum, but the constraint is violated at most* $\frac{4\gamma\alpha A^{max}_{c_i}}{(1-\gamma)^2}R(\mathbb{P}(\Gamma))$*.*

As shown in Cases 1 and 2, there is a trade-off between the optimality of accumulated reward and the safety constraint satisfaction when the allowable constraint violation thresholds δ_{c_i} vary. In particular, when δ_{c_i} is increased, the optimality is improved while the constraint violation increases. As indicated by the optimality-safety trade-off, in the implementation of the proposed algorithm, we choose a large δ_{c_i} when the constraint satisfaction is not required to be strict, and a small $\delta_{c_i} \approx 0$ when the constraint satisfaction is prioritized. The reason for the trade-off is that the constraint function in problem (1) approximate the true constraints $J_{c_i,\tau}(\pi) - d_{i,\tau} \leq 0$ for any π by only knowing the information (the advantage functions $A^{\pi_\phi}_{c_i,\tau}$) at a single policy π_ϕ, and therefore are more conservative than the true constraints, which leads to loss of optimality. To the best of our knowledge, as anytime safety cannot be guaranteed in the existing framework [24, 12], it is the first time to show the trade-off between optimality and safety, and is also the first to provide an optimality bound with the anytime safe guarantee. Moreover, when choosing $\epsilon = 0$, Theorem 1 is reduced to the results in [54] for the unconstrained meta-RL.

Next, we delve into the optimality bound. Consider fixing δ_{c_i} and ϵ and then fixing the upper bound of the constraint violation $J_{c_i,\tau}(\pi^\tau)$. Theorem 1 shows that, the performance of meta-RL is improved when the variance of the task distribution $\mathcal{V}ar^\epsilon(\mathbb{P}(\Gamma))$ is reduced, as π^τ approach the task-specific optimal policy $\pi^\tau_{*,[\epsilon]}$. It corresponds to the intuition of meta-learning, which is that, when the variance of a task distribution is smaller, the tasks are more similar, and then the experience learned from the task distribution works better for new tasks sampled from the task distribution.

6 Experiments

Our experiments aim to validate three claimed benefits of the proposed algorithms for safe meta-RL: (i) superior optimality, i,e, the accumulated rewards of the proposed algorithms can exceed those of baselines; (ii) anytime safety, i,e, all the learned meta-policy and the adapted policies should satisfy the safety constraint; (iii) high computational efficiency for both the meta-training and meta-test.

We conduct experiments on four high-dimensional locomotion scenarios, including Half-Cheetah, Humanoid, Hopper, Swimmer, and three navigation scenarios with collision avoidance, including Point-Circle, Car-Circle-Hazard, and Point-Button in Gym and Safety-Gymnasium libraries [10, 23]. We compare the proposed method with three benchmarks: (a) MAML [16] with constraint penalty; (b) meta-CPO [12]; (c) meta-CRPO [24]. In (a), we add a weighted penalty term for constraint violation to the loss function of the MAML. Note that (c) is originally designed for online safe meta-RL, where tasks are revealed sequentially during the meta-training. So, we use (3) with all training tasks provided before the meta-training and it does not have the meta-training stage (Figures 1 and 2 do not have meta-training for meta-CRPO). For the fairness of the comparison, all the methods have the same data requirements and task settings. More details about the settings of the tasks, algorithm implementation, and hyper-parameters are shown in Appendices D.1 and D.2.

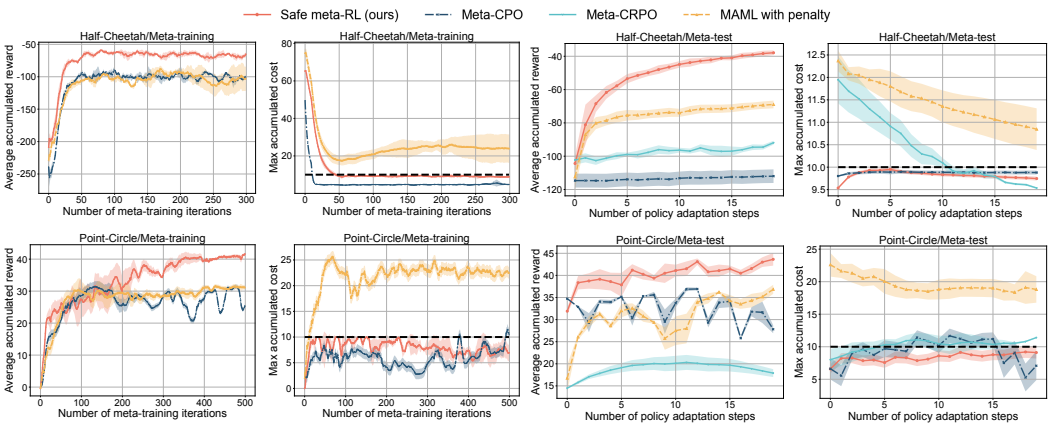


Figure 1: Average accumulated reward (**columns 1 and 3, higher is better**) and maximal accumulated cost (**columns 2 and 4, higher is worse**) across all validation/test tasks during the meta-training (**columns 1 and 2**) and the meta-test (**columns 3 and 4**) in Half-Cheetah (**row 1**) and Point-Circle (**row 2**). The accumulated reward and cost during meta-training are computed on the policy adapted one step from the meta-policy. The black dashed line is the constraint of the accumulated cost (below the line means satisfaction).

Figures 1 and 2 present the experimental results on Half-Cheetah and Point-Circle tasks. Due to page limitations, the results for the other four scenarios are deferred to Appendix D.3.

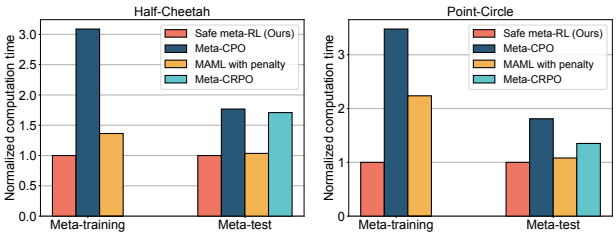


Figure 2: Normalized computation time of the meta-training (per iteration) and meta-test.

Performance on optimality and Safety. Figure 1 illustrates that the proposed safe meta-RL algorithm substantially outperforms all baselines in terms of optimality, achieving approximately 50% higher accumulated rewards than the best-performing baseline during both the meta-training and meta-test phases. Moreover, as depicted in the fourth column of Figure 1, our method ensures anytime safety during meta-testing, i.e., the maximal accumulated costs consistently remain below the prescribed safety thresholds, whereas all baselines experience constraint violations at various adaptation stages.

Efficiency. Figure 2 demonstrates that the proposed algorithm achieves remarkable computational efficiency, reducing the meta-training time by about 70% and the meta-testing time by about 50% relative to meta-CPO. This efficiency gain stems from the closed-form safe policy adaptation and the Hessian-free meta-gradient, which avoid costly second-order computations common in previous meta-RL methods.

Trade-off Verification. Finally, we empirically examine the optimality–safety trade-off in Appendix D.4, confirming our theoretical analysis that relaxing the safety tolerance slightly improves the achievable reward, while strict constraint satisfaction preserves anytime safety with minimal optimality degradation.

7 Conclusion

This paper presents an efficient framework for safe meta-RL that achieves provable anytime safety and near-optimality. By integrating a closed-form one-step safe policy adaptation with a Hessian-free safe meta-policy training scheme, the proposed method ensures zero constraint violation for every exploration policy, guarantees monotonic performance improvement, and significantly reduces computational cost. We provide the first formal analysis establishing both an optimality bound and an explicit safety–optimality trade-off, offering a tunable balance between strict safety enforcement and reward maximization. Empirically, our approach outperforms existing safe meta-RL methods in both optimality and efficiency across diverse locomotion and navigation tasks.

Acknowledgements

This work is partially supported by the National Science Foundation through grants ECCS 1846706 and ECCS 2140175. We would like to thank the reviewers for their constructive and insightful suggestions.

References

[1] Naoki Abe, Prem Melville, Cezar Pendus, Chandan K Reddy, David L Jensen, Vince P Thomas, James J Bennett, Gary F Anderson, Brent R Cooley, and Melissa Kowalczyk. Optimizing debt collections using constrained reinforcement learning. In *Proceedings of ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 75–84, 2010.

[2] Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *International Conference on Machine Learning*, pages 22–31, 2017.

[3] Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *The Journal of Machine Learning Research*, 22(1):4431–4506, 2021.

[4] Eitan Altman. *Constrained Markov decision processes*. Routledge, 2021.

[5] Jacob Beck, Risto Vuorio, Evan Zheran Liu, Zheng Xiong, Luisa Zintgraf, Chelsea Finn, and Shimon Whiteson. A survey of meta-reinforcement learning. *arXiv preprint arXiv:2301.08028*, 2023.

[6] Suneel Belkhale, Rachel Li, Gregory Kahn, Rowan McAllister, Roberto Calandra, and Sergey Levine. Model-based meta-reinforcement learning for flight with suspended payloads. *IEEE Robotics and Automation Letters*, 6(2):1471–1478, 2021.

[7] Quentin Bertrand, Quentin Klopfenstein, Mathurin Massias, Mathieu Blondel, Samuel Vaiter, Alexandre Gramfort, and Joseph Salmon. Implicit differentiation for fast hyperparameter selection in non-smooth convex learning. *Journal of Machine Learning Research*, 23(149):1–43, 2022.

[8] Dimitri P Bertsekas. Nonlinear programming. *Journal of the Operational Research Society*, 48(3):334–334, 1997.

[9] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

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

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

[12] Minjae Cho and Chuangchuang Sun. Constrained meta-reinforcement learning for adaptable safety guarantee with differentiable convex programming. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 20975–20983, 2024.

[13] Yinlam Chow, Mohammad Ghavamzadeh, Lucas Janson, and Marco Pavone. Risk-constrained reinforcement learning with percentile risk criteria. *Journal of Machine Learning Research*, 18(167):1–51, 2018.

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

[15] Dongsheng Ding, Xiaohan Wei, Zhuoran Yang, Zhaoran Wang, and Mihailo Jovanovic. Provably efficient safe exploration via primal-dual policy optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 3304–3312, 2021.

[16] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pages 1126–1135, 2017.

[17] Giorgio Giorgi and Cesare Zuccotti. A tutorial on sensitivity and stability in nonlinear programming and variational inequalities under differentiability assumptions. Technical report, DEM Working Paper Series, 2018.

[18] Yang Guan, Yangang Ren, Qi Sun, Shengbo Eben Li, Haitong Ma, Jingliang Duan, Yifan Dai, and Bo Cheng. Integrated decision and control: toward interpretable and computationally efficient driving intelligence. *IEEE Transactions on Cybernetics*, 53(2):859–873, 2022.

[19] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning*, pages 1861–1870, 2018.

[20] Sandy Huang, Abbas Abdolmaleki, Giulia Vezzani, Philemon Brakel, Daniel J Mankowitz, Michael Neunert, Steven Bohez, Yuval Tassa, Nicolas Heess, Martin Riedmiller, et al. A constrained multi-objective reinforcement learning framework. In *Conference on Robot Learning*, pages 883–893. PMLR, 2022.

[21] Yanlong Huang. Ekmp: Generalized imitation learning with adaptation, nonlinear hard constraints and obstacle avoidance. *arXiv preprint arXiv:2103.00452*, 2021.

[22] Alfredo N Iusem. On the convergence properties of the projected gradient method for convex optimization. *Computational & Applied Mathematics*, 22:37–52, 2003.

[23] Jiaming Ji, Borong Zhang, Jiayi Zhou, Xuehai Pan, Weidong Huang, Ruiyang Sun, Yiran Geng, Yifan Zhong, Josef Dai, and Yaodong Yang. Safety gymnasium: A unified safe reinforcement learning benchmark. *Advances in Neural Information Processing Systems*, 36, 2023.

[24] Vanshaj Khattar, Yuhao Ding, Javad Lavaei, and Ming Jin. A CMDP-within-online framework for meta-safe reinforcement learning. In *International Conference on Learning Representations*, 2023.

[25] Konwoo Kim, Gokul Swamy, Zuxin Liu, Ding Zhao, Sanjiban Choudhury, and Steven Z Wu. Learning shared safety constraints from multi-task demonstrations. *Advances in Neural Information Processing Systems*, 36, 2023.

[26] Dennis Lee, Haoran Tang, Jeffrey Zhang, Huazhe Xu, Trevor Darrell, and Pieter Abbeel. Modular architecture for starcraft ii with deep reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 14, pages 187–193, 2018.

[27] Joonho Lee, Jemin Hwangbo, Lorenz Wellhausen, Vladlen Koltun, and Marco Hutter. Learning quadrupedal locomotion over challenging terrain. *Science Robotics*, 5(47):5986, 2020.

[28] Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies. *Journal of Machine Learning Research*, 17(39):1–40, 2016.

[29] Boyi Liu, Qi Cai, Zhuoran Yang, and Zhaoran Wang. Neural trust region/proximal policy optimization attains globally optimal policy. *Advances in Neural Information Processing Systems*, 32, 2019.

[30] Hao Liu, Richard Socher, and Caiming Xiong. Taming maml: Efficient unbiased meta-reinforcement learning. In *International Conference on Machine Learning*, pages 4061–4071, 2019.

[31] Risheng Liu, Yaohua Liu, Shangzhi Zeng, and Jin Zhang. Towards gradient-based bilevel optimization with non-convex followers and beyond. *Advances in Neural Information Processing Systems*, 34:8662–8675, 2021.

[32] Shicheng Liu and Minghui Zhu. Meta inverse constrained reinforcement learning: Convergence guarantee and generalization analysis. In *The Twelfth International Conference on Learning Representations*.

[33] Shicheng Liu and Minghui Zhu. Distributed inverse constrained reinforcement learning for multi-agent systems. *Advances in Neural Information Processing Systems*, 35:33444–33456, 2022.

[34] Zuxin Liu, Zhepeng Cen, Vladislav Isenbaev, Wei Liu, Steven Wu, Bo Li, and Ding Zhao. Constrained variational policy optimization for safe reinforcement learning. In *International Conference on Machine Learning*, pages 13644–13668. PMLR, 2022.

[35] Gabriel B Margolis, Tao Chen, Kartik Paigwar, Xiang Fu, Donghyun Kim, Sang Bae Kim, and Pulkit Agrawal. Learning to jump from pixels. In *Annual Conference on Robot Learning*, 2021.

[36] Gabriel B Margolis, Ge Yang, Kartik Paigwar, Tao Chen, and Pulkit Agrawal. Rapid locomotion via reinforcement learning. *The International Journal of Robotics Research*, 43(4):572–587, 2024.

[37] 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):529–533, 2015.

[38] Anusha Nagabandi, Ignasi Clavera, Simin Liu, Ronald S Fearing, Pieter Abbeel, Sergey Levine, and Chelsea Finn. Learning to adapt in dynamic, real-world environments through meta-reinforcement learning. In *International Conference on Learning Representations*, 2018.

[39] Santiago Paternain, Miguel Calvo-Fullana, Luiz FO Chamon, and Alejandro Ribeiro. Safe policies for reinforcement learning via primal-dual methods. *IEEE Transactions on Automatic Control*, 68(3):1321–1336, 2022.

[40] David W Peterson. A review of constraint qualifications in finite-dimensional spaces. *Siam Review*, 15(3):639–654, 1973.

[41] Nicholas Polosky, Bruno C Da Silva, Madalina Fiterau, and Jithin Jagannath. Constrained offline policy optimization. In *International Conference on Machine Learning*, pages 17801–17810, 2022.

[42] Aravind Rajeswaran, Chelsea Finn, Sham M Kakade, and Sergey Levine. Meta-learning with implicit gradients. *Advances in Neural Information Processing Systems*, 32, 2019.

[43] Alex Ray, Joshua Achiam, and Dario Amodei. Benchmarking safe exploration in deep reinforcement learning. 2019.

[44] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International Conference on Machine Learning*, pages 1889–1897, 2015.

[45] John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*, 2015.

[46] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, and Marc Lanctot. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

[47] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[48] Chen Tessler, Daniel J Mankowitz, and Shie Mannor. Reward constrained policy optimization. In *International Conference on Learning Representations*, 2018.

[49] Akifumi Wachi and Yanan Sui. Safe reinforcement learning in constrained markov decision processes. In *International Conference on Machine Learning*, pages 9797–9806, 2020.

[50] Lingxiao Wang, Qi Cai, Zhuoran Yang, and Zhaoran Wang. On the global optimality of model-agnostic meta-learning. In *International Conference on Machine Learning*, pages 9837–9846, 2020.

[51] Siyuan Xu and Minghui Zhu. Meta value learning for fast policy-centric optimal motion planning. *Robotics Science and Systems*, 2022.

[52] Siyuan Xu and Minghui Zhu. Efficient gradient approximation method for constrained bilevel optimization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(10):12509–12517, 2023.

[53] Siyuan Xu and Minghui Zhu. Online constrained meta-learning: Provable guarantees for generalization. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[54] Siyuan Xu and Minghui Zhu. Meta-reinforcement learning with universal policy adaptation: Provable near-optimality under all-task optimum comparator. *arXiv preprint arXiv:2410.09728*, 2024.

[55] Tengyu Xu, Yingbin Liang, and Guanghui Lan. CRPO: A new approach for safe reinforcement learning with convergence guarantee. In *International Conference on Machine Learning*, pages 11480–11491, 2021.

[56] Tsung-Yen Yang, Justinian Rosca, Karthik Narasimhan, and Peter J Ramadge. Projection-based constrained policy optimization. In *International Conference on Learning Representations*, 2019.

[57] Dongjie Yu, Haitong Ma, Shengbo Li, and Jianyu Chen. Reachability constrained reinforcement learning. In *International Conference on Machine Learning*, pages 25636–25655, 2022.

[58] Ming Yu, Zhuoran Yang, Mladen Kolar, and Zhaoran Wang. Convergent policy optimization for safe reinforcement learning. *Advances in Neural Information Processing Systems*, 32, 2019.

[59] Zhenyuan Yuan, Siyuan Xu, and Minghui Zhu. All-time safety and sample-efficient meta update for online safe meta reinforcement learning under markov task transition. *Machine Learning*, 114(8):173, 2025.

[60] Jesse Zhang, Brian Cheung, Chelsea Finn, Sergey Levine, and Dinesh Jayaraman. Cautious adaptation for reinforcement learning in safety-critical settings. In *International Conference on Machine Learning*, pages 11055–11065. PMLR, 2020.

[61] Yiming Zhang, Quan Vuong, and Keith Ross. First order constrained optimization in policy space. *Advances in Neural Information Processing Systems*, 33:15338–15349, 2020.

Appendix for "Efficient Safe Meta-Reinforcement Learning: Provable Near-Optimality and Anytime Safety"

A Related Works

Safety metrics in safe RL. Safe RL aims to handle the safety requirements in the practical applications of RL. Safe RL typically applies two categories of safety metrics. The first metric is used in CMDP [4] and is applied in [48, 13, 15, 11, 2, 56, 41, 33, 32]. It introduces costs associated with state-action pairs based on MDP, and the agent is defined as safe when the expected accumulated costs satisfy given safety constraints. The second metric is stay in the safety region [49, 57, 39], which is stricter than the first metric. Specifically, the agent is safe when it remains in a desired safe set for any sampled trajectory. In this paper, we consider the anytime safety during policy adaptation, where each policy is required during the exploration of an unknown MDP. It is naturally infeasible to guarantee anytime safety under the second safety metric, as the action to remain in the safety region is unknown before the exploration. In contrast, the agent could be safe under the first safety metric even if it visits some undesired states. As a result, we consider the first safety metric.

Solutions of CMDPs. The solutions of the CMDPs can be categorized into (i) penalty function [18], (ii) primal-dual approaches [48, 13, 58, 15, 11], (iii) trust-region approaches [2, 56, 61, 34]. Existing works theoretically establish the safety guarantee for both primal-dual approaches [13, 58, 15] and trust-region approaches [2]. The primal-dual approaches update the dual variables and the policy simultaneously. Therefore, they gradually reduce the total cost below the required threshold by multiple policy optimization steps and can only establish the safety guarantee for the final convergent policy and cannot guarantee anytime safety during policy optimization. Therefore, they cannot meet the anytime safety requirement during policy adaptation in the safe meta-RL problems, i.e., the safety constraints are satisfied during each step of policy adaptation. In contrast, trust-region approaches constrain the policy within a safe policy set, potentially ensuring safety for every policy during the policy optimization process. However, the computational complexity of existing trust-region approaches is high, especially when applied to the safe meta-RL problem. The safety policy adaptation in this paper belongs to the category of trust-region approaches. On the other hand, we propose a novel safe policy adaptation method to address the computational inefficiency issue.

Cautious adaptation and safe meta-RL. Cautious adaptation [60] and safe meta-RL both consider to learn prior knowledge to improve the safety level of the adaptations in new environments. On the other hand, cautious adaptation considers the out-of-distribution exploration with the prior learned safety knowledge. The safe meta-RL focuses on in-distribution few-shot learning with safety constraints. Therefore, the safe meta-RL requires less exploration data during adaptation than cautious adaptation, but is limited to in-distribution tasks and less generalizable than cautious adaptation.

Safe meta-RL v.s. multitask/multi-objective safe RL methods. Safe meta-RL, multi-task safe RL [25], and multi-objective safe RL [20] all consider the multiple tasks in the safe RL setting. However, the biggest difference between meta-safe RL and multi-task/multi-objective safe RL is that the agent in meta-safe RL is required to adapt to a new and unknown environment under few-shot data collection. Therefore, the policy adaptation algorithm is the most important part of meta-safe RL. This paper designs a novel policy adaptation algorithm that holds several benefits for the few-shot policy adaptation that the existing methods do not hold. In contrast, the multi-task/multi-objective safe RL learns the policies for multiple tasks during the training stage, where the policy adaptation is not required. Therefore, the multi-task/multi-objective can borrow the existing policy optimization methods and do not need to design a new one.

B Discussion of the relations between CPO [2] and the safe policy adaptation by problem (1)

The safe policy adaptation $\mathcal{A}^s$ in (1) is inspired by the derivation of CPO, the first optimization problem in Section 5.3 of [2], and replaces the term $\sqrt{D_{KL}\left(\pi(\cdot|s)\|\pi_\phi(\cdot|s)\right)}$ in the objective and the constraint functions of the optimization problem by $D_{KL}\left(\pi(\cdot|s)\|\pi_\phi(\cdot|s)\right)$. Similarly, we derive the inequalities in Lemma 1 replace the term $\max_s D_{KL}(\pi'(\cdot|s)||\pi(\cdot|s)$ in Theorem 1 in [44] and replace the term $\sqrt{\mathbb{E}_{s\sim\nu^\pi_\tau}\left[D_{KL}(\pi'(\cdot|s)||\pi(\cdot|s))\right]}$ in Corollary 3 in [2] by $\mathbb{E}_{s\sim\nu^\pi_\tau}\left[D_{KL}(\pi'(\cdot|s)||\pi(\cdot|s))\right]$ in the right-hand side of the inequalities.

The modification from [2] to the safe policy adaptation $\mathcal{A}^s$ holds two benefits: (i) performance guarantee and (ii) computational efficiency. First, as Corollary 3 in [2] enables the feasibility, the monotonic improvement, and the constraint satisfaction to hold for the solution of the first optimization problem in Section 5.3 of [2], Lemma 1 enables the feasibility, the monotonic improvement, and the constraint satisfaction to hold for the safe policy adaptation $\mathcal{A}^s$. Second, the modification to the safe policy adaptation $\mathcal{A}^s$ enables us to derive its closed-form solution, which significantly reduces the computational complexity of the meta-safe RL algorithm, as mentioned in Section 4.1. On the other hand, one cannot derive the closed-form solution for the first optimization problem in Section 5.3 of [2], and the computational complexity is high. Paper [2] solves an approximate problem to mitigate the issue, but the computational complexity is still high, meanwhile, the safety constraint violation cannot be avoided in theory and also usually appears in practice.

C Comparisons between the proposed safe policy adaptation method and existing Lagrangian-based safe RL algorithms

The Lagrangian-based policy optimization algorithm, such as RCPO [48], PPO-Lagrangian [43] and CRPO [55] used in meta-CRPO [24], has been widely used to solve safe RL. However, although both the proposed method in Section 4.1 and the primal-dual method in RCPO, PPO-Lagrangian, and CRPO, are Lagrangian-based safe policy optimization algorithms, they are different. The primal-dual method is much worse than the proposed method and is not suitable for this safe meta-RL problem.

The method in Section 4.1 is to solve the safe policy adaptation problem in (1). As mentioned in Section 3.1, the safe policy adaptation (1) holds several benefits similar to CPO, including the safety guarantee for a single policy optimization step (using data collected on a single policy) and the monotonic improvement. Moreover, we derive the closed-form solution under certain Lagrangian multipliers for the optimization problem (1). Based on the derived closed-form solution of (1) (shown in Proposition 1), we can use the method shown in (3) and (7) to solve the safe policy adaptation problem in (1), which significantly reduces the computational complexity during the meta-training.

In contrast, RCPO and PPO-Lagrangian do not hold any of the benefits shown in CPO and the proposed algorithm. First, RCPO and PPO-Lagrangian use the gradient ascent steps on the Lagrangian, which do not have the safety guarantee and the monotonic improvement in each policy optimization step, and therefore cannot guarantee anytime safety in the meta-test stage. Moreover, there is no closed-form solution for the policy optimization step in RCPO and PPO-Lagrangian, which leads the high computational complexity during the meta-training.

D Experimental Supplements

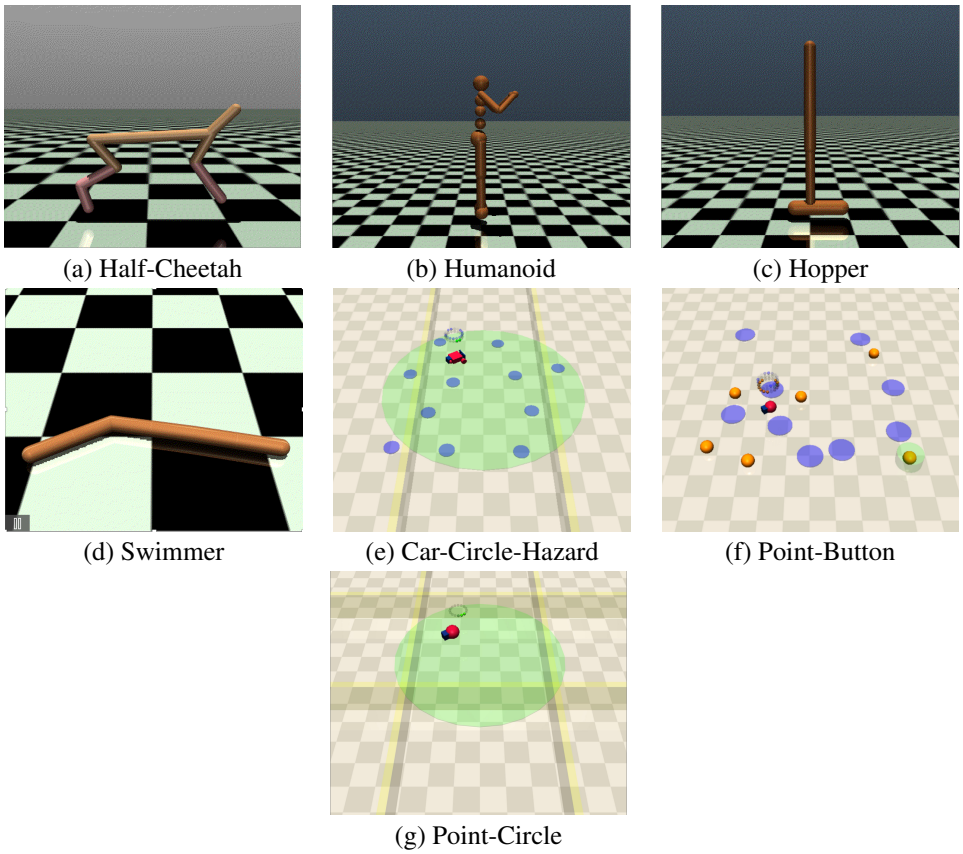
All experiments are executed on a computer with a 5.20 GHz Intel Core i12 CPU.

D.1 Task settings

We conduct experiments on totally seven scenarios, which include four high-dimensional locomotion scenarios (Half-Cheetah, Humanoid, Hopper, and Swimmer) in Gym library [10], and three navigation scenarios with collision avoidance (Point-Circle, Car-Circle-Hazard, and Point-Button) in Safety-Gymnasium library [23]. The scenarios are visually illustrated in Figure 3. We use the task setups similar to those used in previous works on meta-RL and safe meta-RL [12, 16, 24]. We provide the details of the task setups as follows.

Half-Cheetah. Half-Cheetah (Figure 3.a) has a 17-dimensional state space and a 6-dimensional action space. In the experiment of Half-Cheetah, the reward is the negative absolute value between the agent's current velocity and a goal velocity, where the goal velocity characterizes the task. The task distribution is defined by the distribution of the goal velocity, which is a uniform distribution from 0.0 to 2.0. The cost is defined by $h_{\text{cheetah}} - h_0 \leq d_\tau$, i.e. the cost is positive when its head is higher than h_0.

Humanoid. Humanoid (Figure 3.b) has a 376-dimensional observation space and a 17-dimensional action space. In the experiment of Humanoid, the reward is set as $v_y \sin\theta + v_x \cos\theta$, where v_x and v_y are the velocities along the x-axis and y-axis, and θ is the walking direction of the humanoid. So



(a) Half-Cheetah (b) Humanoid (c) Hopper

(d) Swimmer (e) Car-Circle-Hazard (f) Point-Button

(g) Point-Circle

Figure 3: High-dimensional locomotion tasks and navigation tasks with collision avoidance.

the reward is the velocity along the direction θ. The task is characterized by the walking direction θ, which is sampled uniformly from 0 to $\pi/2$. The cost is defined by the control cost of the humanoid robot, i.e., $\sum_i c_i^2$, where c_i is the torque imposed on each component.

Hopper. Hopper (Figure 3.c) has a 12-dimensional state space and a 3-dimensional action space. In the experiment of Hopper, the reward is the negative absolute value between the agent's current velocity and a goal velocity, where the goal velocity characterizes the task. The task distribution is defined by the distribution of the goal velocity, which is a uniform distribution from 0.0 to 1.0. The cost is defined by the control cost of the robot.

Swimmer. Swimmer (Figure 3.d) has a 8-dimensional state space and a 2-dimensional action space. In the experiment of Swimmer, for different tasks, we add a Gaussian noise to the state transition, and the variance is uniformly sampled from 0.0 to 0.5 for different tasks; we use the reward defined as the negative absolute value between the agent's current velocity and a goal velocity, which is a uniform distribution from 0.0 to 1.0, we used the cost defined by the control cost of the swimmer robot, i.e., $w \sum_i c_i^2$, where c_i is the torque imposed on each component and the weight w is sampled uniformly from 0.5 to 1.

Point-Circle. Point-Circle (Figure 3.e) has a 28-dimensional state space and a 2-dimensional action space. In the experiment of Point-Circle, a positive reward is given when the agent runs in a circle, and a positive cost is given when the agent does not stay within the safe region. The setting of the safe region characterizes the task. The task distribution is defined by the distribution of the circle radius and the wall distance. The circle radius is a uniform distribution from 1.0 to 1.5 and the wall distance is a uniform distribution from 0.55 to 0.75.

Car-Circle-Hazard. Car-Circle-Hazard (Figure 3.f) has a 60-dimensional state space and a 2-dimensional action space. In the experiment of Car-Circle-Hazard, a positive reward is given when the agent runs in a circle, and a positive cost is given when the agent does not stay within the safe

Table 2: Hyper-parameter setting in $\mathcal{A}^s$

scenario	λ, λ_{c_1}	d_τ	δ_{c_1}
Half-Cheetah	1.0	10.0	0.0
Humanoid	5.0	20.0	0.0
Hopper	1.0	5.0	0.0
Swimmer	0.2	5.0	0.0
Point-Circle	0.5	10.0	0.0
Car-Circle-Hazard	0.5	10.0	0.0
Point-Button	0.5	10.0	0.0

region or collides with Hazards. The setting of the safe region and the hazards characterize the task. The task distribution is defined by the distribution of the circle radius, the distribution of the positions, and the distribution of the number of hazards. The circle radius is a uniform distribution from 0.7 to 1.0 and the number of hazards is a uniform distribution from 3 to 7. the distribution of the position of the hazard is a uniform distribution over the safety space.

Point-Button. Point-Button (Figure 3.g) has a 56-dimensional state space and a 2-dimensional action space. In the experiment of Point-Button, a positive reward is given when the agent touches a goal button, and a positive cost is given when it does not stay within the safe region and touches any no-goal button or hazards. The setting of the buttons and the hazards characterize the task. The task distribution is defined by the distribution of the number and the positions of buttons and the number and the positions of hazards. Both the number of buttons and the number of hazards is a uniform distribution from 6 to 10, and the distributions of positions of buttons and hazards are uniform distributions over the safety space.

D.2 Algorithm settings

We apply Algorithm 4. We consider the policy as a Gaussian distribution, where the neural network produces the means and variances of the actions. The neural network policy has two hidden layers of size 64, with tanh nonlinearities. The horizon is 200, with 40 rollouts per policy adaptation step for all problems in the high-dimensional locomotion scenarios. The horizon is 500, with 10 rollouts per policy adaptation step for all problems in the navigation scenarios. The discount factor $\gamma = 0.99$. In each iteration, we sample 10 tasks from the task distribution. Therefore, for each meta-training iteration, the number of the sampled state-action pairs is 50k or 80k. The models are trained for up to 300 meta-iterations in the meta-training. Therefore, the overall number of sampled state-action pairs is from 15M or 24M. The meta-policy is tested on 20 tasks and is adapted by 20 iterations for each task in the meta-test. For the TRPO in meta-parameter optimization, we use the KL-divergence constraint as $\delta = 1e-3$. We set $\lambda = \lambda_{c_1}$ in the safe policy adaptation $\mathcal{A}^s$ in problem (1). Table 2 shows the setting of λ and d_τ in $\mathcal{A}^s$ for each scenario.

We compare the proposed method with three benchmarks: (a) MAML [16] with constraint penalty, (b) meta-CPO [12], and meta-CRPO [24]. For all methods, we run each algorithm 5 times, including meta-training and meta-test, and show the mean and standard deviation of the evaluation quantities.

D.3 Supplemental results

Figures 4 and 5 show the experimental results in Humanoid, Hopper, Car-Circle-Hazard, and Point-Button. Note that meta-CRPO is not designed for offline optimization of meta-policy, and then there is no meta-training result for the approach. Due to the high dimension of the Humanoid tasks, the meta-training of meta-CPO is too slow (10 times slower than the proposed method) in Humanoid tasks. It is extremely time-consuming to run the meta-training of meta-CPO multiple times on humanoid tasks and draw its figure. So the result of meta-CPO is not shown in Fig 4.

Figure 4 shows that the proposed safe meta-RL algorithm significantly outperforms all the baseline methods regarding the optimality, i.e. the accumulated reward during both the meta-training and the meta-test in all the scenarios. Moreover, it shows that the proposed algorithms achieve anytime safety during the meta-test, i.e., the maximal accumulated costs always satisfy the constraints, while

the baselines cannot achieve it. Figure 5 shows that our algorithm is much more efficient than the baselines in both the meta-training and meta-test stages.

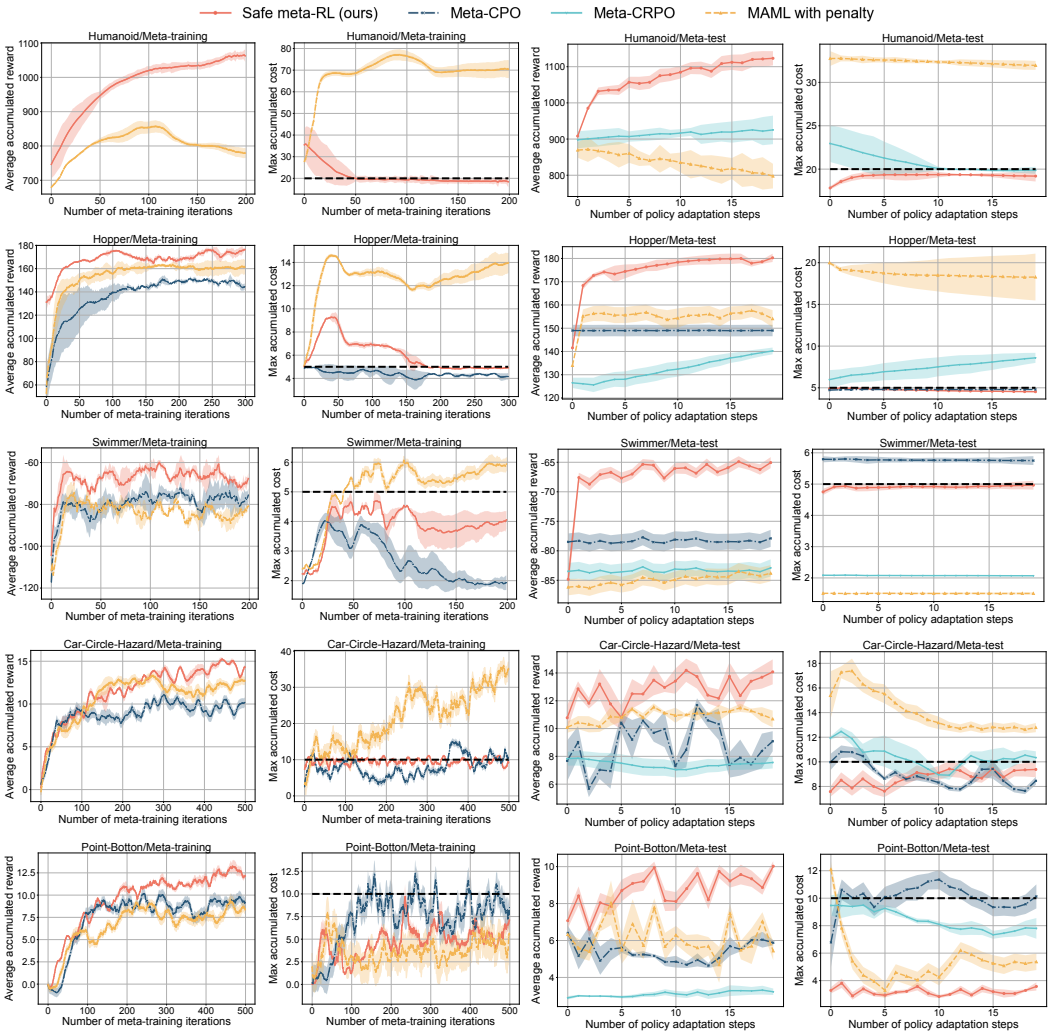


Figure 4: Average accumulated reward (**columns 1 and 3, higher is better**) and maximal accumulated cost (**columns 2 and 4, higher is worse**) across all validation/test tasks during the meta-training (**columns 1 and 2**) and the meta-test (**columns 3 and 4**) in Humanoid (**row 1**), Hopper (**row 2**), Swimmer (**row 3**), Car-Circle-Hazard (**row 4**), Point-Botton (**row 5**). The accumulated reward and cost during meta-training are computed on the policy adapted one step from the meta-policy. The black dashed line is the constraint of the accumulated cost (below the line means satisfaction).

D.4 Experimental results on the trade-off between optimality and constraint satisfaction

To investigate the influence of the allowable constraint violation constant δ_{c_i}, in experiments, we conduct the experiments with $\delta_{c_i} = 0.0$, 1.0, 2.0 and 3.0, on two environments, including Half-cheetah and Car-Circle-Hazard. The results are shown in Figure 6.

As stated in Section 5.2, the theoretical result shows a trade-off between the optimality and the safety constraint satisfaction when the allowable constraint violation thresholds δ_{c_i} vary. In particular, when δ_{c_i} is increased, the optimality is improved while the constraint violation increases. This statement is verified by Figure 6. Specifically, especially in Car-Circle-Hazard, when the allowable constraint violation threshold δ_{c_i} varies from 0.0 to 3.0, the performance is improved but the constraint violation is increased in both the meta-training and the meta-test. Therefore, as indicated in both theoretical results in Section 5.2 and the experimental results in Figure 6, we choose a large δ_{c_i} when the

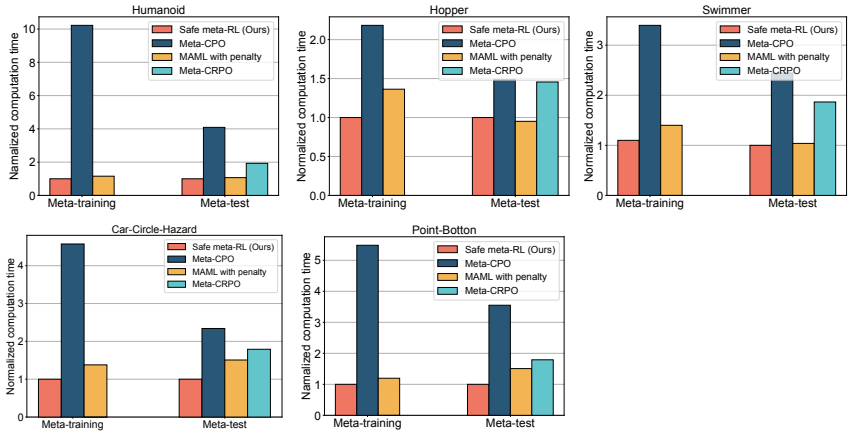


Figure 5: Normalized computation time of the meta-training and the meta-test in Humanoid, Hopper, Swimmer, Car-Circle-Hazard, and Point-Botton.

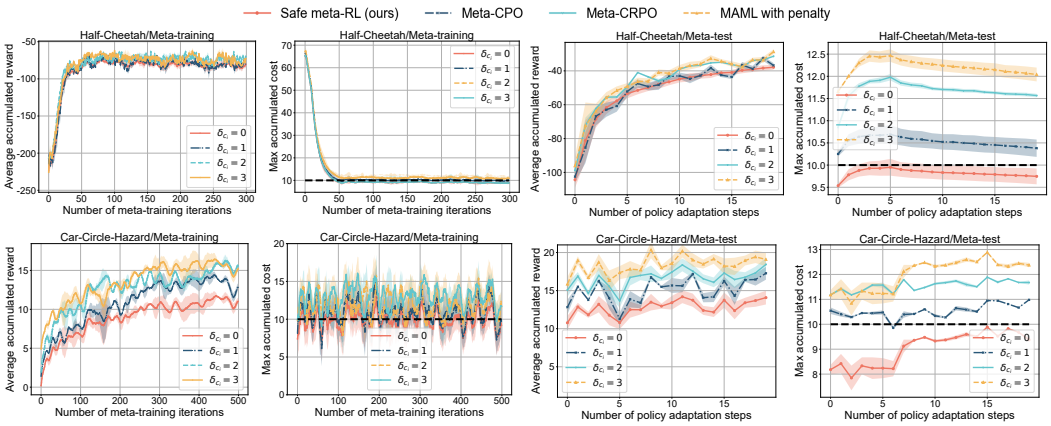


Figure 6: Average accumulated reward (**columns 1 and 3, higher is better**) and maximal accumulated cost (**columns 2 and 4, higher is worse**) across all validation/test tasks during the meta-training (**columns 1 and 2**) and the meta-test (**columns 3 and 4**) in Half-Cheetah (**row 1**) and Car-Circle-Hazard (**row 2**). The accumulated reward and cost during meta-training are computed on the policy adapted one step from the meta-policy. The black dashed line is the constraint of the accumulated cost (below the line means satisfaction).

constraint satisfaction is not required to be strict, and a small $\delta_{c_i} \to 0$ when the constraint satisfaction is prioritized.

D.5 Selection of hyper-parameter

For the hyper-parameter λ, λ_{c_i}, we set $\lambda = \lambda_{c_i}$ and tune them such that, the KL divergence of initial policy π and the adapted policy π' solved from the safe policy adaptation problem (1) is close to 0.03. If the KL divergence is too large, the objective and constraint functions of problem (1) are not good approximations to the accumulated reward/cost functions, as indicated by Lemma 1. If the KL divergence is too small, the policy adaptation step of problem (1) is too small.

E Algorithm Supplement

E.1 An optional algorithm for solving problem (3)

Algorithm 2 states the algorithm for the safe policy adaptation. We apply the projected gradient descent (PGD) to solve the optimization problem (3) to obtain the Lagrangian multipliers $\{u^*_{c_i,\tau}\}_{i=1}^p$, then the closed-form solution of problem (1) is immediately obtained. The gradient of the objective function $\bar{L}(u)$ of problem (3) w.r.t u (used in line 4 of Algorithm 2) can be stated as

$$\nabla_{u_i} \bar{L}(u) = -\mathbb{E}_{s\sim\nu_\tau^{\pi_\phi}}[\mathbb{E}_{a\sim\pi^u(\cdot|s)}[A^{\pi_\phi}_{c_i,\tau}(s,a)] + (1-\gamma)\lambda_{c_i} D_{KL}\left(\pi^u(\cdot|s)\|\pi_\phi(\cdot|s)\right)] + d'_{i,\tau}, \quad (7)$$

where π^u and $d'_{i,\tau}$ are defined in Proposition 2, and then the gradient step is projected to $\mathbb{R}^p_{\geq 0}$. The computation in (7) is derived based on the dual method shown in Proposition 6.1.1 in [8], which is simplified compared with direct computation by the chain rule. The derivation is shown in Appendix F.2.3. As the optimization problem (3) is the dual problem of (1) and is always convex, the PGD method in Algorithm 2 can guarantee convergence to the global optimum [22]. Due to the low dimensionality of the decision variables of problem (3) (the dimension of the Lagrangian multipliers $\{u^*_{c_i,\tau}\}_{i=1}^p$ is the constraint number p) and the simplicity of gradient computation, the computational complexity of Algorithm 2 is much lower than directly solving problem (1).

Algorithm 2 Safe policy adaptation algorithm

Require: Meta-policy π_ϕ; Advantage functions $A_\tau^{\pi_\phi}$ and $A_{c_i,\tau}^{\pi_\phi}$; step size β.
1: $u_i = 0$ for all $i \in 1, \cdots, p$
2: **for** $n = 1, \cdots, N$ **do**
3: Compute $\pi^u(\cdot|s) \propto \exp(f_\phi(s,\cdot) + (\lambda + (1-\gamma)\sum_{i=1}^p u_i\lambda_{c_i})^{-1}(A_\tau^{\pi_\phi}(s,\cdot) - \sum_{i=1}^p u_i A_{c_i,\tau}^{\pi_\phi}(s,\cdot)))$
4: $u_i \leftarrow \max\{0, u_i - \beta\nabla_{u_i}\bar{L}(u)\}$ for each $i = 1, \cdots, p$, where $\nabla_{u_i}L(u)$ is shown in (7)
5: **end for**
6: $u^*_{c_i,\tau} = u_i$ for all $i = 1, \cdots, p$
7: $\pi^\tau(\cdot|s) \propto \exp(f_\phi(s,\cdot) + (\lambda + (1-\gamma)\sum_{i=1}^p u^*_{c_i,\tau}\lambda_{c_i})^{-1}(A_\tau^{\pi_\phi}(s,\cdot) - \sum_{i=1}^p u^*_{c_i,\tau}A_{c_i,\tau}^{\pi_\phi}(s,\cdot)))$
8: Return $\{u^*_{c_i,\tau}\}_{i=1}^p$, π^τ

E.2 An alternative algorithm implementation

When the proposed algorithms are applied to high-dimensional continuous state and action spaces, we provide Algorithms 3 and 4, an alternative algorithm implementation of Algorithms 1 and 2. Compared with Algorithms 1 and 2, Algorithms 3 and 4 avoid approximating $A_\tau^{\pi_{\phi_n}}$ and $A_{c_i,\tau}^{\pi_{\phi_n}}$ during the meta-training, since it is costly to approximate the value functions $V_\tau^{\pi_{\phi_n}}$ and $V_{c_i,\tau}^{\pi_{\phi_n}}$ by neural networks and use GAE [45] to estimate the advantage functions $A_\tau^{\pi_{\phi_n}}$ and $A_{c_i,\tau}^{\pi_{\phi_n}}$ for each sampled task. Instead, Algorithms 3 and 4 only require to approximate $Q_\tau^{\pi_{\phi_n}}$ and $Q_{c_i,\tau}^{\pi_{\phi_n}}$, which can be estimated by Monte-Carlo sampling.

More specifically, in line 3 of Algorithm 3 replace

$$\pi^u(\cdot|s) \propto \exp(f_\phi(s,\cdot) + (\lambda + (1-\gamma)\sum_{i=1}^p u_i\lambda_{c_i})^{-1}(A_\tau^{\pi_\phi}(s,\cdot) - \sum\nolimits_{i=1}^p u_i A_{c_i,\tau}^{\pi_\phi}(s,\cdot)))$$

in line 3 of Algorithm 2 by

$$\pi^u(\cdot|s) \propto \exp(f_\phi(s,\cdot) + (\lambda + (1-\gamma)\sum_{i=1}^p u_i\lambda_{c_i})^{-1}(Q_\tau^{\pi_\phi}(s,\cdot) - \sum\nolimits_{i=1}^p u_i Q_{c_i,\tau}^{\pi_\phi}(s,\cdot))). \quad (8)$$

These two equations are equivalent, where the Q function replaces the A function. Similarly, line 10 of Algorithm 3 is also equivalent to line 7 of Algorithm 2.

Line 11 in Algorithm 4 is equivalent to line 11 of Algorithm 1, where the Q function also replaces the A function. The left problem is how to solve the optimization problem (3) and obtain the the Lagrangian multipliers $u^*_{c_i,\tau}(\pi_{\phi_n})$ only using the Q functions.

We show the solution next. The gradient of the objective function $\bar{L}(u)$ in problem (3) w.r.t u is

$$\nabla_{u_i}\bar{L}(u) = -\mathbb{E}_{s\sim\nu_\tau^{\pi_\phi}}[\mathbb{E}_{a\sim\pi^u(\cdot|s)}[A_{c_i,\tau}^{\pi_\phi}(s,a)] + (1-\gamma)\lambda_{c_i}D_{KL}\left(\pi^u(\cdot|s)\|\pi_\phi(\cdot|s)\right)] + d'_{i,\tau},$$

as shown in (7). Notice that the value of $\nabla_{u_i}\bar{L}(u)$ is the constraint function in the optimization problem (1),

$$-\mathbb{E}_{s\sim\nu_\tau^{\pi_\phi}}[\mathbb{E}_{a\sim\pi(\cdot|s)}[A_{c_i,\tau}^{\pi_\phi}(s,a)] + (1-\gamma)\lambda_{c_i}D_{KL}\left(\pi(\cdot|s)\|\pi_\phi(\cdot|s)\right)] + d'_{i,\tau},$$

when $\pi = \pi^u$. Moreover, the constraint function in problem (1) is already designed as a replacement of $-J_{c_i,\tau}(\pi) + d_{i,\tau} + \delta_{c_i}$ and it is cheaper to compute than $-J_{c_i,\tau}(\pi) + d_{i,\tau} + \delta_{c_i}$ for arbitrary π in problem (1). However, in the problem of approximating $\nabla_{u_i}\bar{L}(u)$, thanks to the derived closed-form

Algorithm 3 Safe policy adaptation algorithm with the first-order approximation

Require: Meta-policy π_ϕ; Advantage functions $Q_\tau^{\pi_\phi}$ and $Q_{c_i,\tau}^{\pi_\phi}$; step size β.
1: $u_i = 0$ for all $i \in 1, \cdots, p$
2: **for** $n = 1, \cdots, N$ **do**
3: Compute $\pi^u(\cdot|s) \propto \exp(f_\phi(s,\cdot) + (\lambda + (1-\gamma)\sum_{i=1}^p u_i\lambda_{c_i})^{-1}(Q_\tau^{\pi_\phi}(s,\cdot) - \sum_{i=1}^p u_i Q_{c_i,\tau}^{\pi_\phi}(s,\cdot)))$
4: **for** $i = 1, \cdots, p$ **do**
5: $u_i \leftarrow \max\{0, u_i - \beta\nabla_{u_i}\bar{L}(u)\}$ where $\nabla_{u_i}L(u)$ is shown in (9)
6: **end for**
7: **end for**
8: $u_{c_i,\tau}^* = u_i$ for all $i = 1, \cdots, p$
9: $\pi^\tau(\cdot|s) \propto \exp(f_\phi(s,\cdot) + (\lambda + (1-\gamma)\sum_{i=1}^p u_{c_i,\tau}^*\lambda_{c_i})^{-1}(Q_\tau^{\pi_\phi}(s,\cdot) - \sum_{i=1}^p u_{c_i,\tau}^* Q_{c_i,\tau}^{\pi_\phi}(s,\cdot)))$
10: Return $\{u_{c_i,\tau}^*\}_{i=1}^p$, π^τ

Algorithm 4 An alternative algorithm of meta-training

Require: Initial meta-policy π_{ϕ_0};
1: **for** $n = 0, \cdots, N$ **do**
2: Sample a task τ with the CMDP $\mathcal{M}_\tau$ from the task distribution $\mathbb{P}(\Gamma)$
3: Evaluate $J_{c_i,\tau}(\pi_{\phi_n})$, $Q_\tau^{\pi_{\phi_n}}(\cdot,\cdot)$ and $Q_{c_i,\tau}^{\pi_{\phi_n}}(\cdot,\cdot)$ for the current meta-policy π_{ϕ_n} on task τ
4: **if** $J_{c_i,\tau}(\pi_{\phi_n}) \le d_{i,\tau} + \delta_{c_i}, \forall i = 1, \cdots, p$ **then**
5: Obtain the task-specific policy π^τ and the Lagrangian multipliers $u_{c_i,\tau}^*(\pi_{\phi_n})$ by Algorithm 3 with the meta-policy π_{ϕ_n}
6: Evaluate $Q_\tau^{\pi^\tau}(\cdot,\cdot)$ for the task-specific policy π^τ on task τ
7: Compute the meta-gradient $\nabla_\phi J_\tau(\pi^\tau) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim\nu_\tau^{\pi^\tau}, a\sim\pi^\tau(\cdot|s)}[\nabla_\phi f_{\phi_n}(s,a)Q_\tau^{\pi^\tau}(s,a)]$
8: Take a step of TRPO [44] with using $\nabla_\phi J_\tau(\pi^\tau)$ towards maximize $J_\tau(\pi^\tau)$ to obtain ϕ_{n+1}
9: **else**
10: Choose any $i_n \in \{1, \cdots, p\}$ such that $J_{C_{i_n}}(\pi_{\phi_n}) > d_{i_n,\tau} + \delta_{c_{i_n}}$
11: Compute the policy gradient $\nabla_\phi J_{C_{i_n},\tau}(\pi_{\phi_n}) \propto \mathbb{E}_{s\sim\nu_\tau^{\pi_{\phi_n}}, a\sim\pi_{\phi_n}(\cdot|s)}[\nabla_\phi f_{\phi_n}(s,a)Q_{C_{i_n},\tau}^{\pi_{\phi_n}}(s,a)]$.
12: Take a step of TRPO [44] with using $\nabla_\phi J_{C_{i_n},\tau}(\pi_{\phi_n})$ towards minimize $J_{C_{i_n},\tau}(\pi_\phi)$ to obtain ϕ_{n+1}
13: **end if**
14: **end for**

π^τ as π^u shown in (8), using the original one $-J_{c_i,\tau}(\pi^u) + d_{i,\tau} + \delta_{c_i}$ becomes cheaper. So, we directly use $-J_{c_i,\tau}(\pi^u) + d_{i,\tau} + \delta_{c_i}$. Therefore, we have

$$\nabla_{u_i}\bar{L}(u) \approx (1-\gamma)(-J_{c_i,\tau}(\pi^u) + d_{i,\tau} + \delta_{c_i}). \tag{9}$$

Next, we use the first-order approximation to approximate $-J_{c_i,\tau}(\pi^u) + d_{i,\tau} + \delta_{c_i}$. Assume the policy π^u is parameterized by π_{θ_u}, then

$$\begin{aligned}
&\frac{1}{1-\gamma}\nabla_{u_i}\bar{L}(u) \approx -J_{c_i,\tau}(\pi^u) + d_{i,\tau} + \delta_{c_i} \\
&\approx -\left(\nabla_\phi^\top J_{c_i,\tau}(\pi_\phi)(\theta_\phi - \phi) + J_{c_i,\tau}(\pi_\phi)\right) + d_{i,\tau} + \delta_{c_i} \\
&= -\frac{1}{1-\gamma}\mathbb{E}_{s\sim\nu_\tau^{\pi_\phi}, a\sim\pi_\phi(\cdot|s)}[\nabla_u^\top \ln\pi_\phi(a|s)Q_{c_i,\tau}^{\pi_\phi}(s,a)](\theta_u - \phi) - J_{c_i,\tau}(\pi_\phi) + d_{i,\tau} + \delta_{c_i}
\end{aligned}$$

Then,

$$\nabla_{u_i}\bar{L}(u) \approx -\mathbb{E}_{s\sim\nu_\tau^{\pi_\phi}, a\sim\pi_\phi(\cdot|s)}[\nabla_u^\top \ln\pi_\phi(a|s)Q_{c_i,\tau}^{\pi_\phi}(s,a)](\theta_u - \phi) + (1-\gamma)(J_{c_i,\tau}(\pi_\phi) - d_{i,\tau} - \delta_{c_i}), \tag{10}$$

In this way, we replace all the estimations of the A function with the estimations of the Q functions, without the requirement of extra data collection.

E.3 Action sampling in algorithm implementation

In Algorithms 2 and 1, we need to sample actions from

$$\pi^u(\cdot|s) \propto \exp(f_\phi(s,\cdot) + \eta^{-1}(Q_\tau^{\pi_\phi}(s,\cdot) - \sum\nolimits_{i=1}^p u_i Q_{c_i,\tau}^{\pi_\phi}(s,\cdot))). \tag{11}$$

When the action space is discrete (no matter whether the state space is discrete or continuous), it is trivial to do the sampling. When the action space is high-dimensional and continuous, it is not easy to do the sampling. Here, we show two solutions. In the implementation of Algorithms 2 and 1, we apply the second solution.

E.3.1 The first solution

Similar to many widely used RL algorithm implementations, such as [44], we also consider the policy parameterized by a Gaussian distribution, i.e.,

$$\pi_\phi(a|s) = \frac{\exp\left(f_\phi(s,a)\right)}{\int_{a'}\exp\left(f_\phi(s,a')\right)da'} = A_1\exp\left(-\frac{(a-g_\phi(s))^2}{2\delta_\phi^2}\right),$$

where $f_\phi = -\frac{(a-g_\phi(s))^2}{2\delta_\phi^2}$ and g_ϕ is a neural network with the input s. So the policy is a softmax policy.

For the policy in (11), we have

$$\pi^u(a|s) = A_2\exp\left(-\frac{(a-g_\phi(s))^2}{2\delta_\phi^2} - \eta^{-1}\frac{(a-g_Q(s))^2}{2\delta_Q^2}\right).$$

Here, $Q_\tau^{\pi_\phi}(s,a) - \sum_{i=1}^p u_i Q_{c_i,\tau}^{\pi_\phi}(s,a)$ is approximated by $-\frac{(a-g_Q(s))^2}{2\delta_Q^2} + C(s)$ where $g_Q(s)$ and $C(s)$ are neural networks with the input s.

Then,

$$\pi^u(a|s) = A_3\exp\left(-\frac{\left(a-(\frac{\eta\delta_Q^2}{\eta\delta_\phi^2+\delta_Q^2}g_\phi(s)+\frac{\delta_\phi^2}{\eta\delta_\phi^2+\delta_Q^2}g_Q(s))\right)^2}{2\frac{\delta_\phi^2\delta_Q^2}{\eta\delta_\phi^2+\delta_Q^2}}\right), \tag{12}$$

i.e., the $\pi^u(a|s)$ is Gaussian with the mean is $\frac{\eta\delta_Q^2}{\eta\delta_\phi^2+\delta_Q^2}g_\phi(s)+\frac{\delta_\phi^2}{\eta\delta_\phi^2+\delta_Q^2}g_Q(s)$ and the standard deviation is $\sqrt{\frac{\delta_\phi^2\delta_Q^2}{\eta\delta_\phi^2+\delta_Q^2}}$. This can be sampled by many code libraries directly.

We can also treat the approximate function $-\frac{(a-g_Q(s))^2}{2\delta_Q^2}$ as $A_\tau^{\pi_\phi}(s,a) - \sum_{i=1}^p u_i A_{c_i,\tau}^{\pi_\phi}(s,a)$ and used in Algorithms (2) and (1), which take $\pi^u(\cdot|s) \propto \exp(f_\phi(s,\cdot)+\eta^{-1}(A_\tau^{\pi_\phi}(s,\cdot)-\sum_{i=1}^p u_i A_{c_i,\tau}^{\pi_\phi}(s,\cdot)))$.

E.3.2 The second solution

In the second solution, we also consider the policy parameterized by a Gaussian distribution, i.e.,

$$\pi_\phi(a|s) = \frac{\exp\left(f_\phi(s,a)\right)}{\int_{a'}\exp\left(f_\phi(s,a')\right)da'} = A_1\exp\left(-\frac{(a-g_\phi(s))^2}{2\delta_\phi^2}\right),$$

where $f_\phi = -\frac{(a-g_\phi(s))^2}{2\delta_\phi^2}$ and g_ϕ is a neural network with the input s.

We use the policy parameterized by θ to approximate the policy $\pi^u(\cdot|s) \propto \exp(f_\phi(s,\cdot) + \eta^{-1}(Q_\tau^{\pi_\phi}(s,\cdot) - \sum_{i=1}^p u_i Q_{c_i,\tau}^{\pi_\phi}(s,\cdot)))$, by minimizing the expected KL-divergence, i.e.,

$$\min_\theta loss(\theta) = \mathbb{E}_{s\sim\nu_\tau^{\pi_\phi}}\left[D_{KL}\left(\pi_\theta(\cdot|s)\,\|\frac{\exp(f_\phi(s,\cdot)+\eta^{-1}(Q_\tau^{\pi_\phi}(s,\cdot)-\sum_{i=1}^p u_i Q_{c_i,\tau}^{\pi_\phi}(s,\cdot)))}{Z_\phi(s)}\right)\right].$$

As shown in [19], the problem is equivalent to $\min_\theta loss(\theta) =$

$$\mathbb{E}_{s\sim\nu_\tau^{\pi_\phi},a\sim\pi_\theta(\cdot|s)}\left[\ln\pi_\theta(a|s) - \left(f_\phi(s,a)+\eta^{-1}(Q_\tau^{\pi_\phi}(s,a)-\sum\nolimits_{i=1}^p u_i Q_{c_i,\tau}^{\pi_\phi}(s,a))\right)\right].$$

This optimization problem can be restated as

$$\min_\theta \mathbb{E}_{s\sim\nu_\tau^{\pi_\phi},a\sim\pi_\phi(\cdot|s)}\left[\frac{\pi_\theta(\cdot|s)}{\pi_\phi(\cdot|s)}\left(\ln\pi_\theta(a|s) - \left(f_\phi(s,a)+\eta^{-1}(Q_\tau^{\pi_\phi}(s,a)-\sum\nolimits_{i=1}^p u_i Q_{c_i,\tau}^{\pi_\phi}(s,a))\right)\right)\right].$$

Therefore, we do not need more data to approximate the expectation $\mathbb{E}_{s\sim\nu_\tau^{\pi_\phi},a\sim\pi_\phi(\cdot|s)}$. Similarly, we can also use π_θ to approximate $\pi^u(\cdot|s) \propto \exp(f_\phi(s,\cdot) + (\lambda + (1-\gamma)\sum_{i=1}^p u_i\lambda_{c_i})^{-1}(A_\tau^{\pi_\phi}(s,\cdot) - \sum_{i=1}^p u_i A_{c_i,\tau}^{\pi_\phi}(s,\cdot)))$.

F Analysis and Proof

F.1 Auxiliary results

Lemma 2 (Policy gradient [47, 3])**.** *Let π_θ be the parameterized policy with the parameter θ. It holds that*

$$\begin{aligned}\nabla_\theta J_\tau(\pi_\theta) =& \frac{1}{1-\gamma}\mathbb{E}_{s\sim\nu_\tau^{\pi_\theta},a\sim\pi_\theta(\cdot|s)}\left[\nabla_\theta \ln \pi_\theta(a|s) Q_\tau^{\pi_\theta}(s,a)\right] \\ =& \frac{1}{1-\gamma}\mathbb{E}_{s\sim\nu_\tau^{\pi_\theta},a\sim\pi_\theta(\cdot|s)}\left[\nabla_\theta \ln \pi_\theta(a|s) A_\tau^{\pi_\theta}(s,a)\right].\end{aligned}$$

Lemma 3 (Policy gradient of the softmax policy)**.** *For the softmax policy π_θ as $\pi_\theta(a|s) = \frac{\exp(f_\theta(s,a))}{\sum_{a'\in\mathcal{A}}\exp(f_\theta(s,a'))}$ (in discrete action space $\mathcal{A}$) or $\pi_\theta(a|s) \triangleq \frac{\exp(f_\theta(s,a))}{\int_{\mathcal{A}}\exp(f_\theta(s,a'))da'}$ (in continuous action space $\mathcal{A}$), $\forall(s,a)\in\mathcal{S}\times\mathcal{A}$. It holds that*

$$\nabla_\theta J_\tau(\pi_\theta) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim\nu_\tau^{\pi_\theta},a\sim\pi_\theta(\cdot|s)}\left[\nabla_\theta f_\theta(s,a) A_\tau^{\pi_\theta}(s,a)\right]. \tag{13}$$

Proof. We prove it under the discrete action space $\mathcal{A}$. The proof under the continuous action space $\mathcal{A}$ is similar.

From Lemma 2, we have

$$\begin{aligned}&\nabla_\theta J_\tau(\pi_\theta) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim\nu_\tau^{\pi_\theta},a\sim\pi_\theta(\cdot|s)}\left[\nabla_\theta \ln \pi_\theta(a|s) A_\tau^{\pi_\theta}(s,a)\right] \\ &= \frac{1}{1-\gamma}\mathbb{E}_{s\sim\nu_\tau^{\pi_\theta},a\sim\pi_\theta(\cdot|s)}\left[\nabla_\theta \ln\left(\frac{\exp(f_\theta(s,a))}{\sum_{a'\in\mathcal{A}}\exp(f_\theta(s,a'))}\right) A_\tau^{\pi_\theta}(s,a)\right] \\ &= \frac{1}{1-\gamma}\mathbb{E}_{s\sim\nu_\tau^{\pi_\theta},a\sim\pi_\theta(\cdot|s)}\left[\nabla_\theta f_\theta(s,a) - \nabla_\theta \ln\left(\sum_{a'\in\mathcal{A}}\exp(f_\theta(s,a'))\right) A_\tau^{\pi_\theta}(s,a)\right]\end{aligned}$$

Here, $\nabla_\theta \ln\left(\sum_{a'\in\mathcal{A}}\exp(f_\theta(s,a'))\right)$ is independent with a, then $\nabla_\theta J_\tau(\pi_\theta)$

$$\begin{aligned}=& \frac{1}{1-\gamma}\mathbb{E}_{s\sim\nu_\tau^{\pi_\theta},a\sim\pi_\theta(\cdot|s)}\left[\nabla_\theta f_\theta(s,a) - \nabla_\theta \ln\left(\sum_{a'\in\mathcal{A}}\exp(f_\theta(s,a'))\right) A_\tau^{\pi_\theta}(s,a)\right] \\ =& \frac{1}{1-\gamma}\mathbb{E}_{s\sim\nu_\tau^{\pi_\theta},a\sim\pi_\theta(\cdot|s)}\left[\nabla_\theta f_\theta(s,a) A_\tau^{\pi_\theta}(s,a)\right] - \\ & \frac{1}{1-\gamma}\mathbb{E}_{s\sim\nu_\tau^{\pi_\theta}}\left[\nabla_\theta \ln\left(\sum_{a'\in\mathcal{A}}\exp(f_\theta(s,a'))\right)\mathbb{E}_{a\sim\pi_\theta(\cdot|s)} A_\tau^{\pi_\theta}(s,a)\right].\end{aligned}$$

Since $\mathbb{E}_{a\sim\pi_\theta(\cdot|s)} A_\tau^{\pi_\theta}(s,a) = \mathbb{E}_{a\sim\pi_\theta(\cdot|s)}[Q_\tau^{\pi_\theta}(s,a)] - V_\tau^{\pi_\theta}(s) = 0$. Then,

$$\nabla_\theta J_\tau(\pi_\theta) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim\nu_\tau^{\pi_\theta},a\sim\pi_\theta(\cdot|s)}\left[\nabla_\theta f_\theta(s,a) A_\tau^{\pi_\theta}(s,a)\right].$$

□

F.2 Proofs of closed-form solution of safe policy adaptation

F.2.1 Proof of Proposition 1

We provide the complete statement of Proposition 1 as the following Proposition 5.

Proposition 5. *When the softmax policy π_ϕ satisfies $J_{c_i,\tau}(\pi_\phi) \leq d_{i,\tau} + \delta_{c_i}, \forall i = 1, \cdots, p$, the solution π^τ of the optimization problem* (1) *exists. Suppose an appropriate constraint qualification (to be stipulated) holds at π^τ, there exists $\{u^*_{c_i,\tau}\}_{i=1}^p$ with $u^*_{c_i,\tau} \geq 0$, such that*

$$\pi^\tau(\cdot|s) \propto \exp\left(f_\phi(s,\cdot) + \eta^{-1}(A^{\pi_\phi}_\tau(s,\cdot) - \sum_{i=1}^p u^*_{c_i,\tau} A^{\pi_\phi}_{c_i,\tau}(s,\cdot))\right),\ \forall s \in \mathcal{S},$$

i.e.,

$$\pi^\tau(a|s) = \frac{\exp\left(f_\phi(s,a) + \left(\lambda + \sum_{i=1}^p u^*_{c_i,\tau}\lambda_{c_i}\right)^{-1}\left(A^{\pi_\phi}_\tau(s,a) - \sum_{i=1}^p u^*_{c_i,\tau} A^{\pi_\phi}_{c_i,\tau}(s,a)\right)\right)}{\sum_{a\in\mathcal{A}}\exp\left(f_\phi(s,a') + \eta^{-1}\left(A^{\pi_\phi}_\tau(s,a') - \sum_{i=1}^p u^*_{c_i,\tau} A^{\pi_\phi}_{c_i,\tau}(s,a')\right)\right)},$$

in discrete action space $\mathcal{A}$, or

$$\pi^\tau(a|s) = \frac{\exp\left(f_\phi(s,a) + \left(\lambda + \sum_{i=1}^p u^*_{c_i,\tau}\lambda_{c_i}\right)^{-1}\left(A^{\pi_\phi}_\tau(s,a) - \sum_{i=1}^p u^*_{c_i,\tau} A^{\pi_\phi}_{c_i,\tau}(s,a)\right)\right)}{\int_{a'}\exp\left(f_\phi(s,a') + \eta^{-1}\left(A^{\pi_\phi}_\tau(s,a') - \sum_{i=1}^p u^*_{c_i,\tau} A^{\pi_\phi}_{c_i,\tau}(s,a')\right)\right)da'},$$

in continuous action space $\mathcal{A}$, where $\eta = (1-\gamma)\lambda + \sum_{i=1}^p u^*_{c_i,\tau}\lambda_{c_i}$.

There are many constraint qualifications where each of them assures the validity of the proposition, including but not limited to Mangasarian-Fromovitz constraint qualification (MFCQ), linear independence constraint qualification (LICQ), and Slater's condition (SC) [17]. Refer to [40] for more validated constraint qualifications.

The assumption that one constraint qualification holds at π^τ is mild. For example, if there exists a policy π such that $\forall i$

$$J_{c_i,\tau}(\pi_\phi) + \mathop{\mathbb{E}}_{\substack{s\sim\nu^{\pi_\phi}_\tau \\ a\sim\pi(\cdot|s)}}\left[\frac{A^{\pi_\phi}_{c_i,\tau}(s,a)}{1-\gamma}\right] + \lambda_{c_i}\,\mathbb{E}_{s\sim\nu^{\pi_\phi}_\tau}\left[D_{KL}\left(\pi(\cdot|s)\|\pi_\phi(\cdot|s)\right)\right] < d_{i,\tau} + \delta_{c_i}, \tag{14}$$

then the Slater's condition holds. Note that when $\pi = \pi_\phi$, we have $J_{c_i,\tau}(\pi_\phi) + \mathop{\mathbb{E}}_{\substack{s\sim\nu^{\pi_\phi}_\tau \\ a\sim\pi(\cdot|s)}}\left[\frac{A^{\pi_\phi}_{c_i,\tau}(s,a)}{1-\gamma}\right] + \lambda_{c_i}\,\mathbb{E}_{s\sim\nu^{\pi_\phi}_\tau}\left[D_{KL}\left(\pi(\cdot|s)\|\pi_\phi(\cdot|s)\right)\right] \leq d_{i,\tau} + \delta_{c_i}$. It usually exists a π near π_ϕ such that (14) holds or the π_ϕ itself can assure (14) holds. Next, we prove the proposition.

Proofs of Proposition 5. The optimization problem (1) can be restated as

$$\begin{aligned}
&\mathop{\text{argmin}}_{\pi\in\Pi} \ -\mathop{\mathbb{E}}_{\substack{s\sim\nu^{\pi_\phi}_\tau \\ a\sim\pi(\cdot|s)}}\left[A^{\pi_\phi}_\tau(s,a)\right] + \lambda\,\mathbb{E}_{s\sim\nu^{\pi_\phi}_\tau}\left[D_{KL}\left(\pi(\cdot|s)\|\pi_\phi(\cdot|s)\right)\right], \\
&\text{s.t.} \mathop{\mathbb{E}}_{\substack{s\sim\nu^{\pi_\phi}_\tau \\ a\sim\pi(\cdot|s)}}\left[A^{\pi_\phi}_{c_i,\tau}(s,a)\right] + \lambda'_{c_i}\,\mathbb{E}_{s\sim\nu^{\pi_\phi}_\tau}\left[D_{KL}\left(\pi(\cdot|s)\|\pi_\phi(\cdot|s)\right)\right] \leq d'_{i,\tau},\ i = 1,\cdots,p,
\end{aligned}$$

where the constants $\lambda'_{c_i} \triangleq (1-\gamma)\lambda_{c_i}$, and $d'_{i,\tau} \triangleq (1-\gamma)(d_{i,\tau} + \delta_{c_i} - J_{c_i,\tau}(\pi_\phi))$.

First, we consider the discrete state-action space $\mathcal{S}\times\mathcal{A}$. Considering the probability at each state-action pair $\pi(a|s)$ as the decision variable, the minimization is taken over the probability simplex $\left\{\pi(\cdot|s) : 0 \leq \pi(a|s) \leq 1, \sum_{a\in\mathcal{A}}\pi(a|s) = 1\right\}$. Then the optimization problem is formally stated as

$$\begin{aligned}
&\mathop{\text{argmin}}_{\pi}\ \mathbb{E}_{s\sim\nu^{\pi_\phi}_\tau}\left[\sum_{a\in\mathcal{A}} -\pi(a|s)A^{\pi_\phi}_\tau(s,a) + \lambda D_{KL}\left(\pi(\cdot|s)\|\pi_\phi(\cdot|s)\right)\right], \\
&\text{s.t. } \mathbb{E}_{s\sim\nu^{\pi_\phi}_\tau}\left[\sum_{a\in\mathcal{A}}\pi(a|s)A^{\pi_\phi}_{c_i,\tau}(s,a) + \lambda'_{c_i}D_{KL}\left(\pi(\cdot|s)\|\pi_\phi(\cdot|s)\right)\right] \leq d'_{i,\tau},\ i = 1,\cdots,p, \\
&\quad\sum_{a\in\mathcal{A}}\pi(a|s) = 1 \text{ for any } s\in\mathcal{S}, \\
&\quad\pi(a|s) \leq 1 \text{ for any } a\in\mathcal{A}, s\in\mathcal{S}, \\
&\quad-\pi(a|s) \leq 0 \text{ for any } a\in\mathcal{A}, s\in\mathcal{S}.
\end{aligned} \tag{15}$$

Since $\pi_\phi \in \Pi_\tau^C$, we have $d'_{i,\tau} = (1-\gamma)(d_{i,\tau} + \delta_{c_i} - J_{c_i,\tau}(\pi_\phi)) \geq 0$, the solution of (15) exists.

According to Theorem 1 in [17] and theorems in [8, 9], since the constraint qualification holds, the Karush-Kuhn-Tucker (KKT) conditions hold at π^τ, i.e., there exists Lagrangian multipliers $\{u^*_{c_i,\tau}\}_{i=1}^p$, $u_0^*(s)$ for all $s \in \mathcal{S}$, $u_1^*(s,a)$ and $u_2^*(s,a)$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, such that

$$u^*_{c_i,\tau} \geq 0, \forall i = 1, \cdots, p,$$

$$u_1^*(s,a) \geq 0, u_2^*(s,a) \geq 0, \ \forall (s,a) \in \mathcal{S} \times \mathcal{A}, \tag{16}$$

$$\mathbb{E}_{s\sim\nu_\tau^{\pi_\phi}}\left[\sum_{a\in\mathcal{A}} \pi^\tau(a|s) A^{\pi_\phi}_{c_i,\tau}(s,a) + \lambda'_{c_i} D_{KL}\left(\pi^\tau(\cdot|s)\|\pi_\phi(\cdot|s)\right)\right] - d'_{i,\tau} \leq 0, \ \forall i = 1, \cdots, p,$$

$$\pi^\tau(s,a) \geq 0, \pi^\tau(s,a) \leq 1, \forall (s,a) \in \mathcal{S} \times \mathcal{A}, \tag{17}$$

$$\sum_{a\in\mathcal{A}} \pi^\tau(a|s) = 1, \ \forall s \in \mathcal{S}, \tag{18}$$

$$u^*_{c_i,\tau}\left(\mathbb{E}_{s\sim\nu_\tau^{\pi_\phi}}\left[\sum_{a\in\mathcal{A}} \pi^\tau(a|s) A^{\pi_\phi}_{c_i,\tau}(s,a) + \lambda'_{c_i} D_{KL}\left(\pi^\tau(\cdot|s)\|\pi_\phi(\cdot|s)\right)\right] - d'_{i,\tau}\right) = 0,$$

$$u_1^*(s,a)(\pi^\tau(s,a) - 1) = 0, \forall (s,a) \in \mathcal{S} \times \mathcal{A}, \tag{19}$$

$$-u_2^*(s,a)\pi^\tau(s,a) = 0, \forall (s,a) \in \mathcal{S} \times \mathcal{A}, \tag{20}$$

$$\nabla_\pi L(\pi^\tau, \{u^*_{c_i,\tau}\}_{i=1}^p, u_0^*, u_1^*, u_2^*) = 0, \tag{21}$$

where

$$\begin{aligned} L(\pi, \{u^*_{c_i,\tau}\}_{i=1}^p, u_0^*, u_1^*, u_2^*)) &\triangleq \mathbb{E}_{s\sim\nu_\tau^{\pi_\phi}}\left[\sum_{a\in\mathcal{A}} -\pi(a|s) A_\tau^{\pi_\phi}(s,a) + \lambda D_{KL}\left(\pi(\cdot|s)\|\pi_\phi(\cdot|s)\right)\right] \\ &+ \sum_{i=1}^p u^*_{c_i,\tau}\left(\mathbb{E}_{s\sim\nu_\tau^{\pi_\phi}}\left[\sum_{a\in\mathcal{A}} \pi(a|s) A^{\pi_\phi}_{c_i,\tau}(s,a) + \lambda'_{c_i} D_{KL}\left(\pi(\cdot|s)\|\pi_\phi(\cdot|s)\right)\right] - d'_{i,\tau}\right) \\ &+ \sum_{s\in\mathcal{S}} u_0^*(s)\left(\sum_{a\in\mathcal{A}} \pi(a|s) - 1\right) + \sum_{s\in\mathcal{S}}\sum_{s\in\mathcal{S}} u_1^*(s,a)(\pi(s,a) - 1) - u_2^*(s,a)\pi(s,a). \end{aligned} \tag{22}$$

Note that (16) (17) (18) (19) (20)(21) constitute the KKT condition for the following optimization problem:

$$\begin{aligned} \operatorname*{argmin}_\pi \ & \mathbb{E}_{s\sim\nu_\tau^{\pi_\phi}}\left[\sum_{a\in\mathcal{A}} \pi(a|s)\left(-A_\tau^{\pi_\phi}(s,a) + \sum_{i=1}^p u^*_{c_i,\tau} A^{\pi_\phi}_{c_i,\tau}(s,a)\right)\right. \\ & \left. + \left(\lambda + \sum_{i=1}^p u^*_{c_i,\tau}\lambda'_{c_i}\right) D_{KL}\left(\pi(\cdot|s)\|\pi_\phi(\cdot|s)\right)\right] - \sum_{i=1}^p u^*_{c_i,\tau} d'_{i,\tau} \\ \text{s.t.} \ & \sum_{a\in\mathcal{A}} \pi(a|s) = 1 \text{ for any } s \in \mathcal{S}, \\ & \pi(a|s) \leq 1 \text{ for any } a \in \mathcal{A}, s \in \mathcal{S}, \\ & -\pi(a|s) \leq 0 \text{ for any } a \in \mathcal{A}, s \in \mathcal{S}. \end{aligned} \tag{23}$$

i.e., the KKT condition for the optimization problem (23) holds at π^τ with Lagrangian multipliers $u_0^*(s)$, $u_1^*(s,a)$ and $u_2^*(s,a)$. Here, $\{u^*_{c_i,\tau}\}_{i=1}^p$ are constants for the problem.

Since the terms $-\mathbb{E}_{s\sim\nu_\tau^{\pi_\phi}}\left[\sum_{a\in\mathcal{A}} \pi(a|s) A_\tau^{\pi_\phi}(s,a)\right]$ and $\mathbb{E}_{s\sim\nu_\tau^{\pi_\phi}}\left[\sum_{a\in\mathcal{A}} \pi(a|s) A^{\pi_\phi}_{c_i,\tau}(s,a)\right]$ are linear; the term $\mathbb{E}_{s\sim\nu_\tau^{\pi_\phi}}\left[D_{KL}\left(\pi(\cdot|s)\|\pi_\phi(\cdot|s)\right)\right]$ is convex, the optimization problem (23) is convex. Moreover, since all the constraint functions are affine, the Slater's condition holds naturally for the optimization problem (23), as shown in [9]. Therefore, the strong duality holds. Then, π^τ is the optimal solution for (23).

In (23), we can omit the term $-\sum_{i=1}^{p} u^*_{c_i,\tau} d'_{i,\tau}$ and keep the solution unchanged. Next, we borrow the conclusion of Proposition 3.1 in [29], we have $\pi^\tau(a|s) =$

$$\frac{\exp\left(f_\phi(s,a) + \left(\lambda + \sum_{i=1}^{p} u^*_{c_i,\tau}\lambda'_{c_i}\right)^{-1}\left(A^{\pi_\phi}_\tau(s,a) - \sum_{i=1}^{p} u^*_{c_i,\tau} A^{\pi_\phi}_{c_i,\tau}(s,a)\right)\right)}{\sum_{a\in\mathcal{A}}\exp\left(f_\phi(s,a') + \left(\lambda + \sum_{i=1}^{p} u^*_{c_i,\tau}\lambda'_{c_i}\right)^{-1}\left(A^{\pi_\phi}_\tau(s,a') - \sum_{i=1}^{p} u^*_{c_i,\tau} A^{\pi_\phi}_{c_i,\tau}(s,a')\right)\right)},$$

i.e.,

$$\pi^\tau(\cdot|s) \propto \exp\left(f_\phi(s,\cdot) + (\lambda + \sum_{i=1}^{p} u^*_{c_i,\tau}\lambda'_{c_i})^{-1}(A^{\pi_\phi}_\tau(s,\cdot) - \sum_{i=1}^{p} u^*_{c_i,\tau} A^{\pi_\phi}_{c_i,\tau}(s,\cdot))\right),$$

for all $s \in \mathcal{S}$. Since $\lambda'_{c_i} = (1-\gamma)\lambda_{c_i}$, the proof is done.

□

F.2.2 Proof of Proposition 2

Proof of Proposition 2. For the Lagrangian multiplier variables u, u_0, u_1, u_2, we denote the solution of $\min_\pi L(\pi, u, u_0, u_1, u_2)$ as $\pi^{\{u,u_0,u_1,u_2\}}$ (L is shown in (22)), i.e.,

$$\pi^{\{u,u_0,u_1,u_2\}} = \arg\min_\pi L(\pi, u, u_0, u_1, u_2).$$

From the proof of Proposition 1, we have the strong duality for the optimization problem (15) holds. Then, we have $\{u^*, u_0^*, u_1^*, u_2^*\} =$

$$\arg\max_{\{u,u_0,u_1,u_2\}} L(\pi^{\{u,u_0,u_1,u_2\}}, u, u_0, u_1, u_2), \text{ s.t. } u \geq 0, u_1 \geq 0, u_2 \geq 0. \tag{24}$$

Next, from the above optimization problem, we set u_0, u_1, u_2 as $u_0^*(u)$, $u_1^*(u)$, $u_2^*(u)$ in (24), where $u_0^*(u)$, $u_1^*(u)$, $u_2^*(u)$ are the solution of dual variable (Lagrangian multiplier solution) of the following problem:

$$\begin{aligned}
\operatorname*{argmin}_\pi \mathbb{E}_{s\sim\nu^{\pi_\phi}_\tau}\Bigg[&\sum_{a\in\mathcal{A}}\pi(a|s)\left(-A^{\pi_\phi}_\tau(s,a) + \sum_{i=1}^{p} u_i A^{\pi_\phi}_{c_i,\tau}(s,a)\right) \\
&+ \left(\lambda + (1-\gamma)\sum_{i=1}^{p} u_i\lambda_{c_i}\right) D_{KL}\left(\pi(\cdot|s)\|\pi_\phi(\cdot|s)\right)\Bigg] - \sum_{i=1}^{p} u_i d'_{i,\tau} \\
\text{s.t. } &\sum_{a\in\mathcal{A}}\pi(a|s) = 1 \text{ for any } s \in \mathcal{S}, \\
&\pi(a|s) \leq 1 \text{ for any } a \in \mathcal{A}, s \in \mathcal{S}, \\
&-\pi(a|s) \leq 0 \text{ for any } a \in \mathcal{A}, s \in \mathcal{S}.
\end{aligned} \tag{25}$$

We have

$$u^* = \arg\max_u L(\pi^{\{u,u_0^*(u),u_1^*(u),u_2^*(u)\}}, u, u_0^*(u), u_1^*(u), u_2^*(u)), \text{ s.t. } u \geq 0. \tag{26}$$

Similar to solution of (23), we have the solution of (25) is π^u, where $\pi^u(\cdot|s) \propto \exp(f_\phi(s,\cdot) + (\sum_{i=1}^{p} u_i\lambda_{c_i})^{-1}(A^{\pi_\phi}_\tau(s,\cdot) - \sum_{i=1}^{p} u_i A^{\pi_\phi}_{c_i,\tau}(s,\cdot)))$. Moreover, from the strong duality of the optimization problem (25) (linear inequality constraints), we have

$$\pi^{\{u,u_0^*(u),u_1^*(u),u_2^*(u)\}} = \arg\min_\pi L(\pi, u, u_0^*(u), u_1^*(u), u_2^*(u)) = \pi^u. \tag{27}$$

Therefore,

$$u^* = \arg\max_u L(\pi^u, u, u_0^*(u), u_1^*(u), u_2^*(u)), \text{ s.t. } u \geq 0.$$

Moreover, we know

$$\sum_{s\in\mathcal{S}} u_0^*(u)(s)\left(\sum_{a\in\mathcal{A}}\pi^u(a|s) - 1\right) + \sum_{s\in\mathcal{S}}\sum_{s\in\mathcal{S}} u_1^*(u)(s,a)(\pi^u(s,a) - 1) - u_2^*(u)(s,a)\pi^u(s,a) = 0.$$

Form (26) and (22), we have

$$u^* = \max_u \mathbb{E}_{s\sim\nu_\tau^{\pi_\phi}, a\sim\pi^u(\cdot|s)}[-A_\tau^{\pi_\phi}(s,a) + \sum_{i=1}^p u_i A_{c_i,\tau}^{\pi_\phi}(s,a)] + (\lambda + \sum_{i=1}^p u_i \lambda'_{c_i})$$
$$\mathbb{E}_{s\sim\nu_\tau^{\pi_\phi}}\left[D_{KL}\left(\pi^u(\cdot|s)\|\pi_\phi(\cdot|s)\right)\right] - \sum_{i=1}^p u_i(1-\gamma)(d_{i,\tau} + \delta_{c_i} - J_{c_i,\tau}(\pi_\phi))$$
$$\text{s.t. } u_i \geq 0,\ \forall i = 1, \cdots, p.$$

Then, the proof is done. □

F.2.3 Deviation of gradient w.r.t. the dual variables

We derive the gradient of $\bar{L}$ w.r.t. the dual variables u for (7). Let

$$\hat{L}(u,\pi^u) \triangleq \mathop{\mathbb{E}}_{\substack{s\sim\nu_\tau^{\pi_\phi}\\ a\sim\pi^u}}[A_\tau^{\pi_\phi}(s,a) - \sum_{i=1}^p u_i A_{c_i,\tau}^{\pi_\phi}(s,a)]$$
$$-(\lambda + (1-\gamma)\sum_{i=1}^p u_i\lambda_{c_i})\,\mathbb{E}_{s\sim\nu_\tau^{\pi_\phi}}\left[D_{KL}\left(\pi^u(\cdot|s)\|\pi_\phi(\cdot|s)\right)\right] + \sum_{i=1}^p u_i d'_{i,\tau}$$

where $d'_{i,\tau} \triangleq (1-\gamma)(d_{i,\tau} + \delta_{c_i} - J_{c_i,\tau}(\pi_\phi))$. Then,

$$\nabla_u \bar{L}(u) = \nabla_1 \hat{L}(u,\pi^u) + \nabla_u \pi^u \nabla_2 \hat{L}(u,\pi^u)$$

Consider $\nabla_2 \hat{L}(u,\pi^u)$. From (27), we have

$$\pi^{\{u,u_0^*(u),u_1^*(u),u_2^*(u)\}} = \arg\min_\pi L(\pi,u,u_0^*(u),u_1^*(u),u_2^*(u)) = \pi^u$$

where L is shown in (22) and $u_0^*(u)$, $u_1^*(u)$, $u_2^*(u)$ are the solution of dual variable of (25). Then

$$\nabla_1 L(\pi^u,u,u_0^*(u),u_1^*(u),u_2^*(u)) = 0.$$

Moreover, we know

$$\sum_{s\in\mathcal{S}} u_0^*(u)(s)\left(\sum_{a\in\mathcal{A}}\pi^u(a|s) - 1\right) + \sum_{s\in\mathcal{S}}\sum_{s\in\mathcal{S}} u_1^*(u)(s,a)(\pi^u(s,a)-1) - u_2^*(u)(s,a)\pi^u(s,a) = 0.$$

Thus,

$$\nabla_2 \hat{L}(u,\pi^u) = \nabla_1 L(\pi^u,u,u_0^*(u),u_1^*(u),u_2^*(u)) = 0.$$

Then, we have

$$\nabla_u \bar{L}(u) = \nabla_1 \hat{L}(u,\pi^u).$$

Therefore,

$$\nabla_{u_i}\bar{L}(u) = -\mathbb{E}_{s\sim\nu_\tau^{\pi_\phi}}[\mathbb{E}_{a\sim\pi^u(\cdot|s)}[A_{c_i,\tau}^{\pi_\phi}(s,a)] + (1-\gamma)\lambda_{c_i} D_{KL}\left(\pi^u(\cdot|s)\|\pi_\phi(\cdot|s)\right)] + d'_{i,\tau}.$$

F.3 Meta-Gradient

F.3.1 Computation of meta-gradient

Proposition 6. *Let $\pi^\tau = \mathcal{A}^s(\pi_\phi,\Lambda,\Delta,\tau)$. Suppose all the assumptions in Proposition* (5) *hold. Suppose the LICQ and the strict complementary slackness condition (SCSC) [17, 52] for the optimization problem* (3.1) *holds at π^τ. Then, $\nabla_\phi J_\tau(\pi^\tau)$ exists and*

$$\nabla_\phi J_\tau(\pi^\tau) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim\nu_\tau^{\pi^\tau},a\sim\pi^\tau(\cdot|s)}[\big(\nabla_\phi\eta(\pi_\phi)^{-1}\bar{Q}_\tau^{\pi_\phi}(s,a)$$
$$+\eta(\pi_\phi)^{-1}\nabla_\phi\bar{Q}_\tau^{\pi_\phi}(s,a) + \nabla_\phi f_\phi(s,a)\big)\,Q_\tau^{\pi^\tau}(s,a)],$$

where $\eta(\pi_\phi) \triangleq \lambda + (1-\gamma)\sum_{i=1}^p u_{c_i,\tau}^*(\pi_\phi)\lambda_{c_i}$, and $\bar{Q}_\tau^{\pi_\phi} \triangleq Q_\tau^{\pi_\phi} - \sum_{i=1}^p u_{c_i,\tau}^*(\pi_\phi)Q_{c_i,\tau}^{\pi_\phi}$.

Proof. For any meta-policy π_ϕ, the objective function of the optimization problem (3.1) is strongly concave and the constraint function is convex. The LICQ and the SCSC hold at π^τ. According to Theorem 2 in [52], $\nabla_\phi J_\tau(\pi^\tau)$ exists.

We have

$$\pi^\tau(\cdot|s) \propto \exp(f_\phi(s,\cdot) + \eta(\pi_\phi)^{-1}(A_\tau^{\pi_\phi}(s,\cdot) - \sum\nolimits_{i=1}^{p} u_{c_i,\tau}^* A_{c_i,\tau}^{\pi_\phi}(s,\cdot)))$$

is equivalent to

$$\pi^\tau(\cdot|s) \propto \exp(f_\phi(s,\cdot) + \eta(\pi_\phi)^{-1}(Q_\tau^{\pi_\phi}(s,\cdot) - \sum\nolimits_{i=1}^{p} u_{c_i,\tau}^* Q_{c_i,\tau}^{\pi_\phi}(s,\cdot))).$$

From Lemma 3, we have

$$\begin{aligned}\nabla_\phi J_\tau(\pi^\tau) =& \frac{1}{1-\gamma}\mathbb{E}_{s\sim\nu_\tau^{\pi^\tau}, a\sim\pi^\tau(\cdot|s)}[\nabla_\phi \left(\eta(\pi_\phi)^{-1}\bar{Q}_\tau^{\pi_\phi}(s,a) + f_\phi(s,a)\right) Q_\tau^{\pi^\tau}(s,a)] \\ =& \frac{1}{1-\gamma}\mathbb{E}_{s\sim\nu_\tau^{\pi^\tau}, a\sim\pi^\tau(\cdot|s)}[\big(\nabla_\phi\eta(\pi_\phi)^{-1}\bar{Q}_\tau^{\pi_\phi}(s,a) \\ & +\eta(\pi_\phi)^{-1}\nabla_\phi\bar{Q}_\tau^{\pi_\phi}(s,a) + \nabla_\phi f_\phi(s,a)\big) Q_\tau^{\pi^\tau}(s,a)].\end{aligned}$$

□

F.3.2 Computation of $\nabla_\phi Q_\tau^{\pi_\phi}(s,a)$

We have

$$\nabla_\phi Q_\tau^{\pi_\phi}(s,a) = \frac{\gamma}{1-\gamma}\cdot\mathbb{E}_{(s',a')\sim\sigma_{\tau,\pi_\phi}^{(s,a)}}\left[\nabla_\phi f_\phi\left(s',a'\right) Q_\tau^{\pi_\phi}\left(s',a'\right)\right]. \tag{28}$$

where the state-action visitation probability $\sigma_{\tau,\pi_\theta}^{(s,a)}$ initialized at $(s,a)\in\mathcal{S}\times\mathcal{A}$ is defined by

$$\sigma_{\tau,\pi_\phi}^{(s,a)}\left(s',a'\right) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t\mathbb{P}\left(s_t=s', a_t=a'|\pi_\phi, s_0\sim P_\tau(\cdot|s,a)\right).$$

Proof. As shown in [50],

$$\begin{aligned}\nabla_\phi Q_\tau^{\pi_\phi}(s,a) &= \nabla_\phi\left((1-\gamma)\cdot r_\tau(s,a) + \gamma\cdot\mathbb{E}_{s'\sim P_\tau(\cdot|s,a)}\left[V_\tau^{\pi_\phi}\left(s'\right)\right]\right) \\ &= \frac{\gamma}{1-\gamma}\cdot\mathbb{E}_{(s',a')\sim\sigma_{\tau,\pi_\phi}^{(s,a)}}\left[\nabla_\phi\ln\pi_\phi\left(a'|s'\right)\cdot Q_\tau^{\pi_\phi}\left(s',a'\right)\right].\end{aligned}$$

By Lemma 3, from (13), we can obtain (28). □

F.3.3 Gradient of Lagrangian multipliers

We show the existence and the computation of $\nabla_\phi u_{c_i,\tau}^*(\pi_\phi)$ in the following proposition.

Proposition 7. *Let $\pi^\tau = \mathcal{A}^s(\pi_\phi,\Lambda,\Delta,\tau)$. Suppose all the assumptions in Proposition* (5) *hold. Suppose the LICQ and the strict complementary slackness condition (SCSC) [17, 52] for the optimization problem* (3.1) *holds at π^τ. Then, the Lagrangian multipliers $u_{c_i,\tau}^*(\pi_\phi)$ is unique for any given π_ϕ, $\nabla_\phi u_{c_i,\tau}^*(\pi_\phi)$ exists. For $i\in\{1,\cdots,p\}$, if $u_{c_i,\tau}^*(\pi_\phi)=0$, then $\nabla_\phi u_{c_i,\tau}^*(\pi_\phi)=0$. Let $\bar{u}_{c_i,\tau}^*(\pi_\phi)$ be the vector includes all all $i\in\{1,\cdots,p\}$ with $u_{c_i,\tau}^*(\pi_\phi)>0$,*

$$\nabla_\phi u_{c_i,\tau}^*(\pi_\phi) = -\nabla_\phi\nabla_{\bar{u}}\hat{L}(\bar{u},\phi)\nabla_{\bar{u}}^2\hat{L}(\bar{u},\phi)^{-1}$$

where $\hat{L}(\bar{u},\phi) = \mathbb{E}[A_\tau^{\pi_\phi}(s,a) - \sum_{i=1}^{p} u_i A_{c_i,\tau}^{\pi_\phi}(s,a)] - \eta^u\ \mathbb{E}_{s\sim\nu_\tau^{\pi_\phi}}\left[D_{KL}\left(\pi^u(\cdot|s)\|\pi_\phi(\cdot|s)\right)\right] + \sum_{i=1}^{p} u_i(d_{i,\tau}+\delta_{c_i} - J_{c_i,\tau}(\pi_\phi))$.

Proof. For any meta-policy π_ϕ, the objective function of the optimization problem (3.1) is strongly concave and the constraint function is convex. The LICQ and the SCSC hold at π^τ. According to Theorem 2 in [52], the Lagrangian multipliers $u_{c_i,\tau}^*(\pi_\phi)$ is unique for any given π_ϕ and $\nabla_\phi u_{c_i,\tau}^*(\pi_\phi)$ exists. The computation is shown in [52]. For all $i\in\{1,\cdots,p\}$ with $u_{c_i,\tau}^*(\pi_\phi)=0$, we have $\nabla_\phi u_{c_i,\tau}^*(\pi_\phi)=0$.

□

F.4 Optimality and constraint satisfaction analysis

F.4.1 Lemmas for optimality and safe analysis

Lemma 4. *For any task τ, and any policies π and $\pi' \in \Pi$, the following bound holds:*

$$\frac{1}{1-\gamma}\mathop{\mathbb{E}}_{\substack{s\sim\nu_\tau^\pi \\ a\sim\pi'(\cdot|s)}}[A_\tau^\pi(s,a)] - C_\tau^\pi(\pi') \le J_\tau(\pi') - J_\tau(\pi) \le \frac{1}{1-\gamma}\mathop{\mathbb{E}}_{\substack{s\sim\nu_\tau^\pi \\ a\sim\pi'(\cdot|s)}}[A_\tau^\pi(s,a)] + C_\tau^\pi(\pi') \quad (29)$$

where

$$C_\tau^\pi(\pi') = \frac{4\gamma \max_{s,a} A_\tau^\pi(s,a)}{(1-\gamma)^2} D_{TV}^{max}(\pi||\pi')\mathbb{E}_{s\sim\nu_\tau^\pi}\left[D_{TV}(\pi(\cdot|s)||\pi'(\cdot|s))\right].$$

Here, we define $D_{TV}(\pi(\cdot|s)||\pi'(\cdot|s)) \triangleq \frac{1}{2}\sum_{a\in\mathcal{A}}|\pi(a|s) - \pi'(a|s)|$ *and* $D_{TV}^{max}(\pi||\pi') \triangleq \max_{s\in\mathcal{S}} D_{TV}(\pi(\cdot|s)||\pi'(\cdot|s))$.

The inequalities (29) *also holds for each $i = 1, \cdots, p$, when A_τ^π and $A_\tau^{\pi'}$ are replaced by $A_{c_i,\tau}^\pi$ and $A_{c_i,\tau}^{\pi'}$, $\max_{s,a} A_\tau^\pi(s,a)$ is replaced by $\max_{s,a} A_{c_i,\tau}^\pi(s,a)$, J_τ is replaced by $J_{c_i,\tau}$.*

Proof. The proof follows similar lines of Theorem 1 in [44] and Corollary 1 and 2 in [2]. For the sake of self-containedness, we provide the complete proof.

Let P_τ^π is a matrix where $P_\tau^\pi(i,j) = \mathbb{E}_{a\sim\pi(\cdot|s_i)}P_\tau(s_j|s_i,a)$ and $P_\tau^{\pi'}$ is a matrix where $P_\tau^{\pi'}(i,j) = \mathbb{E}_{a\sim\pi'(\cdot|s_i)}P_\tau(s_j|s_i,a)$. Let $G = (1 + \gamma P_\tau^\pi + (\gamma P_\tau^\pi)^2 + \ldots) = (1-\gamma P_\tau^\pi)^{-1}$, and similarly $\tilde{G} = (1 + \gamma P^{\pi'_\tau} + (\gamma P^{\pi'_\tau})^2 + \ldots) = (1-\gamma P^{\pi'_\tau})^{-1}$. Let ρ be a density vector on state space and r_τ is a reward function vector on state space, thus $r_\tau^\top \rho$ is a scalar meaning the expected reward under density ρ. Note that $J_\tau(\pi) = r_\tau^\top G\rho_\tau$, and $J_\tau(\pi') = r_\tau^\top \tilde{G}\rho_\tau$. Here, ρ_τ is the initial state distribution for task τ. Let $\Delta = P_\tau^{\pi'} - P_\tau^\pi$.

Follow the proof in Appendix B in [44], we have

$$G^{-1} - \tilde{G}^{-1} = (1-\gamma P_\pi) - (1-\gamma P_{\tilde{\pi}}) = \gamma\Delta.$$

Left multiply by $\tilde{G}$ and right multiply by G,

$$\tilde{G} = \gamma\tilde{G}\Delta G + G. \quad (30)$$

Left multiply by G and right multiply by $\tilde{G}$,

$$\tilde{G} = \gamma G\Delta\tilde{G} + G. \quad (31)$$

Substituting the right-hand side in (30) into $\tilde{G}$ in (31), then

$$\tilde{G} = G + \gamma G\Delta G + \gamma^2 G\Delta\tilde{G}\Delta G.$$

So we have

$$J_\tau(\pi') - J_\tau(\pi) = r_\tau^\top(\tilde{G} - G)\rho_\tau = \gamma r_\tau^\top G\Delta G\rho_\tau + \gamma^2 r_\tau^\top G\Delta\tilde{G}\Delta G\rho_\tau. \quad (32)$$

Note that $r_\tau^\top G = v_\tau^{\pi\top}$, where v is the value function on the state space. We also have $G\rho_\tau = \frac{1}{1-\gamma}\nu_\tau^\pi$, where ν_τ^π is the state visitation distribution vector. So,

$$J_\tau(\tilde{\pi}) - J_\tau(\pi) = r_\tau^\top(\tilde{G} - G)\rho_\tau = \frac{\gamma}{1-\gamma}v_\tau^{\pi\top}\Delta\nu_\tau^\pi + \frac{\gamma^2}{1-\gamma}v_\tau^{\pi\top}\Delta\tilde{G}\Delta\nu_\tau^\pi.$$

Consider the first term $\frac{\gamma}{1-\gamma}v_\tau^{\pi\top}\Delta\nu_\tau^\pi$, similar to Equation (50) in [44], we have

$$\begin{aligned}
&\gamma v_\tau^{\pi\top}\Delta\nu_\tau^\pi = v_\tau^{\pi\top}(P_\tau^{\pi'} - P_\tau^\pi)\nu_\tau^\pi \\
=&\sum_s \nu_\tau^\pi(s)\sum_{s'}\sum_a(\pi'(a|s) - \pi(a|s))P_\tau(s'|s,a)\,\gamma v_\tau^\pi(s') \\
=&\sum_s \nu_\tau^\pi(s)\sum_a(\pi'(a|s) - \pi(a|s))\left[r(s) + \sum_{s'}P_\tau(s'|s,a)\,\gamma v_\tau^\pi(s') - v(s)\right] \\
=&\sum_s \nu_\tau^\pi(s)\sum_a(\pi'(a|s) - \pi(a|s))A_\tau^\pi(s,a)
\end{aligned} \quad (33)$$

Since we have $\sum_a \pi(a|s)A_\tau^\pi(s,a) = 0$, we have

$$\gamma {v_\tau^\pi}^\top \Delta \nu_\tau^\pi = \sum_s \nu_\tau^\pi(s) \sum_a \pi'(a|s) A_\tau^\pi(s,a) = \mathop{\mathbb{E}}_{\substack{s\sim\nu_\tau^\pi \\ a\sim\pi'(\cdot|s)}} \left[A_\tau^\pi(s,a)\right].$$

Combine (32) and the above equation, we have the following for the second term:

$$\frac{\gamma^2}{1-\gamma} {v_\tau^\pi}^\top \Delta \tilde{G} \Delta \nu_\tau^\pi = J_\tau(\pi') - J_\tau(\pi) - \frac{1}{1-\gamma} \mathop{\mathbb{E}}_{\substack{s\sim\nu_\tau^\pi \\ a\sim\pi'(\cdot|s)}} \left[A_\tau^\pi(s,a)\right].$$

Then we need to show

$$\left| \frac{\gamma^2}{1-\gamma} {v_\tau^\pi}^\top \Delta \tilde{G} \Delta \nu_\tau^\pi \right| \le C_\tau^\pi(\pi').$$

First,

$$\begin{aligned}
&\left| \frac{\gamma^2}{1-\gamma} {v_\tau^\pi}^\top \Delta \tilde{G} \Delta \nu_\tau^\pi \right| \\
\le &\left| \frac{\gamma^2}{1-\gamma} \left({v_\tau^\pi}^\top \Delta \right)_{\mathcal{S}^v} \left(\tilde{G} \Delta \nu_\tau^\pi \right)_{\mathcal{S}^v} \right| + \left| \frac{\gamma^2}{1-\gamma} \left({v_\tau^\pi}^\top \Delta \right)_{\mathcal{S}/\mathcal{S}^v} \left(\tilde{G} \Delta \nu_\tau^\pi \right)_{\mathcal{S}/\mathcal{S}^v} \right|
\end{aligned}$$

By Hölder's inequality,

$$\left| \frac{\gamma^2}{1-\gamma} {v_\tau^\pi}^\top \Delta \tilde{G} \Delta \nu_\tau^\pi \right| \le \frac{\gamma}{1-\gamma} \| \gamma {v_\tau^\pi}^\top \Delta \|_\infty \| \tilde{G} \Delta \nu_\tau^\pi \|_1.$$

Similar to (33), each element in the vector $\gamma {v_\tau^\pi}^\top \Delta$ is $\sum_a (\pi'(a|s) - \pi(a|s)) A_\tau^\pi(s,a)$, then we have

$$\left\| \gamma {v_\tau^\pi}^\top \Delta \right\|_\infty \le \max_{s\in\mathcal{S}} \sum_a |\pi'(a|s) - \pi(a|s)| A_\tau^\pi(s,a) \le 2 \max_{s,a} A_\tau^\pi(s,a) D_{TV}^{max}(\pi||\pi').$$

From the Lemma 3 of [2], we have

$$\| \tilde{G} \Delta \nu_\tau^\pi \|_1 \le \frac{2}{1-\gamma} \mathbb{E}_{s\sim\nu_\tau^\pi} \left[D_{TV}(\pi(\cdot|s)||\pi'(\cdot|s)) \right].$$

Therefore, we have

$$\begin{aligned}
&\left| \frac{\gamma^2}{1-\gamma} {v_\tau^\pi}^\top \Delta \tilde{G} \Delta \nu_\tau^\pi \right| \le C_\tau^\pi(\pi') \\
= &\frac{4\gamma \max_{s,a} A_\tau^\pi(s,a)}{(1-\gamma)^2} D_{TV}^{max}(\pi||\pi') \mathbb{E}_{s\sim\nu_\tau^\pi} \left[D_{TV}(\pi(\cdot|s)||\pi'(\cdot|s)) \right]
\end{aligned}$$

Then the bounds hold.

□

Lemma 5 (Restatement of Lemma 1)**.** *For any task τ, and any policies π and $\pi' \in \Pi$ with $\max_{s\in\mathcal{S}} D_{TV}(\pi||\pi') \le \alpha$ $\mathbb{E}_{s\sim\nu_\tau^\pi}[D_{TV}(\pi(\cdot|s)\,||\pi'(\cdot|s))]\}$, the following bound holds:*

$$J_\tau(\pi') - J_\tau(\pi) \le \frac{1}{1-\gamma} \mathop{\mathbb{E}}_{\substack{s\sim\nu_\tau^\pi \\ a\sim\pi'(\cdot|s)}} \left[A_\tau^\pi(s,a)\right] + \frac{2\gamma\alpha A^{max}}{(1-\gamma)^2} \mathbb{E}_{s\sim\nu_\tau^\pi} \left[D_{KL}(\pi'(\cdot|s)||\pi(\cdot|s)) \right]$$

and

$$J_\tau(\pi') - J_\tau(\pi) \ge \frac{1}{1-\gamma} \mathop{\mathbb{E}}_{\substack{s\sim\nu_\tau^\pi \\ a\sim\pi'(\cdot|s)}} \left[A_\tau^\pi(s,a)\right] - \frac{2\gamma\alpha A^{max}}{(1-\gamma)^2} \mathbb{E}_{s\sim\nu_\tau^\pi} \left[D_{KL}(\pi'(\cdot|s)||\pi(\cdot|s)) \right].$$

These two inequalities also holds for each $i = 1, \cdots, p$, when A_τ^π and $A_\tau^{\pi'}$ are replaced by $A_{c_i,\tau}^\pi$ and $A_{c_i,\tau}^{\pi'}$, A^{max} is replaced by $A_{c_i}^{max}$, J_τ is replaced by $J_{c_i,\tau}$.

Lemma 5 is a variant of Theorem 1 in [44] and Corollary 1 and 2 in [2]. The difference is that, the inequalities in Lemma 5 replace the term $\max_s D_{KL}(\pi'(\cdot|s)||\pi(\cdot|s)$ in Theorem 1 in [44] and replace the term $\sqrt{\mathbb{E}_{s\sim\nu_\tau^\pi}\left[D_{KL}(\pi'(\cdot|s)||\pi(\cdot|s))\right]}$ in Corollary 1 and 2 in [2] by $\mathbb{E}_{s\sim\nu_\tau^\pi}\left[D_{KL}(\pi'(\cdot|s)||\pi(\cdot|s))\right]$ in the right-hand side of the inequalities.

Proof. We show the first inequality. The second inequality can be proven similarly. From Lemma 4,

$$\begin{aligned}
&J_\tau(\pi') - J_\tau(\pi) - \frac{1}{1-\gamma}\mathop{\mathbb{E}}_{\substack{s\sim\nu_\tau^\pi\\ a\sim\pi'(\cdot|s)}}\left[A_\tau^\pi(s,a)\right]\\
\leq&\frac{4\gamma\max_{s,a}A_\tau^\pi(s,a)}{(1-\gamma)^2}D_{TV}^{max}(\pi||\pi')\mathbb{E}_{s\sim\nu_\tau^\pi}\left[D_{TV}(\pi(\cdot|s)||\pi'(\cdot|s))\right].
\end{aligned}$$

We have $D_{TV}^{max}(\pi||\pi') \leq \alpha\mathbb{E}_{s\sim\nu_\tau^\pi}\left[D_{TV}(\pi(\cdot|s)||\pi'(\cdot|s))\right]$. Therefore, we have

$$J_\tau(\pi')-J_\tau(\pi)-\frac{1}{1-\gamma}\mathop{\mathbb{E}}_{\substack{s\sim\nu_\tau^\pi\\ a\sim\pi'(\cdot|s)}}\left[A_\tau^\pi(s,a)\right] \leq \frac{4\gamma\alpha\max_{s,a}A_\tau^\pi(s,a)}{(1-\gamma)^2}\mathbb{E}_{s\sim\nu_\tau^\pi}\left[D_{TV}(\pi(\cdot|s)||\pi'(\cdot|s))\right]^2.$$

From Jensen's inequality, we have

$$\mathbb{E}_{s\sim\nu_\tau^\pi}\left[D_{TV}\left(\pi(\cdot|s)\|\pi'(\cdot|s)\right)\right]^2 \leq \mathbb{E}_{s\sim\nu_\tau^\pi}\left[D_{TV}^2\left(\pi(\cdot|s)\|\pi'(\cdot|s)\right)\right].$$

From the above inequalities, we have

$$J_\tau(\pi') - J_\tau(\pi) - \frac{1}{1-\gamma}\mathop{\mathbb{E}}_{\substack{s\sim\nu_\tau^\pi\\ a\sim\pi'(\cdot|s)}}\left[A_\tau^\pi(s,a)\right] \leq \frac{4\gamma\alpha A^{max}}{(1-\gamma)^2}\mathbb{E}_{s\sim\nu_\tau^\pi}\left[D_{TV}^2(\pi(\cdot|s)||\pi'(\cdot|s))\right]. \quad (34)$$

From [14], we have

$$D_{TV}^2(\pi(\cdot|s)||\pi'(\cdot|s)) \leq \frac{1}{2}D_{KL}(\pi'(\cdot|s)||\pi(\cdot|s)).$$

Therefore,

$$J_\tau(\pi') - J_\tau(\pi) \leq \frac{1}{1-\gamma}\mathop{\mathbb{E}}_{\substack{s\sim\nu_\tau^\pi\\ a\sim\pi'(\cdot|s)}}\left[A_\tau^\pi(s,a)\right] + \frac{2\gamma\alpha A^{max}}{(1-\gamma)^2}\mathbb{E}_{s\sim\nu_\tau^\pi}\left[D_{KL}(\pi'(\cdot|s)||\pi(\cdot|s))\right]$$

□

F.4.2 Proof of Propostion 4

Before we prove Proposition 4, we first show a lemma.

Lemma 6. *There exists a constant α such that, for any bounded ϕ, $\max_{s\in\mathcal{S}} D_{TV}(\pi_\phi||\pi^\tau) \leq \alpha\ \mathbb{E}_{s\sim\nu_\tau^{\pi_\phi}}\left[D_{TV}(\pi_\phi(\cdot|s)\ ||\pi^\tau(\cdot|s))\right]\}$ where $\pi^\tau = \mathcal{A}^s(\pi_\phi,\Lambda,\Delta,\tau)$ with $\lambda \geq \frac{2\gamma\alpha A^{max}}{1-\gamma}$ and $\lambda_{c_i} \geq \frac{2\gamma\alpha A_{c_i}^{max}}{(1-\gamma)^2}$ for each i.*

Proof. Consider $\pi^\tau = \mathcal{A}^s(\pi_\phi,\Lambda,\Delta,\tau)$ with $\lambda \geq \frac{2\gamma\alpha A^{max}}{1-\gamma}$ and $\lambda_{c_i} \geq \frac{2\gamma\alpha A_{c_i}^{max}}{(1-\gamma)^2}$ for each i. From Lemma 1, we have

$$\pi^\tau(\cdot|s) \propto \exp(f_\phi(s,\cdot) + (\lambda + (1-\gamma)\sum_{i=1}^p u_{c_i,\tau}^*\lambda_{c_i})^{-1}(A_\tau^{\pi_\phi}(s,\cdot) - \textstyle\sum_{i=1}^p u_{c_i,\tau}^* A_{c_i,\tau}^{\pi_\phi}(s,\cdot))).$$

When $\lambda \geq \frac{2\gamma\alpha A^{max}}{1-\gamma}$ and $\lambda_{c_i} \geq \frac{2\gamma\alpha A_{c_i}^{max}}{(1-\gamma)^2}$,

$$(\lambda + (1-\gamma)\sum_{i=1}^p u_{c_i,\tau}^*\lambda_{c_i})^{-1} \leq \frac{1-\gamma}{2\alpha\gamma}\frac{1}{A^{max} + \sum_{i=1}^p u_{c_i,\tau}^* A_{c_i}^{max}}.$$

Then,

$$\left|(\lambda+(1-\gamma)\sum_{i=1}^{p}u^*_{c_i,\tau}\lambda_{c_i})^{-1}(A^{\pi_\phi}_\tau(s,\cdot)-\textstyle\sum_{i=1}^{p}u^*_{c_i,\tau}A^{\pi_\phi}_{c_i,\tau}(s,\cdot)))\right|\leq\frac{1-\gamma}{2\alpha\gamma}.$$

Next, we denote $\lambda+(1-\gamma)\sum_{i=1}^{p}u^*_{c_i,\tau}\lambda_{c_i}$ as λ_{new}, denote $(A^{\pi_\phi}_\tau(s,\cdot)-\sum_{i=1}^{p}u^*_{c_i,\tau}A^{\pi_\phi}_{c_i,\tau}(s,\cdot))$ as $A_{new}(s,\cdot)$, and denote $A^{max}+\sum_{i=1}^{p}u^*_{c_i,\tau}A^{max}_{c_i}$ as A^{max}_{new}. Then, we have $|\lambda^{-1}_{new}A_{new}(s,\cdot)|\leq\frac{1-\gamma}{2\alpha\gamma}$.

Consider α is sufficiently large, then $\frac{1-\gamma}{2\alpha\gamma}$ could be sufficiently small. From

$$\pi^\tau(\cdot|s)\propto\exp(f_\phi(s,\cdot)+\lambda^{-1}_{new}A_{new}(s,\cdot)),$$

we have

$$\begin{aligned}\pi^\tau(\cdot|s)&=\frac{\exp\left(f_\phi(s,a)+\lambda^{-1}_{new}A_{new}(s,\cdot)\right)}{\sum_{a\in\mathcal{A}}\exp\left(f_\phi(s,a')+\lambda^{-1}_{new}A_{new}(s,\cdot)\right)}=\frac{\exp\left(f_\phi(s,a)+\lambda^{-1}_{new}A_{new}(s,\cdot)\right)}{\sum_{a\in\mathcal{A}}\exp\left(f_\phi(s,a')\right)}\\&=\pi_\phi(\cdot|s)\exp(\lambda^{-1}_{new}A_{new}(s,\cdot))=\pi_\phi(\cdot|s)(1+\lambda^{-1}_{new}A_{new}(s,\cdot))\end{aligned}$$

Therefore,

$$D_{TV}(\pi_\phi(\cdot|s)||\pi^\tau(\cdot|s))=\frac{1}{2}\sum_{a\in\mathcal{A}}|\pi_\phi(a|s)-\pi^\tau(a|s)|=\frac{1}{2}\lambda^{-1}_{new}\sum_{a\in\mathcal{A}}|A_{new}(s,a)|\pi_\phi(a|s)$$

For a policy with a bounded parameter ϕ, $\pi_\phi(a|s)$ is non-trivial larger than 0. Then $\sum_{a\in\mathcal{A}}|A_{new}(s,a)|\pi_\phi(a|s)$ is non-trivial larger than 0 if there exists a with $|A_{new}(s,a)|>0$. If $|A_{new}(s,a)|=0$ for any $s\in\mathcal{S}$ and $a\in\mathcal{A}$, then α could any constant. If there is $s\in\mathcal{S}$ and $a\in\mathcal{A}$ $|A_{new}(s,a)|>0$, since $A_{new}(s,a)$ is continuous, there exists a closed set S, such that $\sum_{a\in\mathcal{A}}|A_{new}(s,a)|\pi_\phi(a|s)$ is non-trivial larger than 0. Therefore, there exists a constant α such that

$$\max_{s\in\mathcal{S}}D_{TV}(\pi_\phi||\pi^\tau)\leq\alpha\mathbb{E}_{s\sim\nu^{\pi_\phi}_\tau}[D_{TV}(\pi_\phi(\cdot|s)||\pi^\tau(\cdot|s))]\}.$$

□

Next, we prove Proposition 4.

Proof of Propostion 4. From Lemma 1 and Lemma 6, we have

$$J_\tau(\pi)\leq J_\tau(\pi_\phi)+\mathbb{E}_{s\sim\nu^\pi_\tau,a\sim\pi(\cdot|s)}\left[\frac{A^{\pi_\phi}_\tau(s,a)}{1-\gamma}\right]+\frac{2\gamma\alpha A^{max}}{(1-\gamma)^2}\mathbb{E}_{s\sim\nu^{\pi_\phi}_\tau}\left[D_{KL}(\pi(\cdot|s)||\pi_\phi(\cdot|s))\right]$$

Since $\lambda_{c_i}\geq\frac{2\gamma\alpha A^{max}_{c_i}}{(1-\gamma)^2}$, we have

$$\begin{aligned}&J_{c_i,\tau}(\pi^\tau)\\&\leq J_{c_i,\tau}\left(\pi_\phi\right)+\mathop{\mathbb{E}}_{\substack{s\sim\nu^{\pi_\phi}_\tau\\a\sim\pi^\tau(\cdot|s)}}\left[\frac{A^{\pi_\phi}_{c_i,\tau}(s,a)}{1-\gamma}\right]+\frac{2\gamma\alpha A^{max}_{c_i}}{(1-\gamma)^2}\,\mathbb{E}_{s\sim\nu^{\pi_\phi}_\tau}\left[D_{KL}\left(\pi^\tau(\cdot|s)\|\pi_\phi(\cdot|s)\right)\right]\\&\leq J_{c_i,\tau}\left(\pi_\phi\right)+\mathop{\mathbb{E}}_{\substack{s\sim\nu^{\pi_\phi}_\tau\\a\sim\pi^\tau(\cdot|s)}}\left[\frac{A^{\pi_\phi}_{c_i,\tau}(s,a)}{1-\gamma}\right]+\lambda_{c_i}\,\mathbb{E}_{s\sim\nu^{\pi_\phi}_\tau}\left[D_{KL}\left(\pi^\tau(\cdot|s)\|\pi_\phi(\cdot|s)\right)\right]\\&\leq d_{i,\tau}+\delta_{c_i}.\end{aligned}$$

Also, we have

$$J_\tau(\pi)\geq J_\tau(\pi_\phi)+\mathbb{E}_{s\sim\nu^\pi_\tau,a\sim\pi(\cdot|s)}\left[\frac{A^{\pi_\phi}_\tau(s,a)}{1-\gamma}\right]-\frac{2\gamma\alpha A^{max}}{(1-\gamma)^2}\mathbb{E}_{s\sim\nu^{\pi_\phi}_\tau}\left[D_{KL}(\pi(\cdot|s)||\pi_\phi(\cdot|s))\right]$$

Since $\lambda\geq\frac{2\gamma\alpha A^{max}}{1-\gamma}$, we have

$$J_\tau(\pi)\geq J_\tau(\pi_\phi)+\mathbb{E}_{s\sim\nu^\pi_\tau,a\sim\pi(\cdot|s)}\left[\frac{A^{\pi_\phi}_\tau(s,a)}{1-\gamma}\right]-\frac{\lambda}{1-\gamma}\mathbb{E}_{s\sim\nu^{\pi_\phi}_\tau}\left[D_{KL}(\pi(\cdot|s)||\pi_\phi(\cdot|s))\right].$$

For the solution π^τ of problem (1), we have

$$
\begin{aligned}
J_\tau(\pi^\tau) &\geq \mathop{\mathbb{E}}_{\substack{s\sim\nu_\tau^{\pi_\phi}\\ a\sim\pi^\tau(\cdot|s)}}\left[\frac{A_\tau^{\pi_\phi}(s,a)}{1-\gamma}\right] - \frac{\lambda}{1-\gamma}\,\mathbb{E}_{s\sim\nu_\tau^{\pi_\phi}}\left[D_{KL}\left(\pi^\tau(\cdot|s)\|\pi_\phi(\cdot|s)\right)\right] + J_\tau(\pi_\phi) \\
&= \max_{\pi\in\Pi_\tau^C} \mathop{\mathbb{E}}_{\substack{s\sim\nu_\tau^{\pi_\phi}\\ a\sim\pi(\cdot|s)}}\left[\frac{A_\tau^{\pi_\phi}(s,a)}{1-\gamma}\right] - \frac{\lambda}{1-\gamma}\,\mathbb{E}_{s\sim\nu_\tau^{\pi_\phi}}\left[D_{KL}\left(\pi(\cdot|s)\|\pi_\phi(\cdot|s)\right)\right] + J_\tau(\pi_\phi) \\
&\geq \mathop{\mathbb{E}}_{\substack{s\sim\nu_\tau^{\pi_\phi}\\ a\sim\pi_\phi(\cdot|s)}}\left[\frac{A_\tau^{\pi_\phi}(s,a)}{1-\gamma}\right] - \frac{\lambda}{1-\gamma}\,\mathbb{E}_{s\sim\nu_\tau^{\pi_\phi}}\left[D_{KL}\left(\pi_\phi(\cdot|s)\|\pi_\phi(\cdot|s)\right)\right] + J_\tau(\pi_\phi) = J_\tau(\pi_\phi).
\end{aligned}
$$

where Π_τ^C is the feasible set of problem (1). The last inequality comes from $\pi^\phi \in \Pi_\tau^C$. □

F.4.3 Proof of Theorem 1

Recall the notations defined in Section 5.2 and used in this section: the optimal task-specific policy π_*^τ for task τ as

$$\pi_*^\tau \triangleq \text{argmax}_{\pi\in\Pi}\ J_\tau(\pi) \text{ s.t. } J_{c_i,\tau}(\pi) \leq d_{i,\tau};$$

the conservative task-specific optimal policy $\pi_{*,[\epsilon]}^\tau$, which is optimal for τ under conservative safety constraints, i.e.,

$$\pi_{*,[\epsilon]}^\tau \triangleq \text{argmax}_{\pi\in\Pi}\ J_\tau(\pi) \text{ s.t. } J_{c_i,\tau}(\pi) \leq d_{i,\tau} - \epsilon,$$

where the conservative constant $\epsilon \geq 0$; the task variance

$$\mathcal{V}ar(\mathbb{P}(\Gamma)) \triangleq \min_\phi \mathbb{E}_{\tau\sim\mathbb{P}(\Gamma)}\mathbb{E}_{s\sim\nu_\tau^{\pi_\phi}}[D_{KL}(\pi_*^\tau(\cdot|s)||\pi_\phi(\cdot|s))];$$

the task variance under the conservative safety constraints

$$\mathcal{V}ar^\epsilon(\mathbb{P}(\Gamma)) \triangleq \min_\phi \mathbb{E}_{\tau\sim\mathbb{P}(\Gamma)}\mathbb{E}_{s\sim\nu_\tau^{\pi_\phi}}[D_{KL}(\pi_{*,[\epsilon]}^\tau(\cdot|s)||\pi_\phi(\cdot|s))],$$

and its minimal point

$$\hat{\phi}^{[\epsilon]} \triangleq \arg\min_\phi \mathbb{E}_{\tau\sim\mathbb{P}(\Gamma)}\mathbb{E}_{s\sim\nu_\tau^{\pi_\phi}}[D_{KL}(\pi_{*,[\epsilon]}^\tau(\cdot|s)||\pi_\phi(\cdot|s))],$$

the radius of the task distribution $\mathbb{P}(\Gamma)$

$$R(\mathbb{P}(\Gamma)) \triangleq \max_{\tau\in\Gamma,\epsilon\in E} \mathbb{E}_{s\sim\nu_\tau^{\pi_{\hat{\phi}^{[\epsilon]}}}}[D_{KL}(\pi_{*,[\epsilon]}^\tau(\cdot|s)||\pi_{\hat{\phi}^{[\epsilon]}}(\cdot|s))],$$

where the set E is defined by $\{\epsilon \geq 0 : \pi_{*,[\epsilon]}^\tau \text{ exists for all } \tau \in \Gamma\}$.

We also define

$$R^{[\epsilon]}(\mathbb{P}(\Gamma)) \triangleq \max_{\tau\in\Gamma} \mathbb{E}_{s\sim\nu_\tau^{\pi_{\hat{\phi}^{[\epsilon]}}}}[D_{KL}(\pi_{*,[\epsilon]}^\tau(\cdot|s)||\pi_{\hat{\phi}^{[\epsilon]}}(\cdot|s))].$$

We first show some lemmas for the proof of Theorem 1.

Lemma 7. *For any ϵ, the policy $\pi_{*,[\epsilon]}^\tau$ belongs to the set $\{\pi \in \Pi : J_{c_i,\tau}(\pi_{\hat{\phi}^{[\epsilon]}}) + \mathop{\mathbb{E}}_{\substack{s\sim\nu_\tau^{\pi_{\hat{\phi}^{[\epsilon]}}}\\ a\sim\pi(\cdot|s)}}\left[\frac{A_{c_i,\tau}^{\pi_{\hat{\phi}^{[\epsilon]}}}(s,a)}{1-\gamma}\right] + \frac{2\gamma\alpha A_{c_i}^{max}}{(1-\gamma)^2}\mathbb{E}_{s\sim\nu_\tau^{\pi_{\hat{\phi}^{[\epsilon]}}}}[D_{KL}(\pi(\cdot|s)||\pi_{\hat{\phi}^{[\epsilon]}}(\cdot|s))] \leq d_{i,\tau} - \epsilon + \frac{4\gamma\alpha A_{c_i}^{max}}{(1-\gamma)^2}R(\mathbb{P}(\Gamma))$ for all $i = 1,\cdots,p\}$*

Proof. From the second inequality in Lemma 1, $J_{c_i,\tau}(\pi_{*,[\epsilon]}^\tau) \geq$

$$J_{c_i,\tau}(\pi_{\hat{\phi}^{[\epsilon]}}) + \frac{1}{1-\gamma}\mathop{\mathbb{E}}_{\substack{s\sim\nu_\tau^{\pi_{\hat{\phi}^{[\epsilon]}}}\\ a\sim\pi_{*,[\epsilon]}^\tau(\cdot|s)}}\left[A_{c_i,\tau}^{\pi_{\hat{\phi}^{[\epsilon]}}}(s,a)\right] - \frac{2\gamma\alpha A_{c_i}^{max}}{(1-\gamma)^2}\mathbb{E}_{s\sim\nu_\tau^{\pi_{\hat{\phi}^{[\epsilon]}}}}\left[D_{KL}(\pi_{*,[\epsilon]}^\tau(\cdot|s)||\pi_{\hat{\phi}^{[\epsilon]}}(\cdot|s))\right].$$

Since $J_{c_i,\tau}(\pi^{\tau}_{*,[\epsilon]}) \leq d_{i,\tau} - \epsilon$, we have

$$J_{c_i,\tau}(\pi_{\hat{\phi}^{[\epsilon]}}) + \frac{1}{1-\gamma} \mathop{\mathbb{E}}_{\substack{s\sim\nu_\tau^{\pi_{\hat{\phi}^{[\epsilon]}}} \\ a\sim\pi^{\tau}_{*,[\epsilon]}(\cdot|s)}} \left[A^{\pi_{\hat{\phi}^{[\epsilon]}}}_{c_i,\tau}(s,a)\right]$$
$$- \frac{2\gamma\alpha A^{max}_{c_i}}{(1-\gamma)^2} \mathbb{E}_{s\sim\nu_\tau^{\pi_{\hat{\phi}^{[\epsilon]}}}} \left[D_{KL}(\pi^{\tau}_{*,[\epsilon]}(\cdot|s)||\pi_{\hat{\phi}^{[\epsilon]}}(\cdot|s))\right] \leq d_{i,\tau} - \epsilon.$$

Then,

$$\begin{aligned}
&J_{c_i,\tau}(\pi_{\hat{\phi}^{[\epsilon]}}) + \mathop{\mathbb{E}}_{\substack{s\sim\nu_\tau^{\pi_{\hat{\phi}^{[\epsilon]}}} \\ a\sim\pi^{\tau}_{*,[\epsilon]}(\cdot|s)}} \left[\frac{A^{\pi_{\hat{\phi}^{[\epsilon]}}}_{c_i,\tau}(s,a)}{1-\gamma}\right] + \frac{2\gamma\alpha A^{max}_{c_i}}{(1-\gamma)^2} \mathbb{E}_{s\sim\nu_\tau^{\pi_{\hat{\phi}^{[\epsilon]}}}} \left[D_{KL}(\pi^{\tau}_{*,[\epsilon]}(\cdot|s)||\pi_{\hat{\phi}^{[\epsilon]}}(\cdot|s))\right] \\
&\leq d_{i,\tau} - \epsilon + \frac{4\gamma\alpha A^{max}_{c_i}}{(1-\gamma)^2} \mathbb{E}_{s\sim\nu_\tau^{\pi_{\hat{\phi}^{[\epsilon]}}}} \left[D_{KL}(\pi^{\tau}_{*,[\epsilon]}(\cdot|s)||\pi_{\hat{\phi}^{[\epsilon]}}(\cdot|s))\right] \\
&\leq d_{i,\tau} - \epsilon + \frac{4\gamma\alpha A^{max}_{c_i}}{(1-\gamma)^2} R^{[\epsilon]}(\mathbb{P}(\Gamma)) \\
&\leq d_{i,\tau} - \epsilon + \frac{4\gamma\alpha A^{max}_{c_i}}{(1-\gamma)^2} R(\mathbb{P}(\Gamma)).
\end{aligned}$$

□

Lemma 8. *We have*

$$\pi_{\hat{\phi}^{[\epsilon]}} \in \left\{\pi \in \Pi : J_{c_i,\tau}(\pi) \leq d_{i,\tau} - \epsilon + \frac{4\gamma\alpha A^{max}_{c_i}}{(1-\gamma)^2} R(\mathbb{P}(\Gamma)) \text{ for all } i = 1, \cdots, p \text{ and } \tau \in \Gamma\right\}.$$

Proof. From the second inequality in Lemma 1, $J_{c_i,\tau}(\pi^{\tau}_{*,[\epsilon]}) \geq$

$$J_{c_i,\tau}(\pi_{\hat{\phi}^{[\epsilon]}}) + \frac{1}{1-\gamma} \mathop{\mathbb{E}}_{\substack{s\sim\nu_\tau^{\pi_{\hat{\phi}^{[\epsilon]}}} \\ a\sim\pi^{\tau}_{*,[\epsilon]}(\cdot|s)}} \left[A^{\pi_{\hat{\phi}^{[\epsilon]}}}_{c_i,\tau}(s,a)\right] - \frac{2\gamma\alpha A^{max}_{c_i}}{(1-\gamma)^2} \mathbb{E}_{s\sim\nu_\tau^{\pi_{\hat{\phi}^{[\epsilon]}}}} \left[D_{KL}(\pi^{\tau}_{*,[\epsilon]}(\cdot|s)||\pi_{\hat{\phi}^{[\epsilon]}}(\cdot|s))\right].$$

Since $J_{c_i,\tau}(\pi^{\tau}_{*,[\epsilon]}) \leq d_{i,\tau} - \epsilon$, we have

$$\begin{aligned}
&J_{c_i,\tau}(\pi_{\hat{\phi}^{[\epsilon]}}) + \frac{1}{1-\gamma} \mathop{\mathbb{E}}_{\substack{s\sim\nu_\tau^{\pi_{\hat{\phi}^{[\epsilon]}}} \\ a\sim\pi^{\tau}_{*,[\epsilon]}(\cdot|s)}} \left[A^{\pi_{\hat{\phi}^{[\epsilon]}}}_{c_i,\tau}(s,a)\right] \\
&\leq d_{i,\tau} - \epsilon + \frac{2\gamma\alpha A^{max}_{c_i}}{(1-\gamma)^2} \mathbb{E}_{s\sim\nu_\tau^{\pi_{\hat{\phi}^{[\epsilon]}}}} \left[D_{KL}(\pi^{\tau}_{*,[\epsilon]}(\cdot|s)||\pi_{\hat{\phi}^{[\epsilon]}}(\cdot|s))\right].
\end{aligned}$$

Also, from (33) and the proof of Lemma 1, we have

$$\begin{aligned}
&\frac{1}{1-\gamma} \mathop{\mathbb{E}}_{\substack{s\sim\nu_\tau^{\pi_{\hat{\phi}^{[\epsilon]}}} \\ a\sim\pi^{\tau}_{*,[\epsilon]}(\cdot|s)}} \left[A^{\pi_{\hat{\phi}^{[\epsilon]}}}_{c_i,\tau}(s,a)\right] \\
&= \frac{1}{1-\gamma} \sum_s \nu_\tau^{\pi_{\hat{\phi}^{[\epsilon]}}}(s) \sum_a (\pi^{\tau}_{*,[\epsilon]}(a|s) - \pi_{\hat{\phi}^{[\epsilon]}}(a|s)) A^{\pi_{\hat{\phi}^{[\epsilon]}}}_{c_i,\tau}(s,a) \\
&\leq \frac{2\alpha A^{max}_{c_i}}{1-\gamma} D^{max}_{TV}(\pi^{\tau}_{*,[\epsilon]}(\cdot|s)||\pi_{\hat{\phi}^{[\epsilon]}}(\cdot|s))^2 \\
&\leq \frac{4\alpha A^{max}_{c_i}}{1-\gamma} \mathbb{E}_{s\sim\nu_\tau^{\pi_{\hat{\phi}^{[\epsilon]}}}} \left[D_{KL}(\pi^{\tau}_{*,[\epsilon]}(\cdot|s)||\pi_{\hat{\phi}^{[\epsilon]}}(\cdot|s))\right]
\end{aligned}$$

Then,

$$
\begin{aligned}
&J_{c_i,\tau}(\pi_{\hat{\phi}^{[\epsilon]}}) + \frac{1}{1-\gamma} \mathop{\mathbb{E}}_{\substack{s\sim\nu_\tau^{\pi_{\hat{\phi}^{[\epsilon]}}} \\ a\sim\pi^\tau_{*,[\epsilon]}(\cdot|s)}} \left[A_{c_i,\tau}^{\pi_{\hat{\phi}^{[\epsilon]}}}(s,a)\right] \\
&\leq d_{i,\tau} - \epsilon + \frac{2\alpha A_{c_i}^{max}}{(1-\gamma)^2} \mathbb{E}_{s\sim\nu_\tau^{\pi_{\hat{\phi}^{[\epsilon]}}}} \left[D_{KL}(\pi^\tau_{*,[\epsilon]}(\cdot|s)||\pi_{\hat{\phi}^{[\epsilon]}}(\cdot|s))\right] \\
&\leq d_{i,\tau} - \epsilon + \frac{4\gamma\alpha A_{c_i}^{max}}{(1-\gamma)^2} R^{[\epsilon]}(\mathbb{P}(\Gamma)) \\
&\leq d_{i,\tau} - \epsilon + \frac{4\gamma\alpha A_{c_i}^{max}}{(1-\gamma)^2} R(\mathbb{P}(\Gamma)).
\end{aligned}
$$

Here, we assume $\gamma \geq 0.5$, which is commonly used.

□

Theorem 2. *Suppose that Assumptions 1 holds. Let* $\pi^\tau(\hat{\phi}^{[\epsilon]}) = \mathcal{A}^s(\pi_{\hat{\phi}^{[\epsilon]}}, \Lambda, \Delta, \tau)$ *with* $\lambda = \frac{2\gamma\alpha A^{max}}{1-\gamma}$, $\lambda_{c_i} = \frac{2\gamma\alpha A_{c_i}^{max}}{(1-\gamma)^2}$ *and* $\delta_{c_i} = \frac{2\gamma\alpha A_{c_i}^{max}}{(1-\gamma)^2} R(\mathbb{P}(\Gamma)) - \epsilon$. *We have*

$$
\mathbb{E}_{\tau\sim\mathbb{P}(\Gamma)}[J_\tau(\pi^\tau_{*,[\epsilon]}) - J_\tau(\mathcal{A}^s(\pi_{\phi^*}, \Lambda, \Delta, \tau))] \leq \frac{4\gamma\alpha A^{max}}{(1-\gamma)^2} \mathcal{V}ar^\epsilon(\mathbb{P}(\Gamma)).
$$

Proof. From Lemma 7, we have that $\pi^\tau_{*,[\epsilon]} \in \Pi^B$, where $\Pi^B \triangleq \{\pi \in \Pi : J_{c_i,\tau}(\pi_{\hat{\phi}^{[\epsilon]}}) + \frac{1}{1-\gamma} \mathop{\mathbb{E}}_{\substack{s\sim\nu_\tau^{\pi_{\hat{\phi}^{[\epsilon]}}} \\ a\sim\pi(\cdot|s)}} \left[A_{c_i,\tau}^{\pi_{\hat{\phi}^{[\epsilon]}}}(s,a)\right] + \lambda_{c_i,\tau} \mathbb{E}_{s\sim\nu_\tau^{\pi_{\hat{\phi}^{[\epsilon]}}}} [D_{KL}(\pi(\cdot|s)||\pi_{\hat{\phi}^{[\epsilon]}}(\cdot|s))] \leq d_{i,\tau} + \delta_{c_i}, \forall i\}$.

Also, $\pi^\tau(\hat{\phi}^{[\epsilon]}) \in \Pi^B$. Therefore, from the definition of $\mathcal{A}^s$ in problem (1), we have

$$
\mathop{\mathbb{E}}_{\substack{s\sim\nu_\tau^{\pi_{\hat{\phi}^{[\epsilon]}}} \\ a\sim\pi^\tau(\hat{\phi}^{[\epsilon]})(\cdot|s)}} \left[A_\tau^{\pi_{\hat{\phi}^{[\epsilon]}}}(s,a)\right] - \lambda \bar{D}_{KL}(\pi^\tau(\hat{\phi}^{[\epsilon]}), \pi_{\hat{\phi}^{[\epsilon]}}) \geq \mathop{\mathbb{E}}_{\substack{s\sim\nu_\tau^{\pi_{\hat{\phi}^{[\epsilon]}}} \\ a\sim\pi^\tau_{*,[\epsilon]}(\cdot|s)}} \left[A_\tau^{\pi_{\hat{\phi}^{[\epsilon]}}}(s,a)\right] - \lambda \bar{D}_{KL}(\pi^\tau_{*,[\epsilon]}, \pi_{\hat{\phi}^{[\epsilon]}}),
$$

where we use $\bar{D}_{KL}(\pi_1(\cdot|s), \pi_2(\cdot|s))$ to represent $\mathbb{E}_{s\sim\nu_\tau^{\pi_2}}[D_{KL}(\pi_1(\cdot|s), \pi_2(\cdot|s))]$.

From the second inequality in Lemma 1 and the above inequality,

$$
\begin{aligned}
J_\tau(\pi^\tau(\hat{\phi}^{[\epsilon]})) - J_\tau(\pi_{\hat{\phi}^{[\epsilon]}}) \geq& \frac{1}{1-\gamma} \mathop{\mathbb{E}}_{\substack{s\sim\nu_\tau^{\pi_{\hat{\phi}^{[\epsilon]}}} \\ a\sim\pi^\tau(\hat{\phi}^{[\epsilon]})(\cdot|s)}} \left[A_\tau^{\pi_{\hat{\phi}^{[\epsilon]}}}(s,a)\right] - \frac{\lambda}{1-\gamma} \bar{D}_{KL}(\pi^\tau(\hat{\phi}^{[\epsilon]}), \pi_{\hat{\phi}^{[\epsilon]}}) \\
\geq& \frac{1}{1-\gamma} \mathop{\mathbb{E}}_{\substack{s\sim\nu_\tau^{\pi_{\hat{\phi}^{[\epsilon]}}} \\ a\sim\pi^\tau_{*,[\epsilon]}(\cdot|s)}} \left[A_\tau^{\pi_{\hat{\phi}^{[\epsilon]}}}(s,a)\right] - \frac{\lambda}{1-\gamma} \bar{D}_{KL}(\pi^\tau_{*,[\epsilon]}, \pi_{\hat{\phi}^{[\epsilon]}}).
\end{aligned}
$$

From the first inequality in Lemma 1,

$$
J_\tau(\pi^\tau_{*,[\epsilon]}) - J_\tau(\pi_{\hat{\phi}^{[\epsilon]}}) \leq \frac{1}{1-\gamma} \mathop{\mathbb{E}}_{\substack{s\sim\nu_\tau^{\pi_{\hat{\phi}^{[\epsilon]}}} \\ a\sim\pi^\tau_{*,[\epsilon]}(\cdot|s)}} \left[A_\tau^{\pi_{\hat{\phi}^{[\epsilon]}}}(s,a)\right] + \frac{2\gamma\alpha A^{max}}{(1-\gamma)^2} \bar{D}_{KL}(\pi^\tau_{*,[\epsilon]}, \pi_{\hat{\phi}^{[\epsilon]}}).
$$

From the last two inequalities,

$$
J_\tau(\pi^\tau(\hat{\phi}^{[\epsilon]})) - J_\tau(\pi^\tau_{*,[\epsilon]}) \geq -(\frac{2\gamma\alpha A^{max}}{(1-\gamma)^2} + \frac{\lambda}{1-\gamma}) \bar{D}_{KL}(\pi^\tau_{*,[\epsilon]}, \pi_{\hat{\phi}^{[\epsilon]}}),
$$

i.e.,

$$
J_\tau(\pi^\tau_{*,[\epsilon]}) - J_\tau(\mathcal{A}^s(\pi_{\hat{\phi}^{[\epsilon]}}, \Lambda, \Delta, \tau)) \leq (\frac{2\gamma\alpha A^{max}}{(1-\gamma)^2} + \frac{\lambda}{1-\gamma}) \bar{D}_{KL}(\pi^\tau_{*,[\epsilon]}, \pi_{\hat{\phi}^{[\epsilon]}}).
$$

Then,

$$\begin{aligned}
&\mathbb{E}_{\tau\sim\mathbb{P}(\Gamma)}[J_\tau(\pi^\tau_{*,[\epsilon]}) - J_\tau(\mathcal{A}^s(\pi_{\hat{\phi}^{[\epsilon]}}, \Lambda, \Delta, \tau))] \\
&\leq (\frac{2\gamma\alpha A^{max}}{(1-\gamma)^2} + \frac{\lambda}{1-\gamma})\mathbb{E}_{\tau\sim\mathbb{P}(\Gamma)}[\bar{D}_{KL}(\pi^\tau_{*,[\epsilon]}, \pi_{\hat{\phi}^{[\epsilon]}})] \\
&= (\frac{2\gamma\alpha A^{max}}{(1-\gamma)^2} + \frac{\lambda}{1-\gamma})\mathcal{V}ar^\epsilon(\mathbb{P}(\Gamma)).
\end{aligned}$$

Moreover, from Lemma 8,

$$\pi_{\hat{\phi}^{[\epsilon]}} \in \Pi^C \triangleq \{\pi \in \Pi : J_{c_i,\tau}(\pi) \leq d_{i,\tau} + \delta_{c_i} \text{ for all } i = 1, \cdots, p \text{ and } \tau \in \Gamma\}.$$

From the definition of ϕ^*, we have

$$\begin{aligned}
\mathbb{E}_{\tau\sim\mathbb{P}(\Gamma)}[J_\tau(\mathcal{A}^s(\pi_{\phi^*}, \Lambda, \Delta, \tau))] &\geq \max_{\pi\in\Pi^C} \mathbb{E}_{\tau\sim\mathbb{P}(\Gamma)}[J_\tau(\mathcal{A}^s(\pi, \Lambda, \Delta, \tau))] \\
&\geq \mathbb{E}_{\tau\sim\mathbb{P}(\Gamma)}[J_\tau(\mathcal{A}^s(\pi_{\hat{\phi}^{[\epsilon]}}, \Lambda, \Delta, \tau))]
\end{aligned}$$

Then, we have

$$\begin{aligned}
&\mathbb{E}_{\tau\sim\mathbb{P}(\Gamma)}[J_\tau(\pi^\tau_{*,[\epsilon]}) - J_\tau(\mathcal{A}^s(\pi_{\phi^*}, \Lambda, \Delta, \tau))] \\
&\leq \mathbb{E}_{\tau\sim\mathbb{P}(\Gamma)}[J_\tau(\pi^\tau_{*,[\epsilon]}) - J_\tau(\mathcal{A}^s(\pi_{\hat{\phi}^{[\epsilon]}}, \Lambda, \Delta, \tau))] \\
&\leq (\frac{2\gamma\alpha A^{max}}{(1-\gamma)^2} + \frac{\lambda}{1-\gamma})\mathcal{V}ar^\epsilon(\mathbb{P}(\Gamma)) \\
&\leq \frac{4\gamma\alpha A^{max}}{(1-\gamma)^2}\mathcal{V}ar^\epsilon(\mathbb{P}(\Gamma)).
\end{aligned}$$

□

Proof of Theorem 1. Theorem 1 is proven by combining Theorem 2 with Corrolary 1. □

G Discussion on the Tightness of the Derived Lower Bound in Theorem 1

Notice that the meta-safe RL aims to extract common knowledge from multiple existing RL tasks. The tasks in the task distribution are usually correlated, and their near-optimal policies usually produce similar actions on a large part of the state space. For example, in the experiments of navigation scenarios with collision avoidance, although optimal policies under different environments should produce different actions when meeting different obstacles, the actions in free space should be similar. As $\mathcal{V}ar(\mathbb{P}(\Gamma))$ is defined as $\min_\phi \mathbb{E}_{\tau\sim\mathbb{P}(\Gamma)}\mathbb{E}_{s\sim\nu^{\pi_\phi}_\tau}[D_{KL}(\pi^\tau_*(\cdot|s)||\pi_\phi(\cdot|s))]$, which computes the expectation over the entire state space, it could be small, especially when the state space is large. Moreover, $\mathcal{V}ar(\mathbb{P}(\Gamma)) \triangleq \min_\phi \mathbb{E}_{\tau\sim\mathbb{P}(\Gamma)}\mathbb{E}_{s\sim\nu^{\pi_\phi}_\tau}[D_{KL}(\pi^\tau_*(\cdot|s)||\pi_\phi(\cdot|s))]$ is the minimal average distance between policies. According to our experiment, 0.05 is a sufficiently large number for the quantity for the KL divergence between two policies and the value of $\mathcal{V}ar(\mathbb{P}(\Gamma))$ under this KL divergence is about $\frac{1}{4} \times 0.05$.

Next, we roughly evaluate the quantity of the optimality bound $\frac{4\gamma\alpha A^{max}}{(1-\gamma)^2}\mathcal{V}ar^\epsilon(\mathbb{P}(\Gamma))$ and show it can be relatively tight. We set $\gamma = 0.99$, the reward/cost $r \in [0, 1]$ and $c \in [0, 1]$. The max advantage function value $A^{\max} \leq 1$ is generally not very small. However, as indicated in the proof of Lemma F.4.1, we actually can replace $A^{\max}$ by $\max_{s,a} A^{\pi^*_\phi}_\tau(s, a)$, which is computed on the meta-policy and is usually much smaller than 1, we set it as 0.05. Then, the optimality bound of $\frac{4\gamma\alpha A^{max}}{(1-\gamma)^2}\mathcal{V}ar^\epsilon(\mathbb{P}(\Gamma))$ is about 25. Note that, the quantity of the total reward/cost is about $\sum_{n=1}^\infty \gamma^n = \frac{1}{\gamma} = 100$. Therefore, the bound is about 25% of the total reward/cost, which is relatively tight.

It is standard in RL and safe RL to derive a lower bound with an order of quantity similar to the optimality bound $\frac{4\gamma\alpha A^{max}}{(1-\gamma)^2}\mathcal{V}ar^\epsilon(\mathbb{P}(\Gamma))$ in this paper. For example, in TRPO [44] and CPO [2], the lower bounds also include the term $\frac{A^{max}}{(1-\gamma)^2}D_{KL}$. Moreover, for conciseness and broad applicability, we have to derive a general optimality bound, which makes the bound looser. When the used condition and target problem are specified, a tighter optimality bound can be derived. For example, when a

large task distribution is specified, i.e., $\mathcal{V}ar(\mathbb{P}(\Gamma))$ is large, we can also make some inequalities (such as those in lines 1573 to 1585) tighter. However, when the target problem is required to be specified, the generality of the optimality bound will be lost and it will become harder to understand and less intuitive.

H Limitations and Future Works

In this paper, we consider the safety metric of CMDP, i.e., the expected accumulated costs satisfy the given safety threshold. This metric is generally less rigorous than the safe control research, where safety is defined as persistently satisfying certain state constraints. A future work is establishing a safe meta-RL algorithm with the rigorous safety metric. Another limitation is that we assume the solution of problem (2) exists, i.e., there exists a policy such that it is safe for all tasks as the initial policy for policy adaptation steps. A future work is to release this assumption and identify a safe task-specific meta-policy for each given task.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction, including the main contribution statement and related works, accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We provide the limitations in the Appendix H.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: All proofs are provided in Appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide all details of the information needed to reproduce the main experimental results in the experiment section and in Appendix B.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide open access to the data and code with sufficient instructions in the supplemental material.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide all training details in Appendix B.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We provide it in the section of the experiment.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Justification: We provide the information at the beginning of the Appendix B.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: the NeurIPS Code of Ethics is followed.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: This paper presents work whose goal is to advance the field of Machine Learning. There is no potential societal consequence.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: The paper does not use existing assets.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: This research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.