Offline Constrained Reinforcement Learning under Partial Data Coverage

Seokmin Ko
KAIST
komin0407@kaist.ac.kr

Ambuj Tewari
University of Michigan
tewaria@umich.edu

Kihyuk Hong
KAIST
kihyukh@kaist.ac.kr

Abstract

We study offline constrained reinforcement learning with general function approximation in discounted constrained Markov decision processes. Prior methods either require full data coverage to evaluate intermediate policies, lack oracle efficiency, or require knowledge of the data-generating distribution to extract a policy from a learned density ratio. We propose PDOCRL, an oracle-efficient primal-dual algorithm based on a Lagrangian decomposition of the occupancy-measure linear program. We use a density ratio to estimate the Lagrangian from offline samples and reparameterize a copy of the occupancy measure to make the policy an explicit optimization variable. PDOCRL therefore optimizes the policy directly, without a separate extraction step or knowledge of the data-generating distribution. Exploiting the linear structure of the decomposed Lagrangian, PDOCRL uses only standard oracles for policy optimization, online linear optimization, and linear minimization. We assume all-policy action-value realizability and partial coverage of an optimal feasible policy with a realizable density ratio. Under Slater's condition and standard boundedness assumptions, PDOCRL returns a near-optimal, near-feasible policy with a $\widetilde{\mathcal{O}}(\epsilon^{-2})$ sample guarantee. Experiments on BulletGym tasks using DSRL datasets show that PDOCRL is competitive with strong offline constrained RL baselines.

1 Introduction

Offline constrained reinforcement learning aims to learn, from a fixed dataset, a policy that achieves high reward while satisfying safety or resource constraints. We study this problem in the framework of discounted constrained Markov decision processes (CMDPs)[1], with one primary reward and several auxiliary signals whose discounted returns must exceed prescribed thresholds. This formulation is natural when online exploration is risky or expensive but logged data are available, as in autonomous driving, healthcare, and robotics [11, 16].

The main obstacle in offline CMDPs is data coverage: reliable optimization is only possible for policies whose state-action occupancy is sufficiently supported by the dataset. In unconstrained offline RL, recent pessimistic actor-critic methods show that one can still obtain meaningful guarantees under insufficient coverage by updating policies against pessimistic critics [3, 46].

In the constrained setting, however, applying the pessimistic critic idea within a primal-dual framework introduces an additional obstruction. A primal-dual method must update dual multipliers based on estimates of constraint violations, which requires evaluating the constraint values of the current policy. As discussed by Hong et al. [10], this entails accurately estimating constraint values for *intermediate policies* encountered during optimization. Under partial data coverage, these intermediate policies may not be supported by the dataset, making the dual updates unreliable and effectively reintroducing strong coverage requirements.

This raises the following question:

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Table 1: Algorithms for offline constrained RL. Red entries indicate an undesirable property

Algorithm	Partial coverage	Oracle efficient	Requires μ_D or π_{ref}	Function Approximation
PDCA [10]	No	No	No	$\forall\pi : Q^\pi \in \mathcal{Q}, w^\pi \in \mathcal{W}$
WSAC [34]	Yes	No	Yes	$w^{\pi_{\text{ref}}} \in \mathcal{W}, \forall\pi : Q^\pi \in \mathcal{Q}$
MBCL [13]	No	Yes	No	$\forall\pi, Q : \mathcal{T}^\pi Q \in \mathcal{Q}$
POCC [43]	Yes	Yes	Yes	$w^* \in \mathcal{W}, \forall w \in \mathcal{W} : x_w \in \mathcal{X}$
PDOCRL (Ours)	Yes	Yes	No	$w^* \in \mathcal{W}, \forall\pi : Q^\pi \in \mathcal{Q}$

Can we design an offline CMDP algorithm that avoids evaluating unsupported intermediate policies, while remaining provably sample efficient under partial data coverage and computationally efficient with access to standard oracles?

Here and throughout, we use *oracle efficient* in the sense that the algorithm performs only a polynomial number of elementary operations and calls to basic optimization oracles for the function classes. Specifically, no-regret online linear optimization, no-regret policy optimization, and linear minimization over given function classes, rather than requiring a generic, intractable optimization such as general min-max optimization over these classes.

In this paper, we answer this question affirmatively, but not via an actor-critic method. Instead, we adopt the linear-programming (LP) view of CMDPs, which optimizes over occupancy measures or density ratios rather than repeatedly evaluating unsupported intermediate policies. This viewpoint is well matched to partial coverage, but under general function approximation it exposes two distinct challenges. First, restricted saddle-point formulations can admit *spurious saddle points*: even when the function classes contain an optimal feasible policy, its occupancy measure, and the optimal action-value functions for the relevant Lagrangian reward combinations, a saddle point need not correspond to an optimal policy. Second, existing LP-based methods typically recover a policy from a learned occupancy measure or density ratio using the data distribution μ_D [14, 15, 31, 42, 43], which is problematic in practice.

We address these two issues separately. For the first, we show that a naive restricted saddle formulation is indeed insufficient: spurious saddle points can persist even when the class realizes the optimal solution and the optimal action-value functions for all Lagrangian reward combinations. We then identify a condition under which this pathology disappears.

The policy-extraction issue is handled separately by a decomposed LP formulation and a reparameterization that make the policy an explicit optimization variable. This lets us optimize policies directly, rather than first learning an occupancy measure and then extracting a policy using the unknown data distribution μ_D. The resulting method, PDOCRL, is an oracle-efficient primal-dual algorithm for the decomposed empirical saddle problem. Under our realizability and coverage assumptions, a near saddle point found by PDOCRL yields a near-optimal and near-feasible policy with a $\widetilde{\mathcal{O}}(\epsilon^{-2})$ sample complexity guarantee under partial coverage. As summarized in Table 1, this separates PDOCRL from prior offline constrained RL methods: PDCA [10] and WSAC [34] require solving intractable min-max problems, while POCC [43] requires μ_D and access to another auxiliary function class. Thus PDOCRL is the only method in our comparison that simultaneously provides partial-coverage guarantees, oracle efficiency, and no dependence on μ_D or a reference policy.

We also evaluate PDOCRL on deep offline safe RL benchmarks. On BulletGym tasks with DSRL datasets [7, 19], it is competitive with strong baselines such as BCQ-Lag, BEAR-Lag, CPQ, CDT, and COptiDICE [15, 18, 37].

1.1 Related work

Offline constrained RL with function approximation Several recent works study offline constrained RL with function approximation through policy-value, actor-critic, or primal-dual formulations [10, 13, 34]. Batch constrained policy learning and primal-dual-critic methods require estimating the reward and constraint values of intermediate policies, which typically leads to full-coverage or similarly strong coverage assumptions [10, 13]. The weighted actor-critic method of Wei et al. [34]

handles partial coverage by using importance-weighted critics, but its formulation depends on a reference policy or the data distribution and is not oracle efficient.

Density-ratio and DICE methods A closely related line of work uses marginalized importance sampling to rewrite target-policy or occupancy-measure objectives as expectations under the offline data distribution. This idea appears in off-policy evaluation and DICE-style distribution-correction methods [17, 24, 25, 33, 36, 39, 44, 45]. It has also been adapted to offline policy optimization through methods such as AlgaeDICE and OptiDICE [14, 26]. These methods motivate our density-ratio parameterization $\mu = w\mu_D$, but they do not by themselves resolve the constrained offline-RL issues of oracle-efficient optimization and policy recovery without access to the unknown data distribution.

LP-based offline RL under partial coverage On the theory side, there are LP- and density-ratio-based analyses for offline RL under partial coverage [2, 29, 31, 42, 46]. These methods are naturally aligned with partial coverage because they reason about covered occupancies rather than requiring accurate evaluation of arbitrary intermediate policies. However, relative to PDOCRL, prior LP/density-ratio analyses typically rely on regularization of the saddle-point problem, auxiliary function classes or occupancy-validity constraints, or a separate policy-extraction step that depends on the unknown data distribution μ_D or on a reference policy.

Constrained LP-based offline RL In the constrained setting, COptiDICE is the closest empirical DICE-style method, extending stationary-distribution correction to offline constrained RL [15]. The closest LP-based partial-coverage theory result is the primal-dual method of Zhang et al. [43], which handles offline convex CMDPs under partial coverage. In contrast to PDOCRL, these constrained LP/density-ratio approaches either use an additional policy-extraction or weighted behavior-cloning step, or require auxiliary structure beyond the density-ratio and critic classes considered here.

2 Preliminaries

Notations For functions $F : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ and $G : \mathcal{S} \to \mathbb{R}$, we denote their vector representations as $\boldsymbol{F} \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|}$ and $\boldsymbol{G} \in \mathbb{R}^{|\mathcal{S}|}$, respectively. For a function $Q : \mathcal{S}\times\mathcal{A} \to \mathbb{R}$ and a policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$, we write $Q(s,\pi) = \sum_a \pi(a \mid s)Q(s,a)$. Given reward functions $r_0, r_1, \ldots, r_I$, and $\boldsymbol{\lambda} \in \mathbb{R}^I$, we write $r_{\boldsymbol{\lambda}} = r_0 + \sum_{i=1}^I \lambda_i r_i$. We write $B\Delta^I = \{\lambda \in \mathbb{R}^I_+ : \|\lambda\|_1 \le B\}$. We say $(\widehat{x}, \widehat{y}) \in \mathcal{X} \times \mathcal{Y}$ is a ξ-near saddle point of $L(x,y)$ if $L(x,\widehat{y}) \le L(\widehat{x},y) + \xi$ for all $(x,y) \in \mathcal{X}\times\mathcal{Y}$.

2.1 Constrained reinforcement learning

We formulate the constrained RL as a discounted CMDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r_0, \{r_i\}_{i=1}^I, \gamma, d_0)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $P(\cdot \mid s,a)$ is the transition kernel, $\gamma \in (0,1)$ is the discount factor, d_0 is the initial-state distribution, $r_0 : \mathcal{S}\times\mathcal{A} \to [0,1]$ is the primary reward, and $r_1, \ldots, r_I : \mathcal{S}\times\mathcal{A} \to [0,1]$ are auxiliary reward signals defining the constraints. For simplicity, we assume that d_0 is concentrated on a single start state s_0.

A stationary policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ and the transition kernel P induce a trajectory distribution P^π over $(s_0, a_0, s_1, a_1, \ldots)$, with expectation denoted by $\mathbb{E}^\pi$. Action-value functions are defined as

$$Q_i^\pi(s,a) \coloneqq \mathbb{E}^\pi[\textstyle\sum_{t=0}^\infty \gamma^t r_i(s_t,a_t) \mid s_0 = s, a_0 = a], \qquad i = 0,1,\ldots,I.$$

Define the discounted return for each reward signal by

$$J_i(\pi) \coloneqq \mathbb{E}^\pi[\textstyle\sum_{t=0}^\infty \gamma^t r_i(s_t,a_t)], \qquad i = 0,1,\ldots,I.$$

The CMDP objective, given user-specified thresholds $\tau_1, \ldots, \tau_I$, is

$$\max_\pi \; J_0(\pi) \qquad \text{subject to} \qquad (1-\gamma)J_i(\pi) \ge \tau_i, \quad i = 1,\ldots,I. \tag{1}$$

Occupancy measures and LP formulation The linear programming (LP) formulation [20] for solving (1) aims to find an *occupancy measure* that maximizes value, rather than optimizing directly over the policy space. The occupancy measure of a policy π, denoted $\mu^\pi \in \Delta(\mathcal{S}\times\mathcal{A})$, is defined as:

$$\mu^\pi(s,a) \coloneqq (1-\gamma)\textstyle\sum_{t=0}^\infty \gamma^t P^\pi(s_t = s, a_t = a).$$

It satisfies $\langle \boldsymbol{\mu}^\pi, \boldsymbol{r}_i \rangle = (1-\gamma) J_i(\pi)$, $i = 0, 1, \ldots, I$. Let $\boldsymbol{E}$ denote the linear operator defined by $[\boldsymbol{E}^\top \boldsymbol{\mu}](s) = \sum_a \mu(s,a)$, and let $[\boldsymbol{P}^\top \boldsymbol{\mu}](s') = \sum_{s,a} P(s' \mid s,a)\mu(s,a)$. Then (1) is equivalent to the linear program [30]

$$\begin{aligned} \max_{\boldsymbol{\mu} \geq 0} \quad & \langle \boldsymbol{\mu}, \boldsymbol{r}_0 \rangle \\ \text{subject to} \quad & \boldsymbol{E}^\top \boldsymbol{\mu} = (1-\gamma)\boldsymbol{d}_0 + \gamma \boldsymbol{P}^\top \boldsymbol{\mu}, \\ & \langle \boldsymbol{\mu}, \boldsymbol{r}_i \rangle \geq \tau_i, \qquad i = 1, \ldots, I. \end{aligned} \tag{2}$$

The first constraint, known as the Bellman flow constraint, ensures $\boldsymbol{\mu}$ corresponds to the occupancy measure of some policy. We call such a $\boldsymbol{\mu}$ an *admissible* occupancy measure.

Policy extraction Given an occupancy measure μ, not necessarily admissible, we can extract a policy denoted $\pi(\mu)$ by normalizing μ at each state:

$$[\pi(\mu)](a|s) \coloneqq \tfrac{\mu(s,a)}{\sum_{a'} \mu(s,a')} \text{ if } \textstyle\sum_{a'} \mu(s,a') > 0, \quad \tfrac{1}{|\mathcal{A}|} \text{ otherwise.} \tag{3}$$

Given an admissible occupancy measure μ, we can recover the policy that induces μ by the policy extraction rule. Hence, we can find an optimal policy by extracting from the optimal occupancy measure μ^* that solves the linear program (2).

2.2 Offline dataset and concentrability coefficient

We observe an offline dataset $\mathcal{D} = \{(s_j, a_j, s'_j)\}_{j=1}^n$, where (s_j, a_j) are sampled i.i.d. from an unknown data distribution μ_D and $s'_j \sim P(\cdot \mid s_j, a_j)$. We define the concentrability coefficient of a policy π with respect to μ_D as

$$C^\pi \coloneqq \|\mu^\pi / \mu_D\|_\infty = \max_{s,a} \mu^\pi(s,a)/\mu_D(s,a),$$

which measures how well the dataset covers the state-action pairs visited by the policy π. Intuitively, if C^π is large, then the dataset does not cover the state-action pairs visited by π, making it difficult to learn about π from the dataset.

3 Main results for offline constrained RL

We start from the LP formulation of the offline CMDP in (2). Introducing a Lagrange multipliers $\boldsymbol{V} \in \mathbb{R}^{|\mathcal{S}|}$ for the Bellman flow constraints and $\boldsymbol{\lambda} \in \mathbb{R}_+^I$ for the safety constraints gives the Lagrangian

$$L(\boldsymbol{\mu}; \boldsymbol{V}, \boldsymbol{\lambda}) \coloneqq \langle \boldsymbol{\mu}, \boldsymbol{r} \rangle + \langle \boldsymbol{V}, (1-\gamma)\boldsymbol{d}_0 + \gamma \boldsymbol{P}^\top \boldsymbol{\mu} - \boldsymbol{E}^\top \boldsymbol{\mu} \rangle + \textstyle\sum_{i=1}^I \lambda_i (\langle \boldsymbol{\mu}, \boldsymbol{r}_i \rangle - \tau_i), \tag{4}$$

which is linear in the occupancy measure μ. Writing $\mu(s,a) = w(s,a)\mu_D(s,a)$ turns the μ-dependent term into an expectation under the data distribution μ_D, allowing it to be estimated directly from the offline dataset by sample averaging:

$$\widehat{L}(\boldsymbol{w}; \boldsymbol{V}, \boldsymbol{\lambda}) \coloneqq (1-\gamma)V(s_0) + \tfrac{1}{n} \textstyle\sum_{j=1}^n w(s_j, a_j) \left(r_{\boldsymbol{\lambda}}(s_j, a_j) + \gamma V(s'_j) - V(s_j) \right) - \sum_{i=1}^I \lambda_i \tau_i, \tag{5}$$

where $\boldsymbol{r}_{\boldsymbol{\lambda}} = \boldsymbol{r}_0 + \sum_{i=1}^I \lambda_i \boldsymbol{r}_i$ is the Lagrangian reward. If μ_D were known, a natural approach would be to optimize over the density ratio w, form the induced occupancy measure $\mu = w\mu_D$, and then recover a policy by the statewise normalization in using (3). Appendix B shows that this approach already admits provable guarantees. In practice, however, μ_D may be unknown, and policy extraction from w becomes nontrivial.

Existing empirical methods [14, 15] address this issue by adding an importance-weighted behavior-cloning step motivated by the identity $\pi^* = \operatorname{argmax}_\pi \mathbb{E}_{(s,a)\sim\mu^D}[w^{\pi^*}(s,a) \log \pi(a|s)]$:

$$\widehat{\pi} = \operatorname{argmax}_\pi \tfrac{1}{n} \textstyle\sum_{j=1}^n \widehat{w}(s_j, a_j) \log \pi(a_j|s_j), \tag{6}$$

where $\widehat{w}$ is the learned density ratio. However, this approach suffers from a fundamental statistical limitation: empirical policy extraction from density ratios can be unreliable under finite samples, even when the underlying density ratio is well-behaved. The following proposition formalizes this failure mode. See Appendix A for a proof.

Proposition 1. *There exist a MDP $(\mathcal{S}, \mathcal{A}, P, r, \gamma, d_0)$ and a policy class Π that contains an optimal policy with $|\mathcal{S}| = |\mathcal{A}| = |\Pi| = 2$, data distribution μ_D, and a density-ratio estimate $\widehat{w}$ such that:*

(i) Policy $\widehat{\pi} = \pi(\widehat{w}\,\mu_D)$ extracted from $\widehat{w}$ with the knowledge of μ_D is optimal.
(ii) The policy extraction $\widehat{\pi} = \operatorname{argmax}_{\pi\in\Pi} \mathbb{E}_{\mu_D}[\widehat{w}(s,a)\log\pi(a|s)]$ is optimal.
(iii) But the empirical policy extraction (6) returns a suboptimal policy with constant probability.

This negative result suggests that policy extraction from density ratios is inherently fragile. It motivates designing algorithms that optimize over policies directly, without requiring a separate policy-extraction step. In the next section, we introduce a Lagrangian decomposition and reparameterization that make the policy an explicit optimization variable, thereby avoiding the policy-extraction step.

3.1 Lagrangian decomposition and reparameterization

The key difficulty is that the occupancy measure variable μ in the Lagrangian formulation (4) plays two roles: Lagrangian estimation and policy extraction. For Lagrangian estimation, we use the change-of-measure $\mu = w\mu_D$ which allows the Lagrangian to be expressed as an expectation under the data distribution μ_D. After the change of variables, however, policy extraction has to be done on $w\mu_D(s,a)$, and hence requires knowledge of the unknown data distribution μ_D.

Lagrangian decomposition addresses this issue by introducing a copy variable ν of μ and separating these two roles: μ is used for Lagrangian estimation through the change of measure, while ν is used for policy extraction. The decomposition is obtained by adding a constraint $\nu = \mu$ to the original LP formulation (2) and decomposing the Bellman flow constraint:

$$\begin{aligned} \max_{\boldsymbol{\mu}\geq 0,\ \boldsymbol{\nu}\geq 0} \quad & \langle \boldsymbol{\mu}, \boldsymbol{r}_0 \rangle \\ \text{subject to} \quad & \boldsymbol{E}^\top \boldsymbol{\nu} = (1-\gamma)\boldsymbol{d}_0 + \gamma \boldsymbol{P}^\top \boldsymbol{\mu}, \\ & \boldsymbol{\mu} = \boldsymbol{\nu}, \\ & \langle \boldsymbol{\mu}, \boldsymbol{r}_i \rangle \geq \tau_i, \qquad i = 1, \ldots, I. \end{aligned} \tag{7}$$

Introducing the multiplier $\boldsymbol{Q} \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|}$ for the constraint $\boldsymbol{\mu} = \boldsymbol{\nu}$, we get the Lagrangian function

$$L(\boldsymbol{\mu}, \boldsymbol{\nu}; \boldsymbol{V}, \boldsymbol{Q}, \boldsymbol{\lambda}) = (1-\gamma)\langle \boldsymbol{d}_0, \boldsymbol{V} \rangle + \langle \boldsymbol{\mu}, \boldsymbol{r}_{\boldsymbol{\lambda}} + \gamma \boldsymbol{P}\boldsymbol{V} - \boldsymbol{Q} \rangle + \langle \boldsymbol{\nu}, \boldsymbol{Q} - \boldsymbol{E}\boldsymbol{V} \rangle - \textstyle\sum_{i=1}^{I} \lambda_i \tau_i. \tag{8}$$

Lagrangian decomposition [8, 32] is a classical method for decomposing an optimization problem into subproblems by introducing copies of optimization variables. It is first used in the context of MDP by Mehta et al. [21]. It has since been used in RL in several settings [5, 10, 27, 28].

One could now change variables by writing $w = \mu/\mu_D$, estimate the decomposed Lagrangian as $\widehat{L}(w, \nu; V, Q, \lambda)$, find a saddle point, and then extract a policy from ν. But a simple reparameterization trick allows us to simplify the optimization problem by introducing a policy variable. Specifically, consider a policy π induced by ν. Then, $\nu(s,a) = \pi(a|s)\nu(s)$ and the decomposed Bellman flow constraint suggests the following reparameterization of ν in terms of μ and π:

$$\nu_{\mu,\pi}(s,a) \coloneqq \pi(a \mid s)\big((1-\gamma)d_0(s) + \gamma[\boldsymbol{P}^\top \boldsymbol{\mu}](s)\big).$$

By construction, $\nu = \nu_{\mu,\pi}$ satisfies the decomposed Bellman flow constraint for every μ and π, and the policy extracted from $\nu_{\mu,\pi}$ is exactly π. Substituting $\nu = \nu_{\mu,\pi}$ into the Lagrangian and using the notation $Q_\pi(s) = \sum_{a'} \pi(a'|s)Q(s,a')$ yield

$$L(\boldsymbol{\mu}, \pi; \boldsymbol{Q}, \boldsymbol{\lambda}) = (1-\gamma)\langle \boldsymbol{Q}_\pi, \boldsymbol{d}_0 \rangle + \langle \boldsymbol{\mu}, \boldsymbol{r}_{\boldsymbol{\lambda}} + \gamma \boldsymbol{P}\boldsymbol{Q}_\pi - \boldsymbol{Q} \rangle - \textstyle\sum_{i=1}^{I} \lambda_i \tau_i, \tag{9}$$

The reparameterized decomposed Lagrangian was first introduced by Gabbianelli et al. [5] for offline RL under linear MDPs. In that setting, the linear structure further allows μ and Q to be represented by linear functions of features, reducing the optimization to one over feature weights. In the general function approximation setting, however, the restricted saddle-point problem over general function classes for μ and Q can admit spurious saddle points that do not correspond to optimal policies, as discussed next, requiring additional assumptions to rule out this pathology.

3.2 Spurious saddle points and assumptions for saddle-point optimality

The next proposition shows that merely realizing an optimal policy, its occupancy measure, and its associated action-value functions is not sufficient to rule this out.

Proposition 2. *There exists a CMDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r_0, \{r_i\}_{i=1}^I, \gamma, d_0)$, function classes $\mathcal{U} \subseteq \mathbb{R}_+^{\mathcal{S}\times\mathcal{A}}$, $\mathcal{Q} \subseteq \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$, and a policy class Π such that:*

(i) there exists an optimal feasible policy $\pi^* \in \Pi$ with occupancy measure $\mu^* \in \mathcal{U}$;
(ii) for all $\alpha \in \mathbb{R}_+^I$, $Q_\alpha^* \in \mathcal{Q}$, where Q_α^* is the optimal value function for $r_0 + \sum_{i=1}^I \alpha_i r_i$;
(iii) nevertheless, there exists $\widehat{\pi} \in \Pi$ that is not optimal for the CMDP such that $(\widehat{\mu}, \widehat{\pi}; \widehat{Q}, \widehat{\boldsymbol{\lambda}})$ is a saddle point of the reparameterized Lagrangian L over $(\mathcal{U} \times \Pi) \times (\mathcal{Q} \times \mathbb{R}_+^I)$.

To deal with the problem of spurious saddle points in the decomposed formulation, we first impose the following Slater-type condition to ensure the existence of a feasible policy that satisfies the constraints with a positive margin:

Assumption A (Slater's condition)**.** There exist a constant $\varphi > 0$ and a policy π that satisfy $(1-\gamma)J_i(\pi) \geq \tau_i + \varphi$ for all $i = 1, \ldots, I$.

Under a mild assumption that a feasible policy exists, the thresholds τ_i can be adjusted downward by φ to ensure Slater's condition holds with margin φ, if necessary. This Slater-type condition is standard in constrained RL analyses, where it is commonly used to obtain strong duality and to bound the optimal dual variable [10, 22, 40]. The following lemma demonstrates that, under this assumption, an optimal dual variable $\boldsymbol{\lambda}^*$ is bounded.

Lemma 3. *Consider a constrained optimization problem* (1) *with thresholds $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_I)$. Suppose the problem satisfies Assumption A with margin $\varphi > 0$. Then, the optimal dual variable $\boldsymbol{\lambda}^*$ of the problem satisfies $\|\boldsymbol{\lambda}^*\|_1 \leq \frac{1}{\varphi}$.*

Now, to deal with spurious saddle points, we impose the following all-policy value function realizability assumption.

Assumption B (All-Policy Value Function Realizability)**.** For every policy $\pi \in \Pi$, we have $Q_0^\pi \in \mathcal{Q}$ and $Q_0^\pi + (1 + \frac{1}{\varphi})Q_i^\pi \in \mathcal{Q}$ for all $i = 0, 1, \ldots, I$.

This is a strong all-policy realizability condition, but it is standard in iterative offline actor-critic analyses with general function approximation, where the critic must evaluate the policies generated by the actor [3, 34, 46]. Single-policy realizability is possible in methods that solve saddle/minimax problems such as PRO-RL [42] and CORAL [31], but these methods are not oracle-efficient in our sense. With the stronger all-policy condition the lemma below shows spurious saddle points vanish. We leave the question of whether weaker assumptions for iterative methods can rule out spurious saddle points to future work.

Lemma 4. *Assume Assumption A and B. Consider function classes $\mathcal{U} \subseteq \mathbb{R}_+^{|\mathcal{S}\times\mathcal{A}|}$ and $\mathcal{Q} \subseteq \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|}$. Let Π be a policy class containing an optimal feasible policy π^*, and let $\mu^* = \mu^{\pi^*} \in \mathcal{U}$. If $(\widehat{\mu}, \widehat{\pi}; \widehat{Q}, \widehat{\lambda})$ is a saddle point of L over $(\mathcal{U} \times \Pi) \times (\mathcal{Q} \times (1 + \frac{1}{\varphi})\Delta^I)$, then $\widehat{\pi}$ is an optimal feasible policy.*

3.3 Lagrangian estimation and minimax algorithm

As discussed previously, we can change variable $w = \mu/\mu_D$ to turn the μ-dependent term in the Lagrangian (9) into an expectation under μ_D, which can be estimated by sample averaging from the offline dataset as:

$$\widehat{L}(\boldsymbol{w}, \pi; \boldsymbol{Q}, \boldsymbol{\lambda}) \coloneqq (1-\gamma)Q(s_0, \pi) + \frac{1}{n}\sum_{j=1}^n w(s_j, a_j)\Big(r_{\boldsymbol{\lambda}}(s_j, a_j) + \gamma Q(s_j', \pi) - Q(s_j, a_j)\Big) - \sum_{i=1}^I \lambda_i \tau_i.$$

Motivated by this estimator, we first consider the following idealized empirical saddle-point problem:

$$\max_{w\in\mathcal{W}, \pi\in\Pi} \min_{Q\in\mathcal{Q}, \lambda\in(1+\frac{1}{\varphi})\Delta^I} \widehat{L}(w, \pi; Q, \lambda), \tag{10}$$

and output the policy component $\widehat{\pi}$ of a saddle point $(\widehat{w}, \widehat{\pi}; \widehat{Q}, \widehat{\lambda})$. Although solving (10) exactly is generally computationally intractable, it provides a clean benchmark for the analysis: under suitable realizability and boundedness assumptions, an exact saddle point of $\widehat{L}$ yields a near-optimal and near-feasible policy. We first require that the density ratio of an optimal feasible policy belong to $\mathcal{W}$.

Assumption C. For an optimal policy π^*, we have $w^{\pi^*} = \mu^{\pi^*}/\mu_D \in \mathcal{W}$.

Algorithm 1: PDOCRL

Input: Dataset $\mathcal{D} = \{(s_j, a_j, s'_j)\}_{j=1}^n$, function classes $\mathcal{W}$, $\mathcal{Q}$ and Π, no-regret online linear-optimization oracle $\texttt{OLO}_{\mathcal{W}}$, no-regret policy optimization oracle $\texttt{PO}_{\Pi,\mathcal{Q}}$, linear minimization oracles $\texttt{LMO}_{\mathcal{Q}}$ and $\texttt{LMO}_{B\Delta^I}$, dual bound $B > 0$, number of iterations T.

for $t = 1, \dots, T$ **do**

$w_t \leftarrow \texttt{OLO}_{\mathcal{W}}(\{\nabla_w \widehat{L}(\,\cdot\,, \pi_\tau; Q_\tau, \lambda_\tau)\}_{\tau=1}^{t-1})$.

$\pi_t \leftarrow \texttt{PO}_{\Pi,\mathcal{Q}}(Q_1, \dots, Q_{t-1})$.

$\lambda_t \leftarrow \texttt{LMO}_{B\Delta^I}(\widehat{L}(w_t, \pi_t; Q_{t-1}, \,\cdot\,))$.

$Q_t \leftarrow \texttt{LMO}_{\mathcal{Q}}(\widehat{L}(w_t, \pi_t; \,\cdot\,, \lambda_t))$.

Return: Uniform$(\pi_1, \dots, \pi_T)$

To control the statistical error of $\widehat{L}$, we also impose standard boundedness assumptions on the density-ratio and Q-function classes.

Assumption D. For all $w \in \mathcal{W}$, $\|w\|_\infty \le C^*$ and for all $Q \in \mathcal{Q}$, $\|Q\|_\infty \le (2 + \frac{1}{\varphi})/(1-\gamma)$.

The bound $(2+\frac{1}{\varphi})/(1-\gamma)$ on the Q-functions is natural since the Lagrangian reward $r_{\boldsymbol{\lambda}}$ is bounded by $2+\frac{1}{\varphi}$. The bound C^* on $\mathcal{W}$, together with the assumption $w^{\pi^*} \in \mathcal{W}$, implies $C^{\pi^*} = \|w^{\pi^*}\|_\infty \le C^*$. Thus, positing a density-ratio class $\mathcal{W}$ bounded by C^* amounts to assuming that the learner knows an upper bound on the concentrability coefficient of the optimal policy. Such an assumption is common in offline RL with function approximation. Appendix C.2 contains the guarantee for $\widehat{\pi}$ obtained by the minimax algorithm (10).

3.4 Oracle-efficient primal-dual algorithm

We have not yet addressed the computational challenges of finding the saddle point. In this section, we propose an oracle-efficient primal-dual algorithm that solves the saddle point problem (10) by iteratively optimizing for the primal variables w, π and the dual variables Q, λ, using standard oracles.

The w-player employs a no-regret online linear optimization algorithm over $\mathcal{W}$ defined as follows.

Definition 5. An online algorithm $\texttt{OLO}_{\mathcal{W}}$ is called a *no-regret online linear optimization oracle* for the function class $\mathcal{W} \subseteq \mathbb{R}_+^{\mathcal{S}\times\mathcal{A}}$, if for every sequence $c_1, \dots, c_T \in \mathcal{W}$, the algorithm sequentially outputs $w_t = \texttt{OLO}_{\mathcal{W}}(c_1, \dots, c_{t-1}) \in \mathcal{W}$, $t = 1, \dots, T$, such that, for every comparator $w \in \mathcal{W}$,

$$\tfrac{1}{T}\textstyle\sum_{t=1}^T \langle c_t, w - w_t\rangle \to 0, \quad \text{as } T \to \infty.$$

For example, with an assumption that $\mathcal{W}$ is convex, the projected gradient ascent can serve as the no-regret oracle, as discussed in Appendix D.2.

The π-player employs a no-regret policy optimization oracle over Π defined as follows.

Definition 6. An online algorithm $\texttt{PO}_{\Pi,\mathcal{Q}}$ is called a *no-regret policy optimization oracle* for the π-player, function class $\mathcal{Q} \subseteq \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ and policy class $\Pi \subseteq (\Delta(\mathcal{A}))^{\mathcal{S}}$, if for every sequence $Q_1, \dots, Q_T \in \mathcal{Q}$, the algorithm sequentially outputs policies $\pi_t = \texttt{PO}_{\Pi,\mathcal{Q}}(Q_1, \dots, Q_{t-1}) \in \Pi$, $t = 1, \dots, T$, such that, for every comparator policy $\pi \in \Pi$,

$$\tfrac{1}{T}\textstyle\sum_{t=1}^T \mathbb{E}_\pi[Q_t(s,\pi) - Q_t(s,\pi_t)] \to 0, \quad \text{as } T \to \infty,$$

where $\mathbb{E}_\pi$ denotes the expectation over $s \sim \sum_a \mu^\pi(\cdot, a)$, the state occupancy measure of π.

As discussed by Wei et al. [34], there exist many methods that can serve as the no-regret oracle, for example, the mirror descent approach [6] and natural policy gradient approach [12] of the form $\pi_t(\cdot|s) \propto \pi_{t-1}(\cdot|s)\exp(\alpha Q_{t-1}(s,\cdot))$ for some learning rate $\alpha > 0$.

For the Q-player, since the Q-dependent part of $\widehat{L}$ is linear in Q, we use a linear minimization oracle $\texttt{LMO}_{\mathcal{Q}}(c) \in \operatorname{argmin}_{Q\in\mathcal{Q}}\langle c, Q\rangle$. Similarly, since the λ-dependent part of $\widehat{L}$ is linear in λ, we use a linear minimization oracle $\texttt{LMO}_{B\Delta^I}(c) \in \operatorname{argmin}_{\lambda\in B\Delta^I}\langle c, \lambda\rangle$.

The full algorithm is described in Algorithm 1. The algorithm runs for T iterations and returns a policy sampled uniformly at random from $\pi_1, \ldots, \pi_T$. The policy returned by the algorithm has the following guarantee. For the exact statement and the proof, see Theorem 17 in Appendix C.

Theorem 7 (Informal)**.** *Under Assumptions A, B, C and D, running Algorithm 1 with dual bound $B = 1 + \frac{1}{\varphi}$ and T large enough, the output policy $\widehat{\pi}$ satisfies*

$$(1-\gamma)J_0(\widehat{\pi}) \geq (1-\gamma)J_0(\pi^*) - \varepsilon_n, \quad (1-\gamma)J_i(\widehat{\pi}) \geq \tau_i - \varepsilon_n, \;\; i = 1, \ldots, I.$$

with probability at least $1-\delta$, where $\varepsilon_n = \widetilde{\mathcal{O}}(C^* \sqrt{\log(I|\mathcal{W}||\Pi||\mathcal{Q}|/\delta)/n})$.

3.5 Adaptation to offline unconstrained RL setting

The algorithms we have discussed so far for the constrained RL setting can be adapted to the unconstrained offline RL setting by removing the dual variable λ and the associated terms in the Lagrangian, and the analysis can be adapted accordingly.

In fact, with a careful analysis on the guarantee of the saddle point of the Lagrangian, we can get a stronger guarantee in the unconstrained setting. Specifically, we can compete with any comparator policy covered by $\mathcal{W}$, instead of just the optimal policy as in the constrained setting. The adapted algorithm is shown in Algorithm 2, and the guarantee is stated in Theorem 8.

Algorithm 2: PDORL

Input: Dataset $\mathcal{D} = \{(s_j, a_j, s'_j)\}_{j=1}^n$, function classes $\mathcal{W}$, $\mathcal{Q}$ and Π, no-regret online linear-optimization oracle $\texttt{OLO}_{\mathcal{W}}$, no-regret policy optimization oracle $\texttt{PO}_{\Pi,\mathcal{Q}}$, linear minimization oracles $\texttt{LMO}_{\mathcal{Q}}$, number of iterations T.

for $t = 1, \ldots, T$ **do**
- $w_t \leftarrow \texttt{OLO}_{\mathcal{W}}(\{\nabla_w \widehat{L}(\,\cdot\,, \pi_\tau; Q_\tau)\}_{\tau=1}^{t-1})$.
- $\pi_t \leftarrow \texttt{PO}_{\Pi,\mathcal{Q}}(Q_1, \ldots, Q_{t-1})$.
- $Q_t \leftarrow \texttt{LMO}_{\mathcal{Q}}(\nabla_Q \widehat{L}(w_t, \pi_t; \,\cdot\,))$.

Return: Uniform$(\pi_1, \ldots, \pi_T)$

Theorem 8 (Informal)**.** *Assume $Q^\pi \in \mathcal{Q}$ for all $\pi \in \Pi$, and $w^{\pi_c} \in \mathcal{W}$ for some comparator policy $\pi_c \in \Pi$, and $\|w\|_\infty \leq C$ for all $w \in \mathcal{W}$ and $\|Q\|_\infty \leq 1/(1-\gamma)$ for all $Q \in \mathcal{Q}$. Running Algorithm 2 with T large enough, the output policy $\widehat{\pi}$ satisfies*

$$(1-\gamma)J(\widehat{\pi}) \geq (1-\gamma)J(\pi_c) - \varepsilon_n$$

with probability at least $1-\delta$, where $\varepsilon_n = \mathcal{O}(\frac{C}{1-\gamma}\sqrt{\log(|\mathcal{W}||\Pi||\mathcal{Q}|/\delta)/n})$.

4 Experiments

We compare PDOCRL with offline constrained RL algorithms [19], assessing constraint violation and reward maximization.[1] For each task, training is performed for 100,000 gradient steps using three cost thresholds and three random seeds. Each trained model is then evaluated over 20 episodes, and we report the average return. See Appendix F for details on the experimental setup.

Our empirical evaluation primarily examines constraint satisfaction, with reward performance assessed in conjunction with feasibility, rather than claiming a uniform improvement in the reward–cost trade-off. The experiments serve as a proof of concept for direct policy optimization under the proposed formulation. The practical implementation in Algorithm 3 uses gradient-based surrogates for the optimization oracles and evaluates the last-iterate policy. Consequently, the theoretical guarantees for Algorithm 1 do not directly apply to the implementation evaluated here. Appendix F.2 discusses these distinctions.

Benchmark and datasets The experiments are conducted on the BulletGym benchmark [7], consisting of continuous-control tasks with explicit safety constraints. We use the offline datasets

[1] Source code: `https://github.com/komin0407/PDOCRL`

provided by the DSRL package [19], which offers D4RL-styled [4] datasets for safe RL tasks. For each BulletGym task, the dataset is collected by training multiple safe RL expert policies under varying cost thresholds, then merging the resulting trajectories. We follow the standard evaluation protocol of [19], reporting normalized reward and normalized cost, where the normalized cost is defined as the ratio of the evaluated cost return to the cost threshold.

Baselines The baselines are taken from the benchmark paper Liu et al. [19]. BC-All and BC-Safe [18, 38] are imitation-learning-based methods. CDT [18] uses a transformer architecture conditioned on return and cost. BCQ-Lag [37], BEAR-Lag [37], and CPQ [37] are Q-learning-based methods, while COptiDICE [15] is a LP-based approach, most closely related to our method.

Evaluation metrics We evaluate each method using the normalized reward return and normalized cost return. The reward is normalized by the empirical minimum and maximum reward values of each task, making the results comparable across different environments. The cost is normalized by the target cost threshold, so that a value below 1 indicates that the safety constraint is satisfied. For simplicity, we refer to the normalized reward return and normalized cost return as *reward* and *cost*, respectively. For more details, refer to F.1.

Table 2: Evaluation results of the normalized reward and cost on BulletGym tasks. The cost threshold is 1. ↑: higher reward is better. ↓: lower cost (up to threshold 1) is better. **Bold**: Safe agents (normalized cost ≤ 1). Gray: Unsafe agents. **Blue**: Safe agent with the highest reward.

Task	BC-All		BC-Safe		CDT		BCQ-Lag		BEAR-Lag		CPQ		COptiDICE		**PDOCRL (Ours)**	
	Rew.↑	Cost↓	Rew.↑	Cost↓	Rew.↑	Cost↓	Rew.↑	Cost↓	Rew.↑	Cost↓	Rew.↑	Cost↓	Rew.↑	Cost↓	Rew.↑	Cost↓
BallRun	$0.60_{\pm 0.10}$	$5.08_{\pm 0.74}$	$0.27_{\pm 0.14}$	$1.46_{\pm 0.39}$	$0.39_{\pm 0.09}$	$1.16_{\pm 0.19}$	$0.76_{\pm 0.01}$	$3.91_{\pm 0.35}$	$-0.47_{\pm 0.00}$	$5.03_{\pm 0.00}$	$0.22_{\pm 0.00}$	$1.27_{\pm 0.12}$	$0.59_{\pm 0.00}$	$3.52_{\pm 0.00}$	$\mathbf{0.48}_{\pm 0.33}$	$\mathbf{0.86}_{\pm 0.82}$
CarRun	$\mathbf{0.97}_{\pm 0.02}$	$\mathbf{0.33}_{\pm 0.05}$	$\mathbf{0.94}_{\pm 0.00}$	$\mathbf{0.22}_{\pm 0.02}$	$\mathbf{0.99}_{\pm 0.01}$	$\mathbf{0.65}_{\pm 0.31}$	$\mathbf{0.94}_{\pm 0.01}$	$\mathbf{0.15}_{\pm 0.91}$	$0.68_{\pm 0.01}$	$7.78_{\pm 0.09}$	$0.95_{\pm 0.01}$	$1.79_{\pm 0.18}$	$\mathbf{0.87}_{\pm 0.00}$	$\mathbf{0.00}_{\pm 0.00}$	$\mathbf{0.80}_{\pm 0.08}$	$\mathbf{0.11}_{\pm 0.15}$
DroneRun	$0.24_{\pm 0.02}$	$2.13_{\pm 0.62}$	$\mathbf{0.28}_{\pm 0.25}$	$\mathbf{0.74}_{\pm 0.07}$	$\mathbf{0.63}_{\pm 0.04}$	$\mathbf{0.79}_{\pm 0.68}$	$0.72_{\pm 0.12}$	$5.54_{\pm 0.81}$	$0.42_{\pm 0.10}$	$2.47_{\pm 0.34}$	$0.33_{\pm 0.10}$	$3.52_{\pm 0.58}$	$0.67_{\pm 0.02}$	$4.15_{\pm 0.10}$	$\mathbf{0.10}_{\pm 0.15}$	$\mathbf{0.84}_{\pm 1.01}$
AntRun	$0.72_{\pm 0.06}$	$2.93_{\pm 2.40}$	$0.65_{\pm 0.15}$	$1.09_{\pm 0.84}$	$\mathbf{0.72}_{\pm 0.04}$	$\mathbf{0.91}_{\pm 0.42}$	$0.76_{\pm 0.07}$	$5.11_{\pm 2.39}$	$\mathbf{0.15}_{\pm 0.02}$	$\mathbf{0.73}_{\pm 0.07}$	$\mathbf{0.03}_{\pm 0.02}$	$\mathbf{0.02}_{\pm 0.09}$	$\mathbf{0.61}_{\pm 0.01}$	$\mathbf{0.94}_{\pm 0.69}$	$\mathbf{0.03}_{\pm 0.02}$	$\mathbf{0.00}_{\pm 0.00}$
BallCircle	$0.74_{\pm 0.15}$	$4.71_{\pm 1.79}$	$\mathbf{0.52}_{\pm 0.08}$	$\mathbf{0.65}_{\pm 0.17}$	$0.77_{\pm 0.06}$	$1.07_{\pm 0.27}$	$0.69_{\pm 0.11}$	$2.36_{\pm 1.04}$	$0.86_{\pm 0.18}$	$3.09_{\pm 1.53}$	$\mathbf{0.64}_{\pm 0.01}$	$\mathbf{0.76}_{\pm 0.00}$	$0.70_{\pm 0.04}$	$2.61_{\pm 0.79}$	$\mathbf{0.61}_{\pm 0.15}$	$\mathbf{0.21}_{\pm 0.31}$
CarCircle	$0.58_{\pm 0.25}$	$3.74_{\pm 2.20}$	$\mathbf{0.50}_{\pm 0.22}$	$\mathbf{0.84}_{\pm 0.67}$	$\mathbf{0.75}_{\pm 0.06}$	$\mathbf{0.95}_{\pm 0.61}$	$0.63_{\pm 0.19}$	$1.89_{\pm 1.37}$	$0.74_{\pm 0.10}$	$2.18_{\pm 1.33}$	$\mathbf{0.71}_{\pm 0.02}$	$\mathbf{0.33}_{\pm 0.00}$	$0.49_{\pm 0.05}$	$3.14_{\pm 2.98}$	$\mathbf{0.63}_{\pm 0.05}$	$\mathbf{0.11}_{\pm 0.21}$
DroneCircle	$0.72_{\pm 0.04}$	$3.03_{\pm 0.29}$	$\mathbf{0.56}_{\pm 0.18}$	$\mathbf{0.57}_{\pm 0.27}$	$\mathbf{0.63}_{\pm 0.07}$	$\mathbf{0.98}_{\pm 0.27}$	$0.80_{\pm 0.07}$	$3.07_{\pm 0.89}$	$0.78_{\pm 0.04}$	$3.68_{\pm 0.44}$	$-0.22_{\pm 0.05}$	$1.28_{\pm 0.97}$	$0.26_{\pm 0.03}$	$1.02_{\pm 0.46}$	$\mathbf{-0.26}_{\pm 0.00}$	$\mathbf{0.26}_{\pm 0.39}$
AntCircle	$0.58_{\pm 0.19}$	$4.90_{\pm 3.55}$	$\mathbf{0.40}_{\pm 0.16}$	$\mathbf{0.96}_{\pm 2.67}$	$0.54_{\pm 0.20}$	$1.78_{\pm 4.33}$	$0.58_{\pm 0.25}$	$2.87_{\pm 3.08}$	$0.65_{\pm 0.20}$	$5.48_{\pm 3.33}$	$\mathbf{0.00}_{\pm 0.00}$	$\mathbf{0.00}_{\pm 0.00}$	$0.17_{\pm 0.10}$	$5.04_{\pm 6.74}$	$\mathbf{0.00}_{\pm 0.00}$	$\mathbf{0.00}_{\pm 0.00}$
Average	$0.64_{\pm 0.25}$	$3.36_{\pm 3.31}$	$\mathbf{0.52}_{\pm 0.27}$	$\mathbf{0.82}_{\pm 1.37}$	$0.68_{\pm 0.19}$	$1.04_{\pm 1.65}$	$0.74_{\pm 0.25}$	$3.11_{\pm 3.55}$	$0.48_{\pm 0.27}$	$3.80_{\pm 3.95}$	$0.33_{\pm 0.29}$	$1.12_{\pm 1.85}$	$0.55_{\pm 0.24}$	$2.55_{\pm 3.62}$	$\mathbf{0.30}_{\pm 0.38}$	$\mathbf{0.30}_{\pm 0.60}$

In the results reported in Table 2, PDOCRL has a mean normalized cost below the threshold of one on all eight BulletGym tasks. This observation concerns the reported averages over seeds and cost thresholds; it does not imply constraint satisfaction in every individual run or evaluation episode. PDOCRL achieves competitive rewards on some tasks, but its reward performance is lower than that of several baselines on others. The results therefore provide stronger empirical support for constraint satisfaction than for a uniform improvement in the reward–cost trade-off, and indicate that the practical method can be conservative.

4.1 Ablation study

We conduct an ablation study to demonstrate the effectiveness of the reparameterization trick used in our algorithm, which avoids the policy extraction step and eliminates the need for knowledge of the data-generating distribution. The ablated variant obtains $\widehat{w}$ by optimizing for the estimated Lagrangian (5), then using the policy extraction step (3) to obtain a policy.

As shown in Table 3, PDOCRL shows better average reward and cost performance than the variant that requires policy extraction, suggesting that the policy extraction step can be a meaningful bottleneck in offline constrained reinforcement learning.

5 Conclusion

We proposed PDOCRL, an oracle-efficient algorithm for offline constrained RL with general function approximation under partial data coverage. Our analysis shows that restricted Lagrangian saddle-point formulations can admit spurious solutions, and that Slater's condition together with all-policy action-value realizability rules out this pathology. Building on this guarantee, PDOCRL combines Lagrangian decomposition with an occupancy-ratio reparameterization to optimize policies directly, without access to the data-generating distribution. A natural direction for future work is to relax the all-policy realizability condition, ideally toward a single-policy realizability assumption.

Table 3: Ablation study on BulletGym tasks. We compare PDOCRL (Ours) with an ablated variant, *Policy extraction required*. ↑: higher reward is better. ↓: lower cost is better.

Task	**PDOCRL (Ours)**		Policy extraction required	
	Reward ↑	Cost ↓	Reward ↑	Cost ↓
BallRun	$0.48_{\pm 0.33}$	$0.86_{\pm 0.82}$	$-1.06_{\pm 0.15}$	$2.69_{\pm 1.44}$
CarRun	$0.80_{\pm 0.08}$	$0.11_{\pm 0.15}$	$-0.59_{\pm 0.26}$	$4.63_{\pm 2.51}$
DroneRun	$0.10_{\pm 0.15}$	$0.84_{\pm 1.01}$	$-0.26_{\pm 0.09}$	$0.35_{\pm 1.00}$
AntRun	$0.03_{\pm 0.02}$	$0.00_{\pm 0.00}$	$0.21_{\pm 0.06}$	$0.01_{\pm 0.02}$
BallCircle	$0.61_{\pm 0.15}$	$0.21_{\pm 0.31}$	$0.01_{\pm 0.04}$	$2.19_{\pm 2.92}$
CarCircle	$0.63_{\pm 0.05}$	$0.11_{\pm 0.21}$	$-0.04_{\pm 0.05}$	$3.87_{\pm 4.65}$
DroneCircle	$-0.26_{\pm 0.00}$	$0.26_{\pm 0.39}$	$-0.11_{\pm 0.13}$	$1.30_{\pm 0.60}$
AntCircle	$0.00_{\pm 0.00}$	$0.00_{\pm 0.00}$	$0.06_{\pm 0.02}$	$0.65_{\pm 0.35}$
Average	$0.30_{\pm 0.38}$	$0.30_{\pm 0.60}$	$-0.22_{\pm 0.41}$	$1.96_{\pm 2.74}$

Limitations. Our theoretical guarantees apply to the oracle-based Algorithm 1 and its uniform-mixture output policy under the stated assumptions. They do not directly cover the practical implementation evaluated in our experiments, which uses neural function approximation, gradient-based optimization surrogates, target networks, clipped importance weights, and a last-iterate policy. The empirical results should therefore be interpreted as a proof of concept for the proposed formulation rather than as a verification of the theoretical guarantees for the neural implementation. The results also reveal a tendency toward conservative behavior: low normalized costs can be accompanied by substantially reduced rewards, and the reported averages do not establish constraint satisfaction for every individual run or evaluation episode.

Acknowledgment

KH and SK were supported by the National Research Foundation of Korea (NRF), funded by the Ministry of Science and ICT (Grant No. RS-2022-NR068758, RS-2026-25471598).

References

[1] Eitan Altman. *Constrained Markov Decision Processes*. Vol. 7. CRC Press, 1999.

[2] Jinglin Chen and Nan Jiang. "Offline reinforcement learning under value and density-ratio realizability: The power of gaps". In: *Uncertainty in Artificial Intelligence*. PMLR. 2022, pp. 378–388.

[3] Ching-An Cheng, Tengyang Xie, Nan Jiang, and Alekh Agarwal. "Adversarially trained actor critic for offline reinforcement learning". In: *International Conference on Machine Learning*. PMLR. 2022, pp. 3852–3878.

[4] Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. "D4rl: Datasets for deep data-driven reinforcement learning". In: *arXiv preprint arXiv:2004.07219* (2020).

[5] Germano Gabbianelli, Gergely Neu, Matteo Papini, and Nneka M Okolo. "Offline primal-dual reinforcement learning for linear mdps". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2024, pp. 3169–3177.

[6] Matthieu Geist, Bruno Scherrer, and Olivier Pietquin. "A theory of regularized markov decision processes". In: *International conference on machine learning*. PMLR. 2019, pp. 2160–2169.

[7] Sven Gronauer. "Bullet-safety-gym: A framework for constrained reinforcement learning". In: (2022).

[8] Monique Guignard and Siwhan Kim. "Lagrangean decomposition: A model yielding stronger Lagrangean bounds". In: *Mathematical programming* 39.2 (1987), pp. 215–228.

[9] L.K. Hansen and P. Salamon. "Neural network ensembles". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 12.10 (1990), pp. 993–1001.

[10] Kihyuk Hong and Ambuj Tewari. "A Primal-Dual Algorithm for Offline Constrained Reinforcement Learning with Linear MDPs". In: *International Conference on Machine Learning*. PMLR. 2024, pp. 18711–18737.

[11] Nan Jiang and Tengyang Xie. "Offline reinforcement learning in large state spaces: Algorithms and guarantees". In: *Statistical Science* (2024).

[12] Sham M Kakade. "A natural policy gradient". In: *Advances in neural information processing systems* 14 (2001).

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

[14] 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*. PMLR. 2021, pp. 6120–6130.

[15] Jongmin Lee, Cosmin Paduraru, Daniel J Mankowitz, Nicolas Heess, Doina Precup, Kee-Eung Kim, and Arthur Guez. "COptiDICE: Offline constrained reinforcement learning via stationary distribution correction estimation". In: *International Conference on Learning Representations*. 2022.

[16] Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. "Offline reinforcement learning: Tutorial, review, and perspectives on open problems". In: *arXiv preprint arXiv:2005.01643* (2020).

[17] Qiang Liu, Lihong Li, Ziyang Tang, and Dengyong Zhou. "Breaking the curse of horizon: Infinite-horizon off-policy estimation". In: *Advances in Neural Information Processing Systems* 31 (2018), pp. 5356–5366.

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

[19] Zuxin Liu et al. *Datasets and Benchmarks for Offline Safe Reinforcement Learning*. 2023. arXiv: 2306.09303 [cs.LG]. URL: https://arxiv.org/abs/2306.09303.

[20] Alan S Manne. "Linear programming and sequential decisions". In: *Management Science* 6.3 (1960), pp. 259–267.

[21] Prashant Mehta and Sean Meyn. "Q-learning and Pontryagin's minimum principle". In: *Proceedings of the 48h IEEE Conference on Decision and Control (CDC) held jointly with 2009 28th Chinese Control Conference*. IEEE. 2009, pp. 3598–3605.

[22] Washim Uddin Mondal and Vaneet Aggarwal. "Sample-Efficient Constrained Reinforcement Learning with General Parameterization". In: *Advances in Neural Information Processing Systems*. Vol. 37. 2024, pp. 68380–68405.

[23] Rémi Munos and Csaba Szepesvári. "Finite-Time Bounds for Fitted Value Iteration." In: *Journal of Machine Learning Research* 9.5 (2008).

[24] Ofir Nachum, Yinlam Chow, Bo Dai, and Lihong Li. "DualDICE: Behavior-agnostic estimation of discounted stationary distribution corrections". In: *Advances in Neural Information Processing Systems* 32 (2019), pp. 2315–2325.

[25] Ofir Nachum and Bo Dai. "Reinforcement learning via Fenchel-Rockafellar duality". In: *arXiv preprint arXiv:2001.01866* (2020).

[26] Ofir Nachum, Bo Dai, Ilya Kostrikov, Yinlam Chow, Lihong Li, and Dale Schuurmans. "AlgaeDICE: Policy gradient from arbitrary experience". In: *arXiv preprint arXiv:1912.02074* (2019).

[27] Gergely Neu and Nneka Okolo. "Efficient global planning in large MDPs via stochastic primal-dual optimization". In: *International Conference on Algorithmic Learning Theory*. PMLR. 2023, pp. 1101–1123.

[28] Gergely Neu and Nneka Okolo. "Offline RL via Feature-Occupancy Gradient Ascent". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2025.

[29] Asuman E Ozdaglar, Sarath Pattathil, Jiawei Zhang, and Kaiqing Zhang. "Revisiting the linear-programming framework for offline RL with general function approximation". In: *International Conference on Machine Learning*. PMLR. 2023, pp. 26769–26791.

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

[31] Paria Rashidinejad, Hanlin Zhu, Kunhe Yang, Stuart Russell, and Jiantao Jiao. "Optimal conservative offline RL with general function approximation via augmented Lagrangian". In: *International Conference on Learning Representations*. 2023.

[32] Fred Shepardson and Roy E Marsten. "A Lagrangean relaxation algorithm for the two duty period scheduling problem". In: *Management Science* 26.3 (1980), pp. 274–281.

[33] Masatoshi Uehara, Jiawei Huang, and Nan Jiang. "Minimax weight and Q-function learning for off-policy evaluation". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 9659–9668.

[34] Honghao Wei, Xiyue Peng, Arnob Ghosh, and Xin Liu. "Adversarially trained weighted actor-critic for safe offline reinforcement learning". In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 52806–52835.

[35] Tengyang Xie, Ching-An Cheng, Nan Jiang, Paul Mineiro, and Alekh Agarwal. "Bellman-consistent pessimism for offline reinforcement learning". In: *Advances in neural information processing systems* 34 (2021), pp. 6683–6694.

[36] Tengyang Xie, Yifei Ma, and Yu-Xiang Wang. "Towards optimal off-policy evaluation for reinforcement learning with marginalized importance sampling". In: *Advances in Neural Information Processing Systems* 32 (2019), pp. 9665–9675.

[37] Haoran Xu, Xianyuan Zhan, and Xiangyu Zhu. "Constraints penalized q-learning for safe offline reinforcement learning". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 36. 8. 2022, pp. 8753–8760.

[38] Mengdi Xu, Zuxin Liu, Peide Huang, Wenhao Ding, Zhepeng Cen, Bo Li, and Ding Zhao. "Trustworthy reinforcement learning against intrinsic vulnerabilities: Robustness, safety, and generalizability". In: *arXiv preprint arXiv:2209.08025* (2022).

[39] Mengjiao Yang, Ofir Nachum, Bo Dai, Lihong Li, and Dale Schuurmans. "Off-policy evaluation via the regularized Lagrangian". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 6551–6561.

[40] Donghao Ying, Mengzi Amy Guo, Hyunin Lee, Yuhao Ding, Javad Lavaei, and Zuo-Jun Max Shen. "Policy-Based Primal-Dual Methods for Concave CMDPs with Variance Reduction". In: *Journal of Artificial Intelligence Research* 83.31 (2025).

[41] Andrea Zanette. "When is realizability sufficient for off-policy reinforcement learning?" In: *International Conference on Machine Learning*. PMLR. 2023, pp. 40637–40668.

[42] Wenhao Zhan, Baihe Huang, Audrey Huang, Nan Jiang, and Jason Lee. "Offline reinforcement learning with realizability and single-policy concentrability". In: *Conference on Learning Theory*. PMLR. 2022, pp. 2730–2775.

[43] Haobo Zhang, Xiyue Peng, Honghao Wei, and Xin Liu. "Safe and Efficient: A Primal-Dual Method for Offline Convex CMDPs under Partial Data Coverage". In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 34239–34269.

[44] Ruiyi Zhang, Bo Dai, Lihong Li, and Dale Schuurmans. "GenDICE: Generalized offline estimation of stationary values". In: *International Conference on Learning Representations*. 2020.

[45] Shangtong Zhang, Bo Liu, and Shimon Whiteson. "GradientDICE: Rethinking generalized offline estimation of stationary values". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 11194–11203.

[46] Hanlin Zhu, Paria Rashidinejad, and Jiantao Jiao. "Importance weighted actor-critic for optimal conservative offline reinforcement learning". In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 49579–49602.

A Proof of Proposition 1

We construct a single one-step contextual bandit ($\gamma = 0$), data distribution, and policy class that work for all n. Only the learned ratio $\widehat{w}_n$ depends on n. Let $\mathcal{S} = \{c, r\}$, $\mathcal{A} = \{0, 1\}$, $d_0(c) = d_0(r) = \frac{1}{2}$. The behavior policy π_B is uniform such that the induced occupancy measure is $\mu_D(s, a) = \frac{1}{4}$ for every $(s, a) \in \mathcal{S} \times \mathcal{A}$. The reward function is $r(c, 0) = r(c, 1) = 0$ and $r(r, 1) = 1, R(r, 0) = 0$. Therefore, for any policy π, $J(\pi) = \mathbb{E}_\pi[r(s, a)] = \frac{1}{2}\pi(1 \mid r)$ and it follows that any policy π with $\pi(1|r) = 1$ is optimal with optimal value $J(\pi) = \frac{1}{2}$.

Let $\Pi = \{\pi^*, \widetilde{\pi}\}$ where $\pi^\star$ is an optimal policy with $\pi^\star(0 \mid c) = \frac{1}{8}$, $\pi^\star(1 \mid c) = \frac{7}{8}$, $\pi^\star(1 \mid r) = 1$, $\pi^\star(0 \mid r) = 0$; and $\widetilde{\pi}$ is $\widetilde{\pi}(0 \mid c) = \frac{7}{8}$, $\widetilde{\pi}(1 \mid c) = \frac{1}{8}$, $\widetilde{\pi}(1 \mid r) = \frac{1}{7}$, $\widetilde{\pi}(0 \mid r) = \frac{6}{7}$. Then, $J(\pi^*) = \frac{1}{2}$ and $J(\widetilde{\pi}) = \frac{1}{14}$.

We first verify boundedness of the true optimal density ratio. Since this is a one-step contextual bandit, $\mu^{\pi^\star}(s, a) = \mathbb{P}(S = s)\pi^\star(a \mid s) = \frac{1}{2}\pi^\star(a \mid s)$. Because $\mu_D(s, a) = 1/4$, we have $w^\star(s, a) = \frac{\mu^{\pi^\star}(s,a)}{\mu_D(s,a)} = 2\pi^\star(a \mid s)$. Therefore, $w^\star(c, 0) = \frac{1}{4}$, $w^\star(c, 1) = \frac{7}{4}$, and $w^\star(r, 1) = 2$, $w^\star(r, 0) = 0$. Thus $0 \le w^\star(s, a) \le 2$ for all (s, a).

Now, consider a learned density-ratio $\widehat{w}_n$ with $\widehat{w}_n(c, 0) = 2 - \frac{1}{n}$, $\widehat{w}_n(c, 1) = \frac{1}{n}$, $\widehat{w}_n(r, 1) = 2$ and $\widehat{w}_n(r, 0) = 0$. Then, $0 \le \widehat{w}_n(s, a) \le 2$ for all (s, a). Define $\widehat{\mu}_n(s, a) := \mu_D(s, a)\widehat{w}_n(s, a)$. Since $\mu_D(s, a) = 1/4$, we have $\widehat{\mu}_n(c, 0) = \frac{1}{4}\left(2 - \frac{1}{n}\right)$, $\widehat{\mu}_n(c, 1) = \frac{1}{4n}$, $\widehat{\mu}_n(r, 1) = \frac{1}{2}$, and $\widehat{\mu}_n(r, 0) = 0$. Hence, the policy extracted from $\widehat{\mu}$ puts weight 1 on state r, and is optimal, showing the first claim.

Next we analyze the population weighted-log extraction over the finite class Π:

$$\operatorname*{argmax}_{\pi \in \Pi} L_n(\pi) \coloneqq \mathbb{E}_{\mu_D}[\widehat{w}_n \log \pi(a|s)].$$

It can be shown that

$$L_n(\pi^*) - L_n(\widetilde{\pi}) = \frac{\log 7}{2n} > 0,$$

Hence the population weighted-log extraction over Π uniquely selects $\pi^\star$, showing the second claim.

It remains to show that the empirical weighted-log extraction from the dataset $\mathcal{D}$

$$\operatorname*{argmax}_{\pi \in \Pi} \widehat{L}_n(\pi) \coloneqq \frac{1}{n}\sum_{i=1}^{n} \widehat{w}_n(s_i, a_i) \log \pi(a_i|s_i),$$

selects the bad policy with constant probability. Define the counts

$$N_{s,a} := \sum_{i=1}^{n} \mathbb{I}\{(s_i, a_i) = (s, a)\}.$$

Then,

$$\begin{aligned}
\widehat{L}_n(\pi^\star) - \widehat{L}_n(\widetilde{\pi}) &= \frac{1}{n}\sum_{i=1}^{n} \widehat{w}_n(S_i, A_i) \log \frac{\pi^\star(A_i \mid S_i)}{\widetilde{\pi}(A_i \mid S_i)} \\
&= \frac{\log 7}{n}\left[-2(N_{c0} - N_{r1}) + \frac{1}{n}(N_{c0} + N_{c1})\right] \\
&\le \frac{\log 7}{n}\left[-2(N_{c0} - N_{r1}) + 1\right],
\end{aligned}$$

where the inequality follows since $N_{c0} - N_{r1} \le n$. Hence, $\widehat{L}_n(\pi^*) < \widehat{L}_n(\widetilde{\pi})$ and the empirical policy extraction procedure will select the suboptimal $\widehat{\pi}$, as long as $N_{c0} - N_{r1} \ge 1$.

Now, we lower-bound $\mathbb{P}(N_{c0} - N_{r1} \ge 1)$. For each i, define

$$X_i := \mathbb{I}\{(s_i, a_i) = (c, 0)\} - \mathbb{I}\{(s_i, a_i) = (r, 1)\}.$$

Because μ_D is uniform over the four state-action pairs, we have

$$\mathbb{P}(X_i = 1) = \frac{1}{4}, \qquad \mathbb{P}(X_i = -1) = \frac{1}{4}, \qquad \mathbb{P}(X_i = 0) = \frac{1}{2}.$$

Let $X := \sum_{i=1}^{n} X_i = N_{c0} - N_{r1}$. The distribution of X is symmetric around zero, because X_i and $-X_i$ have the same distribution. Therefore, $\mathbb{P}(X > 0) = \mathbb{P}(X < 0)$ and it follows that

$$\mathbb{P}(X > 0) = \frac{1 - \mathbb{P}(X = 0)}{2}.$$

We now prove that $\mathbb{P}(X = 0) \leq 1/2$. Let $K := \sum_{i=1}^{n} \mathbf{1}\{|X_i| = 1\}$. Then $K \sim \text{Binomial}\left(n, \frac{1}{2}\right)$. Conditional on $K = k$, the k nonzero signs are i.i.d. Rademacher random variables. If $k = 0$, then $X = 0$ with probability 1. If $k = 1$, then $X \neq 0$ with probability 1. If $k \geq 2$, then

$$\mathbb{P}(X = 0 \mid K = k) \leq \frac{1}{2}.$$

Indeed, for odd k, the conditional probability is 0. For even $k \geq 2$,

$$\mathbb{P}(X = 0 \mid K = k) = \binom{k}{k/2} 2^{-k}.$$

For $k = 2$, this equals $1/2$, and for every even $k \geq 4$, the central binomial probability is strictly smaller than $1/2$. Hence

$$\begin{aligned}
\mathbb{P}(X = 0) &= \sum_{k=0}^{n} \mathbb{P}(X = 0 \mid K = k)\mathbb{P}(K = k) \\
&\leq \mathbb{P}(K = 0) + \frac{1}{2}\mathbb{P}(K \geq 2) \\
&= 2^{-n} + \frac{1}{2}\left(1 - 2^{-n} - n2^{-n}\right) \\
&= \frac{1}{2} + \frac{1 - n}{2^{n+1}} \leq \frac{1}{2},
\end{aligned}$$

where the second equality uses $\mathbb{P}(K = 0) = 2^{-n}$ and $\mathbb{P}(K = 1) = n2^{-n}$. Consequently,

$$\mathbb{P}(N_{c0} - N_{r1} > 0) = \mathbb{P}(X > 0) = \frac{1 - \mathbb{P}(X = 0)}{2} \geq \frac{1}{4}.$$

Hence, the suboptimal policy $\widetilde{\pi}$ is chosen with probability at least $1/4$, completing the proof.

B Known data distribution case

Introducing Lagrangian multipliers $\boldsymbol{V} \in \mathbb{R}^{|\mathcal{S}|}$ for the Bellman flow constraints and $\boldsymbol{\lambda} \in \mathbb{R}^{I}_{+}$ for the safety constraints in the linear program, the Lagrangian function is

$$L(\boldsymbol{\mu}, \boldsymbol{V}, \boldsymbol{\lambda}) \coloneqq \langle \boldsymbol{\mu}, \boldsymbol{r} \rangle + \langle \boldsymbol{V}, (1 - \gamma)\boldsymbol{d}_0 + \gamma \boldsymbol{P}^{\top}\boldsymbol{\mu} - \boldsymbol{E}^{\top}\boldsymbol{\mu} \rangle + \sum_{i=1}^{I} \lambda_i(\langle \boldsymbol{\mu}, \boldsymbol{r}_i \rangle - \tau_i),$$

In the general function approximation setting, the restricted saddle-point problem over $\mathcal{F}(\mathcal{U}) \times (\mathcal{V} \times \mathbb{R}^{I}_{+})$ can admit spurious saddle points that do not correspond to optimal policies. The below proposition shows that merely realizing an optimal policy, its occupancy measure, and its associated action-value functions is not sufficient to rule this out.

Proposition 9. *There exists a CMDP $(\mathcal{S}, \mathcal{A}, r_0, \{r_i\}_{i=1}^{I}, \gamma, d_0)$ and function classes $\mathcal{U} \subseteq \mathbb{R}_{+}^{\mathcal{S}\times\mathcal{A}}$ and $\mathcal{V} \subseteq \mathbb{R}^{\mathcal{S}}$ with $\mu^* \in \mathcal{U}$ and $V^* \in \mathcal{V}$ with the following property. The function classes satisfy realizability assumption: $\mu^* \in \mathcal{U}$, $V_i^* \in \mathcal{V}$ for all $i = 0, 1, \ldots, I$. However, there exists a saddle point $(\widehat{\mu}; \widehat{V}, \widehat{\lambda})$ of the Lagrangian function L associated with the linear programming formulation of the MDP over $\mathcal{U} \times (\mathcal{V} \times \mathbb{R}^{I}_{+})$, where the extracted policy $\widehat{\pi}$ is not optimal.*

Proof of Proposition 9. We explicitly construct a CMDP $(\mathcal{S}, \mathcal{A}, r_0, \{r_i\}_{i=1}^{I}, \gamma, d_0)$ and function classes $\mathcal{U}$, $\mathcal{V}$ as follows; see Figure 1 for an illustration.

For notational simplicity, we identify the restricted primal class $\mathcal{U}$ with the set of admissible occupancy measures, and construct it so that

$$\mathcal{U} = \{\mu^*, \widetilde{\mu}\}.$$

We explicitly construct a CMDP $(\mathcal{S}, \mathcal{A}, r_0, \{r_i\}_{i=1}^I, \gamma, d_0)$ and function classes $\mathcal{U}, \mathcal{V}$ as follows.

Let $I = 1$, $\mathcal{S} = \{s_0, l_1, r_1, l_2, r_2\}$, and $\mathcal{A} = \{L, R\}$. Let the objective reward r_0 be defined by

$$r_0(l_1, a) = 1, \qquad r_0(l_2, a) = 4 \qquad \text{for all } a \in \mathcal{A},$$

and all other entries of r_0 are zero. We choose the constraint reward to be identical:

$$r_1 := r_0, \qquad \tau_1 := 0.$$

Let $\gamma = 1/2$ and $d_0(s_0) = 1$, so that the process starts from s_0 deterministically.

The states l_1, l_2, and r_2 are absorbing, i.e.,

$$P(s \mid s, L) = P(s \mid s, R) = 1 \qquad \text{for } s \in \{l_1, l_2, r_2\}.$$

At state s_0, action L transitions uniformly to the next layer:

$$P(l_1 \mid s_0, L) = P(r_1 \mid s_0, L) = 1/2.$$

At state r_1, action L also transitions uniformly:

$$P(l_2 \mid r_1, L) = P(r_2 \mid r_1, L) = 1/2.$$

On the other hand, action R is biased to the right:

$$P(l_1 \mid s_0, R) = P(l_2 \mid r_1, R) = 1/4, \qquad P(r_1 \mid s_0, R) = P(r_2 \mid r_1, R) = 3/4.$$

Since $r_1 = r_0 \geq 0$ and $\mu \geq 0$, the safety constraint

$$\langle \mu, r_1 \rangle \geq \tau_1 = 0$$

is redundant for every nonnegative μ. Hence the set of optimal policies of this CMDP coincides with that of the underlying unconstrained MDP with reward r_0.

Let $V^* := V_0^* = V_1^*$. A direct calculation shows that

$$V^*(s_0) = 1, \qquad V^*(l_1) = 2, \qquad V^*(r_1) = 2, \qquad V^*(l_2) = 8, \qquad V^*(r_2) = 0.$$

Moreover, an optimal occupancy measure is induced by the policy that always chooses action L:

$$\mu^*(s_0, L) = \frac{1}{2}, \qquad \mu^*(l_1, L) = \frac{1}{4}, \qquad \mu^*(r_1, L) = \frac{1}{8}, \qquad \mu^*(l_2, L) = \mu^*(r_2, L) = \frac{1}{16},$$

and all other entries are zero.

The CMDP Lagrangian is

$$L(\mu; V, \lambda) = \langle \mu, r_0 \rangle + \langle V, (1-\gamma) d_0 + \gamma P^\top \mu - E^\top \mu \rangle + \lambda(\langle \mu, r_1 \rangle - \tau_1).$$

At $(V^*, 0)$, this becomes

$$L(\mu; V^*, 0) = \langle \mu, r_0 \rangle + \langle V^*, (1-\gamma) d_0 + \gamma P^\top \mu - E^\top \mu \rangle.$$

Using Bellman optimality of V^*, every action has zero Bellman residual except (r_1, R), for which

$$r_0(r_1, R) + \gamma \sum_{s'} P(s' \mid r_1, R) V^*(s') - V^*(r_1) = 0 + \frac{1}{2}\Big(\frac{1}{4} \cdot 8 + \frac{3}{4} \cdot 0\Big) - 2 = -1.$$

Therefore,

$$L(\mu; V^*, 0) = (1-\gamma) V^*(s_0) - \mu(r_1, R) = \frac{1}{2} V^*(s_0) - \mu(r_1, R).$$

Hence $L(\mu; V^*, 0)$ depends on μ only through $\mu(r_1, R)$, and any μ with $\mu(r_1, R) = 0$ is a best response to $(V^*, 0)$.

Now define $\widetilde{\mu}$ by

$$\widetilde{\mu}(s_0, R) = \frac{1}{2}, \qquad \widetilde{\mu}(l_1, L) = \widetilde{\mu}(l_1, R) = \frac{1}{4},$$

and all other entries zero. Then $\widetilde{\mu}(r_1, R) = 0$, so $\widetilde{\mu}$ is also a maximizer of $L(\cdot; V^*, 0)$ over $\mathcal{F}(\mathcal{U})$. Let

$$\mathcal{V} = \{V^*\}.$$

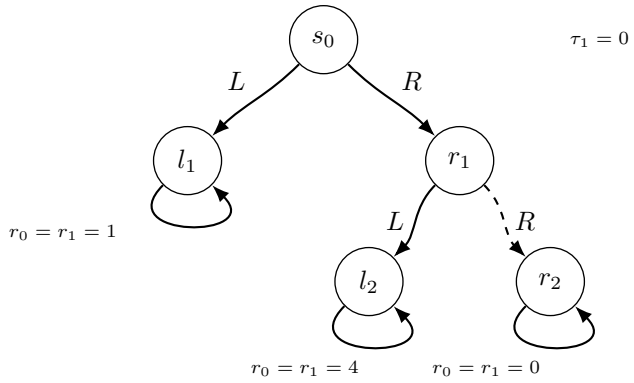


Figure 1: Counterexample for Proposition 9. Here $r_1 = r_0$ and $\tau_1 = 0$, so the constraint is redundant. Nevertheless, the restricted saddle point $(\widetilde{\mu}; V^*, 0)$ induces a suboptimal extracted policy.

Since $r_1 = r_0 \geq 0$ and $\tau_1 = 0$, for any $\lambda \geq 0$,

$$L(\widetilde{\mu}; V^*, \lambda) = L(\widetilde{\mu}; V^*, 0) + \lambda \langle \widetilde{\mu}, r_0 \rangle \geq L(\widetilde{\mu}; V^*, 0).$$

Thus $\lambda = 0$ is a minimizing dual variable for $\widetilde{\mu}$. Because $\mathcal{V}$ is a singleton and $\widetilde{\mu}$ maximizes $L(\cdot; V^*, 0)$, we conclude that

$$(\widetilde{\mu}; V^*, 0)$$

is a saddle point of L over $\mathcal{U} \times (\mathcal{V} \times \mathbb{R}_+)$.

The realizability condition also holds. By construction, $\mu^* \in \mathcal{U}$, and since $r_1 = r_0$, we have

$$V_1^* = V_0^* = V^* \in \mathcal{V}.$$

Finally, we show that the policy extracted from $\widetilde{\mu}$ is not optimal. Define the extracted policy by

$$\pi_\mu(a \mid s) = \begin{cases} \dfrac{\mu(s,a)}{\sum_{a'} \mu(s,a')}, & \text{if } \sum_{a'} \mu(s,a') > 0, \\ \mathbf{1}\{a = R\}, & \text{otherwise.} \end{cases}$$

Under this rule, the policy $\widetilde{\pi} := \pi_{\widetilde{\mu}}$ chooses action R at s_0, and since $\widetilde{\mu}(r_1, L) = \widetilde{\mu}(r_1, R) = 0$, it also chooses the default action R at r_1. Therefore,

$$V_{r_0}^{\widetilde{\pi}}(r_1) = \frac{1}{2}\Big(\frac{1}{4}V^*(l_2) + \frac{3}{4}V^*(r_2)\Big) = \frac{1}{2}\Big(\frac{1}{4} \cdot 8 + \frac{3}{4} \cdot 0\Big) = 1,$$

and hence

$$V_{r_0}^{\widetilde{\pi}}(s_0) = \frac{1}{2}\Big(\frac{1}{4}V^*(l_1) + \frac{3}{4}V_{r_0}^{\widetilde{\pi}}(r_1)\Big) = \frac{1}{2}\Big(\frac{1}{4} \cdot 2 + \frac{3}{4} \cdot 1\Big) = \frac{5}{8} < 1 = V^*(s_0).$$

Thus $\widetilde{\pi}$ is not optimal, even though $(\widetilde{\mu}; V^*, 0)$ is a saddle point of the restricted CMDP Lagrangian. This proves the proposition. □

This proposition suggests that finding a saddle point of L over $\mathcal{U} \times (\mathcal{V} \times \lambda)$ may not give an optimal solution even when $\mathcal{U} \times (\mathcal{V} \times \lambda)$ contains an optimal solution. As a workaround, Zhan et al. [42] add a regularization term $-\alpha \mathbb{E}_\mu[f(\mu/\mu_D)]$ to enforce strong concavity of L in terms of μ. However, this approach leads to sample complexity for finding ε near optimal policy scales with $1/\varepsilon^4$. Rather than regularizing the saddle problem, we make explicit a stronger realizability requirement on the function class $\mathcal{V}$:

Assumption E (All-Policy Value Function Realizability)**.** For every policy $\pi \in \Pi$, we have $V_0^\pi \in \mathcal{V}$ and $V_0^\pi + (1 + \frac{1}{\varphi})V_i^\pi \in \mathcal{V}$ for all $i = 0, 1, \ldots, I$.

Compared with prior LP-based analyses that only require realizability of an optimal value function [42], this is a stronger all-policy assumption. It is precisely this strengthening that rules out spurious saddle points in the restricted saddle problem.

With this stronger assumption on $\mathcal{V}$, we can show that a saddle point of L over $\mathcal{U} \times \mathcal{V}$ is optimal, as formally stated in the following:

Lemma 10. *Let $B \coloneqq 1 + \frac{1}{\varphi}$. Assume Assumption A and E, and let $\Lambda \coloneqq \{\lambda \in \mathbb{R}_+^I : \|\lambda\|_1 \le B\}$. Consider function classes $\mathcal{U} \subseteq \mathbb{R}_+^{|\mathcal{S}\times\mathcal{A}|}$ and $\mathcal{V} \subseteq \mathbb{R}^{|\mathcal{S}|}$. If $(\widehat{\mu}, ; \widehat{V}, \widehat{\lambda})$ is a saddle point of L over $\mathcal{U} \times (\mathcal{V} \times \Lambda)$, then $\widehat{\pi}$ is an optimal feasible policy.*

Proof. Let $\boldsymbol{e}_i \in \mathbb{R}^I$ denote the i-th standard basis vector. Using the reparameterized form of the Lagrangian, for any $\mu \in \mathcal{U}$, the Bellman equations imply

$$L(\mu; V_0^\pi, \mathbf{0}) = (1-\gamma)J_0(\pi),$$

and, since $V_0^\pi + BV_i^\pi$ is the value function of π under reward $r_0 + Br_i$,

$$L(\mu; V_0^\pi + BV_i^\pi, Be_i) = (1-\gamma)J_0(\pi) + B\big((1-\gamma)J_i(\pi) - \tau_i\big)$$

for each $i = 1, \ldots, I$.

Next, since $\mu^* = \mu^{\pi^*}$, for any $V \in \mathcal{V}$ and $\lambda \in \Lambda$,

$$L(\mu^*; V, \lambda) = (1-\gamma)J_0(\pi^*) + \sum_{i=1}^I \lambda_i\big((1-\gamma)J_i(\pi^*) - \tau_i\big).$$

Since π^* is feasible, the second term is nonnegative, and hence

$$L(\mu^*; V, \lambda) \ge (1-\gamma)J_0(\pi^*).$$

By the saddle-point property, choosing $(V, \lambda) = (V_0^{\widehat{\pi}}, 0)$ gives

$$(1-\gamma)J_0(\pi^*) \le L(\mu^*; \widehat{V}, \widehat{\lambda}) \le L(\widehat{\mu}, ; V_0^{\widehat{\pi}}, 0) = (1-\gamma)J_0(\widehat{\pi}).$$

Thus $\widehat{\pi}$ attains objective value at least that of the optimal feasible policy π^*.

It remains to show feasibility of $\widehat{\pi}$. Let λ^* be an optimal dual variable of the original CMDP. By Lemma 3, $\|\lambda^*\|_1 \le \frac{1}{\varphi} < B$. Define

$$m \coloneqq \min_{i=1,\ldots,I}\big((1-\gamma)J_i(\widehat{\pi}) - \tau_i\big),$$

and let $j \in \arg\min_{i=1,\ldots,I}\big((1-\gamma)J_i(\widehat{\pi}) - \tau_i\big)$.

By strong duality for the original CMDP and optimality of λ^*, we have

$$(1-\gamma)J_0(\pi^*) \ge (1-\gamma)J_0(\widehat{\pi}) + \sum_{i=1}^I \lambda_i^*\big((1-\gamma)J_i(\widehat{\pi}) - \tau_i\big) \ge (1-\gamma)J_0(\widehat{\pi}) + m\|\lambda^*\|_1.$$

On the other hand, applying the saddle-point property with $(V, \lambda) = (V_0^{\widehat{\pi}} + BV_j^{\widehat{\pi}}, B\boldsymbol{e}_j)$ yields

$$(1-\gamma)J_0(\pi^*) \le L(\widehat{\mu}, ; V_0^{\widehat{\pi}} + BV_j^{\widehat{\pi}}, B\boldsymbol{e}_j) = (1-\gamma)J_0(\widehat{\pi}) + Bm.$$

Combining the previous two displays, we obtain

$$m\|\lambda^*\|_1 \le Bm.$$

If $m < 0$, then dividing by m reverses the inequality and gives $\|\lambda^*\|_1 \ge B$, which contradicts $\|\lambda^*\|_1 < B$. Hence $m \ge 0$, that is,

$$(1-\gamma)J_i(\widehat{\pi}) \ge \tau_i, \qquad i = 1, \ldots, I.$$

So $\widehat{\pi}$ is feasible.

We have shown that $\widehat{\pi}$ is feasible and that $J_0(\widehat{\pi}) \ge J_0(\pi^*)$. Since π^* is an optimal feasible policy, it follows that $\widehat{\pi}$ is also optimal. □

With the result in the proposition above, we consider an algorithm that finds a saddle point $(\widehat{\mu}, \widehat{V}, \widehat{\lambda})$ that solves $\max_{\mu\in\mathcal{U}} \min_{V\in\mathcal{V}, \lambda\in\mathbb{R}_+^I} L(\mu, V, \lambda)$, then returns the policy $\widehat{\pi} = \pi(\widehat{\mu})$ extracted from $\widehat{\mu}$. Since computing the Lagrangian function requires the knowledge of the transition probability kernel P, we need to estimate the Lagrangian function. To facilitate the estimation, we change the measure

from μ to μ_D by multiplying the importance weight $w = \mu/\mu_D$ and express the Lagrangian in terms of w instead of μ as follows.

$$L(\boldsymbol{w}, \boldsymbol{V}, \boldsymbol{\lambda}) \coloneqq (1-\gamma)V(s_0) + \mathbb{E}_{\mu_D}[(w(s,a)(r(s,a)+\gamma[PV](s,a) - V(s))] + \sum_{i=1}^{I} \lambda_i(\langle \boldsymbol{\mu}, \boldsymbol{r}_i \rangle - \tau_i).$$

Analogous to Lemma 10, we can show that if a function class $\mathcal{W}$ for the MIW contains the optimal $w^* = \mu^*/\mu_D$ and a function class $\mathcal{V}$ contains $V_0^\pi + (1+\frac{1}{\varphi})V_i^\pi$ for all policies $\pi \in \Pi$, then a saddle point of $L(w, V, \lambda)$ over $\mathcal{W} \times \mathcal{V} \times \Lambda$ is optimal. In addition to Assumption E, we make the standard optimal-density-ratio realizability assumption on $\mathcal{W}$ (Assumption C).

With the realizability assumptions, we consider an algorithm that finds a saddle point of the estimated Lagrangian:

$$\begin{aligned} &\max_{w \in \mathcal{W}} \min_{V \in \mathcal{V},\, \lambda \in \Lambda} \widehat{L}(w; V, \lambda) \\ &:= (1-\gamma)V(s_0) + \frac{1}{n}\sum_{j=1}^{n} w(s_j, a_j)\Big(r_{\boldsymbol{\lambda}}(s_j, a_j) + \gamma V(s_j') - V(s_j)\Big) - \sum_{i=1}^{I} \lambda_i \tau_i. \end{aligned} \tag{11}$$

where $r_{\boldsymbol{\lambda}} = r_0 + \sum_{i=1}^{I} \lambda_i r_i$, and $\widehat{L}(w, V, \lambda)$ is an unbiased estimate of $L(w, V, \lambda)$. After finding a saddle point $(\widehat{w}, \widehat{V}, \widehat{\lambda})$, we need to extract policy from $\widehat{w}$. In general, the policy extraction from an importance weight $w \in \mathcal{W}$ can be done by the policy extraction from the corresponding occupancy measure $\mu = w \cdot \mu_D$ and using (3):

$$[\pi(w)](a|s) = \frac{w(s,a)\mu_D(s,a)}{\sum_{a'} w(s,a')\mu_D(s,a')} \text{ if } \sum_{a'} w(s,a')\mu_D(s,a') > 0, \quad \frac{1}{|\mathcal{A}|} \text{ otherwise.} \tag{12}$$

The policy $\pi(\widehat{w})$ extracted from a solution $(\widehat{w}, \widehat{V}, \widehat{\lambda})$ of the saddle point problem (11) gives the guarantee. We show the guarantee below.

Lemma 11. *Consider the empirical Lagrangian estimate*

$$\widehat{L}(w; V, \lambda) = (1-\gamma)V(s_0) + \frac{1}{n}\sum_{j=1}^{n} w(s_j, a_j)\Big(r_{\boldsymbol{\lambda}}(s_j, a_j) + \gamma V(s_j') - V(s_j)\Big) - \sum_{i=1}^{I} \lambda_i \tau_i.$$

where $r_{\boldsymbol{\lambda}} = r_0 + \sum_{i=1}^{I} \lambda_i r_i$. Given a function class $\mathcal{W}$ and $\mathcal{V}$ for w and V, with boundedness condition $\|w\|_\infty \le C$ for all $w \in \mathcal{W}$, $\|V\|_\infty \le \frac{1}{1-\gamma}$ for all $V \in \mathcal{V}$, we have with probability at least $1-\delta$ that, for all $w \in \mathcal{W}$, $V \in \mathcal{V}$ and $\lambda \in B\Delta^I$

$$\left|L(w; V, \lambda) - \widehat{L}(w; V, \lambda)\right| \le \epsilon_n$$

where $\epsilon_n = \mathcal{O}\Big(\Big(1 + \frac{C}{1-\gamma}(1+B)\Big)\sqrt{\log\big((I+1)|\mathcal{W}||\mathcal{V}|/\delta\big)/n}\Big)$.

Proof. Fix $w \in \mathcal{W}$, $V \in \mathcal{V}$, and $\lambda \in \{0, B\boldsymbol{e}_1, \dots, B\boldsymbol{e}_I\}$. Define

$$Z_j \coloneqq (1-\gamma)V(s_0) + w(s_j, a_j)\Big(r_0(s_j, a_j) + \sum_{i=1}^{I} \lambda_i r_i(s_j, a_j) + \gamma V(s_j') - V(s_j)\Big) - \sum_{i=1}^{I} \lambda_i \tau_i.$$

Then

$$\widehat{L}(w; V, \lambda) = \frac{1}{n}\sum_{j=1}^{n} Z_j, \qquad L(w; V, \lambda) = \mathbb{E}[Z_j].$$

We now bound $|Z_j|$ uniformly. Using $\|V\|_\infty \le \frac{1}{1-\gamma}$, we have

$$|(1-\gamma)V(s_0)| \le 1.$$

Next, since $\|w\|_\infty \le C$, $\|r_0\|_\infty \le 1$, $\|r_i\|_\infty \le 1$, and $\|\lambda\|_1 \le B$,

$$\left| r_0(s_j, a_j) + \sum_{i=1}^{I} \lambda_i r_i(s_j, a_j) \right| \le 1 + B.$$

Also,

$$|\gamma V(s'_j) - V(s_j)| \le \gamma \|V\|_\infty + \|V\|_\infty \le \frac{1+\gamma}{1-\gamma} \le \frac{2}{1-\gamma}.$$

Therefore,

$$\left| w(s_j, a_j) \Big(r_0(s_j, a_j) + \sum_{i=1}^{I} \lambda_i r_i(s_j, a_j) + \gamma V(s'_j) - V(s_j) \Big) \right| \le C \left(1 + B + \frac{2}{1-\gamma} \right).$$

Finally,

$$\left| \sum_{i=1}^{I} \lambda_i \tau_i \right| \le \|\lambda\|_1 \|\tau\|_\infty \le B \|\tau\|_\infty.$$

Hence,

$$|Z_j| \le 1 + C \left(1 + B + \frac{2}{1-\gamma} \right) + B\|\tau\|_\infty = \mathcal{O}\left(1 + \frac{C}{1-\gamma}(1+B) \right).$$

Since $Z_1, \ldots, Z_n$ are i.i.d. and uniformly bounded, Hoeffding's inequality implies that for any fixed triple (w, V, λ),

$$\left| L(w; V, \lambda) - \widehat{L}(w; V, \lambda) \right| = \left| \mathbb{E}[Z_j] - \frac{1}{n} \sum_{j=1}^{n} Z_j \right| \le \mathcal{O}\left(\left(1 + \frac{C}{1-\gamma}(1+B) \right) \sqrt{\frac{\log(1/\delta)}{n}} \right)$$

with probability at least $1 - \delta$.

Applying a union bound over all triples in $\mathcal{W} \times \mathcal{V} \times \{0, B\boldsymbol{e}_1, \ldots, B\boldsymbol{e}_I\}$ yields that, with probability at least $1 - \delta$, the above bound holds uniformly for all $w \in \mathcal{W}$, $V \in \mathcal{V}$, and $\lambda \in \{0, B\boldsymbol{e}_1, \ldots, B\boldsymbol{e}_I\}$:

$$\left| L(w; V, \lambda) - \widehat{L}(w; V, \lambda) \right| \le \mathcal{O}\left(\left(1 + \frac{C}{1-\gamma}(1+B) \right) \sqrt{\frac{\log\big((I+1)|\mathcal{W}||\mathcal{V}|/\delta\big)}{n}} \right).$$

Now, for any $\lambda \in B\Delta_I$, we can express $\lambda = B \sum_{i=1}^{I} \alpha_i \boldsymbol{e}_i$ for some $\alpha_0, \alpha_1, \ldots, \alpha_I \ge 0$ with $\sum_{i=0}^{I} \alpha_i = 1$. Since $L(w, V, \lambda)$ and $\widehat{L}(w, V, \lambda)$ are affine in λ, we have

$$L(w, V, \lambda) = \alpha_0 L(w, V, 0) + \sum_{i=1}^{I} \alpha_i L(w, V, \boldsymbol{e}_i), \quad \widehat{L}(w, V, \lambda) = \alpha_0 \widehat{L}(w, V, 0) + \sum_{i=1}^{I} \alpha_i \widehat{L}(w, V, \boldsymbol{e}_i),$$

and the uniform concentration inequality we obtained for $\lambda \in \{0, B\boldsymbol{e}_1, \ldots, B\boldsymbol{e}_I\}$ gives concentration inequality for all $\lambda \in B\Delta^I$. This completes the proof. □

Theorem 12. *Let $\mathcal{W}$ be a function class such that $\|w\|_\infty \le C$ and $\mathcal{V}$ be a function class such that $\|V\|_\infty \le \frac{1}{1-\gamma}$. Under Assumptions A, C, and E, if $(\widehat{w}; \widehat{V}, \widehat{\lambda})$ is a saddle point that solves*

$$\max_{w \in \mathcal{W}} \min_{V \in \mathcal{V},\, \lambda \in (1+\frac{1}{\varphi})\Delta^I} \widehat{L}(w; V, \lambda),$$

then the extracted policy $\widehat{\pi} = \pi(\widehat{w})$ satisfies, with probability at least $1 - \delta$,

$$J_0(\widehat{\pi}) \ge J_0(\pi^*) - \frac{2\varepsilon_n}{1-\gamma}, \qquad (1-\gamma) J_i(\widehat{\pi}) \ge \tau_i - \frac{4\varepsilon_n}{B}, \qquad i = 1, \ldots, I,$$

where

$$\varepsilon_n = \mathcal{O}\left(\left(1 + \frac{C}{1-\gamma}(1+B) \right) \sqrt{\frac{\log\big((I+1)|\mathcal{W}||\mathcal{V}|/\delta\big)}{n}} \right).$$

Proof. Let

$$\mathcal{E} := \left\{ \sup_{w\in\mathcal{W},\, V\in\mathcal{V},\, \lambda\in(1+\frac{1}{\varphi})\Delta^I} \left|L(w,V,\lambda) - \widehat{L}(w,V,\lambda)\right| \le \varepsilon_n \right\}.$$

By Lemma 11, $\mathbb{P}(\mathcal{E}) \ge 1-\delta$. We work under the event $\mathcal{E}$.

Let $w^* = \mu^{\pi^*}/\mu_D \in \mathcal{W}$, which exists by Assumption C. Since π^* is feasible, the corresponding occupancy measure μ^{π^*} satisfies

$$\langle \mu^{\pi^*}, r_i\rangle = (1-\gamma)J_i(\pi^*) \ge \tau_i, \qquad i = 1,\dots,I.$$

Moreover, since μ^{π^*} satisfies the Bellman flow constraint, for every $V \in \mathcal{V}$ and every $\lambda \in \Lambda_B$ we have

$$L(w^*,V,\lambda) = (1-\gamma)J_0(\pi^*) + \sum_{i=1}^I \lambda_i\big((1-\gamma)J_i(\pi^*) - \tau_i\big) \ge (1-\gamma)J_0(\pi^*).$$

Also, exactly as in the proof of Lemma 10, for any $w \in \mathcal{W}$ and any $\lambda \in \Lambda_B$,

$$L\big(w, V_\lambda^{\pi(w)}, \lambda\big) = (1-\gamma)J_0(\pi(w)) + \sum_{i=1}^I \lambda_i\big((1-\gamma)J_i(\pi(w)) - \tau_i\big),$$

where $\pi(w)$ is the policy extracted from w.

Now let $\widehat{\pi} = \pi(\widehat{w})$. For any $\lambda \in \Lambda_B$, since $(\widehat{w}, \widehat{V}, \widehat{\lambda})$ is a saddle point of $\widehat{L}$, we obtain

$$\begin{aligned}
(1-\gamma)J_0(\pi^*) &\le L(w^*, \widehat{V}, \widehat{\lambda}) \\
&\le \widehat{L}(w^*, \widehat{V}, \widehat{\lambda}) + \varepsilon_n \\
&\le \widehat{L}(\widehat{w}, \widehat{V}, \widehat{\lambda}) + \varepsilon_n \\
&\le \widehat{L}(\widehat{w}, V_\lambda^{\widehat{\pi}}, \lambda) + \varepsilon_n \\
&\le L(\widehat{w}, V_\lambda^{\widehat{\pi}}, \lambda) + 2\varepsilon_n \\
&= (1-\gamma)J_0(\widehat{\pi}) + \sum_{i=1}^I \lambda_i\big((1-\gamma)J_i(\widehat{\pi}) - \tau_i\big) + 2\varepsilon_n.
\end{aligned}$$

Hence, for all $\lambda \in \Lambda_B$,

$$(1-\gamma)J_0(\pi^*) \le (1-\gamma)J_0(\widehat{\pi}) + \sum_{i=1}^I \lambda_i\big((1-\gamma)J_i(\widehat{\pi}) - \tau_i\big) + 2\varepsilon_n.$$

Setting $\lambda = 0$ gives

$$(1-\gamma)J_0(\pi^*) \le (1-\gamma)J_0(\widehat{\pi}) + 2\varepsilon_n,$$

so

$$J_0(\widehat{\pi}) \ge J_0(\pi^*) - \frac{2\varepsilon_n}{1-\gamma}.$$

Next, fix any $i \in \{1,\dots,I\}$ and choose $\lambda = Be_i \in \Lambda_B$. Then

$$(1-\gamma)J_0(\pi^*) \le (1-\gamma)J_0(\widehat{\pi}) + B\big((1-\gamma)J_i(\widehat{\pi}) - \tau_i\big) + 2\varepsilon_n.$$

Using the objective bound just proved,

$$(1-\gamma)J_0(\pi^*) - (1-\gamma)J_0(\widehat{\pi}) \le 2\varepsilon_n,$$

and therefore

$$B\big((1-\gamma)J_i(\widehat{\pi}) - \tau_i\big) \ge -4\varepsilon_n.$$

Equivalently,

$$(1-\gamma)J_i(\widehat{\pi}) \ge \tau_i - \frac{4\varepsilon_n}{B}, \qquad i = 1,\dots,I.$$

This completes the proof. □

C Unknown data distribution case

C.1 Proof of Proposition 2

For the proof, recall the reparameterized form of the Lagrangian:

$$L(\boldsymbol{\mu}, \pi; \boldsymbol{Q}, \boldsymbol{\lambda}) = (1-\gamma)\langle \boldsymbol{Q}_\pi, \boldsymbol{d}_0 \rangle + \langle \boldsymbol{\mu}, \boldsymbol{r}_{\boldsymbol{\lambda}} + \gamma \boldsymbol{P}\boldsymbol{Q}_\pi - \boldsymbol{Q} \rangle - \sum_{i=1}^{I} \lambda_i \tau_i.$$

Proof. Consider a single-constraint CMDP with $\mathcal{S} = \{s_0, l_1, r_1, l_2, r_2\}$, $\mathcal{A} = \{L, R\}$, $\gamma = \frac{1}{2}$ and $d_0 = \delta_{s_0}$. The transition kernel is $P(l_1 \mid s_0, L) = 1$, $P(r_1 \mid s_0, R) = 1$, $P(l_2 \mid r_1, L) = 1$, $P(r_2 \mid r_1, R) = 1$, and $P(l_1 \mid l_1, a) = P(l_2 \mid l_2, a) = P(r_2 \mid r_2, a) = 1$ for all $a \in \mathcal{A}$, such that the states l_1, l_2, r_2 are absorbing under both actions. The reward functions are $r_0(l_1, a) = 1$, $r_0(l_2, a) = 2$ for all $a \in \mathcal{A}$ and $r_0(s, a) = 0$ for all other state-action pairs. Set $r_1 = r_0$ and $\tau_1 = 0$, such that every policy is feasible.

The optimal action-value function Q_0^* for reward r_0 can be computed by solving the Bellman optimality equations: $Q_0^*(l_1, a) = 2$, $Q_0^*(l_2, a) = 4$, $Q_0^*(r_2, a) = 0$ for all $a \in \mathcal{A}$. Also, $Q_0^*(r_1, L) = 2$, $Q_0^*(r_1, R) = 0$, $Q_0^*(s_0, L) = 1$, $Q_0^*(s_0, R) = 1$. $Q_0^*(s_0, L) = 1$, $Q_0^*(l_1, L) = 2$.

Now consider two policies π^* and $\widehat{\pi}$ defined as follows: $\pi^*(s) = L$ for all $s \in \mathcal{S}$ and $\widehat{\pi}(s_0) = \widehat{\pi}(r_1) = R$, $\widehat{\pi}(l_1) = \widehat{\pi}(l_2) = \widehat{\pi}(r_2) = L$. The policy π^* is optimal, because it moves from s_0 to l_1 and then stays there forever, achieving $J_0(\pi^*) = 1$. By contrast, $\widehat{\pi}$ moves from s_0 to r_1, then from r_1 to r_2, and all rewards on this trajectory are zero. Hence $J_0(\widehat{\pi}) = 0$. Therefore $\widehat{\pi}$ is not optimal.

Let $\mu^* = \mu^{\pi^*}$ be the occupancy measure of π^*. Since under π^* the process is at (s_0, L) at time 0, and at (l_1, L) at every time $t \geq 1$, $\mu^*(s_0, L) = (1-\gamma) = \frac{1}{2}$, $\mu^*(l_1, L) = (1-\gamma)\sum_{t=1}^{\infty} \gamma^t = \frac{1}{2}$, and all other coordinates of μ^* are zero.

Define $\mathcal{U} = \{\mu^*\}$, $\Pi = \{\pi^*, \widehat{\pi}\}$, and $\mathcal{Q} = \{(1+\alpha)Q_0^* : \alpha \geq 0\}$. Then, since $r_1 = r_0$, for every $\alpha \geq 0$, $Q^*_{r_0 + \alpha r_1} = (1+\alpha)Q_0^* \in \mathcal{Q}$. Hence, the realizability properties in item (i) and (ii) are satisfied.

We now show that $(\mu^*, \widehat{\pi}; Q_0^*, 0)$ is a saddle point. First, we verify primal optimality at $(Q_0^*, 0)$. Since μ^* is supported only on (s_0, L) and (l_1, L), and since for both $\pi \in \Pi$ we have $Q^*_{0,\pi}(s_0) = 1$ and $Q^*_{0,\pi}(l_1) = 2$, the Bellman residuals on the support of μ^* are

$$r_0(s_0, L) + \gamma Q^*_{0,\pi}(l_1) - Q_0^*(s_0, L) = 0 + \frac{1}{2} \cdot 2 - 1 = 0,$$

and

$$r_0(l_1, L) + \gamma Q^*_{0,\pi}(l_1) - Q_0^*(l_1, L) = 1 + \frac{1}{2} \cdot 2 - 2 = 0.$$

Hence, for every $\pi \in \Pi$,

$$L(\mu^*, \pi; Q_0^*, 0) = (1-\gamma)Q^*_{0,\pi}(s_0) = \frac{1}{2}.$$

Because $\mathcal{U}$ is a singleton, both (μ^*, π^*) and $(\mu^*, \widehat{\pi})$ maximize $L(\mu, \pi; Q_0^*, 0)$ over $\mathcal{U} \times \Pi$. In particular, $(\mu^*, \widehat{\pi})$ is a primal maximizer.

Second, we verify dual optimality at $(\mu^*, \widehat{\pi})$. Fix arbitrary $\alpha \geq 0$ and $\lambda \geq 0$, and consider $Q_\alpha^* = (1+\alpha)Q_0^*$. Since $\widehat{\pi}(l_1) = L$, we have

$$Q^*_{\alpha,\widehat{\pi}}(s_0) = Q_\alpha^*(s_0, R) = 1 + \alpha, \qquad Q^*_{\alpha,\widehat{\pi}}(l_1) = Q_\alpha^*(l_1, L) = 2(1+\alpha).$$

Again using that μ^* is supported only on (s_0, L) and (l_1, L), we get

$$(1+\lambda)r_0(s_0, L) + \gamma Q^*_{\alpha,\widehat{\pi}}(l_1) - Q_\alpha^*(s_0, L) = 0 + \frac{1}{2} \cdot 2(1+\alpha) - (1+\alpha) = 0,$$

and

$$(1+\lambda)r_0(l_1, L) + \gamma Q^*_{\alpha,\widehat{\pi}}(l_1) - Q_\alpha^*(l_1, L) = (1+\lambda) + \frac{1}{2} \cdot 2(1+\alpha) - 2(1+\alpha) = \lambda - \alpha.$$

Therefore,

$$L(\mu^*, \widehat{\pi}; Q^*_\alpha, \lambda) = (1-\gamma)Q^*_{\alpha,\widehat{\pi}}(s_0) + \mu^*(l_1, L)(\lambda - \alpha) = \frac{1}{2}(1+\alpha) + \frac{1}{2}(\lambda - \alpha) = \frac{1}{2}(1+\lambda).$$

Thus

$$L(\mu^*, \widehat{\pi}; Q^*_\alpha, \lambda) \geq \frac{1}{2} = L(\mu^*, \widehat{\pi}; Q^*_0, 0) \qquad \text{for all } \alpha \geq 0,\ \lambda \geq 0.$$

So $(Q^*_0, 0)$ is a dual minimizer over $\mathcal{Q} \times \mathbb{R}_+$.

Combining the previous two parts, $(\mu^*, \widehat{\pi}; Q^*_0, 0)$ is a saddle point of the reparameterized Lagrangian over $(\mathcal{U} \times \Pi) \times (\mathcal{Q} \times \mathbb{R}_+)$. But $\widehat{\pi}$ is not optimal, since $J_0(\widehat{\pi}) = 0 < 1 = J_0(\pi^*)$. □

C.2 Analysis of saddle points of lagrangian function

Recall the Lagrangian function:

$$\begin{aligned} L(\boldsymbol{\mu}, \pi; \boldsymbol{Q}, \boldsymbol{\lambda}) &= (1-\gamma)\langle \boldsymbol{Q}_\pi, \boldsymbol{d}_0 \rangle + \langle \boldsymbol{\mu}, \boldsymbol{r}_{\boldsymbol{\lambda}} + \gamma \boldsymbol{P}\boldsymbol{Q}_\pi - \boldsymbol{Q} \rangle - \sum_{i=1}^I \lambda_i \tau_i \\ &= \langle \boldsymbol{Q}_\pi, (1-\gamma)\boldsymbol{d}_0 + \gamma \boldsymbol{P}^\top \boldsymbol{\mu} \rangle + \langle \boldsymbol{\mu}, \boldsymbol{r}_{\boldsymbol{\lambda}} - \boldsymbol{Q} \rangle - \sum_{i=1}^I \lambda_i \tau_i \end{aligned}$$

Change of variable $\boldsymbol{w} = \boldsymbol{\mu}/\boldsymbol{\mu}_D$ gives the equivalent form of the Lagrangian:

$$L(\boldsymbol{w}, \pi; \boldsymbol{Q}, \boldsymbol{\lambda}) = (1-\gamma)\langle \boldsymbol{Q}_\pi, \boldsymbol{d}_0 \rangle + \langle \boldsymbol{w}, \boldsymbol{r}_{\boldsymbol{\lambda}} + \gamma \boldsymbol{P}\boldsymbol{Q}_\pi - \boldsymbol{Q} \rangle_{\boldsymbol{\mu}_D} - \sum_{i=1}^I \lambda_i \tau_i,$$

where $\langle w, f \rangle_{\mu_D} = \sum_{s,a} w(s,a)\mu_D(s,a)f(s,a)$. We first show useful identities of the Lagrangian function, which will be used in the proof of Lemma 14.

Lemma 13. *For every π, Q, and λ, we have*

$$L(w^\pi, \pi; \boldsymbol{Q}, \boldsymbol{\lambda}) = (1-\gamma)J_0(\pi) + \sum_{i=1}^I \lambda_i((1-\gamma)J_i(\pi) - \tau_i),$$

and

$$L(w, \pi; \boldsymbol{Q}^\pi_{\boldsymbol{\lambda}}, \boldsymbol{\lambda}) = (1-\gamma)J_0(\pi) + \sum_{i=1}^I \lambda_i((1-\gamma)J_i(\pi) - \tau_i),$$

where $\boldsymbol{Q}^\pi_{\boldsymbol{\lambda}}$ is the value function of policy π with respect to the reward $\boldsymbol{r}_{\boldsymbol{\lambda}}$.

Proof. To show the first identity, note that

$$\begin{aligned} L(w^\pi, \pi; \boldsymbol{Q}, \boldsymbol{\lambda}) &= (1-\gamma)\langle \boldsymbol{Q}_\pi, \boldsymbol{d}_0 \rangle + \langle \boldsymbol{w}^\pi, \boldsymbol{r}_{\boldsymbol{\lambda}} + \gamma \boldsymbol{P}\boldsymbol{Q}_\pi - \boldsymbol{Q} \rangle_{\boldsymbol{\mu}_D} - \sum_{i=1}^I \lambda_i \tau_i \\ &= (1-\gamma)\langle \boldsymbol{Q}_\pi, \boldsymbol{d}_0 \rangle + \langle \boldsymbol{\mu}^\pi, \boldsymbol{r}_{\boldsymbol{\lambda}} + \gamma \boldsymbol{P}\boldsymbol{Q}_\pi - \boldsymbol{Q} \rangle - \sum_{i=1}^I \lambda_i \tau_i \\ &= \langle \boldsymbol{Q}_\pi, (1-\gamma)\boldsymbol{d}_0 + \gamma \boldsymbol{P}^\top \boldsymbol{\mu}^\pi \rangle - \langle \boldsymbol{\mu}^\pi, \boldsymbol{Q} \rangle + \langle \boldsymbol{\mu}^\pi, \boldsymbol{r}_{\boldsymbol{\lambda}} \rangle - \sum_{i=1}^I \lambda_i \tau_i. \end{aligned}$$

By the Bellman flow equation, we have $(1-\gamma)\boldsymbol{d}_0 + \gamma \boldsymbol{P}^\top \boldsymbol{\mu}^\pi = \boldsymbol{d}^\pi$ where $\boldsymbol{d}^\pi$ is the state occupancy measure of policy π. Since

$$\langle \boldsymbol{Q}_\pi, \boldsymbol{d}^\pi \rangle = \sum_{s,a} Q(s,a)\pi(a|s)d^\pi(s) = \sum_{s,a} Q(s,a)\mu^\pi(s,a) = \langle \boldsymbol{\mu}^\pi, \boldsymbol{Q} \rangle,$$

we have

$$L(w^\pi, \pi; \boldsymbol{Q}, \boldsymbol{\lambda}) = \langle \boldsymbol{\mu}^\pi, \boldsymbol{r}_{\boldsymbol{\lambda}} \rangle - \sum_{i=1}^I \lambda_i \tau_i = (1-\gamma)J_0(\pi) + \sum_{i=1}^I \lambda_i((1-\gamma)J_i(\pi) - \tau_i),$$

proving the first identity. To show the second identity, note that if $\boldsymbol{Q}$ is a value function of some policy π with respect to the reward $\boldsymbol{r}_{\boldsymbol{\lambda}}$, then the Bellman equation implies that $\boldsymbol{r}_{\boldsymbol{\lambda}} + \gamma \boldsymbol{P}\boldsymbol{Q}_\pi - \boldsymbol{Q} = 0$, which gives the second identity □

The following lemma shows objective bound and feasibility bound of a near-saddle point of L over a restricted domain.

Lemma 14. *Assume Assumption A,B, C and D. Suppose $(\widehat{w}, \widehat{\pi}; \widehat{Q}, \widehat{\lambda})$ is a ξ-near saddle point of L over $(\mathcal{W} \times \Pi) \times (\mathcal{Q} \times \Lambda_B)$, where $B := 1 + \frac{1}{\varphi}$ and $\Lambda_B := B\Delta^I$. Then*

$$(1-\gamma)J_0(\widehat{\pi}) \geq (1-\gamma)J_0(\pi^*) - \xi \qquad \textit{and} \qquad (1-\gamma)J_i(\widehat{\pi}) \geq \tau_i - \xi, \quad i = 1, \ldots, I,$$

where $\pi^* \in \Pi$ is an optimal feasible policy with $w^{\pi^*} \in \mathcal{W}$.

Proof. Fix a comparator policy $\pi \in \Pi$ that is feasible and satisfies $w^\pi \in \mathcal{W}$. Since $(\widehat{w}, \widehat{\pi}; \widehat{Q}, \widehat{\lambda})$ is a ξ-near saddle point, applying the near-saddle inequality with (w^{π^*}, π^*) and $(Q_0^{\widehat{\pi}}, 0)$ gives

$$L(w^{\pi^*}, \pi^*; \widehat{Q}, \widehat{\lambda}) \leq L(\widehat{w}, \widehat{\pi}; Q_0^{\widehat{\pi}}, 0) + \xi.$$

Using the first identity in Lemma 13 and the fact that π^* is feasible, we have

$$L(w^{\pi^*}, \pi^*; \widehat{Q}, \widehat{\lambda}) = (1-\gamma)J_0(\pi^*) + \sum_{i=1}^I \widehat{\lambda}_i((1-\gamma)J_i(\pi^*) - \tau_i) \geq (1-\gamma)J_0(\pi^*).$$

On the other hand, the second identity in Lemma 13 implies

$$L(\widehat{w}, \widehat{\pi}; Q_0^{\widehat{\pi}}, 0) = (1-\gamma)J_0(\widehat{\pi}).$$

Hence, the saddle inequality gives

$$(1-\gamma)J_0(\widehat{\pi}) \geq (1-\gamma)J_0(\pi^*) - \xi,$$

which proves near-optimality of the objective.

For feasibility, let $(\mu^*, \pi^*; Q^*, \lambda^*)$ be a saddle point of the unrestricted population Lagrangian for the original CMDP. By Lemma 3, $\|\lambda^*\|_1 \leq \varphi^{-1}$. Define

$$m := \min_{i \in [I]} \big((1-\gamma)J_i(\widehat{\pi}) - \tau_i\big).$$

Since $w^{\widehat{\pi}}$ is admissible for the unrestricted primal problem, although it might not be in $\mathcal{W}$, we have

$$\begin{aligned}
(1-\gamma)J_0(\pi^*) &= L(\mu^*, \pi^*; Q^*, \lambda^*) \\
&\geq L(w^{\widehat{\pi}}, \widehat{\pi}; Q^*, \lambda^*) \\
&= (1-\gamma)J_0(\widehat{\pi}) + \sum_{i=1}^I \lambda_i^*((1-\gamma)J_i(\widehat{\pi}) - \tau_i) \\
&\geq (1-\gamma)J_0(\widehat{\pi}) + m\|\lambda^*\|_1,
\end{aligned}$$

where the last equality is by the first identity in Lemma 13, and the last inequality is by the definition of m and the fact that $\lambda_i^* \geq 0$. Rearranging, we get

$$(1-\gamma)\big(J_0(\pi^*) - J_0(\widehat{\pi})\big) \geq m\|\lambda^*\|_1.$$

Now choose $j \in \arg\min_{i \in [I]} \big((1-\gamma)J_i(\widehat{\pi}) - \tau_i\big)$, so that $m = (1-\gamma)J_j(\widehat{\pi}) - \tau_j$. Apply the near-saddle inequality with (w^{π^*}, π^*) and $(Q_{0,j}^{\widehat{\pi}}, Be_j)$:

$$L(w^{\pi^*}, \pi^*; \widehat{Q}, \widehat{\lambda}) \leq L(\widehat{w}, \widehat{\pi}; Q_{0,j}^{\widehat{\pi}}, Be_j) + \xi.$$

As before, feasibility of π^* implies

$$L(w^{\pi^*}, \pi^*; \widehat{Q}, \widehat{\lambda}) = (1-\gamma)J_0(\pi^*) + \sum_{i=1}^I \widehat{\lambda}_i((1-\gamma)J_i(\pi^*) - \tau_i) \geq (1-\gamma)J_0(\pi^*),$$

where the first inequality uses the first identity in Lemma 13, while the second identity in Lemma 13 gives

$$L(\widehat{w},\widehat{\pi};Q^{\widehat{\pi}}_{0,j},Be_j)=(1-\gamma)J_0(\widehat{\pi})+B\big((1-\gamma)J_j(\widehat{\pi})-\tau_j\big)=(1-\gamma)J_0(\widehat{\pi})+Bm.$$

Combining and rearranging, we get

$$(1-\gamma)\big(J_0(\pi^*)-J_0(\widehat{\pi})\big)\le Bm+\xi.$$

Combining with $(1-\gamma)(J_0(\pi^*)-J_0(\widehat{\pi}))\ge m\|\lambda^*\|_1$ obtained previously, we get

$$m\|\lambda^*\|_1\le Bm+\xi.$$

If $m\ge 0$, then $(1-\gamma)J_i(\widehat{\pi})\ge\tau_i$ for all i, so there is nothing to prove. If $m<0$, then

$$m(B-\|\lambda^*\|_1)\ge-\xi.$$

Since $B=1+\varphi^{-1}$ and $\|\lambda^*\|_1\le\varphi^{-1}$, we have $B-\|\lambda^*\|_1\ge 1$, and therefore $m\ge-\xi$. Thus

$$(1-\gamma)J_i(\widehat{\pi})\ge\tau_i-\xi,\qquad i=1,\dots,I.$$

□

Proof of Lemma 4. This is a direct consequence of Lemma 14 with $\xi=0$. □

Theorem 15. *Assume Assumptions A, B, C and D. Let $B\coloneqq 1+\frac{1}{\varphi}$ where φ is the Slater constant defined in Assumption A. Let $\pi^*\in\Pi$ be an optimal feasible policy assumed in Assumption C such that $w^{\pi^*}\in\mathcal{W}$. If $(\widehat{w},\widehat{\pi};\widehat{Q},\widehat{\lambda})$ is a saddle point of*

$$\max_{w\in\mathcal{W},\ \pi\in\Pi}\ \min_{Q\in\mathcal{Q},\ \lambda\in\Lambda_B}\ \widehat{L}(w,\pi;Q,\lambda),$$

where $\Lambda_B=B\Delta^I$, then with probability at least $1-\delta$,

$$(1-\gamma)J_0(\widehat{\pi})\ge(1-\gamma)J_0(\pi^*)-\varepsilon_n,\qquad(1-\gamma)J_i(\widehat{\pi})\ge\tau_i-\varepsilon_n,\quad i=1,\dots,I,$$

where

$$\varepsilon_n=\mathcal{O}\left(\big(C^*(1+B+\frac{1}{1-\gamma})+B\big)\sqrt{\frac{\log((I+1)|\mathcal{W}|\,|\Pi|\,|\mathcal{Q}|/\delta)}{n}}\right).$$

We first prove the corresponding concentration bound for the Lagrangian estimate defined in (10):

$$\widehat{L}(w,\pi;Q,\boldsymbol{\lambda})\coloneqq(1-\gamma)Q(s_0,\pi)+\frac{1}{n}\sum_{j=1}^{n}w(s_j,a_j)(r_{\boldsymbol{\lambda}}(s_j,a_j)+\gamma Q(s'_j,\pi)-Q(s_j,a_j))-\sum_{i=1}^{I}\lambda_i\tau_i,$$

where $r_{\boldsymbol{\lambda}}(s,a)=r_0(s,a)+\sum_{i=1}^{I}\lambda_i r_i(s,a)$.

Lemma 16. *Consider the empirical Lagrangian estimate*

$$\widehat{L}(w,\pi;Q,\boldsymbol{\lambda})\coloneqq(1-\gamma)Q(s_0,\pi)+\frac{1}{n}\sum_{j=1}^{n}w(s_j,a_j)(r_{\boldsymbol{\lambda}}(s_j,a_j)+\gamma Q(s'_j,\pi)-Q(s_j,a_j))-\sum_{i=1}^{I}\lambda_i\tau_i.$$

Given a function class $\mathcal{W}$, Π and $\mathcal{Q}$, with boundedness conditions $\|w\|_\infty\le C$ for all $w\in\mathcal{W}$, $\|Q\|_\infty\le\frac{1}{1-\gamma}$ for all $Q\in\mathcal{Q}$, we have with probability at least $1-\delta$ that, for all $w\in\mathcal{W}$, $\pi\in\Pi$, $Q\in\mathcal{Q}$ and $\lambda\in B\Delta^I$,

$$|L(w,\pi;Q,\lambda)-\widehat{L}(w,\pi;Q,\lambda)|\le\varepsilon_n,$$

where

$$\varepsilon_n=\mathcal{O}\left(\big(C(1+B+\frac{1}{1-\gamma})+B\big)\sqrt{\frac{\log((I+1)|\mathcal{W}|\,|\Pi|\,|\mathcal{Q}|/\delta)}{n}}\right).$$

On the same event, the bound also holds uniformly for all $w\in\mathcal{W}$, $\pi\in\Pi$, $Q\in\operatorname{conv}(\mathcal{Q})$, and $\lambda\in B\Delta^I$.

Proof. Fix $w \in \mathcal{W}$, $\pi \in \Pi$, $Q \in \mathcal{Q}$, and $\boldsymbol{\lambda} \in \{0, B\boldsymbol{e}_1, \ldots, B\boldsymbol{e}_I\}$. Since the term $(1-\gamma)Q(s_0, \pi)$ is deterministic, write

$$Z_j \coloneqq w(s_j, a_j)(r_{\boldsymbol{\lambda}}(s_j, a_j) + \gamma Q(s'_j, \pi) - Q(s_j, a_j)) - \sum_{i=1}^{I} \lambda_i \tau_i.$$

Then

$$L(w, \pi; Q, \lambda) - \widehat{L}(w, \pi; Q, \lambda) = \frac{1}{n} \sum_{j=1}^{n} (\mathbb{E}[Z_j] - Z_j).$$

Using $r_0, r_i, \tau_i \in [0, 1]$, $\|w\|_\infty \leq C$, $\|\boldsymbol{\lambda}\|_1 \leq B$, and $\|Q\|_\infty \leq \frac{1}{1-\gamma}$, we have

$$|Z_j| \leq C(1 + B + 2\frac{1}{1-\gamma}) + B = \mathcal{O}\big(C(1 + B + \frac{1}{1-\gamma}) + B\big).$$

Hence, Hoeffding's inequality gives

$$\big|L(w, \pi; Q, \lambda) - \widehat{L}(w, \pi; Q, \lambda)\big| \leq \mathcal{O}\left(\big(C(1 + B + \frac{1}{1-\gamma}) + B\big)\sqrt{\frac{\log(1/\delta)}{n}}\right).$$

Applying a union bound over $\mathcal{W} \times \Pi \times \mathcal{Q} \times \{0, B\boldsymbol{e}_1, \ldots, B\boldsymbol{e}_I\}$, we get that with probability at least $1 - \delta$,

$$\big|L(w, \pi; Q, \boldsymbol{\lambda}) - \widehat{L}(w, \pi; Q, \boldsymbol{\lambda})\big| \leq \mathcal{O}\left(\big(C(1 + B + \frac{1}{1-\gamma}) + B\big)\sqrt{\frac{\log((I+1)|\mathcal{W}|\,|\Pi|\,|\mathcal{Q}|/\delta)}{n}}\right).$$

Now, for $\boldsymbol{\lambda} \in B\Delta^I$, note that it can be expressed as $\boldsymbol{\lambda} = B\sum_{i=1}^{I} \alpha_i \boldsymbol{e}_i$ for some $\alpha_0, \alpha_1, \ldots, \alpha_I \geq 0$ with $\sum_{i=0}^{I} \alpha_i = 1$. Since $L(w, \pi; Q, \lambda)$ and $\widehat{L}(w, \pi; Q, \lambda)$ are affine in λ, we have

$$L(w, \pi; Q, \lambda) = \alpha_0 L(w, \pi; Q, 0) + \sum_{i=1}^{I} \alpha_i L(w, \pi; Q, B\boldsymbol{e}_i),$$

and

$$\widehat{L}(w, \pi; Q, \lambda) = \alpha_0 \widehat{L}(w, \pi; Q, 0) + \sum_{i=1}^{I} \alpha_i \widehat{L}(w, \pi; Q, B\boldsymbol{e}_i).$$

The triangle inequality and $\sum_{i=0}^{I} \alpha_i = 1$ therefore extend the same uniform concentration bound to every $\lambda \in B\Delta^I$. To extend the bound in the critic argument, let $\widetilde{Q} = \sum_{k=1}^{K} \beta_k Q_k$ be any finite convex combination of $Q_k \in \mathcal{Q}$, with $\beta_k \geq 0$ and $\sum_{k=1}^{K} \beta_k = 1$. For fixed w, π, and λ, both $L(w, \pi; Q, \lambda)$ and $\widehat{L}(w, \pi; Q, \lambda)$ are affine in Q. Consequently, on the uniform concentration event above,

$$\begin{aligned}
&\big|L(w, \pi; \widetilde{Q}, \lambda) - \widehat{L}(w, \pi; \widetilde{Q}, \lambda)\big| \\
&\quad = \left|\sum_{k=1}^{K} \beta_k \big(L(w, \pi; Q_k, \lambda) - \widehat{L}(w, \pi; Q_k, \lambda)\big)\right| \\
&\quad \leq \sum_{k=1}^{K} \beta_k \big|L(w, \pi; Q_k, \lambda) - \widehat{L}(w, \pi; Q_k, \lambda)\big| \leq \varepsilon_n.
\end{aligned}$$

This holds simultaneously for every such convex combination, so no additional union bound or convexity assumption on $\mathcal{Q}$ is required. This completes the proof.

□

Now, we are ready to prove the theorem.

Proof of Theorem 15. On the event of Lemma 16, we have

$$\sup_{w,\pi,Q,\lambda\in B\Delta^I} |L(w,\pi;Q,\lambda) - \widehat{L}(w,\pi;Q,\lambda)| \le \varepsilon_n.$$

Since $(\widehat{w},\widehat{\pi};\widehat{Q},\widehat{\lambda})$ is a saddle point of $\widehat{L}$, Lemma 23 implies that it is a $2\varepsilon_n$-near saddle point of L over $(\mathcal{W}\times\Pi)\times(\mathcal{Q}\times B\Delta^I)$. Applying Lemma 14 with $\xi = 2\varepsilon_n$ gives

$$(1-\gamma)J_0(\widehat{\pi}) \ge (1-\gamma)J_0(\pi^*) - 2\varepsilon_n, \qquad (1-\gamma)J_i(\widehat{\pi}) \ge \tau_i - 2\varepsilon_n, \quad i = 1,\dots,I.$$

Absorbing the factor 2 into the $\mathcal{O}(\cdot)$ term completes the proof. □

C.3 Analysis of primal-dual algorithm

For the following theorem statement, we use the notation $\epsilon^{\pi}_{\mathrm{opt}}(T)$ to denote the average regret of the no-regret policy optimization oracle for policy π after T iterations, and $\epsilon^{w}_{\mathrm{opt}}(T)$ to denote the average regret of the no-regret linear optimization oracle for the w-player after T iterations. That is,

$$\epsilon^{\pi}_{\mathrm{opt}}(T) = \max_{\pi'\in\Pi}\frac{1}{T}\sum_{t=1}^{T} L(w_t,\pi';Q_t,\lambda_t) - L(w_t,\pi_t;Q_t,\lambda_t),$$

and

$$\epsilon^{w}_{\mathrm{opt}}(T) = \max_{w'\in\mathcal{W}}\frac{1}{T}\sum_{t=1}^{T} L(w',\pi_t;Q_t,\lambda_t) - L(w_t,\pi_t;Q_t,\lambda_t).$$

Theorem 17. *Under Assumptions A, B, C and D, running Algorithm 1 with dual bound $B = 1+\frac{1}{\varphi}$ and T large enough so that $\epsilon^{\pi^*}_{\mathrm{opt}}(T) \le \varepsilon_n$ and $\frac{2(C^*)^2}{\sqrt{T}} + \frac{(1+\frac{1}{\varphi}+\frac{2}{1-\gamma})}{2\sqrt{T}} \le \varepsilon_n$, and $\eta = 1/\sqrt{T}$, we have with probability at least $1-\delta$ that*

$$(1-\gamma)J_0(\widehat{\pi}) \ge (1-\gamma)J_0(\pi^*) - \varepsilon_n, \quad (1-\gamma)J_i(\widehat{\pi}) \ge \tau_i - \varepsilon_n, \;\; i = 1,\dots,I.$$

where $\varepsilon_n = \mathcal{O}(C^*(\frac{1}{\varphi}+\frac{1}{1-\gamma})\sqrt{\frac{\log(I|\mathcal{W}|\,|\Pi|\,|\mathcal{Q}|/\delta)}{n}})$.

Proof. Define $L(\pi,\lambda) = (1-\gamma)J_0(\pi) + \sum_{i=1}^{I}\lambda_i((1-\gamma)J_i(\pi)-\tau_i)$. Let $\widehat{\pi} = \mathrm{Unif}\{\pi_1,\dots,\pi_T\}$ be the policy returned by the algorithm and let $\widehat{\lambda} = \frac{1}{T}\sum_{t=1}^{T}\lambda_t$. Since $\widehat{\pi}$ samples one of $\pi_1,\dots,\pi_T$ uniformly at the beginning of the episode, we have

$$\mathcal{L}(\widehat{\pi},\lambda) = \frac{1}{T}\sum_{t=1}^{T}\mathcal{L}(\pi_t,\lambda).$$

Also, by linearity in λ,

$$\mathcal{L}(\pi^*,\widehat{\lambda}) = \frac{1}{T}\sum_{t=1}^{T}\mathcal{L}(\pi^*,\lambda_t).$$

By Lemma 25, it is enough to show that $L(\pi^*, \widehat{\lambda}) \le L(\widehat{\pi}, \lambda) + \varepsilon_n$ for all $\lambda \in (1 + \frac{1}{\varphi})\Delta^I$. Consider the following decomposition:

$$\begin{aligned}
L(\pi^*, \widehat{\lambda}) - L(\widehat{\pi}, \lambda) &= \frac{1}{T}\sum_{t=1}^{T} L(\pi^*, \lambda_t) - L(\pi_t, \lambda) \\
&= \frac{1}{T}\sum_{t=1}^{T} L(w^*, \pi^*; Q_t, \lambda_t) - L(w_t, \pi_t; Q_\lambda^{\pi_t}, \lambda) \\
&= \frac{1}{T}\sum_{t=1}^{T} L(w^*, \pi^*; Q_t, \lambda_t) - L(w^*, \pi_t; Q_t, \lambda_t) \\
&\qquad + \frac{1}{T}\sum_{t=1}^{T} L(w^*, \pi_t; Q_t, \lambda_t) - L(w_t, \pi_t; Q_t, \lambda_t) \\
&\qquad + \frac{1}{T}\sum_{t=1}^{T} L(w_t, \pi_t; Q_t, \lambda_t) - L(w_t, \pi_t; Q_\lambda^{\pi_t}, \lambda_t) \\
&\qquad + \frac{1}{T}\sum_{t=1}^{T} L(w_t, \pi_t; Q_\lambda^{\pi_t}, \lambda_t) - L(w_t, \pi_t; Q_\lambda^{\pi_t}, \lambda),
\end{aligned}$$

where the second equality is by Lemma 13.

Regret of the π-player Using the uniform concentration bound on the Lagrangian estimate provided in Lemma 16, the summand in the first term can be bounded by

$$\begin{aligned}
L(w^*, \pi^*; Q_t, \lambda_t) - L(w^*, \pi_t; Q_t, \lambda_t) &\le \widehat{L}(w^*, \pi^*; Q_t, \lambda_t) - \widehat{L}(w^*, \pi_t; Q_t, \lambda_t) + 2\varepsilon_n \\
&= (1-\gamma)Q_t(s_0, \pi^* - \pi_t) + \gamma\langle \mu^*, PQ_t(\cdot, \pi^* - \pi_t)\rangle \\
&= \mathbb{E}_{\pi^*}[Q_t(s, \pi^*) - Q_t(s, \pi_t)],
\end{aligned}$$

where the expectation is over $s \sim \sum_a \mu^{\pi^*}(\cdot, a)$. The last equality is by the Bellman flow equation. Since the π-player employs a no-regret policy optimization oracle (Definition 6), it follows that

$$\frac{1}{T}\sum_{t=1}^{T} L(w^*, \pi^*; Q_t, \lambda_t) - L(w^*, \pi_t; Q_t, \lambda_t) = \epsilon_{\mathrm{opt}}^{\pi^*}(T) \to 0,$$

as $T \to \infty$, hence we can choose T large enough so that the average regret of the π-player is at most ε_n.

Regret of the w-player To bound the second summation, note that

$$L(w^*, \pi_t; Q_t, \lambda_t) - L(w_t, \pi_t; Q_t, \lambda_t) \le \widehat{L}(w^*, \pi_t; Q_t, \lambda_t) - \widehat{L}(w_t, \pi_t; Q_t, \lambda_t) + 2\varepsilon_n,$$

and as discussed in Section D.2, we can bound the average regret of the w-player by

$$\frac{1}{T}\sum_{t=1}^{T} \widehat{L}(w^*, \pi_t; Q_t, \lambda_t) - \widehat{L}(w_t, \pi_t; Q_t, \lambda_t) \le \frac{2(C^*)^2}{\eta T} + \frac{\eta(1 + \frac{1}{\varphi} + \frac{2}{1-\gamma})}{2},$$

and the choice $\eta = \frac{1}{\sqrt{T}}$ gives the bound $\frac{2(C^*)^2}{\sqrt{T}} + \frac{(1+\frac{1}{\varphi}+\frac{2}{1-\gamma})}{2\sqrt{T}}$. Choosing T large enough so that this bound is at most ε_n completes the bound for the second term.

Regret of Q-player Fix the comparator $\lambda \in B\Delta^I$ and set $\alpha_i = \lambda_i / B$ for $i = 1, \ldots, I$ and $\alpha_0 = 1 - \sum_{i=1}^{I} \alpha_i$. These coefficients are nonnegative and sum to one. By linearity of action-value functions in the reward signal,

$$Q_\lambda^{\pi_t} = Q_0^{\pi_t} + \sum_{i=1}^{I} \lambda_i Q_i^{\pi_t} = \alpha_0 Q_0^{\pi_t} + \sum_{i=1}^{I} \alpha_i \big(Q_0^{\pi_t} + B Q_i^{\pi_t}\big).$$

Assumption B places each critic in this convex combination in $\mathcal{Q}$, hence $Q_\lambda^{\pi_t} \in \mathrm{conv}(\mathcal{Q})$. The convex-hull extension in Lemma 16 therefore justifies concentration at $Q_\lambda^{\pi_t}$, uniformly over the comparator λ and the data-dependent iterates. The summand in the third term can be bounded by

$$L(w_t, \pi_t; Q_t, \lambda_t) - L(w_t, \pi_t; Q_\lambda^{\pi_t}, \lambda_t) \le \widehat{L}(w_t, \pi_t; Q_t, \lambda_t) - \widehat{L}(w_t, \pi_t; Q_\lambda^{\pi_t}, \lambda_t) + 2\varepsilon_n,$$

The empirical difference is nonpositive because the Q-player minimizes over $\mathcal{Q}$ and $\widehat{L}$ is affine in Q. Indeed, using the coefficients of the comparator λ while keeping the multiplier argument λ_t fixed, we have

$$\begin{aligned}
&\widehat{L}(w_t, \pi_t; Q_\lambda^{\pi_t}, \lambda_t) \\
&= \widehat{L}(w_t, \pi_t; Q_0^{\pi_t} + \textstyle\sum_{i=1}^I \alpha_i B Q_i^{\pi_t}, \lambda_t) \\
&= \alpha_0 \widehat{L}(w_t, \pi_t; Q_0^{\pi_t}, \lambda_t) + \textstyle\sum_{i=1}^I \alpha_i \widehat{L}(w_t, \pi_t; Q_0^{\pi_t} + B Q_i^{\pi_t}, \lambda_t) \\
&\ge \min\{\widehat{L}(w_t, \pi_t; Q_0^{\pi_t}, \lambda_t), \widehat{L}(w_t, \pi_t; Q_0^{\pi_t} + B Q_1^{\pi_t}, \lambda_t), \dots, \widehat{L}(w_t, \pi_t; Q_0^{\pi_t} + B Q_I^{\pi_t}, \lambda_t)\} \\
&\ge \widehat{L}(w_t, \pi_t; Q_t, \lambda_t).
\end{aligned}$$

Regret of λ-player The summand in the fourth term can be bounded by

$$\begin{aligned}
L(w_t, \pi_t; Q_\lambda^{\pi_t}, \lambda_t) - L(w_t, \pi_t; Q_\lambda^{\pi_t}, \lambda) &\le \widehat{L}(w_t, \pi_t; Q_\lambda^{\pi_t}, \lambda_t) - \widehat{L}(w_t, \pi_t; Q_\lambda^{\pi_t}, \lambda) + 2\varepsilon_n \\
&= \widehat{L}(w_t, \pi_t; Q_{t-1}, \lambda_t) - \widehat{L}(w_t, \pi_t; Q_{t-1}, \lambda) + 2\varepsilon_n \\
&\le 2\varepsilon_n,
\end{aligned}$$

where the last inequality holds since the λ-player employs a greedy strategy. The equality is by the fact that the variables Q and λ are decoupled in the Lagrangian estimate $\widehat{L}(w, \pi; Q, \lambda)$.

Combining the four bounds, we get

$$L(\pi^*, \widehat{\lambda}) - L(\widehat{\pi}, \lambda) \le \mathcal{O}(\varepsilon_n),$$

and invoking Lemma 25 completes the proof. □

D Analysis for offline unconstrained RL

In this section, we provide a full analysis of the algorithm PDORL (Algorithm 2 for the offline unconstrained RL. We first present a formal theorem of the guarantee.

Theorem 18. *Assume $Q^\pi \in \mathcal{Q}$ for all $\pi \in \Pi$, and $w^{\pi_c} \in \mathcal{W}$ for some comparator policy $\pi_c \in \Pi$, and $\|w\|_\infty \le C$ for all $w \in \mathcal{W}$ and $\|Q\|_\infty \le 1/(1-\gamma)$ for all $Q \in \mathcal{Q}$. Running Algorithm 2 with learning rate $\eta = 1/\sqrt{T}$ and T large enough so that $\epsilon_{\mathrm{opt}}^{\pi^*}(T) \le \varepsilon_n$, the output policy $\widehat{\pi}$ satisfies*

$$(1-\gamma) J(\widehat{\pi}) \ge (1-\gamma) J(\pi_c) - \varepsilon_n$$

with probability at least $1-\delta$, where $\varepsilon_n = \mathcal{O}(\frac{C}{1-\gamma}\sqrt{\log(|\mathcal{W}||\Pi||\mathcal{Q}|/\delta)/n})$.

The algorithm runs a primal-dual algorithm to find a saddle point of the Lagrangian estimate:

$$\widehat{L}(w, \pi; Q) = (1-\gamma) Q(s_0, \pi) + \frac{1}{n}\sum_{i=1}^n w(s_i, a_i)(r(s_i, a_i) + \gamma Q(s_i', \pi) - Q(s_i, a_i)),$$

which is an unbiased estimator for the Lagrangian

$$L(\boldsymbol{\mu}, \pi; \boldsymbol{Q}) = (1-\gamma)\langle \boldsymbol{Q}_\pi, \boldsymbol{d}_0 \rangle + \langle \boldsymbol{\mu}, \boldsymbol{r} + \gamma \boldsymbol{P}\boldsymbol{Q}_\pi - \boldsymbol{Q} \rangle.$$

We first show a uniform concentration bound for the Lagrangian estimate.

Lemma 19. *Consider a Lagrangian function estimate*

$$\widehat{L}(w, \pi; Q) = (1-\gamma) Q(s_0, \pi) + \frac{1}{n}\sum_{i=1}^n w(s_i, a_i)(r(s_i, a_i) + \gamma Q(s_i', \pi) - Q(s_i, a_i))$$

Given a function class $\mathcal{W}$, Π and $\mathcal{Q}$ for w, π and Q, respectively, with boundedness condition $\|w\|_\infty \le C$ for all $w \in \mathcal{W}$ and $\|Q\|_\infty \le \frac{1}{1-\gamma}$ for all $Q \in \mathcal{Q}$, we have with probability at least $1-\delta$ that, for all $w \in \mathcal{W}$, $\pi \in \Pi$ and $Q \in \mathcal{Q}$,

$$|L(w, \pi; Q) - \widehat{L}(w, \pi; Q)| \le \varepsilon_n.$$

where $\varepsilon_n = \mathcal{O}(\frac{C}{1-\gamma}\sqrt{\log(|\mathcal{W}||\Pi||\mathcal{Q}|/\delta)/n})$.

Proof. The proof follows the proof of Lemma 16, which shows a uniform concentration bound of the Lagrangian estimate for the constrained RL setting. □

Lemma 20. *For every π, Q, and λ, we have*

$$L(w^\pi, \pi; \boldsymbol{Q}) = (1-\gamma)J(\pi),$$

and

$$L(w, \pi; \boldsymbol{Q}^\pi) = (1-\gamma)J(\pi).$$

Proof. It is a direct consequence of the identities under the constrained setting shown in Lemma 13. □

We are now ready to show the theorem.

Proof of Theorem 18. Let $\bar{\pi} = \text{Unif}(\pi_1, \ldots, \pi_T)$ be the policy returned by Algorithm 2. It is a policy that first uniformly randomly draws from $\{\pi_1, \ldots, \pi_T\}$, then follows the policy for the entire trajectory. It follows that $J(\bar{\pi}) = \frac{1}{T}\sum_{t=1}^T J(\pi_t)$, and

$$\begin{aligned}
(1-\gamma)J(\pi_c) - (1-\gamma)J(\bar{\pi}) &= \frac{1}{T}\sum_{t=1}^T L(w^{\pi_c}, \pi_c; Q_t) - L(w_t, \pi_t; Q^{\pi_t}) \\
&= \frac{1}{T}\sum_{t=1}^T L(w^{\pi_c}, \pi_c; Q_t) - L(w^{\pi_c}, \pi_t; Q_t) \\
&\qquad + \frac{1}{T}\sum_{t=1}^T L(w^{\pi_c}, \pi_t; Q_t) - L(w_t, \pi_t; Q_t) \\
&\qquad + \frac{1}{T}\sum_{t=1}^T L(w_t, \pi_t; Q_t) - L(w_t, \pi_t; Q^{\pi_t}),
\end{aligned}$$

where the first equality is by the identities in Lemma 13. Note that the suboptimality of the policy $\bar{\pi}$ is decomposed into average regret terms of the three players. We bound each of the regret terms as follows.

Regret of π-player The one-step regret of π-player can be written as

$$\begin{aligned}
L(w^{\pi_c}, \pi_c; Q_t) - L(w^{\pi_c}, \pi_t; Q_t) &= (1-\gamma)\langle Q_t(\cdot, \pi_c - \pi_t), d_0\rangle + \gamma\langle \mu^{\pi_c}, PQ_t(\cdot, \pi_c - \pi_t)\rangle \\
&= (1-\gamma)\langle Q_t(\cdot, \pi_c - \pi_t), d_0\rangle + \gamma\langle Q_t(\cdot, \pi_c - \pi_t), P^\top \mu_c\rangle \\
&= \langle Q_t(\cdot, \pi_c - \pi_t), (1-\gamma)d_0 + \gamma P^\top \mu^{\pi_c}\rangle \\
&= \langle \boldsymbol{E}^\top \mu^{\pi_c}, Q_t(\cdot, \pi_c - \pi_t)\rangle \\
&= \sum_{s\in\mathcal{S}} \mu^{\pi_c}(s) Q_t(s, \pi_c - \pi_t) \\
&= \mathbb{E}_{\pi_c}[Q_t(\pi_c) - Q_t(\pi_t)],
\end{aligned}$$

where the fourth equality is by the Bellman flow equation, and $\mathbb{E}_{\pi_c}$ is the expectation is over $s \sim \sum_a \mu^{\pi_c}(\cdot, a)$. Since the π-player employs a no-regret policy optimization oracle (Definition 6), the regret is bounded by

$$\frac{1}{T}\sum_{t=1}^T L(w^{\pi_c}, \pi_c; Q_t) - L(w_t, \pi_t; Q^{\pi_t}) = \frac{1}{T}\sum_{t=1}^T \mathbb{E}_{\pi_c}[Q_t(\pi_c) - Q_t(\pi_t)] \le \epsilon^\pi_{\text{opt}}(T) \to 0,$$

as $T \to \infty$. Hence the regret can be made arbitrarily small by increasing T.

Regret of w-player To bound the second summation, note that

$$L(w^{\pi_c}, \pi_t; Q_t) - L(w_t, \pi_t; Q_t) \leq \widehat{L}(w^{\pi_c}, \pi_t; Q_t) - \widehat{L}(w_t, \pi_t; Q_t) + 2\varepsilon_n,$$

and following the discussion in Section D.2, we can bound the average regret of the w-player by

$$\frac{1}{T}\sum_{t=1}^{T} \widehat{L}(w^{\pi_c}, \pi_t; Q_t, \lambda_t) - \widehat{L}(w_t, \pi_t; Q_t, \lambda_t) \leq \frac{2C^2}{\eta T} + \frac{\eta}{1-\gamma},$$

and the choice $\eta = \frac{1}{\sqrt{T}}$ gives the bound $\frac{2C^2}{\sqrt{T}} + \frac{1}{(1-\gamma)\sqrt{T}}$. Choosing T large enough so that this bound is at most ε_n completes the bound for the second term.

Regret of Q-player The one-step regret of Q-player can be bounded as

$$\begin{aligned} L(w_t, \pi_t; Q_t) - L(w_t, \pi_t; Q^{\pi_t}) &\leq \widehat{L}(w_t, \pi_t; Q_t) - \widehat{L}(w_t, \pi_t; Q^{\pi_t}) + 2\varepsilon_n \\ &\leq 2\varepsilon_n \end{aligned}$$

where the first inequality is by the concentration inequality in Lemma 19 and the second inequality is by the fact that Q-player chooses $Q_t \in \mathcal{Q}$ greedily and that $Q^{\pi_t} \in \mathcal{Q}$ by the all-policy realizability assumption. Hence, the regret of Q-player can be bounded by

$$\frac{1}{T}\sum_{t=1}^{T} L(w_t, \pi_t; Q_t) - L(w_t, \pi_t; Q^{\pi_t}) \leq 2\varepsilon_n.$$

Combining the regret bounds completes the proof. □

D.1 Comparison to previous work

In this section, we compare our algorithm, PDORL (Algorithm 2), with prior work on offline reinforcement learning with function approximation, where the performance objective is the infinite-horizon discounted sum of rewards. See Table 4 for the comparison.

Table 4: Comparison of algorithms for offline RL with general function approximation

Algorithm	Data coverage	Oracle efficient	Requires μ_D	Function Approximation	N
FQI [23]	Full	Yes	No	$\mathcal{TQ} \subseteq \mathcal{Q}$	ϵ^{-2}
Minimax [35]	Any π_c	No	No	$\forall \pi : \mathcal{T}^{\pi}\mathcal{Q} \subseteq \mathcal{Q}$	ϵ^{-5}
Minimax [41]	π^*	No	No	$Q^* \in \mathcal{Q}$	ϵ^{-2}
PRO-RL [42]	π^*	No	Yes	$w^* \in \mathcal{W}, V^* \in \mathcal{V}$	ϵ^{-6}
A-Crab [46]	Any π_c	No	No	$w^{\pi_c} \in \mathcal{W}, \forall \pi \in \Pi : Q^{\pi} \in \mathcal{Q}$	ϵ^{-2}
ATAC [3]	Any π_c	No	No	$\forall \pi : \mathcal{T}^{\pi}\mathcal{Q} \subseteq \mathcal{Q}, \pi_c \in \Pi$	ϵ^{-3}
CORAL [31]	π^*	No	Yes	$w^* \in \mathcal{W}, V^* \in \mathcal{V}, \forall w \in \mathcal{W} : u_w \in \mathcal{U}$	ϵ^{-2}
PDORL (Ours)	Any π_c	Yes	No	$\forall \pi \in \Pi : Q^{\pi} \in \mathcal{Q}, w^{\pi_c} \in \mathcal{W}$	ϵ^{-2}

Note that our algorithm is the only oracle-efficient algorithm that achieves sample complexity of ϵ^{-2} under partial data coverage that does not require the knowledge of the data generating distribution μ_D.

D.2 Projected gradient Ascent for w-player

The w-player runs the projected gradient ascent algorithm on $\mathrm{co}(\mathcal{W}$. The uniform concentration inequality for $\widehat{L}(w, \pi; Q, \lambda)$ over $w \in \mathcal{W}$ implies uniform concentration inequality over $w \in (\mathcal{W})$ since $\widehat{L}$ is affine in w. To analyze the guarantee of the projected gradient ascent algorithm employed by the w-player, we use this uniform concentration inequality, and the following standard regret bound for projected gradient ascent.

Lemma 21. *Given $x_1 \in \mathcal{X} \subseteq \mathbb{R}^d$ where $\mathcal{X}$ is convex and $\eta > 0$, define the sequences $x_2, \dots, x_{n+1} \in \mathcal{X}$ and $h_1, \dots, h_n \in \mathbb{R}^d$ such that for $k = 1, \dots, n$,*

$$x_{k+1} = \Pi_{\mathcal{X}}(x_k + \eta h_k)$$

where $\Pi_{\mathcal{X}}(\cdot)$ is a projection onto $\mathcal{X}$. Then, with the assumption that $\|h_k\|_2 \le G$ for all $k = 1, \dots, n$, we have for any $x^* \in \mathcal{X}$:

$$\sum_{k=1}^{n} \langle x^* - x_k, h_k \rangle \le \frac{\|x_1 - x^*\|_2^2}{2\eta} + \frac{\eta n G^2}{2}.$$

Proof. We start by bounding following term:

$$\begin{aligned}
\|x_{k+1} - x^*\|_2^2 &= \|\Pi_{\mathcal{X}}(x_k + \eta h_k) - \Pi_{\mathcal{X}}(x^*)\|_2^2 \\
&\le \|(x_k + \eta h_k) - x^*\|_2^2 \\
&= \|x_k - x^*\|_2^2 + 2\eta \langle x_k - x^*, h_k \rangle + \eta^2 \|h_k\|_2^2.
\end{aligned}$$

Rearranging, we get

$$2\eta \langle x^* - x_k, h_k \rangle \le \|x_k - x^*\|_2^2 - \|x_{k+1} - x^*\|_2^2 + \eta^2 \|h_k\|_2^2.$$

Summing over $k = 1, \dots, n$, we get

$$2\eta \sum_{k=1}^{n} \langle x^* - x_k, h_k \rangle \le \|x_1 - x^*\|_2^2 - \|x_{n+1} - x^*\|_2^2 + \eta^2 \sum_{k=1}^{n} \|h_k\|_2^2 \le \|x_1 - x^*\|_2^2 + \eta^2 n G^2.$$

Rearranging completes the proof. □

We now apply the lemma to the w-player. For the dataset $\mathcal{D} = \{(s_i, a_i, s_i')\}_{i=1}^n$, define the empirical evaluation map

$$E_{\mathcal{D}}(w) \coloneqq \big(w(s_1, a_1), \dots, w(s_n, a_n)\big) \in \mathbb{R}^n$$

and the convex set

$$\widetilde{\mathcal{W}}_{\mathcal{D}} \coloneqq \operatorname{co}\{E_{\mathcal{D}}(w) : w \in \mathcal{W}\} \subseteq \mathbb{R}^n.$$

Since $\widehat{L}(w, \pi; Q, \lambda)$ depends on w only through the evaluation vector $E_{\mathcal{D}}(w)$, we may equivalently let the w-player play directly in $\widetilde{\mathcal{W}}_{\mathcal{D}}$.

At round t, define the vector $g_t \in \mathbb{R}^n$ by

$$[g_t]_i \coloneqq r_{\lambda_t}(s_i, a_i) + \gamma Q_t(s_i', \pi_t) - Q_t(s_i, a_i), \qquad i = 1, \dots, n.$$

The w-player update is

$$u_{t+1} = \Pi_{\widetilde{\mathcal{W}}_{\mathcal{D}}}(u_t + \eta g_t).$$

The w-player aims to maximize the empirical Lagrangian, so for any comparator $w^* \in \mathcal{W}$,

$$\widehat{L}(w^*, \pi_t; Q_t, \lambda_t) - \widehat{L}(w_t, \pi_t; Q_t, \lambda_t) = \langle E_{\mathcal{D}}(w^*) - u_t,\ g_t \rangle.$$

Applying Lemma 21 with $x_t = u_t$, $x^* = E_{\mathcal{D}}(w^*)$, and $h_t = g_t$, we obtain

$$\begin{aligned}
\sum_{t=1}^{T} \big(\widehat{L}(w^*, \pi_t; Q_t, \lambda_t) - \widehat{L}(w_t, \pi_t; Q_t, \lambda_t)\big) &\le \frac{\|E_{\mathcal{D}}(w^*) - u_1\|^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \|g_t\|^2 \\
&\le \frac{2(C^*)^2}{\eta} + \frac{\eta T (B + \frac{2}{1-\gamma})}{2},
\end{aligned}$$

where use the bounds in Assumption D to bound $\|E_{\mathcal{D}}(w^*) - u_1\|^2$ and $\|g_t\|^2$. Choosing $\eta = \frac{C^*}{(B + \frac{2}{1-\gamma})\sqrt{T}}$ gives further bounds $\frac{(B + \frac{2}{1-\gamma})}{\sqrt{T}}$.

E Properties of saddle points

Lemma 22 (Lemma 31 in Zhan et al. [42])**.** *Suppose (x^*, y^*) is a saddle point of $f(x, y)$ over $\mathcal{X} \times \mathcal{Y}$, then for any $\mathcal{X}' \subseteq \mathcal{X}$ and $\mathcal{Y}' \subseteq \mathcal{Y}$, if $(x^*, y^*) \in \mathcal{X}' \times \mathcal{Y}'$, we have:*

$$(x^*, y^*) \in \operatorname*{argmin}_{x \in \mathcal{X}'} \operatorname*{argmax}_{y \in \mathcal{Y}'} f(x, y),$$
$$(x^*, y^*) \in \operatorname*{argmax}_{y \in \mathcal{Y}'} \operatorname*{argmin}_{x \in \mathcal{X}'} f(x, y).$$

Proof of Lemma 3. Let π^* be an optimal policy of the optimization problem $\mathcal{P}(\boldsymbol{\tau})$. Define the dual function $f(Q, \boldsymbol{\lambda}) = \max_{\mu \in \mathbb{R}_+^{\mathcal{S} \times \mathcal{A}}, \nu \in \mathbb{R}_+^{\mathcal{S} \times \mathcal{A}}} L(\mu, \nu; Q, \lambda)$. Let $(\boldsymbol{Q}^*, \boldsymbol{\lambda}^*) = \operatorname{argmin}_{\boldsymbol{Q} \in \mathbb{R}^{\mathcal{S} \times \mathcal{A}}, \boldsymbol{\lambda} \in \mathbb{R}_+^I} f(\boldsymbol{Q}, \boldsymbol{\lambda})$. Trivially, $\lambda_i^* \geq 0$ for all $i = 1, \ldots, I$. Also, by strong duality, we have $f(Q^*, \boldsymbol{\lambda}^*) = (1 - \gamma) J_0(\pi^*)$.

By the definition of the dual function, for any policy π, we have

$$\begin{aligned}(1 - \gamma) J_0(\pi^*) = f(Q^*, \lambda^*) &\geq L(\mu^\pi, \mu^\pi; Q^*, \lambda^*) \\ &= \langle \mu^\pi, r_0 \rangle + \langle \mu^\pi, R^\top \lambda^* \rangle - \langle \lambda^*, \tau \rangle \\ &= (1 - \gamma) J_0(\pi) + (1 - \gamma) \sum_{i=1}^I \lambda_i^* (J_i(\pi) - \frac{\tau_i}{1 - \gamma}). \end{aligned} \tag{13}$$

Let $\widehat{\pi}$ be a feasible policy with $J_i(\widehat{\pi}) \geq (\tau_i + \varphi)/(1 - \gamma)$ for all $i = 1, \ldots, I$. Such a policy exists by the assumption of this lemma. Then, using the display above, we get

$$\begin{aligned}(1 - \gamma) J_0(\pi^*) &\geq (1 - \gamma) J_0(\widehat{\pi}) + (1 - \gamma) \sum_{i=1}^I \lambda_i^* (J_i(\widehat{\pi}) - \frac{\tau_i}{1 - \gamma}) \\ &\geq (1 - \gamma) J_0(\widehat{\pi}) + \varphi \sum_{i=1}^I \lambda_i^* \\ &= (1 - \gamma) J_0(\widehat{\pi}) + \varphi \|\lambda^*\|_1.\end{aligned}$$

Rearranging and using $1/(1 - \gamma) \geq J_0(\pi^*) \geq J_0(\widehat{\pi}) \geq 0$ completes the proof:

$$\|\boldsymbol{\lambda}^*\|_1 \leq \frac{(1 - \gamma) J_0(\pi^*) - (1 - \gamma) J_0(\widehat{\pi})}{\varphi} \leq \frac{1}{\varphi}.$$

□

Lemma 23. *If $(\widehat{x}, \widehat{y})$ is a saddle point of $\widehat{L}(\cdot, \cdot)$ over $\mathcal{X} \times \mathcal{Y}$ and $|L(x, y) - \widehat{L}(x, y)| \leq \xi/2$ for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$, then $(\widehat{x}, \widehat{y})$ is a ξ-near saddle point of $L(\cdot, \cdot)$.*

Proof. Since $(\widehat{x}, \widehat{y})$ is a saddle point of $\widehat{L}(\cdot, \cdot)$, we have for all $(x, y) \in \mathcal{X} \times \mathcal{A}$ that

$$L(x, \widehat{y}) - L(\widehat{x}, y) \leq \widehat{L}(x, \widehat{y}) - \widehat{L}(\widehat{x}, y) + \xi \leq \xi,$$

as required. □

Lemma 24. *Consider the following Lagrangian function*

$$L(\boldsymbol{\mu}, \pi; \boldsymbol{Q}, \boldsymbol{\lambda}) = (1 - \gamma) \langle \boldsymbol{Q}_\pi, \boldsymbol{d}_0 \rangle + \langle \boldsymbol{\mu}, \boldsymbol{r}_{\boldsymbol{\lambda}} + \gamma \boldsymbol{P} \boldsymbol{Q}_\pi - \boldsymbol{Q} \rangle - \sum_{i=1}^I \lambda_i \tau_i,$$

where we use the notation $Q_\pi(s) = \sum_{a'} \pi(a'|s) Q(s, a')$. Then, the Lagrangian function has a saddle point over $(\mathcal{U} \times \Pi) \times (\mathcal{Q} \times \mathbb{R}_+^I)$, where $\mathcal{U}$, Π and $\mathcal{Q}$ contain optimal variables μ^{π^*}, π^* and $Q_0^{\pi^*} + \sum_{i=1}^I \lambda_i^* Q_i^{\pi^*}$ where π^* is an optimal policy and λ^* is an optimal dual variable.

Proof. By linear-programming strong duality, the decomposed Lagrangian (8) admits a saddle point over the unrestricted domains of μ, ν, V, Q, and λ. Moreover, if (μ^*, ν^*) is any optimal primal pair and π^* is the policy extracted from ν^*, then $\nu^* = \nu_{\mu^*,\pi^*}$, so the reparameterized Lagrangian (9) also has a saddle point over the unrestricted space of stationary policies. Hence, when restricting the variables to $(\mathcal{U}, \Pi) \times (\mathcal{Q} \times \mathbb{R}_+^I)$ for some function classes $\mathcal{U}$, $\mathcal{Q}$, policy class Π, saddle points of the restricted Lagrangian are guaranteed to exist as long as the classes contain an optimal primal-dual pair, due to the invariance property of saddle points shown in Lemma 22. □

Lemma 25. *Consider a CMDP* (1) *that satisfies the Slater's condition with margin φ (Assumption A). Define*

$$L(\pi, \lambda) := (1-\gamma)J_0(\pi) + \sum_{i=1}^{I} \lambda_i \big((1-\gamma)J_i(\pi) - \tau_i\big).$$

Let $\pi^\star$ be an optimal feasible policy. Suppose that $\widehat{\pi}$ is a policy and $\widehat{\lambda} \in \Lambda_B := \{\lambda \in \mathbb{R}_+^I : \|\lambda\|_1 \le B\}$ satisfy

$$L(\pi^\star, \widehat{\lambda}) \le L(\widehat{\pi}, \lambda) + \xi \qquad \textit{for all } \lambda \in \Lambda_B.$$

Then, $\widehat{\pi}$ is a near-optimal, near-feasible solution:

$$(1-\gamma)J_0(\widehat{\pi}) \ge (1-\gamma)J_0(\pi^\star) - \xi, \qquad (1-\gamma)J_i(\widehat{\pi}) \ge \tau_i - \xi.$$

for every $i = 1, \dots, I$,

Proof. Since π^* is feasible, we have

$$L(\pi^*, \widehat{\lambda}) = (1-\gamma)J_0(\pi^*) + \sum_{i=1}^{I} \widehat{\lambda}_i((1-\gamma)J_i(\pi^*) - \tau_i) \ge (1-\gamma)J_0(\pi^*).$$

Because $\mathbf{0} \in \Lambda_B$, the assumed inequality with $\boldsymbol{\lambda} = 0$ gives

$$(1-\gamma)J_0(\pi^*) \le L(\pi^*, \widehat{\lambda}) \le L(\widehat{\pi}, 0) + \xi = (1-\gamma)J_0(\widehat{\pi}) + \xi,$$

which shows the near optimality of $\widehat{\pi}$.

It remains to prove near feasibility. Consider the linear program formulation (2) of the CMDP:

$$\begin{aligned} \max_{\boldsymbol{\mu} \ge 0} \quad & \langle \boldsymbol{\mu}, \boldsymbol{r}_0 \rangle \\ \text{subject to} \quad & \boldsymbol{E}^\top \boldsymbol{\mu} = (1-\gamma)\boldsymbol{d}_0 + \gamma \boldsymbol{P}^\top \boldsymbol{\mu}, \\ & \langle \boldsymbol{\mu}, \boldsymbol{r}_i \rangle \ge \tau_i, \qquad i = 1, \dots, I. \end{aligned}$$

By strong duality for this linear program and Lemma 3, there exists a saddle point $(\boldsymbol{\mu}^*, \boldsymbol{\lambda}^*)$ of the corresponding Lagrangian $L(\boldsymbol{\mu}, \boldsymbol{\lambda})$ where $\boldsymbol{\lambda}^* \in \Lambda_B$. Also, it can be easily checked that $L(\pi, \boldsymbol{\lambda}) = L(\mu^\pi, \boldsymbol{\lambda})$. Hence,

$$L(\pi, \boldsymbol{\lambda}^*) = L(\mu^\pi, \boldsymbol{\lambda}^*) \le L(\boldsymbol{\mu}^*, \boldsymbol{\lambda}) = L(\pi^*, \boldsymbol{\lambda}),$$

for every $\boldsymbol{\lambda} \ge \mathbf{0}$, where the inequality follows since $(\boldsymbol{\mu}^*, \boldsymbol{\lambda}^*)$ is a saddle point. In particular, setting $\boldsymbol{\lambda} = \mathbf{0}$ and $\pi = \widehat{\pi}$, we get

$$L(\widehat{\pi}, \lambda^*) \le (1-\gamma)J_0(\pi^*).$$

Now define

$$m := \min_{i=1,\dots,I} (1-\gamma)J_i(\widehat{\pi}) - \tau_i,$$

and choose $j \in \arg\min_i g_i(\widehat{\pi})$, so that $g_j(\widehat{\pi}) = m$. Then, the above inequality gives

$$(1-\gamma)J_0(\pi^*) \ge L(\widehat{\pi}, \lambda^*) = (1-\gamma)J_0(\widehat{\pi}) + \sum_{i=1}^{I} \lambda_i^*((1-\gamma)J_i(\widehat{\pi}) - \tau_i) \ge (1-\gamma)J_0(\widehat{\pi}) + m\|\lambda^*\|_1.$$

On the other hand, since $Be_j \in \Lambda_B$, the assumed inequality gives

$$(1-\gamma)J_0(\pi^*) \le L(\pi^*, \widehat{\lambda}) \le L(\widehat{\pi}, Be_j) + \xi = (1-\gamma)J_0(\widehat{\pi}) + Bm + \xi.$$

Combining the last two displays, we obtain

$$m\|\lambda^*\|_1 \leq Bm + \xi.$$

If $m \geq 0$, then $(1-\gamma)J_i(\widehat{\pi}) - \tau_i \geq 0$ for all i, so $\widehat{\pi}$ is feasible, and there is nothing to prove. Suppose now that $m < 0$. Then, rearranging the above inequality, we get

$$m \geq \frac{-\xi}{B - \|\lambda^*\|_1} \geq -\xi$$

where we use Lemma 3 which gives $\|\lambda^*\|_1 \leq \frac{1}{\varphi} = B - 1$. By the definition of m, this means

$$(1-\gamma)J_i(\widehat{\pi}) - \tau_i \geq -\xi \qquad \text{for every } i = 1, \ldots, I,$$

which shows the desired near-feasibility of $\widehat{\pi}$. □

F Experimental setup

In this section, we explain the evaluation metric, the practical implementation of PDOCRL and the experimental setting of Table 2 and Table 3.

F.1 Evaluation metrics

In the metric setting, we follow a metric formula [19]. Denote R^π, $R_{\text{max}}(\mathcal{T})$ and $R_{\text{min}}(\mathcal{T})$ as empirical total reward evaluation of the policy, maximum, and minimum empirical total reward for task $\mathcal{T}$, respectively. Then the normalized reward is calculated by:

$$R_{\text{normalized}} = \frac{R^\pi - R_{\text{min}}(\mathcal{T})}{R_{\text{max}}(\mathcal{T}) - R_{\text{min}}(\mathcal{T})}$$

The normalized cost is calculated by dividing the total cost evaluation by the cost threshold. Denote C^π and $C_{\text{threshold}}$ as the total cost evaluation and the cost threshold respectively. The normalized cost is calculated by:

$$C_{\text{normalized}} = \frac{C^\pi}{C_{\text{threshold}}}$$

F.2 Practical implementation

We now describe the practical instantiation of PDOCRL for deep RL experiments. We parameterize the importance weight function w, the policy π, and the state-action value function Q as neural networks with parameters θ, ϕ, and ψ, respectively. Since each benchmark task specifies a single cost threshold C, the Lagrangian multiplier reduces to a scalar λ, which we parameterize via an unconstrained scalar λ_τ (e.g., $\lambda = \text{softplus}(\lambda_\tau)$) to ensure non-negativity.

The practical updates are motivated by the decomposed empirical Lagrangian associated with (10):

$$\begin{aligned} &J(\theta, \phi, \psi, \lambda_\tau) \\ &\quad = (1-\gamma)\,\overline{Q}_\psi(\pi_\phi) \;+\; \frac{1}{n}\sum_{j=1}^{n} w_\theta(s_j, a_j)\big[r_\lambda(s_j, a_j) + \gamma\, Q_\psi(s'_j, \pi_\phi) - Q_\psi(s_j, a_j)\big] \;-\; \lambda\, C, \end{aligned}$$

where $\overline{Q}_\psi(\pi_\phi) = \frac{1}{N}\sum_{i=1}^{N} Q_\psi(s_i, \pi_\phi)$ approximates the initial-state term $(1-\gamma)\,Q(s_0, \pi)$ by averaging over all states in the mini-batch, treating the empirical state distribution as a surrogate for the initial state distribution.

We adopt three stabilization techniques common in offline RL. First, to mitigate overestimation of Q, we maintain a target network ψ' [9] that is periodically synchronized with ψ. Second, because the importance weight w_θ can also suffer from overestimation, we clip its output to the interval $[0, 10]$. Third, while Algorithm 1 returns a policy sampled uniformly at random over all iterates, in practice we return the last-iterate policy, which we find to perform well empirically.

The resulting procedure is summarized in Algorithm 3.

Table 5: Comparison between the analyzed oracle-based algorithm and the practical implementation used in the experiments. The theoretical guarantees apply to Algorithm 1 under its stated assumptions and do not directly extend to Algorithm 3.

Component	Algorithm 1 / theory	Practical implementation
Function classes	Classes satisfying the stated realizability and boundedness assumptions; finite classes in the stated analysis	Neural-network parameterizations, for which the required realizability and optimization assumptions are not verified.
Primal updates	No-regret optimization oracles for the weight function and policy	Gradient-based updates, without an established guarantee matching the assumed oracle regret bounds.
Critic update	Linear-minimization oracle over the critic class	Gradient-based critic optimization, used as a practical surrogate for the analyzed oracle.
Multiplier update	Linear minimization over the bounded nonnegative multiplier domain	Numerical updates of a nonnegative multiplier, rather than exact linear minimization.
Output policy	Uniform mixture over iterates, sampled once per episode	Last-iterate policy evaluated at the final checkpoint.
Target networks	Not part of the analyzed algorithm	Used to stabilize training.
Weight clipping	Optimization over the prescribed bounded weight class	Importance weights clipped to $[0, 10]$; this numerical choice does not establish the theoretical assumptions.
Initial-state term	Expectation under the initial-state distribution	Offline dataset states used as a surrogate.

Algorithm 3: Practical PDOCRL

Input: Offline dataset $D = \{(s_0, s, a, r, c, s')_i\}_{i=1}^N$, learning rate η, Polyak averaging coefficient κ.

Initialize parameters $\theta, \phi, \psi, \lambda_\tau$
Initialize target parameters $\psi' \leftarrow \psi$
for $t = 1, 2, \ldots$ **do**
  Sample a mini-batch from D
  `// Ascent on primal variables`
  $\theta \leftarrow \theta + \eta \nabla_\theta J$
  $\phi \leftarrow \phi + \eta \nabla_\phi J$
  `// Descent on dual variables`
  $\psi \leftarrow \psi - \eta \nabla_\psi J$
  $\lambda_\tau \leftarrow \lambda_\tau - \eta \nabla_{\lambda_\tau} J$
  `// Soft target network update`
  $\psi' \leftarrow (1 - \kappa)\, \psi' + \kappa\, \psi$
return last-iterate policy π_ϕ

Relation to the theoretical algorithm. Algorithm 3 is a practical neural-network instantiation motivated by Algorithm 1, rather than an exact implementation of its optimization oracles. It retains the policy as an explicit optimization variable, thereby avoiding a separate policy-extraction step through the unknown data-generating distribution. However, gradient-based updates of the neural networks and the nonnegative multiplier are practical surrogates for the no-regret and linear-minimization oracles assumed in the analysis; their required optimization guarantees are not established here.

The analyzed output policy is a uniform mixture over the policy iterates: an iterate is sampled uniformly at the beginning of an episode and used throughout that episode. In contrast, our experiments evaluate the final policy iterate, following the final-checkpoint evaluation protocol. Target networks and clipping the importance weights to $[0, 10]$ are additional implementation choices used for stabilization. Moreover, the practical initial-state term uses states sampled from the offline dataset as a surrogate for the initial-state distribution, as described above. These choices are not covered by the stated theoretical guarantees. Table 5 summarizes the main distinctions.

Table 6: Hyperparameters for PDOCRL.

Hyperparameter	Value
Hidden sizes (actor / critic / weight)	$[256, 256]$
Number of Q-networks	2
Actor / Q / weight learning rate	3×10^{-4}
Lagrangian learning rate	1×10^{-4}
Batch size	512
Gradient steps	100,000
Discount factor γ	0.99
Target update coefficient τ	0.005
Reward scale	0.1
Cost scale	1.0
Slater margin φ	1.0
Initial λ	1.0
Evaluation episodes	20
Evaluation frequency	every 2,500 steps

F.3 Experimental setting

In this section, we describe the experimental setup for the results reported in Table 2 and Table 3. The baseline results in Table 2 are taken from [19]. For the baselines, all experiments use cost thresholds $C \in \{20, 40, 80\}$ and three random seeds $\in \{0, 10, 20\}$, yielding nine configurations per environment. For PDOCRL, we use the same cost thresholds and with seeds $\in \{0, 5, 10\}$. For each configuration, the final checkpoint is evaluated over 20 episodes for the baselines and 10 episodes for PDOCRL, and the resulting reward and cost are normalized following the protocol of [19]. The values reported in the tables are averages of the normalized reward and cost across all nine configurations.

All baseline algorithms adopt a common architecture of two hidden layers of size $[256, 256]$ with ReLU activations, are trained for 100,000 gradient steps with a batch size of 512, and use an actor learning rate of 10^{-4} and a critic learning rate of 10^{-3}. For Lagrangian-based methods (BCQ-Lag, BEAR-Lag), the Lagrangian multiplier is controlled by a PID controller with gains $(K_P, K_I, K_D) = (0.1, 0.003, 0.001)$. We refer the reader to [19] for further algorithm-specific details.

For PDOCRL, the policy network π_ϕ, the Q-networks Q_ψ, and the importance weight network w_θ each use two hidden layers of size $[256, 256]$ with ReLU activations, matching the baseline architecture. We maintain two Q-networks and take the minimum of their outputs to mitigate overestimation. The target networks are updated via Polyak averaging with coefficient $\tau = 0.005$. The actor, Q-network, and importance weight networks are each optimized with a learning rate of 3×10^{-4}, while the Lagrangian parameter λ_τ uses a learning rate of 10^{-4}. The reward signal is scaled by a factor of 0.1. We set the Slater margin to $\varphi = 1.0$ and initialize the Lagrangian multiplier at $\lambda_0 = 1.0$. Training runs for 100,000 gradient steps with a batch size of 512 and a discount factor $\gamma = 0.99$. The hyperparameters of PDOCRL are summarized in Table 6. For the ablation study, the baseline is identical to PDOCRL except for the policy extraction procedure. Lastly, All PDOCRL experiments were conducted on a local CPU-only machine, without the use of GPU acceleration.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: See the methods section 3.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: See the section section 5

   Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [Yes]

Justification: See the section 3, section B, section C, and section D.

Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: See sections section 4 and section F.

   Guidelines:

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

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

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

Answer: [Yes]

Justification: We provide the code to reproduce our results in section 4

Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: See the section 4 and section F.2

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: See the experiment section section 4

   Guidelines:

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

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

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: See section section F.3.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The research conducted in this paper conforms with the NeurIPS Code of Ethics.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: See the section 1, and section 5

    Guidelines:

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

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

11. **Safeguards**

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

    Answer: [N/A]

    Justification: This paper does not release any models or datasets that pose a high risk for misuse. Therefore, no safeguards are required.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: We cite the works to introduce all assets we deal with.

    Guidelines:

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

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

13. **New assets**

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

    Answer: [N/A]

    Justification: There is no asset

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: There is no human study.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: There is no hu,am study.

    Guidelines:

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

16. **Declaration of LLM usage**

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

Answer: [N/A]

Justification: LLMs were used for editing and writing procedures. They were not utilized for the core methods

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.