Provable Guarantees for Meta-Safe Reinforcement Learning

Vanshaj Khattar
Virginia Tech
Blacksburg, VA 24061
vanshajk@vt.edu

Yuhao Ding
UC Berkeley
Berkeley, CA 94709
yuhao_ding@berkeley.edu

Javad Lavaei
UC Berkeley
Berkeley, CA 94709
lavaei@berkeley.edu

Ming Jin[*]
Virginia Tech
Blacksburg, VA 24061
jinming@vt.edu

Abstract

We study the problem of meta-safe reinforcement learning (meta-SRL) through the CMDP-within-online framework and show that the task-averaged optimality gap and constraint satisfaction improve with task-similarity in the static environment or task-relatedness in the changing environment. Several technical challenges arise when making this framework practical while still having strong theoretical guarantees. To address these challenges, we propose a meta-algorithm that performs inexact online learning on the upper bounds of intra-task optimality gap and constraint violations estimated by off-policy stationary distribution corrections. Furthermore, in settings with dynamically changing task-environment, our approach enables the learning rates to be adapted to intra-task geometry. Furthermore, we address the problem of meta-initialization of the critic network with function approximations. Finally, experiments are conducted to demonstrate the effectiveness of our approach. The proposed theoretical framework is the first to handle the nonconvexity and stochasticity nature of within-task CMDPs, while exploiting inter-task dependency and intra-task geometries for multi-task safe learning.

1 Introduction

The field of meta-reinforcement learning (meta-RL) has recently evolved as one of the promising directions that enables reinforcement learning (RL) agents to learn quickly in dynamically changing environments [43, 73, 104, 84, 51]. Many real-world applications, nevertheless, have safety constraints that should be rarely violated, which existing works do not fully address. These safe RL problems are often modeled as constrained Markov decision processes (CMDPs), where the agent aims to maximize the value function while satisfying given constraints on the trajectory [4]. However, unlike meta-learning, CMDP algorithms are not designed to generalize efficiently over unseen tasks [25, 79, 40, 31, 32, 96, 98, 93, 23, 19]. In this paper, we study how meta-learning can be principally designed to help safe RL algorithms adapt more quickly while satisfying critical constraints.

There are several unique challenges involved in meta-learning for the CMDP settings. First, there are multiple losses incurred at each time step, i.e., the reward and constraints. Since these functions are typically non-convex and coupled through dynamics, adapting existing theories developed for stylized settings such as online convex optimization [50, 83] and supervised learning [26, 77] is not

[*] Corresponding author

Preprint. Under review.

straightforward. Second, it is unrealistic to assume the computation of a globally optimal policy or the corresponding state visitation distribution for CMDPs (unlike online learning [83]); even the characterization of the distance of a suboptimal policy to the set of global policies has been limited to some restricted settings, i.e., either algorithm-dependent [72] or requiring entropy regularization [33]. Thus, classical online learning algorithms that assume exact or unbiased estimator of the loss function do not apply [83]. Third, while the incorporation of a critic network can often improve convergence [94, 49], to understand the role of meta-learning for the critic entails studying the "prescient" form of RL convergence guarantees that explicitly shows the dependency of the critic initialization on the overall convergence. Above all, there is an interplay among nonconvexity and stochasticity nature of the optimization problem, as well as algorithm and generalization considerations, posing significant complexity to leverage inter-task dependency and intra-task geometries.

To this end, we propose a provably low-regret online learning framework that extends the current meta-learning algorithms to the safe RL settings. In view of the aforementioned challenges, our main contributions to the meta-learning and safe RL literature are as follows:

1. **Inexact CMDP-within-online framework:** We propose a novel CMDP-within-online framework where the within-task is CMDP and the meta learner aims to learn the meta-initialization and learning rate. In our framework, the meta learner only requires the inexact optimal policies for each within-task CMDP and the approximate state visitation distributions estimated using collected offline trajectories to construct the upper bounds on the suboptimality gap and constraint violations. An upper bound on these estimation errors is also established.
2. **Adapting to the dynamic regret and intra-task geometry:** We consider adaptive learning rates for objective and constraints in each task, which account for the inherent difficulties of each objective, to minimize dynamic regret across tasks.
3. **Meta-critic initialization under function approximation settings:** We propose to meta initialize the critic network to provably improve within-task convergence guarantees and constraint satisfactions, where both the critic and the actor are instantiated as two-layer neural networks.

Incorporating all these components makes our Meta-safe RL (Meta-SRL) approach highly practical and theoretically appealing for potential adaption to different RL settings.

1.1 Related work

Meta-reinforcement learning: Current state-of-the-art meta-RL includes learning the initial conditions [43], hyperparameters [54], step directions [65] and stepsizes [97], and training recurrent neural networks to embed previous task experience [38] (see also [22] for sim-to-real transfer), with recent developments on improving meta-optimization [82, 66, 85] (see [51] for a review). Recently, [41, 56] provided the theoretical studies on the convergence of model-agnostic meta-RL. However, these works all focus on the unconstrained meta-RL and their local optimality convergence, while our work is the first to obtain provable guarantees for optimality and constraint satisfaction for CMDPs.

Online meta-learning/learning-to-learn (LTL). Most efforts studying initialization-based meta-learning focus on the setting with decomposable within-task loss functions that are often convex [44, 30, 9]; non-convex within-task settings are studied usually for multi-task representation learning [8, 71, 37, 88]. Theoretically, our work is inspired by the Average Regret-Upper-Bound Analysis (ARUBA) strategy [59] for obtaining a meta-procedure, which has been recently extended to learning non-convex piecewise-Lipschitz functions [10]; the main technical advance in our work is in providing the guarantees for CMDPs, which is challenging due to the interplay between the non-convexity and stochasticity of the optimization and the complexity of the within-task safe RL algorithms that involves policy update, critic learning, and the proper choice of stepsizes for reward/constraints.

Inexact online learning. Online learning with access to inexact loss/gradient information has been studied for stochastic zero-biased noise [21, 95, 11, 34], deterministic error/nonzero-biased stochastic noise [11, 34], and adversarial perturbation [81]. Our analysis for static regret uses the formalism of $\epsilon-$subgradient [55, Chap. XI]; for dynamic regret, we extend the work [101] to the inexact setting allowing multiple updates per round and provide improved rates than prior results [11, 34].

Safe RL and CMDP. Direct policy search methods have had substantial empirical successes in solving CMDPs [17, 89, 14, 14, 1, 24] (see, e.g., [45] for a survey of safe RL). Recently, a major progress in understanding the theoretical nonasymptotic global convergence behavior of policy-based methods for CMDPs has also been achieved [25, 79, 40, 31, 32, 96, 98, 93, 23, 69, 68]. However, most of these works only study a single CMDP task and don't seek to make the algorithm perform well on new, potentially related CMDP tasks. In addition, while our work uses the constraint-rectified policy optimization (CRPO) algorithm proposed in [93] as a building block, our framework can be potentially adapted to most of the existing RL literature by making the dependence of guarantees on initial policy/step sizes explicit, e.g., safe exploration [40], regularization [46], off-policy evaluation [39, 86], and offline RL under constraints [61, 92, 62, 87].

2 CMDP-within-online framework

In this section, we introduce the CMDP-within-online framework for the meta-SRL problems. In this framework, a within-task online learning algorithm (such as CRPO [93]) is run on some CMDP task $t \in [T]$, which is encapsulated inside another online learning algorithm (meta-learning algorithm) that learns an initialization policy ϕ_t and learning rate $\alpha_t > 0$ for each within-task algorithm (and also the critic initialization in the function approximation setting). The goal is to learn the optimal meta initialization ϕ^* and learning rate α^* (and also the critic initialization in the function approximation setting) for the within-task algorithm, such that the task-averaged performance is optimized.

2.1 CMDP and the primal approach

Model. For each task $t \in [T]$, a CMDP $\mathcal{M}_t$ is defined by the state space $\mathcal{S}$, the action space $\mathcal{A}$, the discount factor γ, the initial state distribution over the state-space ρ_t, the transition kernel $P_t(s'|s,a) : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$, the reward functions $c_{t,0} : \mathcal{S} \times \mathcal{A} \to [0,1]$ and cost functions $c_{t,i} : \mathcal{S} \times \mathcal{A} \to [0,1]$ for $i = 1, ..., p$. The actions are chosen according to a stochastic policy $\pi_t : \mathcal{S} \to \Delta(\mathcal{A})$ where $\Delta(\mathcal{A})$ is the simplex over the action space. We use $\Delta(\mathcal{A})^{|\mathcal{S}|}$ to denote the simplex over all states $\mathcal{S}$. The initial policy for task t is denoted as $\pi_{t,0}$. The discounted state visitation distribution of a policy π is defined as $\nu_{s_0}^{\pi}(s) := (1-\gamma)\sum_{t=0}^{\infty} \gamma^t P\left(s_t = s \mid \pi, s_0\right)$ and we write $\nu_t^*(s) := \mathbb{E}_{s_0 \sim \rho_t}\left[d_{s_0}^{\pi^*}(s)\right]$ as the visitation distribution when the initial state follows ρ_t at task t.

Policy parameterization. In the tabular setting, we consider the softmax parameterization. For any $w \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{A}|}$, the corresponding softmax policy π_θ is defined as $\pi_\theta(a \mid s) := \frac{\exp(\theta(s,a))}{\sum_{a' \in \mathcal{A}} \exp(\theta(s,a'))}, \forall (s,a) \in \mathcal{S} \times \mathcal{A}$. In the function approximation setting, we parameterize the policy by a two-layer neural network together with the softmax policy $\theta(s,a) := f((s,a);\omega,b) = \frac{1}{\sqrt{W}} \sum_{\iota=1}^{W} b_\iota \cdot \mathrm{ReLU}(\omega_\iota^\top \xi(x,a))$ for any state-action pair (s,a), where $\xi(s,a) \in \mathbb{R}^d$ is the feature vector with $d \geq 2$ and $\|\xi(s,a)\| \leq 1$, $\mathrm{ReLU}(x) = \mathbb{1}(x > 0) \cdot x$, $b = [b_1, \cdots, b_W]^\top \in \mathbb{R}^W$, and $\omega = [\omega_1^\top, \cdots, \omega_W^\top]^\top \in \mathbb{R}^{Wd}$ form the set of parameters θ.

Value function. For a given task t and a policy π, we define the state value function as $V_{t,\pi}^i(s) = \mathbb{E}_t\left[\sum_{t=0}^{\infty} \gamma^t c_{t,i}\left(s_t, a_t, s_{t+1}\right) \mid s_0 = s, \pi\right]$ and the state-action value function as $Q_{t,\pi}^i(s,a) = \mathbb{E}_t\left[\sum_{t=0}^{\infty} \gamma^t c_{t,i}\left(s_t, a_t, s_{t+1}\right) \mid s_0 = s, a_0 = a, \pi\right]$.Furthermore, the expected total reward/cost functions are defined as $J_{t,i}(\pi) = \mathbb{E}_{\rho_t}\left[V_{t,\pi}^i(s)\right] = \mathbb{E}_{\rho_t \cdot \pi}\left[Q_{t,\pi}^i(s,a)\right]$.

CMDP. In each task t, the goal of the agent in safe RL is to solve the following CMDP problem

$$\max_{\pi} J_{t,0}(\pi) \quad \text{s.t.} \quad J_{t,i}(\pi) \leq d_{t,i}, \quad \forall i = 1, ..., p, \tag{1}$$

where $d_{t,i}$ is a fixed limit on the expected total cost $J_{t,i}$. We denote the optimal solution of problem (1) for the task t as π_t^* (which can be non-unique).

Primal approach. In this work, we focus on the primal approach: CRPO [93], for CMDPs since the primal approach does not introduce additional dual variables to optimize and involves less hyperparamter tuning. For example, in the tabular setting with softmax parameterization and the carefully chosen parameters, the suboptimality gap and constraint violation for CMDP task t can be bounded as follows (if the exact $\{Q_{t,\pi}^i\}_{i=0}^p$ are available for all π):

$$\begin{aligned} R_0 &= J_{t,0}(\pi_t^*) - \mathbb{E}[J_{t,0}(\hat{\pi}_t)] \leq \frac{2}{\alpha_t M}\mathbb{E}_{s \sim \nu_t^*}[D(\pi_t^* | \pi_{t,0})] + \frac{4\alpha_t |\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3}, \\ R_i &= \mathbb{E}[J_{t,i}(\hat{\pi}_t)] - d_{t,i} \leq \frac{2}{\alpha_t M}\mathbb{E}_{s \sim \nu_t^*}[D(\pi_t^* | \pi_{t,0})] + \frac{4\alpha_t |\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3}, \forall i = 1, ..., p. \end{aligned} \tag{2}$$

where M is the number of iterations for running CRPO, $\hat{\pi}_t$ is the policy returned by running CRPO for M steps at task t, $D(\cdot|\cdot)$ represents the KL divergence between two policies and α_t is the learning rate of the CRPO at task t.

2.2 Meta-CMDP problem setup

We now consider the lifelong extension of CMDPs in which safe RL tasks arrive one at a time and $t = 1, 2, \dots, T$ now index a sequence of online learning problems. In each single task t, the agent must sequentially optimize the policy $\{\pi_{t,i}\}_{i=0}^{M}$ so that the corresponding sub-optimality and the constraint violation, such as given in (2), grow sub-linearly in M. Beyond the single task, the meta-learner aims to optimize the upper bounds in (2) over the initial policy $\pi_{t,0}$ and the learning rate α_t so that the task-averaged sub-optimality and the task-averaged constraint violation are expected to improve as the meta learner solves more tasks. We will aim to minimize the task-averaged sub-optimality gap and the task-averaged constraint violation defined as follows:

Definition 1. *The* ***task-averaged optimality gap (TAOG)*** $\bar{R}_0$ *and the* ***task-averaged constraint-violation*** *(TACV) of a safe RL algorithm after* T *tasks with* $\{M\}_{t=1}^{T}$ *steps are*

$$\bar{R}_0 = \frac{1}{T}\sum_{t=1}^{T}\left[J_{t,0}(\pi_t^*) - \mathbb{E}[J_{t,0}(\hat{\pi}_t)]\right],\ \bar{R}_i = \frac{1}{T}\sum_{t=1}^{T}\left[\mathbb{E}[J_{t,i}(\hat{\pi}_t)] - d_{t,i}\right],\ \forall i = 1, ..., p, \tag{3}$$

where $\hat{\pi}_t$ *is the policy returned by running some safe RL algorithm for* M *time-steps at task t and the expectation is taken with respect to selecting* $\hat{\pi}_t$.

Note that, unlike in the standard regret and further assumptions on the environment (e.g., [60]), one cannot achieve TAOG and TACV decreasing in T, because the comparator π_t^* is dynamic which can lead to suboptimality or constraint violation at each task t. Furthermore, the average is taken over T and a low TAOG and TACV ensure that the optimality gap or constraint violation of an algorithm is low on average over the tasks compared to that of the optimal within-task parameter. In this work, we aim to design the meta algorithm so that TAOG and TACV are minimized if CRPO is used as the within-task algorithm for CMDPs.

2.3 Task-similarity and task-relatedness

In the meta-SRL, we expect the TAOG and TACV improves with the similarity of the online CMDP tasks. Furthermore, this notion of similarity not only affects the evaluation of the meta learning algorithm, but also impacts the quality of the meta initialization being learned and eventually the performance on an unseen task. We now discuss the notions of similarity for the CMDP tasks.

When the optimal policies $\{\pi_t^*\}_{t=1}^{T}$ are close to a fixed policy, the task-similarity can be measured by $D^2 = \min\limits_{\phi \in \Delta(\mathcal{A})^{|\mathcal{S}|}} \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{s\sim\nu_t^*}[D(\pi_t^*|\phi)]$. This notion of task similarity in the static environment is natural for studying gradient-based meta-learning, as the notion that there exists a meta parameter ϕ from which a good parameter for any individual task is reachable with only a few steps implies that they are all close together. In particular, when the tasks are all identical, i.e., $\{\pi_t^*\}_{t=1}^{T}$ are all equal, we have $D^2 = 0$. Furthermore, in many settings we have a changing environment, so it is natural to study dynamic regret and compare with a sequence of potentially time-varying comparators $\{\psi_t\}_{t=1}^{T}$, [59]. The corresponding task-relatedness can can be measured by $V_\psi^2 = \sum_{t=1}^{T}\frac{1}{T}\mathbb{E}_{s\sim\nu_t^*}[D(\pi_t^*|\psi_t)]$.

In this work, we aim to develop algorithms whose TAOG and TACV scales with the task-similarity or task-relatedness, which implies that our method will do well if these measures are small. To understand the CMDP framework and the impact of task-similarity on the upper bounds of TAOG and TACV for meta-SRL, we first present a simplified result under the ideal setting where $\{\nu_t^*\}_{t=1}^{T}$ and $\{\pi_t^*\}_{t=1}^{T}$ are available after each task t and the task similarity D^{*2} is known.

Lemma 1. *Assume* $\{\nu_t^*\}_{t=1}^{T}$ *and* $\{\pi_t^*\}_{t=1}^{T}$ *are given. For each task t, we run CRPO for* M *iterations with* $\alpha = \frac{(1-\gamma)^{\frac{3}{2}} D^*}{\sqrt{M|\mathcal{S}||\mathcal{A}|}}$. *In addition, the initialization* $\{\pi_{t,0}\}_{t=1}^{T}$ *are determined by playing Follow-the-Regularized-Leader (FTRL) or online mirror descent (OMD) [50] on the functions* $\mathbb{E}_{s\sim\nu_t^*}\left[D\left(\pi_t^*|\cdot\right)\right]$, *for* $t = 1, \dots, T$. *Then, it holds that*

$$\bar{R}_0 \le \mathcal{O}\left(\left(D^* + \frac{1}{D^*\sqrt{T}}\right)\frac{1}{\sqrt{M}}\right), \bar{R}_i \le \mathcal{O}\left(\left(D^* + \frac{1}{D^*\sqrt{T}}\right)\frac{1}{\sqrt{M}}\right),\ \forall i = 1, \dots, p.$$

Lemma 1 shows that the TAOG and TACV scales with the task-similarity D^* and this method will do well if $D^{*2} \ll \log|\mathcal{A}|$ [2]. While the knowledge of D^* can be relaxed [59, 9], running FTRL or OMD on $\{\mathbb{E}_{s\sim\nu_t^*}[D(\pi_t^*|\cdot)]\}_{t=1}^T$ requires the exact knowledge of the optimal policies π_t^* and the induced state distributions ν_t^*, which is not possible in most applications. Second, learning can be made more adaptive by considering different learning rates for different objectives (e.g. some objectives being more important) to help adapt to the dynamic regret and intra-task geometry. Moreover, critic information learned from previous tasks can also be initialized for a new unseen task, to achieve lower regret during the learning. We aim to address these challenges in the next section.

3 Provable guarantees for practical CMDP-within-online framework

3.1 Inexact CMDP-within-online framework

One of the key steps to generalize the online-within-online methodology [9, 3] to meta-SRL is to relax the assumption of accessing the true upper-bounds of intra-task performance by designing algorithms to estimate and update on their inexact versions.

Estimation of upper bounds. Once a CMDP task t is complete, the meta-learner only has access to a suboptimal policy $\hat{\pi}_t$ and the trajectory dataset $\mathcal{D}_t$ produced by some safe RL algorithm. Let $\tilde{\nu}_t$ denote the discounted state visitation distribution induced by policy $\hat{\pi}_t$. To obtain an estimate $\hat{\nu}_t$ from $\mathcal{D}_t$, state-of-the-art methods often rely on estimating discounted state visitation distribution ratios or corrections [48, 67, 47]. However, the main issues are that $\mathcal{D}_t$ is collected by multiple behavior policies during the learning period, and depending on how far these behavior policies are from the target policy, the per-step importance ratios involved in these methods may have large variance, which may result in a detrimental effect on stochastic algorithms. In this work, we use one of the methods from the distribution correction estimation (DICE) family, namely DualDICE [75], which is agnostic to the number or type of behavior policies used and does not involve any per-step importance ratios, thus is less likely to be affected by their high variance. In particular, for each state-action pair (s, a), the method aims to estimate the quantity $\omega_{\pi/\mathcal{D}_t}(s,a) = \frac{d^\pi(s,a)}{d^{\mathcal{D}_t}(s,a)}$, i.e., the likelihood that the target policy π will experience the state-action pair normalized by the probability with which the state-action pair appears in the off-policy data $\mathcal{D}_t$. Thereby, we estimate $\mathbb{E}_{\nu_t^*}[D(\pi_t^*|\pi)]$ with $\mathbb{E}_{\hat{\nu}_t}[D(\hat{\pi}_t|\pi)]$ by plugging in $\hat{\pi}_t$ from the within-task CMDP and $\hat{\nu}_t$ from DualDICE in lieu of the optimal policy π_t^* and the corresponding discounted state visitation distribution ν_t^*.

Bounding the estimation error. We breakdown the error by sources of origin:

$$\begin{aligned}\mathbb{E}_{\nu_t^*}[D(\pi_t^*|\pi)] - \mathbb{E}_{\hat{\nu}_t}[D(\hat{\pi}_t|\pi)] = &\underbrace{\mathbb{E}_{\nu_t^*}[D(\pi_t^*|\pi)] - \mathbb{E}_{\tilde{\nu}_t}[D(\pi_t^*|\pi)]}_{(A)} \\ &+ \underbrace{\mathbb{E}_{\tilde{\nu}_t}[D(\pi_t^*|\pi)] - \mathbb{E}_{\hat{\nu}_t}[D(\pi_t^*|\pi)]}_{(B)} + \underbrace{\mathbb{E}_{\hat{\nu}_t}[D(\pi_t^*|\pi)] - \mathbb{E}_{\hat{\nu}_t}[D(\hat{\pi}_t|\pi)]}_{(C)},\end{aligned} \tag{4}$$

where (A) accounts for the mismatch between the discounted state visitation distributions of an optimal policy π_t^* and a suboptimal one $\hat{\pi}_t$, (B) originates from the estimation error of DualDICE, and (C) is due to the difference between π_t^* and $\hat{\pi}_t$ measured according to $\hat{\pi}_t$. By triangle inequality, we can bound the total error by controlling each term separately. To streamline the presentation, we consider the tabular setting with softmax parameterization. First, we state the following assumptions.

Assumption 1. *For any state $s \in \mathcal{S}$, there exist positive constants C_π, L_π, μ_π such that the following hold: (1) $|D(\pi_t^*(\cdot|s)|\pi_\theta(\cdot|s))|$, $|D(\hat{\pi}_t(\cdot|s)|\pi_\theta(\cdot|s))| \le C_\pi$ for $\theta \in \Theta$; (2) $D(\pi_t^*(\cdot|s)|\pi(\cdot|s))$ is L_g-Lipschitz and L_π-smooth in $\pi(\cdot|s)$; (3) $D(\pi_t^*(\cdot|s)|\pi(\cdot|s))$ is μ_π-strongly convex in $\pi(\cdot|s)$.*

The first two conditions require restricting the policy $\pi(a|s)$ (or the parameterized $\pi_\theta(a|s)$) to be bounded away from zero for all state and action pairs; also the first condition can be satisfied for compact Θ [72, Lemma 27]. The third one requires π^* to be lower bounded; however, it is only used to derive dynamic regret bound (see, e.g., [53]) and not required for static regret. Also, more relaxed conditions are possible (see, e.g., [101, 7]).

[2] If the initial policy π_0 is an uniform distribution over $\mathcal{A}$, we have $\mathbb{E}_{s\sim\nu_t^*}[D(\pi_t^*|\pi_0)] \le \log|\mathcal{A}|$.

To bound (A), we need to control the distance between ν_t^* and $\tilde{\nu}_t$, which can be bounded by the distance between the inducing policy parameters as long as they are Lipschitz continuous [94, Lemma 3]. In addition, the bound on (C) also depends on the distance between policies. In general, controlling the distance between a policy to an optimal policy based on the suboptimality gap requires the optimization to have some curvatures around the optima (e.g., quadratic growth [35] or Hölderian growth [58]). However, to the best of knowledge of the authors, the only available results are an algorithm-dependent PL inequalities for policy gradient [72] or quadratic growth with entropy regularization [33]. Given some mild assumptions on the objective/constraint functions and policy parametrization as follows, we can show that a growth condition holds broadly for any CMDP problems.

Assumption 2. *The functions $J_{t,i}(\cdot)$ for $i = 0, 1, ..., p$ and $t \in [T]$ and parametric policy π_θ are definable in some o-minimal structure [90].*

Theorem 3.1 (KL divergence estimation error bound)**.** *The following bound holds:*

$$\begin{aligned}&|\mathbb{E}_{\nu_t^*}[D(\pi_t^*|\pi)] - \mathbb{E}_{\hat{\nu}_t}[D(\hat{\pi}_t|\pi)]|\\&= \mathcal{O}\left(\psi\left(\frac{1}{\sqrt{M}}\right) + \frac{1}{\sqrt{M}} + \sqrt{\epsilon_{opt}} + \sqrt{\epsilon_{approx}(\mathcal{F}, \mathcal{H})}\right),\end{aligned}$$

where ψ is a strictly increasing continuous function with the property that $\psi(0) = 0$ as specified in Lemma 1, $\epsilon_{approx}(\mathcal{F}, \mathcal{H})$ and ϵ_{opt} are the approximation error and optimization error of DualDICE, defined in (40) *and* (41)*, respectively.*

With the above uniform bound on estimation error, our next step is to develop inexact online gradient descent with regret bounds, which can be used to furnish the overall regret bound of the proposed inexact CMDP-within-online algorithm and the excess risk bound given in Theorems 3.2 and 3.3.

Inexact online learning with static regret bound. Recall that at each iteration t, the meta-algorithm approximates the upper-bound of the KL divergence between the optimal policy and an initial policy, $\mathbb{E}_{\nu_t^*}[D(\pi_t^*|\pi_0)]$ by plugging in a near-optimal policy $\hat{\pi}_t$ and its estimated discounted state visitation distribution $\hat{\nu}_t$, i.e., $\mathbb{E}_{\hat{\nu}_t}[D(\hat{\pi}_t|\pi_0)]$. Let ∇_t and $\hat{\nabla}_t$ denote the gradients of $\mathbb{E}_{\nu_t^*}[D(\pi_t^*|\pi_0)]$ and $\mathbb{E}_{\hat{\nu}_t}[D(\hat{\pi}_t|\pi_0)]$ at π_0, respectively. We have shown that $|\mathbb{E}_{\nu_t^*}[D(\pi_t^*|\pi_{t,0})] - \mathbb{E}_{\hat{\nu}_t}[D(\hat{\pi}_t|\pi_0)]| \leq \epsilon_t$, where ϵ_t is specified by Theorem 3.1. Also, $\hat{\nabla}_t$ is a $2\epsilon_t$-subgradient of $\mathbb{E}_{\nu_t^*}[D(\pi_t^*|\pi_0)]$, and $\|\nabla_t - \hat{\nabla}_t\| \leq C_g\epsilon_t$ for some constant C_g that depends on the smoothness parameter L_π. Thus, as is shown in the appendix, running online projected subgradient descent (OGD) on a sequence of smooth and convex loss functions with $\tilde{\epsilon}_t$-subgradient for T rounds will incur a static regret $\mathcal{O}(\sqrt{T} + \mathcal{E}_T)$, where $\mathcal{E}_T := \sum_{t=1}^T \tilde{\epsilon}_t$ is the cumulative inexactness.

Inexact online learning with dynamic regret bound. To measure the performance of our meta-algorithm in dynamic settings, we analyze the dynamic regret bound, i.e., $U_T := \sum_{t=1}^T f_t(x_t) - \sum_{t=1}^T f_t(x_t^*)$, where $x_t^* \in \arg\min_{x\in\mathcal{X}} f_t(x)$ is a sequence of local minimizers [101]. By exploiting the strong convexity, previous studies have shown that the dynamic regret can be upper bounded by the path-length of the comparator sequence, defined as $\mathcal{P}_T^* := \sum_{t=2}^T \|x_t^* - x_{t-1}^*\|$ that captures the cumulative difference between successive comparators [102]. The bound can be further improved for strongly convex functions as the minimum of the path-length and the squared path-length, $\mathcal{S}_T^* := \sum_{t=2}^T \|x_t^* - x_t\|^2$, which can be much smaller than the path-length [101]. We extend these results to the settings of inexact online gradient descent by also allowing the learner to query the inexact gradient of the function multiple times. As we show in the appendix, online projected gradient descent on a sequence of strongly convex functions with access to multiple ϵ_t-subgradients in each round can be bounded by $\mathcal{O}(\min(\mathcal{S}_T^* + \mathcal{E}_T, \mathcal{P}_T^* + \tilde{\mathcal{E}}_T))$, where $\tilde{\mathcal{E}}_T := \sum_{t=1}^T \sqrt{\epsilon}_t$ is the cumulative square root of inexactness.

3.2 Adapting to the dynamic regret and intra-task geometry

When the task-environment changes dynamically, a fixed meta-initialization ϕ may not be practically appealing, so it is natural to study dynamic regret by comparing with a potentially time-varying sequence ψ_t. Also, the tasks may share some common aspects of the optimization landscape, e.g., the constraints are harder to satisfy than optimizing the rewards, so adapting learning rates based on prior experience may further improve performance. For this purpose, we use adaptive learning rates for each objective. We denote $\alpha_{t,0}$ as the learning rate for the objective (reward), and $\{\alpha_{t,i}\}_{i=1}^p$ as the learning rates for the constraints. Then, the upper bounds for TAOG and TACV is given by:

$$U_t(\pi_{t,0}, \{\alpha_{t,i}\}_{i=0}^p) \coloneqq \frac{c_1^t}{\alpha_{t,0}} \mathbb{E}_{s\sim\nu_t^*}[D(\pi_t^*|\pi_{t,0})] + c_2^t M \sum_{i=0}^p \frac{\alpha_{t,i}^2}{\alpha_{t,0}} + \sqrt{M} \sum_{i=0}^p \frac{c_3^t \alpha_{t,i} + c_4^t \alpha_{t,i}^2}{\alpha_{t,0}},$$

where the constants are given in the appendix. The above result is derived by assuming that $\alpha_{t,i} \geq \alpha_{t,0}$ for all $i = 1, ..., p$. Intuitively, it implies that we tend to prioritize constraint satisfaction over reward maximization. Alternatively, we can get by this assumption at a cost of a slightly more complicated condition for $\{\alpha_{t,i}\}_{i=0}^p$ (see the appendix for discussions). Further, we assume that $\alpha_{t,0}$ is bounded away from 0, so $\{\alpha_{t,i}\}_{i=0}^p \in \Lambda \coloneqq \{\{\alpha'_i\}_{i=0}^p | \alpha'_i \geq \alpha'_0 \geq \zeta, \forall i = 1, ..., p\}$ for some $\zeta > 0$. Note that Λ is a convex set. Overall, the goal of the meta-learner is to make a sequence of decisions $x_t = \{\pi_{t,0} \in \Pi, \{\alpha_{t,i}\}_{i=0}^p \in \Lambda\}$ such that TAOG and TACV are minimized.

To design the adaptive algorithm, we consider the following two parallel sequences of loss functions over initial policy ϕ, $f_t^{init}(\phi) = \mathbb{E}_{\nu_t^*}[D(\pi_t^*|\phi)]$, and learning rates $\{\kappa_i\}_{i=0}^p$:

$$f_t^{sim}(\{\kappa_i\}_{i=0}^p) = \frac{c_1^t \mathbb{E}_{\nu_t^*}[D(\pi_t^*|\pi_{t,0})]}{\kappa_0} + \underbrace{c_2^t M \sum_{i=0}^p \frac{\kappa_i^2}{\kappa_0} + \sqrt{M} \sum_{i=0}^p \frac{c_3^t \kappa_i + c_4^t \kappa_i^2}{\kappa_0}}_{f_t^{rate}(\{\kappa_i\}_{i=0}^p)},$$

where c_i^t for $i = 1, .., 4$ are constants such that $f_t^{sim}(\{\alpha_{t,i}\}_{i=0}^p) = U_t(\pi_{t,0}, \{\alpha_{t,i}\}_{i=0}^p)$ matches the upper bound on the suboptimality/constraint violation of the within-task CRPO. We also denote the inexact versions $\hat{f}_t^{init}(\phi)$ and $\hat{f}_t^{sim}(\{\kappa_i\}_{i=0}^p)$ by replacing $\mathbb{E}_{\nu_t^*}[D(\pi_t^*|\phi)]$ with $\mathbb{E}_{\hat{\nu}_t}[D(\hat{\pi}_t|\phi)]$ in the above. Note that instead of running one online algorithm on U_t, we will run two online algorithms separately for the function sequences $\hat{f}_t^{init}$ and $\hat{f}_t^{sim}$ by taking actions on the initial policy and learning rates, respectively, and the overall regret can be bounded by an expression that depends on the regrets for each sequence. While the idea is inspired by [59], the proof is more involved due to the complicated form of f_t^{sim}. Let INIT and SIM be two algorithms, such that the actions $\pi_{t,0} \coloneqq \text{INIT}(t)$ are taken over $\hat{f}_t^{init}$ and the actions $\{\kappa_i\}_{i=0}^p \coloneqq \text{SIM}(t)$ are taken over $\hat{f}_t^{sim}$; these actions will then be used as policy initialization and learning rates for the next CMDP. We assume the following inexact upper bounds for each algorithm:

1. INIT: inexact dynamic regret with respect to a sequence $\{\psi_t\}_{t=1}^T \leq U_T^{init}(\psi)$,
2. SIM: inexact static regret with respect to $\{\kappa_i\}_{i=0}^p \leq U_T^{sim}(\{\kappa_i\}_{i=0}^p)$.

The following theorem bounds the TAOG and TACV regrets.

Theorem 3.2. *Let each within-task CMDP t run M steps of CRPO, initialized by policy $\pi_{t,0} \coloneqq$ $\text{INIT}(t)$ and learning rates $\{\alpha_{t,i}\}_{i=0}^p \coloneqq \text{SIM}(t)$. Let $\{\kappa_i^*\}_{i=0}^p \coloneqq \arg\min L(\{\kappa_i\}_{i=0}^p)$, where*

$$L(\{\kappa_i\}_{i=0}^p) = U_T^{sim}(\{\kappa_i\}_{i=0}^p) + \frac{U_T^{init}(\{\psi_t\}_{t=1}^T)}{\kappa_0} + \sum_{t=1}^T \left[\frac{f_t^{init}(\psi_t)}{\kappa_0} + f_t^{rate}(\{\kappa_i\}_{i=0}^p)\right], \quad (5)$$

and $\{\psi_t\}_{t=1}^T$ is any comparator sequence. Then, the following bounds on TAOG and TACV hold:

$$\bar{R}_i \leq L(\{\kappa_i^*\}_{i=0}^p), \qquad \forall\, i = 0, ..., p. \quad (6)$$

Since $f_t^{init}(\phi)$ is smooth and strongly convex by Assumption 1 and $f_t^{sim}(\{\kappa_i\}_{i=0}^p)$ is convex and smooth for $\kappa_i \in \Lambda$, we can directly apply regret bounds developed in Sec. 3.1 for running OGD as both INIT and SIM on the inexact losses $\hat{f}_t^{init}(\phi)$ and $\hat{f}_t^{sim}(\{\kappa_i\}_{i=0}^p)$, respectively.Due to the space restriction, we refer the reader to the Appendix for more discussions on how the upper bounds in (6) are related to the task-similarity or the task-relatedness in Section 2.3. Also, note that it is straightforward to extend our method to analyze task transfer bound using standard online-to-batch conversion techniques (see, e.g., [59, 9, 30]).

3.3 Meta-critic initialization under function approximation settings

We now extend the previous results to the function approximation settings and, additionally, introduce the strategy of meta-initialize the critic network. We adopt the setup in [93, 20, 36]; in particular, only ω is updated during the training while keeping b fixed. We use $f((s,a);\omega)$ as $f((s,a);\omega,b)$ for notational simplicity. We adopt the same assumptions as [93, Asm. 2, Asm. 3] (restated in appendix), except for the following.

Assumption 3. *There exists a positive constant $C_0 > 0$, such that for any $\beta \geq 0$, $x \in \mathbb{R}^d$, with policy π and $\|x\|_2 = 1$, it holds that $P(|x^\top \xi(s,a)| \leq \beta) \leq C_0\beta^2$, where $(s,a) \sim \nu_\pi$ is sampled from the discounted state visitation distribution induced by π.*

Assumption 3 implies that the distribution of $\xi(s,a)$ has a uniformly upper bounded probability density over the unit sphere, which can be satisfied for most of the ergodic Markov chain. Note that here the probability upper bound is $C_0\tau^2$, which is slightly stronger than $C_0\tau$ [93, Asm. 1].

With function approximation, the upper bound $U_t(\pi_{t,0}, \{\alpha_{t,i}\}_{i=0}^p)$ for TAOG and TACV is given by:

$$\frac{\bar{c}_1^t}{\alpha_{t,0}}\mathbb{E}_{\nu_t^*}[D(\pi_t^*|\pi_{t,0})] + \bar{c}_2^t\sum_{i=0}^p \frac{\alpha_{t,i}}{\alpha_{t,0}} + \bar{c}_3^t\sum_{i=0}^p \frac{\alpha_{t,i}^2}{\alpha_{t,0}} + \sum_{i=0}^p \sum_{m\in\mathcal{N}_{t,i}} \frac{\bar{c}_4^t\alpha_{t,i}}{\alpha_{t,0}}\|\bar{Q}_{t,\pi_m}^i - Q_{t,\pi_m}^i\|_{\nu_m},$$

where $\bar{Q}_{t,\pi_m}^i(s,a) := f_{t,i}((s,a);\omega_m)$ with ω_m produced by the TD learning algorithm [93, 20], ν_m is the discounted state visitation distribution of policy π_m obtained at the m-th within-task step, and $\bar{c}_i^t$ for $i = 1,2,3,4$ are constants specified in Theorem G.1. Observe that the last term above is due to the critic regret. Thus, to improve this term by meta-initializing the critic, we need to reveal its dependence on the critic initialization at the start of each task. On the high level, our strategy is to link the TD learning between two consecutive steps of policy update by warm-start. Specifically, recall that after each m-th step of policy update, we will perform K iterations of critic TD evaluation. Let $\omega_{t,m}^{i,k}$ denote the critic parameter for reward/constraint i at the k-th iteration after the m-th step of policy update. For any step $1 < m \leq M$, instead of starting anew, the critic can inherit the parameter from the last iteration at the previous step step, i.e. $\omega_{t,m}^{i,0} \leftarrow \bar{\omega}_{t,m-1}^i$, where $\bar{\omega}_{t,m-1}^i = \frac{1}{K}\sum_{k=1}^K \omega_{t,m-1}^{i,k}$. Let $\omega_{t,m}^{i,*}$ represent the true critic parameter for constraint i corresponding to the policy at the m-th update step. We make the following assumption on the stationary feature covariance matrix $\Sigma_m = \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \nu_m(s,a)\xi(s,a)\xi(s,a)^\top$ following [13].

Assumption 4. *The minimum eigenvalue of Σ_m is bounded below for all $m = 1,...,M$, i.e., $\lambda_{\min}(\Sigma_m) \geq 1/c_e^2$ for some positive constant $c_e > 0$.*

Then, as shown in Lemma 25, the critic error can be controlled by:

$$\|\bar{Q}_{t,\pi_m}^i - Q_{t,\pi_m}^i\|_{\nu_m} \leq c_e c_1(K,W)^{m+1}\|\omega_{t,0}^{i,*} - \omega_{t,0}^i\|_2 + \bar{c}_2(K,W), \tag{7}$$

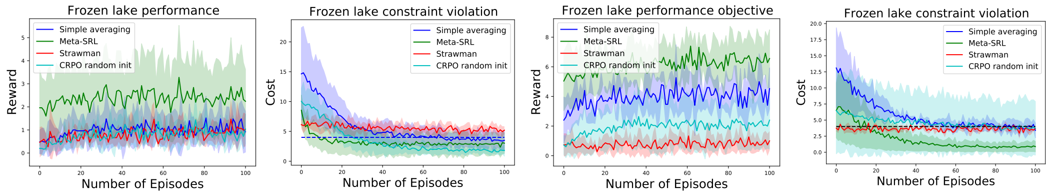
where $\omega_{t,0}^i$ is the critic initialization and $\omega_{t,0}^{i,*}$ is the optimal critic corresponding to the initial policy at the beginning of task t for constraint i, W is the width of the NN, and the term $c_1(K,W)$ diminishes at the order of $\Theta(K^{-1/4})$, and $\Theta(W^{-1/2})$.

Compared to the previous section, we need to introduce a new sequence for critic online learning. Let $\bar{f}_t^{init,a}(\phi) = \mathbb{E}_{\nu_t^*}[D(\pi_t^*|\phi)]$, $\bar{f}_t^{c,i}(\omega_t^i) = \sum_{m\in\mathcal{N}_{t,i}} \bar{c}_4^t c_1(K,L)^{m+1}\|\omega_{t,0}^{i,*} - \omega_t^i\|_2$, where $\mathcal{N}_{t,i}$ is the set of times that CRPO updates the constraints i in task t, and $\bar{f}_t^{sim}(\{\kappa_i\}_{i=0}^p)$ be given by

$$\frac{\bar{c}_1^t}{\kappa_{t,0}}\mathbb{E}_{\nu^*}[D(\pi_t^*|\pi_{t,0})]+\bar{c}_2^t\sum_{i=0}^p \frac{\kappa_{t,i}}{\kappa_{t,0}}+\bar{c}_3^t\sum_{i=0}^p \frac{\kappa_{t,i}^2}{\kappa_{t,0}}+\sum_{i=0}^p \sum_{m\in\mathcal{N}_{t,i}} \frac{\kappa_{t,i}\bar{c}_4^t c_e c_1(K,W)^{m+1}}{\kappa_{t,0}}\|\omega_{t,0}^{i,*}-\omega_{t,0}^i\|_2, \tag{8}$$

and their inexact versions $\hat{f}_t^{init,a}(\phi)$, $\hat{f}_t^{c,i}(\omega_t^i)$ and $\hat{f}_t^{sim}(\{\kappa_i\}_{i=0}^p)$ by replacing $\mathbb{E}_{\nu_t^*}[D(\pi_t^*|\phi)]$ and $\omega_{t,0}^{i,*}$ with $\mathbb{E}_{\hat{\nu}_t}[D(\hat{\pi}_t|\phi)]$ and $\omega_{t-1,M}^i$ in the above. We will run INIT^a, INIT^{c,i}, and SIM on the function sequences $\{\hat{f}_t^{init,a}\}_t$, $\{\hat{f}_t^{c,i}\}_t$ and $\{\hat{f}_t^{sim}\}_t$ in parallel to meta-initialize the actor, critic (for $i = 0,...,p$), and learning rates. Further, we assume that INIT^a has an inexact dynamic regret $U_T^{init,a}(\{\psi_t\}_{t=1}^T)$ against $\{\psi_t\}_{t=1}^T$; INIT^{c,i} has an inexact dynamic regret $U_T^{c,i}(\{\omega_t^i\}_{t=1}^T)$ against $\{\omega_t^i\}_{t=1}^T$ for $i = 0,...,p$; and SIM has an inexact static regret $U_T^{sim}(\{\kappa_i\}_{i=0}^p)$ against $\{\kappa_i\}_{i=0}^p \in \Lambda$. We derive the following upper bounds for TAOG and TACV under the function approximation setting.

Theorem 3.3. *Let each within-task CMDP t run M steps of CRPO, initialized by policy $\pi_{t,0} :=$ INIT$^a(t)$, $\omega_{t,0}^i :=$ INIT$^{c,i}(t)$ for $i = 0,...,p$, and learning rates $\{\alpha_{t,i}\}_{i=0}^p :=$ SIM(t). Let*



(a) (b) (c) (d)

Figure 1: Frozen lake results for reward maximization [*(a)*, *(c)*] and constraint violations [*(b)*, *(d)*] when the task-relatedness is high [*(a)*, *(b)*] or low [*(c)*, *(d)*], among meta-SRL and baselines. Black dashed line represents the averaged thresholds for the constraint violations.

$\{\kappa_i^*\}_{i=0}^p := \arg\min L(\{\kappa_i\}_{i=0}^p)$, *where*

$$L(\{\kappa_i\}_{i=0}^p) = \frac{U_T^{init,a}(\{\pi_t^*\}_{t=1}^T)}{\kappa_0} + \sum_{i=0}^p \frac{\kappa_i U_T^{c,i}(\{\omega_{t,0}^{i,*}\}_{t=1}^T)}{\kappa_0 M} + \sum_{t=1}^T \sum_{i=0}^p \frac{\bar{c}_2^t \kappa_i}{\kappa_0}. \tag{9}$$

Then, the following bounds on TAOG and TACV hold:

$$\bar{R}_i \leq U_T^{sim}(\{\kappa_i^*\}_{i=0}^p) + L(\{\kappa_i^*\}_{i=0}^p). \tag{10}$$

We remark that $L(\{\kappa_i\}_{i=0}^p)$ captures the dynamic regret bounds for both the actor and critic updates; by minimizing over L, we also link the overall upper bound to the static regret of learning rates. Thus, the overall regret of our meta-algorithm can be improved with increased task-relatedness.

4 Experiments

In this section, we show the effectiveness of the proposed meta-SRL method and compare with simple averaging (i.e., initialize with the average of learned policies from past CMDPs), Strawman (i.e., initialize with the learned policy from the latest CMDP), and random initialization strategies as done in CRPO. Different CMDPs are generated using a probability distribution over the parameters of CMDPs (e.g., rewards, transition dynamics), similar to the latent CMDP model [60, 22].

We consider both the Frozen lake and Acrobot environments from the OpenAI gym under constrained settings [18] (results for Acrobot are similar and discussed in the appendix, along with descriptions of the environments). For Frozen lake, we randomly generate $T = 20$ different orientations as tasks over the probability of a state being frozen or a hole, and evaluate the performance for the scenarios with high task-similarity (low variance for the latent CMDP distribution) or low task-similarity (high variance for the latent CMDP distribution).

We can observe from Figure 1 that meta-SRL achieves higher rewards and lower constraint violations more quickly than baseline initializations. Simple averaging does well on reward maximization, but note that taking the average $\bar{\pi}_t := \frac{1}{t}\sum_{t'=1}^t \hat{\pi}_{t'}$ is not following-the-average-leader, which would require taking the *weighted sum* of policies by their stationary distributions. Indeed, for Frozen lake, different locations of the hole can result in different stationary distributions—it is more sensible to put higher weights on policies that frequently visit a particular state, since it implies that the corresponding strategies can have substantial impact on rewards and constraint violations. Moreover, the main benefit of Meta-SRL can be observed from Figure 1 (b) and (d), where the Meta-SRL is able to achieve constraint satisfaction faster than baselines, even in the case of lower task-similarity conditions. This illustrates the benefit of incorporating stationary distribution correction estimation and critic meta-initialization.

5 Conclusion and future directions

This paper introduced a novel framework, meta-SRL, for meta-learning over CMDPs. The proposed framework does not assume access to globally optimal policies from the training tasks, and instead performs online learning over inexact within-task bounds estimated by stationary distribution correction. Moreover, strategies for learning rates adaptation and meta-critic initialization are designed to further exploit task-relatedness and intra-task geometry.

While the present study represents a first step in this important direction, more works are needed to further understand the limits of the approach and verify its practicality in various domains. For future research, one interesting direction is to design meta-SRL with zero constraint violation by introducing pessimism in the face of constraints [69, 68]. Other possible directions can be to improve exploration using regularization [27, 46, 91], consider generalization under adversarial scenarios [78, 70], nonstationary environments [32], and multi-agent settings [57, 29].

References

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

[2] Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98):1–76, 2021.

[3] Pierre Alquier, Massimiliano Pontil, et al. Regret bounds for lifelong learning. In *Artificial Intelligence and Statistics*, pages 261–269. PMLR, 2017.

[4] Eitan Altman. *Constrained Markov decision processes: stochastic modeling*. Routledge, 1999.

[5] Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning*, pages 322–332. PMLR, 2019.

[6] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2):235–256, 2002.

[7] Dheeraj Baby and Yu-Xiang Wang. Optimal dynamic regret in proper online learning with strongly convex losses and beyond. *arXiv preprint arXiv:2201.08905*, 2022.

[8] Maria-Florina Balcan, Avrim Blum, and Santosh Vempala. Efficient representations for lifelong learning and autoencoding. In *Conference on Learning Theory*, pages 191–210. PMLR, 2015.

[9] Maria-Florina Balcan, Mikhail Khodak, and Ameet Talwalkar. Provable guarantees for gradient-based meta-learning. In *International Conference on Machine Learning*, pages 424–433. PMLR, 2019.

[10] Maria-Florina F Balcan, Mikhail Khodak, Dravyansh Sharma, and Ameet Talwalkar. Learning-to-learn non-convex piecewise-lipschitz functions. *Advances in Neural Information Processing Systems*, 34, 2021.

[11] Amrit Singh Bedi, Paban Sarma, and Ketan Rajawat. Tracking moving agents via inexact online gradient descent algorithm. *IEEE Journal of Selected Topics in Signal Processing*, 12(1):202–217, 2018.

[12] Omar Besbes, Yonatan Gur, and Assaf Zeevi. Non-stationary stochastic optimization. *Operations research*, 63(5):1227–1244, 2015.

[13] Jalaj Bhandari, Daniel Russo, and Raghav Singal. A finite time analysis of temporal difference learning with linear function approximation. In *Conference on learning theory*, pages 1691–1692. PMLR, 2018.

[14] Shalabh Bhatnagar and K Lakshmanan. An online actor–critic algorithm with function approximation for constrained markov decision processes. *Journal of Optimization Theory and Applications*, 153(3):688–708, 2012.

[15] Jérôme Bolte, Aris Daniilidis, Adrian Lewis, and Masahiro Shiota. Clarke subgradients of stratifiable functions. *SIAM Journal on Optimization*, 18(2):556–572, 2007.

[16] Jérôme Bolte, Aris Daniilidis, Olivier Ley, and Laurent Mazet. Characterizations of łojasiewicz inequalities: subgradient flows, talweg, convexity. *Transactions of the American Mathematical Society*, 362(6):3319–3363, 2010.

[17] Vivek S Borkar. An actor-critic algorithm for constrained markov decision processes. *Systems & control letters*, 54(3):207–213, 2005.

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

[19] Archana Bura, Aria HasanzadeZonuzy, Dileep Kalathil, Srinivas Shakkottai, and Jean-Francois Chamberland. Safe exploration for constrained reinforcement learning with provable guarantees. *arXiv preprint arXiv:2112.00885*, 2021.

[20] Qi Cai, Zhuoran Yang, Jason D Lee, and Zhaoran Wang. Neural temporal-difference learning converges to global optima. *Advances in Neural Information Processing Systems*, 32, 2019.

[21] Nicolo Cesa-Bianchi, Shai Shalev-Shwartz, and Ohad Shamir. Online learning of noisy data. *IEEE Transactions on Information Theory*, 57(12):7907–7931, 2011.

[22] Xiaoyu Chen, Jiachen Hu, Chi Jin, Lihong Li, and Liwei Wang. Understanding domain randomization for sim-to-real transfer. *arXiv preprint arXiv:2110.03239*, 2021.

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

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

[25] Yinlam Chow, Ofir Nachum, Edgar A Duéñez-Guzmán, and Mohammad Ghavamzadeh. A Lyapunov-based approach to safe reinforcement learning. In *Advances in Neural Information Processing Systems*, 2018.

[26] Pádraig Cunningham, Matthieu Cord, and Sarah Jane Delany. Supervised learning. In *Machine learning techniques for multimedia*, pages 21–49. Springer, 2008.

[27] Bo Dai, Albert Shaw, Lihong Li, Lin Xiao, Niao He, Zhen Liu, Jianshu Chen, and Le Song. Sbeed: Convergent reinforcement learning with nonlinear function approximation. In *International Conference on Machine Learning*, pages 1125–1134. PMLR, 2018.

[28] Damek Davis, Dmitriy Drusvyatskiy, Sham Kakade, and Jason D Lee. Stochastic subgradient method converges on tame functions. *Foundations of computational mathematics*, 20(1):119–154, 2020.

[29] Frits De Nijs, Erwin Walraven, Mathijs De Weerdt, and Matthijs Spaan. Constrained multiagent markov decision processes: A taxonomy of problems and algorithms. *Journal of Artificial Intelligence Research*, 70:955–1001, 2021.

[30] Giulia Denevi, Carlo Ciliberto, Riccardo Grazzi, and Massimiliano Pontil. Learning-to-learn stochastic gradient descent with biased regularization. In *International Conference on Machine Learning*, pages 1566–1575. PMLR, 2019.

[31] 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. PMLR, 2021.

[32] Yuhao Ding and Javad Lavaei. Provably efficient primal-dual reinforcement learning for CMDPs with non-stationary objectives and constraints. *arXiv preprint arXiv:2201.11965*, 2022.

[33] Yuhao Ding, Junzi Zhang, and Javad Lavaei. Beyond exact gradients: Convergence of stochastic soft-max policy gradient methods with entropy regularization. *arXiv preprint arXiv:2110.10117*, 2021.

[34] Rishabh Dixit, Amrit Singh Bedi, Ruchi Tripathi, and Ketan Rajawat. Online learning with inexact proximal online gradient descent algorithms. *IEEE Transactions on Signal Processing*, 67(5):1338–1352, 2019.

[35] Dmitriy Drusvyatskiy and Adrian S Lewis. Error bounds, quadratic growth, and linear convergence of proximal methods. *Mathematics of Operations Research*, 43(3):919–948, 2018.

[36] Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. *arXiv preprint arXiv:1810.02054*, 2018.

[37] Simon Shaolei Du, Wei Hu, Sham M Kakade, Jason D Lee, and Qi Lei. Few-shot learning via learning the representation, provably. In *International Conference on Learning Representations*, 2020.

[38] Yan Duan, John Schulman, Xi Chen, Peter L Bartlett, Ilya Sutskever, and Pieter Abbeel. RL^2: Fast reinforcement learning via slow reinforcement learning. *arXiv preprint arXiv:1611.02779*, 2016.

[39] Yaqi Duan, Zeyu Jia, and Mengdi Wang. Minimax-optimal off-policy evaluation with linear function approximation. In *International Conference on Machine Learning*, pages 2701–2709. PMLR, 2020.

[40] Yonathan Efroni, Shie Mannor, and Matteo Pirotta. Exploration-exploitation in constrained MDPs. *arXiv preprint arXiv:2003.02189*, 2020.

[41] Alireza Fallah, Kristian Georgiev, Aryan Mokhtari, and Asuman Ozdaglar. On the convergence theory of debiased model-agnostic meta-reinforcement learning. *Advances in Neural Information Processing Systems*, 34, 2021.

[42] Jianqing Fan, Cong Ma, and Yiqiao Zhong. A selective overview of deep learning. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 36(2):264, 2021.

[43] 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. PMLR, 2017.

[44] Chelsea Finn, Aravind Rajeswaran, Sham Kakade, and Sergey Levine. Online meta-learning. In *International Conference on Machine Learning*, pages 1920–1930. PMLR, 2019.

[45] Javier Garcıa and Fernando Fernández. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1):1437–1480, 2015.

[46] Matthieu Geist, Bruno Scherrer, and Olivier Pietquin. A theory of regularized markov decision processes. In *International Conference on Machine Learning*, pages 2160–2169. PMLR, 2019.

[47] Carles Gelada and Marc G Bellemare. Off-policy deep reinforcement learning by bootstrapping the covariate shift. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3647–3655, 2019.

[48] Assaf Hallak and Shie Mannor. Consistent on-line off-policy evaluation. In *International Conference on Machine Learning*, pages 1372–1383. PMLR, 2017.

[49] Hado Hasselt. Double Q-learning. *Advances in neural information processing systems*, 23, 2010.

[50] Elad Hazan et al. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.

[51] Timothy Hospedales, Antreas Antoniou, Paul Micaelli, and Amos Storkey. Meta-learning in neural networks: A survey. *arXiv preprint arXiv:2004.05439*, 2020.

[52] Alexander D Ioffe. An invitation to tame optimization. *SIAM Journal on Optimization*, 19(4):1894–1917, 2009.

[53] Ali Jadbabaie, Alexander Rakhlin, Shahin Shahrampour, and Karthik Sridharan. Online optimization: Competing with dynamic comparators. In *Artificial Intelligence and Statistics*, pages 398–406. PMLR, 2015.

[54] Max Jaderberg, Wojciech M Czarnecki, Iain Dunning, Luke Marris, Guy Lever, Antonio Garcia Castaneda, Charles Beattie, Neil C Rabinowitz, Ari S Morcos, Avraham Ruderman, et al. Human-level performance in 3d multiplayer games with population-based reinforcement learning. *Science*, 364(6443):859–865, 2019.

[55] HU Jean-Baptiste. Convex analysis and minimization algorithms: advanced theory and bundle methods, 2010.

[56] Kaiyi Ji, Junjie Yang, and Yingbin Liang. Theoretical convergence of multi-step model-agnostic meta-learning. *Journal of Machine Learning Research*, 23(29):1–41, 2022.

[57] Chi Jin, Qinghua Liu, and Tiancheng Yu. The power of exploiter: Provable multi-agent rl in large state spaces. *arXiv preprint arXiv:2106.03352*, 2021.

[58] Patrick R Johnstone and Pierre Moulin. Faster subgradient methods for functions with hölderian growth. *Mathematical Programming*, 180(1):417–450, 2020.

[59] Mikhail Khodak, Maria-Florina F Balcan, and Ameet S Talwalkar. Adaptive gradient-based meta-learning methods. *Advances in Neural Information Processing Systems*, 32, 2019.

[60] Jeongyeol Kwon, Yonathan Efroni, Constantine Caramanis, and Shie Mannor. RL for latent MDPs: Regret guarantees and a lower bound. *Advances in Neural Information Processing Systems*, 34, 2021.

[61] Hoang Le, Cameron Voloshin, and Yisong Yue. Batch policy learning under constraints. In *International Conference on Machine Learning*, pages 3703–3712. PMLR, 2019.

[62] Jongmin Lee, Wonseok Jeon, Byungjun Lee, Joelle Pineau, and Kee-Eung Kim. Optidice: Offline policy optimization via stationary distribution correction estimation. In *International Conference on Machine Learning*, pages 6120–6130. PMLR, 2021.

[63] David A Levin and Yuval Peres. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.

[64] Yuanzhi Li and Yingyu Liang. Learning overparameterized neural networks via stochastic gradient descent on structured data. *Advances in Neural Information Processing Systems*, 31, 2018.

[65] Zhenguo Li, Fengwei Zhou, Fei Chen, and Hang Li. Meta-SGD: Learning to learn quickly for few-shot learning. *arXiv preprint arXiv:1707.09835*, 2017.

[66] Hao Liu, Richard Socher, and Caiming Xiong. Taming maml: Efficient unbiased meta-reinforcement learning. In *International conference on machine learning*, pages 4061–4071. PMLR, 2019.

[67] Qiang Liu, Lihong Li, Ziyang Tang, and Dengyong Zhou. Breaking the curse of horizon: Infinite-horizon off-policy estimation. *Advances in Neural Information Processing Systems*, 31, 2018.

[68] Tao Liu, Ruida Zhou, Dileep Kalathil, Panganamala Kumar, and Chao Tian. Learning policies with zero or bounded constraint violation for constrained MDPs. *Advances in Neural Information Processing Systems*, 34, 2021.

[69] Tao Liu, Ruida Zhou, Dileep Kalathil, PR Kumar, and Chao Tian. Fast global convergence of policy optimization for constrained MDPs. *arXiv preprint arXiv:2111.00552*, 2021.

[70] Thodoris Lykouris, Max Simchowitz, Alex Slivkins, and Wen Sun. Corruption-robust exploration in episodic reinforcement learning. In *Conference on Learning Theory*, pages 3242–3245. PMLR, 2021.

[71] Andreas Maurer, Massimiliano Pontil, and Bernardino Romera-Paredes. The benefit of multitask representation learning. *Journal of Machine Learning Research*, 17(81):1–32, 2016.

[72] Jincheng Mei, Chenjun Xiao, Csaba Szepesvari, and Dale Schuurmans. On the global convergence rates of softmax policy gradient methods. In *International Conference on Machine Learning*, pages 6820–6829. PMLR, 2020.

[73] Eric Mitchell, Rafael Rafailov, Xue Bin Peng, Sergey Levine, and Chelsea Finn. Offline meta-reinforcement learning with advantage weighting. In *International Conference on Machine Learning*, pages 7780–7791. PMLR, 2021.

[74] Aryan Mokhtari, Shahin Shahrampour, Ali Jadbabaie, and Alejandro Ribeiro. Online optimization in dynamic environments: Improved regret rates for strongly convex problems. In *2016 IEEE 55th Conference on Decision and Control*, pages 7195–7201. IEEE, 2016.

[75] Ofir Nachum, Yinlam Chow, Bo Dai, and Lihong Li. Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections. *Advances in Neural Information Processing Systems*, 32, 2019.

[76] Behnam Neyshabur, Zhiyuan Li, Srinadh Bhojanapalli, Yann LeCun, and Nathan Srebro. Towards understanding the role of over-parametrization in generalization of neural networks. In *International Conference on Learning Representations (ICLR)*, 2019.

[77] Richard Nock and Aditya Menon. Supervised learning: No loss no cry. In *International Conference on Machine Learning*, pages 7370–7380. PMLR, 2020.

[78] Xinlei Pan, Daniel Seita, Yang Gao, and John Canny. Risk averse robust adversarial reinforcement learning. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 8522–8528. IEEE, 2019.

[79] 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*, 2022.

[80] Ali Rahimi and Benjamin Recht. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. *Advances in neural information processing systems*, 21, 2008.

[81] Alon Resler and Yishay Mansour. Adversarial online learning with noise. In *International Conference on Machine Learning*, pages 5429–5437. PMLR, 2019.

[82] Jonas Rothfuss, Dennis Lee, Ignasi Clavera, Tamim Asfour, and Pieter Abbeel. Promp: Proximal meta-policy search. *arXiv preprint arXiv:1810.06784*, 2018.

[83] Shai Shalev-Shwartz et al. Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194, 2012.

[84] Shagun Sodhani, Amy Zhang, and Joelle Pineau. Multi-task reinforcement learning with context-based representations. In *International Conference on Machine Learning*, pages 9767–9779. PMLR, 2021.

[85] Xingyou Song, Wenbo Gao, Yuxiang Yang, Krzysztof Choromanski, Aldo Pacchiano, and Yunhao Tang. Es-maml: Simple hessian-free meta learning. *arXiv preprint arXiv:1910.01215*, 2019.

[86] Guy Tennenholtz, Uri Shalit, and Shie Mannor. Off-policy evaluation in partially observable environments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10276–10283, 2020.

[87] Philip S Thomas, Joelle Pineau, Romain Laroche, et al. Multi-objective spibb: Seldonian offline policy improvement with safety constraints in finite MDPs. *Advances in Neural Information Processing Systems*, 34, 2021.

[88] Nilesh Tripuraneni, Michael Jordan, and Chi Jin. On the theory of transfer learning: The importance of task diversity. *Advances in Neural Information Processing Systems*, 33:7852–7862, 2020.

[89] Eiji Uchibe and Kenji Doya. Constrained reinforcement learning from intrinsic and extrinsic rewards. In *2007 IEEE 6th International Conference on Development and Learning*, pages 163–168. IEEE, 2007.

[90] Lou Van den Dries and Chris Miller. Geometric categories and o-minimal structures. *Duke Mathematical Journal*, 84(2):497–540, 1996.

[91] Nino Vieillard, Tadashi Kozuno, Bruno Scherrer, Olivier Pietquin, Rémi Munos, and Matthieu Geist. Leverage the average: an analysis of kl regularization in reinforcement learning. *Advances in Neural Information Processing Systems*, 33:12163–12174, 2020.

[92] Runzhe Wu, Yufeng Zhang, Zhuoran Yang, and Zhaoran Wang. Offline constrained multi-objective reinforcement learning via pessimistic dual value iteration. *Advances in Neural Information Processing Systems*, 34, 2021.

[93] 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. PMLR, 2021.

[94] Tengyu Xu, Zhe Wang, and Yingbin Liang. Improving sample complexity bounds for (natural) actor-critic algorithms. *Advances in Neural Information Processing Systems*, 33:4358–4369, 2020.

[95] Tianbao Yang, Lijun Zhang, Rong Jin, and Jinfeng Yi. Tracking slowly moving clairvoyant: Optimal dynamic regret of online learning with true and noisy gradient. In *International Conference on Machine Learning*, pages 449–457. PMLR, 2016.

[96] Donghao Ying, Yuhao Ding, and Javad Lavaei. A dual approach to constrained markov decision processes with entropy regularization. *25th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2022.

[97] Kenny Young, Baoxiang Wang, and Matthew E Taylor. Metatrace: Online step-size tuning by meta-gradient descent for reinforcement learning control. *arXiv preprint arXiv:1805.04514*, 2018.

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

[99] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

[100] Junyu Zhang, Chengzhuo Ni, Csaba Szepesvari, Mengdi Wang, et al. On the convergence and sample efficiency of variance-reduced policy gradient method. *Advances in Neural Information Processing Systems*, 34, 2021.

[101] Lijun Zhang, Tianbao Yang, Jinfeng Yi, Rong Jin, and Zhi-Hua Zhou. Improved dynamic regret for non-degenerate functions. *Advances in Neural Information Processing Systems*, 30, 2017.

[102] Peng Zhao, Yu-Jie Zhang, Lijun Zhang, and Zhi-Hua Zhou. Dynamic regret of convex and smooth functions. *Advances in Neural Information Processing Systems*, 33:12510–12520, 2020.

[103] Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th international conference on machine learning (icml-03)*, pages 928–936, 2003.

[104] Luisa M Zintgraf, Leo Feng, Cong Lu, Maximilian Igl, Kristian Hartikainen, Katja Hofmann, and Shimon Whiteson. Exploration in approximate hyper-state space for meta reinforcement learning. In *International Conference on Machine Learning*, pages 12991–13001. PMLR, 2021.

[105] Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. Stochastic gradient descent optimizes over-parameterized deep relu networks. *arXiv preprint arXiv:1811.08888*, 2018.

Appendix

In this section, we start with a brief recapitulation of the CRPO algorithm in Sec. A, which will be our focus as the exemplary within-task safe RL algorithm. We also introduce notations therein that will be used in later analysis. In Sec. B, we give the pseudo-code of our inexact CMDP-within-online algorithm, with further discussions on key aspects. Sec. C provides the proof for Sec. 2 of the main paper, which focuses on an elementary but yet illustrative example of CMDP-within-online approach. We start by providing a simplified proof to help the reader understand the main approach of CRPO, and then demonstrate the potential improvement by exploiting inter-task relatedness (Lemma 3). Sec. D contains the key developments in extending approaches of online learning, specifically online gradient descent, to the case of inexact loss functions. We start with some preliminaries on ϵ-subgradient (Sec. D.1). Then, we conduct the analysis for static regret (Thm. D.1) and dynamic regret (Thm. D.2). In Sec. E, we provide the detailed analysis on the KL divergence estimation error bound, which contributes to one of our main contributions in understanding the key aspects of the proposed inexact CMDP-within-online framework. Our development leverages the seminal results developed for tame geometry, which we briefly review in Sec. E.1. We also briefly set up the notations and recall basic properties of subgradient flow systems E.2. Through a series of bounds, the final result is obtained in Thm. E.1. We then proceed with providing proofs for Sec. 3.2 and Sec. 3.3. In Sec. F, we first extend the analysis of CRPO to the case of adaptive learning rates (Sec. F.1). Then, we provide the proof for Thm. 3.2 in Sec. F.2. The developments are paralleled in Sec. G for the case of function approximations, where we discuss the key ideas for meta-initialization of the critic (Sec. G.2), and provide the proof for Thm. 3.3. Experimental details are provided in Sec. H.

A CRPO Algorithm and notations

We provide some preliminaries and notations for the CRPO algorithm for the sake of completeness. CRPO [93] is a primal-based CMDP algorithm, which performs policy optimization (natural gradient ascent on the reward) when constraints are not violated, or constraint minimization (natural gradient descent on the constraint function) for the corresponding violated constraint. There are three crucial components in the overall strategy to solve the CMDP problem (1):

1. **Policy evaluation:** In each step m of task t, for a certain policy $\pi_{t,m}$, the Q-value functions Q^i_{t,π_m} are estimated for the reward ($i = 0$) and constraints, for $i = 1, ..., p$.
2. **Estimation of constraint violation:** Once the Q-estimates $\bar{Q}^i_{t,\pi_m}(s, a)$ are obtained, then a weighted average is taken to estimate expected constraint violation $\bar{J}_{t,i}(\pi_{t,m})$ under a given policy $\pi_{t,m}$.
3. **Policy optimization:** After the constraint estimation, it is checked if the expected constraint violation $\bar{J}^i_{t,\pi_m}$ exceeds the given safety threshold i.e., if $\bar{J}_{t,i}(\pi_{t,m}) \leq d_{t,i} + \eta_t$ for all $i = 1, \ldots, p$. If yes, then one step of gradient ascent is done on policy to maximize the objective. If it is not satisfied, then one step of gradient descent is done to minimize one of the unsatisfied constraint.

The set of time steps the policy optimization for reward maximization takes place is denoted by $\mathcal{N}_{t,0}$, and the set of time steps reward minimization takes place is denoted by $\mathcal{N}_{t,i}$. Thus $|\mathcal{N}_{t,0}| + \sum_{i=1}^{p} |\mathcal{N}_{t,i}| = M$ for any task t.

Discrete state-action space: In the discrete state-action space, CRPO employs softmax parametrized policies . In the tabular setting, we consider the softmax $\theta \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{A}|}$, the corresponding softmax policy π_θ is defined as $\pi_\theta(a|s) := \frac{\exp(\theta(s,a))}{\sum_{a' \in \mathcal{A}} \exp(\theta(s,a'))}, \forall (s, a) \in \mathcal{S} \times \mathcal{A}$. For the critic value estimation for all objectives, TD-learning is employed [13]. The Q-function for objective i is denoted by $Q^i_{t,\pi}$ for some policy π and the Q-function parameters are denoted by ω. Learning rate for the TD-learning is denoted by β. There is a total of K iterations are done for the TD learning, and $k \in \{1, ..., K\}$ denotes the index of iteration.

Continuous state space: For the continuous state space, the policy is parameterized using a two-layer neural network together with the softmax policy $\theta(s, a) := f((s, a); \omega, b) = \frac{1}{\sqrt{W}} \sum_{\iota=1}^{W} b_\iota \cdot \mathrm{ReLU}(\omega_\iota^\top \xi(x, a))$ for any state-action pair (s, a), ι is the index for the width of the neural network,

Algorithm 1: Inexact CMDP-within-online framework (note that here CRPO [93] is considered as the within-task safe RL algorithm)

```
Initialize actor policy $\phi_1$, learning rates $\{\alpha_{1,i}\}_{i=0,\dots,p}$, and critic parameters $\{\omega_1^i\}_{i\in[p]}$
for task $t \in [T]$ do
    Run CRPO with initializations for actor policy $\phi_t$, critic network parameters $\{\omega_t^i\}_{i\in[p]}$, and learning rates $\{\alpha_{t,i}\}_{i=0,\dots,p}$, and obtain a policy $\hat{\pi}_t$
    Estimate the discounted state visitation distribution $\hat{\nu}_t$ of $\hat{\pi}_t$ based on trajectory data collected within task $t$ with DualDICE [75]
    Run one or multiple steps of OGD on
        (a) $INIT^a$: $\hat{f}_t^{init,a}(\phi)$.
        (b) SIM: $\hat{f}_t^{sim}(\{\kappa_i\}_{i=0}^p)$
        (c) $INIT^{c,i}$: $\hat{f}_t^{c,i}(\omega_t^i)$ for $i = 1, ..., p$ (in the case of critic meta-initialization)
    to obtain $\phi_{t+1}$, $\{\omega_{t+1}^i\}_{i\in[p]}$, and $\{\alpha_{t+1,i}\}_{i=0,\dots,p}$.
end for
```

$\xi(s,a) \in \mathbb{R}^d$ is the feature vector with $d \geq 2$ and $\|\xi(s,a)\| \leq 1$, $\mathrm{ReLU}(x) = \mathbb{1}(x > 0) \cdot x$, $b = [b_1, \cdots, b_W]^\top \in \mathbb{R}^W$, and $\omega = [\omega_1^\top, \cdots, \omega_W^\top]^\top \in \mathbb{R}^{Wd}$ form the set of parameters θ. We denote the state action-pair as $x = (s,a)$ and $x' = (s', a')$. We denote the stationary distribution induced by π_θ as ν_{π_θ}. We denote the TD error at iteration k by $\delta_k(x, x', \omega_k)$. Stochastic semi-gradient is denoted as $g_k(\omega_k^i) = \delta_k(x_k, x_k', \omega_k^i)\nabla_\omega f(x, \omega_k^i)$, and full semi-gradient is denoted as $\bar{g}_k(\omega_k^i) = \mathbb{E}_{\mu_\pi}[\delta_k(x, x', \omega_k^i)\nabla_\omega f(x, \omega_k^i)]$. The weighted norm is defined as $\|f\|_\mu = \sqrt{\int f(x)^2 d\mu(x)}$ for any distribution μ over X. We denote the empirical average of the Q-estimate by $\bar{J}_{t,i}(\omega_{t,m}^i) = \frac{1}{N}\sum_{j=1}^N f_i((s_j, a_j), \bar{\omega}_K^i)$, where N is a parameter of choice and $\bar{\omega}_K = \frac{1}{K}\sum_{k=0}^{K-1} \omega_k$ denotes the average of the parameters from $k = 0$ to $K - 1$.. For simplicity, we will denote ω_k^i as ω_k in some proofs if it is clear from the context.

B Inexact CMDP-within-online Algorithm

Algorithm 1 presents the inexact-CMDP-within-online algorithm for meta-SRL. The first step in the algorithm is to initialize with some random actor policy ϕ_1, critic network parameters $\{\omega_1^i\}_{i\in[p]}$, and the learning rates $\{\alpha_{1,i}\}_{i=0,\dots,p}$ for the first task. Then, for each task t, a within-task algorithm (CRPO) is run for M steps to obtain a policy $\hat{\pi}_t$. The discounted state visitation distribution $\hat{\nu}_t$ induced by $\hat{\pi}_t$ is then estimated using the trajectory data collected within task t. Afterwards, an inexact-OGD method is run on the new loss functions to update the meta-initialization policy ϕ_{t+1}, critic parameters $\{\omega_{t+1}^i\}_{i\in[p]}$ and the learning rates $\{\alpha_{t+1,i}\}_{i=0,\dots,p}$. The online learning loop is iterated for all tasks $t \in [T]$.

C Proof in Section 2

We first present a proof for the results in Equation (2):

Lemma 2. *For CRPO [93] with the softmax parameterization and the exact function estimation, if we have*

$$\eta_t \geq \frac{2}{\alpha M}\left(\mathbb{E}_{s\sim\nu_t^*}\left[D\left(\pi_t^* | \pi_{t,0}\right)\right] + \frac{2M\alpha^2|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3}\right), \tag{11}$$

then the following holds

1. $\mathcal{N}_{t,0} \neq \emptyset$, *i.e.,* $\hat{\pi}_t$ *is well-defined,*
2. $J_{t,0}\left(\pi_t^*\right) - J_{t,0}\left(\hat{\pi}_t\right) \leq \eta_t$.
3. $J_{t,i}\left(\pi_t^*\right) - J_{t,i}\left(\hat{\pi}_t\right) \leq \eta_t$, *for* $i = 1, \dots, p$.

Proof. The following inequality holds due to Lemma 7 in [93]:

$$\alpha \sum_{m\in\mathcal{N}_{t,0}} \left(J_{t,0}\left(\pi_t^*\right) - J_{t,0}\left(\pi_{t,m}\right)\right) + \alpha\eta_t \sum_{i=1}^{p} |\mathcal{N}_{t,i}| \leq \mathbb{E}_{s\sim\nu_t^*}\left[D\left(\pi_t^*|\pi_{t,0}\right)\right] + \frac{2M\alpha^2|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3}. \quad (12)$$

We first verify item 1. If $\mathcal{N}_{t,0} = \emptyset$, then $\sum_{i=1}^{p}|\mathcal{N}_{t,i}| = M$, and (12) implies that

$$\alpha\eta_t M \leq \mathbb{E}_{s\sim\nu_t^*}\left[D\left(\pi_t^*|\pi_{t,0}\right)\right] + \frac{2M\alpha^2|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3}$$

which contradicts (11). Thus, we must have $\mathcal{N}_{t,0} \neq \emptyset$.

We then proceed to verify item 2. If $\sum_{m\in\mathcal{N}_{t,0}} \left(J_{t,0}\left(\pi_t^*\right) - J_{t,0}\left(\pi_{t,m}\right)\right) > \eta_t|\mathcal{N}_{t,0}|$, then (12) implies that

$$\alpha\eta_t M \leq \mathbb{E}_{s\sim\nu_t^*}\left[D\left(\pi_t^*|\pi_{t,0}\right)\right] + \frac{2M\alpha^2|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3},$$

which contradicts (11). Hence, the item 2 holds.

Finally, the item 3 holds obviously since $\hat{\pi}_t$ is sampled from $\mathcal{N}_{t,0}$. This completes the proof. □

We now prove Lemma 1 in Section 2.

Lemma 3. *Assume $\{\nu_t^*\}_{t=1}^T$ and $\{\pi_t^*\}_{t=1}^T$ are given. For each task t, we run CRPO for M iterations with $\alpha = \frac{(1-\gamma)^{\frac{3}{2}}D^*}{\sqrt{M|\mathcal{S}||\mathcal{A}|}}$. In addition, the initialization $\{\pi_{t,0}\}_{t=1}^T$ are determined by playing Follow-the-Regularized-Leader (FTRL) or online mirror descent (OMD) [50] on the functions $\mathbb{E}_{s\sim\nu_t^*}\left[D\left(\pi_t^*|\cdot\right)\right]$, for $t = 1,\ldots,T$. Then, it holds that*

$$\bar{R}_0 \leq \mathcal{O}\left(\left(D^* + \frac{1}{D^*\sqrt{T}}\right)\frac{1}{\sqrt{M}}\right), \bar{R}_i \leq \mathcal{O}\left(\left(D^* + \frac{1}{D^*\sqrt{T}}\right)\frac{1}{\sqrt{M}}\right),\ \forall i = 1,\ldots,p.$$

Proof. By the within-task guarantee (2) for CMDP, we know that $\bar{R}_0$ and $\{\bar{R}_i\}_{i=1}^p$ are well-defined. In addition, it holds that

$$\begin{aligned}\bar{R}_0 \leq & \frac{1}{T}\sum_{t=1}^{T}\left(\frac{2\mathbb{E}_{s\sim\nu_t^*}\left[D\left(\pi_t^*|\pi_{t,0}\right)\right]}{\alpha M} + \frac{\alpha|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3}\right)\\ = & \frac{1}{T}\sum_{t=1}^{T}\left(\frac{\mathbb{E}_{s\sim\nu_t^*}\left[D_{\mathrm{KL}}\left(\pi_t^*\|\phi_t\right)\right] - \mathbb{E}_{s\sim\nu_t^*}\left[D\left(\pi_t^*|\phi^*\right)\right]}{\alpha M}\right)\\ & + \frac{1}{T}\sum_{t=1}^{T}\left(\frac{\mathbb{E}_{s\sim\nu_t^*}\left[D\left(\pi_t^*|\phi^*\right)\right]}{\alpha M} + \frac{2\alpha|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3}\right).\end{aligned}$$

where $\phi_t = \pi_{t,0}$. The first inequality follows from the choice of η_t. The key step is the last step, which splits the total loss into the loss of the meta-update algorithm and the the loss if we had always initialized at ϕ^*.

Since $\{\nu_t^*\}_{t=1}^T$ and $\{\pi_t^*\}_{t=1}^T$ are available, and each $\mathbb{E}_{s\sim\nu_t^*}\left[D\left(\pi_t^*|\cdot\right)\right]$ is strongly convex, and since each ϕ_t is determined by playing FTL or OGD, the following term is upper bounded by a sublinear term:

$$\frac{1}{T}\sum_{t=1}^{T}\left(\frac{\mathbb{E}_{s\sim\nu_t^*}\left[D\left(\pi_t^*|\phi_t\right)\right] - \mathbb{E}_{s\sim\nu_t^*}\left[D\left(\pi_t^*|\phi^*\right)\right]}{\alpha M}\right) \leq \frac{1}{\alpha M\sqrt{T}}.$$

Since $\phi^* = \arg\min_\phi \sum_{t=1}^T \mathbb{E}_{s\sim\nu_t^*}\left[D\left(\pi_t^*|\phi\right)\right]$, by the definition of D^*, we have $\mathbb{E}_{s\sim\nu_t^*}\left[D\left(\pi_t^*|\phi\right)\right] \leq D^{*2}$. Thus, by substituting the definition of ϕ^* and $\alpha = \frac{(1-\gamma)^{\frac{3}{2}}D^*}{\sqrt{M|\mathcal{S}||\mathcal{A}|}}$, it holds that

$$\frac{1}{T}\sum_{t=1}^{T}\left(\frac{\mathbb{E}_{s\sim\nu_t^*}\left[D\left(\pi_t^*|\phi^*\right)\right]}{\alpha M} + \frac{2\alpha|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3}\right) \leq \frac{D^{*2}}{\alpha M} + \frac{2\alpha|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3} \leq \frac{D^*\sqrt{|\mathcal{S}||\mathcal{A}|}}{\sqrt{M(1-\gamma)^{-3}}}.$$

The bound for $\bar{R}_i$ can be derived similarly. □

D Inexact online gradient descent

D.1 Basics for ϵ-subgradient

We start with some basics for ϵ-subdifferential used in the subsequent analysis. This material is based on [55, Chap. XI]. Throughout this section, we consider a convex, closed, and proper function $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ with domain $\mathrm{Dom}(f)$. We always consider a positive $\epsilon > 0$.

Definition 2 (ϵ-subgradient [55]). *Given $\hat{x} \in \mathrm{Dom}(f)$, the vector $u \in \mathbb{R}^d$ is called ϵ-subgradient of f at $\hat{x}$ when the following property holds for any $x \in \mathbb{R}^d$:*

$$f(x) \geq f(\hat{x}) + \langle u, x - \hat{x} \rangle - \epsilon.$$

The set of all ϵ-subgradients of f at $\hat{x}$ is the ϵ-subdifferential of f at $\hat{x}$, denoted by $\partial_\epsilon f(\hat{x})$.

In view of the exact subdifferential $\partial f(x)$, $\partial_\epsilon f(\hat{x})$ can be called an approximate subdifferential, which is a set-valued function with a convex graph. For practical use, $\partial_\epsilon f(\hat{x})$ can be used to characterize the ϵ-solution to a convex minimization problem.

Lemma 4. *([55, Thm. 1.1.5]) The following two properties are equivalent.*

$$0 \in \partial_\epsilon f(\hat{x}) \iff f(\hat{x}) \leq f(x) + \epsilon, \qquad \textit{for all } x \in \mathbb{R}^d.$$

One useful result that stems directly from the definition is to link the ϵ-subdifferential of two uniformly close functions (e.g., an expectation of a function and its empirical version).

Lemma 5. *Consider two convex functions f and g, with the property that $\|f - g\|_\infty \leq \epsilon$. Then, for any $x \in \mathbb{R}^d$ and $u \in \partial f(x)$ in the subdifferential of f at x, it is also in the 2ϵ-subdifferential of g at x, i.e., $u \in \partial_{2\epsilon} g(x)$.*

Proof. The proof follows directly by convexity and the uniform condition:

$$\begin{aligned} g(y) &\geq f(y) - \epsilon \\ &\geq f(x) + \langle s, y - x \rangle - \epsilon \\ &\geq g(x) + \langle s, y - x \rangle - 2\epsilon, \end{aligned}$$

where the second inequality is by convexity of f, and the first and last inequalities are due to the supremum norm condition. □

Our next result is concerned about bounding the distance (measured in ℓ_2 norm) between the true gradient and the ϵ-subgradient of the function, assuming the function is differentiable and smooth.

Lemma 6. *Suppose a function f is convex, differentiable, and L-smooth over $\mathrm{Dom}(f)$, and $u \in \partial_\epsilon f(x)$ is an ϵ-subgradient of f at $x \in \mathrm{Dom}(f)$. Then,*

$$\|u - \nabla f(x)\|_2^2 \leq \frac{2\epsilon}{2C_1 - C_1^2 L},$$

for any $C_1 \in \{c \in (0, \frac{2}{L}) : x + c(u - \nabla f(x)) \in \mathrm{Dom}(f)\}$. In particular, if $x + \frac{1}{L}(u - \nabla f(x)) \in \mathrm{Dom}(f)$, then $\|u - \nabla f(x)\|_2^2 \leq 2\epsilon L$.

Proof. Since s is an ϵ-gradient, $f(y) \geq f(x) + \langle u, y - x \rangle - \epsilon$ for all $y \in \mathrm{Dom}(f)$. Thus,

$$\begin{aligned} 0 &\leq f(x) - f(y) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|^2 \\ &\leq \langle \nabla f(x) - u, y - x \rangle + \frac{1}{2}\|y - x\|^2 + \epsilon \end{aligned}$$

Choose $y = x + c(u - \nabla f(x))$ for $c \in (0, \frac{2}{L})$ such that $x + c(u - \nabla f(x)) \in \mathrm{Dom}(f)$, we have that

$$\|u - \nabla f(x)\|_2^2 \leq \frac{2\epsilon}{2c - c^2 L}.$$

□

Algorithm 2: Inexact OGD Algorithm

```
Input: Learning rate $\alpha$, $x_1 = 0$
for $t = 1, .., T$ do
    Incur loss $\ell_t(x_t)$ and compute $\epsilon$-gradient $\hat{\nabla}_t \ell_t(x_t)$
    $x_{t+1} = P_X(x_t - \alpha \hat{\nabla}_t \ell_t(x_t))$
end for
```

The smoothness condition in the above seems necessary, as we can construct counterexamples that drive the distance of an ϵ-subgradient and its exact counterpart arbitrarily large without the smoothness condition. In fact, it is known that the set-valued mapping $(x, \epsilon) \to \partial_\epsilon f(x)$ is inner semi-continuous for a Lipschitz-continuous f, which is implied by the fact that the distance (using the Hausdorff distance for sets) between any two subdifferential $\partial_\epsilon f(x)$ and $\partial_{\epsilon'} f(x')$ for all $x, x' \in \mathbb{R}^d$ and ϵ, ϵ' is positive, and shown to be bounded by $\mathcal{O}\left(\frac{1}{\min\{\epsilon,\epsilon'\}}(\|x - x'\| + |\epsilon - \epsilon'|\right)$ [55, Thm. 4.1.3]. While the exact gradient can be interpreted as ϵ-subgradient driving $\epsilon \to 0^+$, the above bound is vacuous in this case.

D.2 Static regret for the inexact OGD algorithm

In the following, we consider the online learning setup, where a sequence of loss functions $\{\ell_t\}_{t\in[T]}$ are revealed sequentially, and the performance of the OGD algorithm (see Algorithm 2) is measured against a static decision in hindsight:

$$\text{(static regret)} \quad \sum_{t=1}^{T} \ell_t(x_t) - \min_{x\in X} \sum_{t=1}^{T} \ell_t(x) \tag{13}$$

where $\{x_t \in X\}_{t\in[T]}$ is a sequence of actions played by the online algorithm. For simplicity, we assume that X belongs to the domains of ℓ_t for all $t \in [T]$. Furthermore, we define the following cumulative inexact error bounds:

$$\mathcal{E}_T \coloneqq \sum_{t=1}^{T} \epsilon_t, \tag{14}$$

where ϵ_t corresponds to the inexactness of the ϵ_t-subgradient in each round of OGD.

Theorem D.1 (Static regret bound for the inexact OGD)**.** *Assume that $\{\ell_t\}_{t\in[T]}$ are convex and L_2-smooth, with bounded gradient, i.e., $\|\nabla \ell_t(x)\|_2 \leq L_1$ for all $t \in [T]$ and all $x \in X$. Then, for any comparator $x \in X$, with the stepsize $\alpha \coloneqq \frac{\|x\|}{L_1\sqrt{2T}}$, we have that*

$$\sum_{t=1}^{T} \ell_t(x_t) - \sum_{t=1}^{T} \ell_t(x) \leq L_1\|x\|\sqrt{\frac{2}{T}} + \frac{1 + \frac{\sqrt{2}cL_1L_2\|x\|}{\sqrt{T}}}{T} \sum_{t=1}^{T} \epsilon_t,$$

where ϵ_t is the amount of inexactness at each step t.

Proof. By convexity and the fact that $\hat{\nabla}_t$ is an ϵ_t-subgradient of ℓ_t at x_t, we have that

$$\ell_t(x_t) - \ell_t(x) \leq \langle \hat{\nabla}_t, x_t - x \rangle + \epsilon_t, \quad \forall x \in X$$

Hence, summing over $t = 1, ..., T$, we get

$$\frac{1}{T}\sum_{t=1}^{T} \ell_t(x_t) - \ell_t(x) \leq \frac{1}{T}\sum_{t=1}^{T} \langle \hat{\nabla}_t, x_t - x \rangle + \epsilon_t.$$

To bound the RHS, observe that

$$\begin{aligned} \|x_{t+1} - x\|^2 &\leq \|x_t - \alpha\hat{\nabla}_t - x\|^2 \\ &= \|x_t - x\|^2 - 2\alpha\langle x_t - x, \hat{\nabla}_t \rangle + \alpha^2\|\hat{\nabla}_t\|^2, \end{aligned}$$

where the first inequality is due to the OGD update rule and the nonexpansiveness of the projection operator. Thus, rearranging the terms and exploiting the telescopic sum over $t \in [T]$, we have that

$$\begin{aligned}\sum_{t=1}^{T} \langle x_t - x, \hat{\nabla}_t \rangle &\leq \frac{1}{2\alpha}(\|x_1 - x\|^2 - \|x_{T+1} - x\|^2) + \frac{\alpha}{2}\sum_{t=1}^{T} \|\hat{\nabla}_t\|^2 \\ &\leq \frac{1}{2\alpha}\|x_1 - x\|^2 + \frac{\alpha}{2}\sum_{t=1}^{T} \|\hat{\nabla}_t\|^2.\end{aligned}$$

Furthermore, since ℓ_t is L_2-smooth with bounded gradient, and $\hat{\nabla}_t$ is an ϵ_t-gradient for any $t \in [T]$, by Lemma 6, the following holds:

$$\begin{aligned}\|\hat{\nabla}_t\|^2 &\leq 2\|\nabla_t\|^2 + 2\|\nabla_t - \hat{\nabla}_t\|^2 \\ &\leq 2L_1^2 + 2cL_2\epsilon_t,\end{aligned}$$

where the constant c is specified by Lemma 6. Hence, combining the above relations, we get

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T} \ell_t(x_t) - \ell_t(x) &\leq \frac{1}{2\alpha T}\|x_1 - x\|^2 + \frac{1}{T}\sum_{t=1}^{T}\left(\frac{\alpha}{2}\|\hat{\nabla}_t\|^2 + \epsilon_t\right) \\ &\leq \frac{1}{2\alpha T}\|x_1 - x\|^2 + \alpha L_1^2 + \left(\frac{\alpha c L_2 + 1}{T}\right)\sum_{t=1}^{T}\epsilon_t.\end{aligned}$$

Let $\alpha = \frac{\|x\|}{L_1\sqrt{2T}}$, then we get the RHS as

$$L_1\|x\|\sqrt{\frac{2}{T}} + \frac{1 + \frac{\sqrt{2}cL_1L_2\|x\|}{\sqrt{T}}}{T}\sum_{t=1}^{T}\epsilon_t.$$

□

Remark 1. *We can relax the dependence of setting the stepsize on T by using a standard doubling trick (first proposed in [6], see also, e.g., [9, 59]).*

D.3 Dynamic regret for the inexact OGD algorithm

In the following, we consider a stronger notion of regret that measures the performance of the OGD algorithm (see Algorithm 2) against a dynamically changing sequence in hindsight (see, e.g., [103, 53, 101]):

$$\text{(dynamic regret)} \quad \sum_{t=1}^{T} \ell_t(x_t) - \sum_{t=1}^{T} \ell_t(x_t^*) \tag{15}$$

where $x_t^* \in \arg\min_{x \in X} \ell_t(x)$ is the optimal decision for the loss ℓ_t. It is well-known that in the worst-case, it is impossible to achieve a sub-linear dynamic regret bound, due to the arbitrary fluctuation in the functions [103, 12, 95]. Thus, it is common to upper bound the dynamic regret in terms of certain regularity of the comparator sequence. One possible regularity condition is the path-length of the comparator sequence [103, 53]:

$$\mathcal{P}_T := \sum_{t=2}^{T} \|x_t^* - x_{t-1}^*\|, \tag{16}$$

which captures the cumulative Euclidean norm of the difference between successive comparators (note that we will use $\|\cdot\|$ for the Euclidean norm, unless otherwise specified). The path-length measure is also the regularity condition used in existing inexact OGD literature [11, 34]. However, as remarked in [101], a potentially tighter bound can be achieved by examining the squared path-length measure:

$$\mathcal{S}_T := \sum_{t=2}^{T} \|x_t^* - x_{t-1}^*\|^2, \tag{17}$$

which can be much smaller than $\mathcal{P}_T$ when the local variations are small. For example, when $\|x_t^* - x_{t-1}^*\| = \Theta(1/\sqrt{T}$ for all $t \in [T]$, we have $\mathcal{P}_T = \Theta(\sqrt{T})$ but $\mathcal{S}_T = \Theta(1)$. In this section, we provide analysis with respect to both measures for strongly convex and smooth functions. Furthermore, we propose to apply inexact OGD multiple times in each round, and demonstrate that the dynamic regret is reduced from $\mathcal{O}(\mathcal{P}_T + \mathcal{E}_T)$ to $\mathcal{O}(\min\{\mathcal{P}_T + \mathcal{E}_T, \mathcal{S}_T + \tilde{\mathcal{E}}_T\})$, where

$$\tilde{\mathcal{E}}_T \coloneqq \sum_{t=1}^{T} \sqrt{\epsilon_t}$$

is the cumulative square root inexactness bounds. Note that our results improve over existing bounds for inexact online learning [11, 34] and can be regarded as a generalization of [101] to the inexact settings. We start with a result that will be used in later analysis.

Lemma 7. *Assume that $f : X \to \mathbb{R}$ is λ-strongly convex and L-smooth, and let $x^* = \underset{x \in X}{\arg\min} f(x)$ be the unique optimal solution. Let $v = P_X(x - \alpha \hat{\nabla} f(x))$, where $\hat{\nabla} f(u) \in \partial_\epsilon f(u)$ and $\alpha \leq \frac{1}{2L}$, we have that*

$$\|v - x^*\|^2 \leq \frac{1}{\lambda\alpha + 1}\|x^* - x\|^2 + \frac{c\alpha + 2L\alpha}{\lambda L\alpha + L}\epsilon,$$

where the constant c is specified in Lemma 6.

Proof. By the update rule, we have that

$$v = \underset{x' \in X}{\arg\min} f(x) + \langle \hat{\nabla} f(x), x' - x \rangle + \frac{1}{2\alpha}\|x' - x\|^2. \tag{18}$$

By strong convexity of the objective above,

$$\langle \hat{\nabla} f(x), v - x \rangle + \frac{1}{2\alpha}\|v - x\|^2 \leq \langle \hat{\nabla} f(x), x^* - x \rangle + \frac{1}{2\alpha}\|x^* - x\|^2 - \frac{1}{2\alpha}\|v - x^*\|^2. \tag{19}$$

Since, $f(x)$ is λ-strongly convex and L-smooth, we have that

$$f(x^*) - \frac{\lambda}{2}\|x^* - x\|^2 \geq f(x) + \langle \nabla f(x), x^* - x \rangle, \tag{20}$$

and

$$f(x^*) \leq f(x) + \langle \nabla f(x), x^* - x \rangle + \frac{L}{2}\|x^* - x\|^2. \tag{21}$$

Also, since $\hat{\nabla} f(x)$ is an ϵ-subgradient, we can write

$$f(x^*) \geq f(x) + \langle \hat{\nabla} f(x), x^* - x \rangle - \epsilon. \tag{22}$$

Combining (20), (21) and (22), we have that

$$f(x^*) + \frac{L - \lambda}{2}\|x^* - x\|^2 \geq f(x) + \langle \hat{\nabla} f(x), x^* - x \rangle - \epsilon.$$

Algorithm 3: Inexact Online Multiple Gradient Descent Algorithm

```
Input: Learning rate $\alpha$, $x_1 = 0$
for $t = 1, .., T$ do
    Incur loss $\ell_t(x_t)$
    $z_t^1 = x_t$
    for k = 1,...,K do
        $z_t^{k+1} = P_X(x_t - \alpha\hat{\nabla}\ell_t(z_t^k))$
    end for
    $x_{t+1} = z_t^{K+1}$
end for
```

Combining the above relations, we have that

$$
\begin{aligned}
f(v) &\leq f(x) + \langle \nabla f(x), v - x\rangle + \frac{L}{2}\|v - x\|^2 \\
&= f(x) + \langle \hat{\nabla} f(x), v - x\rangle + \frac{L}{2}\|v - x\|^2 + \langle \nabla f(x) - \hat{\nabla} f(x), v - x\rangle \\
&\overset{(i)}{\leq} f(x) + \langle \hat{\nabla} f(x), x^* - x\rangle + \left(\frac{L}{2} - \frac{1}{2\alpha}\right)\|v - x\|^2 \\
&\qquad\qquad + \frac{1}{2\alpha}\|x^* - x\|^2 - \frac{1}{2\alpha}\|v - x^*\|^2 + \langle \nabla f(x) - \hat{\nabla} f(x), v - x\rangle \\
&\overset{(ii)}{\leq} f(x^*) + \left(\frac{L}{2} - \frac{1}{2\alpha}\right)\|v - x\|^2 + \frac{1}{2\alpha}\|x^* - x\|^2 \\
&\qquad\qquad - \frac{1}{2\alpha}\|v - x^*\|^2 + \langle \nabla f(x) - \hat{\nabla} f(x), v - x\rangle + \epsilon \\
&\overset{(iii)}{\leq} f(v) - \left(\frac{\lambda}{2} + \frac{1}{2\alpha}\right)\|v - x^*\|^2 + \left(\frac{L}{2} - \frac{1}{2\alpha}\right)\|v - x\|^2 \\
&\qquad\qquad + \frac{1}{2\alpha}\|x^* - x\|^2 + \langle \nabla f(x) - \hat{\nabla} f(x), v - x\rangle + \epsilon \\
&\overset{(iv)}{\leq} f(v) - \left(\frac{\lambda}{2} + \frac{1}{2\alpha}\right)\|v - x^*\|^2 + \left(\frac{L}{2} - \frac{1}{2\alpha}\right)\|v - x\|^2 \\
&\qquad\qquad + \frac{1}{2\alpha}\|x^* - x\|^2 + \|\nabla f(x) - \hat{\nabla} f(x)\|\|v - x\| + \epsilon \\
&\overset{(v)}{\leq} f(v) - \left(\frac{\lambda}{2} + \frac{1}{2\alpha}\right)\|v - x^*\|^2 \\
&\qquad\qquad + \left(\frac{L}{2} - \frac{1}{2\alpha} + \frac{\kappa}{2}\right)\|v - x\|^2 + \frac{1}{2\alpha}\|x^* - x\|^2 + \left(\frac{c}{2\kappa} + 1\right)\epsilon,
\end{aligned}
$$

where the first inequality is due to L-smoothness, (i) follows from (19), (ii) is due to convexity, (iii) is due to strong convexity, (iv) follows from Cauchy-Schwarz inequality, and (v) is due to the inequality $ab \leq \frac{1}{2\kappa}a^2 + \frac{\kappa}{2}b^2$ for $a, b \geq 0$ and $\kappa > 0$ and the constant c comes from Lemma 6. Choosing $\kappa = L$, $\alpha \leq \frac{1}{2L}$, and rearranging the above, we have then proved the claim. □

Theorem D.2 (Dynamic regret for inexact OGD with multiple updates)**.** *Assume that $\ell_t : X \to \mathbb{R}$ is λ-strongly convex, L_1-Lipschitz, and L_2-smooth for all $t \in [T]$. By setting $\alpha \leq \frac{1}{2L_2}$, $K := \left\lceil \frac{\ln 2}{\ln(1+\lambda\alpha)} \right\rceil$, then, for any $\beta > 0$, we have that*

$$
\begin{aligned}
\sum_{t=1}^{T} \ell_t(x_t) - \ell_t(x_t^*) \leq \min\Bigg(&C_1\|x_1 - x_1^*\|^2 + C_2\mathcal{E}_T + C_3 S_T^* + \frac{1}{2\beta}\sum_{t=1}^{T}\|\nabla \ell_t(x_t^*)\|^2, \\
&C_4\|x_1 - x_1^*\| + C_5\sum_{t=1}^{T}\sqrt{\epsilon_t} + C_4 P_T^*\Bigg),
\end{aligned}
$$

where $C_1 = 2(L_2+\beta)$*,* $C_2 = (L_2+\beta)\frac{3c\alpha+6\alpha L_2}{2\lambda\alpha L_2}$*,* $C_3 = 3(L_2+\beta)$*,* $C_4 = \frac{2L_1}{2-\sqrt{2}}$ *and* $C_5 = \frac{2L_1}{2-\sqrt{2}}\sqrt{\frac{c\alpha+2L_2\alpha}{2\alpha\lambda L_2}}$*.*

Proof. The proof has two parts, where we use different techniques to bound the dynamic regret by $\mathcal{S}_T$ and $\mathcal{E}_T$, as well as $\mathcal{P}_T$ and $\tilde{\mathcal{E}}_T$. Then the final bound is obtained by taking the minimum between the two bounds.

Bounding the dynamic regret by $\mathcal{S}_T$ and $\mathcal{E}_T$. Since ℓ_t is L_2-smooth, we have that

$$\ell_t(x_t) - \ell_t(x_t^*) \le \langle \nabla \ell_t(x_t^*), x_t - x_t^* \rangle + \frac{L_2}{2}\|x_t - x_t^*\|^2 \tag{23}$$

$$\le \|\nabla \ell_t(x_t^*)\|\|x_t - x_t^*\| + \frac{L_2}{2}\|x_t - x_t^*\|^2 \tag{24}$$

$$\le \frac{1}{2\beta}\|\nabla \ell_t(x_t^*)\|^2 + \frac{L_2+\beta}{2}\|x_t - x_t^*\|^2, \tag{25}$$

where the second inequality is due to Cauchy–Schwartz and the third inequality is due to $ab \le \frac{1}{2\beta}a^2 + \frac{\beta}{2}b^2$ for $a, b \ge 0$ and $\beta > 0$.

Now, using $\|x-y\|^2 \le (1+\iota)\|x-z\|^2 + \left(1+\frac{1}{\iota}\right)\|z-y\|^2$, we can bound

$$\sum_{t=1}^T \|x_t - x_t^*\|^2 \le \|x_1 - x_1^*\|^2 + \sum_{t=2}^T (1+\iota)\|x_t - x_{t-1}^*\|^2 + \left(1+\frac{1}{\iota}\right)\|x_t^* - x_{t-1}^*\|^2. \tag{26}$$

Recall the updating rule $z_{t-1}^{j+1} = P_X(z_{t-1}^j - \alpha\hat{\nabla} f_{t-1}(z_{t-1}^j))$, $j = 1, ..., K$; then, we can write that

$$\|x_t - x_{t-1}^*\|^2 = \|z_{t-1}^{K+1} - x_{t-1}^*\|^2 \tag{27}$$

$$\le \left(\frac{1}{\lambda\alpha+1}\right)^K \|x_{t-1} - x_{t-1}^*\|^2 + \frac{1-\left(\frac{1}{\lambda\alpha+1}\right)^K}{1-\frac{1}{\lambda\alpha+1}}\frac{c\alpha+2L_2\alpha}{\lambda L_2\alpha + L_2}\epsilon_{t-1},$$

where we recursively apply the result from Lemma 7. Thus, by plugging in (27) into (26), and using the definitions of $\mathcal{S}_T$ and $\mathcal{S}_T$, we have that

$$\sum_{t=1}^T \|x_t - x_t^*\|^2 \le \|x_1 - x_1^*\|^2 + (1+\iota)\left(\frac{1}{\lambda\alpha+1}\right)^K \sum_{t=1}^T \|x_t - x_t^*\|^2 \tag{28}$$

$$+ (1+\iota)\frac{1-\left(\frac{1}{\lambda\alpha+1}\right)^K}{1-\frac{1}{\lambda\alpha+1}}\frac{c\alpha+2L_2\alpha}{\lambda L_2\alpha+L_2}\mathcal{E}_T + \left(1+\frac{1}{\iota}\right)\mathcal{S}_T.$$

Rearranging the terms, the above relation implies that

$$\sum_{t=1}^T \|x_t - x_t^*\|^2 \le \frac{(1+\lambda\alpha)^K}{(1+\lambda\alpha)^K - (1+\iota)}\|x_1 - x_1^*\|^2 + \left(1+\frac{1}{\iota}\right)\frac{(1+\lambda\alpha)^K}{(1+\lambda\alpha)^K-(1+\iota)}\mathcal{S}_T$$

$$+ (1+\iota)\frac{(1+\lambda\alpha)^K - 1}{(1+\lambda\alpha)^K - (1+\iota)}\frac{c\alpha+2L_2\alpha}{\lambda L_2\alpha}\mathcal{E}_T$$

Let $\iota = \frac{1}{2}$ and choose $K = \lceil \frac{\log 2}{\log(1+\lambda\alpha)} \rceil$, we have

$$\sum_{t=1}^T \|x_t - x_t^*\|^2 \le 4\|x_1 - x_1^*\|^2 + \frac{3c\alpha + 6L_2\alpha}{\lambda\alpha L_2}\mathcal{E}_T + 6\mathcal{S}_T.$$

Combine the above with (25), and summing over $t \in [T]$, we have that

$$\sum_{t=1}^T \ell_t(x_t) - \ell_t(x_t^*)$$

$$\le \frac{1}{2\beta}\sum_{t=1}^T \|\nabla \ell_t(x_t^*)\|^2 + 3(L_2+\beta)\mathcal{S}_T + (L_2+\beta)\frac{3c\alpha+6L\alpha}{2\lambda\alpha L}\mathcal{E}_T + 2(L_2+\beta)\|x_1 - x_1^*\|^2,$$

which holds true for any positive $\beta > 0$.

Bounding the dynamic regret by $\mathcal{P}_T$ and $\tilde{\mathcal{E}}_T$. By (28) and the choice of $K = \lceil \frac{\log 2}{\log(1+\lambda\alpha)} \rceil$, we have that:

$$\|x_t - x_{t-1}^*\|^2 \leq \frac{1}{2}\|x_{t-1} - x_{t-1}^*\|^2 + \frac{c\alpha + 2L_2\alpha}{2\alpha\lambda L_2}\epsilon_{t-1}.$$

Thus,

$$\begin{aligned} \|x_t - x_{t-1}^*\| &\leq \sqrt{\frac{1}{2}\|x_{t-1} - x_{t-1}^*\|^2 + \frac{c\alpha + 2L_2\alpha}{2\alpha\lambda L_2}\epsilon_{t-1}} \\ &\leq \frac{1}{\sqrt{2}}\|x_{t-1} - x_{t-1}^*\| + \sqrt{\frac{c\alpha + 2L_2\alpha}{2\alpha\lambda L_2}}\sqrt{\epsilon_{t-1}}, \end{aligned} \tag{29}$$

where the last inequlity follows from $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$. Due to the bounded gradient assumption, we have that

$$\sum_{t=1}^{T} \ell_t(x_t) - \ell_t(x_t^*) \leq L_1 \sum_{t=1}^{T} \|x_t - x_t^*\| \tag{30}$$

To bound $\sum_{t=1}^{T} \|x_t - x_t^*\|$, notice that

$$\begin{aligned} \sum_{t=1}^{T} \|x_t - x_t^*\| &\leq \|x_1 - x_1^*\| + \sum_{t=2}^{T} \|x_t - x_{t-1}^*\| + \|x_{t-1}^* - x_t^*\| \\ &\leq \|x_1 - x_1^*\| + \frac{1}{\sqrt{2}}\sum_{t=1}^{T} \|x_t - x_t^*\| + \sqrt{\frac{c\alpha + 2L_2\alpha}{2\alpha\lambda L_2}}\tilde{\mathcal{E}}_T + \mathcal{P}_T, \end{aligned}$$

which implies that

$$\sum_{t=1}^{T} \|x_t - x_t^*\| \leq \frac{2}{2-\sqrt{2}}\|x_1 - x_1^*\| + \frac{2}{2-\sqrt{2}}\sqrt{\frac{c\alpha + 2L_2\alpha}{2\alpha\lambda L_2}}\tilde{\mathcal{E}}_T + \frac{2}{2-\sqrt{2}}\mathcal{P}_T.$$

Plugging the above in (30) proves the claim. □

In the above result, the number of OGD updates per round is on the order of $\mathcal{O}(L_2/\alpha)$, where L_2/α is the condition number of each loss function. Below, we also provide a dynamic regret bound for standard OGD (single update per round); as a result, we only provide the bound in terms of $\mathcal{P}_T$ (similar to [53, 74].

E KL divergence estimation error bound

We recall the following notations. For each task t, the initial state distribution is denoted by ρ_t, the state distribution for the optimal policy π_t^* is given by ν_t^*, the state distribution for the policy $\hat{\pi}_t$ is denoted by $\tilde{\nu}_t$, and the state distribution estimated using the trajectory sample dataset $\mathcal{D}_t$ is denoted as $\hat{\nu}_t$.

In the main paper, we breakdown the KL divergence estimation error by the sources of origin:

$$\begin{aligned} \mathbb{E}_{\nu_t^*}[D(\pi_t^*|\pi)] - \mathbb{E}_{\hat{\nu}_t}[D(\hat{\pi}_t|\pi)] = &\underbrace{\mathbb{E}_{\nu_t^*}[D(\pi_t^*|\pi)] - \mathbb{E}_{\tilde{\nu}_t}[D(\pi_t^*|\pi)]}_{(A)} \\ &+ \underbrace{\mathbb{E}_{\tilde{\nu}_t}[D(\pi_t^*|\pi)] - \mathbb{E}_{\hat{\nu}_t}[D(\pi_t^*|\pi)]}_{(B)} + \underbrace{\mathbb{E}_{\hat{\nu}_t}[D(\pi_t^*|\pi)] - \mathbb{E}_{\hat{\nu}_t}[D(\hat{\pi}_t|\pi)]}_{(C)}, \end{aligned} \tag{31}$$

where (A) accounts for the mismatch between the discounted state visitation distributions of an optimal policy π_t^* and a suboptimal one $\hat{\pi}_t$, (B) originates from the estimation error of DualDICE, and (C) is due to the difference between π_t^* and $\hat{\pi}_t$ measured according to $\hat{\pi}_t$. By triangle inequality, we can bound the total error by controlling each term separately. To streamline the presentation, we consider the tabular setting with softmax parameterization.

To bound (A), we need to control the distance between ν_t^* and $\tilde{\nu}_t$, which can be bounded by the distance between the inducing policy parameters as long as they are Lipschitz continuous [94, Lemma 3]. In addition, the bound on (C) also depends on the distance between policies. In general, controlling the distance between a policy to an optimal policy based on the suboptimality gap requires the optimization to have some curvatures around the optima (e.g., quadratic growth [35] or Hölderian growth [58]). However, to the best of knowledge of the authors, the only available results are an algorithm-dependent PL inequalities for policy gradient [72] or quadratic growth with entropy regularization [33].

E.1 Preliminaries on tame geometry

For the sake of completeness, let us recall some fundamental concepts/results in tame geometry, which allows us to study the global geometry of the solution maps of a wide range of optimization problems, which will be used in bounding the estimation error for the KL divergence. More information can be found in [28, 90]. Recall that a class of functions on a bounded set is called C^p smooth when it possesses the uniformly bounded partial derivatives up to order p.

Definition 3 (Whitney Stratification). *A Whitney C^k stratification of a set I is a partition of I into finitely many nonempty C^k manifolds, called strata, satisfying the following compatibility conditions:*

1. *For any two strata I_a and I_b, the implication $I_a \cap I_b \neq \emptyset$ implies that $I_a \subset \mathrm{cl} I_b$ holds, where $\mathrm{cl} I_b$ denotes the closure of the set I_b.*
2. *For any sequence of points x_k in a stratum I_a, converging to a point $x^\star$ in a stratum I_b, if the corresponding normal vectors $v_k \in N_{I_a}(x_k)$ converge to a vector v, then the inclusion $v \in N_{I_b}(x^\star)$ holds. Here $N_{I_a}(x_k)$ denotes the normal cone to I_a at x_k.*

Roughly speaking, stratification is a locally finite partition of a given set into differentiable manifolds, which fit together in a regular manner (property 1 in Def. 3). Whitney stratification as defined above is a special type of stratification for which the strata are such that their tangent spaces (as viewed from normal cones) also fit regularly (property 2).

There are several paths to verifying Whitney stratifiability. For instance, one can show that the function under study belongs to one of the well-known function classes, such as semialgebraic functions [28], whose members are known to be Whitney stratifiable. However, to study the solution function of a general convex optimization problem, we need a far-reaching axiomatic extension of semialgebraic sets to classes of functions definable on "o-minimal structures," which are very general classes and share several attractive analytic features as semialgebraic sets, including Whitney stratifiability [28, 90].

Definition 4 (o-minimal structure). *[90] An o-minimal structure is defined as a sequence of Boolean algebras O_v of subsets of $\mathbb{R}^v$, such that for each $n_v \in \mathbb{N}$, the following properties hold:*

1. *If some set X belongs to O_v, then $X \times \mathbb{R}$ belong to O_{v+1}.*
2. *Let $P_{proj} : \mathbb{R}^v \times \mathbb{R} \to \mathbb{R}^v$ denote the coordinate projection operator onto $\mathbb{R}^v$, then for any X in O_{v+1}, the set $P_{proj}(X)$ belongs to O_v.*
3. *O_v contains all sets of the form $\{x \in \mathbb{R}^v : \ y(x) = 0\}$, where $y(x)$ is a polynomial in $\mathbb{R}^v$.*
4. *The elements of O_1 are exactly the finite unions of intervals (possibly infinite) and points.*

Then all the sets that belong to O_v are called definable in the o-minimal structure.

Definable sets have broader applicability than semialgebraic sets (in the sense that the latter is a special kind of definable sets) but enjoys the same, remarkable stability property: the composition of definable mappings (including sum, inf-convolution, and several other classical operations of analysis involving a finite number of definable objects) in some o-minimal structure remains in the same structure. We will crucially exploit these properties in the following sections.

E.2 Basic properties of subgradient flow systems

We also recall some basic definitions and properties of the subgradient flow system (see, e.g., [16, Thm. 13]). Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a proper lower semicontinuous function.

Definition 5 (Subgradient flow system). *For every $x \in \operatorname{dom}(f)$, there exists a unique absolutely continuous curve (called trajectory or subgradient curve) $\theta(\tau) : [0, +\infty) \to \mathbb{R}^d$ that satisfies*

$$\begin{cases} \dot{\theta}(\tau) \in -\partial f(\theta(\tau)) & \text{a.e. on} \quad (0, +\infty) \\ \theta(0) = \theta_0 \in \operatorname{dom}(f). \end{cases} \tag{32}$$

Moreover, the trajectory also satisfies the following properties [16, Thm. 13]:

1. $\theta(\tau) \in \operatorname{dom}(\partial f)$ for all $\tau \in (0, +\infty)$.
2. For all $\tau > 0$, the right derivative $\dot{\theta}(\tau^+)$ is well defined and equal to

   $$\dot{\theta}(\tau^+) = -\partial^0 f(\theta(\tau)),$$

   where $\partial^0 f(\theta)$ is the minimum norm subgradient in $\partial f(\theta)$. In particular, we have that $\dot{\theta}(\tau) = -\partial^0 f(\theta(\tau))$, for almost all τ.

E.3 Bounding the distance $\|\hat{\theta}_t - \theta_t^*\|$

Recall Assumption 2, which requires that the objective/constraint functions and policy parametrization are definable in some o-minimal structure [90]. This is a mild assumption as practically all functions from real-world applications, including deep neural networks, are definable in some o-minimal structure [28]; also, the composition of mappings, along with the sum, inf-convolution, and several other classical operations of analysis involving a finite number of definable objects in some o-minimal structure remains in the same structure [90]. The far reaching consequence of definability, exploited in this study, is that definable sets and functions admit, for each $k \geq 1$, a C^k–Whitney stratification with finitely many strata (see, for instance, [90, Result 4.8]). This remarkable property, combined with the result that any stratifiable functions enjoys a nonsmooth Kurdyka–Łojasiewicz inequality [15], provides the foundation to bound the distance $\|\pi_t^* - \hat{\pi}_t\|$ by the suboptimality gap. Note that without further specifications, π_t^* is understood as one of optimal policies that are closest to the policy $\hat{\pi}_t$ (i.e., the projection of $\hat{\pi}_t$ onto the optimal policy set).

We start with the following elementary result. Here and throughout the section, we use $\mathcal{F}_{t,\tilde{d}} = \{\pi_{t,\theta} : J_{t,i}(\pi_{t,\theta}) \leq \tilde{d}_{t,i}\}$ to denote the feasible set with upper bounds $\tilde{d}$. Note that $\mathcal{F}_{t,d}$ is the original feasible set. We also let $\mathbb{I}_{\mathcal{F}_{t,\tilde{d}}}(\cdot)$ be the indicator function for the set $\mathcal{F}_{t,\tilde{d}}$.

Lemma 8. *The function (with variable θ) $J_{t,0}(\pi_{t,\theta}) + \mathbb{I}_{\mathcal{F}_{t,\tilde{d}}}(\pi_{t,\theta})$, where $\tilde{d}_t$ is any vector such that $\mathcal{F}_{t,\tilde{d}}$ is non-empty, is definable.*

Proof. Since $J_{t,i}(\cdot)$ is definable for $i = 1, ..., p$, by the rule of composition, which is due to the definable counterpart of the Tarski-Seidenberg theorem, $J_{t,i}(\pi_{t,\theta}) - d_{t,i}$ is definable for $i = 1, ..., p$. Thus, $\mathcal{F}_{t,\tilde{d}}$ is definable on the same o-minimal structure by the definition. Furthermore, $\mathbb{I}_{\mathcal{F}_{t,\tilde{d}}}(\cdot)$ is definable as the indicator of $\mathcal{F}_{t,\tilde{d}}$. The definability of $J_{t,0}(\pi_{t,\theta})$ follows similarly. Since definability is preserved under addition, the function $J_{t,0}(\pi_{t,\theta}) + \mathbb{I}_{\mathcal{F}_{t,\tilde{d}}}(\pi_{t,\theta})$ is definable. □

For convenience of the reader, we restate the result for non-smooth Kurdyka–Łojasiewicz (KL) inequality from [15, Thm. 14].

Proposition 1 (Non-smooth Kurdyka–Łojasiewicz inequality). *Let f be a lower semicontinuous definable function. There esists $\rho > 0$, a strictly increasing continuous definable function $\psi : [0, \rho] \to (0, \infty)$ which is C^1 smooth on $(0, \rho)$, with $\psi(0) = 0$, and a continuous definable function $\mathcal{X} : \mathbb{R}_+ \to (0, \rho)$ such that*

$$\|\partial^0 f(x)\| \geq \frac{1}{\psi'(|f(x)|)},$$

whenever $0 < |f(x)| \leq \mathcal{X}(\|x\|)$.

Let θ and θ_t^* denote the parameters of a policy π_θ and π_t^*, respectively. Directly bounding the distance between θ and θ_t^* is difficult, because π may be infeasible (this is even true for $\hat{\pi}_t$, since it is only guaranteed to approximately satisfy the constraints), i.e., $\theta \notin \mathcal{F}_{t,d}$. Thus, the typical approach of

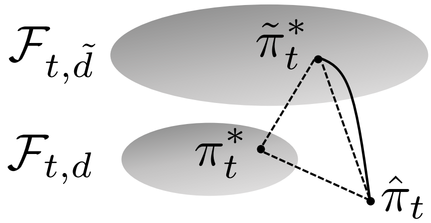


Figure 2: To bound the distance between π_t^* and $\hat{\pi}_t$, we first bound the distance between $\tilde{\pi}_t^*$ and the optimal policy with respect to a larger feasible set $\mathcal{F}_{t,\tilde{d}}$ by an argument based on subgradient flow curve. Note that $\hat{\pi}_t \in \mathcal{F}_{t,\tilde{d}}$ may be infeasible with respect to the original set of constraints but feasible with respect to the relaxed constraints. We then bound the distance between the optimal policies π_t^* and $\tilde{\pi}_t^*$, which correspond to the original feasible set $\mathcal{F}_{t,d}$ and the enlarged set $\mathcal{F}_{t,\tilde{d}}$. By triangle inequality, we can then derive the desired bound on the distance between π_t^* and $\hat{\pi}_t$. Note that for better visualization, we vertically separate the sets $\mathcal{F}_{t,d}$ and $\mathcal{F}_{t,\tilde{d}}$, which also aims to indicate that in general the optimal solution $\tilde{\pi}_t^*$ has a higher objective than π_t^* due to the relaxed constraints.

following the subgradient flow of $J_{t,0}(\pi_\theta) + \mathbb{I}_{\mathcal{F}_{t,d}}(\pi_\theta)$ to reach θ_t^* is not applicable. The idea is to enlarge the feasible set $\mathcal{F}_{t,d}$ by increasing the violation threshold $\tilde{d}_{t,i} \geq d_{t,i} + \delta$, for any $\delta > 0$, such that with high probability, $\theta \in \mathcal{F}_{t,\tilde{d}}$. Then by following the subgradient flow for $J_{t,0}(\pi_\theta) + \mathbb{I}_{\tilde{\mathcal{F}}_t}(\pi_\theta)$, we can arrive at a critical point $\tilde{\theta}_t^*$ (corresponding to the policy $\tilde{\pi}_t^*$), which is most likely different from θ_t^*. It then remains to bound the distance between θ_t^* and $\tilde{\theta}_t^*$, which is possible due to the preservation of definability through inf projection. This is the roadmap we will follow. A graphical illustration of the approach is shown in Fig. 2.

Bounding the term $\|\theta_t^* - \tilde{\theta}_t^*\|$. In this part, we will bound the term $\|\theta_t^* - \tilde{\theta}_t^*\|$, which will be used to bound $\|\pi_t^* - \tilde{\pi}_t^*\|$. Firstly, we will prove that the parameter θ_t, which represents the solution map of an optimization with definable objective and constraints, is definable.

Proposition 2. *Let $\theta_t(d) \in \arg\min \{J_{t,0}(\pi_{t,\theta}),\ s.t.\ J_{t,i}(\pi_{t,\theta}) \leq d_{t,i}, \forall i = 1, ..., p\}$ be the solution map of the constraint parameters d. Then, the function $\theta_t(d)$ is continuous and definable. Furthermore, there exists a finite partition of the space such that the restriction of $\theta_t(d)$ to each partition is C^p smooth.*

Proof. First, it can be seen that the solution map $\theta_t(d) \in \arg\min J_{t,0}(\pi_{t,\theta}) + \mathbb{I}_{\mathcal{F}_{t,d}}(\pi_{t,\theta})$. Let $\phi_t(d) = \min J_{t,0}(\pi_{t,\theta}) + \mathbb{I}_{\mathcal{F}_{t,d}}(\pi_{t,\theta})$ be the optimal value function. Since $J_{t,0}(\pi_{t,\theta}) + \mathbb{I}_{\mathcal{F}_{t,d}}(\pi_{t,\theta})$ is definable by Proposition 8, and definability is preserved under inf projection, $\phi_t(d)$ is definable. Since $\theta_t(d) = \{\theta : J_{t,0}(\pi_{t,\theta}) + \mathbb{I}_{\mathcal{F}_{t,d}}(\pi_{t,\theta}) = \phi_t(d)\}$, by the Tarski-Seidenberg Theorem, $\theta_t(d)$ is definable. The continuity property follows directly from Berge's Maximum theorem.

Following the discussion of Whitney stratifications in [15], since the graph of a definable function is Whitney stratifiable, we can construct a partition by projecting the stratification into the function domain, which will be a Whitney C^p-stratification by the constant rank theorem. Furthermore, the restriction of the definable function to each stratum is C^p-smooth. Alternatively, we can directly use the fact that for any definable function, there exists a C^p-decomposition which has a finite number of cells, and the restriction to each cell is C^p-smooth [90]. This completes the proof. □

Now that we have proved that the function $\theta_t(d)$ is definable, we can obtain the bound $\|\theta_t^* - \tilde{\theta}_t^*\|$. Intuitively, our proof exploits the fact that continuous and definable functions exhibit controlled behaviors along any path, even if it crosses over finite number of Whitney strata.

Lemma 9. *For any $\tilde{d}$ such that $\mathcal{F}_{t,\tilde{d}}$ is non-empty, the following holds:*

$$\|\theta_t^* - \tilde{\theta}_t^*\| = \|\theta(d) - \theta(\tilde{d})\| = \mathcal{O}(\|d - \tilde{d}\|).$$

Proof. Since every smooth function over a bounded set is Lipschitz, let us denote L_d as the maximum of the Lipschitz constants for all the cells of the Whitney stratification of $\theta_t(d)$. Let $d(\lambda) = \lambda d + (1 - \lambda)\tilde{d}$, where $0 \leq \lambda \leq 1$, be the curve that connects between d and $\tilde{d}$. Also, let $0 = \lambda_1 \leq ... \leq \lambda_n = 1$

be the partition such that $\theta_t(d(\lambda))$ belongs to one cell for all $\lambda_i < \lambda < \lambda_{i+1}$ for $i = 1, ..., n-1$. We know that $n < \infty$ since $d(\theta_t)$ is Whitney stratifiable. Thus,

$$\begin{aligned}
\|\theta_t(d) - \theta_t(\tilde{d})\| &\leq \sum_{i=1}^{n-1} \|\theta_t(d(\lambda_i)) - \theta_t(d(\lambda_{i+1}))\| \\
&\leq L_d \sum_{i=1}^{n-1} \|d(\lambda_i) - d(\lambda_{i+1})\| \\
&\leq L_d \|d - \tilde{d}\| \sum_{i=1}^{n-1} |\lambda_{i+1} - \lambda_i| \\
&= L_d \|d - \tilde{d}\|
\end{aligned}$$

where the first inequality is due to triangle inequality, the second inequality is due to Lipschitz continuity, the third inequality is due to the definition of $d(\lambda)$, the first equality is due to the non-decreasing sequence of λ_i. □

Bounding the term $\|\tilde{\theta}_t^* - \hat{\theta}_t\|$. Recall that $\hat{\theta}_t$ is the parameter for $\hat{\pi}_t$ (output of within-task CRPO), and $\tilde{\theta}_t^*$ is the parameter for an optimal solution with an enlarged feasible set $\mathcal{F}_{t,\tilde{d}}$. In this subsection, we will obtain the upper bound for the term $\|\tilde{\theta}_t^* - \hat{\theta}_t\|$. Let $f(\theta, \tilde{d}) = J_{t,0}(\pi_\theta) + \mathbb{I}_{\mathcal{F}_{t,\tilde{d}}}(\pi_\theta)$, and choose $\tilde{d}_i = \mathcal{O}(1/\sqrt{M})$ for $i = 1, ..., p$, which coincides with the upper bound on constraint violation for within-task CRPO such that $\hat{\theta}_t \in \mathcal{F}_{t,\tilde{d}}$ with high probability. In the next result, we will condition on this high probability event.

Lemma 10. *With the choice of $\tilde{d} = d + \delta$, where $\delta = \mathcal{O}(1/\sqrt{M})$ coincides with the upper bound on constraint violation for within-task CRPO such that $\hat{\theta}_t \in \mathcal{F}_{t,\tilde{d}}$, the following holds:*

$$\|\tilde{\theta}_t^* - \hat{\theta}_t\| \leq \mathcal{O}\left(\psi\left(1/\sqrt{M}\right)\right),$$

where ψ is a strictly increasing continuous function with the property that $\psi(0) = 0$ as specified in Lemma 1.

Proof. Without loss of generality, consider $f(\theta) = J_{t,0}(\pi_\theta) + \mathbb{I}_{\mathcal{F}_{t,\tilde{d}}}(\pi_\theta) + c$, where $c := -\inf J_{t,0}(\pi_\theta) + \mathbb{I}_{\tilde{\mathcal{F}}_t}(\pi_\theta)$, so that the minimal value of f is translated to 0. For simplicity, also assume that $f(\hat{\theta}_t) \leq \mathcal{X}(\|\hat{\theta}_t\|)$. Note that this assumption can be relaxed by using the concept of "curves of maximal slope" at the cost of slightly more complicated analysis and bounds [52].

Now, consider a subgradient flow $\dot{\theta}(\tau) \in -\partial f(\theta(\tau))$ (see Definition 5), initialized at $\theta(0) = \hat{\theta}_t$ then, for any $0 \leq s' < s$, we have that

$$\begin{aligned}
\psi\big(f(\theta(s'))\big) - \psi\big(f(\theta(s))\big) &= \int_s^{s'} \frac{d}{d\tau} \psi\big(f(\theta(\tau))\big) d\tau \\
&= \int_{s'}^{s} \psi'\big(f(\theta(\tau))\big) \|\dot{\theta}(\tau)\|^2 d\tau \\
&\geq \int_{s'}^{s} \|\dot{\theta}(\tau)\| d\tau \\
&\geq \left\| \int_{s'}^{s} \dot{\theta}(\tau) d\tau \right\| \\
&= \|\theta(s) - \theta(s')\|
\end{aligned}$$

where the second equality is due to the property of the subgradient flow (see Sec. E.2), the first inequality is due to $\|\partial^0(\psi f)\big(\theta(\tau)\big)\| \geq 1$ from Proposition 1, and the second inequality is due to the triangle inequality. Thus, by taking $s' = 0$ and $s \to \infty$, we have shown that

$$\psi(f(\hat{\theta}_t)) \geq \|\hat{\theta}_t - \tilde{\theta}_t^*\|.$$

Therefore,

$$\begin{aligned}\|\hat{\theta}_t - \tilde{\theta}_t^*\| &\le \psi(f(\hat{\theta}_t) - f(\tilde{\theta}_t^*)) \\ &\overset{(i)}{\le} \psi(J_{t,0}(\hat{\pi}_t) - J_{t,0}(\tilde{\pi}_t^*)),\end{aligned}$$

where the first inequality is due to the optimality of $\tilde{\pi}_t^*$, and (i) follows since both $\hat{\pi}_t$ and $\tilde{\pi}_t^*$ are feasible for $\mathcal{F}_{t,\tilde{d}}$. Note that the suboptimality bound can be split as

$$\psi(J_{t,0}(\hat{\pi}_t) - J_{t,0}(\tilde{\pi}_t^*)) = \psi\bigg(J_{t,0}(\hat{\pi}_t) - J_{t,0}(\pi_t^*) + J_{t,0}(\pi_t^*) - J_{t,0}(\tilde{\pi}_t^*)\bigg).$$

By CRPO, we can bound the value difference $J_{t,0}(\hat{\pi}_t) - J_{t,0}(\pi_t^*) \le \mathcal{O}(1/\sqrt{M})$. Moreover, Since the value function is Lipschitz [94, Lemma 4], the value difference $J_{t,0}(\pi_t^*) - J_{t,0}(\tilde{\pi}_t^*)$ can be bounded by the distance $\|\theta_t^* - \tilde{\theta}_t^*\|$, which is bounded again by $\mathcal{O}(1/\sqrt{M})$ according to Lemma 9. Hence, recognizing that ψ is strictly increasing, we have proved the claim. □

Bounding the term $\|\theta_t^* - \hat{\theta}_t\|$. Finally, we are able to bound the term of our original interests.

Lemma 11. *Under assumption 2, the following holds:*

$$\|\theta_t^* - \hat{\theta}_t\| \le \mathcal{O}\left(\psi\left(\frac{1}{\sqrt{M}}\right) + \frac{1}{\sqrt{M}}\right),$$

where ψ is a strictly increasing continuous function with the property that $\psi(0) = 0$ as specified in Lemma 1.

Proof. The claim follows directly from Lemmas 9 and 10 and the triangle inequality. □

Remark 2. *Note that our strategy to bound $\|\theta_t^* - \hat{\theta}_t\|$ is algorithmic-agnostic as it only relies on the optimization landscape. The only place we rely on the algorithm is to bound the suboptimality gap, which is then converted to a bound on $\|\tilde{\theta}_t^* - \hat{\theta}_t\|$ in Lemma 10. Also, the enlargement of feasible set should be viewed as a proof technique and has no implications to the algorithm design. Indeed, the motivation for the enlargement is to properly design a subgradient flow system. Thus, the result of Lemma 11 is not conditioned on how the enlargement is performed. Also, note that definability is used differently in Lemmas 9 and 10. In the former case, we exploit the Whitney stratification property to provide an upper bound, while in the latter case, we exploit the KL property to obtain a lower bound, hence they serve different purposes.*

E.4 Bounding term *(A)*: $|\mathbb{E}_{\nu_t^*}[D(\pi_t^*|\pi)] - \mathbb{E}_{\tilde{\nu}_t}[D(\pi_t^*|\pi)]|$

The result from Lemma 11 can be used directly to provide bounds for *(A)* and *(C)*. We start with the term *(A)*.

Lemma 12. *The following bound holds:*

$$|\mathbb{E}_{s\sim\nu_t^*}[D(\pi_t^*|\pi)] - \mathbb{E}_{s\sim\tilde{\nu}_t}[D(\pi_t^*|\pi)]| = \mathcal{O}\left(\psi\left(\frac{1}{\sqrt{M}}\right) + \frac{1}{\sqrt{M}}\right) \tag{33}$$

where ψ is a strictly increasing continuous function with the property that $\psi(0) = 0$ as specified in Lemma 1.

Proof.

$$\begin{aligned}&|\mathbb{E}_{s\sim\nu_t^*}[D(\pi_t^*|\pi)] - \mathbb{E}_{s\sim\tilde{\nu}_t}[D(\pi_t^*|\pi)]| \\ &= \left|\sum_{s\in\mathcal{S}_t}\big(\nu_t^*(s) - \tilde{\nu}_t(s)\big)D(\pi_t^*(s)|\pi(s))\right| \\ &\le C_\pi\|\nu_t^* - \tilde{\nu}_t\|_1 \\ &\le 2C_\pi C_\nu\|\theta_t^* - \hat{\theta}_t\|_2\end{aligned}$$

where the first equality is by definition, the first inequality is due to Assumption 1, and the second inequality is due to [94, Lem. 3], which also specifies the constant C_ν, and [63, Prop. 4.2]. The result then follows by recalling the result from Lemma 11. □

E.5 Bounding term (C): $|\mathbb{E}_{\hat{\nu}_t}[D(\pi_t^*|\pi)] - \mathbb{E}_{\hat{\nu}_t}[D(\hat{\pi}_t|\pi)]|$

Similarly, we can prove the upper bound for the error term (C).

Lemma 13. *The following bound holds:*

$$|\mathbb{E}_{s\sim\hat{\nu}_t}[D(\pi_t^*|\pi)] - \mathbb{E}_{s\sim\hat{\nu}_t}[D(\hat{\pi}_t|\pi)]| = \mathcal{O}\left(\psi\left(\frac{1}{\sqrt{M}}\right) + \frac{1}{\sqrt{M}}\right) \tag{34}$$

Proof.

$$\begin{aligned}
&|\mathbb{E}_{s\sim\hat{\nu}_t}[D(\pi_t^*|\pi)] - \mathbb{E}_{s\sim\hat{\nu}_t}[D(\hat{\pi}_t|\pi)]| \\
&= \left|\sum_{s\in\mathcal{S}} \hat{\nu}_t(s)\Big(D(\pi_t^*|\pi) - D(\hat{\pi}_t|\pi)\Big)\right| \\
&\le \sum_{s\in\mathcal{S}} \hat{\nu}_t(s)\Big|D(\pi_t^*|\pi) - D(\hat{\pi}_t|\pi)\Big| \\
&\le L_g \sum_{s\in\mathcal{S}} \hat{\nu}_t(s)\|\pi_t^*(s) - \hat{\pi}_t(s)\|_2 \\
&\le L_g \sum_{s\in\mathcal{S}} \hat{\nu}_t(s)\|\theta_t^* - \hat{\theta}_t - c'1\|_\infty \\
&\le L_g \|\theta_t^* - \hat{\theta}_t\|_2
\end{aligned}$$

where the first inequality is due to the non-negativity of $\hat{\nu}_t(s)$ and triangle inequality, the second inequality is due to Assumption 1, the third inequality holds for any constant c' and is due to [72, Lem. 24], and the last inequality is due to $\sum_{s\in\mathcal{S}} \hat{\nu}_t(s) = 1$ and by choosing $c' = 0$. The result then follows by recalling the result from Lemma 11. □

E.6 Bounding term (B): $|\mathbb{E}_{\tilde{\nu}_t}[D(\pi_t^*|\pi)] - \mathbb{E}_{\hat{\nu}_t}[D(\pi_t^*|\pi)]|$

Now, we will upper bound the error term (B). The proof follows DualDICE [75]. We introduce the following notations. Let $\hat{\mathbb{E}}_{d^{\mathcal{D}_t}}$ denote an average of empirical samples where $\{s_i, a_i, r_i, s_i'\}_{i=1}^N \sim d^{\mathcal{D}_t}$, and ρ_t be the initial state distribution for the CMDP task t. The number of data points $N = \mathcal{O}(M^{1+1/\sigma})$, where σ is any positive number $\sigma \in (0, 1)$. Note that the additional factor of $\mathcal{O}(M^{1/\sigma})$ results from the critic evaluation per policy update (see [93, Thm. 1]). We will roughly bound $N = \mathcal{O}(M^2)$ in the following to simplify the presentation. The stationary distribution correction factor is denoted as $w_{\hat{\pi}_t/\mathcal{D}_t}(s,a) = \frac{\tilde{\nu}_t(s,a)}{d^{\mathcal{D}_t}(s,a)}$.

We make the following regularity assumption on the distribution $d^{\mathcal{D}_t}$ with respect to the target policy $\hat{\pi}_t$ [75, Asm. 1].

Assumption 5 (Reference distribution property)**.** *For any (s,a), $\tilde{\nu}_t(s,a) > 0$ implies that $d^{\mathcal{D}_t}(s,a) > 0$. Furthermore, the correction terms are bounded by some finite constant C_ω: $\|\omega_{\tilde{\nu}_t/\mathcal{D}_t}\|_\infty \le C_\omega$.*

For convenience, we recapitulate the key points from DualDICE, where we also omit the task dependence t (i.e., we use $d^{\mathcal{D}}$, π, and ρ in lieu of $d^{\mathcal{D}_t}$, $\hat{\pi}_t$, and ρ_t, respectively). The objective function is given by

$$\begin{aligned}
J(z,\zeta) = \; & \mathbb{E}_{(s,a,s')\sim d^{\mathcal{D}}, a'\sim\pi(s')}\left[(z(s,a) - \gamma z(s',a'))\zeta(s,a) - \zeta(s,a)^2/2\right] \qquad (35)\\
& - (1-\gamma)\,\mathbb{E}_{s_0\sim\beta, a_0\sim\pi(s_0)}\left[z(s_0,a_0)\right]. \qquad (36)
\end{aligned}$$

The objective in the form prior to introduction of ζ is denoted as $J(z)$:

$$J(z) = \frac{1}{2}\mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}\left[(z - \mathcal{B}^\pi z)(s,a)^2\right] - (1-\gamma)\,\mathbb{E}_{s_0\sim\beta, a_0\sim\pi(s_0)}\left[z(s_0,a_0)\right]. \tag{37}$$

Let $\hat{J}(z,\zeta)$ denotes the empirical surrogate of $J(z,\zeta)$ with optimal solution as $(\hat{z}^*, \hat{\zeta}^*)$. We denote $z_{\mathcal{F}}^* = \arg\min_{z\in\mathcal{F}} J(z)$ and $z^* = \arg\min_{z:S\times A\to\,\mathbb{R}} J(z)$. We denote $L(z) = \max_{\zeta\in\mathcal{H}} J(z,\zeta)$ and $\hat{L}(z) = \max_{\zeta\in\mathcal{H}} \hat{J}(z,\zeta)$ as the primal objectives, and $\ell(\zeta) = \min_{z\in\mathcal{F}} J(z,\zeta)$, $\hat{\ell}(\zeta) =$

$\min_{z\in\mathcal{F}} \hat{J}(z,\zeta)$ as the dual objectives. We apply some optimization algorithm OPT for optimizing $\hat{J}(z,\zeta)$ with samples $\{s_i, a_i, r_i, s_i'\}_{i=1}^N \sim d^{\mathcal{D}}$, $\{s_0^i\}_{i=1}^N \sim \beta$, and target actions $a_i' \sim \pi(s_i'), a_0^i \sim \pi(s_0^i)$ for $i = 1, \ldots, N$. The outputs of OPT is denoted by $(\hat{z}, \hat{\zeta})$. We also make the following definitions to capture the error of approximation with $\mathcal{F}$ for z and $\mathcal{H}$ for ζ in optimizing $\hat{J}(z,\zeta)$:

$$\epsilon_{approx}(\mathcal{F}) \coloneqq \sup_{z\in S\times A\to\ \mathbb{R}} \inf_{z_{\mathcal{F}}\in\mathcal{F}} \left(\|z_{\mathcal{F}} - z\|_{\mathcal{D},1} + \|z_{\mathcal{F}} - z\|_{\rho\pi,1}\right) \tag{38}$$

$$\epsilon_{approx}(\mathcal{H}) \coloneqq \sup_{\zeta\in S\times A\to\ \mathbb{R}} \inf_{\zeta_{\mathcal{H}}\in\mathcal{H}} \left(\|\zeta_{\mathcal{H}} - \zeta\|_{\mathcal{D},1} + \|\zeta_{\mathcal{H}} - \zeta\|_{\rho\pi,1}\right) \tag{39}$$

$$\epsilon_{approx}(\mathcal{F},\mathcal{H}) \coloneqq \epsilon_{approx}(\mathcal{F}) + \epsilon_{approx}(\mathcal{H}) \tag{40}$$

We also define

$$\epsilon_{opt} \coloneqq \|\hat{\zeta} - \hat{\zeta}^*\|_{\mathcal{D}_t}^2 + \left\|\left(\hat{z}^* - \hat{\mathcal{B}}^{\pi}\hat{z}^*\ \right) - \left(\hat{z} - \hat{\mathcal{B}}^{\pi}\hat{z}\right)\right\|_{\mathcal{D}_t}^2 \tag{41}$$

as the optimization error of OPT from DualDICE.

Lemma 14. *By estimating $\hat{\nu}_t$ with DualDICE, the following bound holds:*

$$|\mathbb{E}_{\tilde{\nu}_t}[D(\pi_t^*|\pi)] - \mathbb{E}_{\hat{\nu}_t}[D(\pi_t^*|\pi)]| = \mathcal{O}\left(\sqrt{\frac{1}{M} + \epsilon_{opt} + \epsilon_{approx}(\mathcal{F},\mathcal{H})}\right),$$

Proof. We begin with the following decomposition:

$$\begin{aligned}
&\left(\mathbb{E}_{\tilde{\nu}_t}[D(\pi_t^*|\pi)] - \mathbb{E}_{\hat{\nu}_t}[D(\pi_t^*|\pi)]\right)^2 \leq \\
&\quad \underbrace{2\left(\hat{\mathbb{E}}_{d^{\mathcal{D}_t}} \sum_a \left((\hat{z} - \hat{\mathcal{B}}^{\hat{\pi}_t}\hat{z})(s,a) - (\hat{z}^* - \hat{\mathcal{B}}^{\hat{\pi}_t}\hat{z}^*)(s,a)\right) D(\pi_t^*|\pi)\right)^2}_{\epsilon_1} \\
&\quad + \underbrace{2\left(\hat{\mathbb{E}}_{d^{\mathcal{D}_t}} \sum_a (\hat{z}^* - \hat{\mathcal{B}}^{\hat{\pi}_t}\hat{z}^*)(s,a) D(\pi_t^*|\pi) - \mathbb{E}_{d^{\mathcal{D}_t}} \sum_a \omega_{\hat{\pi}_t/\mathcal{D}_t}(s,a) D(\pi_t^*|\pi))\right)^2}_{\epsilon_2}.
\end{aligned}$$

We will bound each term above separately.

$$\begin{aligned}
\epsilon_1 &\leq 2C_\pi^2 \left(\hat{\mathbb{E}}_{d^{\mathcal{D}_t}} \sum_a \left(\hat{z} - \hat{\mathcal{B}}^{\hat{\pi}_t}\hat{z}\right)(s,a) - \left(\hat{z}^* - \hat{\mathcal{B}}^{\hat{\pi}_t}\hat{z}^*\right)(s,a)\right)^2 \\
&\leq 2C_\pi^2 \left(\underbrace{\|\hat{\zeta} - \hat{\zeta}^*\|_{\mathcal{D}_t}^2 + \left\|\left(\hat{z}^* - \hat{\mathcal{B}}^{\pi}\hat{z}^*\ \right) - \left(\hat{z} - \hat{\mathcal{B}}^{\pi}\hat{z}\right)\right\|_{\mathcal{D}_t}^2}_{\epsilon_{opt}}\right),
\end{aligned}$$

where the error ϵ_{opt} is induced by the optimization OPT. The error term ϵ_2 can be decomposed as

$$\begin{aligned}
\epsilon_2 \leq& \underbrace{2\,C_\pi^2 \left(\hat{\mathbb{E}}_{d^{\mathcal{D}_t}} \sum_a \left(\hat{z}^* - \hat{\mathcal{B}}^{\hat{\pi}_t}\hat{z}^*\right)(s,a) - \mathbb{E}_{d^{\mathcal{D}_t}} \sum_a \left(\hat{z}^* - \mathcal{B}^{\hat{\pi}_t}\hat{z}^*\right)(s,a)\right)^2}_{\epsilon_{stat}} \\
&+ 2C_\pi^2 \left(\mathbb{E}_{d^{\mathcal{D}_t}} \sum_a \left(\left(\hat{z}^* - \mathcal{B}^{\hat{\pi}_t}\hat{z}^*\right)(s,a) - \omega_{\hat{\pi}_t/\mathcal{D}_t}(s,a)\right)\right)^2 \qquad (42) \\
=& 2\epsilon_{stat} + 2C_\pi^2 \left(\mathbb{E}_{d^{\mathcal{D}_t}} \sum_a \left(\left(\hat{z}^* - \mathcal{B}^{\hat{\pi}_t}\hat{z}^*\right)(s,a) - \left(z^* - \mathcal{B}^{\hat{\pi}_t}z^*\right)(s,a)\right)\right)^2,
\end{aligned}$$

where the equality is due to the result that $z^* - \mathcal{B}^{\hat{\pi}_t}z^*(s,a) = \omega_{\hat{\pi}_t/\mathcal{D}_t}(s,a)$ (see [75, Eq. 17]) and ϵ_{stat} is the error due to the finite number error. By [75, Lem. 7], $\epsilon_{stat} = \mathcal{O}\left(\frac{\log M + \log\frac{1}{\delta}}{M^2}\right)$ with probability at least $1-\delta$, where we use the bound on the number of data as $\mathcal{O}(M^2)$. To bound the second term, use the fact that $J(z)$ as defined in (37) is 1-strongly convex. Hence,

$$\begin{aligned}&\Big(\mathbb{E}_{d^{\mathcal{D}_t}}\sum_a\big(\big(\hat{z}^*-\mathcal{B}^{\hat{\pi}_t}\hat{z}^*\big)(s,a)-\big(z^*-\mathcal{B}^{\hat{\pi}_t}z^*\big)(s,a)\big)\Big)^2\\&\leq\|\big(\hat{z}^*-\mathcal{B}^{\hat{\pi}_t}\hat{z}^*\big)-\big(z^*-\mathcal{B}^{\hat{\pi}_t}z^*\big)\|^2_{\mathcal{D}_t}\\&\leq 2\big(J(\hat{z}^*)-J(z^*)\big)\end{aligned}$$

where (i) follows from [75, Section D.1] $\epsilon_{approx}(\mathcal{F})$ is the error due to the approximation with $\mathcal{F}$ for z, $\epsilon_{approx}(\mathcal{H})$ is the error due to the approximation with $\mathcal{H}$ for ζ, and ϵ_{est} is the estimation error, and L is the Lipschitz constant for f.

To bound the error between $J(\hat{z}^*)$ and $J(z^*)$, we use the decomposition suggested in [75]:

$$J(\hat{z}^*)-J(z^*)=\underbrace{J(\hat{z}^*)-L(\hat{z}^*)}_{(i)}+\underbrace{L(\hat{z}^*)-L(z^*_{\mathcal{F}})}_{(ii)}+\underbrace{L(z^*_{\mathcal{F}})-J(z^*_{\mathcal{F}})}_{(iii)}+\underbrace{J(z^*_{\mathcal{F}})-J(z^*)}_{(iv)},\quad(43)$$

where $(i)\leq\frac{2C_\omega}{1-\gamma}\|\zeta^*_{\mathcal{H}}-\zeta^*\|_{\mathcal{D}_t,1}\leq\frac{2C_\omega}{1-\gamma}\epsilon_{approx}(\mathcal{H})$, $(ii)=\mathcal{O}\left(\frac{\sqrt{\log M+\log\frac{1}{\delta}}}{M}\right)$ by [75, Lem. 6] (by also plugging in $N=\mathcal{O}(M^2)$), $(iii)\leq 0$ by definition, and $(iv)=\mathcal{O}(\epsilon_{approx}(\mathcal{F}))$. Note that we refer the reader to [75, Sec. D.1] for the above bounds. Therefore, we can bound $J(\hat{z}^*)-J(z^*)$ on the order of $\mathcal{O}\left(\epsilon_{approx}(\mathcal{H})+\epsilon_{approx}(\mathcal{F})+\frac{\sqrt{\log M+\log\frac{1}{\delta}}}{M}\right)$.

Combining the above relations, while noting that $\frac{1}{\sqrt{M}}$ decreases slower than $\frac{1}{M}$ in terms of M and is thus kept as the upper bound, we have shown the result. □

E.7 Putting it together: bounding the KL divergence estimation error

Theorem E.1 (KL divergence estimation error bound)**.** *The following bound holds:*

$$\begin{aligned}&|\mathbb{E}_{\nu^*_t}[D(\pi^*_t|\pi)]-\mathbb{E}_{\hat{\nu}_t}[D(\hat{\pi}_t|\pi)]|\\&=\mathcal{O}\Big(\psi\left(\frac{1}{\sqrt{M}}\right)+\frac{1}{\sqrt{M}}+\sqrt{\epsilon_{opt}}+\sqrt{\epsilon_{approx}(\mathcal{F},\mathcal{H})}\Big),\end{aligned}$$

where ψ is a strictly increasing continuous function with the property that $\psi(0)=0$ as specified in Lemma 1, $\epsilon_{approx}(\mathcal{F},\mathcal{H})$ is defined in (40)*, and ϵ_{opt} is defined in* (41)*.*

Proof. The result follows by combining the upper bounds for the error terms (A), (B) and (C), as specified by Lemmas 12, 14, and 13. We also apply the elementary inequality $\sqrt{a+b+c}\leq\sqrt{a}+\sqrt{b}+\sqrt{c}$ to further simplify the bound. □

Remark 3. *The bound above depends on the number of iterations M per task on different ways. By increasing M, we can expect to reduce the suboptimality gap, which can help reduce the distance between $\hat{\pi}_t$ to the optimal set of policies. Also, increasing M results in a larger dataset used to estimate its stationary distribution offline by DualDICE, which reduces the estimation error. The bound indicates that the only terms that do not vanish as we increase the number of iterations per task are those due to the inherent optimization error ϵ_{opt} and function approximation error $\epsilon_{approx}(\mathcal{F},\mathcal{H})$. In the case those terms are negligible (which are possible in view of recent breakthrough in over-parametrized deep learning [99, 76, 64, 105, 5], see also [42] for a survey), then the KL divergence estimation can be driven to arbitrary accuracy.*

F Proofs for Section 3.2

F.1 TAOG and TACV bounds for CRPO with adaptive learning rates

This section presents the task-averaged regret upper bounds for the CRPO when the adaptive learning rates are used for each objective (i.e., reward and constraints). We denote the learning rate for the reward as $\alpha_{t,0}$ and for the constraints as $\alpha_{t,i}$ for $i=1,...,p$. We also recall that $d_{t,i}$ is the constraint upper bound for $i = 1, ..., p$ and η_t is the tolerance for constraint violation (i.e., increasing the upper bound to $d_{t,i} + \eta_t$). For a single run of CRPO in task t, we denote $\mathcal{N}_{t,0}$ as the set of time steps the reward is maximized and $\mathcal{N}_{t,i}$ as the set of time steps constraint i is minimized. The Q-function in the CRPO algorithm is learned through TD learning with the total number of iterations denoted by K_{in}. The Q-function of objective i for policy $\pi_{t,m}$ at time step m is denoted by $Q^i_{t,m}$, and the estimated Q-function is denoted by $\bar{Q}^i_{t,m}$. The maximum value for both rewards and constraints is assumed to be c_{max}.

With all notations for CRPO in place, we present the following result, which extends [93] to the case of objective-specific learning rates.

Lemma 15. *For the CRPO algorithm in the tabular settings with learning rates $\{\alpha_{t,i}\}_{i=0,1,...,p}$, the following bound holds:*

$$
\begin{aligned}
&\alpha_{t,0} \sum_{m \in \mathcal{N}_{t,0}} \Big(J_{t,0}(\pi_t^*) - J_{t,0}(\pi_{t,m}) \Big) + \eta_t \sum_{i=1}^{p} \alpha_{t,i} |\mathcal{N}_{t,i}| \\
&\leq \mathbb{E}_{s \sim \nu_t^*}[D(\pi_t^* || \pi_{t,0})] + \frac{2c_{max}^2 |\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3} \sum_{i=0}^{p} \alpha_{t,i}^2 |\mathcal{N}_{t,i}| \\
&+ \sum_{i=0}^{p} \sum_{m \in \mathcal{N}_{t,i}} \frac{\alpha_{t,i}(3 + (1-\gamma)^2 + 3\alpha_{t,i} c_{max})}{(1-\gamma)^2} \|Q^i_{t,m} - \bar{Q}^i_{t,m}\|_2
\end{aligned}
$$

Proof. If $m \in \mathcal{N}_{t,0}$, by [93, Lemma 7], we have that:

$$
\begin{aligned}
\alpha_{t,0}\big(J_{t,0}(\pi_t^*) - J_{t,0}(\pi_{t,m})\big) \leq \mathbb{E}_{\nu_t^*}[D(\pi_t^* | \pi_{t,m}) - D(\pi_t^* | \pi_{t,m+1})] + \frac{2c_{max}^2 |\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3} \alpha_{t,0}^2 \\
+ \frac{3\alpha_{t,0}(1 + \alpha_{t,0} c_{max})}{(1-\gamma)^2} \|Q^0_{t,m} - \bar{Q}^0_{t,m}\|_2. \qquad (44)
\end{aligned}
$$

Similarly, if $m \in \mathcal{N}_{t,i}$, we can write

$$
\begin{aligned}
\alpha_{t,i}\big(J_{t,i}(\pi_{t,m}) - J_{t,i}(\pi_t^*)\big) \leq \mathbb{E}_{\nu_t^*}[D(\pi_t^* | \pi_{t,m}) - D(\pi_t^* | \pi_{t,m+1})] + \frac{2c_{max}^2 |\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3} \alpha_{t,i}^2 \\
+ \frac{3\alpha_{t,i}(1 + \alpha_{t,i} c_{max})}{(1-\gamma)^2} \|Q^i_{t,m} - \bar{Q}^i_{t,m}\|_2. \qquad (45)
\end{aligned}
$$

Taking the summation of (44) and (45) from $m = 0$ to $M - 1$, we get

$$
\begin{aligned}
&\alpha_{t,0} \sum_{m \in \mathcal{N}_{t,0}} \big(J_{t,0}(\pi_t^*) - J_{t,0}(\pi_{t,m}) \big) + \sum_{i=1}^{p} \sum_{m \in \mathcal{N}_{t,i}} \alpha_{t,i}\big(J_{t,i}(\pi_{t,m}) - J_{t,i}(\pi_t^*)\big) \\
&\leq \mathbb{E}_{\nu_t^*}[D(\pi_t^* | \pi_{t,0})] + \frac{2c_{max}^2 |\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3} \sum_{i=0}^{p} \alpha_{t,i}^2 |\mathcal{N}_{t,i}| \\
&+ \sum_{i=0}^{p} \sum_{m \in \mathcal{N}_{t,i}} \frac{3\alpha_{t,i}(1 + \alpha_{t,i} c_{max})}{(1-\gamma)^2} \|Q^i_{t,m} - \bar{Q}^i_{t,m}\|_2
\end{aligned}
$$

Since $J_{t,i}(\pi_{t,m}) - J_{t,i}(\pi_t^*) \geq \eta_t - \|Q^i_{t,m} - \bar{Q}^i_{t,m}\|$ [93, Eq. 15], by rearranging the terms above we obtain the result. □

Next, we study the condition on the maximum constraint violation threshold η_t and how it affects $\mathcal{N}_{t,0}$ and the upper bounds for TAOG and TACV. We make the following assumption to proceed.

Assumption 6. *Assume that $\sum_{m \in \mathcal{N}_{t,0}} J_{t,0}(\pi_t^*) - J_{t,0}(\pi_{t,m}) \geq c_J$ for some $c_J \in (-\frac{1}{2}\alpha_{t,0}\eta_t M, 0]$.*

The assumption above indicates that the policies in $\mathcal{N}_{t,0}$ do not have rewards higher than the optimal policy by more than $\frac{1}{2}\alpha_{t,0}\eta_t$ on average. Note that it is indeed possible to have rewards higher than the optimal policy if the corresponding policy does not satisfy some safety constraints (i.e., infeasible policy). However, it is not a strong assumption since we are comparing with the optimal policy. The above assumption is not present in [93], which invalidates one of its derivation steps (in particular, [93, Thm. 3]), and is thus introduced to rectify the proof.

Lemma 16. *For the CRPO algorithm, choose $\alpha_{t,i} \geq \alpha_{t,0}$. Suppose that the following condition holds:*

$$\frac{1}{2}\eta_t M \alpha_{t,0} \geq \mathbb{E}_{\nu_t^*}[D(\pi_t^*|\pi_{t,0})] + \frac{2c_{max}^2|\mathcal{S}||\mathcal{A}|M}{(1-\gamma)^3}\sum_{i=0}^{p}\alpha_{t,i}^2 + \sum_{i=0}^{p}\sum_{m\in\mathcal{N}_{t,i}}\frac{\alpha_{t,i}(3+(1-\gamma)^2+3\alpha_{t,i}c_{max})}{(1-\gamma)^2}\|Q_{t,m}^i - \bar{Q}_{t,m}^i\|_2. \tag{46}$$

Then, we have that $\mathcal{N}_{t,0} \neq \emptyset$, i.e., $\hat{\pi}_t$ is well-defined; also, one the following two statements must hold,

1. $|\mathcal{N}_{t,0}| \geq M/2$,
2. $\sum_{m\in\mathcal{N}_{t,0}}\Big(J_{t,0}(\pi_t^*) - J_{t,0}(\pi_{t,m})\Big) \leq 0$.

Under assumption 6, we also have the following holds:

$$|\mathcal{N}_{t,0}| \geq \left(\frac{1}{2} - \kappa\right)M$$

for some $\kappa \in (0, \frac{1}{2})$.

Proof. The proof for $\mathcal{N}_{t,0} \neq \emptyset$ follows directly from [93, Lem. 9]. For the second statement, we consider the case that $\sum_{m\in\mathcal{N}_{t,0}}\Big(J_{t,0}(\pi_t^*) - J_{t,0}(\pi_{t,m})\Big) > 0$. From Lemma 15, it implies that

$$\eta_t\sum_{i=1}^{p}\alpha_{t,i}|\mathcal{N}_{t,i}| \leq \mathbb{E}_{s\sim\nu^*}[D(\pi_t^*|\pi_{t,0})] + \frac{2c_{max}^2|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3}\sum_{i=0}^{p}\alpha_{t,i}^2|\mathcal{N}_{t,i}| + \sum_{i=0}^{p}\sum_{m\in\mathcal{N}_{t,i}}\frac{\alpha_{t,i}(3+(1-\gamma)^2+3\alpha_{t,i}c_{max})}{(1-\gamma)^2}\|Q_{t,m}^i - \bar{Q}_{t,m}^i\|_2.$$

Suppose that $|\mathcal{N}_{t,0}| < M/2$, then $\sum_{i=1}^{p}|\mathcal{N}_{t,i}| > M/2$. Since $\alpha_{t,i} \geq \alpha_{t,0}$, we have that

$$\frac{1}{2}\alpha_{t,0}\eta_t M < \mathbb{E}_{s\sim\nu^*}[D(\pi_t^*|\pi_{t,0})] + \frac{2c_{max}^2|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3}\sum_{i=0}^{p}\alpha_{t,i}^2|\mathcal{N}_{t,i}| + \sum_{i=0}^{p}\sum_{m\in\mathcal{N}_{t,i}}\frac{\alpha_{t,i}(3+(1-\gamma)^2+3\alpha_{t,i}c_{max})}{(1-\gamma)^2}\|Q_{t,m}^i - \bar{Q}_{t,m}^i\|_2,$$

which contradicts (46). Hence, we must have $|\mathcal{N}_{t,0}| \geq M/2$.

Next, we show that $|\mathcal{N}_{t,0}| \geq \left(\frac{1}{2} - \kappa\right)M$ for some $\kappa \in (0, \frac{1}{2})$. Under assumption 6 and by Lemma 15, we have that

$$\eta_t\sum_{i=1}^{p}\alpha_{t,i}|\mathcal{N}_{t,i}| \leq \mathbb{E}_{s\sim\nu^*}[D(\pi_t^*|\pi_{t,0})] + \frac{2c_{max}^2|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3}\sum_{i=0}^{p}\alpha_{t,i}^2|\mathcal{N}_{t,i}| + \sum_{i=0}^{p}\sum_{m\in\mathcal{N}_{t,i}}\frac{\alpha_{t,i}(3+(1-\gamma)^2+3\alpha_{t,i}c_{max})}{(1-\gamma)^2}\|Q_{t,m}^i - \bar{Q}_{t,m}^i\|_2 - \alpha_{t,0}c_J.$$

Choose $\kappa := \frac{-c_J}{\alpha_{t,0}\eta_t M}$. Since $-c_J < \frac{1}{2}\alpha_{t,0}\eta_t M$ by assumption, we have that $\kappa \in (0, \frac{1}{2})$. Consider the case that $|\mathcal{N}_{t,0}| < \big(\frac{1}{2} - \kappa\big)M$, which implies that $\sum_{i=1}^{p} |\mathcal{N}_{t,i}| > \big(\frac{1}{2} + \kappa\big)M$. This implies that

$$\Big(\frac{1}{2} + \kappa\Big)\alpha_{t,0}\eta_t M \leq \mathbb{E}_{s\sim\nu^*}[D(\pi_t^*|\pi_{t,0})] + \frac{2c_{max}^2|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3}\sum_{i=0}^{p}\alpha_{t,i}^2|\mathcal{N}_{t,i}|$$
$$+\sum_{i=0}^{p}\sum_{m\in\mathcal{N}_{t,i}}\frac{\alpha_{t,i}(3+(1-\gamma)^2+3\alpha_{t,i}c_{max})}{(1-\gamma)^2}\|Q_{t,m}^i - \bar{Q}_{t,m}^i\|_2,$$

which again contradicts (46). Hence, we must have $|\mathcal{N}_{t,0}| \geq \big(\frac{1}{2} - \kappa\big)M$. □

Now, we prove the upper bound of suboptimality and constraint violation per task.

Lemma 17. *Let the violation tolerance be chosen as:*

$$\eta_t = \frac{2\mathbb{E}_{s\sim\nu^*}[D(\pi_t^*||\pi_{t,0})]}{M\alpha_{t,0}} + \frac{4c_{max}^2|\mathcal{S}||\mathcal{A}|}{\alpha_{t,0}(1-\gamma)^3}\sum_{i=0}^{p}\alpha_{t,i}^2 + \sum_{i=0}^{p}\frac{2\alpha_{t,i}(3+(1-\gamma)^2+3\alpha_{t,i}c_{max})}{\sqrt{M}\alpha_{t,0}(1-\gamma)^2},$$

Then, the following holds

$$U_t(\pi_{t,0}, \{\alpha_{t,i}\}_{i=0}^p) = \frac{c_1^t}{\alpha_{t,0}M}\mathbb{E}_{s\sim\nu_t^*}[D(\pi_t^*|\pi_{t,0})] + c_2^t\sum_{i=0}^{p}\frac{\alpha_{t,i}^2}{\alpha_{t,0}} + \sum_{i=0}^{p}\frac{c_3^t\alpha_{t,i} + c_4^t\alpha_{t,i}^2}{\alpha_{t,0}\sqrt{M}} \tag{47}$$

$$U_{t,i}(\pi_{t,0}, \{\alpha_{t,i}\}_{i=0}^p) = U_t(\pi_{t,0}, \{\alpha_{t,i}\}_{i=0}^p) + \frac{c_5^t}{\sqrt{M}}, \tag{48}$$

where $c_1^t = 2$, $c_2^t = \frac{4c_{max}^2|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3 M}$, $c_3^t = \frac{3+(1-\gamma)^2}{(1-\gamma)^2}$, $c_4^t = \frac{3c_{max}}{(1-\gamma)^2}$, *and* $c_5^t = \frac{2\sqrt{(1-\gamma)|\mathcal{S}||\mathcal{A}|}}{1-2\kappa}$.

Proof. From Lemma 15, we have that

$$\alpha_{t,0}\sum_{m\in\mathcal{N}_{t,0}}\Big(J_{t,0}(\pi_t^*) - J_{t,0}(\pi_{t,m})\Big) + \eta_t\sum_{i=1}^{p}\alpha_{t,i}|\mathcal{N}_{t,i}|$$
$$\leq \mathbb{E}_{s\sim\nu_t^*}[D(\pi_t^*|\pi_{t,0})] + \frac{2c_{max}^2|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3}\sum_{i=0}^{p}\alpha_{t,i}^2|\mathcal{N}_{t,i}|$$
$$+\sum_{i=0}^{p}\sum_{m\in\mathcal{N}_{t,i}}\frac{\alpha_{t,i}(3+(1-\gamma)^2+3\alpha_{t,i}c_{max})}{(1-\gamma)^2}\|Q_{t,\pi_m}^i - \bar{Q}_{t,\omega_m}^i\|_2$$

If $\sum_{m\in\mathcal{N}_{t,0}}\Big(J_{t,0}(\pi_t^*) - J_{t,0}(\pi_{t,m})\Big) \leq 0$, then $J_{t,0}(\pi_t^*) - \mathbb{E}[J_{t,0}(\hat{\pi}_t)] \leq 0$. If $\sum_{m\in\mathcal{N}_{t,0}}\Big(J_{t,0}(\pi_t^*) - J_{t,0}(\pi_{t,m})\Big) > 0$, then, by Lemma 16, we have $|\mathcal{N}_{t,0}| \geq M/2$. Hence,

$$J_{t,0}(\pi_t^*) - \mathbb{E}[J_{t,0}(\hat{\pi}_t)] = \frac{1}{|\mathcal{N}_{t,0}|}\sum_{m\in\mathcal{N}_{t,0}}[J_{t,0}(\pi_t^*) - J_{t,0}(\pi_{t,m})]$$
$$\leq \frac{2}{\alpha_{t,0}M}\mathbb{E}_{\nu_t^*}[D(\pi_t^*|\pi_{t,0})] + c_2^t\sum_{i=0}^{p}\frac{\alpha_{t,i}^2}{\alpha_{t,0}} + \sum_{i=0}^{p}\sum_{m\in\mathcal{N}_{t,i}}\frac{(c_3^t\alpha_{t,i} + c_4^t\alpha_{t,i}^2)}{M\alpha_{t,0}}\|Q_{t,\pi_m}^i - \bar{Q}_{t,\omega_m}^i\|_2,$$
$$\leq \frac{2}{\alpha_{t,0}M}\mathbb{E}_{\nu_t^*}[D(\pi_t^*|\pi_{t,0})] + c_2^t\sum_{i=0}^{p}\frac{\alpha_{t,i}^2}{\alpha_{t,0}} + \sum_{i=0}^{p}\sum_{m\in\mathcal{N}_{t,i}}\frac{(c_3^t\alpha_{t,i} + c_4^t\alpha_{t,i}^2)}{\sqrt{M}\alpha_{t,0}}$$

where the last inequality is due to the choice of $K_{in} = \Theta(K^{1/\sigma}\log^{2/\sigma}(|\mathcal{S}|^2|\mathcal{S}|^2K^{1+2/\sigma}/\delta))$ as specified by [93, Lem. 8] for critic evaluations. Here, the constants are chosen as $c_1^t = 2$, $c_2^t = \frac{4c_{max}^2|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3 M}$, $c_3^t = \frac{3+(1-\gamma)^2}{(1-\gamma)^2}$, $c_4^t = \frac{3c_{max}}{(1-\gamma)^2}$.

For constraint violation, consider any $i = 1, ..., p$, we have

$$
\begin{aligned}
\mathbb{E}[J_{t,i}(\hat{\pi}_t)] - d_{t,i} &= \frac{1}{|\mathcal{N}_{t,0}|} \sum_{m \in \mathcal{N}_{t,0}} J_{t,i}(\pi_{t,m}) - d_{t,i} \\
&\leq \frac{1}{|\mathcal{N}_{t,0}|} \sum_{m \in \mathcal{N}_{t,0}} \left(\bar{J}_{t,i}(\pi_{t,m}) - d_{t,i}\right) + \frac{1}{|\mathcal{N}_{t,0}|} \sum_{m \in \mathcal{N}_{t,0}} |\bar{J}_{t,i}(\pi_{t,m}) - J_{t,i}(\pi_{t,m})| \\
&\leq \eta_t + \frac{1}{|\mathcal{N}_{t,0}|} \sum_{i=0}^{p} \sum_{m \in \mathcal{N}_{t,i}} \|Q^i_{t,\pi_m} - \bar{Q}^i_{t,\pi_m}\|_2 \\
&\leq \eta_t + \frac{2}{(1-2\kappa)M} \sum_{i=0}^{p} \sum_{m \in \mathcal{N}_{t,i}} \|Q^i_{t,\pi_m} - \bar{Q}^i_{t,\pi_m}\|_2
\end{aligned}
$$

where the first inequality is due to triangle inequality, the second inequality is by the design of the CRPO algorithm, and the third inequality is due to Lemma 16, where $\kappa \in (0, \frac{1}{2})$. By the choice of K_{in}, we have that $\sum_{i=0}^{p} \sum_{m \in \mathcal{N}_{t,i}} \|Q^i_{t,\pi_m} - \bar{Q}^i_{t,\pi_m}\|_2 \leq \sqrt{(1-\gamma)|\mathcal{S}||\mathcal{A}|M}$. Plugging the value of η_t obtains the result. □

Finally, we are able to provide the following bounds on TAOG and TACV in the case of adaptive learning rates.

Theorem F.1 (Bounds on TAOG and TACV)**.** *Suppose we run CRPO algorithm for M steps per task t with learning rates $\{\alpha_{t,i}\}_{i=0,...,p}$. Then, after T tasks, the TAOG $\bar{R}_0$ is given by*

$$
\bar{R}_0 = \frac{1}{T} \sum_{t=1}^{T} \left[\frac{c_1^t}{\alpha_{t,0} M} \mathbb{E}_{s \sim \nu_t^*}[D(\pi_t^* | \pi_{t,0})] + c_2^t \sum_{i=0}^{p} \frac{\alpha_{t,i}^2}{\alpha_{t,0}} + \sum_{i=0}^{p} \frac{c_3^t \alpha_{t,i} + c_4^t \alpha_{t,i}^2}{\alpha_{t,0}\sqrt{M}} \right], \tag{49}
$$

and the TACV $\bar{R}_i$ is given by

$$
\bar{R}_i = \frac{1}{T} \sum_{t=1}^{T} \left[\frac{c_1^t}{\alpha_{t,0} M} \mathbb{E}_{s \sim \nu_t^*}[D(\pi_t^* | \pi_{t,0})] + c_2^t \sum_{i=0}^{p} \frac{\alpha_{t,i}^2}{\alpha_{t,0}} + \sum_{i=0}^{p} \frac{c_3^t \alpha_{t,i} + c_4^t \alpha_{t,i}^2}{\alpha_{t,0}\sqrt{M}} + \frac{c_5^t}{\sqrt{M}} \right], \tag{50}
$$

for $i = 1, ..., p$, where $c_1^t = 2$, $c_2^t = \frac{4c_{max}^2 |\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3 M}$, $c_3^t = \frac{3+(1-\gamma)^2}{(1-\gamma)^2}$, $c_4^t = \frac{3c_{max}}{(1-\gamma)^2}$, and $c_5^t = \frac{2\sqrt{(1-\gamma)|\mathcal{S}||\mathcal{A}|}}{1-2\kappa}$.

Proof. The proof follows directly by summing the results (47) and (48) over $t = 1, ..., T$. □

F.2 Adapting to the dynamic regret and intra-task geometry

We restate the theorem below for convenience.

Theorem F.2. *Let each within-task CMDP t run M steps of CRPO, initialized by policy $\pi_{t,0} :=$ INIT(t) and learning rates $\{\alpha_{t,i}\}_{i=0}^{p} :=$ SIM(t). Let $\{\kappa_i^*\}_{i=0}^{p} := \arg\min L(\{\kappa_i\}_{i=0}^{p})$, where*

$$
L(\{\kappa_i\}_{i=0}^{p}) = U_T^{sim}(\{\kappa_i\}_{i=0}^{p}) + \frac{U_T^{init}(\{\psi_t\}_{t=1}^{T})}{\kappa_0} + \sum_{t=1}^{T} \left[\frac{f_t^{init}(\psi_t)}{\kappa_0} + f_t^{rate}(\{\kappa_i\}_{i=0}^{p}) \right], \tag{51}
$$

and $\{\psi_t\}_{t=1}^{T}$ is any comparator sequence. Then, the following bounds on TAOG and TACV hold:

$$
\bar{R}_i \leq L(\{\kappa_i^*\}_{i=0}^{p}), \qquad \forall\, i = 0, ..., p. \tag{52}
$$

Proof. The idea of the proof is to freeze the learning rates first to obtain a dynamic regret bound based on policy initialization, and then optimize over the learning rates to obtain a tighter characterization. Also, since TAOG and TACV only differ by a bias term that does not depend on either the learning rates nor the initial policy, we can treat them indistinguishably. In particular,

$$\sum_{t=1}^{T} U_t(\pi_{t,0}, \{\alpha_{t,i}\}_{i=0}^{p}) = \sum_{t=1}^{T} f_t^{sim}(\{\alpha_{t,i}\}_{i=0}^{p}) \tag{53}$$

$$\leq \min_{\{\kappa_i\}_{i=0}^{p}} \quad U_T^{sim}(\{\kappa_i\}_{i=0}^{p}) + \sum_{t=1}^{T} f_t^{sim}(\{\kappa_i\}_{i=0}^{p}) \tag{54}$$

$$\leq \min_{\{\kappa_i\}_{i=0}^{p}} \quad U_T^{sim}(\{\kappa_i\}_{i=0}^{p}) + \frac{U_T^{init}(\Psi)}{\kappa_0} + \sum_{t=1}^{T} \left[\frac{f_t^{init}(\psi_t)}{\kappa_0} + f_t^{rate}(\{\kappa_i\}_{i=0}^{p}) \right]. \tag{55}$$

where $\Psi := \{\psi_t\}_{t=1}^{T}$. Let

$$L(\{\kappa_i\}_{i=0}^{p}) = U_T^{sim}(\{\kappa_i\}_{i=0}^{p}) + \frac{U_T^{init}(\Psi)}{\kappa_0} + \sum_{t=1}^{T} \left[\frac{f_t^{init}(\psi_t)}{\kappa_0} + f_t^{rate}(\{\kappa_i\}_{i=0}^{p}) \right] \tag{56}$$

and define

$$\{\kappa_i^*\}_{i=0}^{p} = \arg\min L(\{\kappa_i\}_{i=0}^{p}). \tag{57}$$

Thus, plugging $\{\kappa_i^*\}_{i=0}^{p}$ in (55) results in (52). □

Remark 4. *The above bound is developed for a general comparator sequence and online algorithms. To get more insights, let*

$$\tilde{L}(\{\kappa_i\}_{i=0}^{p}) = \frac{U_T^{init}(\Psi)}{\kappa_0} + \sum_{t=1}^{T} \left[\frac{f_t^{init}(\psi_t)}{\kappa_0} + f_t^{rate}(\{\kappa_i\}_{i=0}^{p}) \right] \tag{58}$$

and choose $\kappa_0' = (1-\gamma)^{1.5}/\sqrt{|\mathcal{S}||\mathcal{A}|M}$ as the original CRPO. Note that the difference between L and $\tilde{L}$ is that the latter does not consider the effect of $U_T^{sim}(\{\kappa_i\}_{i=0}^{p})$. Now, define

$$\{\kappa_i'\}_{i=1}^{p} = \underset{\kappa_i \geq \kappa_0, i=1,\ldots,p}{\arg\min} \quad \tilde{L}(\kappa_0'\{\kappa_i\}_{i=1}^{p}). \tag{59}$$

To proceed, using the KKT optimality conditions, we have that

$$\frac{\partial \tilde{L}}{\partial \kappa_i} = \sum_{t=1}^{T} \left(\frac{2c_2^t M \kappa_i}{\kappa_0} + \frac{(c_3^t + 2c_4^t \kappa_i)\sqrt{M}}{\kappa_0} \right). \tag{60}$$

Equation (60) *implies that $\frac{\partial \tilde{L}}{\partial \kappa_i} > 0$ is always positive, and given the constraint assumption that $\kappa_i' \geq \kappa_0'$, we can infer that the optimal solution must be taken at the boundary, i.e., $\kappa_i' = \kappa_0'$ for $i = 1, .., p$. Note that in this case $\{\kappa_i'\}_{i=0}^{p}$ recovers the original CRPO algorithm (no rate adaptation). Since $\{\kappa_i'\}_{i=0}^{p}$ is also a feasible solution to* (57)*, we can see that adaptively choosing the learning rates is provably better than the unadaptive version, and the reason is exactly attributed to the fact that we are considering the effect of $U_T^{sim}(\{\kappa_i\}_{i=0}^{p})$, which accounts for the static regret of choosing the learning rates.*

Remark 5. *Note that we have corrected a mistake in the original submission.*

G Proofs for Section 3.3

G.1 TAOG and TACV bounds for CRPO under function approximation settings

In this section, we will derive the TAOG and TACV regret bounds for CRPO under function approximation settings. Specifically, we parameterize the policy by a two-layer neural network $\theta(s,a) := f((s,a);\omega,b) = \frac{1}{\sqrt{W}} \sum_{\iota=1}^{W} b_\iota \cdot \mathrm{ReLU}(\omega_\iota^\top \xi(x,a))$ for any state-action pair (s,a), where ι is the index of the neuron within the hidden layer, $\xi(s,a) \in \mathbb{R}^d$ is the feature vector with $d \geq 2$ and $\|\xi(s,a)\| \leq 1$, $\mathrm{ReLU}(x) = \mathbb{1}(x > 0) \cdot x$, $b = [b_1, \cdots, b_W]^\top \in \mathbb{R}^W$, and $\omega = [\omega_1^\top, \cdots, \omega_W^\top]^\top \in \mathbb{R}^{Wd}$ is the weight matrix with $\|\omega_\iota\|_2^2 \geq d_1$ for $\iota \in [W]$ and $\|\omega' - \omega\|_2 \leq d_\omega$ for any two weight matrices ω' and ω (i.e., the diameter of the space of weight matrices is bounded).

For simplicity, we use x to denote the state action-pair as (s, a). Recall that the stationary distribution induced by π_ω is ν_{π_ω}. We also let $\delta_k(x, x', \omega_k)$ to be the TD error at iteration k (within the critic update loop). The indicator function $\mathbb{1}(\cdot)$ takes the value of 1 if the inside expression is true and 0 otherwise.

To establish the convergence rate of the TD with high probability, we begin by extending [93] to the case of adaptive learning rates. In the following we let $\omega_k = [\omega_{k,1}^\top, \cdots, \omega_{k,W}^\top]^\top$ be the weight matrix at iteration k (within the critic update loop).

Lemma 18. *Suppose that Assumption 3 holds, then, for any policy π and iteration index $k \geq 0$, the following holds:*

$$\mathbb{E}_{\nu_\pi}\left[\frac{1}{W}\sum_{\iota=1}^{W}\left|\mathbb{1}\big(\omega_{k,\iota}^\top \xi(x) > 0\big) - \mathbb{1}\big(\omega_{0,\iota}^\top \xi(x) > 0\big)\right|\right] \leq \frac{C_0}{d_1^2 W}\|\omega_k - \omega_0\|_2^2. \tag{61}$$

Proof. If $\mathbb{1}\big(\omega_{k,\iota}^\top \xi(x) > 0\big) = \mathbb{1}\big(\omega_{0,\iota}^\top \xi(x) > 0\big)$, then the corresponding term in the LHS of (61) can be ignored. We only consider the case that $\mathbb{1}\big(\omega_{k,\iota}^\top \xi(x) > 0\big) \neq \mathbb{1}\big(\omega_{0,\iota}^\top \xi(x) > 0\big)$, which implies that:

$$|\omega_{0,\iota}^\top \xi(x)| \leq |\omega_{k,\iota}^\top \xi(x) - \omega_{0,\iota}^\top \xi(x)| \leq \|\omega_{k,\iota} - \omega_{0,\iota}\|_2, \tag{62}$$

where the last inequality is due to Cauchy-Schwarz. Thus, the above implies that:

$$\left|\mathbb{1}\big(\omega_{k,\iota}^\top \xi(x) > 0\big) - \mathbb{1}\big(\omega_{0,\iota}^\top \xi(x) > 0\big)\right| \leq \mathbb{1}\big(|\omega_{0,\iota}^\top \xi(x)| \leq \|\omega_{k,\iota} - \omega_{0,\iota}\|_2\big). \tag{63}$$

Using the above relations, we can obtain the following upper bound:

$$\begin{aligned}
&\mathbb{E}_{\mu_\pi}\left[\frac{1}{W}\sum_{\iota=1}^{W}\left|\mathbb{1}\big(\omega_{k,\iota}^\top \xi(x) > 0\big) - \mathbb{1}\big(\omega_{0,\iota}^\top \xi(x) > 0\big)\right|\right] \\
&\leq \mathbb{E}_{\mu_\pi}\left[\frac{1}{W}\sum_{\iota=1}^{W}\mathbb{1}\big(|\omega_{0,\iota}^\top \xi(x)| \leq \|\omega_{k,\iota} - \omega_{0,\iota}\|_2\big)\right] \\
&= \frac{1}{W}\sum_{\iota=1}^{W}\mathbb{P}_{\mu_\pi}\big(|\omega_{0,\iota}^\top \xi(x)| \leq \|\omega_{0,\iota} - \omega_{0,\iota}\|_2\big) \\
&\overset{(i)}{\leq} \frac{C_0}{W}\sum_{\iota=1}^{W}\frac{\|\omega_{k,\iota} - \omega_{0,\iota}\|_2^2}{\|\omega_{0,\iota}\|_2^2} \\
&\overset{(ii)}{\leq} \frac{C_0}{W d_1^2}\|\omega_k - \omega_0\|_2^2,
\end{aligned}$$

where the inequality (i) follows from Assumption 3, and inequality (ii) follows from the fact that $\|\omega_{0,\iota}\|_2^2 \geq d_1$. □

Lemma 19. *Suppose that Assumption 3 holds, then, for any policy π and iteration index k, the following will hold:*

$$\mathbb{E}_{\nu_\pi}\left[|f\big((s,a);\omega_k\big) - f_0\big((s,a);\omega_k\big)|^2\right] \leq \frac{4 d_\omega^2 C_0}{W d_1^2}\|\omega_k - \omega_0\|_2^2. \tag{64}$$

Proof. We proceed as follows

$$\begin{aligned}
&|f\big((s,a);\omega_k\big) - f_0\big((s,a);\omega_k\big)| \\
&= \frac{1}{\sqrt{W}}\left|\sum_{\iota=1}^{W}\Big(\mathbb{1}\big(\omega_{k,\iota}^\top \xi(x) > 0\big) - \mathbb{1}\big(\omega_{0,\iota}^\top \xi(x) > 0\big)\Big) b_\iota \omega_{k,\iota}^\top \xi(x)\right| \\
&\leq \frac{1}{\sqrt{W}}\sum_{\iota=1}^{W}\left|\Big(\mathbb{1}\big(\omega_{k,\iota}^\top \xi(x) > 0\big) - \mathbb{1}\big(\omega_{0,\iota}^\top \xi(x) > 0\big)\Big)\right| |b_\iota| \|\omega_{k,\iota}^\top \xi(x)\|_2 \\
&\leq \frac{1}{\sqrt{W}}\sum_{\iota=1}^{W}\mathbb{1}(|\omega_{0,\iota}^\top \xi(x)| \leq \|\omega_{k,\iota} - \omega_{0,l}\|_2)\|\omega_{k,\iota}^\top \xi(x)\|_2 \\
&\leq \frac{1}{\sqrt{W}}\sum_{\iota=1}^{W}\mathbb{1}(|\omega_{0,\iota}^\top \xi(x)| \leq \|\omega_{k,\iota} - \omega_{0,\iota}\|_2)\big(\|\omega_{0,\iota} - \omega_{k,\iota}\|_2 + \|\omega_{0,\iota}^\top \xi(x)\|_2\big) \\
&\leq \frac{2}{\sqrt{W}}\sum_{\iota=1}^{W}\mathbb{1}(|\omega_{0,\iota}^\top \xi(x)| \leq \|\omega_{k,\iota} - \omega_{0,\iota}\|_2)\|\omega_{0,\iota} - \omega_{k,\iota}\|_2
\end{aligned}$$

where the first inequality is due to Cauchy-Schwarz, the second inequality follows from (63), and the last inequality is due to (62). Taking the square and expectation, we have

$$\begin{aligned}
&\mathbb{E}_{\nu_\pi}\Big[|f\big((s,a);\omega_k\big) - f_0\big((s,a);\omega_k\big)|^2\Big] \\
&\leq \frac{4}{W}\left[\left(\sum_{\iota=1}^{W}\mathbb{1}(|\omega_{0,\iota}^\top \xi(x)| \leq \|\omega_{k,\iota} - \omega_{0,\iota}\|_2)\|\omega_{0,\iota} - \omega_{k,\iota}\|_2\right)^2\right] \\
&\leq \frac{4}{W}\left[\sum_{\iota=1}^{W}\mathbb{1}(|\omega_{0,\iota}^\top \xi(x)| \leq \|\omega_{k,\iota} - \omega_{0,\iota}\|_2)\sum_{\iota=1}^{W}\|\omega_{0,\iota} - \omega_{k,\iota}\|_2^2\right] \\
&= \frac{4d_\omega^2}{W}\mathbb{E}_{\nu_\pi}\left[\sum_{\iota=1}^{W}\mathbb{1}(|\omega_{0,\iota}^\top \xi(x)| \leq \|\omega_{k,\iota} - \omega_{0,\iota}\|_2)\right] \\
&\leq \frac{4d_\omega^2 C_0}{W d_1^2}\|\omega_k - \omega_0\|,
\end{aligned}$$

where second inequality follows from the Hölder's inequality, and the last inequality follows from Assumption 3 and Lemma 18. □

Lemma 20. *Suppose that Assumption 3 holds, then, there exists a positive constant C_g such that for any policy π and the iteration index k, with probability of at least $1-\delta$, we have*

$$\|\bar{g}_k(\omega_k) - \bar{g}_0(\omega_k)\|_2 \leq \frac{9\sqrt{C_0}d_\omega^2}{\sqrt{W}d_1}\sqrt{C_g + \log\frac{1}{\delta}}\|\omega_k - \omega_0\|_2,$$

where $C_g = \frac{1}{72C_0 d_\omega^2}\Big(32C_0 + 36C_0 d_\omega^2 + \frac{42c_{max}^2 C_0}{(1-\gamma)^2}\Big)$.

Proof. By definition, we have

$$\begin{aligned}
&\|\bar{g}_k(\omega_k) - \bar{g}_0(\omega_k)\|_2 \\
&= \|\mathbb{E}_{\nu_\pi}[\delta_k(x,x',\omega_k)\nabla_\omega f(x;\omega_k)] - \mathbb{E}_{\nu_\pi}[\delta_0(x,x',\omega_k)\nabla_\omega f_0(x,\omega_k)]\|_2 \\
&= \|\mathbb{E}_{\nu_\pi}[(\delta_k(x,x',\omega_k) - \delta_0(x,x',\omega_k))\nabla_\omega f(x;\omega_k) + \delta_0(x,x',\omega_k)(\nabla_\omega f(x;\omega_k) - \nabla_\omega f_0(x,\omega_k))\|_2 \\
&\leq \mathbb{E}_{\nu_\pi}[|\delta_k(x,x',\omega_k) - \delta_0(x,x',\omega_k)|\|\nabla_\omega f(x;\omega_k)\|_2 \\
&\qquad\qquad + |\delta_0(x,x',\omega_k)|\|(\nabla_\omega f(x;\omega_k) - \nabla_\omega f_0(x,\omega_k))\|_2] \\
&\leq \mathbb{E}_{\nu_\pi}[|\delta_k(x,x',\omega_k) - \delta_0(x,x',\omega_k)|] + \mathbb{E}_{\nu_\pi}[|\delta_0(x,x',\omega_k)|\|(\nabla_\omega f(x;\omega_k) - \nabla_\omega f_0(x,\omega_k))\|_2],
\end{aligned}$$

where the last inequality follows from the fact that $\|\nabla_\omega f(x, \omega_k)\|_2 \le 1$. Using the above relation, we have

$$\begin{aligned}
&\|\bar{g}_k(\omega_k) - \bar{g}_0(\omega_k)\|_2^2 \\
&\le 2\mathbb{E}_{\nu_\pi}[|\delta_k(x, x', \omega_k) - \delta_0(x, x', \omega_k)|^2] + 2(\mathbb{E}_{\nu_\pi}[|\delta_0(x, x', \omega_k)|\|(\nabla_\omega f(x; \omega_k) - \nabla_\omega f_0(x, \omega_k))\|_2])^2 \\
&\le 2\underbrace{\mathbb{E}_{\nu_\pi}[|\delta_k(x, x', \omega_k) - \delta_0(x, x', \omega_k)|^2]}_{(i)} \\
&\qquad\qquad + 2\underbrace{\mathbb{E}_{\nu_\pi}[|\delta_0(x, x', \omega_k)|^2]}_{(ii)}\underbrace{\mathbb{E}_{\nu_\pi}[\|(\nabla_\omega f(x; \omega_k) - \nabla_\omega f_0(x, \omega_k))\|_2^2]}_{(iii)}.
\end{aligned}$$

To bound the term (i), note that

$$\begin{aligned}
&\mathbb{E}_{\nu_\pi}[|\delta_k(x, x', \omega_k) - \delta_0(x, x', \omega_k)|^2] \\
&\le \mathbb{E}_{\nu_\pi}[|f(x, \omega_k) - f_0(x, \omega_k) - \gamma(f(x', \omega_k) - f_0(x', \omega_k))|^2] \\
&\le 2\mathbb{E}_{\nu_\pi}[|f(x, \omega_k) - f_0(x, \omega_k)|^2] + 2\mathbb{E}_{\nu_\pi}[|f(x', \omega_k) - f_0(x', \omega_k)|^2] \\
&= 4\mathbb{E}_{\nu_\pi}[|f(x, \omega_k) - f_0(x, \omega_k)|^2] \\
&\le \frac{16 d_\omega^2 C_0}{W d_1^2}\|\omega_k - \omega_0\|_2^2, \qquad (65)
\end{aligned}$$

where the last inequality is due to Lemma 19. To bound the term (iii), note that

$$\begin{aligned}
&\|\nabla_\omega f(x; \omega_k) - \nabla_\omega f_0(x, \omega_k)\|_2 \\
&= \frac{1}{\sqrt{W}}\left\|\sum_{\iota=1}^{W}\Big(\mathbb{1}\big(\omega_{k,\iota}^\top \xi(x) > 0\big) - \mathbb{1}\big(\omega_{0,\iota}^\top \xi(x) > 0\big)\Big) b_\iota \omega_{0,\iota}^\top \xi(x)\right\|_2 \\
&\le \frac{1}{\sqrt{W}}\sum_{\iota=1}^{W}\Big|\mathbb{1}\big(\omega_{k,\iota}^\top \xi(x) > 0\big) - \mathbb{1}\big(\omega_{0,\iota}^\top \xi(x) > 0\big)\Big|\|\omega_{0,\iota}\|_2 \\
&\le \frac{1}{\sqrt{W}}\sum_{\iota=1}^{W}\mathbb{1}\big(|\omega_{0,\iota}^\top \xi(x)| \le \|\omega_{k,\iota} - \omega_{0,\iota}\|\big)\|\omega_{0,\iota}\|_2
\end{aligned}$$

where the first inequality is because $|b_\tau| \le 1$ and $\|\xi(x)\|_2 \le 1$, and the second inequality is due (63). Hence,

$$\begin{aligned}
&\mathbb{E}_{\mu_\pi}[\|\nabla_\omega f(x; \omega_k) - \nabla_\omega f_0(x, \omega_k)\|_2^2] \\
&\le \frac{1}{W}\mathbb{E}_{\mu_\pi}\left[\left(\sum_{\iota=1}^{W}\mathbb{1}\big(|\omega_{0,\iota}^\top \xi(x)| \le \|\omega_{k,\iota} - \omega_{0,\iota}\|\big)\right)\left(\sum_{\iota=1}^{W}\|\omega_{0,\iota}\|_2^2\right)\right] \\
&\le \frac{d_\omega^2}{W}\mathbb{E}_{\mu_\pi}\left[\sum_{\iota=1}^{W}\mathbb{1}(|\omega_{0,\iota}^\top \xi(x)| \le \|\omega_{k,\iota} - \omega_{0,\iota}\|_2)\right] \\
&\le \frac{C_0 d_\omega^2}{W d_1^2}\|\omega_k - \omega_0\|_2^2, \qquad (66)
\end{aligned}$$

where the last inequality follows from the Lemma 18. Finally, to bound the term (ii), we recall the bound [93, Lem. 13, Eq. 33], which holds with probability of at least $1 - \delta$:

$$\mathbb{E}_{\nu_\pi}[|\delta_0(x, x', \omega_k)|^2] \le 18 d_\omega^2 + \frac{21 c_{max}^2}{(1-\gamma)^2} + 72 d_\omega^2 \log\left(\frac{1}{\delta}\right) \qquad (67)$$

Hence, combining all the three upper bounds from (66), (65), (67), we get the final upper bound

$$\|\bar{g}_k(\omega_k) - \bar{g}_0(\omega_k)\|_2^2 \le \frac{72 C_0 d_\omega^4}{W d_1^2}\left(C_g + \log\frac{1}{\delta}\right)\|\omega_k - \omega_0\|_2^2, \qquad (68)$$

where the constant $C_g = \frac{1}{72 C_0 d_\omega^2}\left(32 C_0 + 36 C_0 d_\omega^2 + \frac{42 c_{max}^2 C_0}{(1-\gamma)^2}\right)$. By taking square root on both sides, we complete the proof. □

Remark 6. *While the above developments follows closely those of [93], we make a few remarks on the technical differences. First, our analysis reveals the dependence of the bound on the initial parameter ω_0, which is required for exploiting meta-initialization of the critic network. Another key difference arises from the Assumption 3. The main difference can be observed by comparing our bound in Lemma 18 and [93, Lemma 11] (the line directly above [93, Eq. 23]). In particular, the dependence on the term $\|\omega_k - \omega_0\|_2$ is quadratic in our case but linear in [93, Lemma 11]. This difference can be also seen in Lemma 19 and [93, Lemma 12]. As a result, in our Lemma 20, we are able to obtain a dependence on $\|\omega_k - \omega_0\|_2$ that is comparable in order of the RHS, while the dependence would be $\sqrt{\|\omega_k - \omega_0\|_2}$ in the case of [93, Lemma 13]. This difference has major implications for the error propagation in meta-initialization of the critic network, as will be evident in our later analysis.*

We now use these previous results to derive the convergence rate of TD learning, revealing the key dependence on the initialization parameter ω_0. We omit the dependence of the critic parameter on task index t, reward/constraint i, and step dependence m for the notational simplicity, since the results are direct replicates for each case; specifically, we will use ω_k to denote $\omega_{t,m}^{i,k}$ (i.e., the critic parameter for reward/constraint i at the k-th iteration after the m-th step of policy update).

Lemma 21 (Convergence rate of TD with high probability)**.** *Let $\bar{\omega}_K = \frac{1}{K}\sum_{k=0}^{K-1}\omega_k$ denote the average of the parameters from $k = 0$ to $K-1$. Let $\bar{Q}^i_K = f((s,a);\bar{\omega}_K)$ denote the estimate of the true Q-value $Q^i_{\pi_m}$, where π_m is the policy after the m-th step policy update and the Q-value parameter is the last iterate ω_K. Then, with probability of at least $1-\delta$, we have*

$$\|\bar{Q}^i_K(s,a) - Q^i_{\pi_m}(s,a)\|^2_{\nu_\pi} \le c'_1(K,W)\|\omega^* - \omega_0\|_2^2 + c'_2(K,W),$$

where

$$c'_1(K,W) := \frac{72}{(1-\gamma)^2\sqrt{K}} + \frac{12 d_\omega^2 C_0}{W d_1^2}$$
$$c'_2(K,W) := \Theta\left(\frac{\log(1/\delta)}{W}\right) + \Theta\left(\frac{\sqrt{\log(1/\delta)}}{(1-\gamma)^2\sqrt{K}}\right) + \Theta\left(\frac{\sqrt{\log(K/\delta)}}{(1-\gamma)W^{1/4}}\right) + \Theta\left(\frac{\log(K/\delta)}{(1-\gamma)^2\sqrt{W}}\right)$$

Proof. To begin with, we recall the following bound from [93, Eq. 34]:

$$\begin{aligned}\|\omega_{k+1} - \omega^*\|_2^2 \le \|\omega_k - \omega^*\|_2^2 &- [2\beta(1-\gamma) - 12\beta^2]\mathbb{E}_{\nu_\pi}[(f_0(x;\omega_k) - f_0(x;\omega^*))^2] \\ &+ 2\beta(\bar{g}_k(\omega_k) - g_k(\omega_k))^\top(\omega_k - \omega^*) + 4d_\omega\beta\|\bar{g}_0(\omega_k) - \bar{g}_k(\omega_k)\|_2 \\ &+ 3\beta^2\|g_k(\omega_k) - \bar{g}_k(\omega_k)\|_2^2. \end{aligned} \tag{69}$$

Rearranging the above relation yields

$$\begin{aligned}&[2\beta(1-\gamma) - 12\beta^2]\mathbb{E}_{\mu_\pi}\left[\left(f_0((s,a);\omega_k) - f_0((s,a);\omega^*)\right)^2\right] \\ &\le \|\omega_k - \omega^*\|_2^2 - \|\omega_{k+1} - \omega^*\|_2^2 + 2\beta(\bar{g}_k(\omega_k) - g_k(\omega_k))^\top(\omega_k - \omega^*) \\ &\quad + 4d_\omega\beta\|\bar{g}_0(\omega_k) - \bar{g}_k(\omega_k)\|_2 + 3\beta^2\|g_k(\omega_k) - \bar{g}_k(\omega_k)\|_2^2 + 3\beta^2\|\bar{g}_k(\omega_k) - \bar{g}_0(\omega_k)\|_2^2. \end{aligned} \tag{70}$$

Taking the sum of (70) over $k = 0, 1, ..., K-1$ yields

$$\begin{aligned}&[2\beta(1-\gamma) - 12\beta^2]\sum_{k=0}^{K-1}\mathbb{E}_{\nu_\pi}\left[\left(f_0((s,a);\omega_k) - f_0((s,a);\omega^*)\right)^2\right] \\ &\le \|\omega_0 - \omega^*\|_2^2 \\ &\quad + 2\beta\sum_{k=0}^{K-1}(\bar{g}_k(\omega_k) - g_k(\omega_k))^\top(\omega_k - \omega^*) + 4d_\omega\beta\sum_{k=0}^{K-1}\|\bar{g}_0(\omega_k) - \bar{g}_k(\omega_k)\|_2 \\ &\quad + 3\beta^2\sum_{k=0}^{K-1}\|g_k(\omega_k) - \bar{g}_k(\omega_k)\|_2^2 + 3\beta^2\sum_{k=0}^{K-1}\|\bar{g}_k(\omega_k) - \bar{g}_0(\omega_k)\|_2^2. \end{aligned} \tag{71}$$

We will bound all the terms on the RHS as done in the proof of [93, Lemma 1], except the term $\sum_{k=0}^{K-1} \|g_k(\omega_k) - \bar{g}_k(\omega_k)\|_2^2$, which will be bounded by recalling (68) in Lemma 20:

$$\sum_{k=0}^{K-1} \|\bar{g}_k(\omega_k) - \bar{g}_0(\omega_k)\|_2^2 \leq \frac{72C_0 d_\omega^4}{W d_1^2}\left(C_g + \log\frac{1}{\delta_1}\right)\sum_{k=0}^{K-1} \|\omega_k - \omega_0\|_2^2. \tag{72}$$

Combining the above bound with [93, Eqs. 38, 39, 40] and applying the union bound, the following holds with probability of at least $1 - (\delta_1 + \delta_2 + \delta_3 + \delta_4)$ (where $\delta_1, ..., \delta_4$ are associated with the bounds in (72) and [93, Eqs. 38, 39, 40], respectively):

$$\begin{aligned}
&[2\beta(1-\gamma) - 12\beta^2]\sum_{k=0}^{K-1} \mathbb{E}_{\nu_\pi}\left[\left(f_0((s,a);\omega_k) - f_0((s,a);\omega^*)\right)^2\right] \\
&\leq 2\|\omega_0 - \omega^*\|_2^2 + \frac{216\beta^2 C_0 d_\omega^4}{W d_1^2}\left(C_g + \log\frac{1}{\delta_1}\right)\sum_{k=0}^{K-1}\|\omega_k - \omega_0\|_2^2 \\
&\qquad + 10\beta C_\zeta\sqrt{\log\left(\frac{1}{\delta_2}\right)}\sqrt{K} + \beta K \Theta\left(\frac{\sqrt{\log(K/\delta_3)}}{(1-\gamma)W^{1/4}}\right) + \beta^2 K\Theta\left(\frac{\log(K/\delta_4)}{(1-\gamma)^2\sqrt{W}}\right).
\end{aligned} \tag{73}$$

Divide both sides by $[2\beta(1-\gamma) - 12\beta^2]K$ with the stepsize $\beta = \min\{1/\sqrt{K}, (1-\gamma)/12\}$, which implies that $\frac{1}{\sqrt{K}[2\beta(1-\gamma)-12\beta^2]} \leq \frac{12}{(1-\gamma)^2}$. Then with probability of at least $1 - (\delta_1 + \delta_2 + \delta_3 + \delta_4)$, we have

$$\begin{aligned}
&\|f_0((s,a);\bar{\omega}_K) - f_0((s,a);\omega^*)\|_{\nu_\pi}^2 \\
&\leq \frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}_{\nu_\pi}[(f_0((s,a);\omega_k) - f_0((s,a;\omega^*)))^2] \\
&\leq \frac{24\|\omega_0 - \omega^*\|_2^2}{(1-\gamma^2)\sqrt{K}} + \frac{1}{K}\sum_{k=0}^{K-1}\|\omega_k - \omega_0\|_2^2\Theta\left(\frac{\log(1/\delta_1)}{W}\right) \\
&\qquad + \Theta\left(\frac{\sqrt{\log(1/\delta_2)}}{(1-\gamma)^2\sqrt{K}}\right) + \Theta\left(\frac{\sqrt{\log(K/\delta_3)}}{(1-\gamma)W^{1/4}}\right) + \Theta\left(\frac{\log(K/\delta_4)}{(1-\gamma)^2\sqrt{W}}\right).
\end{aligned} \tag{74}$$

Note in the term $\frac{1}{K}\sum_{k=0}^{K-1}\|\omega_k - \omega_0\|_2^2$ above also depends on the initialization ω_0.

Finally, we upper bound the term $\|f((s,a);\bar{\omega}_K) - Q_\pi(s,a)\|_{\nu_\pi}^2$ by decomposing it as follows

$$\begin{aligned}
&\|f((s,a);\bar{\omega}_K) - Q_\pi(s,a)\|_{\nu_\pi}^2 \\
&\leq 3\underbrace{\|f((s,a);\bar{\omega}_K) - f_0((s,a);\bar{\omega}_K)\|_{\nu_\pi}^2}_{(i)} \\
&\qquad + 3\underbrace{\|f_0((s,a);\bar{\omega}_K) - f_0((s,a);\omega^*)\|_{\nu_\pi}^2}_{(ii)} + 3\underbrace{\|f_0((s,a);\omega^*) - Q_\pi(s,a)\|_{\nu_\pi}^2}_{(iii)}
\end{aligned}$$

The term (i) can be bounded by Lemma 19, and (ii) can be bounded by (74). The term (iii) can be bounded by

$$\|f_0((s,a);\omega^*) - Q_\pi(s,a)\|_{\nu_\pi}^2 \leq \frac{1}{1-\gamma}\|f_0((s,a);\omega_\pi^*) - Q_\pi(s,a)\|_{\nu_\pi}^2 \leq \frac{4d_\omega^2 \log\left(\frac{1}{\delta_5}\right)}{W},$$

where the first inequality is due [20], and the second relation holds with probability at least $1 - \delta_5$ [93, Eq. 32] (which is implied by [80]).

Putting it together and let $\delta_1 = \delta_2 = \delta_3 = \delta_4 = \delta_5 = \delta/5$, with probability of at least $1-\delta$, the following holds:

$$
\begin{aligned}
&\|f((s,a);\bar{\omega}_K) - Q_\pi(s,a)\|_{\nu_\pi}^2 \\
&\le \frac{72\|\omega_0-\omega^*\|_2^2}{(1-\gamma^2)\sqrt{K}} + \frac{1}{K}\sum_{k=0}^{K-1}\|\omega_k-\omega_0\|_2^2\Theta\bigg(\frac{\log(1/\delta_1)}{W}\bigg) + \Theta\bigg(\frac{\log(1/\delta_5)}{W}\bigg) \\
&\quad + \Theta\bigg(\frac{\sqrt{\log(1/\delta_2)}}{(1-\gamma)^2\sqrt{K}}\bigg) + \Theta\bigg(\frac{\sqrt{\log(K/\delta_3)}}{(1-\gamma)W^{1/4}}\bigg) + \Theta\bigg(\frac{\log(K/\delta_4)}{(1-\gamma)^2\sqrt{W}}\bigg) \\
&\le \bigg(\frac{72}{(1-\gamma)^2\sqrt{K}} + \frac{12d_\omega^2 C_0}{Wd_1^2}\bigg)\|\omega^*-\omega_0\|_2^2 + \Theta\bigg(\frac{\log(1/\delta)}{W}\bigg) \\
&\quad + \Theta\bigg(\frac{\sqrt{\log(1/\delta)}}{(1-\gamma)^2\sqrt{K}}\bigg) + \Theta\bigg(\frac{\sqrt{\log(K/\delta)}}{(1-\gamma)W^{1/4}}\bigg) + \Theta\bigg(\frac{\log(K/\delta)}{(1-\gamma)^2\sqrt{W}}\bigg)
\end{aligned}
$$

where the last inequality also uses $\|\omega_k-\omega_0\|_2^2 \le d_\omega^2$ for simplicity, since the terms $\|\omega_0-\bar{\omega}_K\|_2^2$ and $\sum_{k=0}^{K-1}\|\omega_k-\omega_0\|_2^2$ share identical gradient in terms of ω_0. This concludes the proof. □

Next, we show the counterpart result of Lemma 15 under function approximation settings with different learning rates. Recall that $\mathcal{N}_{t,i}$ is the set of steps that the CRPO algorithm minimizes the i-th constrain for $i = 1,...,p$ and $\mathcal{N}_{t,0}$ is the set of steps when the reward is maximized. By following the argument similar to that in [2], we can show that $\log(\pi_\omega(a|s))$ is L_f-Lipschitz. Next, we restate the following assumptions [93].

Assumption 7 ([93]). *For the state visitation distribution of the global optimal policy ν_t^*, there exists a constant C_{RN} such that for all policies π, the following holds*

$$
\int_x \bigg(\frac{d\nu_t^*(x)}{d\nu_\pi(x)}\bigg) d\nu_\pi(x) \le C_{RN}^2. \tag{75}
$$

Assumption 8 ([93]). *For any policy π_t, there exists a constant $C_f > 0$ such that for all $k \ge 0$,*

$$
\mathbb{E}_{\rho_t\pi_t}\Big[\exp([\bar{Q}_{t,\pi_m}^i(s,a) - \mathbb{E}\bar{Q}_{t,\pi_m}^i(s,a)]^2/C_f^2)\Big] \le 1.
$$

Lemma 22. *Consider the CRPO algorithm updates for the neural network approximation setting. Let $K = ((1-\gamma)^2\sqrt{W})$, and let the empirical average parameter $N = M\log(2M/\delta)$. Then with probability of at least $1-\delta$, we have*

$$
\begin{aligned}
&\alpha_{t,0}(1-\gamma)\sum_{m\in\mathcal{N}_{t,0}}\Big(J_{t,0}(\pi_t^*) - J_{t,0}(\pi_{t,m})\Big) + (1-\gamma)\eta_t\sum_{i=1}^{p}\alpha_{t,i}|\mathcal{N}_{t,i}| \\
&\le \mathbb{E}_{\nu_t^*}[D(\pi_t^*|\pi_{t,0})] + \frac{8C_{RN}\sqrt{C_0}Z^{1.5}}{\sqrt{d_1}W^{1/4}}\sum_{i=0}^{p}\alpha_{t,i}|\mathcal{N}_{t,i}| + L_f(Z^2+Wd_2^2)\sum_{i=0}^{p}\alpha_{t,i}^2|\mathcal{N}_{t,i}| \\
&+ 3C_{RN}\sum_{i=0}^{p}\sum_{m\in\mathcal{N}_{t,i}}\alpha_{t,i}\|f_{t,i}((s,a);\bar{\omega}_m) - Q_{t,\pi_m}^i(s,a)\|_{\nu_{\pi_m}} + 2C_f(1-\gamma)\sqrt{M}\sum_{i=1}^{p}\alpha_{t,i}
\end{aligned}
$$

Proof. Let $\pi_{t,m})$ be the policy for the m-th update. If $m \in \mathcal{N}_{t,0}$, by [93, Lemma 15], we have that

$$
\begin{aligned}
&\alpha_{t,0}(1-\gamma)\big(J_{t,0}(\pi_t^*) - J_{t,0}(\pi_{t,m})\big) \\
&\le \mathbb{E}_{\nu_t^*}[D(\pi_t^*|\pi_{t,m})] - \mathbb{E}_{\nu_t^*}[D(\pi_t^*|\pi_{t,m+1})] + \frac{8\alpha_{t,0}C_{RN}\sqrt{C_0}d_\omega^{1.5}}{\sqrt{d_1}W^{1/4}} \\
&\quad + \alpha_{t,0}^2L_f(d_\omega^2+Wd_2^2) + 2\alpha_{t,0}C_{RN}\|f_0((s,a);\bar{\omega}_m) - Q_{t,\pi_m}^0(s,a)\|_{\nu_{\pi_m}}.
\end{aligned}
$$

Similarly, if $m \in \mathcal{N}_{t,i}$, we have

$$\begin{aligned}
&\alpha_{t,i}(1-\gamma)\big(J_{t,i}(\pi_{t,m}) - J_{t,i}(\pi_t^*))\big) \\
&\leq \mathbb{E}_{\nu_t^*}[D(\pi_t^*|\pi_{t,m})] - \mathbb{E}_{\nu_t^*}[D(\pi_t^*|\pi_{t,m+1})] + \frac{8\alpha_{t,i}C_{RN}\sqrt{C_0}d_\omega^{1.5}}{\sqrt{d_1}W^{1/4}} \\
&\quad + \alpha_{t,i}^2 L_f(d_\omega^2 + Wd_2^2) + 2\alpha_{t,i}C_{RN}\|f_i((s,a);\bar{\omega}_m) - Q^i_{t,\pi_m}(s,a)\|_{\nu_{\pi_m}}.
\end{aligned}$$

Taking the summation of the above equations from $m = 0$ to M yields

$$\begin{aligned}
&\alpha_{t,0}(1-\gamma)\sum_{m\in\mathcal{N}_{t,0}}\Big(J_{t,0}(\pi_t^*) - J_{t,0}(\pi_{t,m})\Big) + (1-\gamma)\sum_{i=1}^{p}\sum_{m\in\mathcal{N}_{t,i}}\alpha_{t,i}\Big(J_{t,i}(\pi_{t,m}) - J_{t,i}(\pi_t^*)\Big) \\
&\leq \mathbb{E}_{\nu_t^*}[D(\pi_t^*|\pi_{t,0})] + \frac{8C_{RN}\sqrt{C_0}d_\omega^{1.5}}{\sqrt{d_1}W^{1/4}}\sum_{i=0}^{p}\sum_{m\in\mathcal{N}_{t,i}}\alpha_{t,i} + L_f(d_\omega^2 + Wd_2^2)\sum_{i=0}^{p}\sum_{m\in\mathcal{N}_{t,i}}\alpha_{t,i}^2 \\
&+ 3C_{RN}\sum_{i=0}^{p}\sum_{m\in\mathcal{N}_{t,i}}\alpha_{t,i}\|f_{t,i}((s,a);\bar{\omega}_m) - Q^i_{t,\pi_m}(s,a)\|_{\nu_{\pi_m}}.
\end{aligned} \tag{76}$$

Since $\big(J_{t,i}(\pi_{t,m}) - J_{t,i}(\pi_t^*)\big) \geq \eta_t - |\bar{J}_{t,i}(\omega^i_{t,m}) - J_{t,i}(\pi_{t,m})|$ [93, Eq. 57], where $\omega^i_{t,m}$ is the estimated critic parameter for constraint i at time step m, we obtain

$$\begin{aligned}
&\alpha_{t,0}(1-\gamma)\sum_{m\in\mathcal{N}_{t,0}}\Big(J_{t,0}(\pi_t^*) - J_{t,0}(\pi_{t,m})\Big) + (1-\gamma)\eta_t\sum_{i=1}^{p}\alpha_{t,i}|\mathcal{N}_{t,i}| \\
&\leq \mathbb{E}_{\nu_t^*}[D(\pi_t^*|\pi_{t,0})] + \frac{8C_{RN}\sqrt{C_0}d_\omega^{1.5}}{\sqrt{d_1}W^{1/4}}\sum_{i=0}^{p}\alpha_{t,i}|\mathcal{N}_{t,i}| + L_f(d_\omega^2 + Wd_2^2)\sum_{i=0}^{p}\alpha_{t,i}^2|\mathcal{N}_{t,i}| \\
&\quad + 3C_{RN}\sum_{i=0}^{p}\sum_{m\in\mathcal{N}_{t,i}}\alpha_{t,i}\|f_{t,i}((s,a);\bar{\omega}_m) - Q^i_{t,\pi_m}(s,a)\|_{\nu_{\pi_m}} \\
&\quad + (1-\gamma)\sum_{i=1}^{p}\sum_{m\in\mathcal{N}_{t,i}}\alpha_{t,i}\big|\bar{J}_{t,i}(\omega^i_m) - \mathbb{E}[f_{t,i}((s,a);\bar{\omega}_m)]\big| \\
&\leq \mathbb{E}_{\nu_t^*}[D(\pi_t^*|\pi_{t,0})] + \frac{8C_{RN}\sqrt{C_0}d_\omega^{1.5}}{\sqrt{d_1}W^{1/4}}\sum_{i=0}^{p}\alpha_{t,i}|\mathcal{N}_{t,i}| + L_f(d_\omega^2 + Wd_2^2)\sum_{i=0}^{p}\alpha_{t,i}^2|\mathcal{N}_{t,i}| \\
&\quad + 3C_{RN}\sum_{i=0}^{p}\sum_{m\in\mathcal{N}_{t,i}}\alpha_{t,i}\|f_{t,i}((s,a);\bar{\omega}_m) - Q^i_{t,\pi_m}(s,a)\|_{\nu_{\pi_m}} + 2C_f(1-\gamma)\sum_{i=1}^{p}\alpha_{t,i}\sqrt{M},
\end{aligned}$$

where the last inequality follows from $\sum_{m=0}^{M}\big|\bar{J}_{t,i}(\omega^i_m) - \mathbb{E}[f_{t,i}((s,a);\bar{\omega}_m)]\big| \leq 2C_f\sqrt{M}$ [93, Eq. 63]. This concludes the proof. □

Lemma 23. *Suppose $\alpha_{t,i} \geq \alpha_{t,0}$ for all $i = 1, ..., p$, and let the empirical average parameter $N = M\log(2M/\delta)$. If*

$$\begin{aligned}
\frac{1}{2}\alpha_{t,0}(1-\gamma)\eta_t M \geq \mathbb{E}_{\nu_t^*}[D(\pi_t^*|\pi_{t,0})] &+ \frac{8C_{RN}\sqrt{C_0}d_\omega^{1.5}}{\sqrt{d_1}W^{1/4}}\sum_{i=0}^{p}\alpha_{t,i}|\mathcal{N}_{t,i}| \\
&+ L_f(d_\omega^2 + Wd_2^2)\sum_{i=0}^{p}\alpha_{t,i}^2|\mathcal{N}_{t,i}| + 2C_f(1-\gamma)\sqrt{M}\sum_{i=1}^{p}\alpha_{t,i}, \\
&+ 3C_{RN}\sum_{i=0}^{p}\sum_{m\in\mathcal{N}_{t,i}}\alpha_{t,i}\|f_{t,i}((s,a);\bar{\omega}_m) - Q^i_{t,\pi_m}(s,a)\|_{\nu_{\pi_m}}
\end{aligned} \tag{77}$$

then, with probability of at least $1-\delta$, $\mathcal{N}_{t,0} \neq \emptyset$, i.e., $\hat{\pi}_t$ is well-defined; also, one the following two statements must hold, *1.* $|\mathcal{N}_{t,0}| \geq M/2$.

2. $\sum_{m\in\mathcal{N}_{t,0}} \big(J_{t,0}(\pi_t^*) - J_{t,0}(\pi_{t,m})\big) \leq 0$.

Under assumption 6, we also have the following holds:

$$|\mathcal{N}_{t,0}| \geq \left(\frac{1}{2} - \kappa\right) M$$

for some $\kappa \in (0, \frac{1}{2})$.

Proof. We will prove by contradiction [93, Lemma 17]. If $\mathcal{N}_{t,0} = \emptyset$, it implies that $\sum_{i=1}^{p} |\mathcal{N}_{t,i}| = M$. Therefore, by Lemma 22,

$$\begin{aligned}\alpha_{t,0}(1-\gamma)M \leq \mathbb{E}_{\nu_t^*}[D(\pi_t^*|\pi_{t,0})] + \frac{8C_{RN}\sqrt{C_0}d_\omega^{1.5}}{\sqrt{d_1}W^{1/4}}\sum_{i=0}^{p}\alpha_{t,i}|\mathcal{N}_{t,i}| + L_f(d_\omega^2 + Wd_2^2)\sum_{i=0}^{p}\alpha_{t,i}^2|\mathcal{N}_{t,i}| \\ + 3C_{RN}\sum_{i=0}^{p}\sum_{m\in\mathcal{N}_{t,i}}\alpha_{t,i}\|f_{t,i}((s,a);\bar{\omega}_m) - Q_{t,\pi_m}^i(s,a)\|_{\nu_{\pi_m}} + 2C_f(1-\gamma)\sum_{i=1}^{p}\alpha_{t,i}\sqrt{M},\end{aligned}$$

which contradicts (77). Thus, $|\mathcal{N}_{t,0}| \neq \emptyset$ must hold.

To proceed, first, if $\sum_{m\in\mathcal{N}_{t,0}} \big(J_{t,0}(\pi_t^*) - J_{t,0}(\pi_{t,m})\big) \leq 0$, then the second item holds. So we consider the case that $\sum_{m\in\mathcal{N}_{t,0}} \big(J_{t,0}(\pi_t^*) - J_{t,0}(\pi_{t,m})\big) > 0$, then Lemma 22 implies

$$\begin{aligned}(1-\gamma)\eta_t\sum_{i=1}^{p}\alpha_{t,i}|\mathcal{N}_{t,i}| \leq \mathbb{E}_{\nu_t^*}[D(\pi_t^*|\pi_{t,0})] + \frac{8C_{RN}\sqrt{C_0}d_\omega^{1.5}}{\sqrt{d_1}W^{1/4}}\sum_{i=0}^{p}\alpha_{t,i}|\mathcal{N}_{t,i}| \\ + 2C_f(1-\gamma)\sum_{i=1}^{p}\alpha_{t,i}\sqrt{M} + L_f(d_\omega^2 + Wd_2^2)\sum_{i=0}^{p}\alpha_{t,i}^2|\mathcal{N}_{t,i}| \\ + 3C_{RN}\sum_{i=0}^{p}\sum_{m\in\mathcal{N}_{t,i}}\alpha_{t,i}\|f_{t,i}((s,a);\bar{\omega}_m) - Q_{t,\pi_m}^i(s,a)\|_{\nu_{\pi_m}}.\end{aligned}$$

If $|\mathcal{N}_{t,0}| < M/2$, then $\sum_{i=1}^{p}|\mathcal{N}_{t,i}| > M/2$, which implies that the following holds:

$$\begin{aligned}\frac{1}{2}\alpha_{t,0}(1-\gamma)\eta_t M \leq \mathbb{E}_{\nu_t^*}[D(\pi_t^*|\pi_{t,0})] + \frac{8C_{RN}\sqrt{C_0}d_\omega^{1.5}}{\sqrt{d_1}W^{1/4}}\sum_{i=0}^{p}\alpha_{t,i}|\mathcal{N}_{t,i}| \\ + 2C_f(1-\gamma)\sum_{i=1}^{p}\alpha_{t,i}\sqrt{M} + L_f(d_\omega^2 + Wd_2^2)\sum_{i=0}^{p}\alpha_{t,i}^2|\mathcal{N}_{t,i}| \\ + 3C_{RN}\sum_{i=0}^{p}\sum_{m\in\mathcal{N}_{t,i}}\alpha_{t,i}\|f_{t,i}((s,a);\bar{\omega}_m) - Q_{t,\pi_m}^i(s,a)\|_{\nu_{\pi_m}},\end{aligned}$$

which again contradicts (77). Therefore, the first item must hold, i.e., $|\mathcal{N}_{t,0}| \geq M/2$.

The last part follows directly from Lemma 16 and is thus omitted. □

Remark 7. *We can relax the assumption in the main paper that* $\alpha_{t,i} \geq \alpha_{t,0}$*, by having* (77) *satisfied for every constraint* $i \in \{1, ..., p\}$*. This would again lead to a convex set of feasible* $\alpha_{t,i}$ *for* $i = 0, ..., p$*, with slightly more convoluted expressions in subsequent analysis.*

Lemma 24. *Let the violation tolerance be chosen as:*

$$\begin{aligned}\eta_t = \frac{2\mathbb{E}_{\nu_t^*}[D(\pi_t^*|\pi_{t,0})]}{\alpha_{t,0}M(1-\gamma)} + \frac{16C_{RN}\sqrt{C_0}d_\omega^{1.5}}{\sqrt{d_1}W^{1/4}M(1-\gamma)}\sum_{i=0}^{p}\frac{\alpha_{t,i}}{\alpha_{t,0}} \\ + \frac{2L_f(d_\omega^2 + Wd_2^2)}{M(1-\gamma)}\sum_{i=0}^{p}\frac{\alpha_{t,i}^2}{\alpha_{t,0}} + \frac{4C_f}{\sqrt{M}}\sum_{i=1}^{p}\frac{\alpha_{t,i}}{\alpha_{t,0}}, \\ + \frac{6C_{RN}}{\alpha_{t,0}M(1-\gamma)}\sum_{i=0}^{p}\sum_{m\in\mathcal{N}_{t,i}}\alpha_{t,i}\|f_{t,i}((s,a);\bar{\omega}_m) - Q_{t,\pi_m}^i(s,a)\|_{\nu_{\pi_m}}. \qquad (78)\end{aligned}$$

Then, the following holds:

$$U_t(\pi_{t,0}, \{\alpha_{t,i}\}_{i=0}^p) = \eta_t, \tag{79}$$

$$U_{t,i}(\pi_{t,0}, \{\alpha_{t,i}\}_{i=0}^p) = \eta_t + \frac{2C_2}{(1-2\kappa)\sqrt{M}} + \frac{4C_f}{(1-2\kappa)(1-\gamma)^{1.5}W^{1/8}} \log^{1/4}\left(\frac{2(1-\gamma)^2 M\sqrt{W}}{\delta}\right) \tag{80}$$

for $i = 1, ..., p$*.*

Proof. By Lemma 22, with probability of at least $1-\delta$, we have

$$\begin{aligned}
&\alpha_{t,0}(1-\gamma)\sum_{m\in\mathcal{N}_{t,0}}\Big(J_{t,0}(\pi_t^*) - J_{t,0}(\pi_{t,m})\Big) + (1-\gamma)\eta_t\sum_{i=1}^p \alpha_{t,i}|\mathcal{N}_{t,i}| \\
&\le \mathbb{E}_{\nu_t^*}[D(\pi_t^*|\pi_{t,0})] + \frac{8C_{RN}\sqrt{C_0}d_\omega^{1.5}}{\sqrt{d_1}W^{1/4}}\sum_{i=0}^p \alpha_{t,i}|\mathcal{N}_{t,i}| + L_f(d_\omega^2 + Wd_2^2)\sum_{i=0}^p \alpha_{t,i}^2|\mathcal{N}_{t,i}| \\
&\quad + 3C_{RN}\sum_{i=0}^p\sum_{m\in\mathcal{N}_{t,i}} \alpha_{t,i}\|f_{t,i}((s,a);\bar{\omega}_m) - Q_{t,\pi_m}^i(s,a)\|_{\nu_{\pi_m}} + (1-\gamma)2C_f\sqrt{M}\sum_{i=1}^p \alpha_{t,i}.
\end{aligned}$$

If $\sum_{m\in\mathcal{N}_{t,0}}\big(J_{t,0}(\pi_t^*) - J_{t,0}(\pi_{t,m})\big) \ge 0$, then from Lemma 22, we have that $|\mathcal{N}_{t,0}| \ge M/2$, which also implies that $\sum_{i=1}^p |\mathcal{N}_{t,i}| < M/2$. Thus,

$$\begin{aligned}
&J_{t,0}(\pi_t^*) - \mathbb{E}[J_{t,0}(\hat{\pi}_t)] = \frac{1}{|\mathcal{N}_{t,0}|}\sum_{m\in\mathcal{N}_{t,0}}\big(J_{t,0}(\pi_t^*) - J_{t,0}(\pi_{t,m})\big) \\
&\le \frac{2\mathbb{E}_{\nu_t^*}[D(\pi_t^*|\pi_{t,0})]}{\alpha_{t,0}M(1-\gamma)} + \frac{16C_{RN}\sqrt{C_0}d_\omega^{1.5}}{\sqrt{d_1}W^{1/4}(1-\gamma)}\sum_{i=0}^p\frac{\alpha_{t,i}}{\alpha_{t,0}} + \frac{2L_f(d_\omega^2 + Wd_2^2)}{1-\gamma}\sum_{i=0}^p\frac{\alpha_{t,i}^2}{\alpha_{t,0}} \\
&\quad + \frac{6C_{RN}}{\alpha_{t,0}M(1-\gamma)}\sum_{i=0}^p\sum_{m\in\mathcal{N}_{t,i}}\alpha_{t,i}\|f_{t,i}((s,a);\bar{\omega}_m) - Q_{t,\pi_m}^i(s,a)\|_{\nu_{\pi_m}} + \frac{4C_f}{\alpha_{t,0}\sqrt{M}}\sum_{i=1}^p\alpha_{t,i},
\end{aligned}$$

which proves the statement on TAOG.

For constraint violation, consider any $i = 1, ..., p$, we have

$$\begin{aligned}
\mathbb{E}[J_{t,i}(\hat{\pi}_t)] - d_{t,i} &= \frac{1}{|\mathcal{N}_{t,0}|}\sum_{m\in\mathcal{N}_{t,0}} J_{t,i}(\pi_{t,m}) - d_{t,i} \\
&\le \frac{1}{|\mathcal{N}_{t,0}|}\sum_{m\in\mathcal{N}_{t,0}}\big(\bar{J}_{t,i}(\pi_{t,m}) - d_{t,i}\big) + \frac{1}{|\mathcal{N}_{t,0}|}\sum_{m\in\mathcal{N}_{t,0}}|\bar{J}_{t,i}(\pi_{t,m}) - J_{t,i}(\pi_{t,m})| \\
&\le \eta_t + \frac{1}{|\mathcal{N}_{t,0}|}\sum_{m=0}^{K-1}|J_{t,i}(\pi_{t,m}) - \bar{J}_{t,i}(\bar{\omega}_m)| \\
&\le \eta_t + \frac{1}{|\mathcal{N}_{t,0}|}\sum_{m=0}^{K-1}|\bar{J}_{t,i}(\bar{\omega}_m) - \mathbb{E}_{\nu_\pi}[f_{t,i}((s,a);\bar{\omega}_m)]| \\
&\quad + \frac{C_{RN}}{|\mathcal{N}_{t,0}|}\sum_{m\in\mathcal{N}_{t,i}}\|f_{t,i}((s,a);\bar{\omega}_m) - Q_{t,\pi_m}^i(s,a)\|_{\nu_{\pi_m}} \\
&\le \eta_t + \frac{2C_2}{(1-2\kappa)\sqrt{M}} + \frac{4C_f}{(1-2\kappa)(1-\gamma)^{1.5}W^{1/8}}\log^{1/4}\left(\frac{2(1-\gamma)^2M\sqrt{W}}{\delta}\right)
\end{aligned} \tag{81}$$

where the second inequality is by the design of the CRPO algorithm, and the fourth inequality is due to [93, Eqs. 67, 68] (with positive constant C_2) and Lemma 23, with the choice of (78), which satisfies (77). Note that $\kappa \in (0, \frac{1}{2})$ by Lemma 23. Combining (81) and (78) yields the result on TACV. □

Finally, we are able to provide the following bounds on TAOG and TACV in the case of adaptive learning rates.

Theorem G.1 (Bounds on TAOG and TACV with function approximation)**.** *Suppose we run CRPO algorithm for M steps per task t with learning rates $\{\alpha_{t,i}\}_{i=0,...,p}$ and function approximation. Then, after T tasks, the TAOG $\bar{R}_0$ is given by*

$$\bar{R}_0 = \frac{1}{T}\sum_{t=1}^{T}\Bigg[\frac{\bar{c}_1^t \mathbb{E}_{\nu_t^*}[D(\pi_t^*|\pi_{t,0})]}{\alpha_{t,0}} + \bar{c}_2^t\sum_{i=0}^{p}\frac{\alpha_{t,i}}{\alpha_{t,0}} + \bar{c}_3^t\sum_{i=0}^{p}\frac{\alpha_{t,i}^2}{\alpha_{t,0}}, \\ + \bar{c}_4^t\sum_{i=0}^{p}\sum_{m\in\mathcal{N}_{t,i}}\frac{\alpha_{t,i}}{\alpha_{t,0}}\|f_{t,i}((s,a);\bar{\omega}_m) - Q_{t,\pi_m}^i(s,a)\|_{\nu_{\pi_m}}\Bigg], \tag{82}$$

and the TACV $\bar{R}_i$ is given by

$$\bar{R}_i = \frac{1}{T}\sum_{t=1}^{T}\Bigg[\frac{\bar{c}_1^t \mathbb{E}_{\nu_t^*}[D(\pi_t^*|\pi_{t,0})]}{\alpha_{t,0}} + \bar{c}_2^t\sum_{i=0}^{p}\frac{\alpha_{t,i}}{\alpha_{t,0}} + \bar{c}_3^t\sum_{i=0}^{p}\frac{\alpha_{t,i}^2}{\alpha_{t,0}}, \\ + \bar{c}_4^t\sum_{i=0}^{p}\sum_{m\in\mathcal{N}_{t,i}}\frac{\alpha_{t,i}}{\alpha_{t,0}}\|f_{t,i}((s,a);\bar{\omega}_m) - Q_{t,\pi_m}^i(s,a)\|_{\nu_{\pi_m}} + \bar{c}_5^t\Bigg], \tag{83}$$

for $i = 1, ..., p$, where $\bar{c}_1^t = \frac{2}{(1-\gamma)M}$, $\bar{c}_2^t = \frac{16C_{RN}\sqrt{C_0}d_\omega^{1.5}}{\sqrt{d_1}W^{1/4}M(1-\gamma)} + \frac{4C_f}{\sqrt{M}}$, $\bar{c}_3^t = \frac{2L_f(d_\omega^2+Wd_2^2)}{M(1-\gamma)}$, $\bar{c}_4^t = \frac{6C_{RN}}{M(1-\gamma)}$, and $\bar{c}_5^t = \frac{2C_2}{(1-2\kappa)\sqrt{M}} + \frac{4C_f}{(1-2\kappa)(1-\gamma)^{1.5}W^{1/8}}\log^{1/4}\left(\frac{2(1-\gamma)^2M\sqrt{W}}{\delta}\right)$.

Proof. The proof follows directly by summing the results (79) and (80) over $t = 1, ..., T$. □

G.2 Meta-initialization of the critic

This section discusses how properly initializing the critic function can help reduce the task-averaged regret. We start with the following notations. Let $\iota_m = \{i|m \in \mathcal{N}_{t,i}; i \in \{0, ..., p\}\}$ be the constraint or objective being updated at time step m. The term in the TAOG bound (82) (similar for the TACV bound (83)) due to the critic regret is:

$$\sum_{m=1}^{M}\frac{6C_{RN}\alpha_{t,\iota_m}}{\alpha_{t,0}(1-\gamma)M}\|f_{t,\iota_m}((s,a);\bar{\omega}_m) - Q_{t,\pi_m}^{\iota_m}(s,a)\|_{\mu_{\pi_m}} \tag{84}$$

We will control the term $\|f_{t,i}((s,a);\bar{\omega}_m) - Q_{t,\pi_m}^i(s,a)\|_{\mu_{\pi_m}}$ in terms of $\|\omega_{t,0}^{i,*} - \omega_{t,0}^{i,0}\|_2$ for $i = 0, ..., p$, i.e., the initial errors of the critic for reward and every constraint. Recall that after each m-th step of policy update, we will perform K iterations of critic TD evaluation. Let $\omega_{t,m}^{i,k}$ denote the critic parameter for reward/constraint i at the k-th iteration after the m-th step of policy update. For any step $1 < m \leq M$, instead of starting from random initialization, the critic can inherit the learned parameter from the previous time step, i.e. $\omega_{t,m}^{i,0} \leftarrow \bar{\omega}_{t,m-1}^i$, where $\bar{\omega}_{t,m-1}^i = \frac{1}{K}\sum_{k=1}^{K}\omega_{t,m-1}^{i,k}$. Let $\omega_{t,m}^{i,*}$ represent the true critic parameter for constraint i corresponding to the policy at the m-th update step. Then, by Lemma 21, with probability of at least $1-\delta$, the critic error $\|f_{t,i}((s,a);\bar{\omega}_{t,m}^i) - Q_{t,\pi_m}^i(s,a)\|_{\mu_{\pi_m}}$ can be controlled by:

$$c_1(K,W)\|\omega_{t,m}^{i,*} - \bar{\omega}_{t,m-1}^i\|_2 + c_2(K,W), \tag{85}$$

where $c_1(K,W) = \sqrt{c_1'(K,W)}$ and $c_2(K,W) = \sqrt{c_2'(K,W)}$ as specified in Lemma 21, and we plugged in $\bar{\omega}_{t,m-1}^i$ as the initial critic parameter $\omega_{t,m}^{i,0}$ for step m. Note that we used $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for $a, b \geq 0$. Observe that the term $c_1(K,W)$ diminishes at the order of $\Theta(K^{-1/4})$ and $\Theta(W^{-1/2})$ and the term $c_2(K,W)$ diminishes at the order of $\Theta(K^{-1/4})$ and $\Theta(W^{-1/8})$. Recall Assumption 4 on the steady-state feature covariance matrix $\Sigma_m = \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\nu_m(s,a)\xi(s,a)\xi(s,a)^\top$, where ν_m is the steady-state distribution of the policy at step m.

Lemma 25. *Under Assumption 4, the following holds with high probability:*

$$\|f_{t,i}((s,a);\bar{\omega}_m) - Q^i_{t,\pi_m}(s,a)\|_{\mu_{\pi_m}} \leq c_e c_1(K,W)^{m+1}\|\omega^{i,*}_{t,0} - \omega^{i,0}_{t,0}\|_2 + \bar{c}_2(K,W).$$

where $\bar{c}_2(K,W) = c_e \sum_{J=0}^{m} c_1(K,W)^{m-j}(c_2(K,L) + L_Q L_c)$*.*

Proof. We will proceed the proof with proper conditioning on the high probability event. By Lemma 21, the critic error can be controlled by

$$\|f_{t,i}((s,a);\bar{\omega}^i_{t,m}) - Q^i_{t,\pi_m}(s,a)\|_{\mu_{\pi_m}} \leq c_1(K,W)\|\omega^{i,*}_{t,m} - \bar{\omega}^i_{t,m-1}\|_2 + c_2(K,W), \tag{86}$$

where $c_1(K,W) = \sqrt{c'_1(K,W)}$ and $c_2(K,W) = \sqrt{c'_2(K,W)}$ as specified in Lemma 21. We can further bound the term on the RHS using the error decomposition as follows:

$$\begin{aligned}\|\bar{\omega}^i_{t,m-1} - \omega^{i,*}_{t,m}\|_2 &\leq \|\bar{\omega}^i_{t,m-1} - \omega^{i,*}_{t,m-1}\|_2 + \|\omega^{i,*}_{t,m-1} - \omega^{i,*}_{t,m}\|_2 \\ &\leq \|\bar{\omega}^i_{t,m-1} - \omega^{i,*}_{t,m-1}\|_2 + L_Q L_c,\end{aligned} \tag{87}$$

where the second inequality is due to [94, Lem. 4] and [100, Thm. 5.3], where the specifications of constants can be found. Intuitively, if two Lipschitz policies differ only by a gradient step, then the difference of their corresponding value functions is also bounded. By assuming that the steady-state feature covariance matrix has minimum eigenvalue $1/c_e^2 > 0$, we can alternatively upper bound $\|\omega^{i,*}_{t,m} - \bar{\omega}^i_{t,m}\|_2$ by $c_e\|f_{t,i}((s,a);\bar{\omega}_m) - Q^i_{t,\pi_m}(s,a)\|_{\mu_{\pi_m}}$ [13, Lem. 1]. Thus, combining (87) and (86) and telescoping from 0 to m, we have

$$\begin{aligned}&\|\omega^{i,*}_{t,m} - \bar{\omega}^i_{t,m-1}\|_2 \\ &\leq c_e c_1(K,W)^m\|\omega^{i,*}_{t,0} - \bar{\omega}^i_{t,0}\|_2 + c_e\sum_{J=0}^{m-1} c_1(K,W)^{m-1-j}(c_2(K,W) + L_Q L_c)\end{aligned}$$

By applying [13, Lem. 1] again, we can bound $\|f_{t,\iota_m}((s,a);\bar{\omega}_m) - Q^{\iota_m}_{t,\pi_m}(s,a)\|_{\mu_{\pi_m}}$ by:

$$c_e c_1(K,W)^m\|\omega^{i,*}_{t,0} - \bar{\omega}^i_{t,0}\|_2 + c_e\sum_{J=0}^{m-1} c_1(K,W)^{m-1-j}(c_2(K,W) + L_Q L_c). \tag{88}$$

By applying (86) for $m = 1$, we prove the result. □

The goal would be to learn a meta critic, such that the regret due to this bound is minimized.

Theorem G.2. *Let each within-task CMDP t run M steps of CRPO, initialized by policy $\pi_{t,0} := \mathrm{INIT}^a(t)$, $\omega^i_{t,0} := \mathrm{INIT}^{c,i}(t)$ for $i = 0,...,p$, and learning rates $\{\alpha_{t,i}\}_{i=0}^p = \mathrm{SIM}(t)$. Let $\{\kappa_i^*\}_{i=0}^p = \arg\min L(\{\kappa_i\}_{i=0}^p)$, where*

$$L(\{\kappa_i\}_{i=0}^p) = \frac{U_T^{init,a}(\{\pi^*_t\}_{t=1}^T)}{\kappa_0} + \sum_{i=0}^{p}\frac{\kappa_i U_T^{c,i}(\{\omega^{i,*}_{t,0}\}_{t=1}^T)}{\kappa_0 M} + \sum_{t=1}^{T}\sum_{i=0}^{p}\frac{\bar{c}_2^t\kappa_i}{\kappa_0}. \tag{89}$$

Then, the following bounds on TAOG and TACV hold:

$$\bar{R}_i \leq U_T^{sim}(\{\kappa_i^*\}_{i=0}^p) + L(\{\kappa_i^*\}_{i=0}^p), \tag{90}$$

for $i = 0,...,p$.

Proof.

$$
\begin{aligned}
&\sum_{t=1}^{T} U_t(\pi_{t,0}, \{\alpha_{t,i}\}_{i=0}^{p}) \leq \sum_{t=1}^{T} \bar{f}_t^{sim}(\{\alpha_{t,i}\}_{i=0}^{p}) \\
&\leq \min_{\{\kappa_i\}_{i=0}^{p}} \quad U_T^{sim}(\{\kappa_i\}_{i=0}^{p}) + \sum_{t=1}^{T} \bar{f}_t^{sim}(\{\kappa_i\}_{i=0}^{p}) \\
&\leq \min_{\{\kappa_i\}_{i=0}^{p}} \quad U_T^{sim}(\{\kappa_i\}_{i=0}^{p}) + \frac{U_T^{init,a}(\{\psi_t\}_{t=1}^{T})}{\kappa_0} + \sum_{t=1}^{T} \frac{\bar{c}_1^t f_t^{init}(\psi_t)}{\kappa_0 M} \\
&\qquad + \sum_{t=1}^{T} \left[\bar{c}_2^t \sum_{i=0}^{p} \frac{\kappa_i}{\kappa_0} + \sum_{i=0}^{p} \sum_{m \in \mathcal{N}_{t,i}} \frac{\kappa_i \bar{c}_3^t \bar{c}_1 (K,L)^{m+1}}{\kappa_0 M} \|\omega_{t,0}^{i,*} - \omega_{t,0}^{i}\|_2 \right] \\
&\leq \min_{\{\kappa_i\}_{i=0}^{p}} \quad U_T^{sim}(\{\kappa_i\}_{i=0}^{p}) + \frac{U_T^{init,a}(\{\psi_t\}_{t=1}^{T})}{\kappa_0} + \sum_{i=0}^{p} \frac{\kappa_i U_T^{c,i}(\{\omega_t^i\}_{t=1}^{T})}{\kappa_0 M} + \sum_{t=1}^{T} \frac{\bar{c}_1^t f_t^{init}(\psi_t)}{\kappa_0 M} \\
&\qquad + \sum_{t=1}^{T} \left[\bar{c}_2^t \sum_{i=0}^{p} \frac{\kappa_i}{\kappa_0} + \sum_{i=0}^{p} \sum_{m \in \mathcal{N}_{t,i}} \frac{\kappa_i \bar{c}_3^t \bar{c}_1 (K,L)^{m+1}}{\kappa_0 M} \|\omega_{t,0}^{i,*} - \omega_{t}^{i}\|_2 \right]
\end{aligned}
\tag{91}
$$

By choosing $\psi_t = \pi_t^*$, $\omega_t^i = \omega_{t,0}^{i,*}$ as the comparator sequences, we have:

$$
\sum_{t=1}^{T} U_t(\pi_{t,0}, \{\alpha_{t,i}\}_{i=0}^{p}) \tag{92}
$$

$$
\leq \min_{\{\kappa_i\}_{i=0}^{p}} U_T^{sim}(\{\kappa_i\}_{i=0}^{p}) + \underbrace{\frac{U_T^{init,a}(\{\pi_t^*\}_{t=1}^{T})}{\kappa_0} + \sum_{i=0}^{p} \frac{\kappa_i U_T^{c,i}(\{\omega_{t,0}^{i,*}\}_{t=1}^{T})}{\kappa_0 M} + \sum_{t=1}^{T} \sum_{i=0}^{p} \frac{\bar{c}_2^t \kappa_i}{\kappa_0}}_{L(\{\kappa_i\}_{i=0}^{p})} \tag{93}
$$

Plug in $\{\kappa_i^*\}_{i=0}^{p} = \arg\min L(\{\kappa_i\}_{i=0}^{p})$ then proves the claim. □

H Experimental setting

In this section, we describe our experimental setup for the OpenAI gym experiments under constrained settings. We present the performance of meta-SRL on the Acrobot and the Frozen lake under low and high task-similarity conditions.

H.1 Acrobot

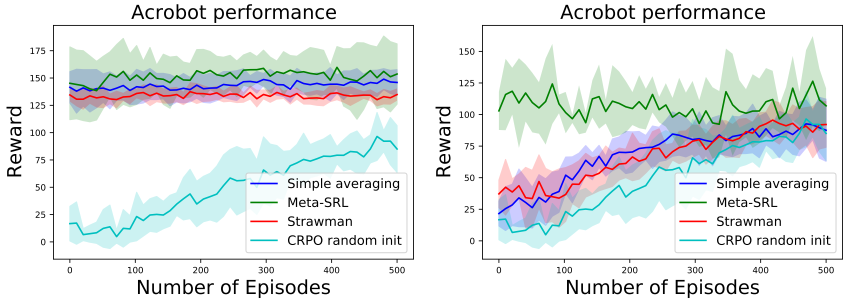
Acrobot is a 2 link robot OpenAI gym environment [18]. It is a continuous state environment, where the agent gets rewarded $+1$ when the end-effector is swung at a specific height. Two constraints are introduced in the problem. The first constraint penalizes the agent with -1 cost if the first link swings in the prohibited direction. The second constraint penalizes the agent with -1 cost if the second link swings in some prohibited direction. We randomly generate $T = 10$ different tasks with a probability distribution which governs the task-similarity of the tasks. We compare the meta-SRL with simple averaging (i.e., initialize with the average of learned policies from past CMDPs), Strawman (i.e., initialize with the learned policy from the latest CMDP), and random initialization strategies as done in CRPO. We do 10 runs on each baseline to get the performance plots with variance.

Following case is considered with both low and high task-similarity.

Task similarity based on height required to achieve reward: In this case, we consider tasks where acrobot needs to achieve different heights to acquire rewards. We denote the height required to get the reward by r_h. For high task-similarity, we create a latent CMDP normal distribution with mean $r_h = 0.5$ (unit: meter) and a 5% standard deviation around the mean. For low task-similarity, we consider two probability distributions with mean $r_h = 0.6$ and $r_h = 0.4$, with a 10% standard deviation around these means. The maximum threshold for costs $d_{t,i}$ for each task is kept constant at the value of 50.

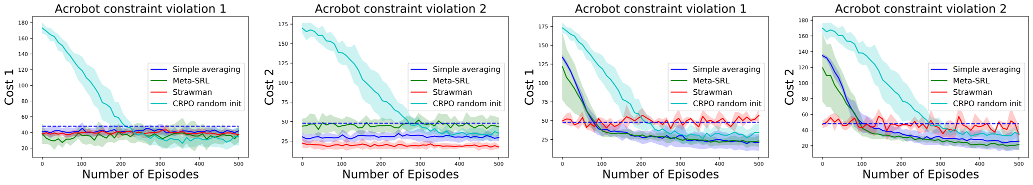
Meta-SRL setup: We choose 10 episodes in each iteration before a policy update, and run 50 update steps of CRPO algorithm in each task.We introduce meta-policy initialization by incorporating stationary distribution correction. We initialize the policy for a new unseen task by taking *weighted* sum of policies by their stationary distributions. We obtain these stationary distributions by running the suboptimal policies obtained from previous tasks for 500 more time steps. For the initialization of the meta-critic parameters, we initialize with the simple average of the parameters of the value functions for cost and reward obtained from the previous training tasks.

Meta-SRL performance on Acrobot: For the acrobot experiments, we see almost similar trend for Meta-SRL as seen in the frozen lake experiments. We consider the tasks where variability is introduced in terms of the height required to achieve the reward as discussed before. We can observe from Figure 3 (a) and (b), that in the case of both low and high task-similarity, the meta-SRL was able to achieve high reward just from the start of the training. We can also observe from Figure 4 (a) and (b), that in the case of high task-similarity, all baselines perform well in ensuring the constraint satisfaction from the start of the training. However, in the case of lower task-similarity, Meta-SRL performs reasonably well on constraint satisfactions compared to other baselines as observed in Figure 4 (c) and (d). We see the benefit of critic meta-initialization and putting higher weights on policies that frequently visit a particular state, since it implies that the corresponding strategies can have substantial impact on rewards and constraint violations. Overall, meta-SRL was able to generalize well over tasks with reasonable task-similarity.



(a) (b)

Figure 3: Acrobot results for reward maximization when the task-similarity (in terms of height required) is high (left plot) or low (right plot) for CMDP tasks.



(a) (b) (c) (d)

Figure 4: Acrobot results for constraint violationsfor different costs when the task-similarity (in terms of height required) is high [*(a), (b)*] or low [*(c), (d)*] for the CMDP tasks. Black dashed line represents the average maximum threshold limit for the constraints.

H.2 Frozen lake experimental setup

For the Frozen lake, we randomly generate $T = 20$ different orientations as tasks over the probability of a state being frozen or a hole, and evaluate the performance for the scenarios with high task-similarity (low variance for the latent CMDP distribution) or low task-similarity (high variance for the latent CMDP distribution). The agent gets rewarded $+1$ when it reaches the goal state, and incurs a cost -1 when it falls into a hole. We choose the constraint threshold $d_{t,i} = 4$. We run CRPO for 100 episodes on each task. The learning rate β for TD learning is taken as 0.5.

High task-similarity setup: In this case, random tasks are generated where the probability of a tile being frozen is kept to be between 0.7 and 0.8. The tasks are similar due to less uncertainty associated with the orientations changing (i.e., high probability of frozen tiles).

Low task-similarity setup: In this case, random tasks are generated where the probability of a tile being frozen is kept to be between 0.45 and 0.55. The tasks are less similar due to high uncertainty associated with the changing orientations.