A Primal-Dual-Critic Algorithm for Offline Constrained Reinforcement Learning

Kihyuk Hong
University of Michigan

Yuhang Li
University of Michigan

Ambuj Tewari
University of Michigan

Abstract

Offline constrained reinforcement learning (RL) aims to learn a policy that maximizes the expected cumulative reward subject to constraints on expected cumulative cost using an existing dataset. In this paper, we propose Primal-Dual-Critic Algorithm (PDCA), a novel algorithm for offline constrained RL with general function approximation. PDCA runs a primal-dual algorithm on the Lagrangian function estimated by critics. The primal player employs a no-regret policy optimization oracle to maximize the Lagrangian estimate and the dual player acts greedily to minimize the Lagrangian estimate. We show that PDCA finds a near saddle point of the Lagrangian, which is nearly optimal for the constrained RL problem. Unlike previous work that requires concentrability and a strong Bellman completeness assumption, PDCA only requires concentrability and realizability assumptions for sample-efficient learning.

1 INTRODUCTION

Offline constrained reinforcement learning (RL) aims to learn a decision making policy that performs well while satisfying safety constraints given a dataset of trajectories collected from historical experiments. It enjoys the benefits of offline RL (Levine et al. 2020): not requiring interaction with the environment enables real-world applications where collecting interaction data is expensive (Kumar et al. 2021; Levine et al. 2018) or dangerous (Tang et al. 2021). It also enjoys the benefits of constrained RL (Altman 1999): being able to specify constraints to the behavior of the agent enables real-world applications with safety concerns (Wang et al. 2020; Gu et al. 2017).

Offline constrained RL with function approximation (e.g., neural networks) is of particular interest because function approximation can encode inductive biases to allow sample-efficient learning in large state spaces. As is the case for offline unconstrained RL (Ozdaglar et al. 2023; Xie et al. 2021a), offline constrained RL with function approximation requires two classes of assumptions for sample-efficient learning.

The first class of assumptions, called *representational assumptions*, requires the learner to have access to a sufficiently rich value function class that models action value functions of policies. The mildest representational assumption is the realizability assumption that requires the action value functions of candidate policies to be captured by the function class. A stronger representational assumption is the Bellman completeness assumption that requires the function class to be closed under the Bellman operator.

The second class of assumptions, called *data coverage assumptions*, requires the offline dataset to be rich enough to cover the state-action distributions induced by target policies. The assumptions address a major challenge in offline RL called *distribution* shift, which refers to the mismatch of the state-action distributions induced by candidate policies and the distribution in the offline dataset. The most commonly used data coverage assumption is concentrability (Munos 2003; Munos 2005), which limits the norm of the ratio of state-action distribution induced by candidate policies to that induced by the behavior policy that generated the offline dataset.

Previous works on offline RL with function approximation require either a strong assumption on data coverage (Xie et al. 2021b) (stronger than concentrability) or a strong representational assumption (Munos et al. 2008; Antos et al. 2008; Xie et al. 2021a; Cheng et al. 2022) (stronger than realizability). Chen et al. (2019) conjectured that concentrability and realizability of value functions are not sufficient for sample-efficient

Proceedings of the 27$^{\text{th}}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

offline RL. Foster et al. (2022) confirmed this by providing an information-theoretic lower bound which shows that concentrability and value function realizability are not sufficient for sample efficient offline RL.

Recently, a line of research on offline unconstrained RL emerged that only requires concentrability and realizability assumptions (Xie et al. 2020; Zhan et al. 2022; Zhu et al. 2023). In particular, they do not require Bellman completeness assumption, which is a strong representational assumption (Zhan et al. 2022; Zanette et al. 2022). They do not contradict the impossibility result by Foster et al. (2022) because they make an additional realizability assumption on the *marginalized importance weights* (MIW; ratio of state-action distribution induced by policy to data distribution). Motivated by their work, we propose a sample-efficient algorithm for offline constrained RL with function approximation that requires concentrability, value function realizability and MIW realizability assumptions. We make the following contributions.

- We show a sample complexity bound that scales with a concentrability measure, $1/\epsilon^2$ and a dimensionality measure of function classes, for finding a nearly optimal policy with suboptimality ϵ that approximately satisfies the cost constraints under the assumptions of value function realizability, concentrability, and MIW realizability of an optimal policy. We do not require Bellman completeness, a strong representational assumption, required by previous work.
- Our algorithm takes as an input a target cost threshold. By using a target cost threshold stricter than the desired threshold, the algorithm can produce a nearly optimal policy that *exactly* satisfies the desired constraints with the same sample complexity.
- We study the case where the function class for MIW is misspecified and does not realize the MIW of an optimal policy. In this case, our algorithm can still find a policy at least as good as any policy of which MIW is realized by the function class but the sample complexity bound is suboptimal and scales with $1/\epsilon^4$.
- Benchmark experiments show the empirical performance of our algorithm generally matches or outperforms the state-of-the-art algorithms COptiDICE and CPQ that produce Markovian policies.

1.1 Related Work

Offline RL without Completeness Assumption There is a recent line of works on offline *unconstrained* RL that removes the Bellman completeness assumption by assuming MIW realizability. Xie et al. (2020) propose a Q-value based algorithm called MABO that learns the optimal Q-value function by solving a minimax optimization problem. They require all-policy realizability of value functions, all-policy concentrability and all-policy marginalized importance weight realizability. Zhan et al. (2022) propose a linear programming based algorithm called PRO-RL that regularizes the objective function to discourage distribution shift. They only require single-policy realizability of both value function and marginalized importance weight, and only require single-policy concentrability. However, their sample complexity is suboptimal ($\sim 1/\epsilon^6$). Zhu et al. (2023) propose an actor-critic based algorithm called A-Crab. They require all-policy value function realizability, single-policy concentrability and single-policy marginalized importance weight realizability.

Offline Constrained RL The only work on provably sample efficient offline constrained RL with function approximation, to the best of our knowledge, is by Le et al. (2019) who propose a provably sample-efficient primal-dual algorithm that uses the fitted-Q iteration algorithm as a subroutine for updating the primal variable and a no-regret online algorithm for updating the dual variable. Their analysis requires all-policy concentrability and Bellman completeness assumptions. Our work improves over Le et al. (2019) by weakening the Bellman completeness assumption.

Practical Algorithms for Offline Constrained RL There are recent works on practical algorithms for offline constrained RL *without* provable guarantees. Lee et al. (2022) propose an algorithm called COptiDICE, which is motivated by the linear programming approach for solving RL. Liu et al. (2023b) propose CDT, an adaptation of the decision transformer framework for offline RL (Chen et al. 2021a) to the offline constrained RL setting. Xu et al. (2022) propose CPQ, a Q-learning based algorithm that penalizes out of distribution actions. Liu et al. (2023a) provide datasets and benchmarks of aforementioned algorithms.

We compare our theoretical guarantees with previous works in Table 1. The first three rows are works on offline *unconstrained* RL with function approximation that do not assume Bellman completeness. The remaining rows are works on offline *constrained* RL with function approximation. The column N shows how the sample complexity bound scales with the error tolerance ϵ. Q^π is the value function for the policy π; w^π is the marginalized importance weight of the policy π; sp is the span function; $\pi^\star$ is the optimal policy. $\mathcal{T}^\pi$ is the Bellman operator and $\forall \pi, \mathcal{T}^\pi f \in \mathcal{F}$ means Bellman completeness. The notations used in the table are formally defined in the next section.

Compared to the work by Le et al. (2019), we relax

Table 1: Comparison of algorithms for offline (constrained) RL with function approximation

Algorithm	Supports constraints	Assumptions			N
		Representation	Data coverage	MIW	
MABO (Xie et al. 2020)	No	$\forall\pi, Q^\pi \in \mathcal{F}$	$\forall\pi, \lVert w^\pi\rVert_{2,\mu} \le C_{\ell_2}$	$\forall\pi, w^\pi \in \mathrm{sp}(\mathcal{W})$	$1/\epsilon^2$
PRO-RL (Zhan et al. 2022)	No	$Q^{\pi^\star} \in \mathcal{F}$	$\lVert w^{\pi^\star}\rVert_\infty \le C^\star_\infty$	$w^{\pi^\star} \in \mathcal{W}$	$1/\epsilon^6$
A-Crab (Zhu et al. 2023)	No	$\forall\pi, Q^\pi \in \mathcal{F}$	$\lVert w^{\pi^\star}\rVert_{2,\mu} \le C^\star_{\ell_2}$	$w^{\pi^\star} \in \mathcal{W}$	$1/\epsilon^2$
MBCL (Le et al. 2019)	Yes	$\forall\pi, f; \mathcal{T}^\pi f \in \mathcal{F}$	$\forall\pi, \lVert w^\pi\rVert_\infty \le C_\infty$		$1/\epsilon^2$
PDCA (Ours)	Yes	$\forall\pi, Q^\pi \in \mathcal{F}$	$\forall\pi, \lVert w^\pi\rVert_{2,\mu} \le C_{\ell_2}$	$w^{\pi^\star} \in \mathcal{W}$	$1/\epsilon^2$

the Bellman completeness assumption at the expense of introducing a MIW realizability of an optimal policy. The MIW realizability is a mild assumption since function class $\mathcal{W}$ only needs to include the MIW of an *optimal* policy. Moreover, we show in Theorem 2 that even when $\mathcal{W}$ does not realize the MIW of an optimal policy, our algorithm can find a policy at least as good as any policy whose MIW is realizable by $\mathcal{W}$. This result allows robustness against misspecification of $\mathcal{W}$.

2 PRELIMINARIES & NOTATIONS

Notation We denote by $\Delta(\mathcal{X})$ the probability simplex over a finite set $\mathcal{X}$. We denote by $\mathbb{R}_+$ the set of nonnegative real numbers. We write $\boldsymbol{\Delta}^I = \{\boldsymbol{x} \in \mathbb{R}^I_+ : \sum_{i=1}^I x_i \le 1\}$. We denote by Unif($\mathcal{X}$) the uniform distribution over $\mathcal{X}$. We write $[N] = \{1, \dots, N\}$ for a natural number N. We write $\mathbf{1} = (1, \dots, 1)$ and $\mathbf{0} = (0, \dots, 0)$. We write $(\cdot)_+ = \max\{0, \cdot\}$.

2.1 Constrained Markov Decision Process

We formulate offline constrained RL using an infinite-horizon discounted constrained Markov decision process (CMDP)(Altman 1999) defined by a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, R, \{C_i\}_{i=1}^I, \gamma, s_0)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition probability kernel, $R : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the reward function, $C_i : \mathcal{S} \times \mathcal{A} \to [0, 1], i = 1, \dots, I$ are the cost functions, $\gamma \in (0, 1)$ is the discount factor, and $s_0 \in \mathcal{S}$ is the initial state. We assume R and C_i, $i = 1, \dots, I$ are known to the learner while P is unknown.

A stationary policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ maps each state to a probability distribution over the action space. Each policy π, together with the transition probability kernel P, induces a discounted occupancy measure $d^\pi : \mathcal{S} \times \mathcal{A} \to [0, 1]$ defined as $d^\pi(s, a) := (1 - \gamma) \sum_{t=0}^\infty \gamma^t P^\pi(s_t = s, a_t = a)$ where P^π is the probability measure on the trajectory $(s_0, a_0, s_1, a_1, \dots)$ induced by the interaction of π and P. The value of a policy π for a function $U : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the expected discounted cumulative values when executing π. It is denoted by $J_U(\pi) := \mathbb{E}^\pi [\sum_{t=0}^\infty \gamma^t U(s_t, a_t)]$ where $\mathbb{E}^\pi$ is the expectation over the randomness of the trajectory $(s_0, a_0, s_1, a_1, \dots)$ induced by π and P. Note that $J_U(\pi) \in [0, \frac{1}{1-\gamma}]$ and $(1-\gamma) J_U(\pi) = \mathbb{E}_\pi[U(s, a)]$ where we use the shorthand $\mathbb{E}_\pi[\cdot]$ for $\mathbb{E}_{(s,a)\sim d^\pi}[\cdot]$. The Q-value function of a policy π for a function $U : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is denoted by $Q^\pi_U(s, a) := \mathbb{E}^\pi [\sum_{t=0}^\infty \gamma^t U(s_t, a_t) \mid s_0 = s, a_0 = a]$.

2.2 Function Approximation

We assume access to a policy class $\Pi \subseteq (\pi : \mathcal{S} \to \Delta(\mathcal{A}))$ consisting of candidate policies. We assume access to a function class $\mathcal{F} \subseteq (\mathcal{S} \times \mathcal{A} \to [0, \frac{1}{1-\gamma}])$ that models the Q-value functions for the reward R and the costs $C_1, \dots, C_I$. We make the following realizability assumption on $\mathcal{F}$.

Assumption A (Value function realizability)**.** *For any policy $\pi \in \Pi$, we have $Q^\pi_R \in \mathcal{F}$ and $Q^\pi_{C_i} \in \mathcal{F}$ for all $i = 1, \dots, I$.*

Unlike Le et al. (2019), we do not assume Bellman completeness that requires $\mathcal{T}^\pi_U f \in \mathcal{F}$ for all $\pi \in \Pi$ and $f \in \mathcal{F}$ where $\mathcal{T}^\pi_U$ is the Bellman operator defined by $(\mathcal{T}^\pi_U f)(s, a) = U(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[f(s', \pi)]$. As Zhan et al. (2022) and Zanette (2022) discuss, it is a strong assumption hard to meet and has an unnatural non-monotone property: adding a function to the function class may make the function class violate Bellman completeness.

2.3 Offline Constrained RL

Offline constrained RL aims to find a policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ among a given policy class Π that maximizes $J_R(\pi)$ while satisfying the constraints $J_{C_i}(\pi) \le \tau_i$ for all $i = 1, \dots, I$, where the thresholds $\tau_i \in [0, \frac{1}{1-\gamma}]$ are given. Offline constrained RL can be written as an optimization problem $\mathcal{P}(\boldsymbol{\tau})$ defined as follows.

Definition 1 (Optimization problem)**.** *Given cost thresholds $\boldsymbol{\tau} = (\tau_1, \dots, \tau_I)$, we denote by $\mathcal{P}(\boldsymbol{\tau})$ the following optimization problem.*

$$\begin{aligned} \max_{\pi \in \mathrm{Conv}(\Pi)} \quad & J_R(\pi) \\ \text{subject to} \quad & J_{C_i}(\pi) \le \tau_i, \quad i = 1, \dots I. \end{aligned} \tag{OPT}$$

As done in Le et al. (2019), instead of optimizing over the policy class Π, we optimize over its convex hull denoted by $\mathrm{Conv}(\Pi)$. The convex hull $\mathrm{Conv}(\Pi)$ contains all policy mixtures of the form $\sum_{j=1}^{m} \beta_j \pi_j$ where $\pi_1, \ldots, \pi_m \in \Pi$, $\beta_j \geq 0$ for $j = 1, \ldots, m$ and $\sum_{j=1}^{m} \beta_j = 1$. A policy mixture $\sum_{j=1}^{m} \beta_j \pi_j$ is executed by sampling a single policy from $\pi_1, \ldots, \pi_m$ according to the distribution $(\beta_1, \ldots, \beta_m)$, and then executing the sampled policy for the entire trajectory. Viewing the problem in the occupancy measure space, (OPT) can be seen as

$$\begin{aligned} \max_{\nu \in \mathrm{Conv}(\mathcal{V})} \quad & \langle \nu, R \rangle \\ \text{subject to} \quad & \langle \nu, C_i \rangle \leq \tau_i, \quad i = 1, \ldots, I \end{aligned} \tag{1}$$

where $\mathcal{V} = \{d^\pi : \pi \in \Pi\}$ is the set of occupancy measures of policies in Π and d^π, R, C_i, $i = 1, \ldots I$ are viewed as a vector in $\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$. Since we define a mixture of policies in the trajectory level, the set of occupancy measures of policies in $\mathrm{Conv}(\Pi)$ is just the convex hull of $\mathcal{V}$. Since the above is an optimization problem in the space of $\mathbb{R}^{|S||A|}$, strong duality holds if we assume the following Slater's condition.

Assumption B (Slater's condition)**.** *There exist a constant $\varphi > 0$ and a policy $\pi \in \Pi$ such that $J_{C_i}(\pi) \leq \tau_i - \frac{\varphi}{1-\gamma}$ for all $i = 1, \ldots, I$. Assume φ is known.*

Slater's condition is a mild assumption commonly made for constrained RL (Le et al. 2019; Chen et al. 2021b; Bai et al. 2022; Ding et al. 2020) for ensuring strong duality of the optimization problem. It is mild because given the knowledge of the feasibility of the problem, we can guarantee that Slater's condition is met by slightly loosening the cost threshold.

2.4 Offline Dataset

In offline constrained RL, we assume access to an offline dataset $\mathcal{D} = \{(s_j, a_j, s_j')\}_{j=1}^{n}$ where (s_j, a_j) are generated i.i.d. from a data distribution d^μ induced by a behavior policy μ and $s_j' \sim P(\cdot \mid s_j, a_j)$. Such an i.i.d. assumption on the offline dataset is commonly made in the offline RL literature (Xie et al. 2021a; Zhan et al. 2022; Chen et al. 2022; Zhu et al. 2023) to facilitate analysis of concentration bounds. We assume the policy class Π contains μ. We assume that the threshold $\boldsymbol{\tau}$ is chosen such that the optimization problem $\mathcal{P}(\boldsymbol{\tau})$ is feasible. However, we do not require the behavior policy μ to be feasible for $\mathcal{P}(\boldsymbol{\tau})$. To limit the distribution shift of policies from the data distribution, we make the following concentrability assumption, where we use the notation $\|\cdot\|_{2,\mu} = \sqrt{\mathbb{E}_\mu[(\cdot)^2]}$.

Assumption C (Concentrability)**.** *For all $\pi \in \Pi$, we have $\|d^\pi / d^\mu\|_{2,\mu} \leq C_{\ell_2}$.*

The concentrability assumption limits distribution shift of candidate policies from the data distribution. Specifically, the occupancy measure induced by a policy in Π is covered by the data distribution d^μ. This assumption is weaker than the assumption made by Le et al. (2019), who assume that the ℓ_∞ norm of the distribution shift of following any nonstationary policy that uses a policy in Π every time step is bounded.

2.5 Marginalized Importance Weight

The notion of marginalized importance weight (MIW) is used extensively in the offline RL literature (Xie et al. 2020; Chen et al. 2022; Zhan et al. 2022; Zhu et al. 2023; Lee et al. 2022; Lee et al. 2021) to correct for the distribution mismatch between a policy π and the behavior policy μ. It is defined as follows.

Definition 2 (Marginalized importance weight)**.** *For a policy π, we define the marginalized importance weight $w^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^+$ as $w^\pi(s, a) \coloneqq \frac{d^\pi(s,a)}{d^\mu(s,a)}$.*

Immediately from the definition of MIW, we get the identity $\mathbb{E}_\pi[(\cdot)] = \mathbb{E}_\mu[w^\pi(s, a)(\cdot)]$, which we frequently use in the analysis. We assume access to a function class $\mathcal{W}$ consisting of functions $w : \mathcal{S} \times \mathcal{A} \to \mathbb{R}_+$ that represents MIW with respect to the offline data distribution d^μ. We assume the following boundedness assumption on $\mathcal{W}$.

Assumption D (Boundedness of $\mathcal{W}$)**.** *Assume $\|w\|_\infty \leq C_\infty$ and $\|w\|_{2,\mu} \leq C_{\ell_2}$ for all $w \in \mathcal{W}$.*

Denote by $\pi^\star$ an optimal policy of the optimization problem (OPT). We assume that the MIW of $\pi^\star$ is realized by $\mathcal{W}$.

Assumption E (Realizability of MIW)**.** *Assume that $w^{\pi^\star} \in \mathcal{W}$ for an optimal policy $\pi^\star$.*

The single-policy realizability of MIW that requires MIW of an optimal policy to be realizable by $\mathcal{W}$ is a weaker assumption than the all-policy realizability of MIW assumption required by Xie et al. (2020). Compared to the set of assumptions made by Le et al. (2019), we replace the strong Bellman completeness assumption with a single-policy realizability of MIW.

3 ALGORITHM & MAIN RESULTS

In this section, we present our algorithm called Primal-Dual-Critic Algorithm (PDCA) and then present our main results on the the sample complexity bound.

3.1 Primal-Dual Algorithm Structure

The Lagrangian of the optimization problem $\mathcal{P}(\boldsymbol{\tau})$ (Definition 1) is $L(\pi, \boldsymbol{\lambda}) = J_R(\pi) + \boldsymbol{\lambda} \cdot (\boldsymbol{\tau} - J_{\boldsymbol{C}}(\pi))$ where we use the notation $J_{\boldsymbol{C}}(\pi) = (J_{C_1}(\pi), \ldots, J_{C_I}(\pi))$. Our algorithm, like MBCL algorithm for offline constrained RL proposed by Le et al. (2019), adopts the primal-dual algorithm structure that updates π and $\boldsymbol{\lambda}$ alternatively. The primal-dual algorithm structure can be seen as a sequential game of length K between the π-player who tries to maximize $L(\cdot, \boldsymbol{\lambda})$ and the λ-player who tries to minimize $L(\pi, \cdot)$. Both players try to minimize their regrets against respective best actions in hindsight.

The key difference of our algorithm from MBCL is that in each round k, the π-player plays before the λ-player, while in MBCL, the order is reversed. Since λ-player sees what π-player plays before playing, the λ-player can act greedily to minimize the regret. On the other hand, the π-player has to use a no-regret policy optimization oracle, defined below, with a sublinear regret against adversarially chosen sequence of λ's.

Definition 3 (No-regret policy optimization oracle)**.** *An algorithm is called a no-regret policy optimization oracle if for any sequence of functions $h_1, \ldots, h_K : \mathcal{S} \times \mathcal{A} \to [-1, 1]$, the sequence of policies $\pi_1, \ldots, \pi_K \in \Pi$ produced by the algorithm satisfies*

$$\frac{1}{K} \sum_{k=1}^{K} \mathbb{E}_\pi[h_k(s, \pi) - h_k(s, \pi_k)] = \epsilon_{opt}(K)$$

with high probability for any $\pi \in \mathrm{Conv}(\Pi)$ where $\epsilon_{opt}(K) \to 0$ as $K \to \infty$.

A well-known instance of the oracle is the natural policy gradient algorithm (Kakade 2001) based on the updates $\pi_{k+1}(a|s) \propto \pi_k(a|s) \exp(\eta h_k(s, a))$.

3.2 Critics for Lagrangian

We want to estimate the Lagrangian function $L(\pi, \boldsymbol{\lambda}) = J_R(\pi) + \boldsymbol{\lambda} \cdot (\boldsymbol{\tau} - J_{\boldsymbol{C}}(\pi))$ for all $\pi \in \Pi$ and $\boldsymbol{\lambda} \in B\boldsymbol{\Delta}^I$. We use critics for $J_R(\pi)$ and $J_{\boldsymbol{C}}(\pi)$ that are inspired by the reward critic proposed by Zhu et al. (2023) for their actor-critic algorithm for offline unconstrained RL. Our critic for $J_U(\pi)$ aims to solve

$$\min_{f \in \mathcal{F}} \; 2\mathcal{E}_\mu(\pi, f; U) \pm A_\mu(\pi, f)$$

where the sign of A_μ is appropriately chosen and

$$\mathcal{E}_\mu(\pi, f; U) := \max_{w \in \mathcal{W}} |\mathbb{E}_\mu[w(s, a)(f - \mathcal{T}_U^\pi f)(s, a)]|$$
$$A_\mu(\pi, f) := \mathbb{E}_\mu[f(s, \pi) - f(s, a)].$$

Here, $\mathcal{T}_U^\pi : \mathbb{R}^{\mathcal{S}\times\mathcal{A}} \to \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ is the Bellman operator with $(\mathcal{T}_U^\pi f)(s, a) = U(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot \mid s, a)}[f(s', \pi)]$ and $f(s, \pi) = \sum_{a \in \mathcal{A}} \pi(a|s) f(s, a)$. Minimizing the first term $\mathcal{E}_\mu$ in the objective function of the critics encourages Bellman-consistency. The second term, with appropriate sign, facilitates the regret bound analysis for the π-player. Since the data distribution of the behavior policy d^μ is unknown, we solve an empirical version $\min_{f \in \mathcal{F}} 2\mathcal{E}_\mathcal{D}(\pi, f; U) + A_\mathcal{D}(\pi, f)$ where

$$\mathcal{E}_\mathcal{D}(\pi, f; U) := \max_{w \in \mathcal{W}} |\mathbb{E}_\mathcal{D}[w(s, a)(f(s, a) - U(s, a) - \gamma f(s', \pi))]|$$
$$A_\mathcal{D}(\pi, f) := \mathbb{E}_\mathcal{D}[f(s, \pi) - f(s, a)].$$

Here, $\mathbb{E}_\mathcal{D}[F(s, a, s')] = \frac{1}{|\mathcal{D}|} \sum_{(s,a,s') \in \mathcal{D}} F(s, a, s')$. The critics for the reward value $J_R(\pi)$ and the cost value $J_{C_i}(\pi)$ are chosen to be

$$\text{Reward critic:} \quad \min_{f \in \mathcal{F}} 2\mathcal{E}_\mathcal{D}(\pi, f; R) + A_\mathcal{D}(\pi, f) \tag{2}$$

$$\text{Cost critic:} \quad \min_{g \in \mathcal{F}} 2\mathcal{E}_\mathcal{D}(\pi, g; C_i) - A_\mathcal{D}(\pi, g). \tag{3}$$

Note that we use the same reward critic as the one used in Zhu et al. (2023), but we negate the term $\mathcal{A}_\mathcal{D}$ for the cost critic.

3.3 Cost Critic for λ-Player

While the π-player optimizes for the Lagrangian estimated by critics (2) and (3) in the previous section, the λ-player optimizes for $\boldsymbol{\lambda} \cdot (\boldsymbol{\tau} - J_{\boldsymbol{C}}(\pi_k))$ estimated by an offline policy evaluation (OPE) oracle.

Definition 4 (OPE oracle)**.** *Let π be a policy and U a utility function. Let $\mathcal{F}$ be a function class that contains the value function Q_U^π. Suppose the dataset $\mathcal{D} = \{(s_j, a_j, s'_j)\}_{j=1}^n$ is generated with a behavior policy μ is a behavior policy with sufficient coverage such that $\|d^\pi/d^\mu\|_{2,\mu} \le C_{\ell_2}$. An algorithm that produces an estimate $h \in \mathbb{R}$ for $J_U(\pi)$ such that*

$$|h - J_U(\pi)| \le \mathcal{O}\left(\frac{C_{\ell_2}}{1-\gamma}\sqrt{\frac{\log(|\mathcal{F}|/\delta)}{n}}\right)$$

with probability at least $1 - \delta$ is called an OPE oracle.

An example of such an oracle is provided by Zanette (2022) whose algorithm produces an estimate with error decreasing at the required rate scaled by a constant factor called the incompleteness factor inherent to $\mathcal{F}$ that measures misalignment between $\mathcal{F}$ and $\mathcal{T}^\pi\mathcal{F}$.

Remark 1. *We use a separate OPE cost critic for estimating the objective function of the λ-player instead of reusing the cost critic (3) used for estimating the Lagrangian for the π-player. This is because the regrets for the two players take different forms. As we show in Section 4.1, the regret of π-player is the sum of $J_{R+\boldsymbol{\lambda}_k\cdot\boldsymbol{C}}(\pi^\star) - J_{R+\boldsymbol{\lambda}_k\cdot\boldsymbol{C}}(\pi_k)$ while the regret of λ-player is the sum of $(\boldsymbol{\lambda}_k - \boldsymbol{\lambda}^\star) \cdot (\boldsymbol{\tau} - J_{\boldsymbol{C}}(\pi_k))$. Since the regret minimizing decision for λ-player depends on the sign of $\tau - J_{C_i}(\pi_k)$, we estimate $J_{C_i}(\pi_k)$ by calling an OPE oracle. This is why we require concentrability of all policies in Π. We leave relaxing to a single-policy concentrability assumption as future work.*

Algorithm 1: Primal-Dual-Critic Algorithm

Input: Dataset $\mathcal{D} = \{(s_j, a_j, s'_j)\}_{j=1}^n$, number of iterations K, cost thresholds $\boldsymbol{\tau}$, bound B, , no-regret policy optimization oracle PO, offline policy evaluation oracle OPE.

Init: π_1: uniform policy.

for $k = 1, \dots, K$ **do**
- $f_k \leftarrow \operatorname{argmin}_{f \in \mathcal{F}} 2\mathcal{E}_{\mathcal{D}}(\pi_k, f; R) + A_{\mathcal{D}}(\pi_k, f)$.
- $g_k^i \leftarrow \operatorname{argmin}_{g \in \mathcal{F}} 2\mathcal{E}_{\mathcal{D}}(\pi_k, g; C_i) - A_{\mathcal{D}}(\pi_k, g)$ for all $i = 1, \dots, I$.
- $h_k^i \leftarrow \mathrm{OPE}(\pi_k, C_i)$ for all $i = 1, \dots, I$.
- $\boldsymbol{\lambda}_k \leftarrow \operatorname{argmin}_{\boldsymbol{\lambda} \in B\boldsymbol{\Delta}^I} \boldsymbol{\lambda} \cdot (\boldsymbol{\tau} - \boldsymbol{h}_k)$.
- $\pi_{k+1} \leftarrow \mathrm{PO}(\pi_k, f_k + \boldsymbol{\lambda}_k \cdot (\boldsymbol{\tau} - \boldsymbol{g}_k))$.

Return: $\bar{\pi} = \mathrm{Unif}(\pi_1, \dots, \pi_K)$

3.4 Proposed Algorithm

We propose an algorithm called Primal-Dual-Critic Algorithm (PDCA) (Algorithm 1) that uses critics for the Lagrangian function and no-regret policy optimization oracle for the π-player and a greedy λ-player. The algorithm takes cost thresholds $\boldsymbol{\tau}$ as an input and runs a primal-dual algorithm on the estimate of Lagrangian $L(\pi, \boldsymbol{\lambda}) = J_R(\pi) + \boldsymbol{\lambda} \cdot (\boldsymbol{\tau} - J_{\boldsymbol{C}}(\pi))$. The algorithm iterates for K steps. In each step k, the algorithm calculates the reward critic f_k and the cost critics $\boldsymbol{g}_k = (g_k^1, \cdots, g_k^I)$ using the offline dataset $\mathcal{D}$. Then, the π-player invokes a no-regret policy optimization oracle on the estimate $f_k + \boldsymbol{\lambda}_k \cdot (\boldsymbol{\tau} - \boldsymbol{g}_k)$. The OPE oracle is used to estimate $J_{\boldsymbol{C}}(\pi_k)$ by $\boldsymbol{h}_k(s_0, \pi_k)$, and the λ-player chooses $\boldsymbol{\lambda}_k$ that minimizes $\boldsymbol{\lambda} \cdot (\boldsymbol{\tau} - \boldsymbol{h}_k)$. After running the iterations, the algorithm returns the policy $\bar{\pi} = \mathrm{Unif}(\pi_1, \dots, \pi_K)$, which is a uniform mixture of the policies $\pi_1, \dots, \pi_K$. The uniform mixture policy initially samples a policy uniformly at random from $\{\pi_1, \dots, \pi_K\}$, and then follows the sampled policy for the entire trajectory.

3.5 Main Results

Our main theoretical result is a sample complexity bound of our algorithm called Primal-Dual-Critic Algorithm (PDCA) (Algorithm 1) for finding a policy that satisfies the constraints approximately and is ϵ-optimal with respect to the optimal policy $\pi^\star$ for $\mathcal{P}(\boldsymbol{\tau})$.

Theorem 1. *Under assumptions A,B,C,D and E, the policy $\bar{\pi}$ returned by the PDCA algorithm (Algorithm 1) with the cost threshold $\boldsymbol{\tau}$, bound $B = 1 + \frac{1}{\varphi}$ and K large enough, satisfies $J_{C_i}(\bar{\pi}) \le \tau_i + \mathcal{O}(\epsilon)$ for all $i = 1, \dots, I$, and $J_R(\bar{\pi}) \ge J_R(\pi^\star) - \mathcal{O}(\epsilon)$ with probability at least $1 - \delta$ where $\pi^\star$ is optimal for $\mathcal{P}(\boldsymbol{\tau})$ as long as*

$$n \ge \Omega\left(\frac{(C_{\ell_2})^2 \log(I|\mathcal{F}||\Pi||\mathcal{W}|/\delta)}{(1-\gamma)^4 \varphi^2 \epsilon^2}\right).$$

The sample complexity bound provided by Le et al. (2019) is $\mathcal{O}(\frac{C_\infty(\dim_F + \log(I/\delta))}{(1-\gamma)^{10}\epsilon^2})$ where $\dim_F$ is a complexity measure of the function class they use for modeling the Lagrangian function, which is analogous to the log cardinality term $\log|\mathcal{F}|$ in our finite function class setting. Compared to their bound, our bound saves a factor of $\frac{1}{(1-\gamma)^6}$ and depends on concentrability coefficient C_{ℓ_2} instead of C_∞. As Zhu et al. (2023) discuss, $(C_{\ell_2})^2 \le C_\infty$ and C_∞ can be arbitrarily larger than $(C_{\ell_2})^2$. Also, our algorithm requires weaker assumptions. While Le et al. (2019) require ℓ_∞ concentrability for all sequences of policies, we require ℓ_2 concentrability for fixed policies. While Le et al. (2019) require Bellman completeness for the value function class, we only require realizability. The only additional assumption we need is the single-policy realizability of the marginalized importance weight.

With different choices of inputs to the PDCA algorithm, we get the following results. See Appendix E for the formal statements and proofs.

Arbitrary Comparator Policy Without the Slater's condition and the MIW realizability assumptions, running PDCA with the cost threshold $\boldsymbol{\tau}$ and the bound $B = \frac{1}{(1-\gamma)\epsilon}$ gives a policy $\bar{\pi}$ that is nearly feasible and satisfies near-optimality ($J_R(\bar{\pi}) \ge J_R(\pi_c) - \mathcal{O}(\epsilon)$) against any comparator policy $\pi_c \in \mathrm{Conv}(\Pi)$ of which MIW is realizable by $\mathcal{W}$. However, the sample complexity bound is of $\mathcal{O}(1/\epsilon^4)$.

Theorem 2. *Under assumptions A,C and D, the policy $\bar{\pi}$ returned by PDCA (Algorithm 1) with the cost threshold $\boldsymbol{\tau}$ and bound $B = \frac{1}{(1-\gamma)\epsilon}$ satisfies $J_{C_i}(\bar{\pi}) \le \tau_i + \mathcal{O}(\epsilon)$ for all $i = 1, \dots, I$, and $J_R(\pi) \ge J_R(\pi_c) - \mathcal{O}(\epsilon)$ with probability at least $1 - \delta$ where $\pi_c \in \mathrm{Conv}(\Pi)$ is any policy of which MIW is realizable by $\mathcal{W}$ as long as*

$$n \ge \Omega\left(\frac{(C_{\ell_2})^2 \log(I|\mathcal{F}||\Pi||\mathcal{W}|/\delta)}{(1-\gamma)^6 \epsilon^4}\right).$$

See Appendix E.2 for a proof.

Exact Feasibility With the same sample complexity as in Theorem 1, running PDCA with a tightened cost threshold $\boldsymbol{\tau} - \eta\mathbf{1}$ where $\eta = \mathcal{O}(\epsilon)$ and the bound $B = \mathcal{O}(\frac{1}{\varphi})$ gives a policy $\bar{\pi}$ that is exactly feasible ($J_{\boldsymbol{C}}(\bar{\pi}) \le \boldsymbol{\tau}$) and nearly optimal ($J_R(\bar{\pi}) \ge J_R(\pi^\star) - \mathcal{O}(\epsilon)$). Exact feasibility can be shown with the following additional MIW realizability assumption.

Assumption F. *Suppose the Slater's condition holds (Assumption B). For some constant $\alpha \in [\frac{\varphi}{c(1-\gamma)}, \frac{\varphi}{1-\gamma}]$ where $c \ge 1$, we have $w^{\pi_\alpha^\star} \in \mathcal{W}$ where $\pi_\alpha^\star$ denotes an optimal policy of the optimization problem $\mathcal{P}(\boldsymbol{\tau} - \alpha\mathbf{1})$.*

With the above assumption, we get the following result.

Theorem 3. *Let $\epsilon \in (0, \frac{1}{2}]$ be given. Under assumptions A,B,C,D,E and F, the policy $\bar{\pi}$ returned by the PDCA algorithm (Algorithm 1) with the cost threshold $\boldsymbol{\tau} - \eta\mathbf{1}$, where $\eta = \varphi\epsilon$, and bound $B = \frac{5}{\varphi}$ satisfies $J_{C_i}(\bar{\pi}) \leq \tau_i$ for all $i = 1, \ldots, I$, and $J_R(\pi) \geq J_R(\pi^\star) - \mathcal{O}(\epsilon)$ with probability at least $1 - \delta$, where $\pi^\star$ is an optimal policy for $\mathcal{P}(\boldsymbol{\tau})$ as long as*

$$n \geq \Omega\left(\frac{(C_{\ell_2})^2 \log(I|\mathcal{F}||\Pi||\mathcal{W}|/\delta)}{(1-\gamma)^4\varphi^2\epsilon^2}\right).$$

See Appendix E.3 for a proof.

4 ANALYSIS

In this section, we provide a proof sketch for Theorem 1. We show in Section 4.1 that PDCA finds a near saddle point of the Lagrangian. We show in Section 4.2 that the near saddle point approximately solves the optimization problem (OPT). We use the notation $a \lesssim b$ to indicate $a \leq b + \zeta(n, K)$ and $a \approx b$ to indicate $a = b + \zeta(n, K)$ where $\zeta(n, K) \to 0$ as $n, K \to \infty$.

4.1 PDCA Finds a Near Saddle Point

We show that PDCA run with thresholds $\boldsymbol{\tau}$ and bound B finds a near saddle point of the Lagrangian $L(\pi, \boldsymbol{\lambda}) = J_R(\pi) + \boldsymbol{\lambda} \cdot (\boldsymbol{\tau} - J_{\boldsymbol{C}}(\pi))$: the policy $\bar{\pi}$ returned by PDCA and $\bar{\boldsymbol{\lambda}} = \frac{1}{K}\sum_{k=1}^K \boldsymbol{\lambda}_k$ for sufficiently large K satisfy

$$L(\pi, \bar{\boldsymbol{\lambda}}) \leq L(\bar{\pi}, \boldsymbol{\lambda}) + \mathcal{O}(\epsilon_{\text{stat}})$$

for all $\pi \in \text{Conv}(\Pi)$ with $w^\pi \in \mathcal{W}$ and $\boldsymbol{\lambda} \in B\boldsymbol{\Delta}^I$ where $\epsilon_{\text{stat}} = \mathcal{O}(1/\sqrt{n})$ is a statistical error term for estimating $\mathcal{E}_\mu$ and A_μ. See Appendix B for the full analysis of ϵ_{stat}. To show that $(\bar{\pi}, \bar{\boldsymbol{\lambda}})$ is a near saddle point, we decompose $L(\pi, \bar{\boldsymbol{\lambda}}) - L(\bar{\pi}, \boldsymbol{\lambda})$ into regrets of the π-player and the λ-player, and bound each regret separately as follows. See Appendix C for full proof.

Regret Bound for π-Player Regret of π-player $\frac{1}{K}\sum_{k=1}^K L(\pi, \boldsymbol{\lambda}_k) - \frac{1}{K}\sum_{k=1}^K L(\pi_k, \boldsymbol{\lambda}_k)$ simplifies to $\frac{1}{K}\sum_{k=1}^K (J_R(\pi) - J_R(\pi_k)) + \frac{1}{K}\sum_{k=1}^K \boldsymbol{\lambda}_k \cdot (J_{\boldsymbol{C}}(\pi_k) - J_{\boldsymbol{C}}(\pi))$. Decomposing $J_R(\pi) - J_R(\pi_k)$ by performance difference lemma (Lemma 12 in Cheng et al. (2022)) and using the properties of the critics give

$$(1-\gamma)(J_R(\pi) - J_R(\pi_k)) \lesssim \mathbb{E}_\pi[f_k(s, \pi) - f_k(s, \pi_k)]$$

which shows the performance difference with respect to the reward function R can be upper bounded by the difference of the reward critic. Similarly, we have

$$(1-\gamma)(J_{C_i}(\pi_k) - J_{C_i}(\pi)) \lesssim \mathbb{E}_\pi[g_k^i(s, \pi_k) - g_k^i(s, \pi)],$$

for all $i = 1, \ldots, I$, and it follows that

$$\begin{aligned}&\tfrac{1}{K}\textstyle\sum_{k=1}^K L(\pi, \boldsymbol{\lambda}_k) - \tfrac{1}{K}\sum_{k=1}^K L(\pi_k, \boldsymbol{\lambda}_k) \\ &\quad\lesssim \tfrac{1}{K}\textstyle\sum_{k=1}^K \mathbb{E}_\pi[z_k(s, \pi) - z_k(s, \pi_k)] \leq \tfrac{1}{K}\epsilon_{\text{opt}}(K)\end{aligned}$$

where $z_k = f_k + \boldsymbol{\lambda}_k \cdot (\boldsymbol{\tau} - \boldsymbol{g}_k)$ is the input to policy optimization oracle (Definition 3) used by π-player; the last inequality is by the property of the oracle.

Regret Bound for λ-Player The regret of the λ-player $\frac{1}{K}\sum_{k=1}^K L(\pi_k, \boldsymbol{\lambda}_k) - L(\bar{\pi}, \boldsymbol{\lambda})$ can be simplified to $\frac{1}{K}\sum_{k=1}^K (\boldsymbol{\lambda}_k - \boldsymbol{\lambda}) \cdot (\boldsymbol{\tau} - J_{\boldsymbol{C}}(\pi_k))$. Recall that PDCA calls OPE oracle to estimate $J_{C_i}(\pi_k) \approx h_k^i$. Since the λ-player greedily chooses $\boldsymbol{\lambda}_k \in B\boldsymbol{\Delta}^I$ that minimizes $\boldsymbol{\lambda} \cdot (\boldsymbol{\tau} - \boldsymbol{h}_k)$, we have for all $\boldsymbol{\lambda} \in B\boldsymbol{\Delta}^I$ that

$$\begin{aligned}&\tfrac{1}{K}\textstyle\sum_{k=1}^K (\boldsymbol{\lambda}_k - \boldsymbol{\lambda}) \cdot (\boldsymbol{\tau} - J_{\boldsymbol{C}}(\pi_k)) \\ &\qquad\approx \tfrac{1}{K}\textstyle\sum_{k=1}^K (\boldsymbol{\lambda}_k - \boldsymbol{\lambda}) \cdot (\boldsymbol{\tau} - \boldsymbol{h}_k) \leq 0.\end{aligned}$$

4.2 A Near Saddle Point Nearly solves OPT

We can show that if the Slater's condition (Assumption B) holds, then a near saddle point $(\bar{\pi}, \bar{\boldsymbol{\lambda}})$ of $L(\cdot, \cdot)$ that satisfies $L(\pi, \bar{\boldsymbol{\lambda}}) \leq L(\bar{\pi}, \boldsymbol{\lambda})$ for all $\pi \in \text{Conv}(\Pi)$ with $w^\pi \in \mathcal{W}$ and $\boldsymbol{\lambda} \in B\boldsymbol{\Delta}^I$, then

$$\begin{aligned}&J_R(\bar{\pi}) \geq J_R(\pi^\star) - \xi \\ &J_{C_i}(\bar{\pi}) \leq \tau_i + \tfrac{\xi}{B - 1/\varphi}, \quad i = 1, \ldots, I\end{aligned}$$

where $\pi^\star \in \text{Conv}(\Pi)$ is the optimal policy for $\mathcal{P}(\boldsymbol{\tau})$. See Appendix D for the proof of the above result. Combining the results in Section 4.1 and Section 4.2, Theorem 1 follows. See Appendix E for the full proof.

5 EXPERIMENTS

To demonstrate the empirical performance of our algorithm, we compare with MBCL (Le et al. 2019), the only previous work with provable guarantees, and the following practical algorithms for offline constrained RL: COptiDICE (Lee et al. 2022), CDT (Liu et al. 2023b), CQP (Xu et al. 2022). For comparing with MBCL, we use tabular setting since MBCL can only run with discrete action set. For computational efficiency in solving the min-max optimization problems in Line 2,3 of Algorithm 1, we take $\mathcal{W} = [0, C_\infty]^{\mathcal{S}\times\mathcal{A}}$ where the bound C_∞ is treated as a hyperparameter. Such $\mathcal{W}$ reduces $\mathcal{E}_\mathcal{D}(\pi, f; U)$ to

$$\begin{aligned}C_\infty \max\{&\mathbb{E}_\mathcal{D}[(f(s,a) - U(s,a) - \gamma f(s', \pi))_+], \\ &\mathbb{E}_\mathcal{D}[(U(s,a) + \gamma f(s', \pi) - f(s,a))_+]\}.\end{aligned}$$

Following experimental settings of previous works, we use a single cost signal ($I = 1$) for all experiments. For experimental details, see Appendix G.

5.1 Tabular Experiments

Following Lee et al. (2022), we randomly generate tabular CMDP with 10 states and 5 actions and prepare

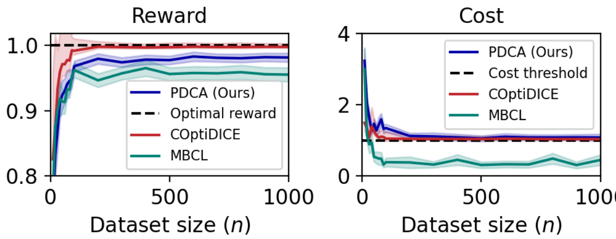

Figure 1: Tabular CMDP experiment

an offline dataset using a data distribution induced by a mixture of uniform policy and the optimal policy. We compare the performance of PDCA to MBCL and COptiDICE on datasets of varying sizes. For each dataset size, we repeat the experiments 10 times and report the average of the reward value and the cost value. Figure 1 shows the result. The shaded region indicates the standard error. Overall, PDCA outperforms MBCL and is comparable to COptiDICE.

5.2 Real-World RL Benchmark Experiments

We follow the experimental setup in Lee et al. (2022) and run the algorithms on 4 environments provided in the Real-World RL (RWRL) suite (Dulac-Arnold et al. 2020). For the benchmark experiments, we use a practical version of PDCA shown in Algorithm 2. We parameterize the function class $\mathcal{F}$ with neural networks. The reward critic uses a neural network f_θ parameterized by θ and each cost critic for the cost C_i uses a neural network $g^i_{\theta^i}$ parameterized by θ^i. For solving the optimization problems, the reward critic uses stochastic gradient descent algorithm with a learning rate η_{fast} on the loss $2\mathcal{E}_\mathcal{D}(f_\theta, \pi) + A_\mathcal{D}(f_\theta, \pi)$. Similarly, the cost critic uses stochastic gradient descent algorithm with the same learning rate η_{fast} on the loss $2\mathcal{E}_\mathcal{D}(f_\theta, \pi) - A_\mathcal{D}(f_\theta, \pi)$. Following the practical version of no-regret policy optimization oracle implemented by Cheng et al. (2022), we use a policy network to parameterize Π. The π-player uses a neural network π_ψ parameterized by ψ and use a stochastic gradient descent algorithm on the loss $-A_\mathcal{D}(f_\theta + \boldsymbol{\lambda}\cdot(\boldsymbol{\tau} - \boldsymbol{g_\theta}), \pi_\psi)$. For the OPE oracle, we use a neural network $h^i_{\vartheta^i}$ parameterized by ϑ^i and use a stochastic gradient descent algorithm on the loss $\mathcal{E}_\mathcal{D}(h^i_{\vartheta^i}, \pi)$ with learning rate η_{fast}. The λ-player acts greedily and chooses $\boldsymbol{\lambda} \in B\boldsymbol{\Delta}^I$ that minimizes $\boldsymbol{\lambda} \cdot (\boldsymbol{\tau} - \boldsymbol{h}_\vartheta(s_0, \pi_\psi))$. See Appendix G.2 for hyperparameter tuning details.

Environments We run experiments on four environments provided in the Real-World RL (RWRL) Benchmark suite (Dulac-Arnold et al. 2020) used by Lee et al. (2022): Cartpole, Walker, Quadruped, and Humanoid. Following Lee et al. (2022), for each environment, we choose the most challenging safety condition among

---

Algorithm 2: Practical Version of PDCA

Input: Dataset $\mathcal{D} = \{(s_j, a_j, s'_j)\}_{j=1}^n$

Init: Network f_θ for reward critic; $g^i_{\theta^i}$, $i = 1, \dots, I$ for cost critics; $h^i_{\vartheta^i}$, $i = 1, \dots, I$ for OPE. Policy network π_ψ.

for $k = 1, 2, \dots, K$ **do**
- Sample a minibatch $\mathcal{D}_{\text{mini}}$ from dataset $\mathcal{D}$.
- `// Update critics`
- $\ell_{\text{reward}}(\theta) = 2\mathcal{E}_{\mathcal{D}_{\text{mini}}}(f_\theta, \pi_\psi) + A_{\mathcal{D}_{\text{mini}}}(f_\theta, \pi_\psi)$.
- $\ell_{\text{cost}}(\theta^i) = 2\mathcal{E}_{\mathcal{D}_{\text{mini}}}(g^i_{\theta^i}, \pi_\psi) - A_{\mathcal{D}_{\text{mini}}}(g^i_{\theta^i}, \pi_\psi)$ for $i = 1, \dots, I$.
- $\ell_{\text{ope}}(\theta^i) = \mathcal{E}_{\mathcal{D}_{\text{mini}}}(g^i_{\theta^i}, \pi_\psi)$ for $i = 1, \dots I$.
- $\theta \leftarrow \text{ADAM}(\pi_\psi, \nabla \ell_{\text{reward}}(\theta), \eta_{\text{fast}})$.
- $\theta^i \leftarrow \text{ADAM}(\pi_\psi, \nabla \ell_{\text{cost}}(\theta^i), \eta_{\text{fast}})$, $i = 1, \dots, I$.
- $\vartheta^i \leftarrow \text{ADAM}(\pi_\psi, \nabla \ell_{\text{ope}}(\vartheta^i), \eta_{\text{fast}})$, $i = 1, \dots, I$.
- `// Update π.`
- $\ell_{\text{actor}}(\psi) = -A_{\mathcal{D}_{\text{mini}}}(f_\theta + \sum_{i=1}^I \lambda_i(\tau_i - g^i_{\theta^i}), \pi_\psi)$.
- $\psi \leftarrow \text{ADAM}(\ell_{\text{actor}}, \eta_{\text{slow}})$.
- `// Update λ.`
- $z_i \leftarrow \tau_i - h^i_{\theta^i}(s_0, \pi_\psi)$, for $i = 1, \dots, I$
- $\lambda_i \leftarrow B$ if $z_i < 0$ otherwise $\lambda_i \leftarrow 0$, $i = 1, \dots, I$.

---

the multiple safety conditions provided by RWRL suite. We give the cost of 1 if the safety condition is violated at each time step. The thresholds on the expected discounted cumulative costs are 0.05 for Cartpole and Walker, and 0.01 for Quadruped and Humanoid. We follow the same safety coefficient parameters (difficulty levels provided by RWRL suite) used by Lee et al. (2022): for Cartpole and Walker we use 0.3, and for Quadruped and Humanoid we use 0.5.

Offline Dataset Generation Since RWRL suite does not provide an offline dataset we generate one for each environment by a policy trained by an online RL algorithm using a reward function penalized by cost function, $R - \lambda C$, where we vary λ. Specifically, for each environment, we choose three different λ values and for each λ, we run the soft actor-critic algorithm (SAC) (Haarnoja et al. 2018) with the reward function $R - \lambda C$. The SAC algorithm is run for 1,000,000 steps. For each policy trained with different λ values, we generate 1,000 trajectories. During trajectory generation, actions are perturbed with Gaussian noise with mean=0 and std=0.15. The three sets of trajectories, one for each λ, are mixed to form an offline dataset consisting of 3,000 trajectories. For the λ values, we use $0.3, 0.8, 1$ for Cartpole, $1, 1.8, 2$ for Walker, $0, 0.1, 0.5$ for Quadruped, $0, 0.4, 0.5$ for Humanoid.

Evaluation At every 1000 iterations, we run the policy online and report the average discounted cumulative reward and cost and their standard errors

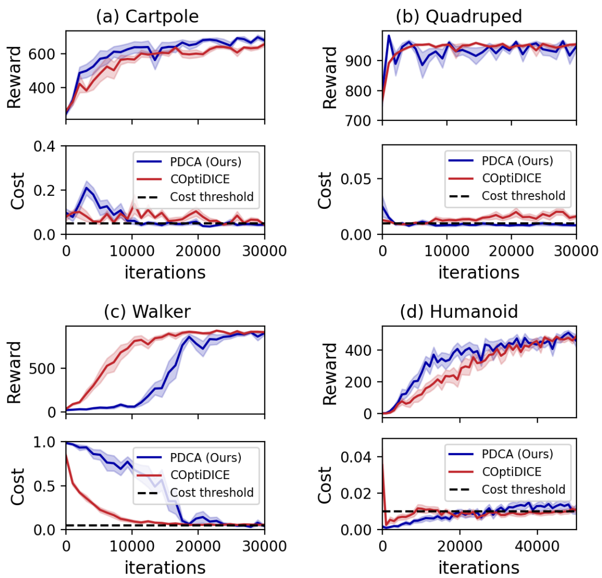

Figure 2: RWRL Benchmark Experiments

of 5 trajectories. Figure 2 shows comparison of the performance of our algorithm and COptiDICE on 4 RWRL environments. The black dotted horizontal line indicates the cost threshold. The blue and red lines indicate the cumulative cost and reward for PDCA and COptiDICE respectively. Overall, the performance of PDCA is comparable to that of COptiDICE.

5.3 Safety Gym Benchmark Experiments

We run PDCA on the bullet safety gym (Gronauer 2022) with offline datasets provided by Liu et al. (2023a) and compare the performance of PDCA to CDT (Liu et al. 2023b), CPQ (Xu et al. 2022) and COptiDICE (Lee et al. 2022). See Appendix G.3 for the details of the offline datasets and hyperparameter tuning procedure.

Environments The Bullet Safety Gym (Gronauer 2022) provides environments based on physics simulator where the agent can move around the physical environment scattered with obstacles. The layout of the obstacles is not fixed and randomly generated in each episode. For our benchmark experiments, we use the three different agents: ball that can move freely on a plane and is controlled by a two-dimensional force vector; car that can control wheel velocities and steering angle; ant that is quadrupedal composed of nine rigid bodies with each leg controlled by two actuators. We use two different tasks. The circle task encourages the agent to move on a circle. The reward signal depends on the speed of the agent and the proximity of the agent to the boundary. Costs are incurred when the agent leaves the circle. The run task rewards the agent for running through an avenue between two safety boundaries. The agent incurs costs when exceeding speed limit.

Table 2: Safety Gym Results (Cost Threshold = 1.00)

Task	CDT	CPQ	C'DICE	PDCA
	R | C	R | C	R | C	R | C
AntCircle	0.54|**1.78**	0.00|0.00	0.17|**5.04**	0.22|**3.53**
AntRun	0.72|0.91	0.03|0.02	0.61|0.94	0.28|0.93
BallCircle	0.77|**1.07**	0.64|0.76	0.70|**2.61**	0.63|**2.29**
BallRun	0.39|**1.16**	0.22|**1.27**	0.59|**3.52**	0.55|**3.38**
CarCircle	0.75|0.95	0.71|0.33	0.49|**3.14**	0.22|**2.42**

Evaluation Following Liu et al. (2023a), we run with cost thresholds set to 10, 20, 40, each with 3 random seeds and report the average performance of the 9 runs. To approximate the uniform mixture of historical policies produced by PDCA, we average the performance of policies taken every 2500 iterations. The performance of each policy is measured by running the policy for 20 episodes and taking the average of the discounted cumulative reward/cost. When reporting, the cost value is normalized so that the cost threshold is scaled to 1. See Table 2 for the results. The column **R** is the reward and **C** the cost averaged over three random seeds and three cost thresholds. Boldfaced numbers indicate cost values that exceed the threshold. The performance of PDCA is generally not dominated by others in the sense that it is not outperformed in terms of both reward and cost violation simultaneously. The transformer based algorithm CDT generally outperforms other algorithms. We believe that this is because CDT learns non-Markovian policies, which may be better suited for the benchmark environments.

6 CONCLUSION

We propose a primal-dual algorithm PDCA for offline constrained RL with function approximation. PDCA is sample-efficient under concentrability, value function realizability and MIW realizability assumptions, which relaxes Bellman completeness assumption required by previous work. PDCA requires all-policy concentrability only to guarantee the concentration bound on the estimates returned by OPE. Relaxing this to the single-policy concentrability assumption is an interesting future work that will likely require using pessimistic estimates for the costs and modifying the strategy of the λ-player to work with pessimistic estimates.

7 ACKNOWLEDGEMENTS

We acknowledge the support of NSF via grant IIS-2007055.

References

Altman, Eitan (1999). *Constrained Markov Decision Processes: Stochastic Modeling.*

Antos, András, Csaba Szepesvári, and Rémi Munos (2008). "Learning near-optimal policies with Bellman-residual minimization based fitted policy iteration and a single sample path". In: *Machine Learning.*

Bai, Qinbo, Amrit Singh Bedi, Mridul Agarwal, Alec Koppel, and Vaneet Aggarwal (2022). "Achieving Zero Constraint Violation for Constrained Reinforcement Learning via Primal-Dual Approach". In: *Proceedings of the AAAI Conference on Artificial Intelligence.*

Boyd, Stephen P. and Lieven Vandenberghe (2004). *Convex optimization.*

Chen, Jinglin and Nan Jiang (2019). "Information-Theoretic Considerations in Batch Reinforcement Learning". In: *Proceedings of the 36th International Conference on Machine Learning.*

– (2022). "Offline reinforcement learning under value and density-ratio realizability: The power of gaps". In: *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence.*

Chen, Lili, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Misha Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch (2021a). "Decision transformer: Reinforcement learning via sequence modeling". In: *Advances in neural information processing systems.*

Chen, Yi, Jing Dong, and Zhaoran Wang (2021b). *A Primal-Dual Approach to Constrained Markov Decision Processes.*

Cheng, Ching-An, Tengyang Xie, Nan Jiang, and Alekh Agarwal (2022). "Adversarially Trained Actor Critic for Offline Reinforcement Learning". In: *Proceedings of the 39th International Conference on Machine Learning.*

Ding, Dongsheng, Kaiqing Zhang, Tamer Basar, and Mihailo Jovanovic (2020). "Natural Policy Gradient Primal-Dual Method for Constrained Markov Decision Processes". In: *Advances in Neural Information Processing Systems.*

Dulac-Arnold, Gabriel, Nir Levine, Daniel J. Mankowitz, Jerry Li, Cosmin Paduraru, Sven Gowal, and Todd Hester (2020). "An empirical investigation of the challenges of real-world reinforcement learning". In.

Foster, Dylan J., Akshay Krishnamurthy, David Simchi-Levi, and Yunzong Xu (2022). *Offline Reinforcement Learning: Fundamental Barriers for Value Function Approximation.*

Gronauer, Sven (2022). *Bullet-Safety-Gym: A Framework for Constrained Reinforcement Learning.* Tech. rep. mediaTUM.

Gu, Shixiang, Ethan Holly, Timothy Lillicrap, and Sergey Levine (2017). "Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates". In: *2017 IEEE International Conference on Robotics and Automation (ICRA).*

Haarnoja, Tuomas, Aurick Zhou, Pieter Abbeel, and Sergey Levine (2018). "Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor". In: *Proceedings of the 35th International Conference on Machine Learning.*

Kakade, Sham and John Langford (2002). "Approximately Optimal Approximate Reinforcement Learning". In: *Proceedings of the Nineteenth International Conference on Machine Learning.*

Kakade, Sham M (2001). "A Natural Policy Gradient". In: *Advances in Neural Information Processing Systems.*

Kumar, Aviral, Anikait Singh, Stephen Tian, Chelsea Finn, and Sergey Levine (2021). "A Workflow for Offline Model-Free Robotic Reinforcement Learning". In.

Le, Hoang, Cameron Voloshin, and Yisong Yue (2019). "Batch Policy Learning under Constraints". In: *Proceedings of the 36th International Conference on Machine Learning.*

Lee, Jongmin, Wonseok Jeon, Byungjun Lee, Joelle Pineau, and Kee-Eung Kim (2021). "OptiDICE: Offline Policy Optimization via Stationary Distribution Correction Estimation". In: *Proceedings of the 38th International Conference on Machine Learning.*

Lee, Jongmin, Cosmin Paduraru, Daniel J. Mankowitz, Nicolas Heess, Doina Precup, Kee-Eung Kim, and Arthur Guez (2022). "COptiDICE: Offline Constrained Reinforcement Learning via Stationary Distribution Correction Estimation". In.

Levine, Sergey, Aviral Kumar, George Tucker, and Justin Fu (2020). *Offline Reinforcement Learning: Tutorial, Review, and Perspectives on Open Problems.*

Levine, Sergey, Peter Pastor, Alex Krizhevsky, Julian Ibarz, and Deirdre Quillen (2018). "Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection". In: *The International Journal of Robotics Research.*

Liu, Zuxin, Zijian Guo, Haohong Lin, Yihang Yao, Jiacheng Zhu, Zhepeng Cen, Hanjiang Hu, Wenhao Yu, Tingnan Zhang, Jie Tan, et al. (2023a). "Datasets and Benchmarks for Offline Safe Reinforcement Learning". In: *arXiv preprint arXiv:2306.09303.*

Liu, Zuxin, Zijian Guo, Yihang Yao, Zhepeng Cen, Wenhao Yu, Tingnan Zhang, and Ding Zhao (2023b). "Constrained decision transformer for offline safe reinforcement learning". In: *arXiv preprint arXiv:2302.07351*.

Munos, Rémi (2003). "Error bounds for approximate policy iteration". In: *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*.

– (2005). "Error bounds for approximate value iteration". In: *Proceedings of the 20th national conference on Artificial intelligence - Volume 2*.

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

Ozdaglar, Asuman E, Sarath Pattathil, Jiawei Zhang, and Kaiqing Zhang (2023). "Revisiting the linear-programming framework for offline rl with general function approximation". In: *International Conference on Machine Learning*. PMLR.

Tang, Shengpu and Jenna Wiens (2021). "Model Selection for Offline Reinforcement Learning: Practical Considerations for Healthcare Settings". In: *Proceedings of the 6th Machine Learning for Healthcare Conference*.

Wang, Wei, Nanpeng Yu, Yuanqi Gao, and Jie Shi (2020). "Safe Off-Policy Deep Reinforcement Learning Algorithm for Volt-VAR Control in Power Distribution Systems". In: *IEEE Transactions on Smart Grid*.

Xie, Tengyang, Ching-An Cheng, Nan Jiang, Paul Mineiro, and Alekh Agarwal (2021a). "Bellman-consistent pessimism for offline reinforcement learning". In: *Advances in neural information processing systems*.

Xie, Tengyang and Nan Jiang (2021b). "Batch Value-function Approximation with Only Realizability". In: *Proceedings of the 38th International Conference on Machine Learning*.

– (2020). "Q* Approximation Schemes for Batch Reinforcement Learning: A Theoretical Comparison". In: *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*.

Xu, Haoran, Xianyuan Zhan, and Xiangyu Zhu (2022). "Constraints penalized q-learning for safe offline reinforcement learning". In: *Proceedings of the AAAI Conference on Artificial Intelligence*.

Zanette, Andrea (2022). *When is Realizability Sufficient for Off-Policy Reinforcement Learning?*

Zanette, Andrea and Martin J. Wainwright (2022). "Bellman Residual Orthogonalization for Offline Reinforcement Learning". In.

Zhan, Wenhao, Baihe Huang, Audrey Huang, Nan Jiang, and Jason Lee (2022). "Offline Reinforcement Learning with Realizability and Single-policy Concentrability". In: *Proceedings of Thirty Fifth Conference on Learning Theory*.

Zhu, Hanlin, Paria Rashidinejad, and Jiantao Jiao (2023). *Importance Weighted Actor-Critic for Optimal Conservative Offline Reinforcement Learning.*

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. *Yes. Section 2 describes problem setting and assumptions. Section 3 describes algorithm.*
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. *Yes. Section 3 (Theorem 1) shows sample complexity and Section 4 gives proof sketch.*
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. *Yes. Provided as supplemntary material.*

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. *Yes. Section 2 gives full set of assumptions. Section 3 (Theorem 1) states the theoretical sample complexity bound and assumptions*
   (b) Complete proofs of all theoretical results. *Yes. Complete proofs are given in appendix.*
   (c) Clear explanations of any assumptions. *Yes. Section 2 explains assumptions required by theoretical results.*

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). *Yes. Code and instructions in the suppllmentary material*
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). *Yes. Training details are given in Section 5. More details are provided in Appendix G*
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). *Yes. See Section 5*
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). *Yes. See first paragraph of Appendix G*

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. *Yes. cited: Liu et al. (2023a)*
   (b) The license information of the assets, if applicable. *Not Applicable*
   (c) New assets either in the supplemental material or as a URL, if applicable. *Not Applicable*
   (d) Information about consent from data providers/curators. *Not Applicable*
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. *Not Applicable*

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. *Not Applicable*
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. *Not Applicable*
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. *Not Applicable*

Supplementary Materials

A PERFORMANCE DIFFERENCE LEMMAS

In this section, we provide two generalizations of the classical performance difference lemma (Kakade et al. 2002). For completeness, we first state the classical performance difference lemma below.

Lemma 1 (Performance Difference Lemma. Kakade et al. (2002))**.**

$$(1-\gamma)(J_U(\widehat{\pi}) - J_U(\pi)) = A_\pi(\widehat{\pi}, Q_U^{\widehat{\pi}}) \tag{4}$$

The following is the first generalization of the performance difference lemma. It decomposes the difference in performance of two policies, where the performance of one of the policies is measured with respect to an arbitrary Q-value function f. The same result is proved as an intermediate step in the proof of Lemma 12 in Cheng et al. (2022). We state it as a separate lemma and provide a simplified proof below.

Lemma 2. *For any functions $U, f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ and any policies $\pi, \widehat{\pi} : \mathcal{S} \to \Delta(\mathcal{A})$, we have*

$$(1-\gamma)(f(s_0, \widehat{\pi}) - J_U(\pi)) = A_\pi(\widehat{\pi}, f) + \mathbb{E}_\pi[(f - \mathcal{T}_U^{\widehat{\pi}} f)(s,a)].$$

Proof. Note that

$$\begin{aligned}
\mathbb{E}_{(s,a)\sim d^\pi}[f(s,\widehat{\pi})] &= (1-\gamma)\mathbb{E}^\pi\left[\sum_{t=0}^{\infty} \gamma^t f(s_t, \widehat{\pi})\right] \\
&= (1-\gamma)\mathbb{E}^\pi\left[f(s_0,\widehat{\pi}) + \sum_{t=0}^{\infty} \gamma^t \mathbb{E}^\pi\left[\gamma f(s_{t+1}, \widehat{\pi}) \mid s_t, a_t\right]\right] \\
&= (1-\gamma) f(s_0, \widehat{\pi}) + (1-\gamma)\mathbb{E}^\pi\left[\sum_{t=0}^{\infty} \gamma^t \mathcal{T}_0^{\widehat{\pi}} f(s_t, a_t)\right] \\
&= (1-\gamma) f(s_0, \widehat{\pi}) + \mathbb{E}_{(s,a)\sim d^\pi}[\mathcal{T}_0^{\widehat{\pi}} f(s,a)].
\end{aligned}$$

Rearranging, we get

$$\begin{aligned}
(1-\gamma) f(s_0, \widehat{\pi}) &= \mathbb{E}_{(s,a)\sim d^\pi}[(f - \mathcal{T}_0^{\widehat{\pi}} f)(s,a)] - \mathbb{E}_{(s,a)\sim d^\pi}[f(s,a) - f(s,\widehat{\pi})] \\
&= \mathbb{E}_{(s,a)\sim d^\pi}[(f - \mathcal{T}_0^{\widehat{\pi}} f)(s,a)] + A_\pi(\widehat{\pi}, f).
\end{aligned}$$

Using $(1-\gamma)J_U(\pi) = \mathbb{E}_{(s,a)\sim d^\pi}[U(s,a)]$, we get

$$\begin{aligned}
(1-\gamma)(f(s_0,\widehat{\pi}) - J_U(\pi)) &= \mathbb{E}_\pi[(f - \mathcal{T}_0^{\widehat{\pi}} f)(s,a)] + A_\pi(\widehat{\pi}, f) - \mathbb{E}_\pi[U(s,a)] \\
&= A_\pi(\widehat{\pi}, f) + \mathbb{E}_\pi[(f - \mathcal{T}_U^{\widehat{\pi}} f)(s,a)].
\end{aligned}$$

□

Note that when we set $f = Q_U^{\widehat{\pi}}$ in the lemma above, we recover the classical performance difference lemma. Now, we state the second generalization of the performance difference lemma. The same lemma is stated and proved in Cheng et al. (2022) and also used in Zhu et al. (2023). We state the lemma and provide a simpler proof below.

Lemma 3 (Performance difference decomposition. Lemma 12 in Cheng et al. (2022))**.** *For any policies* $\pi, \widehat{\pi}, \mu : \mathcal{S} \to \Delta(\mathcal{A})$ *and any functions* $U : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ *and* $f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$*, we have*

$$\begin{aligned}(1-\gamma)(J_U(\pi) - J_U(\widehat{\pi})) = \mathbb{E}_\mu[(f - \mathcal{T}_U^{\widehat{\pi}} f)(s,a)] + \mathbb{E}_\pi[(\mathcal{T}_U^{\widehat{\pi}} f - f)(s,a)] \\ + \mathbb{E}_\pi[f(s,\pi) - f(s,\widehat{\pi})] + A_\mu(\widehat{\pi}, f) - A_\mu(\widehat{\pi}, Q_U^{\widehat{\pi}}).\end{aligned}$$

Proof. We have

$$\begin{aligned}&(1-\gamma)(J_U(\pi) - J_U(\widehat{\pi})) \\ &\quad = (1-\gamma)(J_U(\pi) - f(s_0,\widehat{\pi})) + (1-\gamma)(f(s_0,\widehat{\pi}) - J_U(\mu)) + (1-\gamma)(J_U(\mu) - J_U(\widehat{\pi})) \\ &\quad = -(A_\pi(\widehat{\pi}, f) + \mathbb{E}_\pi[(f - \mathcal{T}_U^{\widehat{\pi}} f)(s,a)]) + (A_\mu(\widehat{\pi}, f) + \mathbb{E}_\mu[(f - \mathcal{T}_U^{\widehat{\pi}} f)(s,a)]) - A_\mu(\widehat{\pi}, Q_U^{\pi})\end{aligned}$$

where the second inequality uses the generalized performance difference lemma (Lemma 2) for the first two terms and the classical performance difference lemma (Lemma 1) for the third term. Rearranging and observing that $A_\pi(\widehat{\pi}, f) = \mathbb{E}_\pi[f(s,\widehat{\pi}) - f(s,\pi)]$ complete the proof. □

Indeed, the lemma above is a generalization because setting $\mu = \pi$ reduces to the classical performance difference lemma (Lemma 1).

B CONCENTRATION INEQUALITIES

In this section, we provide concentration inequalities for relating $\mathcal{E}_\mu$ and A_μ to the empirical versions $\mathcal{E}_\mathcal{D}$ and $A_\mathcal{D}$ respectively. First, we show a concentration bound on $\mathbb{E}_\mathcal{D}[w(s,a)(f(s,a) - U(s,a) - \gamma f(s',\pi))]$, which will be used to show a concentration bound on $\mathcal{E}_\mathcal{D}(\pi, f, U)$.

Lemma 4 (Concentration of Bellman error)**.** *Let* $w : \mathcal{S} \times \mathcal{A} \to \mathbb{R}_+$ *with* $\|w\| \le C_\infty$ *and* $\|w\|_{2,\mu} \le C_{\ell_2}$*. Let* $f : \mathcal{S} \times \mathcal{A} \to [0, \frac{1}{1-\gamma}]$ *and* $U : \mathcal{S} \times \mathcal{A} \to [0,1]$ *be any functions. Let* $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ *be any policy. With probability at least* $1-\delta$*, we have*

$$\begin{aligned}&|\mathbb{E}_\mu[w(s,a)(f - \mathcal{T}_U^\pi)(s,a)] - \mathbb{E}_\mathcal{D}[w(s,a)(f(s,a) - U(s,a) - \gamma f(s',\pi))]| \\ &\qquad \le \mathcal{O}\left(\frac{C_{\ell_2}}{1-\gamma}\sqrt{\frac{\log(1/\delta)}{n}} + \frac{C_\infty}{1-\gamma}\frac{\log(1/\delta)}{n}\right).\end{aligned}$$

Proof. Define

$$X_j = w(s_j,a_j)(f(s_j,a_j) - U(s_j,a_j) - \gamma f(s_j',\pi)).$$

Note that $\mathbb{E}_\mathcal{D}[w(s,a)(f(s,a) - U(s,a) - \gamma f(s',\pi))] = \frac{1}{n}\sum_{j=1}^n X_j$. By assumption of the data distribution, $X_1, \ldots, X_n$ are i.i.d. By the boundedness assumption on w and f, we have $|X_j| \le \mathcal{O}(C_\infty/(1-\gamma))$. Also, we have

$$\begin{aligned}\mathbb{E}[X_j] &= \mathbb{E}_{(s,a)\sim d^\mu, s'\sim P(\cdot|s,a)}[w(s,a)(f(s,a) - U(s,a) - \gamma f(s',\pi))] \\ &= \mathbb{E}_{(s,a)\sim d^\mu}[\mathbb{E}_{s'\sim P(\cdot|s,a)}[w(s,a)(f(s,a) - U(s,a) - \gamma f(s',\pi))|s,a]] \\ &= \mathbb{E}_\mu[w(s,a)(f - \mathcal{T}_U^\pi f)(s,a)].\end{aligned}$$

By Bernstein's inequality, we have with probability at least $1-\delta$ that

$$\begin{aligned}&\left|\mathbb{E}_\mu[w(s,a)(f - \mathcal{T}_U^\pi f)(s,a)] - \frac{1}{n}\sum_{j=1}^n X_j\right| \\ &\qquad \le \mathcal{O}\left(\sqrt{\frac{\mathrm{Var}_\mu[w(s,a)(f - \mathcal{T}_U^\pi f)(s,a)]\log(1/\delta)}{n}} + \frac{C_\infty \log(1/\delta)}{(1-\gamma)n}\right)\end{aligned}$$

The variance term $\mathrm{Var}_\mu[w(s,a)(f - \mathcal{T}_U^\pi f)(s,a)]$ can be bounded as follows.

$$\begin{aligned}\mathrm{Var}_\mu[w(s,a)(f - \mathcal{T}_U^\pi f)(s,a)] &\le \mathbb{E}_\mu[w(s,a)^2(f - \mathcal{T}_U^\pi f)^2(s,a)] \\ &\le \mathcal{O}(\|w\|_{2,\mu}^2/(1-\gamma)^2) \\ &\le \mathcal{O}((C_{\ell_2})^2/(1-\gamma)^2)\end{aligned}$$

where the second inequality uses the boundedness assumption on f and the last inequality uses the boundedness assumption on w. Hence, we have

$$\left|\mathbb{E}_\mu[w(s,a)(f-\mathcal{T}_U^\pi f)(s,a)] - \frac{1}{n}\sum_{j=1}^n X_j\right| \le \mathcal{O}\left(\frac{C_{\ell_2}}{1-\gamma}\sqrt{\frac{\log(1/\delta)}{n}} + \frac{C_\infty \log(1/\delta)}{(1-\gamma)n}\right).$$

This completes the proof. □

The following lemma relates $\mathcal{E}_\mu$ to $\mathcal{E}_\mathcal{D}$. The proof closely follows that of Lemma 4 in Zhu et al. (2023), which shows the same result for a single reward function.

Lemma 5 (Concentration of Bellman error term)**.** *Under Assumption D, with probability at least $1-\delta$, we have*

$$|\mathcal{E}_\mu(\pi,f;R) - \mathcal{E}_\mathcal{D}(\pi,f;R)| \le \mathcal{O}(\epsilon_{stat})$$
$$|\mathcal{E}_\mu(\pi,g_i;C_i) - \mathcal{E}_\mathcal{D}(\pi,g_i;C_i)| \le \mathcal{O}(\epsilon_{stat}) \quad \text{for all } i=1,\dots,I,$$

for any $\pi\in\Pi$, $f\in\mathcal{F}$ and $g_i\in\mathcal{F}$, $i=1,\dots,I$, where

$$\epsilon_{stat} := \frac{C_{\ell_2}}{1-\gamma}\sqrt{\frac{\log(I|\mathcal{F}||\Pi||\mathcal{W}|/\delta)}{n}} + \frac{C_\infty\log(I|\mathcal{F}||\Pi||\mathcal{W}|/\delta)}{(1-\gamma)n}.$$

Proof. It is enough to show that, with probability $1-\delta$, we have $|\mathcal{E}_\mu(\pi,h;U)-\mathcal{E}_\mathcal{D}(\pi,h;U)| \le \mathcal{O}(\epsilon_{\text{stat}})$ for any policy $\pi\in\Pi$, any function $h\in\mathcal{F}$ and any function $U\in\{R,C_1,\dots,C_I\}$. Fix $\pi\in\Pi$, $h\in\mathcal{F}$ and $U\in\{R,C_1,\dots,C_I\}$, and define

$$X_j(w) = w(s_j,a_j)(h(s_j,a_j) - U(s_j,a_j) - \gamma h(s_j',\pi))$$

where $w\in\mathcal{W}$. By Lemma 4, we have

$$\left|\mathbb{E}_\mu[w(s,a)(h-\mathcal{T}_U^\pi h)(s,a)] - \frac{1}{n}\sum_{j=1}^n X_j(w)\right| \le \mathcal{O}\left(\frac{C_{\ell_2}}{1-\gamma}\sqrt{\frac{\log(1/\delta)}{n}} + \frac{C_\infty \log(1/\delta)}{(1-\gamma)n}\right).$$

Since the inequality above holds for all $w\in\mathcal{W}$, it follows by a union bound over $w\in\mathcal{W}$ that

$$\begin{aligned}
\mathcal{E}_\mu(\pi,h;U) - \mathcal{E}_\mathcal{D}(\pi,h;U) &= \max_{w\in\mathcal{W}} |\mathbb{E}_\mu[w(s,a)(h-\mathcal{T}_U^\pi h)(s,a)]| \\
&\quad - \max_{w\in\mathcal{W}} |\mathbb{E}_\mathcal{D}[w(s,a)(h(s,a)-U(s,a)-\gamma h(s',\pi))]| \\
&\le |\mathbb{E}_\mu[w^\star(s,a)(h-\mathcal{T}_U^\pi h)(s,a)]| - |\mathbb{E}_\mathcal{D}[w^\star(s,a)(h(s,a)-U(s,a)-\gamma h(s',\pi))]| \\
&\le |\mathbb{E}_\mu[w^\star(s,a)(h-\mathcal{T}_U^\pi h)(s,a)] - \mathbb{E}_\mathcal{D}[w^\star(s,a)(h(s,a)-U(s,a)-\gamma h(s',\pi))]| \\
&= \left|\mathbb{E}_\mu[w^\star(s,a)(h-\mathcal{T}_U^\pi h)(s,a)] - \frac{1}{n}\sum_{j=1}^n X_j(w^\star)\right| \\
&\le \mathcal{O}\left(\frac{C_{\ell_2}}{1-\gamma}\sqrt{\frac{\log(|\mathcal{W}|/\delta)}{n}} + \frac{C_\infty\log(|\mathcal{W}|/\delta)}{(1-\gamma)n}\right)
\end{aligned}$$

where in the first inequality, we use the notation $w^\star = \operatorname{argmax}_{w\in\mathcal{W}} |\mathbb{E}_\mu[w(s,a)(f-\mathcal{T}_U^\pi f)(s,a)]|$; the second inequality follows by the identity $|a|-|b| \le |a-b|$; and the last inequality uses the previous result. The bound of $\mathcal{E}_\mathcal{D}(\pi,h;U) - \mathcal{E}_\mu(\pi,h;U)$ follows similarly, and the union bound of the two bounds gives

$$|\mathcal{E}_\mu(\pi,h;U) - \mathcal{E}_\mathcal{D}(\pi,h;U)| \le \mathcal{O}\left(\frac{C_{\ell_2}}{1-\gamma}\sqrt{\frac{\log(|\mathcal{W}|/\delta)}{n}} + \frac{C_\infty\log(|\mathcal{W}|/\delta)}{(1-\gamma)n}\right).$$

A union bound on all $(h,\pi,U)\in\mathcal{F}\times\Pi\times\{R,C_1,\dots,C_I\}$ completes the proof. □

The following lemma relates A_μ to $A_\mathcal{D}$. The proof closely follows that of Lemma 5 in Zhu et al. (2023), which shows the same result for a single reward function.

Lemma 6 (Concentration of the advantage function)**.** *With probability at least $1-\delta$, for any $\pi \in \Pi$, $f \in \mathcal{F}$, we have*

$$|A_\mu(\pi, f) - A_{\mathcal{D}}(\pi, f)| \le \mathcal{O}\left(\sqrt{\frac{\log(|\mathcal{F}||\Pi|/\delta)}{n(1-\gamma)^2}}\right) \le \mathcal{O}(\epsilon_{stat})$$

where ϵ_{stat} is defined in Lemma 5.

Proof. Note that $\mathbb{E}[A_{\mathcal{D}}(\pi, f)] = A_\mu(\pi, f)$ and $|f(s_i, \pi) - f(s_i, a_i)| \le \mathcal{O}(\frac{1}{1-\gamma})$. Fixing $\pi \in \mathcal{F}$ and $f \in \mathcal{F}$ and applying Hoeffding's inequality, we have with probability at least $1-\delta$ that

$$|A_\mu(\pi, f) - A_{\mathcal{D}}(\pi, f)| \le \mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n(1-\gamma)^2}}\right).$$

Applying union bound on $(\pi, f) \in \Pi \times \mathcal{F}$ completes the proof. □

C PDCA PRODUCES A NEAR SADDLE POINT

In this section, we show that our algorithm PDCA (Algorithm 1) produces a near saddle point.

Lemma 7. *Consider the policy $\bar{\pi}$ produced by the algorithm PDCA (Algorithm 1) with threshold $\boldsymbol{\tau}$ and bound B. Let $\bar{\boldsymbol{\lambda}} := \frac{1}{K}\sum_{k=1}^{K} \boldsymbol{\lambda}_k$ where $\boldsymbol{\lambda}_1, \dots, \boldsymbol{\lambda}_K \in B\boldsymbol{\Delta}^I$ is the sequence of Lagrange multipliers produced by the λ-player in the algorithm. Then, under Assumption A,C and D, with probability at least $1-\delta$, we have for any $\pi \in Conv(\Pi)$ satisfying $w^\pi \in \mathcal{W}$ and any $\boldsymbol{\lambda} \in B\boldsymbol{\Delta}^I$ that*

$$L(\pi, \bar{\boldsymbol{\lambda}}) \le L(\bar{\pi}, \boldsymbol{\lambda}) + \frac{1+2B}{(1-\gamma)^2}\epsilon_{opt}(K) + \mathcal{O}\left(\frac{B\epsilon_{stat}}{1-\gamma}\right).$$

We prove Lemma 7 by bounding the regrets of the π-player and λ-player against their best actions in hindsight.

C.1 Bounding the Regret of the π-Player

Lemma 8. *Under Assumption A and D, with probability at least $1-\delta$, the sequences $\pi_1, \dots, \pi_K \in \Pi$ and $\boldsymbol{\lambda}_1, \dots, \boldsymbol{\lambda}_K \in B\boldsymbol{\Delta}^I$ produced by the algorithm* PDCA *using the cost threshold $\boldsymbol{\tau}$ and the bound B satisfy*

$$L(\pi, \bar{\boldsymbol{\lambda}}) - \frac{1}{K}\sum_{k=1}^{K} L(\pi_k, \boldsymbol{\lambda}_k) \le \frac{1+2B}{(1-\gamma)^2}\epsilon_{opt}(K) + \mathcal{O}\left(\frac{B\epsilon_{stat}}{1-\gamma}\right)$$

for all $\pi \in Conv(\Pi)$ with $w^\pi \in \mathcal{W}$ where $L(\pi, \boldsymbol{\lambda}) = J_R(\pi) + \boldsymbol{\lambda} \cdot (\boldsymbol{\tau} - J_{\boldsymbol{C}}(\pi))$.

Proof. Fix a policy $\pi \in \mathrm{Conv}(\Pi)$ satisfying $w^\pi \in \mathcal{W}$. By the definition of $L(\cdot,\cdot)$, we get

$$L(\pi, \bar{\boldsymbol{\lambda}}) - \frac{1}{K}\sum_{k=1}^{K} L(\pi_k, \boldsymbol{\lambda}_k) = \frac{1}{K}\sum_{k=1}^{K}\underbrace{(J_R(\pi) - J_R(\pi_k))}_{(a)} + \frac{1}{K}\sum_{k=1}^{K}\sum_{i=1}^{I}\lambda_k^i\underbrace{(J_{C_i}(\pi_k) - J_{C_i}(\pi))}_{(b)}.$$

By a union bound with δ scaled appropriately, the concentration bounds for $\mathcal{E}_{\mathcal{D}}$ and $A_{\mathcal{D}}$ in Lemma 5 and Lemma 6, and the regret bound of the oracle used by the π-player (Definition 3) hold with probability at least $1-\delta$. For the rest of the proof, we assume these events hold.

Bounding (a) We use the performance difference lemma (Lemma 3) to bound (a) as follows.

$$\begin{aligned}
&(1-\gamma)(J_R(\pi) - J_R(\pi_k)) \\
&\quad= \mathbb{E}_\mu[(f_k - \mathcal{T}_R^{\pi_k} f_k)(s,a)] + \mathbb{E}_\pi[(\mathcal{T}_R^{\pi_k} f_k - f_k)(s,a)] \\
&\qquad + \mathbb{E}_\pi[f_k(s,\pi) - f_k(s,\pi_k)] + A_\mu(\pi_k, f_k) - A_\mu(\pi_k, Q_R^{\pi_k}) \\
&\quad\le 2\mathcal{E}_\mu(\pi_k, f_k; R) + \mathbb{E}_\pi[f_k(s,\pi) - f_k(s,\pi_k)] + A_\mu(\pi_k, f_k) - A_\mu(\pi_k, Q_R^{\pi_k}) \\
&\quad\le \mathbb{E}_\pi[f_k(s,\pi) - f_k(s,\pi_k)] + \underbrace{2\mathcal{E}_{\mathcal{D}}(\pi_k, f_k; R) + A_{\mathcal{D}}(\pi_k, f_k)}_{(\star)} - A_{\mathcal{D}}(\pi_k, Q_R^{\pi_k}) + \mathcal{O}(\epsilon_{\mathrm{stat}})
\end{aligned}$$

where the first inequality follows by $w^\pi \in \mathcal{W}$ which implies $|\mathbb{E}_\pi[(f - \mathcal{T}_U^\pi f)(s,a)]| \le \mathcal{E}_\mu(\pi, f; U)$; and the second inequality follows by the concentration results in Lemma 5 and Lemma 6. Recall that the reward critic chooses $f_k \in \mathcal{F}$ that minimizes $2\mathcal{E}_\mathcal{D}(\pi_k, \cdot; R) + A_\mathcal{D}(\pi_k, \cdot)$. We have $Q_R^{\pi_k} \in \mathcal{F}$ by the realizability assumption (Assumption A). Hence,

$$2\mathcal{E}_\mathcal{D}(\pi_k, f_k; R) + A_\mathcal{D}(\pi_k, f_k) \le 2\mathcal{E}_\mathcal{D}(\pi_k, Q_R^{\pi_k}; R) + A_\mathcal{D}(\pi_k, Q_R^{\pi_k}).$$

Using this inequality for bounding $(\star)$ and continuing the bound of $J_R(\pi) - J_R(\pi_k)$, we get

$$\begin{aligned}(1-\gamma)(J_R(\pi) - J_R(\pi_k)) &\le \mathbb{E}_\pi[f_k(s,\pi) - f_k(s,\pi_k)] + 2\mathcal{E}_\mathcal{D}(\pi_k, Q_R^{\pi_k}; R) + \mathcal{O}(\epsilon_{\text{stat}}) \\ &\le \mathbb{E}_\pi[f_k(s,\pi) - f_k(s,\pi_k)] + 2\mathcal{E}_\mu(\pi_k, Q_R^{\pi_k}; R) + \mathcal{O}(\epsilon_{\text{stat}}) \\ &= \mathbb{E}_\pi[f_k(s,\pi) - f_k(s,\pi_k)] + \mathcal{O}(\epsilon_{\text{stat}})\end{aligned}$$

where the last equality uses the fact that Q_R^π solves $f - \mathcal{T}_R^\pi f = 0$, which gives $\mathcal{E}_\mu(\pi_k, Q_R^{\pi_k}; R) = \max_{w\in\mathcal{W}} |\mathbb{E}_\mu[w(s,a)(Q_R^{\pi_k} - \mathcal{T}_R^{\pi_k} Q_R^{\pi_k})]| = 0$.

Bounding (b) Similarly, we can bound (b) as follows.

$$\begin{aligned}&(1-\gamma)(J_{C_i}(\pi_k) - J_{C_i}(\pi)) \\ &\quad= \mathbb{E}_\mu[(\mathcal{T}_{C_i}^{\pi_k} g_k^i - g_k^i)(s,a)] + \mathbb{E}_\pi[(g_k^i - \mathcal{T}_{C_i}^{\pi_k} g_k^i)(s,a)] \\ &\qquad\quad + \mathbb{E}_\pi[g_k^i(s,\pi_k) - g_k^i(s,\pi)] - A_\mu(\pi_k, g_k^i) + A_\mu(\pi_k, Q_{C_i}^{\pi_k}) \\ &\quad\le 2\mathcal{E}_\mu(\pi_k, g_k^i; C_i) + \mathbb{E}_\pi[g_k^i(s,\pi_k) - g_k^i(s,\pi)] - A_\mu(\pi_k, g_k^i) + A_\mu(\pi_k, Q_{C_i}^{\pi_k}) \\ &\quad\le \mathbb{E}_\pi[g_k^i(s,\pi_k) - g_k^i(s,\pi)] + 2\mathcal{E}_\mathcal{D}(\pi_k, g_k^i; C_i) - A_\mathcal{D}(\pi_k, g_k^i) + A_\mathcal{D}(\pi_k, Q_{C_i}^{\pi_k}) + \mathcal{O}(\epsilon_{\text{stat}}) \\ &\quad\le \mathbb{E}_\pi[g_k^i(s,\pi_k) - g_k^i(s,\pi)] + 2\mathcal{E}_\mathcal{D}(\pi_k, Q_{C_i}^{\pi_k}; C_i) + \mathcal{O}(\epsilon_{\text{stat}}) \\ &\quad\le \mathbb{E}_\pi[g_k^i(s,\pi_k) - g_k^i(s,\pi)] + 2\mathcal{E}_\mu(\pi_k, Q_{C_i}^{\pi_k}; C_i) + \mathcal{O}(\epsilon_{\text{stat}}) \\ &\quad= \mathbb{E}_\pi[g_k^i(s,\pi_k) - g_k^i(s,\pi)] + \mathcal{O}(\epsilon_{\text{stat}})\end{aligned}$$

where the third inequality uses the realizability assumption (Assumption A) for $Q_{C_i}^{\pi_k} \in \mathcal{F}$ and the fact that the cost critic chooses $g_k^i \in \mathcal{F}$ that minimizes $2\mathcal{E}_\mathcal{D}(\pi_k, \cdot; C_i) - A_\mathcal{D}(\pi_k, \cdot)$.

Using the Property of π-Player Using the bounds for (a) and (b) and continuing, we get

$$\begin{aligned}&\frac{1-\gamma}{K}\sum_{k=1}^K (L(\pi, \boldsymbol{\lambda}_k) - L(\pi_k, \boldsymbol{\lambda}_k)) \\ &\qquad= \frac{1-\gamma}{K}\sum_{k=1}^K (J_R(\pi) - J_R(\pi_k)) + \frac{1-\gamma}{K}\sum_{k=1}^K\sum_{i=1}^I \lambda_k^i (J_{C_i}(\pi_k) - J_{C_i}(\pi)) \\ &\qquad\le \frac{1}{K}\sum_{k=1}^K \left(\mathbb{E}_\pi[f_k(s,\pi) - f_k(s,\pi_k)] + \sum_{i=1}^I \lambda_k^i \mathbb{E}_\pi[g_k^i(s,\pi_k) - g_k^i(s,\pi)]\right) + \mathcal{O}(B\epsilon_{\text{stat}}) \\ &\qquad= \frac{1}{K}\sum_{k=1}^K \mathbb{E}_\pi[z_k(s,\pi) - z_k(s,\pi_k)] + \mathcal{O}(B\epsilon_{\text{stat}}) \\ &\qquad\le \frac{1+2B}{1-\gamma}\epsilon_{\text{opt}}(K) + \mathcal{O}(B\epsilon_{\text{stat}})\end{aligned}$$

where $z_k = f_k + \sum_{i=1}^I \lambda_k^i(\tau_i - g_k^i)$ and the last inequality follows by the property of the policy optimization oracle (Definition 3) employed by the π-player and the fact that $|z_k(s,a)| \le \frac{1+2B}{1-\gamma}$ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$. Rearranging completes the proof. □

C.2 Bound the Regret of the λ-Player

Lemma 9. *Under Assumption A, C and D, with probability at least $1-\delta$, the sequences $\pi_1, \dots, \pi_K \in \Pi$ and $\boldsymbol{\lambda}_1, \dots, \boldsymbol{\lambda}_K \in B\boldsymbol{\Delta}^I$ produced by the algorithm* PDCA *using the cost threshold $\boldsymbol{\tau}$ and the bound B satisfy*

$$\frac{1}{K}\sum_{k=1}^K L(\pi_k, \boldsymbol{\lambda}_k) \le \frac{1}{K}\sum_{k=1}^K L(\pi_k, \boldsymbol{\lambda}) + \mathcal{O}\left(\frac{B\epsilon_{stat}}{1-\gamma}\right)$$

for all $\boldsymbol{\lambda} \in B\boldsymbol{\Delta}^I$ *where* $L(\pi, \boldsymbol{\lambda}) = J_R(\pi) + \boldsymbol{\lambda} \cdot (\boldsymbol{\tau} - J_{\boldsymbol{C}}(\pi))$*.*

Proof. Recall that the OPE oracle produces an estimate h for the value of π with respect to a utility function U that satisfies $|J_U(\pi) - h| \le \mathcal{O}\left(\frac{C_{\ell_2}}{1-\gamma}\sqrt{\frac{\log(|\mathcal{F}|/\delta)}{n}}\right) \le \mathcal{O}(\epsilon_{\text{stat}})$ with probability at least $1-\delta$. By applying a union bound on $(\pi, U) \in \Pi \times \{C_1, \ldots, C_I\}$, we have with probability at least $1-\delta$ that

$$(1-\gamma)|J_{C_i}(\pi_k) - h_k^i| \le \mathcal{O}(\epsilon_{\text{stat}})$$

for all $k = 1, \ldots, K$ and all $i = 1, \ldots, I$. Hence,

$$\begin{aligned}\frac{1}{K}\sum_{k=1}^{K} L(\pi_k, \boldsymbol{\lambda}_k) - \frac{1}{K}\sum_{k=1}^{K} L(\pi_k, \boldsymbol{\lambda}) &= \frac{1}{K}\sum_{k=1}^{K}\sum_{i=1}^{I}(\lambda_k^i - \lambda^i)(\tau_i - J_{C_i}(\pi_k)) \\ &\le \frac{1}{K}\sum_{k=1}^{K}\sum_{i=1}^{I}(\lambda_k^i - \lambda^i)(\tau_i - h_k^i) + \mathcal{O}\left(\frac{B\epsilon_{\text{stat}}}{1-\gamma}\right).\end{aligned}$$

The first term in the last expression is $\frac{1}{K}\sum_{k=1}^{K} \boldsymbol{\lambda}_k \cdot (\boldsymbol{\tau} - \boldsymbol{h}_k) - \frac{1}{K}\sum_{k=1}^{K} \boldsymbol{\lambda} \cdot (\boldsymbol{\tau} - \boldsymbol{h}_k) \le 0$ since the λ-player chooses $\boldsymbol{\lambda}_k$ greedily that minimizes $\boldsymbol{\lambda} \mapsto \boldsymbol{\lambda} \cdot (\boldsymbol{\tau} - \boldsymbol{h}_k)$, and we are done. □

C.3 Proof of Lemma 7

Combining the results of Lemma 8 and Lemma 9, we can show that the pair $(\bar{\pi}, \bar{\boldsymbol{\lambda}})$ is approximately a saddle point where $\bar{\pi}$ is the policy returned by PDCA and $\bar{\boldsymbol{\lambda}}$ is the average of the sequence of Lagrange multipliers $\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_K$ produced by PDCA.

Proof of Lemma 7. Fix a policy $\pi \in \text{Conv}(\Pi)$ and a Lagrange multiplier $\boldsymbol{\lambda} \in B\boldsymbol{\Delta}^I$. Then,

$$\begin{aligned}L(\pi, \bar{\boldsymbol{\lambda}}) &\le \frac{1}{K}\sum_{k=1}^{K} L(\pi_k, \boldsymbol{\lambda}_k) + \frac{1+2B}{(1-\gamma)^2}\epsilon_{\text{opt}}(K) + \mathcal{O}\left(\frac{B\epsilon_{\text{stat}}}{1-\gamma}\right) \\ &\le \frac{1}{K}\sum_{k=1}^{K} L(\pi_k, \boldsymbol{\lambda}) + \frac{1+2B}{(1-\gamma)^2}\epsilon_{\text{opt}}(K) + \mathcal{O}\left(\frac{B\epsilon_{\text{stat}}}{1-\gamma}\right)\end{aligned}$$

where the first inequality uses Lemma 8 and the second inequality uses Lemma 9. Observing that $\frac{1}{K}\sum_{k=1}^{K} L(\pi_k, \boldsymbol{\lambda}) = L(\bar{\pi}, \boldsymbol{\lambda})$ by the linearity of $L(\cdot, \boldsymbol{\lambda})$ completes the proof.

D PROPERTIES OF A NEAR SADDLE POINT

In this section, we study the properties of a near saddle point formally defined below.

Definition 5. *We say* $(\bar{x}, \bar{y})$ *is a* ξ*-near saddle point for a function* $L(\cdot,\cdot)$ *with respect to the input space* $\mathcal{X} \times \mathcal{Y}$ *if* $L(x, \bar{y}) \le L(\bar{x}, y) + \xi$ *for all* $x \in \mathcal{X}$ *and* $y \in \mathcal{Y}$*.*

Lemma 10. *Suppose* $(\bar{\pi}, \bar{\boldsymbol{\lambda}})$ *is a* ξ*-near saddle point for* $L(\cdot,\cdot)$ *with respect to* $\widetilde{\Pi} \times B\boldsymbol{\Delta}^I$ *where* $\widetilde{\Pi} \subseteq$ *Conv*(Π) *is a class of mixtures of policies and at least one mixture policy in* $\widetilde{\Pi}$ *is feasible for* $\mathcal{P}(\boldsymbol{\tau})$*. Then, we have*

$$J_R(\bar{\pi}) \ge J_R(\pi_c) - \xi \qquad \text{(Optimality)}$$

$$J_{C_i}(\bar{\pi}) \le \tau_i + \frac{\xi}{B} + \frac{1}{B(1-\gamma)}, \quad \textit{for all } i = 1, \ldots I \qquad \text{(Feasibility)}$$

where π_c *is any feasible policy in* $\widetilde{\Pi}$*.*

Proof. We first prove $J_R(\bar{\pi}) \ge J_R(\pi_c) - \xi$, near optimality of $\bar{\pi}$.

Optimality Since $(\bar{\pi}, \bar{\boldsymbol{\lambda}})$ is a ξ-near saddle point for $L(\cdot,\cdot)$ with respect to $\widetilde{\Pi} \times B\boldsymbol{\Delta}^I$ and $\pi_c \in \widetilde{\Pi}$, we have $L(\pi_c, \bar{\boldsymbol{\lambda}}) \leq L(\bar{\pi}, \boldsymbol{\lambda}) + \xi$ for all $\boldsymbol{\lambda} \in B\boldsymbol{\Delta}^I$. Choosing $\boldsymbol{\lambda} = \mathbf{0}$, we get

$$L(\pi_c, \bar{\boldsymbol{\lambda}}) \leq L(\bar{\pi}, 0) + \xi = J_R(\bar{\pi}) + \xi.$$

Rearranging, we get

$$\begin{aligned} J_R(\bar{\pi}) &\geq J_R(\pi_c) + \bar{\boldsymbol{\lambda}} \cdot (\boldsymbol{\tau} - J_{\boldsymbol{C}}(\pi_c)) - \xi \\ &\geq J_R(\pi_c) - \xi \end{aligned}$$

where the second inequality uses the feasibility of π_c for $\mathcal{P}(\boldsymbol{\tau})$. This proves the near optimality of $\bar{\pi}$ with respect to π_c.

Feasibility Now, to prove near feasibility of $\bar{\pi}$, recall from the proof of the near optimality that $L(\pi_c, \bar{\boldsymbol{\lambda}}) \leq L(\bar{\pi}, \boldsymbol{\lambda}) + \xi$ for all $\boldsymbol{\lambda} \in B\boldsymbol{\Delta}^I$ holds since $(\bar{\pi}, \bar{\boldsymbol{\lambda}})$ is a ξ-near saddle point. Choosing $\boldsymbol{\lambda}$ such that $\lambda_j = B$ for $j = \text{argmin}_{i\in[I]}(\tau_i - J_{C_i}(\bar{\pi}))$ and $\lambda_j = 0$ for other j's, and defining $m = \min_{i\in[I]}(\tau_i - J_{C_i}(\bar{\pi}))$, we get

$$L(\pi_c, \bar{\boldsymbol{\lambda}}) \leq L(\bar{\pi}, \boldsymbol{\lambda}) + \xi = J_R(\bar{\pi}) + Bm + \xi.$$

On the other hand, the feasibility of π_c for $\mathcal{P}(\boldsymbol{\tau})$ gives

$$L(\pi_c, \bar{\boldsymbol{\lambda}}) = J_R(\pi_c) + \bar{\boldsymbol{\lambda}} \cdot (\boldsymbol{\tau} - J_{\boldsymbol{C}}(\pi_c)) \geq J_R(\pi_c).$$

Combining the previous two inequalities, we get

$$Bm + \xi \geq J_R(\pi_c) - J_R(\bar{\pi}) \geq -\frac{1}{1-\gamma}$$

where the last inequality uses $0 \leq J_R(\cdot) \leq \frac{1}{1-\gamma}$. Rearranging, and using the fact that $m \leq \tau_i - J_{C_i}(\bar{\pi})$ for all $i = 1, \ldots, I$, we get

$$\tau_i - J_{C_i}(\bar{\pi}) \geq m \geq -\frac{1}{B(1-\gamma)} - \frac{\xi}{B}$$

for all $i = 1, \ldots, I$. and it follows that

$$J_{C_i}(\bar{\pi}) \leq \tau_i + \frac{\xi}{B} + \frac{1}{B(1-\gamma)}.$$

□

Now, we study the case where PDCA is run with a tightened cost threshold $\boldsymbol{\tau} - \eta\mathbf{1}$ where $\eta \geq 0$. We denote by $L_\eta(\pi, \boldsymbol{\lambda}) = J_R(\pi) + \boldsymbol{\lambda} \cdot (\boldsymbol{\tau} - \eta\mathbf{1} - J_{\boldsymbol{C}}(\pi))$ the Lagrangian for the tightened problem $\mathcal{P}(\boldsymbol{\tau} - \eta\mathbf{1})$. The following lemma shows the property of a ξ-near saddle point for $L_\eta(\cdot,\cdot)$.

Lemma 11. *Assume that Slater's condition (Assumption B) holds and that $\eta < \frac{\phi}{1-\gamma}$ so that $\mathcal{P}(\boldsymbol{\tau} - \eta\mathbf{1})$ also satisfies Slater's condition. Suppose $(\bar{\pi}, \bar{\boldsymbol{\lambda}})$ is a ξ-near saddle point for $L_\eta(\cdot,\cdot)$ with respect to $\widetilde{\Pi} \times B\boldsymbol{\Delta}^I$. Let $(\pi^\star_\eta, \boldsymbol{\lambda}^\star_\eta)$ be a primal-dual solution to $\mathcal{P}(\boldsymbol{\tau} - \eta\mathbf{1})$ and $\pi^\star_\eta \in \widetilde{\Pi}$. Assume $B > \|\boldsymbol{\lambda}^\star_\eta\|_1$. Then, we have*

$$J_R(\bar{\pi}) \geq J_R(\pi^\star_\eta) - \xi \qquad \text{(Optimality)}$$

$$J_{C_i}(\bar{\pi}) \leq \tau_i - \eta + \frac{\xi}{B - \|\boldsymbol{\lambda}^\star_\eta\|_1}, \quad \textit{for all } i = 1, \ldots, I \qquad \text{(Feasibility)}$$

Proof. We first prove near optimality of $\bar{\pi}$.

Optimality Since $(\bar{\pi}, \bar{\boldsymbol{\lambda}})$ is a ξ-near saddle point for $L_\eta(\cdot,\cdot)$ with respect to $\widetilde{\Pi} \times B\boldsymbol{\Delta}^I$ and $\pi^\star_\eta \in \widetilde{\Pi}$, we have $L_\eta(\pi^\star_\eta, \bar{\boldsymbol{\lambda}}) \leq L_\eta(\bar{\pi}, \boldsymbol{\lambda}) + \xi$ for all $\boldsymbol{\lambda} \in B\boldsymbol{\Delta}^I$. Choosing $\boldsymbol{\lambda} = \mathbf{0}$, we get

$$L_\eta(\pi^\star_\eta, \bar{\boldsymbol{\lambda}}) \leq L_\eta(\bar{\pi}, 0) + \xi = J_R(\bar{\pi}) + \xi.$$

Rearranging, we get

$$J_R(\bar{\pi}) \geq J_R(\pi^\star_\eta) + \bar{\boldsymbol{\lambda}} \cdot (\boldsymbol{\tau} - \eta\mathbf{1} - J_{\boldsymbol{C}}(\pi^\star_\eta)) - \xi \geq J_R(\pi^\star_\eta) - \xi$$

where the second inequality uses the feasibility of $\pi^\star_\eta$ for $\mathcal{P}(\boldsymbol{\tau} - \eta\mathbf{1})$. Now, we prove feasibility of $\bar{\pi}$.

Feasibility Recall that $(\pi^\star_\eta, \boldsymbol{\lambda}^\star_\eta)$ is a primal-dual solution to the optimization problem $\mathcal{P}(\boldsymbol{\tau} - \eta\mathbf{1})$ and $L_\eta(\cdot,\cdot)$ is the Lagrangian function corresponding to the problem $\mathcal{P}(\boldsymbol{\tau} - \eta\mathbf{1})$. By strong duality, $(\pi^\star_\eta, \boldsymbol{\lambda}^\star_\eta)$ is a saddle point for $L_\eta(\cdot,\cdot)$ with respect to $\mathrm{Conv}(\Pi) \times \mathbb{R}^I_+$. Hence, we have

$$L_\eta(\bar{\pi}, \boldsymbol{\lambda}^\star_\eta) \le L_\eta(\pi^\star_\eta, \boldsymbol{\lambda}^\star_\eta) = J_R(\pi^\star_\eta) + \boldsymbol{\lambda}^\star_\eta \cdot (\boldsymbol{\tau} - \eta\mathbf{1} - J_{\boldsymbol{C}}(\pi^\star_\eta)) = J_R(\pi^\star_\eta)$$

where the first inequality follows from the fact that $(\pi^\star_\eta, \boldsymbol{\lambda}^\star_\eta)$ is a saddle point of $L_\eta(\cdot,\cdot)$ and the last equality follows from the complementary slackness property of the solution $(\pi^\star_\eta, \boldsymbol{\lambda}^\star_\eta)$. Rearranging, we get

$$J_R(\pi^\star_\eta) - J_R(\bar{\pi}) \ge \boldsymbol{\lambda}^\star_\eta \cdot (\boldsymbol{\tau} - \eta\mathbf{1} - J_{\boldsymbol{C}}(\bar{\pi})) \ge (m - \eta)\|\boldsymbol{\lambda}^\star_\eta\|_1 \tag{5}$$

where we define $m = \min_{i\in[I]}(\tau_i - J_{C_i}(\bar{\pi}))$. Now, to upper bound $J_R(\pi^\star_\eta) - J_R(\bar{\pi})$, we first use the feasibility of $\pi^\star_\eta$ for $\mathcal{P}(\boldsymbol{\tau} - \eta\mathbf{1})$ as follows.

$$L_\eta(\pi^\star_\eta, \bar{\boldsymbol{\lambda}}) = J_R(\pi^\star_\eta) + \bar{\boldsymbol{\lambda}} \cdot (\boldsymbol{\tau} - \eta\mathbf{1} - J_{\boldsymbol{C}}(\pi^\star_\eta)) \ge J_R(\pi^\star_\eta).$$

On the other hand, since $(\bar{\pi}, \bar{\boldsymbol{\lambda}})$ is a ξ-near saddle point for $L_\eta(\cdot,\cdot)$ with respect to $\widetilde{\Pi} \times B\boldsymbol{\Delta}^I$ and $\pi^\star_\eta \in \widetilde{\Pi}$, we have $L_\eta(\pi^\star_\eta, \bar{\boldsymbol{\lambda}}) \le L_\eta(\bar{\pi}, \boldsymbol{\lambda}) + \xi$ for any $\boldsymbol{\lambda} \in B\boldsymbol{\Delta}^I$. By choosing $\boldsymbol{\lambda}$ such that $\lambda_j = B$ for $j = \mathrm{argmin}_{i\in[I]}(\tau_i - J_{C_i}(\bar{\pi}))$ and recalling $m = \min_{i\in[I]}(\tau_i - J_{C_i}(\bar{\pi}))$, we get

$$L_\eta(\pi^\star_\eta, \bar{\boldsymbol{\lambda}}) \le L_\eta(\bar{\pi}, \boldsymbol{\lambda}) + \xi = J_R(\bar{\pi}) + B(m - \eta) + \xi.$$

Combining the previous two results (upper bound and lower bound of $L_\eta(\pi^\star_\eta, \bar{\boldsymbol{\lambda}})$), we get

$$J_R(\pi^\star_\eta) - J_R(\bar{\pi}) \le B(m - \eta) + \xi. \tag{6}$$

Combining the lower bound (5) and the upper bound (6) of $J_R(\pi^\star_\eta) - J_R(\bar{\pi})$ and rearranging, we get

$$m - \eta \ge \frac{-\xi}{B - \|\boldsymbol{\lambda}^\star_\eta\|_1}.$$

Since $\tau_i - \eta - J_{C_i}(\bar{\pi}) \ge m - \eta$ for all $i \in [I]$, rearranging the above gives

$$J_{C_i}(\bar{\pi}) \le \tau_i - \eta + \frac{\xi}{B - \|\boldsymbol{\lambda}^\star_\eta\|_1}$$

for all $i = 1, \ldots, I$. □

Note that the lemma above requires $B > \|\boldsymbol{\lambda}^\star_\eta\|_1$. We will show in Theorem 3 that with Slater's condition, we can upper bound $\|\boldsymbol{\lambda}^\star_\eta\|_1$ so that we can choose B that indeed satisfies $B > \|\boldsymbol{\lambda}^\star_\eta\|_1$.

E PROOF OF MAIN RESULTS

E.1 Proof of Theorem 1

We restate the theorem for convenience:

Theorem 1. *Under assumptions A,B,C,D and E, the policy $\bar{\pi}$ returned by the PDCA algorithm (Algorithm 1) with the cost threshold $\boldsymbol{\tau}$, bound $B = 1 + \frac{1}{\varphi}$ and K large enough, satisfies $J_{C_i}(\bar{\pi}) \le \tau_i + \mathcal{O}(\epsilon)$ for all $i = 1, \ldots, I$, and $J_R(\bar{\pi}) \ge J_R(\pi^\star) - \mathcal{O}(\epsilon)$ with probability at least $1 - \delta$ where $\pi^\star$ is optimal for $\mathcal{P}(\boldsymbol{\tau})$ as long as*

$$n \ge \Omega\left(\frac{(C_{\ell_2})^2 \log(I|\mathcal{F}||\Pi||\mathcal{W}|/\delta)}{(1-\gamma)^4\varphi^2\epsilon^2}\right).$$

Proof. Recall from Lemma 5 that $\epsilon_{\text{stat}} := \frac{C_{\ell_2}}{1-\gamma}\sqrt{\frac{\log(I|\mathcal{F}||\Pi||\mathcal{W}|/\delta)}{n}} + \frac{C_\infty \log(I|\mathcal{F}||\Pi||\mathcal{W}|/\delta)}{(1-\gamma)n}$. The bound on n guarantees

$$\epsilon_{\text{stat}} \le \mathcal{O}((1-\gamma)\varphi\,\epsilon).$$

Invoking Lemma 7 with cost threshold $\boldsymbol{\tau}$ and bound $B = 1 + \frac{1}{\varphi}$, we get with probability at least $1 - \delta$ that

$$L(\pi, \bar{\boldsymbol{\lambda}}) \leq L(\bar{\pi}, \boldsymbol{\lambda}) + \epsilon_{\text{saddle}}$$

for all $\pi \in \text{Conv}(\Pi)$ with $w^\pi \in \mathcal{W}$ and $\boldsymbol{\lambda} \in B\boldsymbol{\Delta}^I$ where $\epsilon_{\text{saddle}} \coloneqq \frac{1+2B}{(1-\gamma)^2}\epsilon_{\text{opt}}(K) + \mathcal{O}\left(\frac{B\epsilon_{\text{stat}}}{1-\gamma}\right)$. Since PDCA chooses K such that $\frac{1+2B}{(1-\gamma)^2}\epsilon_{\text{opt}}(K) \leq \epsilon$ and the bound on n guarantees $\epsilon_{\text{stat}} \leq \mathcal{O}((1-\gamma)\varphi\epsilon)$, we have $\epsilon_{\text{saddle}} \leq \mathcal{O}(\epsilon)$. Hence, invoking Lemma 11 with $\xi = \epsilon_{\text{saddle}}$, $B = 1 + \frac{1}{\varphi}$ and $\eta = 0$, we get

$$J_R(\bar{\pi}) \geq J_R(\pi_c) - \epsilon_{\text{saddle}} \geq J_R(\pi_c) - \mathcal{O}(\epsilon)$$
$$J_{C_i}(\bar{\pi}) \leq \tau_i + \frac{\epsilon_{\text{saddle}}}{1 + \frac{1}{\varphi} - \|\boldsymbol{\lambda}^\star\|_1} \leq \tau_i + \mathcal{O}(\epsilon), \quad i = 1, \dots, I$$

where $\boldsymbol{\lambda}^\star$ is the optimal dual variable for the problem $\mathcal{P}(\boldsymbol{\tau})$ and the last inequality uses $\|\boldsymbol{\lambda}^\star\|_1 \leq \frac{1}{\varphi}$, which follows by the Slater's condition (Assumption B) and Lemma 13.

□

E.2 Result for arbitrary competing policy

Theorem 2. *Under assumptions A,C and D, the policy $\bar{\pi}$ returned by PDCA (Algorithm 1) with the cost threshold $\boldsymbol{\tau}$ and bound $B = \frac{1}{(1-\gamma)\epsilon}$ satisfies $J_{C_i}(\bar{\pi}) \leq \tau_i + \mathcal{O}(\epsilon)$ for all $i = 1, \dots, I$, and $J_R(\pi) \geq J_R(\pi_c) - \mathcal{O}(\epsilon)$ with probability at least $1 - \delta$ where $\pi_c \in Conv(\Pi)$ is any policy of which MIW is realizable by $\mathcal{W}$ as long as*

$$n \geq \Omega\left(\frac{(C_{\ell_2})^2 \log(I|\mathcal{F}||\Pi||\mathcal{W}|/\delta)}{(1-\gamma)^6\epsilon^4}\right).$$

Proof. Recall from Lemma 5 that $\epsilon_{\text{stat}} \coloneqq \frac{C_{\ell_2}}{1-\gamma}\sqrt{\frac{\log(I|\mathcal{F}||\Pi||\mathcal{W}|/\delta)}{n}} + \frac{C_\infty \log(I|\mathcal{F}||\Pi||\mathcal{W}|/\delta)}{(1-\gamma)n}$. The choice $n \geq \Omega\left(\frac{(C_{\ell_2})^2 \log(I|\mathcal{F}||\Pi||\mathcal{W}|/\delta)}{(1-\gamma)^6\epsilon^4}\right)$ guarantees

$$\epsilon_{\text{stat}} \leq \mathcal{O}((1-\gamma)^2\epsilon^2).$$

Invoking Lemma 7 with cost threshold $\boldsymbol{\tau}$ and bound $B = \frac{1}{(1-\gamma)\epsilon}$, we get with probability at least $1 - \delta$ that

$$L(\pi, \bar{\boldsymbol{\lambda}}) \leq L(\bar{\pi}, \boldsymbol{\lambda}) + \epsilon_{\text{saddle}}$$

for all $\pi \in \text{Conv}(\Pi)$ and $\boldsymbol{\lambda} \in \frac{1}{(1-\gamma)\epsilon}\boldsymbol{\Delta}^I$ where $\epsilon_{\text{saddle}} \coloneqq \frac{1+2B}{(1-\gamma)^2}\epsilon^\pi_{\text{opt}}(K) + \frac{2B}{1-\gamma}\epsilon^\lambda_{\text{opt}}(K) + \mathcal{O}\left(\frac{B\epsilon_{\text{stat}}}{1-\gamma}\right)$. Since PDCA chooses K such that $\frac{1+2B}{(1-\gamma)^2}\epsilon^\pi_{\text{opt}}(K) \leq \epsilon$ and $\frac{2B}{1-\gamma}\epsilon^\lambda_{\text{opt}}(K) \leq \epsilon$, and n is chosen to guarantee $\epsilon_{\text{stat}} \leq \mathcal{O}((1-\gamma)^2\epsilon^2)$, we have $\epsilon_{\text{saddle}} \leq \mathcal{O}(\epsilon)$. Hence, invoking Lemma 10 with $\xi = \epsilon_{\text{saddle}}$ and $B = \frac{1}{(1-\gamma)\epsilon}$, we get

$$J_R(\bar{\pi}) \geq J_R(\pi_c) - \epsilon_{\text{saddle}} \geq J_R(\pi_c) - \mathcal{O}(\epsilon)$$
$$J_{C_i}(\bar{\pi}) \leq \tau_i + \frac{\epsilon_{\text{saddle}}}{B} + \frac{1}{B(1-\gamma)} \leq \tau_i + \mathcal{O}(\epsilon), \quad i = 1, \dots, I.$$

□

E.3 Learning policy satisfying constraints exactly

Note that results in previous sections provide a bound on sample complexity for finding a nearly optimal policy that *approximately* satisfies the constraints. In this section, we provide a bound for finding a nearly optimal policy that satisfies the constraints *exactly* by running PDCA with tightened constraints. We need the following additional technical assumption on MIW realizability.

Theorem 3. *Let $\epsilon \in (0, \frac{1}{2}]$ be given. Under assumptions A,B,C,D,E and F, the policy $\bar{\pi}$ returned by the PDCA algorithm (Algorithm 1) with the cost threshold $\boldsymbol{\tau} - \eta\mathbf{1}$, where $\eta = \varphi\epsilon$, and bound $B = \frac{5}{\varphi}$ satisfies $J_{C_i}(\bar{\pi}) \leq \tau_i$ for all $i = 1, \dots, I$, and $J_R(\pi) \geq J_R(\pi^\star) - \mathcal{O}(\epsilon)$ with probability at least $1 - \delta$, where $\pi^\star$ is an optimal policy for $\mathcal{P}(\boldsymbol{\tau})$ as long as*

$$n \geq \Omega\left(\frac{(C_{\ell_2})^2 \log(I|\mathcal{F}||\Pi||\mathcal{W}|/\delta)}{(1-\gamma)^4\varphi^2\epsilon^2}\right).$$

Proof. Recall from Lemma 5 that $\epsilon_{\text{stat}} := \frac{C_{\ell_2}}{1-\gamma}\sqrt{\frac{\log(I|\mathcal{F}||\Pi||\mathcal{W}|/\delta)}{n}} + \frac{C_\infty \log(I|\mathcal{F}||\Pi||\mathcal{W}|/\delta)}{(1-\gamma)n}$. The bound on n in the theorem guarantees $\epsilon_{\text{stat}} \leq \mathcal{O}((1-\gamma)\varphi\,\epsilon)$. Invoking Lemma 7 with cost threshold $\boldsymbol{\tau} - \eta\mathbf{1}$ and bound $B = \frac{5}{\varphi}$, we get with probability at least $1-\delta$ that

$$L_\eta(\pi, \bar{\boldsymbol{\lambda}}) \leq L_\eta(\bar{\pi}, \boldsymbol{\lambda}) + \epsilon_{\text{saddle}} \tag{7}$$

for all $\pi \in \text{Conv}(\Pi)$ satisfying $w^\pi \in \mathcal{W}$ and $\boldsymbol{\lambda} \in \frac{5}{\varphi}\boldsymbol{\Delta}^I$ where $L_\eta(\cdot,\cdot)$ is the Lagrangian for $\mathcal{P}(\boldsymbol{\tau} - \eta\mathbf{1})$ and $\epsilon_{\text{saddle}} := \frac{1+2B}{(1-\gamma)^2}\epsilon_{\text{opt}}(K) + \mathcal{O}\left(\frac{\epsilon_{\text{stat}}}{(1-\gamma)\varphi}\right)$. Since PDCA chooses K such that $\frac{1+2B}{(1-\gamma)^2}\epsilon_{\text{opt}}(K) \leq \epsilon$ and n is chosen to guarantee $\epsilon_{\text{stat}} \leq \mathcal{O}((1-\gamma)\varphi\,\epsilon)$, we have $\epsilon_{\text{saddle}} \leq 2\epsilon$ (with appropriate scaling of n by a universal constant).

Let $\pi^\star$ be an optimal policy for $\mathcal{P}(\boldsymbol{\tau})$ and $\pi^\star_\alpha$ for $\mathcal{P}(\boldsymbol{\tau} - \alpha\mathbf{1})$. By the MIW realizability assumptions E and F, we have $w^{\pi^\star}, w^{\pi^\star_\alpha} \in \mathcal{W}$ and it follows from (7) that

$$L_\eta(\pi^\star, \bar{\boldsymbol{\lambda}}) \leq L_\eta(\bar{\pi}, \boldsymbol{\lambda}) + 2\epsilon \tag{8}$$

$$L_\eta(\pi^\star_\alpha, \bar{\boldsymbol{\lambda}}) \leq L_\eta(\bar{\pi}, \boldsymbol{\lambda}) + 2\epsilon \tag{9}$$

for all $\boldsymbol{\lambda} \in \frac{5}{\varphi}\boldsymbol{\Delta}^I$. Now, we show near optimality and exact feasibility from these inequalities.

Near Optimality Setting $\boldsymbol{\lambda} = \mathbf{0}$ in (8) and rearranging, we get

$$\begin{aligned}
J_R(\bar{\pi}) &\geq J_R(\pi^\star) + \bar{\boldsymbol{\lambda}} \cdot (\boldsymbol{\tau} - \eta\mathbf{1} - J_{\boldsymbol{C}}(\pi^\star)) - 2\epsilon \\
&\geq J_R(\pi^\star) - \eta\|\bar{\boldsymbol{\lambda}}\|_1 - \mathcal{O}(\epsilon) \\
&\geq J_R(\pi^\star) - \mathcal{O}(\epsilon)
\end{aligned}$$

where the second inequality follows by the feasibility of $\pi^\star$ for $\mathcal{P}(\boldsymbol{\tau})$ and $\epsilon_{\text{saddle}} \leq \mathcal{O}(\epsilon)$; the last inequality follows by $\eta\|\bar{\boldsymbol{\lambda}}\|_1 \leq \eta B = \mathcal{O}(\epsilon)$. This proves near optimality of $\bar{\pi}$ Now we prove that $\bar{\pi}$ is (exactly) feasible for $\mathcal{P}(\boldsymbol{\tau})$.

Exact Feasibility Define $m = \min_{i\in[I]}(\tau_i - J_{C_i}(\bar{\pi}))$. If $m \geq 0$ then $\tau_i - J_{C_i}(\bar{\pi}) \geq 0$ for all $i = 1, \ldots, I$ and exact feasibility trivially holds. We only consider the case where $m < 0$. Define a mixture policy $\widetilde{\pi} = (1-\zeta)\pi^\star + \zeta\pi^\star_\alpha$ where $\zeta \in (0,1)$ is to be determined later. Since $L_\eta(\cdot, \bar{\boldsymbol{\lambda}})$ is linear, a linear combination of (8) and (9) with coefficients $1-\zeta$ and ζ respectively, we get

$$L_\eta(\widetilde{\pi}, \bar{\boldsymbol{\lambda}}) \leq L_\eta(\bar{\pi}, \boldsymbol{\lambda}) + 2\epsilon.$$

Choosing $\boldsymbol{\lambda}$ such that $\lambda_j = B$ for $j = \text{argmin}_{i\in[I]}(\tau_i - J_{C_i}(\bar{\pi}))$ and $\lambda_j = 0$ for all other indices, we get

$$\begin{aligned}
L_\eta(\widetilde{\pi}, \bar{\boldsymbol{\lambda}}) &\leq J_R(\bar{\pi}) + \boldsymbol{\lambda} \cdot (\boldsymbol{\tau} - \eta\mathbf{1} - J_{\boldsymbol{C}}(\bar{\pi})) + 2\epsilon \\
&= J_R(\bar{\pi}) + B(m - \eta) + 2\epsilon.
\end{aligned}$$

On the other hand, using the fact that $\widetilde{\pi}$ is feasible for $\mathcal{P}(\boldsymbol{\tau} - \zeta\alpha\mathbf{1})$, we get

$$\begin{aligned}
L_\eta(\widetilde{\pi}, \bar{\boldsymbol{\lambda}}) &= J_R(\widetilde{\pi}) + \bar{\boldsymbol{\lambda}}(\boldsymbol{\tau} - \eta\mathbf{1} - J_{\boldsymbol{C}}(\widetilde{\pi})) \\
&\geq J_R(\widetilde{\pi}) + (\zeta\alpha - \eta)\|\bar{\boldsymbol{\lambda}}\|_1.
\end{aligned}$$

Combining the previous two results (upper bound and lower bound of $L_\eta(\widetilde{\pi}, \bar{\boldsymbol{\lambda}})$) and rearranging, we get

$$J_R(\widetilde{\pi}) - J_R(\bar{\pi}) \leq B(m - \eta) - (\zeta\alpha - \eta)\|\bar{\boldsymbol{\lambda}}\|_1 + 2\epsilon. \tag{10}$$

Now, to get a lower bound of $J_R(\widetilde{\pi}) - J_R(\bar{\pi})$, let $(\widetilde{\pi}^\star, \widetilde{\boldsymbol{\lambda}}^\star)$ be a primal-dual solution of $\mathcal{P}(\boldsymbol{\tau} - \zeta\alpha\mathbf{1})$. Note that $\mathcal{P}(\boldsymbol{\tau} - \zeta\alpha\mathbf{1})$ is feasible by the Slater's condition assumption B and the fact that $\zeta\alpha \in (0, \frac{\varphi}{1-\gamma})$. Since $(\widetilde{\pi}^\star, \widetilde{\boldsymbol{\lambda}}^\star)$ is a saddle point of $L_{\zeta\alpha}(\pi, \boldsymbol{\lambda}) = J_R(\pi) + \boldsymbol{\lambda} \cdot (\boldsymbol{\tau} - \zeta\alpha\mathbf{1} - J_{\boldsymbol{C}}(\pi))$ with respect to $\text{Conv}(\Pi) \times \mathbb{R}^I_+$, we get

$$L_{\zeta\alpha}(\bar{\pi}, \widetilde{\boldsymbol{\lambda}}^\star) \leq L_{\zeta\alpha}(\widetilde{\pi}^\star, \widetilde{\boldsymbol{\lambda}}^\star) = J_R(\widetilde{\pi}^\star) \leq J_R(\pi^\star) \leq \frac{1}{1-\zeta} J_R(\widetilde{\pi})$$

where the equality follows by the complementary slackness property; the second inequality follows since the feasibility set of $\mathcal{P}(\boldsymbol{\tau})$ contains that of $\mathcal{P}(\boldsymbol{\tau}-\zeta\alpha\mathbf{1})$; and the last inequality follows by $J_R(\widetilde{\pi}) = (1-\zeta)J_R(\pi^\star) + \zeta J_R(\pi^\star_\alpha) \geq (1-\zeta)J_R(\pi^\star)$. Rearranging, we get

$$\begin{aligned} J_R(\widetilde{\pi}) - J_R(\bar{\pi}) &\geq -\zeta J_R(\bar{\pi}) + (1-\zeta)\widetilde{\boldsymbol{\lambda}}^\star \cdot (\boldsymbol{\tau} - \zeta\alpha\mathbf{1} - J_{\boldsymbol{C}}(\bar{\pi})) \\ &\geq \frac{-\zeta}{1-\gamma} + (1-\zeta)(m-\zeta\alpha)\|\widetilde{\boldsymbol{\lambda}}^\star\|_1 \end{aligned}$$

where the second inequality follows by $J_R(\cdot) \leq \frac{1}{1-\gamma}$ and the definition of m. Combining with the upper bound of $J_R(\widetilde{\pi}) - J_R(\bar{\pi})$ shown in (10) and rearranging, we get

$$(B - (1-\zeta)\|\widetilde{\boldsymbol{\lambda}}^\star\|_1)m \geq B\eta + (\zeta\alpha - \eta)\|\bar{\boldsymbol{\lambda}}\|_1 - \frac{\zeta}{1-\gamma} - (1-\zeta)\zeta\alpha\|\widetilde{\boldsymbol{\lambda}}^\star\|_1 - 2\epsilon. \tag{11}$$

Now, we choose our parameters as follows.

$$\zeta = (1-\gamma)\epsilon, \quad B = \frac{5c}{\varphi}, \quad \eta = \frac{\varphi\epsilon}{c}.$$

Note that $B\eta = 5\epsilon$ and $\zeta\alpha \geq \eta$. Also, since $\widetilde{\boldsymbol{\lambda}}^\star$ is a dual solution of $\mathcal{P}(\boldsymbol{\tau}-\zeta\alpha\mathbf{1})$, which has a margin of $\frac{\varphi}{1-\gamma} - \zeta\alpha$, Lemma 13 gives $\|\widetilde{\boldsymbol{\lambda}}^\star\|_1 \leq \frac{1}{\varphi-(1-\gamma)\zeta\alpha}$. Hence,

$$\zeta\alpha\|\widetilde{\boldsymbol{\lambda}}^\star\|_1 \leq \frac{\zeta\alpha}{\varphi-(1-\gamma)\zeta\alpha} \leq \frac{1}{1-\gamma}\frac{\zeta\varphi}{\varphi-\varphi\zeta} \leq \frac{2\zeta}{1-\gamma} \leq 2\epsilon$$

where the second inequality uses $\alpha \leq \frac{\varphi}{1-\gamma}$ and the fact that $h(x) = \frac{x}{\varphi - x}$ is increasing for $x \in (0,\varphi)$; and the third inequality uses the fact that $\zeta \leq \frac{1}{2}$. Note that $\|\widetilde{\boldsymbol{\lambda}}^\star\|_1 \leq \frac{\epsilon}{\zeta\alpha} = \frac{1}{(1-\gamma)\alpha} < B$ so that $B - (1-\zeta)\|\widetilde{\boldsymbol{\lambda}}^\star\|_1 > 0$. Hence, the previous result (11) gives

$$\begin{aligned} (B - (1-\zeta)\|\widetilde{\boldsymbol{\lambda}}^\star\|_1)m &\geq B\eta + (\zeta\alpha - \eta)\|\bar{\boldsymbol{\lambda}}\|_1 - \frac{\zeta}{1-\gamma} - (1-\zeta)\zeta\alpha\|\widetilde{\boldsymbol{\lambda}}^\star\|_1 - 2\epsilon \\ &\geq 5\epsilon + 0 - \epsilon - 2\epsilon - 2\epsilon \\ &= 0. \end{aligned}$$

Since $B - (1-\zeta)\|\widetilde{\boldsymbol{\lambda}}^\star\|_1 > 0$, we have $m \geq 0$ which implies $\tau_i - J_{C_i}(\bar{\pi}) \geq 0$ for all $i = 1, \ldots, I$. This completes the proof.

□

F CONVEX OPTIMIZATION

Lemma 12. *Let $(\pi^\star, \lambda^\star)$ be optimal primal dual solutions to the constrained optimization problem $\mathcal{P}(\tau)$. Let $(\widetilde{\pi}^\star, \widetilde{\lambda}^\star)$ be optimal primal dual solutions to the perturbed problem $\mathcal{P}(\widetilde{\boldsymbol{\tau}})$ where $\widetilde{\boldsymbol{\tau}} = \boldsymbol{\tau} - \eta\mathbf{1}$. Then, we have*

$$J_R(\widetilde{\pi}^\star) \geq J_R(\pi^\star) - \eta\|\widetilde{\boldsymbol{\lambda}}^\star\|_1.$$

Proof. The proof follows Chapter 5.6 in Boyd et al. (2004). By strong duality of the optimization problem $\mathcal{P}(\widetilde{\boldsymbol{\tau}})$, we have $J_R(\widetilde{\pi}^\star) = d(\widetilde{\boldsymbol{\lambda}}^\star)$ where $d(\boldsymbol{\lambda}) = \max_{\pi \in \mathrm{Conv}(\Pi)} L(\pi, \boldsymbol{\lambda}; \widetilde{\boldsymbol{\tau}})$ is the dual function of $\mathcal{P}(\widetilde{\boldsymbol{\tau}})$. Hence,

$$\begin{aligned} J_R(\widetilde{\pi}^\star) &= d(\widetilde{\boldsymbol{\lambda}}^\star) \\ &\geq J_R(\pi^\star) + \widetilde{\boldsymbol{\lambda}}^\star \cdot (\widetilde{\boldsymbol{\tau}} - J_{\boldsymbol{C}}(\pi^\star)) \\ &= J_R(\pi^\star) + \widetilde{\boldsymbol{\lambda}}^\star \cdot (\boldsymbol{\tau} - J_{\boldsymbol{C}}(\pi^\star)) - \widetilde{\boldsymbol{\lambda}}^\star \cdot \eta\mathbf{1} \\ &\geq J_R(\pi^\star) - \widetilde{\boldsymbol{\lambda}}^\star \cdot \eta\mathbf{1} \end{aligned}$$

where the first inequality follows from the definition of the dual function $d(\cdot)$ and the second follows by the feasibility of $\pi^\star$ for $\mathcal{P}(\boldsymbol{\tau})$. □

Lemma 13. *Consider a constrained optimization problem $\mathcal{P}(\boldsymbol{\tau})$ with threshold $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_I)$ with $\tau_i > 0$ for all $i = 1, \ldots, I$. Suppose the problem satisfies Slater's condition with margin $\frac{\varphi}{1-\gamma} > 0$, in other words, there exists $\pi \in \Pi$ that satisfies the constraint $J_{C_i}(\pi) \leq \tau_i - \frac{\varphi}{1-\gamma}$ for all $i = 1, \ldots, I$. Then, the optimal dual variable $\boldsymbol{\lambda}^\star$ of the problem satisfies $\|\boldsymbol{\lambda}^\star\|_1 \leq \frac{1}{\varphi}$.*

Proof. Let $\pi^\star$ be an optimal policy of the optimization problem $\mathcal{P}(\boldsymbol{\tau})$. Define the dual function $f(\boldsymbol{\lambda}) = \max_{\pi \in \Pi} J_R(\pi) + \boldsymbol{\lambda} \cdot (\boldsymbol{\tau} - J_{\boldsymbol{C}}(\pi))$. Let $\boldsymbol{\lambda}^\star = \operatorname{argmin}_{\boldsymbol{\lambda} \in \mathbb{R}_+^I} f(\boldsymbol{\lambda})$. Trivially, $\lambda_i^\star \geq 0$ for all $i = 1, \ldots, I$. Also, by strong duality, we have $f(\boldsymbol{\lambda}^\star) = J_R(\pi^\star)$. Let $\widehat{\pi}$ be a feasible policy with $J_{\boldsymbol{C}}(\widehat{\pi}) \leq \boldsymbol{\tau} - \frac{\varphi}{1-\gamma}\mathbf{1}$ where the inequality is component-wise and $\mathbf{1} = (1, \ldots, 1)$. Such a policy exists by the assumption of this lemma. Then,

$$J_R(\pi^\star) = f(\boldsymbol{\lambda}^\star) \geq J_R(\widehat{\pi}) + \boldsymbol{\lambda}^\star \cdot (\boldsymbol{\tau} - J_{\boldsymbol{C}}(\widehat{\pi})) \geq J_R(\widehat{\pi}) + \boldsymbol{\lambda}^\star \cdot \frac{\varphi}{1-\gamma}\mathbf{1} = J_R(\widehat{\pi}) + \frac{\varphi}{1-\gamma}\|\boldsymbol{\lambda}^\star\|_1.$$

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

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

□

G EXPERIMENTS

In this section, we empirically demonstrate the performance of our algorithm PDCA by running it in various environments and comparing the performance to COptiDICE. For the parameter tuning and the experiments, we used an internal cluster of nodes with 20-core 2.40 GHz CPU and Nvidia Tesla V100 GPU. The total amount of computing time was around 600 hours.

G.1 Tabular CMDP Experiments

In this section, we provide details of the experiment run on a randomly generated CMDP discussed in Section 5. We follow a similar experimental protocol as Lee et al. (2022).

CMDP Generation We set the number of states to 10 and the number of actions to 5. The transition probability is randomly generated by drawing from a Dirichlet distribution with all parameters $(1, \ldots, 1)$ for generating each $P(\cdot|s, a)$. We set the number of cost functions I to 1. The reward function R is randomly drawn from $[0, 1]$ uniformly for each $R(s, a)$. The cost function C_1 is randomly drawn from a beta distribution with parameters 0.2, 0.2 for each $C_1(s, a)$. We choose the discount factor $\gamma = 0.8$ and the cost threshold $\tau = 0.5$. We repeat the random generation policy until the cost threshold is not slack for the optimal policy.

π-Player We use the natural policy gradient algorithm with exponential weight updating scheme for the π player (Algorithm 3).

Algorithm 3: Natural policy gradient

Input: Learning rate η, sequence of functions $h_1, \ldots, h_K \in (\mathcal{S} \times \mathcal{A} \to [-1, 1])$

Init: π_1 a uniform policy.

```
for k = 1, ..., K − 1 do
    for s ∈ S do
        π_{k+1}(a|s) ∝ π_k(a|s) exp(η h_k(s, a)) normalized to sum to 1 across a ∈ A.
```

Return: Sequence $\pi_1, \ldots, \pi_K$

Offline Dataset We set behavior policy μ to be a mixture of π_{uniform} and $\pi^\star$ where π_{uniform} is the uniform policy that takes actions uniformly at random from the action space $\mathcal{A}$ at every state, and $\pi^\star$ is the optimal solution to the generated CMDP. We exactly solve for the occupancy measure d^μ of the behavior policy μ. We repeatedly sample the (s, a) pair from d^μ and then sample s' according to $P(\cdot|s, a)$.

Hyperparameters The learning rate for the π-player is chosen using grid search in $\{1, 2, 5, 10\}$. The bound B for the λ-player is chosen by grid search in $\{2, 5, 10\}$. The bound C_∞ for $\mathcal{W}$ is chosen by a grid search in $\{2, 5, 10\}$.

G.2 RWRL Benchmark Experiments

Hyperparameters We do a grid search on $\{0.00005, 0.0001, 0.0003, 0.0005, 0.001\}$ for determining the learning rate η_{fast} for the critics and a grid search on $\{0.00005, 0.0001, 0.0002\}$ for determining the learning rate η_{slow} for the π-player. The chosen learning rates are $\eta_{\text{fast}} = 0.0003$ and $\eta_{\text{slow}} = 0.0001$. We use the batch size 1024. We run $K = 30000$ iterations for Cartpole, Walker, Quadruped environments and $K = 50000$ iterations for the Humanoid environment. For the policy network and the networks for the critics, we use fully-connected neural networks with two hidden layers of width 256.

G.3 Bullet Safety Gym Benchmark Experiments

In this section, we provide details of the experiments run on Bullet Safety Gym benchmark environments.

Offline Datasets We use the offline datasets provided by Liu et al. (2023a). They collect dataset for each environment by merging trajectories generated by algorithms trained with various cost thresholds and hyperparameters. After merging, they run a post-processing of filtering redundant trajectories to ensure a diverse set of trajectories. For details, refer to their paper.

Hyperparameters Following the setup used by Liu et al. (2023a), we set the learning rate η_{fast} for the critics to 0.001 and the learning rate η_{slow} for the π-player to 0.0001. We use the batch size 512. We run $K = 100,000$ iterations. For the policy network and the networks for the critics, we use fully-connected neural networks with two hidden layers of width 256. We do a grid search on $\{2, 5, 10\}$ for the bound B for the λ-player. We do a grid search on $\{2, 5, 10\}$ for the bound C_∞ for $\mathcal{W}$.