A Primal-Dual Algorithm for Offline Constrained Reinforcement Learning with Linear MDPs

Kihyuk Hong [1] **Ambuj Tewari** [1]

Abstract

We study offline reinforcement learning (RL) with linear MDPs under the infinite-horizon discounted setting which aims to learn a policy that maximizes the expected discounted cumulative reward using a pre-collected dataset. Existing algorithms for this setting either require a uniform data coverage assumptions or are computationally inefficient for finding an ϵ-optimal policy with $\mathcal{O}(\epsilon^{-2})$ sample complexity. In this paper, we propose a primal dual algorithm for offline RL with linear MDPs in the infinite-horizon discounted setting. Our algorithm is the first computationally efficient algorithm in this setting that achieves sample complexity of $\mathcal{O}(\epsilon^{-2})$ with partial data coverage assumption. Our work is an improvement upon a recent work that requires $\mathcal{O}(\epsilon^{-4})$ samples. Moreover, we extend our algorithm to work in the offline constrained RL setting that enforces constraints on additional reward signals.

1. Introduction

We study the offline constrained reinforcement learning (RL) setting where a dataset of trajectories collected previously is given and the goal is to learn a decision making policy that performs well with respect to a reward signal while satisfying constraints on additional reward signals. The setting is applicable to real-world problems that have safety concerns. Learning from a previously collected dataset without interacting with the environment, a key property of offline RL (Levine et al., 2020), is useful in real-world problems where interacting with the environment is expensive or dangerous (Kumar et al., 2021; Tang & Wiens, 2021; Levine et al., 2018). Enforcing constraints on additional reward signals, a key property of constrained RL (Altman, 2021), is useful for applications with safety concerns (Wang et al., 2019; Brunke et al., 2022).

A challenge in offline RL is *distribution shift* (Levine et al., 2020), a mismatch of the state-action distribution in the offline dataset to the state-action distributions induced by candidate policies. For sample-efficient learning, offline RL requires the data distribution of the target policy to be covered by the offline dataset (Chen & Jiang, 2019). A uniform data coverage assumption (Antos et al., 2007) is a convenient, but a strong assumption that requires the offline dataset to cover state-action distributions induced by all policies. Recent works study offline RL with partial data coverage assumption that only requires the offline dataset to cover state-action distribution induced by a single target policy (Jin et al., 2021).

Another challenge in offline RL, which is also a challenge in online RL, is that many practical problems have large state spaces, making sample efficient learning difficult. For sample efficient learning in large state space, we need to assume a structure in the problem. In this paper, we study the linear MDP setting (Jin et al., 2020) that assumes the transition probability matrix and the reward function have linear structures. This setting ensures the value function is linear in a low-dimensional representation of state-action pairs, allowing sample-efficient learning. To the best of our knowledge, none of the previous works on offline RL for linear MDPs provides with partial data coverage provide a computationally efficient algorithm with $\mathcal{O}(\epsilon^{-2})$ sample complexity. In this paper, we introduce a novel algorithm that achieves this. Furthermore, we extend to the offline *constrained* RL setting that allows specifying constraints on additional reward signals.

1.1. Related Work

In Table 1, we compare our work to previous works. The column N shows how the sample complexity bound scales with the error tolerance ϵ. The first five algorithms are for offline RL with general function approximation. The algorithms can be reduced to the linear function approximation setting by taking a value function class consisting of linear functions. The sixth algorithm is for offline RL with linear function approximation. The last two algorithms are

[1]Department of Statistics, University of Michigan. Correspondence to: Kihyuk Hong <kihyukh@umich.edu>, Ambuj Tewari <tewaria@umich.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

Table 1. Comparison of algorithms for offline (constrained) RL

Setting	Algorithm	Partial coverage	Computationally efficient	Support constraints	N
General	FQI (Munos & Szepesvári, 2008)	No	Yes	No	ϵ^{-2}
General	CBPL (Le et al., 2019)	No	Yes	Yes	ϵ^{-2}
General	Minimax (Xie et al., 2021)	Yes	No	No	ϵ^{-2}
General	CPPO (Uehara & Sun, 2022)	Yes	No	No	ϵ^{-2}
General	Minimax (Zanette, 2023)	No	No	No	ϵ^{-2}
Linear	PSPI (Xie et al., 2021)	Yes	Yes	No	ϵ^{-5}
Linear MDPs	Primal-Dual (Gabbianelli et al., 2024)	Yes	Yes	No	ϵ^{-4}
Linear MDPs	Primal-Dual (**Ours**)	Yes	Yes	Yes	ϵ^{-2}

results on offline RL with linear MDPs, which is a special case of the linear function approximation setting. The computational efficiency of algorithms for the general function approximation setting is judged based on the efficiency when applied to linear function class. As the table shows, our algorithm is the first computationally efficient algorithm with sample complexity $\mathcal{O}(\epsilon^{-2})$ for finding ϵ-optimal policy under partial data coverage assumption. Moreover, our algorithm supports constraints on additional reward signals.

Offline RL with General Function Approximation Offline RL with general function approximation is widely studied in the discounted infinite-horizon setting. When casting the linear function approximation setting to the general function approximation setting, we get the realizability and Bellman completeness for free when using linear function class since the value function under linear function approximation is linear. In Table 1, we only compared works on general function approximation that assumes realizability, Bellman completeness and data coverage. There are other works that relax Bellman completeness assumption at the cost of introducing another assumption. For example, Xie & Jiang (2020); Zhan et al. (2022); Zhu et al. (2023); Hong et al. (2023) relax Bellman completeness assumption and introduce marginalized importance weight assumption.

Offline RL with Episodic Setting Offline RL with linear function approximation has been studied in the finite-horizon episodic setting. Zanette et al. (2021) propose a computationally efficient actor-critic algorithm with pessimism to achieve $\mathcal{O}(\epsilon^{-2})$ sample complexity under partial data coverage. Jin et al. (2021) propose a computationally efficient value iteration based algorithm with pessimism to achieve $\mathcal{O}(\epsilon^{-2})$ sample complexity under partial data coverage. However, they require the knowledge of the covariance matrix induced by the state-action data distribution. Although their results are computationally efficient and work under partial data coverage, they do not apply to the infinite-horizon discounted setting. Wu et al. (2021) study offline constrained RL with a more general way of specifying constraints. Their focus is on episodic setting with linear mixture MDP.

2. Preliminaries

Notations We denote by $\Delta(\mathcal{X})$ the probability simplex over a finite set $\mathcal{X}$. We write $\mathbf{\Delta}^I = \{\boldsymbol{x} \in \mathbb{R}_+^I : \sum_{i=1}^I x_i \leq 1\}$. We write $\mathbb{B}_d(B) = \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\|_2 \leq B\}$. Given a matrix A, denote by $A^\dagger$ its pseudoinverse.

We consider a Markov decision process (MDP) $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma, \nu_0)$ 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 probability transition kernel, $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$, is the reward function, γ is the discount factor and ν_0 is the initial state distribution. We assume that initial state is fixed to s_0 for simplicity. We assume $\mathcal{S}$ and $\mathcal{A}$ are finite, but potentially very large. We assume the reward function r is deterministic and known to the learner. The probability transition kernel P is unknown to the learner.

The interaction protocol between the learner and the MDP is as follows. The learner interacts with the MDP starting from the initial state $s_0 \in \mathcal{S}$. At each step $t = 0, 1, \ldots,$ the learner chooses an action $a_t \in \mathcal{A}$ and observes the reward $r(s_t, a_t)$ and the next state s_{t+1}. The next state s_{t+1} is drawn by the environment from $P(\cdot|s_t, a_t)$.

We define the normalized expected cumulative rewards

$$J(\pi) \coloneqq (1-\gamma)\mathbb{E}^\pi \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \right]$$

where $\mathbb{E}^\pi$ is the expectation with respect to the distribution of the trajectory $(s_0, a_0, s_1, a_1, \ldots)$ induced by the interaction of the probability transition P and the policy π. The normalizing factor $1 - \gamma$ makes $J(\pi) \in [0, 1]$ for all $i = 0, \ldots, I$.

The goal of RL is to find a policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ that maximizes the reward.

2.1. Linear MDP

For sample-efficient learning for arbitrarily large state space, we assume access to a feature mapping $\boldsymbol{\varphi} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ that reduces the dimension of the problem as follows.

Assumption A (Linear MDP)**.** We assume that the transition and the reward functions can be expressed as a linear function of a *known* feature map $\boldsymbol{\varphi} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ such that

$$r(s,a) = \langle \boldsymbol{\varphi}(s,a), \boldsymbol{\theta} \rangle, \quad P(s'|s,a) = \langle \boldsymbol{\varphi}(s,a), \boldsymbol{\psi}(s') \rangle$$

for all $(s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$ where $\boldsymbol{\theta} \in \mathbb{R}^d$ is a *known* parameter for the reward function and $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_d)$ is a vector of d *unknown* (signed) measures on $\mathcal{S}$.

The linear MDP assumption is widely studied in the RL literature for studying theoretical properties of RL with function approximation (Jin et al., 2020). As is commonly done in works on linear MDPs, we further make the following boundedness assumptions. Without loss of generality (see Appendix A in Wei et al. (2021) for justification), we assume

$$\|\boldsymbol{\varphi}(s,a)\|_2 \leq 1, \quad \|\boldsymbol{\theta}\|_2 \leq \sqrt{d}$$

for all $(s,a) \in \mathcal{S} \times \mathcal{A}$. We further make a technical assumption, also made by Wagenmaker et al. (2022), that for some constant D_ψ,

$$\||\boldsymbol{\psi}|(\mathcal{S})\|_2 \leq D_\psi \sqrt{d}$$

where $|\boldsymbol{\psi}|(\mathcal{S}) = \sum_{s \in \mathcal{S}} (|\psi_1(s)|, \ldots, |\psi_d(s)|)$. This assumption holds, for example, when ψ_i are probability measures on $\mathcal{S}$, in which case the assumption holds with $D_\psi = 1$.

The linear structure implies a low-dimensional factorization of key quantities as we discuss below. Let $\boldsymbol{P} \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|\times|\mathcal{S}|}$ be the matrix representation of the probability transition kernel P with $(\boldsymbol{P})_{(s,a),s'} = P(s'|s,a)$ for $(s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$. Then, the linear structure gives

$$\boldsymbol{P} = \boldsymbol{\Phi}\boldsymbol{\Psi}$$

where $\boldsymbol{\Psi} \in \mathbb{R}^{d \times |\mathcal{S}|}$ is the *unknown* matrix of all measures with rows $(\psi_i(s'))_{s' \in \mathcal{S}}$ and $\boldsymbol{\Phi} \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|\times d}$ is the *known* matrix of all feature vectors with rows $(\boldsymbol{\varphi}(s,a))_{(s,a)\in\mathcal{S}\times\mathcal{A}}$.

Let Q^π be the action value function of a policy π with respect to a reward function $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is defined as follows.

$$Q^\pi(s,a) = \mathbb{E}^\pi \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) | s_0 = s, a_0 = a \right].$$

It is the expected discounted cumulative reward starting from the state-action pair (s,a) and then executing the policy π every time step. Similarly, the state value function of a policy π with respect to a reward function r is defined as

$$V^\pi(s) = \mathbb{E}^\pi \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) | s_0 = s \right].$$

It is the expected discounted cumulative reward starting from the state s and then executing the policy π every time step.

We use $\boldsymbol{Q}^\pi \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|}$ to denote the matrix representation of the function Q^π such that $(\boldsymbol{Q}^\pi)_{s,a} = Q^\pi(s,a)$ and $\boldsymbol{V}^\pi \in \mathbb{R}^{|\mathcal{S}|}$ the vector representation of the function V^π such that $\boldsymbol{V}^\pi_s = V^\pi(s)$. With these notations, the well known Bellman equation $Q^\pi(s,a) = r(s,a) + \gamma P V^\pi(s,a)$ (see e.g. Puterman (2014)) can be written as

$$\boldsymbol{Q}^\pi = \boldsymbol{r} + \gamma \boldsymbol{P}\boldsymbol{V}^\pi = \boldsymbol{\Phi}(\boldsymbol{\theta} + \gamma \boldsymbol{\Psi}\boldsymbol{V}^\pi) = \boldsymbol{\Phi}\boldsymbol{\zeta}^\pi \tag{1}$$

where $\boldsymbol{r} \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|}$ is the matrix representation of the reward function r and we define $\boldsymbol{\zeta}^\pi := \boldsymbol{\theta} + \gamma \boldsymbol{\Psi}\boldsymbol{V}^\pi \in \mathbb{R}^d$. This shows that the action value function is linear in the feature vector:

$$Q^\pi(s,a) = \langle \boldsymbol{\varphi}(s,a), \boldsymbol{\zeta}^\pi \rangle.$$

Due to the boundedness assumptions $\|\boldsymbol{\theta}\|_2 \leq \sqrt{d}$ and $\||\boldsymbol{\psi}|(\mathcal{S})\|_2 \leq D_\psi \sqrt{d}$, and the fact that $V^\pi(s) \in [0, \frac{1}{1-\gamma}]$, the norm of the parameter $\boldsymbol{\zeta}^\pi$ is bounded by

$$\|\boldsymbol{\zeta}^\pi\|_2 \leq \sqrt{d} + \frac{\gamma D_\psi \sqrt{d}}{1-\gamma} = \mathcal{O}\left(\frac{D_\psi \sqrt{d}}{1-\gamma} \right).$$

We define $D_\zeta := \sqrt{d} + \frac{\gamma D_\psi \sqrt{d}}{1-\gamma}$.

2.2. Offline Learning and Data Coverage

We consider the offline learning setting where the agent has access to a dataset $\mathcal{D} = (s_j, a_j, s'_j)_{j=1}^n$. The pairs (s_j, a_j), $j = 1, \ldots, n$ are assumed to be i.i.d. samples from a distribution $\mu_B \in \Delta(\mathcal{S} \times \mathcal{A})$ and each s'_j is sampled from $P(\cdot \mid s_j, a_j)$. Such an i.i.d. assumption on the offline dataset is commonly made in the offline RL literature (Xie et al., 2021; Zhan et al., 2022; Chen & Jiang, 2022; Zhu et al., 2023) to facilitate the analysis of concentration bounds.

A major challenge in sample efficient offline RL is *distribution shift*, which refers to the mismatch of state-action distribution of the offline dataset and the target (optimal) policy. For sample efficient learning, we require an assumption on data coverage that guarantees the distribution of target policy is covered by the offline dataset. A common data coverage assumption in offline RL is concentrability assumption that limits the ratio of occupancy measure of target policy to that of behavior policy. The normalized occupancy measure of a policy π is defined as

$$\mu^\pi(s,a) = (1-\gamma)\mathbb{E}^\pi \left[\sum_{t=0}^{\infty} \gamma^t \mathbb{I}\{s_t = s, a_t = a\} \right].$$

Roughly, it is the normalized count of the visitation of state-action pair (s,a) when executing the policy π. It is normalized to ensure $\mu^\pi(s,a)$ is a probability measure on $\mathcal{S} \times \mathcal{A}$ so

that $\sum_{s,a} \mu(s,a) = 1$. We use the notation $\boldsymbol{\mu}^\pi \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|}$ to denote the matrix representation of the function $\mu^\pi(s,a)$. A commonly used concentrability assumption is as follows.

Assumption B (Concentrability)**.** For an optimal policy π^*, we have

$$\frac{\mu^*(s,a)}{\mu_B(s,a)} \le C^*$$

for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$ with $\mu_B(s,a) > 0$ where we write $\mu^* = \mu^{\pi^*}$. The bound C^* is known to the learner.

Concentrability assumption is widely used in offline RL with general function approximation (Munos, 2003; 2005). The assumption requires the ratio $\mu^*(s,a)/\mu_B(s,a)$ to be bounded for all $(s,a) \in \mathcal{S}\times\mathcal{A}$ for which $\mu_B(s,a)$ is positive. As discussed by Gabbianelli et al. (2024), in the linear function approximation setting with access to a feature map, we can define data coverage in the feature space rather than in the state-action space. We defer discussion on our result that uses data coverage assumption in feature space to Section 4.1.

3. Algorithm Design

Our algorithm is motivated by the linear programming formulation of the reinforcement learning problem.

$$\begin{aligned} \max_{\boldsymbol{\mu}\ge\mathbf{0}} \quad & \langle \boldsymbol{r}, \boldsymbol{\mu}\rangle \\ \text{subject to} \quad & \boldsymbol{E}^T\boldsymbol{\mu} = (1-\gamma)\boldsymbol{\nu}_0 + \gamma \boldsymbol{P}^T\boldsymbol{\mu}. \end{aligned}$$

Here, $\boldsymbol{E} \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|\times|\mathcal{S}|}$ denotes the matrix with $\boldsymbol{E}_{(s,a),s'} = \mathbb{I}\{s = s'\}$ and $\boldsymbol{\nu}_0$ denotes the initial state distribution, which is assumed to be $\boldsymbol{e}_{s_0}$. Note that the sth entry of $\boldsymbol{E}^T\boldsymbol{\mu}$ is $\sum_a \mu(s,a)$, which is the sum of $\mu(s,\cdot)$ over all possible values of a. The optimization variable $\boldsymbol{\mu} \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|}$ has the interpretation of the normalized occupancy measure. The objective function has the interpretation of the value of a policy, which can be seen by

$$J(\pi) = (1-\gamma)\mathbb{E}^\pi\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)\right] = \langle \boldsymbol{r}, \boldsymbol{\mu}^\pi\rangle.$$

The constraint $\boldsymbol{E}^T\boldsymbol{\mu} = (1-\gamma)\boldsymbol{\nu}_0 + \gamma\boldsymbol{P}^T\boldsymbol{\mu}$, called Bellman flow constraint, makes sure that $\boldsymbol{\mu}$ is a permissible occupancy measure in the sense that there exists a policy π that induces the measure, i.e., $\mu = \mu^\pi$ for some policy π.

We use $\boldsymbol{r} = \boldsymbol{\Phi}\boldsymbol{\theta}$ and $\boldsymbol{P} = \boldsymbol{\Phi}\boldsymbol{\Psi}$, which hold by the linear MDP assumption (Assumption A), to rewrite the linear program as

$$\begin{aligned} \max_{\boldsymbol{\mu}\ge 0} \quad & \langle \boldsymbol{\theta}, \boldsymbol{\Phi}^T\boldsymbol{\mu}\rangle \\ \text{subject to} \quad & \boldsymbol{E}^T\boldsymbol{\mu} = (1-\gamma)\boldsymbol{\nu}_0 + \gamma\boldsymbol{\Psi}^T\boldsymbol{\Phi}^T\boldsymbol{\mu} \end{aligned}$$

Note that the optimization variable $\boldsymbol{\mu} \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|}$ is high-dimensional that depends on the size of $\mathcal{S}$. Following Gabbianelli et al. (2024), with the goal of computational and statistical efficiency, we introduce a low-dimensional optimization variable $\boldsymbol{\lambda} = \boldsymbol{\Phi}^T\boldsymbol{\mu} \in \mathbb{R}^d$, which has the interpretation of the average occupancy in the feature space. With the reparametrization, the optimization problem becomes

$$\begin{aligned} \max_{\boldsymbol{\mu}\ge\mathbf{0},\boldsymbol{\lambda}} \quad & \langle \boldsymbol{\theta}, \boldsymbol{\lambda}\rangle \\ \text{subject to} \quad & \boldsymbol{E}^T\boldsymbol{\mu} = (1-\gamma)\boldsymbol{\nu}_0 + \gamma\boldsymbol{\Psi}^T\boldsymbol{\lambda} \\ & \boldsymbol{\lambda} = \boldsymbol{\Phi}^T\boldsymbol{\mu}. \end{aligned}$$

The dual of the linear program above is

$$\begin{aligned} \min_{\boldsymbol{v},\boldsymbol{\zeta}} \quad & (1-\gamma)\langle \boldsymbol{\nu}_0, \boldsymbol{v}\rangle \\ \text{subject to} \quad & \boldsymbol{\zeta} = \boldsymbol{\theta} + \gamma\boldsymbol{\Psi}\boldsymbol{v} \\ & \boldsymbol{E}\boldsymbol{v} \ge \boldsymbol{\Phi}\boldsymbol{\zeta}. \end{aligned}$$

The dual variable $\boldsymbol{v} \in \mathbb{R}^{|\mathcal{S}|}$ has the interpretation of the vector representation of the state value function and $\boldsymbol{\zeta} \in \mathbb{R}^d$ the parameter such that $\boldsymbol{\Phi}\boldsymbol{\zeta} \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|}$ is the vector representation of the state-action value function. The Lagrangian associated to this pair of linear programs is

$$\begin{aligned} & L(\boldsymbol{\lambda}, \boldsymbol{\mu}; \boldsymbol{v}, \boldsymbol{\zeta}) \\ & = (1-\gamma)\langle \boldsymbol{\nu}_0, \boldsymbol{v}\rangle + \langle \boldsymbol{\lambda}, \boldsymbol{\theta} + \gamma\boldsymbol{\Psi}\boldsymbol{v} - \boldsymbol{\zeta}\rangle + \langle \boldsymbol{\mu}, \boldsymbol{\Phi}\boldsymbol{\zeta} - \boldsymbol{E}\boldsymbol{v}\rangle \\ & = \langle \boldsymbol{\lambda}, \boldsymbol{\theta}\rangle + \langle \boldsymbol{v}, (1-\gamma)\boldsymbol{\nu}_0 + \gamma\boldsymbol{\Psi}^T\boldsymbol{\lambda} - \boldsymbol{E}^T\boldsymbol{\mu}\rangle \\ & \quad + \langle \boldsymbol{\zeta}, \boldsymbol{\Phi}^T\boldsymbol{\mu} - \boldsymbol{\lambda}\rangle. \end{aligned}$$

Note that the optimization variables $\boldsymbol{\lambda}, \boldsymbol{\zeta} \in \mathbb{R}^d$ are low-dimensional, but $\boldsymbol{\mu} \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|}$ and $\boldsymbol{v} \in \mathbb{R}^{|\mathcal{S}|}$ are not. With the goal of running a primal-dual algorithm on the Lagrangian using only low-dimensional variables, we introduce policy variable π and parameterize $\boldsymbol{\mu}$ and $\boldsymbol{v}$, following Gabbianelli et al. (2024), by

$$\mu_{\boldsymbol{\lambda},\pi}(s,a) = \pi(a|s)\left[(1-\gamma)\nu_0(s) + \gamma\langle \psi(s), \boldsymbol{\lambda}\rangle\right] \quad (2)$$

$$v_{\boldsymbol{\zeta},\pi}(s) = \sum_a \pi(a|s)\langle \boldsymbol{\zeta}, \boldsymbol{\varphi}(s,a)\rangle. \quad (3)$$

The choice of $\boldsymbol{\mu}_{\boldsymbol{\lambda},\pi}$ makes the Bellman flow constraint $\boldsymbol{E}^T\boldsymbol{\mu}_{\boldsymbol{\lambda},\pi} = (1-\gamma)\boldsymbol{\nu}_0 + \gamma\boldsymbol{\Psi}^T\boldsymbol{\lambda}$ of the primal problem satisfied. Also, the choice of $\boldsymbol{v}_{\boldsymbol{\zeta},\pi}$ makes $\langle \boldsymbol{\mu}, \boldsymbol{\Phi}\boldsymbol{\zeta} - \boldsymbol{E}\boldsymbol{v}_{\boldsymbol{\zeta},\pi}\rangle = 0$. Using the above parameterization, the Lagrangian can be rewritten in terms of $\boldsymbol{\zeta}, \boldsymbol{\lambda}, \pi$ as follows:

$$\begin{aligned} & f(\boldsymbol{\lambda}, \boldsymbol{\zeta}, \pi) \\ & \quad = \langle \boldsymbol{\lambda}, \boldsymbol{\theta}_0\rangle + \langle \boldsymbol{\zeta}, \boldsymbol{\Phi}^T\boldsymbol{\mu}_{\boldsymbol{\lambda},\pi} - \boldsymbol{\lambda}\rangle \quad (4) \\ & \quad = (1-\gamma)\langle \boldsymbol{\nu}_0, \boldsymbol{v}_{\boldsymbol{\zeta},\pi}\rangle + \langle \boldsymbol{\lambda}, \boldsymbol{\theta}_0 + \gamma\boldsymbol{\Psi}\boldsymbol{v}_{\boldsymbol{\zeta},\pi} - \boldsymbol{\zeta}\rangle. \quad (5) \end{aligned}$$

At the cost of having to keep track of π, we can now run a primal-dual algorithm on the low-dimensional variables $\boldsymbol{\zeta}$,

$\boldsymbol{\lambda}$. The introduction of π in the equation does not make the algorithm inefficient because we can only keep track of the distribution $\pi(s|a)$ for state-action pairs that appear in the dataset. See Appendix C for detail.

Previous work on offline linear MDP (Gabbianelli et al., 2024) runs primal-dual algorithm on the variables $\boldsymbol{\zeta}$ and $\boldsymbol{\beta} = \boldsymbol{\Lambda}^{\dagger}\boldsymbol{\lambda}$ by estimating the gradient of the Lagrangian with respect to the variables. Their algorithm requires running gradient descent algorithm on $\boldsymbol{\zeta}$ for every gradient ascent step of $\boldsymbol{\beta}$, leading to a double-loop algorithm structure. Since each gradient descent/ascent step requires fresh copy of independent data, the double-loop algorithm leads to sample complexity of $\mathcal{O}(\epsilon^{-4})$. We sidestep the need of the double-loop structure and obtain $\mathcal{O}(\epsilon^{-2})$ sample complexity by restricting the values of $\boldsymbol{\lambda}$ to a carefully designed confidence set that allows estimating the gradient uniformly over the choices of $\boldsymbol{\lambda}$, $\boldsymbol{\zeta}$ and π. We outline the argument in the following section.

3.1. Analysis

For a given policy π, recall that $\boldsymbol{\zeta}^{\pi} \in \mathbb{R}^d$ is the parameter satisfying $\boldsymbol{Q}^{\pi} = \boldsymbol{\Phi}\boldsymbol{\zeta}^{\pi}$. It can be shown that for any $\boldsymbol{\lambda} \in \mathbb{R}^d$,

$$f(\boldsymbol{\zeta}^{\pi}, \boldsymbol{\lambda}, \pi) = J(\pi).$$

Also, defining $\boldsymbol{\lambda}^{\pi} = \boldsymbol{\Phi}^T\boldsymbol{\mu}^{\pi}$, which has the interpretation of the average occupancy in the feature space when executing the policy π, it can be shown that for any $\boldsymbol{\zeta} \in \mathbb{R}^d$,

$$f(\boldsymbol{\zeta}, \boldsymbol{\lambda}^{\pi}, \pi) = J(\pi).$$

See Appendix E.2 for proofs. Hence, for any sequences $\{\pi_t\}$, $\{\boldsymbol{\theta}_t\} \subset \mathbb{R}^d$ and $\{\boldsymbol{\lambda}_t\} \subset \mathbb{R}^d$, we have

$$\begin{aligned} J(\pi^*) - J(\pi_t) &= f(\boldsymbol{\zeta}_t, \boldsymbol{\lambda}^*, \pi^*) - f(\boldsymbol{\zeta}^{\pi_t}, \boldsymbol{\lambda}_t, \pi_t) \\ &= \underbrace{(f(\boldsymbol{\zeta}_t, \boldsymbol{\lambda}^*, \pi^*) - f(\boldsymbol{\zeta}_t, \boldsymbol{\lambda}^*, \pi_t))}_{\textsc{Reg}_t^{\pi}} \\ &\quad + \underbrace{(f(\boldsymbol{\zeta}_t, \boldsymbol{\lambda}^*, \pi_t) - f(\boldsymbol{\zeta}_t, \boldsymbol{\lambda}_t, \pi_t))}_{\textsc{Reg}_t^{\lambda}} \\ &\quad + \underbrace{(f(\boldsymbol{\zeta}_t, \boldsymbol{\lambda}_t, \pi_t) - f(\boldsymbol{\zeta}^{\pi_t}, \boldsymbol{\lambda}_t, \pi_t))}_{\textsc{Reg}_t^{\zeta}} \end{aligned}$$

where we use the notation $\boldsymbol{\lambda}^* = \boldsymbol{\lambda}^{\pi^*}$.

Note that the suboptimality $J(\pi^*) - J(\pi_t)$ is decomposed into regrets of the three players. As long as we show that the sums of the three regrets over $t = 1, \ldots, T$ are sublinear in T, we obtain $\frac{1}{T}\sum_{t=1}^{T} J(\pi^*) - J(\pi_t) = J(\pi^*) - J(\bar{\pi}) = o(1)$ where $\bar{\pi} = \mathrm{Unif}(\pi_1, \ldots, \pi_T)$ is the mixture policy that chooses a policy among $\pi_1, \ldots, \pi_T$ uniformly at random and runs the chosen policy for the entire trajectory.

In the rest of the section, we sketch analyses of bounding the regrets of the three players. These analyses will motivate our algorithm presented in Section 4.

3.2. Bounding Regret of π-player

Using Equation (5), the regret of π-player simplifies to

$$\begin{aligned} \mathrm{Reg}_t^{\pi} &= f(\boldsymbol{\zeta}_t, \boldsymbol{\lambda}^*, \pi^*) - f(\boldsymbol{\zeta}_t, \boldsymbol{\lambda}^*, \pi_t) \\ &= \langle \boldsymbol{\nu}^*, \boldsymbol{v}_{\boldsymbol{\zeta}_t, \pi^*} - \boldsymbol{v}_{\boldsymbol{\zeta}_t, \pi_t} \rangle \\ &= \langle \boldsymbol{\nu}^*, \textstyle\sum_a (\pi^*(a|\cdot) - \pi_t(a|\cdot)) \langle \boldsymbol{\zeta}_t, \boldsymbol{\varphi}(\cdot, a) \rangle \rangle. \end{aligned}$$

where we define $\boldsymbol{\nu}^{\pi} = (1-\gamma)\boldsymbol{\nu}_0 + \gamma\boldsymbol{\Psi}^T\boldsymbol{\lambda}^{\pi}$ as the state occupancy measure induced by π and write $\boldsymbol{\nu}^* = \boldsymbol{\nu}^{\pi^*}$. The regret can be bounded if π-player updates its policy using an exponentiation algorithm (Zanette et al., 2021)

$$\pi_{t+1} = \sigma\left(\alpha \sum_{i=1}^{t} \boldsymbol{\Phi}\boldsymbol{\zeta}_i\right)$$

where $\sigma(\boldsymbol{q})$ for $\boldsymbol{q} \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|}$ is a softmax policy with

$$\sigma(\boldsymbol{q})(a|s) := \frac{\exp(q(s,a))}{\sum_{a'} \exp(q(s,a'))}.$$

Based on the standard mirror descent analysis by Gabbianelli et al. (2024) (Appendix D.1) we can show that, choosing $\alpha = \mathcal{O}((1-\gamma)\sqrt{\log|\mathcal{A}|/(dT)})$ gives

$$\frac{1}{T}\sum_{t=1}^{T} \textsc{Reg}_t^{\pi} \le \mathcal{O}\left(\frac{1}{1-\gamma}\sqrt{(d\log|\mathcal{A}|)/T}\right)$$

which vanishes as T increases. Consequently, choosing T to be at least $\Omega(\frac{d\log|\mathcal{A}|}{(1-\gamma)^2\epsilon^2})$ gives $\frac{1}{T}\sum_{t=1}^{T}\mathrm{Reg}_t^{\pi} \le \epsilon$. Note that when the exponentiation algorithm is employed, the π-player does not need to know the value of $\boldsymbol{\zeta}_t$ when choosing π_t, allowing the π-player to play before the ζ-player. Another benefit of the exponentiation algorithm is that the policy chosen by the π-player is restricted to the softmax function class $\Pi(D_{\pi})$ where $\Pi(\cdot)$ is defined as

$$\Pi(B) := \{\sigma(\boldsymbol{\Phi}\boldsymbol{z}) : \boldsymbol{z} \in \mathbb{B}_d(B)\}. \tag{6}$$

and $D_{\pi} := \alpha T D_{\zeta}$. The restriction allows statistically efficient estimation of quantities that depend on policies in $\Pi(B)$ via covering argument on $\Pi(B)$, as we will see in later sections.

3.3. Bounding Regret of ζ-player

Using Equation (4), the regret of ζ-player simplifies to

$$\begin{aligned} \textsc{Reg}_t^{\zeta} &= f(\boldsymbol{\zeta}_t, \boldsymbol{\lambda}_t, \pi_t) - f(\boldsymbol{\zeta}^{\pi_t}, \boldsymbol{\lambda}_t, \pi_t) \\ &= \langle \boldsymbol{\zeta}_t - \boldsymbol{\zeta}^{\pi_t}, \boldsymbol{\Phi}^T\boldsymbol{\mu}_{\boldsymbol{\lambda}_t, \pi_t} - \boldsymbol{\lambda}_t \rangle. \end{aligned}$$

Recall that $\boldsymbol{\mu}_{\boldsymbol{\lambda},\pi} = \pi \circ \boldsymbol{E}[(1-\gamma)\boldsymbol{\nu}_0 + \gamma\boldsymbol{\Psi}^T\boldsymbol{\lambda}]$. The only unknown quantity in the regret is $\boldsymbol{\Psi}^T\boldsymbol{\lambda} \in \mathbb{R}^{|\mathcal{S}|}$. Note that $\boldsymbol{\Psi}^T\boldsymbol{\varphi}(s,a) = (P(s'|s,a))_{s'\in\mathcal{S}}$ is a next-state distribution given current state-action pair (s,a), and $\boldsymbol{e}_{s'_k}$ is an unbiased

estimator for $\boldsymbol{\Psi}^T\boldsymbol{\varphi}(s_k,a_k)$. Hence, if $\boldsymbol{\lambda}$ is a linear combination $\sum_{k=1}^n c_k\boldsymbol{\varphi}(s_k,a_k)$, we can construct an unbiased estimator $\sum_{k=1}^n c_k\boldsymbol{e}_{s'_k}$ for $\boldsymbol{\Psi}^T\boldsymbol{\lambda}$. Motivated by this observation, to facilitate the algorithm design for the ζ-player, we will restrict the λ-player to choose a linear combination of feature vectors that appear in the dataset to allow estimating $\boldsymbol{\Psi}^T\boldsymbol{\lambda}$. Specifically, we strict the λ-player to choose $\boldsymbol{\lambda}_t$ from the following set where the bound B will be chosen later.

$$\mathcal{C}_n(B) \coloneqq \left\{\frac{1}{n}\sum_{k=1}^n c_k\boldsymbol{\varphi}(s_k,a_k) : c_1,\dots,c_n \in [-B,B]\right\}. \tag{7}$$

Given the restriction, we can parameterize the value of $\boldsymbol{\lambda}_t$ by the coefficients $\boldsymbol{c}_t \in [-B,B]^n$ for some bound B, and write $\boldsymbol{\lambda}_t = \boldsymbol{\lambda}(\boldsymbol{c}_t)$ where we define

$$\boldsymbol{\lambda}(\boldsymbol{c}) \coloneqq \frac{1}{n}\sum_{k=1}^n c_k\boldsymbol{\varphi}(s_k,a_k)$$

Following the previous discussion, we define the estimates for $\boldsymbol{\Psi}^T\boldsymbol{\lambda}(c)$ and $\boldsymbol{\mu}_{\boldsymbol{\lambda}(\boldsymbol{c}),\pi}$ parameterized by $\boldsymbol{c} \in [-B,B]^n$:

$$\widehat{\boldsymbol{\Psi}^T\boldsymbol{\lambda}}(\boldsymbol{c}) \coloneqq \frac{1}{n}\sum_{k=1}^n c_k\boldsymbol{e}_{s'_k}$$

$$\widehat{\boldsymbol{\mu}}_{\boldsymbol{\lambda}(\boldsymbol{c}),\pi} \coloneqq \pi \circ \boldsymbol{E}[(1-\gamma)\boldsymbol{\nu}_0 + \gamma\widehat{\boldsymbol{\Psi}^T\boldsymbol{\lambda}}(\boldsymbol{c})].$$

These estimates enjoy the following concentration bound, which can be shown using matrix Bernstein inequality. See Appendix D.2 for a proof.

Lemma 3.1. *For a fixed $\boldsymbol{\lambda}(\boldsymbol{c}) = \frac{1}{n}\sum_{k=1}^n c_k\boldsymbol{\varphi}(s_k,a_k)$ with $|c_k| \le B$ for $k=1,\dots,n$, and a policy π, we have*

$$\|\boldsymbol{\Phi}^T\boldsymbol{\mu}_{\boldsymbol{\lambda}(\boldsymbol{c}),\pi} - \boldsymbol{\Phi}^T\widehat{\boldsymbol{\mu}}_{\boldsymbol{\lambda}(\boldsymbol{c}),\pi}\|_2 \le \mathcal{O}\left(B\sqrt{\frac{\log(d/\delta)}{n}}\right)$$

with probability at least $1-\delta$ conditional on the data of state-action pairs $\{(s_k,a_k)\}_{k=1}^n$.

For estimating the regret $\langle \boldsymbol{\zeta}_t - \boldsymbol{\zeta}^{\pi_t}, \boldsymbol{\Phi}^T\boldsymbol{\mu}_{\boldsymbol{\lambda}(\boldsymbol{c}_t),\pi_t} - \boldsymbol{\lambda}(\boldsymbol{c}_t)\rangle$, we need a uniform concentration bound on the estimates $\boldsymbol{\Phi}^T\widehat{\boldsymbol{\mu}}_{\boldsymbol{\lambda}(\boldsymbol{c}_t),\pi}$ over $\boldsymbol{\lambda}(\boldsymbol{c})$ and π. The restriction on the π-player to choose a policy in the softmax function class defined in (6) allows converting the concentration bound for a fixed policy π in Lemma 3.1 to a uniform concentration bound over all policies in the softmax function class via a covering argument. The conversion is possible due to the fact that the log covering number for the softmax function class is bounded by $\widetilde{\mathcal{O}}(d)$ (see Lemma A.3 in Appendix A). However, such a conversion to a uniform concentration bound over all $\boldsymbol{\lambda}(\boldsymbol{c})$ for $\boldsymbol{c} \in [-B,B]^n$ is elusive since a naive covering argument on the space of parameters $[-B,B]^n$ will give a log covering number bound of $\mathcal{O}(n)$. To sidestep this issue, we exploit the fact that $\mathcal{C}_n(B)$ can be spanned by a set of spanners $\{\boldsymbol{\varphi}(s_j,a_j)\}_{j\in\mathcal{I}}$ for some index set $\mathcal{I} \subseteq \{1,\dots,n\}$ of size at most d. This can be seen by the following lemma by Awerbuch & Kleinberg (2008).

Lemma 3.2 (Barycentric spanner)**.** *Let $\mathcal{K} \subseteq \mathbb{R}^d$ be compact set. Then, there exists a spanner $\{\boldsymbol{\phi}_1,\dots\boldsymbol{\phi}_d\} \subset \mathcal{K}$ such that any vector $\boldsymbol{x} \in \mathcal{K}$ can be represented as $\boldsymbol{x} = \sum_{i=1}^d c_i\boldsymbol{\phi}_i$ where $c_i \in [-C,C]$ for all $i=1,\dots,d$. Such a spanner is called a C-approximate barycentric spanner for $\mathcal{K}$. If $\mathcal{K}$ is finite, we can find a C-approximate barycentric spanner in time complexity $\mathcal{O}(nd^2\log_C d)$.*

Applying this lemma, we can compute a 2-approximate barycentric spanner $\{\boldsymbol{\varphi}(s_j,a_j)\}_{j\in\mathcal{I}}$ for $\{\boldsymbol{\varphi}(s_k,a_k)\}_{k=1}^n$ where $\mathcal{I} \subseteq \{1,\dots,n\}$ is an index set of size d. Given any $\boldsymbol{c} \in [-B,B]^n$, we can convert it to $\boldsymbol{c}' \in [-2B,2B]^n$ with c'_k nonzero only if $k \in \mathcal{I}$ such that $\boldsymbol{\lambda}(\boldsymbol{c}) = \boldsymbol{\lambda}(\boldsymbol{c}')$. This can be seen by

$$\begin{aligned}\boldsymbol{\lambda}(\boldsymbol{c}) &= \frac{1}{n}\sum_{k=1}^n c_k\boldsymbol{\varphi}(s_k,a_k)\\ &= \sum_{j\in\mathcal{I}}\left(\frac{1}{n}\sum_{k=1}^n b_{kj}c_k\right)\boldsymbol{\varphi}(s_j,a_j) = \boldsymbol{\lambda}(\boldsymbol{c}')\end{aligned}$$

where the coefficients $b_{kj} \in [-2,2]$ are such that $\boldsymbol{\varphi}(s_k,a_k) = \sum_{j\in\mathcal{I}} b_{kj}\boldsymbol{\varphi}(s_j,a_j)$, which exist by the fact that $\{\boldsymbol{\varphi}(s_j,a_j)\}_{j\in\mathcal{I}}$ is a 2-approximate barycentric spanner of $\{\boldsymbol{\varphi}(s_k,a_k)\}_{k=1}^n$. We summarize the definition of the conversion from $\boldsymbol{c}$ to $\boldsymbol{c}'$ that satisfies $\boldsymbol{\lambda}(\boldsymbol{c}) = \boldsymbol{\lambda}(\boldsymbol{c}')$.

Definition 3.3. Given a dataset $\{(s_k,a_k,s'_k)\}_{k=1}^n$, let $\{\boldsymbol{\varphi}(s_j,a_j)\}_{j\in\mathcal{I}}$ be a 2-approximate barycentric spanner for $\{\boldsymbol{\varphi}(s_k,a_k)\}_{k=1}^n$ with $|\mathcal{I}| \le d$. We define the conversion of $\boldsymbol{c} \in \mathbb{R}^n$ to $\boldsymbol{c}' \in \mathbb{R}^n$ as

$$c'_j = \begin{cases}\frac{1}{n}\sum_{k=1}^n b_{kj}c_k & \text{if } j \in \mathcal{I}\\ 0 & \text{otherwise.}\end{cases} \tag{8}$$

where b_{kj} are the coefficients such that $\boldsymbol{\varphi}(s_k,a_k) = \sum_{j\in\mathcal{I}} b_{kj}\boldsymbol{\varphi}(s_j,a_j)$ with $b_{kj} = 0$ for $j \notin \mathcal{I}$.

Given $\boldsymbol{c}_t \in [-B,B]^n$ such that $\boldsymbol{\lambda}(\boldsymbol{c}_t) \in \mathcal{C}_n(B)$, let $\boldsymbol{c}'_t \in [-2B,2B]^n$ be the conversion such that $\boldsymbol{\lambda}(\boldsymbol{c}'_t) = \boldsymbol{\lambda}(\boldsymbol{c}_t)$ with only the coefficients with indices in $\mathcal{I}$ nonzero. The converted coefficients $\boldsymbol{c}'_t \in \mathbb{R}^n$ live in a low dimensional space $\{\boldsymbol{c}' \in [-2B,2B]^n : c'_j = 0 \text{ if } j \notin \mathcal{I}\}$ with the log covering number of $\mathcal{O}(d)$. To use a covering argument, let $\boldsymbol{c}''_t$ be the covering center closest to $\boldsymbol{c}'_t$. Then we can decompose the regret of the ζ-player as

$$\begin{aligned}\mathrm{REG}_t^\zeta &= \langle \boldsymbol{\zeta}_t - \boldsymbol{\zeta}^{\pi_t}, \boldsymbol{\Phi}^T\boldsymbol{\mu}_{\boldsymbol{\lambda}(\boldsymbol{c}_t),\pi_t} - \boldsymbol{\lambda}(\boldsymbol{c}_t)\rangle\\ &= \langle \boldsymbol{\zeta}_t - \boldsymbol{\zeta}^{\pi_t}, \boldsymbol{\Phi}^T\boldsymbol{\mu}_{\boldsymbol{\lambda}(\boldsymbol{c}'_t),\pi_t} - \boldsymbol{\Phi}^T\boldsymbol{\mu}_{\boldsymbol{\lambda}_t(\boldsymbol{c}''_t),\pi_t}\rangle\\ &\quad + \langle \boldsymbol{\zeta}_t - \boldsymbol{\zeta}^{\pi_t}, \boldsymbol{\Phi}^T\boldsymbol{\mu}_{\boldsymbol{\lambda}(\boldsymbol{c}''_t),\pi_t} - \boldsymbol{\Phi}^T\widehat{\boldsymbol{\mu}}_{\boldsymbol{\lambda}(\boldsymbol{c}''_t),\pi_t}\rangle\\ &\quad + \langle \boldsymbol{\zeta}_t - \boldsymbol{\zeta}^{\pi_t}, \boldsymbol{\Phi}^T\widehat{\boldsymbol{\mu}}_{\boldsymbol{\lambda}(\boldsymbol{c}''_t),\pi_t} - \boldsymbol{\Phi}^T\widehat{\boldsymbol{\mu}}_{\boldsymbol{\lambda}(\boldsymbol{c}'_t),\pi_t}\rangle\\ &\quad + \langle \boldsymbol{\zeta}_t - \boldsymbol{\zeta}^{\pi_t}, \boldsymbol{\Phi}^T\widehat{\boldsymbol{\mu}}_{\boldsymbol{\lambda}(\boldsymbol{c}'_t),\pi_t} - \boldsymbol{\lambda}(\boldsymbol{c}'_t)\rangle.\end{aligned}$$

The first term can be bounded since $\boldsymbol{\lambda}(c'_t) \approx \boldsymbol{\lambda}_t(\boldsymbol{c}''_t)$. The second term can be bounded using a union bound of the concentration inequalities on $\boldsymbol{\Phi}^T \widehat{\boldsymbol{\mu}}_{\boldsymbol{\lambda}(\boldsymbol{c}''),\pi}$ over $\boldsymbol{c}''$ in the cover of $[-2B, 2B]^d$. The third term can be bounded since $c'_t \approx c''_t$. The last term, interpreted as a regret of the ζ-player against a dynamic action $\boldsymbol{\zeta}^{\pi_t}$, can be bounded by a greedy ζ-player that minimizes $\langle \cdot, \boldsymbol{\Phi}^T \widehat{\boldsymbol{\mu}}_{\boldsymbol{\lambda}(\boldsymbol{c}'_t),\pi_t} - \boldsymbol{\lambda}(\boldsymbol{c}'_t)\rangle$. The greedy strategy requires ζ-player to play after λ-player and π-player. The bounds lead to

$$\frac{1}{T}\sum_{t=1}^{T} \textsc{Reg}_t^{\zeta} \le \mathcal{O}\left(\frac{Bd}{1-\gamma}\sqrt{\frac{\log(Bdn/\delta)}{n}}\right).$$

In summary, we can bound the regret of ζ-player if ζ-player plays $\boldsymbol{\zeta}_t \in \mathbb{B}_d(D_\zeta)$ that minimizes $\langle \cdot, \boldsymbol{\Phi}^T \widehat{\boldsymbol{\mu}}_{\boldsymbol{\lambda}(\boldsymbol{c}'_t),\pi_t} - \boldsymbol{\lambda}(\boldsymbol{c}'_t)\rangle$. The greedy strategy requires ζ-player to play after λ-player and π-player. See Appendix D.2 for detailed analysis.

3.4. Bounding Regret of λ-player

Using Equation (5), the regret of λ-player simplifies to

$$\begin{aligned}\textsc{Reg}_t^{\lambda} &= f(\boldsymbol{\zeta}_t, \boldsymbol{\lambda}^*, \pi_t) - f(\boldsymbol{\zeta}_t, \boldsymbol{\lambda}_t, \pi_t)\\ &= \langle \boldsymbol{\lambda}^* - \boldsymbol{\lambda}_t, \underbrace{\boldsymbol{\theta} + \gamma \boldsymbol{\Psi} \boldsymbol{v}_{\boldsymbol{\zeta}_t,\pi_t} - \boldsymbol{\zeta}_t}_{=\boldsymbol{\xi}_t}\rangle\end{aligned}$$

The sum of $\textsc{Reg}_t^{\lambda}$ over $t = 1, \dots, T$ is the regret of the λ-player against a fixed action λ^* where the reward function at time t is $\langle \cdot, \boldsymbol{\xi}_t\rangle$. From the previous section, we require λ player to play before ζ-player, whose play affects $\boldsymbol{\xi}_t$. Hence, the decision of λ-player at time t must be made before the knowledge of $\boldsymbol{\xi}_t$. Assuming for now that $\boldsymbol{\xi}_t$ is known ($\boldsymbol{\xi}_t$ is in fact unknown and needs to be estimated since $\boldsymbol{\Psi}$ is unknown), the regret of λ-player can be made sublinear in T by employing a no-regret online convex optimization oracle (defined below) on $\mathcal{C}_n(B)$ as long as $\boldsymbol{\lambda}^* \in \mathcal{C}_n(B)$.

Definition 3.4. An algorithm is called a *no-regret online convex optimization oracle* with respect to a convex set $\mathcal{C}$ if, for any sequence of convex functions $h_1, \dots, h_T : \mathbb{R}^d \to [-1, 1]$ and for any $\boldsymbol{\lambda} \in \mathcal{C}$, the sequence of vectors $\boldsymbol{\lambda}_1, \dots, \boldsymbol{\lambda}_T \in \mathcal{C}$ produced by the algorithm satisfies

$$\frac{1}{T}\sum_{t=1}^{T} h_t(\boldsymbol{\lambda}_t) - h_t(\boldsymbol{\lambda}) \le \epsilon_{\text{opt}}^{\lambda}(T)$$

for some $\epsilon_{\text{opt}}^{\lambda}(T) > 0$ that converges to 0 as $T \to \infty$.

The online gradient descent algorithm (Hazan et al., 2016) is an example of a computationally efficient online convex optimization oracle. Employing a no-regret online convex optimization oracle with convex set $\mathcal{C}_n(B)$ on the sequence of functions $\langle \cdot, \boldsymbol{\xi}_t\rangle$, the λ-player can enjoy a sublinear regret against any fixed $\boldsymbol{\lambda} \in \mathcal{C}_n(B)$. However, $\boldsymbol{\lambda}^*$ may not lie in $\mathcal{C}_n(B)$ for any B. In fact, we can construct an example where $\boldsymbol{\lambda}^*$ is not in the span of $\{\boldsymbol{\varphi}(s_k, a_k)\}_{k=1}^n$ with probability at least $1/2$ (Lemma B.2). To sidestep this problem, we show $\boldsymbol{\lambda}^*$ can be approximated by a vector $\widehat{\boldsymbol{\lambda}}^*$ in $\mathcal{C}_n(C^*)$:

Lemma 3.5. *Under the concentrability assumption B, there exists $\widehat{\boldsymbol{\lambda}}^* \in \mathbb{R}^d$ of the form $\widehat{\boldsymbol{\lambda}}^* = \frac{1}{n}\sum_{k=1}^n c_k \boldsymbol{\varphi}(s_k, a_k)$ with $c_k \in [0, C^*]$, $k = 1, \dots, n$ such that*

$$\|\widehat{\boldsymbol{\lambda}}^* - \boldsymbol{\lambda}^*\|_2 \le \mathcal{O}\left(C^* \sqrt{\frac{\log(d/\delta)}{n}}\right)$$

with probability at least $1 - \delta$.

Note that this lemma is the only place the data coverage assumption is needed for our analysis and this is where we require choosing $B = C^*$. Also, in our algorithm, computing $\widehat{\boldsymbol{\lambda}}^*$ is not needed. Only the existence of such a vector $\widehat{\boldsymbol{\lambda}}^*$ is needed in the analysis. With the lemma above, we can approximate the regret as

$$\begin{aligned}\textsc{Reg}_t^{\lambda} &= \langle \boldsymbol{\lambda}^* - \boldsymbol{\lambda}_t, \boldsymbol{\theta} + \gamma \boldsymbol{\Psi} \boldsymbol{v}_{\boldsymbol{\zeta}_t,\pi_t} - \boldsymbol{\zeta}_t\rangle\\ &\approx \langle \widehat{\boldsymbol{\lambda}}^* - \boldsymbol{\lambda}_t, \boldsymbol{\theta} + \gamma \boldsymbol{\Psi} \boldsymbol{v}_{\boldsymbol{\zeta}_t,\pi_t} - \boldsymbol{\zeta}_t\rangle\end{aligned}$$

and argue that the sum of the above quantity over $t = 1, \dots, T$ is sublinear since the regret against $\widehat{\boldsymbol{\lambda}}^* \in \mathcal{C}_n(C^*)$ is sublinear when employing a no-regret online convex optimization oracle. Now, we deal with the fact that the term $\boldsymbol{\Psi} \boldsymbol{v}_{\boldsymbol{\zeta}_t,\pi_t}$ is unknown and needs to be estimated. Observing that $\langle \boldsymbol{\varphi}(s, a), \boldsymbol{\Psi} \boldsymbol{v}\rangle = \mathbb{E}_{s' \sim P(\cdot|s,a)}[\boldsymbol{v}(s')]$, we can estimate $\boldsymbol{\Psi} \boldsymbol{v} \in \mathbb{R}^d$ for any $\boldsymbol{v} \in \mathbb{R}^{|\mathcal{S}|}$ by regressing $v(s')$ on $\boldsymbol{\varphi}(s, a)$ using the triplets (s, a, s') in the dataset $\mathcal{D}$. Following the literature on linear bandits (Abbasi-Yadkori et al., 2011), we use the regularized least squares estimate

$$\widehat{\boldsymbol{\Psi} \boldsymbol{v}} := (n\widehat{\boldsymbol{\Lambda}}_n + \boldsymbol{I})^{-1} \textstyle\sum_{k=1}^n v(s'_k)\boldsymbol{\varphi}(s_k, a_k) \tag{9}$$

where $\widehat{\boldsymbol{\Lambda}}_n := \frac{1}{n}\sum_{k=1}^n \boldsymbol{\varphi}(s_k, a_k)\boldsymbol{\varphi}(s_k, a_k)^T$ is the empirical Gram matrix. By the well-known result for linear bandits (e.g. Theorem 2 in Abbasi-Yadkori et al. (2011)), we have the following high-probability concentration bound (Lemma B.4) for the estimate $\widehat{\boldsymbol{\Psi} \boldsymbol{v}}$ where $v : \mathcal{S} \to [0, D_v]$:

$$\|\boldsymbol{\Psi} \boldsymbol{v} - \widehat{\boldsymbol{\Psi} \boldsymbol{v}}\|_{n\widehat{\boldsymbol{\Lambda}}_n + \boldsymbol{I}} \le \mathcal{O}\left(D_v \sqrt{d \log(n/\delta)}\right).$$

Since we need concentration bound of $\widehat{\boldsymbol{\Psi} \boldsymbol{v}}_{\boldsymbol{\zeta}_t,\pi_t}$ where $\boldsymbol{v}_{\boldsymbol{\zeta}_t,\pi_t}$ are random, we need a uniform bound over all possible functions $v_{\boldsymbol{\zeta}_t,\pi_t}$. Since the domain of v has cardinality $|\mathcal{S}|$, a naive covering argument on the function space of v will make the bound scale with poly$(|\mathcal{S}|)$. To avoid this, we use a careful covering argument exploiting the fact that π_t is a softmax function parameterized by a d-dimensional vector and $\boldsymbol{\zeta}_t$ are d-dimensional vectors. With covering, we can show the following uniform concentration bound.

Lemma 3.6. *Consider a function class*

$$\mathcal{V} = \{v_{\boldsymbol{\zeta},\pi} \in (\mathcal{S} \to [0, D_v]) : \boldsymbol{\zeta} \in \mathbb{B}(D_\zeta), \pi \in \Pi(D_\pi)\}.$$

With probability at least $1-\delta$, we have

$$\|\boldsymbol{\Psi v} - \widehat{\boldsymbol{\Psi v}}\|_{n\widehat{\boldsymbol{\Lambda}}_n + \boldsymbol{I}} \le \mathcal{O}\left(D_v \sqrt{d \log(D_\zeta D_\pi n/\delta)}\right)$$

uniformly over $\boldsymbol{v} \in \mathcal{V}$ where $\widehat{\boldsymbol{\Psi v}}$ is the least squares estimate defined in (9).

See Lemma 3.6 in the Appendix for detail. With the uniform concentration bound, we can continue bounding the regret of λ-player as follows.

$$\begin{aligned}
&\langle \widehat{\boldsymbol{\lambda}}^* - \boldsymbol{\lambda}_t, \boldsymbol{\theta} + \gamma \boldsymbol{\Psi v}_{\boldsymbol{\zeta}_t, \pi_t} - \boldsymbol{\zeta}_t \rangle \\
&= \langle \widehat{\boldsymbol{\lambda}}^* - \boldsymbol{\lambda}_t, \boldsymbol{\theta} + \gamma \widehat{\boldsymbol{\Psi v}}_{\boldsymbol{\zeta}_t, \pi_t} - \boldsymbol{\zeta}_t \rangle \\
&\quad + \gamma \langle \widehat{\boldsymbol{\lambda}}^* - \boldsymbol{\lambda}_t, \boldsymbol{\Psi v}_{\boldsymbol{\zeta}_t, \pi_t} - \widehat{\boldsymbol{\Psi v}}_{\boldsymbol{\zeta}_t, \pi_t} \rangle.
\end{aligned}$$

The sum of the first term across $t = 1, \dots, T$ can be bounded by employing online convex optimization algorithm. We can bound the second term as follows.

$$\begin{aligned}
&\langle \widehat{\boldsymbol{\lambda}}^* - \boldsymbol{\lambda}_t, \boldsymbol{\Psi v}_{\boldsymbol{\zeta}_t, \pi_t} - \widehat{\boldsymbol{\Psi v}}_{\boldsymbol{\zeta}_t, \pi_t} \rangle \\
&\le \|\widehat{\boldsymbol{\lambda}}^* - \boldsymbol{\lambda}_t\|_{(\widehat{\boldsymbol{\Lambda}}_n + \boldsymbol{I}/n)^{-1}} \|\boldsymbol{\Psi v}_{\boldsymbol{\zeta}_t, \pi_t} - \widehat{\boldsymbol{\Psi v}}_{\boldsymbol{\zeta}_t, \pi_t}\|_{\widehat{\boldsymbol{\Lambda}}_n + \boldsymbol{I}/n} \\
&\le \underbrace{\|\widehat{\boldsymbol{\lambda}}^* - \boldsymbol{\lambda}_t\|_{\widehat{\boldsymbol{\Lambda}}_n^\dagger}}_{(i)} \underbrace{\|\boldsymbol{\Psi v}_{\boldsymbol{\zeta}_t, \pi_t} - \widehat{\boldsymbol{\Psi v}}_{\boldsymbol{\zeta}_t, \pi_t}\|_{\widehat{\boldsymbol{\Lambda}}_n + \boldsymbol{I}/n}}_{(ii)}
\end{aligned}$$

where the second inequality follows since $\widehat{\boldsymbol{\lambda}}^* - \boldsymbol{\lambda}_t$ is in the column space of $\widehat{\boldsymbol{\Lambda}}_n$. (ii) can be bounded using Lemma 3.6. (i) can be bounded by the following technical lemma. See Appendix B.1 for a proof.

Lemma 3.7. *For any $\boldsymbol{\lambda}(\boldsymbol{c}) = \frac{1}{n}\sum_{k=1}^n c_k \boldsymbol{\varphi}(s_k, a_k)$ with $c_k \in [-B, B]$, $k = 1, \dots, n$, we have*

$$\|\boldsymbol{\lambda}(\boldsymbol{c})\|^2_{\widehat{\boldsymbol{\Lambda}}_n^\dagger} \le dB^2.$$

Combining all the bounds, we get the following.

$$\frac{1}{T}\sum_{t=1}^T \mathrm{REG}_t^\lambda \le \widetilde{\mathcal{O}}\left(\frac{C^* d^{3/2}}{1-\gamma}\sqrt{\frac{\log(dnT/\delta)}{n}}\right) + \epsilon_{\text{opt}}^\lambda(T)$$

where $\widetilde{\mathcal{O}}$ hides $\log\log|\mathcal{A}|$. See Appendix D.3 for details.

4. Algorithm and Main Results

Motivated by the analysis in the previous section for bounding regrets of the four players, we present a primal-dual algorithm that proceeds in T steps. At each step, the four players λ-player, ζ-player, w-player, π-player choose actions $\boldsymbol{\lambda}_t$, $\boldsymbol{\zeta}_t$, $\boldsymbol{w}_t$, π_t, respectively. Since the analysis in the previous section requires ζ-player and w-player to act greedily, we choose λ-player and π-player to play $\boldsymbol{\lambda}_t$, π_t, respectively, before ζ-player and w-player play. The regret analysis of the three players in the previous section leads to our main result in the following theorem.

Algorithm 1: Primal-Dual Algorithm for Offline Linear MDPs

Input: Dataset $\mathcal{D} = \{(s_j, a_j, r_j, s_j')\}_{j=1}^n$.
Initialize: π_1 uniform, $\boldsymbol{c}_1' \leftarrow \boldsymbol{0}$, $\alpha \leftarrow \sqrt{\log|\mathcal{A}|/T}$.
for $t = 1, \dots, T$ **do**
  $\boldsymbol{\zeta}_t \leftarrow \operatorname{argmin}_{\boldsymbol{\zeta} \in \mathbb{B}_d(D_\zeta)} \langle \boldsymbol{\zeta}, \boldsymbol{\Phi}^T \widehat{\boldsymbol{\mu}}_{\boldsymbol{\lambda}(\boldsymbol{c}_t'), \pi_t} - \boldsymbol{\lambda}(\boldsymbol{c}_t') \rangle$.
  $\boldsymbol{\lambda}(\boldsymbol{c}_{t+1}) \leftarrow \mathrm{OCO}(\boldsymbol{\theta} - \boldsymbol{\zeta}_t + \gamma \widehat{\boldsymbol{\Psi v}}_{\boldsymbol{\zeta}_t, \pi_t}; \mathcal{C}_n(C^*))$.
  Convert $\boldsymbol{c}_{t+1}$ to $\boldsymbol{c}_{t+1}'$ using Definition 3.3.
  $\pi_{t+1} \leftarrow \sigma(\alpha \sum_{i=1}^t \boldsymbol{\Phi}\boldsymbol{\zeta}_i)$.
Return: $\bar{\pi} = \mathrm{Unif}(\pi_1, \dots, \pi_T)$

Theorem 4.1. *Under Assumptions A and B, as long as T is at least $\Omega(\frac{d \log|\mathcal{A}|}{(1-\gamma)^2\epsilon^2})$, the policy $\bar{\pi}$ produced by Algorithm 1 satisfies $J(\bar{\pi}) \ge J(\pi^*) - \epsilon$ with probability at least $1-\delta$ for sample size*

$$n = \mathcal{O}\left(\frac{(C^*)^2 d^3 \log(dn(\log|\mathcal{A}|)/(\delta\epsilon(1-\gamma)))}{(1-\gamma)^2\epsilon^2}\right).$$

Our work is an improvement over the work by Gabbianelli et al. (2024) who give $\widetilde{\mathcal{O}}(\frac{(C^*)^2 d^2 \log|\mathcal{A}|}{(1-\gamma)^4\epsilon^2})$ sample complexity.

4.1. Result on Feature Coverage Assumptions

The discussion so far uses the concentrability assumption (Assumption B) that requires $\mu^*(s,a)/\mu(s,a) \le C^*$ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$. We present a result that requires a feature coverage assumption instead. We use the definition of feature coverage used in Gabbianelli et al. (2024):

Assumption C (Feature coverage)**.** For an optimal policy π^*, we have

$$(\boldsymbol{\lambda}^*)^T (\boldsymbol{\Lambda}^\dagger)^2 \boldsymbol{\lambda}^* \le C^* \quad \text{and} \quad \boldsymbol{\lambda}^* \in \mathrm{Col}(\boldsymbol{\Lambda})$$

where $\boldsymbol{\lambda}^* := \mathbb{E}_{\mu^*}[\boldsymbol{\varphi}(s,a)]$ and $\mathrm{Col}(\cdot)$ is the column space.

The feature coverage assumption requires $\boldsymbol{\lambda}^*$, the expected occupancy of the target policy in the feature space, to be covered by covariance matrix induced by the data distribution μ_B. Under the feature coverage assumption, $\boldsymbol{\lambda}^*$ can be approximated by a linear combination of $\boldsymbol{\varphi}(s_k, a_k)$, $k = 1, \dots, n$ (Lemma B.3). This result is analogous to Lemma 3.5 that uses concentrability assumption instead. It follows that the result in Theorem 4.1 with the concentrability assumption (Assumption B) replaced by the feature coverage assumption (Assumption C). A limitation of our work is that we use a stronger notion of feature coverage compared to the one used by Gabbianelli et al. (2024), who assume $(\boldsymbol{\lambda}^*)^T(\boldsymbol{\Lambda}^\dagger)\boldsymbol{\lambda}^*$ is bounded. However, they require the knowledge of $\boldsymbol{\Lambda}$ and their sample complexity is $\widetilde{\mathcal{O}}(\epsilon^{-4})$. We leave design and analysis of algorithm using the weaker notion of feature coverage to future work.

5. Extension to Offline Constrained RL

We now consider a *constrained* Markov decision process (CMDP) $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, \{r_i\}_{i=0}^I, \gamma, \nu_0)$. It is the same setting as the MDP setting except that now we have multiple reward functions $r_i, i = 0, \dots, I$. We define the normalized expected cumulative rewards for $r_0, \dots, r_I$:

$$J_i(\pi) := (1-\gamma)\mathbb{E}^\pi \left[\sum_{t=0}^{\infty} \gamma^t r_i(s_t, a_t) \right].$$

Constrained RL aims to find a policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ that maximizes the reward signal r_0 subject to the constraints on other reward signals $r_i, i = 1, \dots, I$. Specifically, given thresholds $\boldsymbol{\tau} = (\tau_1, \dots, \tau_I)$, the goal is to find π that solves the following optimization problem denoted by $\mathcal{P}(\boldsymbol{\tau})$.

$$\begin{aligned} \max_{\pi} \quad & J_0(\pi) \\ \text{subject to} \quad & J_i(\pi) \geq \tau_i, \quad i = 1, \dots, I. \end{aligned} \tag{OPT}$$

We assume the following Slater's condition, a commonly made assumption in constrained RL (Le et al., 2019; Chen et al., 2021; Bai et al., 2023; Ding et al., 2020) for ensuring strong duality of the optimization problem.

Assumption D (Slater's condition)**.** There exist a constant $\phi > 0$ and a policy π such that $J_i(\pi) \geq \tau_i + \phi$ for all $i = 1, \dots, I$. Assume ϕ is known.

As discussed in Hong et al. (2023), Slater's condition is a mild assumption since given the knowledge of the feasibility of the problem, we can guarantee that Slater's condition is met by slightly loosening the cost threshold. For sample efficient learning for arbitrarily large state space, we assume the following linear structure on the CMDP.

Assumption E (Linear CMDP)**.** We assume that the transition and the reward functions can be expressed as a linear function of a *known* feature map $\boldsymbol{\varphi} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ such that

$$r_i(s,a) = \langle \boldsymbol{\varphi}(s,a), \boldsymbol{\theta}_i \rangle, \quad P(s'|s,a) = \langle \boldsymbol{\varphi}(s,a), \boldsymbol{\psi}(s') \rangle$$

for all $(s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$ and $i = 1, \dots, I$, where $\boldsymbol{\theta}_i \in \mathbb{R}^d$ are *known* parameters and $\boldsymbol{\psi} = (\psi_1, \dots, \psi_d)$ is a vector of d *unknown* (signed) measures on $\mathcal{S}$.

Similarly to the linear MDP setting, we require the data coverage assumption (Assumption B) where the optimal policy π^* is optimal for the optimization problem $\mathcal{P}(\boldsymbol{\tau})$.

Our algorithm for the linear CMDP setting is motivated by the linear programming formulation of the constrained reinforcement learning problem (OPT):

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

and its dual

$$\begin{aligned} \min_{\boldsymbol{w} \geq \mathbf{0}, \boldsymbol{v}, \boldsymbol{\zeta}} \quad & (1-\gamma)\langle \boldsymbol{\nu}_0, \boldsymbol{v} \rangle - \langle \boldsymbol{w}, \boldsymbol{\tau} \rangle \\ \text{subject to} \quad & \boldsymbol{\zeta} = \boldsymbol{\theta}_0 + \boldsymbol{\Theta}\boldsymbol{w} + \gamma \boldsymbol{\Psi}\boldsymbol{v} \\ & \boldsymbol{E}\boldsymbol{v} \geq \boldsymbol{\Phi}\boldsymbol{\zeta}. \end{aligned}$$

where we write $\boldsymbol{\Theta} = \begin{bmatrix} \boldsymbol{\theta}_1 & \cdots & \boldsymbol{\theta}_I \end{bmatrix} \in \mathbb{R}^{d \times I}$.

Algorithm 2: Primal-Dual Algorithm for Offline Linear CMDPs

Input: Dataset $\mathcal{D} = \{(s_j, a_j, r_j, s_j')\}_{j=1}^n$, D_w, $\boldsymbol{\tau}$

Initialize: π_1 uniform, $\boldsymbol{c}_1' \leftarrow \mathbf{0}$, $\alpha \leftarrow \sqrt{\log|\mathcal{A}|/T}$.

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

  $\boldsymbol{\zeta}_t \leftarrow \operatorname{argmin}_{\boldsymbol{\zeta} \in \mathbb{B}_d(D_\zeta)} \langle \boldsymbol{\zeta}, \boldsymbol{\Phi}^T \widehat{\boldsymbol{\mu}}_{\boldsymbol{\lambda}(\boldsymbol{c}_t'), \pi_t} - \boldsymbol{\lambda}(\boldsymbol{c}_t') \rangle$.

  $\boldsymbol{w}_t \leftarrow \operatorname{argmin}_{\boldsymbol{w} \in D_w \boldsymbol{\Delta}^I} \langle \boldsymbol{w}, \boldsymbol{\tau} - \boldsymbol{\Theta}^T \boldsymbol{\lambda}_t \rangle$.

  $\boldsymbol{\lambda}(\boldsymbol{c}_{t+1}) \leftarrow$
  $\mathrm{OCO}(\boldsymbol{\theta}_0 - \boldsymbol{\zeta}_t + \boldsymbol{\Theta}\boldsymbol{w}_t + \gamma \widehat{\boldsymbol{\Psi}\boldsymbol{v}}_{\boldsymbol{\zeta}_t, \pi_t}; \mathcal{C}_n(C^*))$.

  Convert $\boldsymbol{c}_{t+1}$ to $\boldsymbol{c}_{t+1}'$ using Definition 3.3.

  $\pi_{t+1} \leftarrow \sigma(\alpha \sum_{i=1}^t \boldsymbol{\Phi}\boldsymbol{\zeta}_i)$.

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

The structure of our algorithm for the constrained RL setting is similar to that for the unconstrained RL setting. The difference is that we add the w-player that adjusts the weights on the rewards $r_1, \dots, r_I$. Closely following the analysis for the unconstrained setting, we can show the following sample complexity for the constrained setting.

Theorem 5.1. *Under Assumptions B,D and E, the policy $\bar{\pi}$ produced by Algorithm 1 with threshold $\boldsymbol{\tau}$ and $D_w = 1 + \frac{1}{\phi}$ and T at least $\Omega(\frac{d \log|\mathcal{A}|}{(1-\gamma)^2 \epsilon^2})$ and large enough such that $\epsilon_{opt}^\lambda(T) \leq \epsilon$ satisfies $J_0(\bar{\pi}) \geq J_0(\pi^*) - \epsilon$ and $J_i(\bar{\pi}) \geq \tau_i - \epsilon$ with probability at least $1 - \delta$ with sample size*

$$n = \mathcal{O}\left(\frac{(C^*)^2 d^3 \log(dn(\log|\mathcal{A}|)/(\delta\phi\epsilon(1-\gamma)))}{(1-\gamma)^2 \phi^2 \epsilon^2} \right).$$

See Appendix E for details. By tightening the input thresholds for Algorithm 1 to $\boldsymbol{\tau} + \phi\epsilon\mathbf{1}$ and assuming a two-policy feature coverage assumption, we can show that the output policy $\bar{\pi}$ is ϵ-optimal and satisfies the constraints exactly, i.e., $J_i(\bar{\pi}) \geq \tau_i$, $i = 1, \dots, I$. See Appendix E.4 for details.

6. Conclusion

In this paper, we propose a computationally efficient primal dual algorithm for offline constrained RL with linear function approximation under partial data coverage. Our algorithm is the first computationally efficient algorithm to achieve $\mathcal{O}(\epsilon^{-2})$ sample complexity under partial data coverage. For the partial data coverage assumption, we use the notion of feature coverage. An interesting future work would be to design an algorithm that allows using a weaker notion of feature coverage in the sample complexity bound.

Acknowledgement

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

Impact Statement

This paper presents work whose goal is to advance the field of Reinforcement Learning Theory. There are potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

Altman, E. *Constrained Markov decision processes*. Routledge, 2021.

Antos, A., Szepesvári, C., and Munos, R. Fitted q-iteration in continuous action-space mdps. *Advances in neural information processing systems*, 20, 2007.

Awerbuch, B. and Kleinberg, R. Online linear optimization and adaptive routing. *Journal of Computer and System Sciences*, 74(1):97–114, 2008.

Bai, Q., Bedi, A. S., Agarwal, M., Koppel, A., and Aggarwal, V. Achieving zero constraint violation for concave utility constrained reinforcement learning via primal-dual approach. *Journal of Artificial Intelligence Research*, 78: 975–1016, 2023.

Brunke, L., Greeff, M., Hall, A. W., Yuan, Z., Zhou, S., Panerati, J., and Schoellig, A. P. Safe learning in robotics: From learning-based control to safe reinforcement learning. *Annual Review of Control, Robotics, and Autonomous Systems*, 5:411–444, 2022.

Chen, J. and Jiang, N. Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning*, pp. 1042–1051. PMLR, 2019.

Chen, J. and Jiang, N. Offline reinforcement learning under value and density-ratio realizability: the power of gaps. In *Uncertainty in Artificial Intelligence*, pp. 378–388. PMLR, 2022.

Chen, Y., Dong, J., and Wang, Z. A primal-dual approach to constrained markov decision processes. *arXiv preprint arXiv:2101.10895*, 2021.

Ding, D., Zhang, K., Basar, T., and Jovanovic, M. Natural policy gradient primal-dual method for constrained markov decision processes. *Advances in Neural Information Processing Systems*, 33:8378–8390, 2020.

Gabbianelli, G., Neu, G., Papini, M., and Okolo, N. M. Offline primal-dual reinforcement learning for linear mdps. In *International Conference on Artificial Intelligence and Statistics*, pp. 3169–3177. PMLR, 2024.

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

Hong, K., Li, Y., and Tewari, A. A primal-dual-critic algorithm for offline constrained reinforcement learning. *arXiv preprint arXiv:2306.07818*, 2023.

Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pp. 2137–2143. PMLR, 2020.

Jin, Y., Yang, Z., and Wang, Z. Is pessimism provably efficient for offline rl? In *International Conference on Machine Learning*, pp. 5084–5096. PMLR, 2021.

Kumar, A., Singh, A., Tian, S., Finn, C., and Levine, S. A workflow for offline model-free robotic reinforcement learning. *arXiv preprint arXiv:2109.10813*, 2021.

Le, H., Voloshin, C., and Yue, Y. Batch policy learning under constraints. In *International Conference on Machine Learning*, pp. 3703–3712. PMLR, 2019.

Levine, S., Pastor, P., Krizhevsky, A., Ibarz, J., and Quillen, D. Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. *The International journal of robotics research*, 37(4-5):421–436, 2018.

Levine, S., Kumar, A., Tucker, G., and Fu, J. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

Munos, R. Error bounds for approximate policy iteration. In *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, pp. 560–567, 2003.

Munos, R. Error bounds for approximate value iteration. In *Proceedings of the National Conference on Artificial Intelligence*, volume 20, pp. 1006. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2005.

Munos, R. and Szepesvári, C. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9 (5), 2008.

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

Tang, S. and Wiens, J. Model selection for offline reinforcement learning: Practical considerations for healthcare settings. In *Machine Learning for Healthcare Conference*, pp. 2–35. PMLR, 2021.

Uehara, M. and Sun, W. Pessimistic model-based offline reinforcement learning under partial coverage. In *International Conference on Learning Representations*, 2022.

Wagenmaker, A. J., Chen, Y., Simchowitz, M., Du, S., and Jamieson, K. First-order regret in reinforcement learning with linear function approximation: A robust estimation approach. In *International Conference on Machine Learning*, pp. 22384–22429. PMLR, 2022.

Wainwright, M. J. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

Wang, W., Yu, N., Gao, Y., and Shi, J. Safe off-policy deep reinforcement learning algorithm for volt-var control in power distribution systems. *IEEE Transactions on Smart Grid*, 11(4):3008–3018, 2019.

Wei, C.-Y., Jahromi, M. J., Luo, H., and Jain, R. Learning infinite-horizon average-reward mdps with linear function approximation. In *International Conference on Artificial Intelligence and Statistics*, pp. 3007–3015. PMLR, 2021.

Wu, R., Zhang, Y., Yang, Z., and Wang, Z. Offline constrained multi-objective reinforcement learning via pessimistic dual value iteration. *Advances in Neural Information Processing Systems*, 34:25439–25451, 2021.

Xie, T. and Jiang, N. Q* approximation schemes for batch reinforcement learning: A theoretical comparison. In *Conference on Uncertainty in Artificial Intelligence*, pp. 550–559. PMLR, 2020.

Xie, T., Cheng, C.-A., Jiang, N., Mineiro, P., and Agarwal, A. Bellman-consistent pessimism for offline reinforcement learning. *Advances in neural information processing systems*, 34:6683–6694, 2021.

Zanette, A. When is realizability sufficient for off-policy reinforcement learning? In *International Conference on Machine Learning*, pp. 40637–40668. PMLR, 2023.

Zanette, A., Wainwright, M. J., and Brunskill, E. Provable benefits of actor-critic methods for offline reinforcement learning. *Advances in neural information processing systems*, 34:13626–13640, 2021.

Zhan, W., Huang, B., Huang, A., Jiang, N., and Lee, J. Offline reinforcement learning with realizability and single-policy concentrability. In *Conference on Learning Theory*, pp. 2730–2775. PMLR, 2022.

Zhu, H., Rashidinejad, P., and Jiao, J. Importance weighted actor-critic for optimal conservative offline reinforcement learning. *arXiv preprint arXiv:2301.12714*, 2023.

A. Covering

Lemma A.1 (Covering balls. e.g. Wainwright (2019))**.** *For any $\epsilon \in (0,1)$, we have*

$$\log \mathcal{N}(\mathbb{B}_d(r), \|\cdot\|_\infty, \epsilon) \leq d \log\left(1 + \frac{2r}{\epsilon}\right).$$

Lemma A.2 (Lemma 7 in Zanette et al. (2021))**.** *Consider a feature mapping $\boldsymbol{\varphi} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ such that $\|\boldsymbol{\varphi}(s,a)\|_2 \leq 1$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$. Then for all $s \in \mathcal{S}$, we have*

$$\sum_{a \in \mathcal{A}} |\pi_{\boldsymbol{\theta}'}(a|s) - \pi_{\boldsymbol{\theta}}(a|s)| \leq 8\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2$$

for any pair $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbb{R}^d$ such that $\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2 \leq \frac{1}{2}$.

Lemma A.3 (Covering softmax function class. Lemma 6 in Zanette et al. (2021))**.** *For any $\epsilon \in (0,1)$, we have*

$$\log \mathcal{N}(\Pi(B), \|\cdot\|_{\infty,1}, \epsilon) \leq d \log\left(1 + \frac{16B}{\epsilon}\right)$$

where the norm $\|\cdot\|_{\infty,1}$ is defined by

$$\|\pi - \pi'\|_{\infty,1} := \sup_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} |\pi(a|s) - \pi'(a|s)|.$$

Lemma A.4 (Covering number bound for the space of v)**.** *Consider the function class*

$$\mathcal{V} = \{v_{\boldsymbol{\zeta},\pi} : \boldsymbol{\zeta} \in \mathbb{B}(D_\zeta), \pi \in \Pi(D_\pi)\}$$

where $v_{\boldsymbol{\zeta},\pi} : \mathcal{S} \to \mathbb{R}$ is defined by $v_{\boldsymbol{\zeta},\pi}(s) = \sum_a \pi(a|s)\langle \boldsymbol{\zeta}, \varphi(s,a)\rangle$. Then,

$$\mathcal{N}(\mathcal{V}, \|\cdot\|_\infty, \epsilon) \leq \mathcal{N}(\mathbb{B}(D_\zeta), \|\cdot\|_2, \epsilon/2) \times \mathcal{N}(\Pi(D_\pi), \|\cdot\|_{\infty,1}, \epsilon/(2D_\zeta)).$$

and it follows that

$$\log \mathcal{N}(\mathcal{V}, \|\cdot\|_\infty, \epsilon) \leq \mathcal{O}(d \log(D_\zeta D_\pi/\epsilon)).$$

Proof. Consider $\mathcal{C}_v = \{v_{\boldsymbol{\zeta},\pi} \in (\mathcal{S} \to [0, \frac{1}{1-\gamma}]) : \boldsymbol{\zeta} \in \mathcal{C}_\zeta, \pi \in \mathcal{C}_\pi\}$ where $\mathcal{C}_\zeta$ is an $\epsilon/2$-cover of $\mathbb{B}(D_\zeta)$ with respect to $\|\cdot\|_2$ and $\mathcal{C}_\pi$ is an $\epsilon/(2D_\zeta)$-cover of $\Pi(D_\pi)$ with respect to $\|\cdot\|_{\infty,1}$. Such covers with $|\mathcal{C}_\zeta| \leq (1 + 4D_\zeta/\epsilon)^d$ and $|\mathcal{C}_\pi| \leq (1 + 32D_\zeta D_\pi/\epsilon)^d$ exist by previous lemmas. Consider any $v_{\boldsymbol{\zeta},\pi} \in \mathcal{V}$. Then, there exists $\boldsymbol{\zeta}' \in \mathcal{C}_\zeta$ and $\pi' \in \mathcal{C}_\pi$ with $\|\boldsymbol{\zeta} - \boldsymbol{\zeta}'\|_2 \leq \epsilon/2$ and $\|\pi - \pi'\|_{\infty,1} = \sup_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} |\pi(a|s) - \pi'(a|s)| \leq \epsilon/(2D_\zeta)$. Then for any $s \in \mathcal{S}$, $v_{\boldsymbol{\zeta}',\pi'} \in \mathcal{C}_v$ satisfies

$$\begin{aligned}
|v_{\boldsymbol{\zeta},\pi}(s) - v_{\boldsymbol{\zeta}',\pi'}(s)| &= |\sum_a \pi(a|s)\langle \boldsymbol{\zeta}, \varphi(s,a)\rangle - \sum_a \pi'(a|s)\langle \boldsymbol{\zeta}', \varphi(s,a)\rangle| \\
&= |\sum_a (\pi(a|s) - \pi'(a|s))\langle \boldsymbol{\zeta}, \varphi(s,a)\rangle + \pi'(a|s)\langle \boldsymbol{\zeta} - \boldsymbol{\zeta}', \varphi(s,a)\rangle| \\
&\leq D_\zeta \sum_a |\pi(a|s) - \pi'(a|s)| + \sum_a \pi'(a|s)\epsilon/2 \\
&\leq \epsilon.
\end{aligned}$$

It follows that $\mathcal{C}_v$ is an ϵ-cover of $\mathcal{V}$ with respect to $\|\cdot\|_\infty$ with $|\mathcal{C}_v| = |\mathcal{C}_\zeta||\mathcal{C}_\pi|$ and we are done. □

B. Concentration Inequalities

Lemma B.1 (Matrix Bernstein)**.** *Consider a finite sequence $\{S_k\}$ of independent, random matrices with common dimension $d_1 \times d_2$. Assume that $\mathbb{E}S_k = 0$ and $\|S_k\| \leq L$ for each index k. Let $Z = \sum_k S_k$ and define $v(Z) := \max\{\|\mathbb{E}[ZZ^T]\|, \|\mathbb{E}[Z^TZ]\|\} = \max\{\|\sum_k \mathbb{E}[S_kS_k^T]\|, \|\sum_k \mathbb{E}[S_k^TS_k]\|\}$. Then,*

$$P(\|Z\| \geq t) \leq (d_1 + d_2) \exp\left(\frac{-t^2/2}{v(Z) + Lt/3}\right)$$

and it follows that with probability at least $1-\delta$*, we have*

$$\|Z\| \leq \frac{2L\log((d_1+d_2)/\delta)}{3} + \sqrt{2v(Z)\log((d_1+d_2)/\delta)}.$$

Lemma B.2. *There exists an example where* $\boldsymbol{\lambda}^* = \mathbb{E}_{\mu^*}[\boldsymbol{\varphi}(s,a)]$ *is not in the span of* $\boldsymbol{\varphi}(s_1,a_1),\ldots,\boldsymbol{\varphi}(s_n,a_n)$ *with probability at least* $1/2$*.*

Proof. Consider the case where $\mathcal{S}=\{s\}$, $\mathcal{A}=\{a_1,a_2\}$, $d=2$, and $\boldsymbol{\varphi}(s,a_1)=\boldsymbol{e}_1$ and $\boldsymbol{\varphi}(s,a_2)=\boldsymbol{e}_2$. Let $\mu=\mu^*$ and $\mu(s,a_1)=p$ and $\mu(s,a_2)=1-p$. Let F be the event where $\boldsymbol{\lambda}^*$ is not in the span of $\boldsymbol{\varphi}(s_1,a_1),\ldots,\boldsymbol{\varphi}(s_n,a_n)$. Then,

$$P(F) = (1-p)^n \geq \frac{1}{2}$$

as long as we choose $p \leq 1-2^{-1/n}$. □

Proof of Lemma 3.5. Let $c_k = w^*(s_k,a_k) = \mu^*(s_k,a_k)/\mu(s_k,a_k)$, $k=1,\ldots,n$. Note that $\mu(s_k,a_k)>0$, $k=1,\ldots,n$ must hold, otherwise such s_k,a_k cannot be sampled. By the concentrability assumption, we have $c_k\in[0,C^*]$, $k=1,\ldots,n$. Let $\boldsymbol{z}_k = c_k\boldsymbol{\varphi}(s_k,a_k)$ for $k=1,\ldots,n$. Then, $\|\boldsymbol{z}_k\|\leq C^*$ and

$$\mathbb{E}[\boldsymbol{z}_k] = \mathbb{E}_{(s,a)\sim\mu}[w^*(s,a)\boldsymbol{\varphi}(s,a)] = \mathbb{E}_{(s,a)\sim\mu}\left[\frac{\mu^*(s,a)}{\mu(s,a)}\boldsymbol{\varphi}(s,a)\right] = \mathbb{E}_{(s,a)\sim\mu^*}[\boldsymbol{\varphi}(s,a)] = \boldsymbol{\lambda}^*.$$

Define $\boldsymbol{S}_k = \boldsymbol{z}_k - \boldsymbol{\lambda}^*$, $k=1,\ldots,n$ Then, $\mathbb{E}[\boldsymbol{S}_k]=0$ and $\|\boldsymbol{S}_k\|_2 \leq \|\boldsymbol{z}_k\|_2 + \mathbb{E}[\|\boldsymbol{z}_k\|_2] \leq 2C^*$ and $\|\mathbb{E}[\boldsymbol{S}_k^T\boldsymbol{S}_k]\|_2 \leq \mathbb{E}[\boldsymbol{z}_k^T\boldsymbol{z}_k] \leq (C^*)^2$ and $\|\mathbb{E}[\boldsymbol{S}_k\boldsymbol{S}_k^T]\|_2 \leq (C^*)^2$. Applying matrix Bernstein inequality (Lemma B.1) on $\{\boldsymbol{S}_k\}_{k=1}^n$, we have

$$\|\frac{1}{n}\sum_{k=1}^n \boldsymbol{S}_k\|_2 = \|\frac{1}{n}w^*(s_k,a_k)\boldsymbol{\varphi}(s_k,a_k) - \boldsymbol{\lambda}^*\|_2 \leq \frac{4C^*\log((d+1)/\delta)}{3n} + \sqrt{\frac{8(C^*)^2\log((d+1)/\delta)}{n}}$$

with probability at least $1-\delta$ and the result follows. □

Lemma B.3. *Under the feature coverage assumption C, there exists* $\widehat{\boldsymbol{\lambda}}^*\in\mathbb{R}^d$ *of the form* $\widehat{\boldsymbol{\lambda}}^* = \frac{1}{n}\sum_{k=1}^n c_k\boldsymbol{\varphi}(s_k,a_k)$ *with* $c_k\in[0,C^*]$, $k=1,\ldots,n$ *such that*

$$\|\widehat{\boldsymbol{\lambda}}^* - \boldsymbol{\lambda}^*\|_2 \leq \mathcal{O}\left(C^*\sqrt{\frac{\log(d/\delta)}{n}}\right)$$

with probability at least $1-\delta$*.*

Proof. Since $\boldsymbol{\lambda}^*\in\mathrm{Col}(\boldsymbol{\Lambda})$, we have $\boldsymbol{\lambda}^* = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^\dagger\boldsymbol{\lambda}^*$ and it follows that

$$\begin{aligned}\boldsymbol{\lambda}^* &= \boldsymbol{\Lambda}\boldsymbol{\Lambda}^\dagger\boldsymbol{\lambda}^* \\ &= \mathbb{E}[\boldsymbol{\varphi}(s,a)\boldsymbol{\varphi}(s,a)^T\boldsymbol{\Lambda}^\dagger\boldsymbol{\lambda}^*] \\ &= \mathbb{E}[\langle\boldsymbol{\varphi}(s,a),\boldsymbol{\Lambda}^\dagger\boldsymbol{\lambda}^*\rangle\boldsymbol{\varphi}(s,a)].\end{aligned}$$

Let $c_k = \langle\boldsymbol{\varphi}(s_k,a_k),\boldsymbol{\Lambda}^\dagger\boldsymbol{\lambda}^*\rangle$. Then, by Cauchy-Schwartz, we have

$$|c_k| \leq \|\boldsymbol{\varphi}(s_k,a_k)\|_2\|\boldsymbol{\Lambda}^\dagger\boldsymbol{\lambda}^*\|_2 \leq C^*$$

where the last inequality follows by the feature coverage assumption. Using matrix Bernstein inequality as is done in the proof of Lemma 3.5, the result follows. □

B.1. Proof of Lemma 3.7

Proof of Lemma 3.7. Consider $\boldsymbol{\lambda} = \frac{1}{n}\sum_{k=1}^{n} c_k \boldsymbol{\varphi}(s_k, a_k)$ with $|c_k| \leq B$, $k = 1, \ldots, n$. Let $\widehat{\boldsymbol{\Lambda}}_n = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^T$ be the eigendecomposition of $\widehat{\boldsymbol{\Lambda}}_n = \frac{1}{n}\sum_{k=1}^{n} \boldsymbol{\varphi}(s_k, a_k)\boldsymbol{\varphi}(s_k, a_k)^T$ where $\boldsymbol{D} = \text{diag}(d_1, \ldots, d_d)$ with $d_1 \geq \cdots \geq d_d \geq 0$. Then, we have $\boldsymbol{D} = \boldsymbol{U}^T\widehat{\boldsymbol{\Lambda}}_n\boldsymbol{U} = \frac{1}{n}\sum_{k=1}^{n} \boldsymbol{U}^T\boldsymbol{\varphi}(s_k, a_k)\boldsymbol{\varphi}(s_k, a_k)^T\boldsymbol{U}$ and it follows that $d_i = \frac{1}{n}\sum_{k=1}^{n}\langle \boldsymbol{u}_i, \boldsymbol{\varphi}(s_k, a_k)\rangle^2$ where d_i is the ith diagonal entry of $\boldsymbol{D}$. Also,

$$
\begin{aligned}
\boldsymbol{\lambda} &= \frac{1}{n}\sum_{k=1}^{n} c_k \boldsymbol{\varphi}(s_k, a_k) \\
&= \frac{1}{n}\sum_{k=1}^{n} c_k \sum_{i=1}^{d} \langle \boldsymbol{\varphi}(s_k, a_k), \boldsymbol{u}_i\rangle \boldsymbol{u}_i \\
&= \sum_{i=1}^{d}\left(\frac{1}{n}\sum_{k=1}^{n} c_k \langle \boldsymbol{\varphi}(s_k, a_k), \boldsymbol{u}_i\rangle\right)\boldsymbol{u}_i.
\end{aligned}
$$

where the second equality follows by $\boldsymbol{x} = \boldsymbol{U}\boldsymbol{U}^T\boldsymbol{x} = \sum_{i=1}^{d}\langle \boldsymbol{x}, \boldsymbol{u}_i\rangle \boldsymbol{u}_i$ for any vector $\boldsymbol{x} \in \mathbb{R}^d$. So,

$$
\begin{aligned}
\boldsymbol{\lambda}^T\widehat{\boldsymbol{\Lambda}}_n^{\dagger}\boldsymbol{\lambda} &= \sum_{i=1}^{d'}\left(\frac{1}{n}\sum_{k=1}^{n} c_k \langle \boldsymbol{\varphi}(s_k, a_k), \boldsymbol{u}_i\rangle\right)^2 / d_i \\
&= \frac{1}{n}\sum_{i=1}^{d'}\left(\sum_{k=1}^{n} c_k \langle \boldsymbol{\varphi}(s_k, a_k), \boldsymbol{u}_i\rangle\right)^2 \Bigg/ \left(\sum_{k=1}^{n}\langle \boldsymbol{\varphi}(s_k, a_k), \boldsymbol{u}_i\rangle^2\right) \\
&\leq \frac{1}{n}\sum_{i=1}^{d'}\left(\sum_{k=1}^{n} c_k^2\right) \\
&\leq dB^2
\end{aligned}
$$

where d' is the number of strictly positive diagonal entries in $\boldsymbol{D}$ and the first inequality follows by Cauchy-Schwartz. □

B.2. Proof of Lemma 3.6

Lemma B.4. *Let $v : \mathcal{S} \to [0, D_v]$. With probability at least $1 - \delta$, we have*

$$
\|\boldsymbol{\Psi}\boldsymbol{v} - \widehat{\boldsymbol{\Psi}\boldsymbol{v}}\|_{n\widehat{\boldsymbol{\Lambda}}_n + \boldsymbol{I}} \leq \mathcal{O}\left(D_v\sqrt{d\log(n/\delta)}\right)
$$

where $\widehat{\boldsymbol{\Psi}\boldsymbol{v}}$ is the least squares estimate defined in (9).

Proof. Note that $\|\boldsymbol{\Psi}\boldsymbol{v}\|_2 \leq D_v\sqrt{d}$ by the boundedness assumption on $\boldsymbol{\Psi}$. The result follows directly from Theorem 2 in Abbasi-Yadkori et al. (2011). □

Proof of Lemma 3.6. Let $\mathcal{C}$ be an $(1/n)$-cover on $\mathcal{V}$. By Lemma A.4, such a cover with $\log|\mathcal{C}| \leq \mathcal{O}(d\log(D_\zeta D_\pi n))$ exists. Applying a union bound over $\mathcal{C}$ and using the concentration bound in Lemma B.4, we get

$$
\left\|\boldsymbol{\Psi}\boldsymbol{v} - \widehat{\boldsymbol{\Psi}\boldsymbol{v}}\right\|_{n\widehat{\boldsymbol{\Lambda}}_n + \boldsymbol{I}} \leq \mathcal{O}\left(D_v\sqrt{d\log(D_\zeta D_\pi n/\delta)}\right) \tag{10}
$$

for all $\boldsymbol{v} \in \mathcal{C}$ with probability at least $1 - \delta$. For any $\boldsymbol{v} \in \mathcal{V}$, we can find v' in the cover that satisfies $\|\boldsymbol{v} - \boldsymbol{v}'\|_\infty \leq 1/n$. Hence,

$$
\begin{aligned}
\left\|\boldsymbol{\Psi}\boldsymbol{v} - \widehat{\boldsymbol{\Psi}\boldsymbol{v}}\right\|_{n\widehat{\boldsymbol{\Lambda}}_n + \boldsymbol{I}} &\leq \|\boldsymbol{\Psi}(\boldsymbol{v} - \boldsymbol{v}') + \boldsymbol{\Psi}\boldsymbol{v}' - \widehat{\boldsymbol{\Psi}\boldsymbol{v}'} + \widehat{\boldsymbol{\Psi}\boldsymbol{v}'} - \widehat{\boldsymbol{\Psi}\boldsymbol{v}}\|_{n\widehat{\boldsymbol{\Lambda}}_n + \boldsymbol{I}} \\
&\leq \|\boldsymbol{\Psi}(\boldsymbol{v} - \boldsymbol{v}')\|_{n\widehat{\boldsymbol{\Lambda}}_n + \boldsymbol{I}} + \left\|(n\widehat{\boldsymbol{\Lambda}}_n + \boldsymbol{I})^{-1}\sum_{k=1}^{n}(v'(s_k') - v(s_k'))\boldsymbol{\varphi}(s_k, a_k)\right\|_{n\widehat{\boldsymbol{\Lambda}}_n + \boldsymbol{I}} \\
&\quad + \mathcal{O}\left(D_v\sqrt{d\log(D_\zeta D_\pi n/\delta)}\right).
\end{aligned}
$$

The first term can be bounded using the boundedness assumption on $\boldsymbol{\Psi}$ by

$$\|\boldsymbol{\Psi}(\boldsymbol{v}-\boldsymbol{v}')\|^2_{n\widehat{\boldsymbol{\Lambda}}_n+\boldsymbol{I}} \leq \|\boldsymbol{\Psi}(\boldsymbol{v}-\boldsymbol{v}')\|_2^2 \|n\widehat{\boldsymbol{\Lambda}}_n+\boldsymbol{I}\|_2 \leq d/n^2 \cdot (1+n) \leq \mathcal{O}(1)$$

as long as $n \geq d$. The second term can be bounded by

$$\begin{aligned}
&\left\|(n\widehat{\boldsymbol{\Lambda}}+\boldsymbol{I})^{-1}\sum_{k=1}^n (v'(s_k')-v(s_k'))\boldsymbol{\varphi}(s_k,a_k)\right\|^2_{n\widehat{\boldsymbol{\Lambda}}+\boldsymbol{I}} \\
&\quad= \sum_{k=1}^n (v'(s_k')-v(s_k'))\boldsymbol{\varphi}(s_k,a_k)^T(n\widehat{\boldsymbol{\Lambda}}+\boldsymbol{I})^{-1}\sum_{k=1}^n (v'(s_k')-v(s_k'))\boldsymbol{\varphi}(s_k,a_k) \\
&\quad\leq \sum_{k=1}^n \|(v'(s_k')-v(s_k'))\boldsymbol{\varphi}(s_k,a_k)\|_2^2 \\
&\quad\leq \sum_{k=1}^n (v'(s_k')-v(s_k'))^2 \\
&\quad\leq 1
\end{aligned}$$

where the first inequality uses $n\widehat{\boldsymbol{\Lambda}}+\boldsymbol{I} \succcurlyeq \boldsymbol{I}$. The result follows. □

C. Computational Efficiency

In this section, we explain why our algorithms are computationally efficient by showing that the algorithms only require computing quantities for states that appear in the offline dataset to compute the policy π_t at each step. This is how we avoid computation complexity that scales with the size of the state space.

Recall that $\pi_t = \sigma(\alpha \sum_{i=1}^{t-1} \boldsymbol{\Phi}\boldsymbol{\zeta}_i)$ and by definition of $\sigma(\cdot)$,

$$\pi_t(a|s) = \frac{\exp(\alpha \sum_{i=1}^{t-1} \boldsymbol{\varphi}(s,a)^T\boldsymbol{\zeta}_i)}{\sum_{a'} \exp(\alpha \sum_{i=1}^{t-1} \boldsymbol{\varphi}(s,a')^T\boldsymbol{\zeta}_i)}.$$

We argue that the algorithm only needs to compute $\pi_t(a|s)$ for the states s that appear as the next state in the dataset $\mathcal{D}$. There are two parts where the object π_t is used in the algorithm:

Line 3 in Algorithm 1 and Line 4 in Algorithm 2 In these lines, the object π_t is used to compute

$$\widehat{\boldsymbol{\Psi}\boldsymbol{v}}_{\boldsymbol{\zeta}_t,\pi_t} = (n\widehat{\Lambda}+I)^{-1}\sum_{k=1}^n v_{\boldsymbol{\zeta}_t,\pi_t}(s_k')\boldsymbol{\varphi}(s_k,a_k)$$

where $\boldsymbol{v}_{\boldsymbol{\zeta}_t,\pi_t}(s_k') = (n\widehat{\Lambda}+I)^{-1}\sum_a \pi_t(a|s_k')\langle \boldsymbol{\zeta}_t, \boldsymbol{\varphi}(s_k',a)\rangle$. As we claimed, we only need to compute $\pi_t(\cdot|s_k')$ for s_k' that appear in the dataset $\mathcal{D}$.

Line 2 in Algorithm 1 and Line 2 in Algorithm 2 In this lines, the object π_t is used to compute $\boldsymbol{\Phi}^T\widehat{\boldsymbol{\mu}}_{\boldsymbol{\lambda}_t,\pi_t}$ for $\boldsymbol{\lambda}_t$ of the form $\boldsymbol{\lambda}_t = \frac{1}{n}\sum_{k=1}^n c_k\boldsymbol{\varphi}(s_k,a_k)$. By definition,

$$\widehat{\mu}_{\lambda,\pi} = \pi \circ E[(1-\gamma)\nu_0 + \gamma\widehat{\Psi\lambda} = \pi \circ E[(1-\gamma)e_{s_0} + \gamma\frac{1}{n}\sum_{k=1}^n c_k e_{s_k'}]$$

and it follows that

$$\begin{aligned}
\boldsymbol{\Phi}^T\widehat{\boldsymbol{\mu}}_{\boldsymbol{\lambda}_t,\pi_t} &= (1-\gamma)\boldsymbol{\Phi}^T(\pi_t \circ \boldsymbol{E}\boldsymbol{e}_{s_0}) + \gamma\frac{1}{n}\sum_{k=1}^n c_k\boldsymbol{\Phi}^T(\pi_t \circ \boldsymbol{E}\boldsymbol{e}_{s_k'}) \\
&= (1-\gamma)\sum_a \pi_t(s_0,a)\boldsymbol{\varphi}(s_0,a) + \gamma\frac{1}{n}\sum_{k=1}^n c_k\sum_a \pi_t(a|s_k')\boldsymbol{\varphi}(s_k',a).
\end{aligned}$$

Again, we only need to compute $\pi_t(\cdot|s_k')$ for s_k' that appears in $\mathcal{D}$.

D. Details in Offline Unconstrained RL Setting

D.1. Bounding the Regret of π-Player

Lemma D.1 (Mirror Descent, Lemma D.2 in Gabbianelli et al. (2024))**.** *Let $q_1, \dots, q_T$ be a sequence of functions from $\mathcal{S} \times \mathcal{A}$ to $\mathbb{R}$ with $\|q_t\|_\infty \le D_q$ for $t = 1, \dots, T$. Given an initial policy $\pi_1 : \mathcal{S} \to \Delta(\mathcal{A})$ and a learning rate $\alpha > 0$, define the sequence of policies $\pi_2, \dots, \pi_{T+1}$ such that*

$$\pi_{t+1}(a|s) \propto \pi_t(a|s) \exp(\alpha q_t(s,a)).$$

Then, for any comparator policy π^*, we have

$$\sum_{t=1}^{T} \sum_{s \in \mathcal{S}} \nu^{\pi^*}(s) \langle \pi^*(\cdot|s) - \pi_t(\cdot|s), q_t(s,\cdot) \rangle \le \frac{\mathcal{H}(\pi^* \| \pi_1)}{\alpha} + \frac{\alpha T D_q^2}{2}$$

where $\mathcal{H}(\pi \| \pi') \coloneqq \sum_{s \in \mathcal{S}} \nu^{\pi}(s) \mathcal{D}(\pi(\cdot|s) \| \pi'(\cdot|s))$ is the conditional entropy.

Lemma D.2 (Lemma B.3 in Gabbianelli et al. (2024))**.** *The sequence of policies $\pi_1, \dots, \pi_T$ produced by an exponentiation algorithm $\pi_{t+1} = \sigma(\alpha \sum_{i=1}^{t} \boldsymbol{\Phi} \boldsymbol{\zeta}_i)$ satisfies*

$$\sum_{t=1}^{T} \sum_{s \in \mathcal{S}} \nu^{\pi^*}(s) \sum_{a \in \mathcal{A}} (\pi^*(a|s) - \pi_t(a|s)) \langle \boldsymbol{\zeta}_t, \boldsymbol{\varphi}(s,a) \rangle \le \frac{\log |\mathcal{A}|}{\alpha} + \frac{\alpha T D_\varphi^2 D_\zeta^2}{2}$$

where $\|\boldsymbol{\zeta}_t\|_2 \le D_\zeta$, $t = 1, \dots, T$ and $\|\boldsymbol{\varphi}(\cdot,\cdot)\|_2 \le D_\varphi$.

D.2. Bounding the regret of ζ-player

Proof of Lemma 3.1. Recall $\mu_{\boldsymbol{\lambda}(\boldsymbol{c}),\pi}(s,a) = \pi(a|s) \left[(1-\gamma)\nu_0(s) + \gamma(\boldsymbol{\Psi}^T \boldsymbol{\lambda}(\boldsymbol{c}))_s \right]$ where we use the notation $(\boldsymbol{x})_s$ to denote the sth entry of vector $\boldsymbol{x}$. We can write

$$\begin{aligned}
\boldsymbol{\Phi}^T \boldsymbol{\mu}_{\boldsymbol{\lambda}(\boldsymbol{c}),\pi} &= \sum_{s,a} \mu_{\boldsymbol{\lambda}(\boldsymbol{c}),\pi}(s,a) \boldsymbol{\varphi}(s,a) \\
&= \sum_{s,a} \pi(a|s) [(1-\gamma)\nu_0(s) + \gamma(\boldsymbol{\Psi}^T \boldsymbol{\lambda}(\boldsymbol{c}))_s] \boldsymbol{\varphi}(s,a) \\
&= (1-\gamma) \sum_a \pi(a|s_0) \boldsymbol{\varphi}(s_0,a) + \gamma \sum_s (\boldsymbol{\Psi}^T \boldsymbol{\lambda}(\boldsymbol{c}))_s \sum_a \pi(a|s) \boldsymbol{\varphi}(s,a) \\
&= (1-\gamma) \boldsymbol{\varphi}(s_0,\pi) + \gamma \sum_s (\boldsymbol{\Psi}^T \boldsymbol{\lambda}(\boldsymbol{c}))_s \boldsymbol{\varphi}(s,\pi)
\end{aligned}$$

where we use the notation $\boldsymbol{\varphi}(s,\pi) = \sum_a \pi(a|s) \boldsymbol{\varphi}(s,a)$. Recall that $\boldsymbol{\lambda}(\boldsymbol{c}) = \frac{1}{n} \sum_{k=1}^{n} c_k \boldsymbol{\varphi}(s_k, a_k)$ where $c_k \in [-B, B]$, $k = 1, \dots, n$. Following the same argument for expanding $\boldsymbol{\Phi}^T \boldsymbol{\mu}_{\boldsymbol{\lambda}(\boldsymbol{c}),\pi}$, we get

$$\begin{aligned}
\boldsymbol{\Phi}^T \widehat{\boldsymbol{\mu}}_{\boldsymbol{\lambda}(\boldsymbol{c}),\pi} &= (1-\gamma) \boldsymbol{\varphi}(s_0,\pi) + \gamma \sum_s (\widehat{\boldsymbol{\Psi}^T \boldsymbol{\lambda}}(\boldsymbol{c}))_s \boldsymbol{\varphi}(s,\pi) \\
&= (1-\gamma) \boldsymbol{\varphi}(s_0,\pi) + \frac{\gamma}{n} \sum_{k=1}^{n} c_k \boldsymbol{\varphi}(s'_k, \pi).
\end{aligned}$$

Also, using $\boldsymbol{\Psi}^T \boldsymbol{\lambda}(\boldsymbol{c}) = \frac{1}{n} \sum_{k=1}^{n} c_k \boldsymbol{\Psi}^T \boldsymbol{\varphi}(s_k, a_k) = \frac{1}{n} \sum_{k=1}^{n} c_k P(\cdot|s_k, a_k) = \frac{1}{n} \sum_{k=1}^{n} c_k \mathbb{E}[\boldsymbol{e}_{s'_k} | s_k, a_k]$, we get

$$\begin{aligned}
\boldsymbol{\Phi}^T \boldsymbol{\mu}_{\boldsymbol{\lambda}(\boldsymbol{c}),\pi} &= (1-\gamma) \boldsymbol{\varphi}(s_0,\pi) + \gamma \sum_s (\boldsymbol{\Psi}^T \boldsymbol{\lambda}(\boldsymbol{c}))_s \boldsymbol{\varphi}(s,\pi) \\
&= (1-\gamma) \boldsymbol{\varphi}(s_0,\pi) + \frac{\gamma}{n} \sum_{k=1}^{n} c_k \mathbb{E}[\boldsymbol{\varphi}(s'_k, \pi) | s_k, a_k].
\end{aligned}$$

Hence,

$$\|\mathbf{\Phi}^T(\boldsymbol{\mu}_{\boldsymbol{\lambda}(\boldsymbol{c}),\pi} - \widehat{\boldsymbol{\mu}}_{\boldsymbol{\lambda}(\boldsymbol{c}),\pi})\|_2 = \gamma \left\| \frac{1}{n}\sum_{k=1}^{n} c_k(\boldsymbol{\varphi}(s'_k,\pi) - \mathbb{E}[\boldsymbol{\varphi}(s'_k,\pi)|s_k,a_k]) \right\|_2$$
$$\leq \mathcal{O}\left(B\sqrt{\frac{\log(d/\delta)}{n}} \right)$$

where the last inequality uses Matrix Bernstein inequality (Lemma B.1) with $S_k = c_k\boldsymbol{\varphi}(s'_k,\pi) - c_k\mathbb{E}[\boldsymbol{\varphi}(s'_k,\pi)|s_k,a_k]$. □

Lemma D.3. *Given a fixed $\boldsymbol{\lambda} \in \mathcal{C}_n(B)$, we have for all $\pi \in \Pi(D_\pi)$ that*

$$\|\mathbf{\Phi}^T\boldsymbol{\mu}_{\boldsymbol{\lambda},\pi} - \mathbf{\Phi}^T\widehat{\boldsymbol{\mu}}_{\boldsymbol{\lambda},\pi}\|_2 \leq \mathcal{O}\left(B\sqrt{\frac{\log(d/\delta) + d\log(D_\pi dn)}{n}} \right)$$

with probability at least $1-\delta$.

Proof. Consider an ε-cover of $\Pi(D_\pi)$ with covering balls when measuring distances with the norm $\|\pi - \pi'\|_{\infty,1} = \sup_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}|\pi(a|s) - \pi'(a|s)|$. By Lemma A.3, there exists such a cover with log covering number

$$\log \mathcal{N}(\Pi(D_\pi), \|\cdot\|_{\infty,1}, \varepsilon) \leq d\log\left(1 + \frac{16D_\pi}{\varepsilon}\right).$$

Fix any $\pi \in \Pi(D_\pi)$ and consider its nearest cover center π' measuring distances by $\|\cdot\|_{\infty,1}$. Then,

$$\|\mathbf{\Phi}^T\boldsymbol{\mu}_{\boldsymbol{\lambda},\pi} - \mathbf{\Phi}^T\widehat{\boldsymbol{\mu}}_{\boldsymbol{\lambda},\pi}\|_2 \leq \|\underbrace{\mathbf{\Phi}^T\boldsymbol{\mu}_{\boldsymbol{\lambda},\pi} - \mathbf{\Phi}^T\boldsymbol{\mu}_{\boldsymbol{\lambda},\pi'}}_{(i)}\|_2 + \|\underbrace{\mathbf{\Phi}^T\boldsymbol{\mu}_{\boldsymbol{\lambda},\pi'} - \mathbf{\Phi}^T\widehat{\boldsymbol{\mu}}_{\boldsymbol{\lambda},\pi'}}_{(ii)}\|_2 + \|\underbrace{\mathbf{\Phi}^T\widehat{\boldsymbol{\mu}}_{\boldsymbol{\lambda},\pi'} - \mathbf{\Phi}^T\widehat{\boldsymbol{\mu}}_{\boldsymbol{\lambda},\pi}}_{(iii)}\|_2.$$

Note that

$$\begin{aligned}
\mathbf{\Phi}^T\boldsymbol{\mu}_{\boldsymbol{\lambda},\pi} &= \sum_{s,a}\mu_{\boldsymbol{\lambda},\pi}(s,a)\boldsymbol{\varphi}(s,a) \\
&= \sum_{s,a}\pi(a|s)[(1-\gamma)\nu_0(s) + \gamma(\mathbf{\Psi}^T\boldsymbol{\lambda})_s]\boldsymbol{\varphi}(s,a) \\
&= (1-\gamma)\sum_a \pi(a|s_0)\boldsymbol{\varphi}(s_0,a) + \gamma\sum_s(\mathbf{\Psi}^T\boldsymbol{\lambda})_s\sum_a\pi(a|s)\boldsymbol{\varphi}(s,a) \\
&= (1-\gamma)\boldsymbol{\varphi}(s_0,\pi) + \gamma\sum_s(\mathbf{\Psi}^T\boldsymbol{\lambda})_s\boldsymbol{\varphi}(s,\pi),
\end{aligned} \tag{11}$$

where we use the notation $\boldsymbol{\varphi}(s,\pi) = \sum_a \pi(a|s)\boldsymbol{\varphi}(s,a)$. The first term (i) can be bounded by

$$\begin{aligned}
\|\mathbf{\Phi}^T\boldsymbol{\mu}_{\boldsymbol{\lambda},\pi} - \mathbf{\Phi}^T\boldsymbol{\mu}_{\boldsymbol{\lambda},\pi'}\|_2 &\leq (1-\gamma)\|\boldsymbol{\varphi}(s_0,\pi-\pi')\|_2 + \gamma\left\|\sum_s(\mathbf{\Psi}^T\boldsymbol{\lambda})_s\boldsymbol{\varphi}(s,\pi-\pi')\right\|_2 \\
&= (1-\gamma)\left\|\sum_a(\pi(a|s_0) - \pi'(a|s_0))\boldsymbol{\varphi}(s_0,a)\right\|_2 + \gamma\left\|\sum_s(\mathbf{\Phi}^T\boldsymbol{\lambda})_s\sum_a(\pi(a|s) - \pi'(a|s))\boldsymbol{\varphi}(s,a)\right\|_2 \\
&\leq (1-\gamma)\sum_a|\pi(a|s_0) - \pi'(a|s_0)|\|\boldsymbol{\varphi}(s_0,a)\|_2 + \gamma\sum_s(|\mathbf{\Psi}|^T|\boldsymbol{\lambda}|)_s\sum_a|\pi(a|s) - \pi'(a|s)|\|\boldsymbol{\varphi}(s,a)\|_2 \\
&\leq (1-\gamma)\varepsilon + \gamma\varepsilon\mathbf{1}_{\mathcal{S}}^T|\mathbf{\Psi}|^T|\boldsymbol{\lambda}| \\
&\leq (1-\gamma)\varepsilon + \gamma\varepsilon D_\psi\sqrt{d}\|\boldsymbol{\lambda}\|_2 \\
&\leq (1-\gamma)\varepsilon + \gamma\varepsilon\sqrt{d}D_\psi B
\end{aligned}$$

where the second to last inequality uses the boundedness assumption on $\mathbf{\Psi}$ and the last inequality follows by $\|\boldsymbol{\lambda}\|_2 \leq B$.

The second term (ii) can be bounded by a union bound of the concentration inequality in Lemma 3.1 across all π' in the cover, resulting in the following bound

$$\|\mathbf{\Phi}^T \boldsymbol{\mu}_{\boldsymbol{\lambda},\pi'} - \mathbf{\Phi}^T \widehat{\boldsymbol{\mu}}_{\boldsymbol{\lambda},\pi'}\|_2 \leq \mathcal{O}\left(B\sqrt{\frac{\log(d/\delta) + d\log(D_\pi/\varepsilon)}{n}}\right).$$

To bound the third term (iii), note that

$$\begin{aligned}\mathbf{\Phi}^T \widehat{\boldsymbol{\mu}}_{\boldsymbol{\lambda},\pi} &= (1-\gamma)\boldsymbol{\varphi}(s_0,\pi) + \gamma \sum_s (\widehat{\mathbf{\Psi}^T \boldsymbol{\lambda}})_s \boldsymbol{\varphi}(s,\pi) \\ &= (1-\gamma)\boldsymbol{\varphi}(s_0,\pi) + \frac{\gamma}{n}\sum_{k=1}^n c_k \boldsymbol{\varphi}(s'_k,\pi). \end{aligned} \tag{12}$$

where $\boldsymbol{\lambda} = \frac{1}{n}\sum_{k=1}^n c_k \boldsymbol{\varphi}(s_k,a_k)$. Therefore,

$$\begin{aligned}\|\mathbf{\Phi}^T \widehat{\boldsymbol{\mu}}_{\boldsymbol{\lambda},\pi'} - \mathbf{\Phi}^T \widehat{\boldsymbol{\mu}}_{\boldsymbol{\lambda},\pi}\|_2 &\leq (1-\gamma)\|\boldsymbol{\varphi}(s_0,\pi-\pi')\|_2 + \frac{\gamma}{n}\sum_{k=1}^n c_k \|\boldsymbol{\varphi}(s'_k,\pi-\pi')\|_2 \\ &\leq (1-\gamma)\varepsilon + \gamma B\varepsilon\end{aligned}$$

where the last inequality uses $\|\boldsymbol{\varphi}(s,\pi-\pi')\|_2 = \|\sum_a(\pi(a|s)-\pi'(a|s))\boldsymbol{\varphi}(s,a)\|_2 \leq \sum_a |\pi(a|s)-\pi'(a|s)|\|\boldsymbol{\varphi}(s,a)\|_2 \leq \varepsilon$. Combining the bounds of the three terms, we get

$$\|\mathbf{\Phi}^T \boldsymbol{\mu}_{\boldsymbol{\lambda},\pi} - \mathbf{\Phi}^T \widehat{\boldsymbol{\mu}}_{\boldsymbol{\lambda},\pi}\|_2 \leq \mathcal{O}\left(B\sqrt{\frac{\log(d/\delta) + d\log(D_\pi/\varepsilon)}{n}} + \sqrt{d}B\varepsilon\right).$$

Choosing $\varepsilon = 1/\sqrt{dn}$, we get the desired result. □

Lemma D.4. *The sequences $\{\pi_t\}, \{\boldsymbol{\lambda}(\boldsymbol{c}'_t)\}, \{\boldsymbol{\zeta}_t\}$ produced by Algorithm 1 satisfies*

$$\textsc{Reg}_t^\zeta = \langle \boldsymbol{\zeta}_t - \boldsymbol{\zeta}^{\pi_t}, \mathbf{\Phi}^T \boldsymbol{\mu}_{\boldsymbol{\lambda}(\boldsymbol{c}'_t),\pi_t} - \boldsymbol{\lambda}(\boldsymbol{c}'_t)\rangle \leq \mathcal{O}\left(\frac{C^* d}{1-\gamma}\sqrt{\frac{\log(dnT(\log|\mathcal{A}|)/\delta)}{n}}\right)$$

for $t = 1,\ldots,T$ with probability at least $1-\delta$.

Proof. Recall that $\mathcal{I} \subseteq \{1,\ldots,n\}$ is an index set of size d such that $\{\boldsymbol{\varphi}(s_j,a_j)\}_{j\in\mathcal{I}}$ is a 2-approximate barycentric spanner for $\{\boldsymbol{\varphi}(s_k,a_k)\}_{k=1}^n$. Let $\mathcal{C}'_n(C^*) = \{\boldsymbol{c}' \in [-2C^*, 2C^*]^n : c'_j = 0 \text{ if } j \in \mathcal{I}\}$. Consider a ε-cover of $\mathcal{C}'_n(C^*)$ with respect to distance induced by $\|\cdot\|_\infty$ where ε is to be chosen later, and let $\boldsymbol{c}''_t$ be the closest covering center to $\boldsymbol{c}'_t$. There exists a cover with covering number $(1+4C^*/\varepsilon)^d$. We can decompose the regret of ζ-player at step t into

$$\begin{aligned}\textsc{Reg}_t^\zeta &= \langle \boldsymbol{\zeta}_t - \boldsymbol{\zeta}^{\pi_t}, \mathbf{\Phi}^T \boldsymbol{\mu}_{\boldsymbol{\lambda}(\boldsymbol{c}'_t),\pi_t} - \boldsymbol{\lambda}(\boldsymbol{c}'_t)\rangle \\ &= \underbrace{\langle \boldsymbol{\zeta}_t - \boldsymbol{\zeta}^{\pi_t}, \mathbf{\Phi}^T \boldsymbol{\mu}_{\boldsymbol{\lambda}(\boldsymbol{c}'_t),\pi_t} - \mathbf{\Phi}^T \boldsymbol{\mu}_{\boldsymbol{\lambda}_t(\boldsymbol{c}''_t),\pi_t}\rangle}_{(a)} + \underbrace{\langle \boldsymbol{\zeta}_t - \boldsymbol{\zeta}^{\pi_t}, \mathbf{\Phi}^T \boldsymbol{\mu}_{\boldsymbol{\lambda}(\boldsymbol{c}''_t),\pi_t} - \mathbf{\Phi}^T \widehat{\boldsymbol{\mu}}_{\boldsymbol{\lambda}(\boldsymbol{c}''_t),\pi_t}\rangle}_{(b)} \\ &\quad + \underbrace{\langle \boldsymbol{\zeta}_t - \boldsymbol{\zeta}^{\pi_t}, \mathbf{\Phi}^T \widehat{\boldsymbol{\mu}}_{\boldsymbol{\lambda}(\boldsymbol{c}''_t),\pi_t} - \mathbf{\Phi}^T \widehat{\boldsymbol{\mu}}_{\boldsymbol{\lambda}(\boldsymbol{c}'_t),\pi_t}\rangle}_{(c)} + \underbrace{\langle \boldsymbol{\zeta}_t - \boldsymbol{\zeta}^{\pi_t}, \mathbf{\Phi}^T \widehat{\boldsymbol{\mu}}_{\boldsymbol{\lambda}(\boldsymbol{c}'_t),\pi_t} - \boldsymbol{\lambda}(\boldsymbol{c}'_t)\rangle}_{(d)}.\end{aligned}$$

Bounding (a) Recall from equation (11) that

$$\mathbf{\Phi}^T \boldsymbol{\mu}_{\boldsymbol{\lambda},\pi} = (1-\gamma)\boldsymbol{\varphi}(s_0,\pi) + \gamma\sum_s (\mathbf{\Psi}^T\boldsymbol{\lambda})_s \boldsymbol{\varphi}(s,\pi)$$

where we use the notation $\boldsymbol{\varphi}(s,\pi) = \sum_a \pi(a|s)\boldsymbol{\varphi}(s,a)$. Also, since $\|\boldsymbol{c}'_t - \boldsymbol{c}''_t\|_\infty \leq \varepsilon$, we have

$$\|\boldsymbol{\lambda}(\boldsymbol{c}'_t) - \boldsymbol{\lambda}(\boldsymbol{c}''_t)\|_2 = \frac{1}{n}\left\|\sum_{k=1}^n (c'_{tk} - c''_{tk})\boldsymbol{\varphi}(s_k,a_k)\right\|_2 \leq \frac{1}{n}\sum_{k=1}^n |c'_{tk} - c''_{tk}|\|\boldsymbol{\varphi}(s_k,a_k)\|_2 \leq \varepsilon.$$

Hence,

$$\begin{aligned}\|\mathbf{\Phi}^T\boldsymbol{\mu}_{\boldsymbol{\lambda}(\boldsymbol{c}_t'),\pi_t} - \mathbf{\Phi}^T\boldsymbol{\mu}_{\boldsymbol{\lambda}(\boldsymbol{c}_t''),\pi_t}\|_2 &= \gamma\|\sum_{s\in\mathcal{S}}(\mathbf{\Psi}^T(\boldsymbol{\lambda}(\boldsymbol{c}_t') - \boldsymbol{\lambda}(\boldsymbol{c}_t'')))_s\boldsymbol{\varphi}(s,\pi)\| \\ &\le \gamma\sum_{s\in\mathcal{S}}|(\mathbf{\Psi}^T(\boldsymbol{\lambda}(\boldsymbol{c}_t') - \boldsymbol{\lambda}(\boldsymbol{c}_t'')))_s|\|\boldsymbol{\varphi}(s,\pi)\|_2 \\ &\le \gamma\sum_{s\in\mathcal{S}}|(\mathbf{\Psi}^T(\boldsymbol{\lambda}(\boldsymbol{c}_t') - \boldsymbol{\lambda}(\boldsymbol{c}_t'')))_s| \\ &\le \gamma\varepsilon\mathbf{1}_{\mathcal{S}}^T|\mathbf{\Psi}|^T\mathbf{1}_d \\ &\le \gamma\varepsilon D_\psi d\end{aligned}$$

where we use the notation $|\mathbf{\Psi}|$ for the matrix that takes element-wise absolute value of $\mathbf{\Psi}$. The second inequality follows since $\|\boldsymbol{\psi}(s,\pi)\|_2 = \|\sum_a \pi(a|s)\boldsymbol{\varphi}(s,a)\|_2 \le \sum_a \pi(a|s)\|\boldsymbol{\varphi}(s,a)\|_2 \le \sum_a \pi(a|s) = 1$. The last inequality follows by the boundedness assumption on $\mathbf{\Phi}$. Hence, choosing $\varepsilon = 1/\sqrt{dn}$, Term (a) can be bounded by

$$\begin{aligned}\langle\boldsymbol{\zeta}_t - \boldsymbol{\zeta}_{\boldsymbol{w}_t}^{\pi_t}, \mathbf{\Phi}^T\boldsymbol{\mu}_{\boldsymbol{\lambda}_t,\pi_t} - \mathbf{\Phi}^T\boldsymbol{\mu}_{\boldsymbol{\lambda}_t',\pi_t}\rangle &\le \|\boldsymbol{\zeta}_t - \boldsymbol{\zeta}_{\boldsymbol{w}_t}^{\pi_t}\|_2\|\mathbf{\Phi}^T\boldsymbol{\mu}_{\boldsymbol{\lambda}_t,\pi_t} - \mathbf{\Phi}^T\boldsymbol{\mu}_{\boldsymbol{\lambda}_t',\pi_t}\|_2 \\ &\le \frac{2\gamma D_\zeta D_\psi\sqrt{d}}{\sqrt{n}}.\end{aligned}$$

Bounding (b) The second term can be bounded by a union bound of the concentration inequality in Lemma D.3 over a $(1/\sqrt{dn})$-cover of $\mathcal{C}_n'(C^*) = \{\boldsymbol{c}' \in [-2C^*, 2C^*]^n : c_j' = 0 \text{ if } j \in \mathcal{I}\}$, which gives

$$\begin{aligned}\langle\boldsymbol{\zeta}_t - \boldsymbol{\zeta}^{\pi_t}, \mathbf{\Phi}^T\boldsymbol{\mu}_{\boldsymbol{\lambda}_t',\pi_t} - \mathbf{\Phi}^T\widehat{\boldsymbol{\mu}}_{\boldsymbol{\lambda}_t',\pi_t}\rangle &\le \|\boldsymbol{\zeta}_t - \boldsymbol{\zeta}^{\pi_t}\|_2\|\mathbf{\Phi}^T\boldsymbol{\mu}_{\boldsymbol{\lambda}_t',\pi_t} - \mathbf{\Phi}^T\widehat{\boldsymbol{\mu}}_{\boldsymbol{\lambda}_t',\pi_t}\|_2 \\ &\le \mathcal{O}\left(D_\zeta C^*\sqrt{\frac{d\log(D_\pi dn/\delta)}{n}}\right)\end{aligned}$$

Bounding (c) Recall from (12) that $\mathbf{\Phi}^T\widehat{\boldsymbol{\mu}}_{\boldsymbol{\lambda},\pi} = (1-\gamma)\boldsymbol{\varphi}(s_0,\pi) + \frac{\gamma}{n}\sum_{k=1}^n c_k\boldsymbol{\varphi}(s_k',\pi)$. Since $\|c_t' - c_t''\|_\infty \le 1/\sqrt{dn}$, we have

$$\|\mathbf{\Phi}^T\widehat{\boldsymbol{\mu}}_{\boldsymbol{\lambda}(\boldsymbol{c}_t''),\pi_t} - \mathbf{\Phi}^T\widehat{\boldsymbol{\mu}}_{\boldsymbol{\lambda}(\boldsymbol{c}_t'),\pi_t}\|_2 = \frac{\gamma}{n}\left\|\sum_{k=1}^n(c_{tk}'' - c_{tk}')\boldsymbol{\varphi}(s_k',\pi_t)\right\|_2 \le \gamma/\sqrt{dn}.$$

It follows by Cauchy-Schwartz that

$$\langle\boldsymbol{\zeta}_t - \boldsymbol{\zeta}^{\pi_t}, \mathbf{\Phi}^T\widehat{\boldsymbol{\mu}}_{\boldsymbol{\lambda}(\boldsymbol{c}_t''),\pi_t} - \mathbf{\Phi}^T\widehat{\boldsymbol{\mu}}_{\boldsymbol{\lambda}(\boldsymbol{c}_t'),\pi_t}\rangle \le \mathcal{O}(D_\zeta/\sqrt{dn}).$$

Bounding (d) Recall that ζ-player chooses $\boldsymbol{\zeta}_t \in \mathbb{B}_d(D_\zeta)$ greedily that minimizes $\langle\cdot, \mathbf{\Phi}^T\widehat{\boldsymbol{\mu}}_{\boldsymbol{\lambda}(\boldsymbol{c}_t'),\pi_t} - \boldsymbol{\lambda}(\boldsymbol{c}_t')\rangle$ and that $\boldsymbol{\zeta}_{\boldsymbol{w}_t}^{\pi_t} \in \mathbb{B}_d(D_\zeta)$. Hence, the term (d) can be bounded by

$$\langle\boldsymbol{\zeta}_t - \boldsymbol{\zeta}^{\pi_t}, \mathbf{\Phi}^T\widehat{\boldsymbol{\mu}}_{\boldsymbol{\lambda}(\boldsymbol{c}_t'),\pi_t} - \boldsymbol{\lambda}(\boldsymbol{c}_t')\rangle \le 0.$$

Combining all the bounds, and using $D_\zeta := \sqrt{d} + \frac{\gamma D_\psi\sqrt{d}}{1-\gamma} \le \mathcal{O}(\frac{\sqrt{d}}{1-\gamma})$ and $D_\pi = \alpha T D_\zeta \le \mathcal{O}(\sqrt{\log|\mathcal{A}|T})$, we get

$$\textsc{Reg}_t^\zeta \le \mathcal{O}\left(\frac{C^* d}{1-\gamma}\sqrt{\frac{\log(dnT(\log|\mathcal{A}|)/\delta)}{n}}\right).$$

□

D.3. Bounding the Regret of λ-Player

Lemma D.5. *The sequences $\{\pi_t\}, \{\boldsymbol{\lambda}(\boldsymbol{c}_t)\}, \{\boldsymbol{\zeta}_t\}$ produced by Algorithm 1 satisfies*

$$\frac{1}{T}\sum_{t=1}^T \textsc{Reg}_t^\lambda \le \mathcal{O}\left(\frac{C^* d^{3/2}}{1-\gamma}\sqrt{\frac{\log(dnT(\log|\mathcal{A}|)/\delta)}{n}}\right) + \epsilon_{opt}^\lambda(T)$$

with probability at least $1-\delta$.

Proof. The regret of λ-player at step t can be bounded by

$$\begin{aligned}\textsc{Reg}_t^\lambda &= f(\boldsymbol{\zeta}_t, \boldsymbol{\lambda}^*, \pi_t) - f(\boldsymbol{\zeta}_t, \boldsymbol{\lambda}(\boldsymbol{c}_t), \pi_t) \\ &= \langle \boldsymbol{\lambda}^* - \boldsymbol{\lambda}(\boldsymbol{c}_t), \boldsymbol{\theta} + \gamma \boldsymbol{\Psi} \boldsymbol{v}_{\boldsymbol{\zeta}_t, \pi_t} - \boldsymbol{\zeta}_t \rangle \\ &= \langle \widehat{\boldsymbol{\lambda}}^* - \boldsymbol{\lambda}(\boldsymbol{c}_t), \boldsymbol{\theta} + \gamma \widehat{\boldsymbol{\Psi} \boldsymbol{v}}_{\boldsymbol{\zeta}_t, \pi_t} - \boldsymbol{\zeta}_t \rangle + \gamma \langle \widehat{\boldsymbol{\lambda}}^* - \boldsymbol{\lambda}(\boldsymbol{c}_t), \boldsymbol{\Psi} \boldsymbol{v}_{\boldsymbol{\zeta}_t, \pi_t} - \widehat{\boldsymbol{\Psi} \boldsymbol{v}}_{\boldsymbol{\zeta}_t, \pi_t} \rangle + \langle \boldsymbol{\lambda}^* - \widehat{\boldsymbol{\lambda}}^*, \boldsymbol{\theta} + \gamma \boldsymbol{\Psi} \boldsymbol{v}_{\boldsymbol{\zeta}_t, \pi_t} - \boldsymbol{\zeta}_t \rangle.\end{aligned}$$

The average of the first term over $t = 1, \ldots, T$ is $\epsilon_{\text{opt}}^\lambda(T)$ which vanishes as T increases since the λ-player employs a no-regret online convex optimization oracle (Definition 3.4) on the sequence of functions $\langle \cdot, \boldsymbol{\theta} + \gamma \widehat{\boldsymbol{\Psi} \boldsymbol{v}}_{\boldsymbol{\zeta}_t, \pi_t} - \boldsymbol{\zeta}_t \rangle$. The second term can be bounded as follows.

$$\begin{aligned}\langle \widehat{\boldsymbol{\lambda}}^* - \boldsymbol{\lambda}(\boldsymbol{c}_t), \boldsymbol{\Psi} \boldsymbol{v}_{\boldsymbol{\zeta}_t, \pi_t} - \widehat{\boldsymbol{\Psi} \boldsymbol{v}}_{\boldsymbol{\zeta}_t, \pi_t} \rangle &\le \|\widehat{\boldsymbol{\lambda}}^* - \boldsymbol{\lambda}(\boldsymbol{c}_t)\|_{(n\widehat{\boldsymbol{\Lambda}}_n + \boldsymbol{I})^{-1}} \|\boldsymbol{\Psi} \boldsymbol{v}_{\boldsymbol{\zeta}_t, \pi_t} - \widehat{\boldsymbol{\Psi} \boldsymbol{v}}_{\boldsymbol{\zeta}_t, \pi_t}\|_{n\widehat{\boldsymbol{\Lambda}}_n + \boldsymbol{I}} \\ &\le \frac{1}{\sqrt{n}} \|\widehat{\boldsymbol{\lambda}}^* - \boldsymbol{\lambda}(\boldsymbol{c}_t)\|_{\widehat{\boldsymbol{\Lambda}}_n^\dagger} \|\boldsymbol{\Psi} \boldsymbol{v}_{\boldsymbol{\zeta}_t, \pi_t} - \widehat{\boldsymbol{\Psi} \boldsymbol{v}}_{\boldsymbol{\zeta}_t, \pi_t}\|_{n\widehat{\boldsymbol{\Lambda}}_n + \boldsymbol{I}} \\ &\le \frac{2}{\sqrt{n}} C^* \sqrt{d} \cdot \mathcal{O}(D_v \sqrt{d \log(D_\zeta D_\pi n/\delta)}) \\ &\le \mathcal{O}\left(\frac{C^* d^{3/2}}{1-\gamma} \sqrt{\frac{\log(dnT(\log|\mathcal{A}|)/\delta)}{n}} \right)\end{aligned}$$

where the second inequality follows since $n\widehat{\boldsymbol{\Lambda}}_n + \boldsymbol{I} \succcurlyeq n\widehat{\boldsymbol{\Lambda}}_n$ and the fact that both $\widehat{\boldsymbol{\lambda}}^*$ and $\boldsymbol{\lambda}(\boldsymbol{c}_t)$ are in the column space of $\widehat{\boldsymbol{\Lambda}}$; the third inequality follows by Lemma 3.7 and Lemma 3.6 and the fact that the range of $\boldsymbol{v}_{\boldsymbol{\zeta}_t, \pi_t}$ is $[0, D_\zeta]$ so that we can set $D_v = D_\zeta$; the last inequality follows by $D_\zeta \le \mathcal{O}(\frac{\sqrt{d}}{1-\gamma})$ and $D_\pi = \alpha T D_\zeta = \mathcal{O}(\sqrt{T \log|\mathcal{A}|})$. The third term can be bounded by

$$\begin{aligned}\langle \boldsymbol{\lambda}^* - \widehat{\boldsymbol{\lambda}}^*, \boldsymbol{\theta} + \gamma \boldsymbol{\Psi} \boldsymbol{v}_{\boldsymbol{\zeta}_t, \pi_t} - \boldsymbol{\zeta}_t \rangle &\le \|\boldsymbol{\lambda}^* - \widehat{\boldsymbol{\lambda}}^*\|_2 \|\boldsymbol{\theta} + \gamma \boldsymbol{\Psi} \boldsymbol{v}_{\boldsymbol{\zeta}_t, \pi_t} - \boldsymbol{\zeta}_t\|_2 \\ &\le \mathcal{O}\left(C^* \sqrt{\frac{\log(d/\delta)}{n}} \right) \cdot \mathcal{O}(D_\zeta \sqrt{d}) \\ &\le \mathcal{O}\left(\frac{C^* d}{1-\gamma} \sqrt{\frac{\log(d/\delta)}{n}} \right)\end{aligned}$$

where the second inequality follows by Lemma 3.5 and the last inequality follows by the bound $D_\zeta \le \mathcal{O}(\frac{\sqrt{d}}{1-\gamma})$ and the boundedness assumption on $\boldsymbol{\Psi}$. Combining the three bounds completes the proof. □

E. Details in Offline Constrained RL Setting

E.1. Lagrangian Formulation

Recall that in the linear CMDP setting, the optimization problem of interest is

$$\begin{aligned}\max_{\pi} \quad & J_0(\pi) \\ \text{subject to} \quad & J_i(\pi) \ge \tau_i, \quad i = 1, \ldots, I.\end{aligned}$$

which we denote by $\mathcal{P}(\boldsymbol{\tau})$ parameterized by the thresholds $\boldsymbol{\tau} \in \mathbb{R}^I$. We write the Lagrangian function corresponding to the optimization problem $\mathcal{P}(\boldsymbol{\tau})$ as

$$L(\pi, \boldsymbol{w}) \coloneqq J(\pi) + \boldsymbol{w} \cdot (\boldsymbol{J}(\pi) - \boldsymbol{\tau})$$

where $\boldsymbol{J}(\cdot) = (J_1(\cdot), \ldots, J_I(\cdot))$, $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_I)$ and $\boldsymbol{w} \in \mathbb{R}^I$ is the Lagrangian multipliers corresponding to the constraints. The linear programming formulation of the constrained reinforcement learning problem is:

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

Using $\boldsymbol{r}_i = \boldsymbol{\Phi}\boldsymbol{\theta}_i$, $i = 0, \dots, I$, and $\boldsymbol{P} = \boldsymbol{\Phi}\boldsymbol{\Psi}$, which holds by the linear CMDP assumption (Assumption E), the linear program can be written as

$$\begin{aligned} \max_{\boldsymbol{\mu} \geq 0} \quad & \langle \boldsymbol{\theta}_0, \boldsymbol{\Phi}^T \boldsymbol{\mu} \rangle \\ \text{subject to} \quad & \langle \boldsymbol{\theta}_i, \boldsymbol{\Phi}^T \boldsymbol{\mu} \rangle \geq \tau_i, \quad i = 1, \dots, I, \\ & \boldsymbol{E}^T \boldsymbol{\mu} = (1-\gamma)\boldsymbol{\nu}_0 + \gamma \boldsymbol{\Psi}^T \boldsymbol{\Phi}^T \boldsymbol{\mu} \end{aligned}$$

Note that the optimization variable $\boldsymbol{\mu} \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|}$ is high-dimensional that depends on the size of $\mathcal{S}$. With the goal of computational and statistical efficiency, we introduce a low-dimensional optimization variable $\boldsymbol{\lambda} = \boldsymbol{\Phi}^T \boldsymbol{\mu} \in \mathbb{R}^d$, which has the interpretation of the average occupancy in the feature space. With the reparametrization, the optimization problem becomes

$$\begin{aligned} \max_{\boldsymbol{\mu} \geq \boldsymbol{0}, \boldsymbol{\lambda}} \quad & \langle \boldsymbol{\theta}_0, \boldsymbol{\lambda} \rangle \\ \text{subject to} \quad & \langle \boldsymbol{\theta}_i, \boldsymbol{\lambda} \rangle \geq \tau_i, \quad i = 1, \dots, I, \\ & \boldsymbol{E}^T \boldsymbol{\mu} = (1-\gamma)\boldsymbol{\nu}_0 + \gamma \boldsymbol{\Psi}^T \boldsymbol{\lambda} \\ & \boldsymbol{\lambda} = \boldsymbol{\Phi}^T \boldsymbol{\mu}. \end{aligned}$$

The dual of the linear program above is

$$\begin{aligned} \min_{\boldsymbol{w} \geq \boldsymbol{0}, \boldsymbol{v}, \boldsymbol{\zeta}} \quad & (1-\gamma)\langle \boldsymbol{\nu}_0, \boldsymbol{v} \rangle - \langle \boldsymbol{w}, \boldsymbol{\tau} \rangle \\ \text{subject to} \quad & \boldsymbol{\zeta} = \boldsymbol{\theta}_0 + \boldsymbol{\Theta}\boldsymbol{w} + \gamma \boldsymbol{\Psi}\boldsymbol{v} \\ & \boldsymbol{E}\boldsymbol{v} \geq \boldsymbol{\Phi}\boldsymbol{\zeta}. \end{aligned}$$

where we write $\boldsymbol{\Theta} = \begin{bmatrix} \boldsymbol{\theta}_1 & \cdots & \boldsymbol{\theta}_I \end{bmatrix} \in \mathbb{R}^{d \times I}$. The Lagrangian associated to this pair of linear programs is

$$\begin{aligned} L(\boldsymbol{\lambda}, \boldsymbol{\mu}; \boldsymbol{v}, \boldsymbol{w}, \boldsymbol{\zeta}) &= (1-\gamma)\langle \boldsymbol{\nu}_0, \boldsymbol{v} \rangle + \langle \boldsymbol{\lambda}, \boldsymbol{\theta}_0 + \gamma \boldsymbol{\Psi}\boldsymbol{v} - \boldsymbol{\zeta} \rangle + \langle \boldsymbol{\mu}, \boldsymbol{\Phi}\boldsymbol{\zeta} - \boldsymbol{E}\boldsymbol{v} \rangle - \langle \boldsymbol{w}, \boldsymbol{\tau} - \boldsymbol{\Theta}^T \boldsymbol{\lambda} \rangle \\ &= \langle \boldsymbol{\lambda}, \boldsymbol{\theta}_0 \rangle + \langle \boldsymbol{v}, (1-\gamma)\boldsymbol{\nu}_0 + \gamma \boldsymbol{\Psi}^T \boldsymbol{\lambda} - \boldsymbol{E}^T \boldsymbol{\mu} \rangle + \langle \boldsymbol{\zeta}, \boldsymbol{\Phi}^T \boldsymbol{\mu} - \boldsymbol{\lambda} \rangle - \langle \boldsymbol{w}, \boldsymbol{\tau} - \boldsymbol{\Theta}^T \boldsymbol{\lambda} \rangle. \end{aligned}$$

Note that the optimization variables $\boldsymbol{\lambda}, \boldsymbol{\zeta} \in \mathbb{R}^d$ and $\boldsymbol{w} \in \mathbb{R}^I$ are low-dimensional, but $\boldsymbol{\mu} \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|}$ and $\boldsymbol{v} \in \mathbb{R}^{|\mathcal{S}|}$ are not. With the goal of running a primal-dual algorithm on the Lagrangian using only low-dimensional variables, we introduce policy variable π and parameterize $\boldsymbol{\mu}$ and $\boldsymbol{v}$, as was done for the unconstrained RL setting, by

$$\begin{aligned} \mu_{\boldsymbol{\lambda},\pi}(s,a) &= \pi(a|s)\left[(1-\gamma)\nu_0(s) + \gamma\langle \psi(s), \boldsymbol{\lambda} \rangle\right] \\ v_{\boldsymbol{\zeta},\pi}(s) &= \sum_a \pi(a|s)\langle \boldsymbol{\zeta}, \boldsymbol{\varphi}(s,a) \rangle. \end{aligned}$$

Note that the choice of $\boldsymbol{\mu}_{\boldsymbol{\lambda},\pi}$ makes the Bellman flow constraint $\boldsymbol{E}^T\boldsymbol{\mu} = (1-\gamma)\boldsymbol{\nu}_0 + \gamma\boldsymbol{\Psi}^T\boldsymbol{\lambda}$ of the primal problem satisfied. Also, the choice of $\boldsymbol{v}_{\boldsymbol{\zeta},\pi}$ makes $\langle \boldsymbol{\mu}_{\boldsymbol{\lambda},\pi}, \boldsymbol{\Phi}\boldsymbol{\zeta} - \boldsymbol{E}\boldsymbol{v}_{\boldsymbol{\zeta},\pi} \rangle = 0$. Using the above parameterization, the Lagrangian can be rewritten in terms of $\boldsymbol{\zeta}, \boldsymbol{\lambda}, \boldsymbol{w}, \pi$ as follows:

$$g(\boldsymbol{\lambda}, \boldsymbol{\zeta}, \boldsymbol{w}, \pi) = \langle \boldsymbol{\lambda}, \boldsymbol{\theta}_0 \rangle + \langle \boldsymbol{\zeta}, \boldsymbol{\Phi}^T \boldsymbol{\mu}_{\boldsymbol{\lambda},\pi} - \boldsymbol{\lambda} \rangle - \langle \boldsymbol{w}, \boldsymbol{\tau} - \boldsymbol{\Theta}^T \boldsymbol{\lambda} \rangle \tag{13}$$

$$= (1-\gamma)\langle \boldsymbol{\nu}_0, \boldsymbol{v}_{\boldsymbol{\zeta},\pi} \rangle + \langle \boldsymbol{\lambda}, \boldsymbol{\theta}_0 + \gamma\boldsymbol{\Psi}\boldsymbol{v}_{\boldsymbol{\zeta},\pi} - \boldsymbol{\zeta} \rangle - \langle \boldsymbol{w}, \boldsymbol{\tau} - \boldsymbol{\Theta}^T \boldsymbol{\lambda} \rangle. \tag{14}$$

At the cost of having to keep track of π, we can now run a primal-dual algorithm on the low-dimensional variables $\boldsymbol{\zeta}$, $\boldsymbol{\lambda}$ and $\boldsymbol{w}$. As is the case for the unconstrained RL setting, the introduction of π in the equation does not make the algorithm inefficient because we can only keep track of the distribution $\pi(s|a)$ for state-action pairs that appear in the dataset.

E.2. Technical Lemmas on Lagrangian

For a linearized reward function $u = r_0 + \boldsymbol{w} \cdot \boldsymbol{r}$ where we use the notation $\boldsymbol{r}$ to denote the vector of reward functions $r_1, \dots, r_I$ such that $u(s,a) = r_0(s,a) + \sum_{i=1}^I w_i r_i(s,a)$, the Bellman equation becomes

$$\boldsymbol{Q}^\pi_{\boldsymbol{w}} = \boldsymbol{\Phi}(\boldsymbol{\theta}_0 + \boldsymbol{\Theta}\boldsymbol{w} + \gamma\boldsymbol{\Psi}\boldsymbol{V}^\pi_{\boldsymbol{w}}) = \boldsymbol{\Phi}\boldsymbol{\zeta}^\pi_{\boldsymbol{w}} \tag{15}$$

where we write $\boldsymbol{Q}_{\boldsymbol{w}}^{\pi}$ and $\boldsymbol{V}_{\boldsymbol{w}}^{\pi}$ as the value functions of the policy π with respect to the linearized reward function $r_0 + \boldsymbol{w} \cdot \boldsymbol{r}$ and define

$$\boldsymbol{\zeta}_{\boldsymbol{w}}^{\pi} \coloneqq \boldsymbol{\theta}_0 + \boldsymbol{\Theta}\boldsymbol{w} + \gamma \boldsymbol{\Psi} \boldsymbol{V}_{\boldsymbol{w}}^{\pi}.$$

Note that if $\boldsymbol{w} \in D_w \boldsymbol{\Delta}^I$, we have $\|\boldsymbol{\zeta}_{\boldsymbol{w}}^{\pi}\|_2 \leq 1 + D_w + \frac{\gamma\sqrt{d}(1+D_w)}{1-\gamma} = \mathcal{O}\left(\frac{D_w\sqrt{d}}{1-\gamma}\right)$. We define $D_\zeta \coloneqq 1 + D_w + \frac{\gamma\sqrt{d}(1+D_w)}{1-\gamma}$.

Lemma E.1. *Let $\boldsymbol{\zeta}_{\boldsymbol{w}}^{\pi}$ be the parameter that satisfies $\boldsymbol{Q}_{\boldsymbol{w}}^{\pi} = \boldsymbol{\Phi}\boldsymbol{\zeta}_{\boldsymbol{w}}^{\pi}$ for a given $\boldsymbol{w} \in \mathbb{R}^I$ and a policy π. Then,*

$$L(\pi, \boldsymbol{w}) = g(\boldsymbol{\zeta}_{\boldsymbol{w}}^{\pi}, \boldsymbol{\lambda}, \pi, \boldsymbol{w})$$

for all $\boldsymbol{\lambda} \in \mathbb{R}^d$ in the span of $\{\boldsymbol{\varphi}(s,a)\}_{(s,a)\in\mathcal{S}\times\mathcal{A}}$.

Proof. For convenience, define the reward function $u(s,a) = r_0(s,a) + \sum_{i=1}^{I} w_i r_i(s,a)$. By the linear CMDP assumption, we have $\boldsymbol{u} = \boldsymbol{\Phi}(\boldsymbol{\theta}_0 + \boldsymbol{\Theta}\boldsymbol{w})$ where $\boldsymbol{u} \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|}$ is the vector representation of the reward function u. Also, by the definition of $v_{\boldsymbol{\zeta},\pi}$ in (3), we have

$$v_{\boldsymbol{\zeta}_{\boldsymbol{w}}^{\pi},\pi}(s) = \sum_a \pi(a|s)\langle \boldsymbol{\zeta}_{\boldsymbol{w}}^{\pi}, \boldsymbol{\varphi}(s,a)\rangle = \sum_a \pi(a|s) Q_{\boldsymbol{w}}^{\pi}(s,a) = V_{\boldsymbol{w}}^{\pi}(s).$$

Since we assume $\boldsymbol{\lambda} \in \mathbb{R}^d$ is in the span of $\{\boldsymbol{\varphi}(s,a)\}_{(s,a)\in\mathcal{S}\times\mathcal{A}}$, there exists $\boldsymbol{\alpha} \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|}$ such that $\boldsymbol{\lambda} = \boldsymbol{\Phi}^T\boldsymbol{\alpha}$. Hence, using the form (14) of the Lagrangian function, we have

$$\begin{aligned}
g(\boldsymbol{\zeta}_{\boldsymbol{w}}^{\pi}, \boldsymbol{\lambda}, \pi, \boldsymbol{w}) &= (1-\gamma)\langle \boldsymbol{\nu}_0, \boldsymbol{v}_{\boldsymbol{\zeta}_{\boldsymbol{w}}^{\pi}}\rangle + \langle \boldsymbol{\lambda}, \boldsymbol{\theta}_0 + \gamma\boldsymbol{\Psi}\boldsymbol{v}_{\boldsymbol{\zeta}_{\boldsymbol{w}}^{\pi}} - \boldsymbol{\zeta}_{\boldsymbol{w}}^{\pi}\rangle - \langle \boldsymbol{w}, \boldsymbol{\tau} - \boldsymbol{\Theta}^T\boldsymbol{\lambda}\rangle \\
&= (1-\gamma)\langle \boldsymbol{\nu}_0, \boldsymbol{V}_{\boldsymbol{w}}^{\pi}\rangle + \langle \boldsymbol{\lambda}, \boldsymbol{\theta}_0 + \boldsymbol{\Theta}\boldsymbol{w} + \gamma\boldsymbol{\Psi}\boldsymbol{V}_{\boldsymbol{w}}^{\pi} - \boldsymbol{\zeta}_{\boldsymbol{w}}^{\pi}\rangle - \langle \boldsymbol{w}, \boldsymbol{\tau}\rangle \\
&= (1-\gamma)\langle \boldsymbol{\nu}_0, \boldsymbol{V}_{\boldsymbol{w}}^{\pi}\rangle + \langle \boldsymbol{\alpha}, \boldsymbol{\Phi}(\boldsymbol{\theta}_0 + \boldsymbol{\Theta}\boldsymbol{w} + \gamma\boldsymbol{\Psi}\boldsymbol{V}_{\boldsymbol{w}}^{\pi} - \boldsymbol{\zeta}_{\boldsymbol{w}}^{\pi})\rangle - \langle \boldsymbol{w}, \boldsymbol{\tau}\rangle \\
&= (1-\gamma)\langle \boldsymbol{\nu}_0, \boldsymbol{V}_{\boldsymbol{w}}^{\pi}\rangle + \langle \boldsymbol{\alpha}, \boldsymbol{u} + \gamma\boldsymbol{P}\boldsymbol{V}_{\boldsymbol{w}}^{\pi} - \boldsymbol{Q}_{\boldsymbol{w}}^{\pi}\rangle - \langle \boldsymbol{w}, \boldsymbol{\tau}\rangle \\
&= (1-\gamma)\langle \boldsymbol{\nu}_0, \boldsymbol{V}_{\boldsymbol{w}}^{\pi}\rangle - \langle \boldsymbol{w}, \boldsymbol{\tau}\rangle \\
&= L(\pi, \boldsymbol{w})
\end{aligned}$$

where the second to last equality uses the Bellman equation (1) and the last equality is by $L(\pi, \boldsymbol{w}) = J_0(\pi) + \boldsymbol{w}\cdot(\boldsymbol{J}(\pi) - \boldsymbol{\tau})$ and the fact that $J_0(\pi) + \boldsymbol{w}\cdot\boldsymbol{J}(\pi)$ is the value of π with respect to the linearized value function $r_0 + \boldsymbol{w}\cdot\boldsymbol{r}$. □

Lemma E.2. *Under the linear MDP setting, let $\boldsymbol{\zeta}^{\pi}$ be the parameter that satisfies $\boldsymbol{Q}^{\pi} = \boldsymbol{\Phi}\boldsymbol{\zeta}^{\pi}$ for a policy π. Then,*

$$J(\pi) = f(\boldsymbol{\zeta}^{\pi}, \boldsymbol{\lambda}, \pi)$$

for all $\boldsymbol{\lambda} \in \mathbb{R}^d$ in the span of $\{\boldsymbol{\varphi}(s,a)\}_{(s,a)\in\mathcal{S}\times\mathcal{A}}$.

Proof. This is a direct corollary of Lemma E.1, which can be seen by setting $\boldsymbol{w} = \boldsymbol{0}$. □

Lemma E.3. *Given a policy π, let μ^{π} be the occupancy measure induced by π and let $\boldsymbol{\lambda}^{\pi} = \boldsymbol{\Phi}^T\boldsymbol{\mu}^{\pi}$. Then, for any $\boldsymbol{\zeta} \in \mathbb{R}^d$ and any $\boldsymbol{w} \in \mathbb{R}^I$, we have*

$$L(\pi, \boldsymbol{w}) = g(\boldsymbol{\zeta}, \boldsymbol{\lambda}^{\pi}, \boldsymbol{w}, \pi).$$

Proof. By the definition of $\mu_{\boldsymbol{\lambda},\pi}$ in (2), we have

$$\begin{aligned}
\mu_{\boldsymbol{\lambda}^{\pi},\pi}(s,a) &= \pi(a|s)\left[(1-\gamma)\nu_0(s) + \gamma\langle \boldsymbol{\psi}(s), \boldsymbol{\lambda}^{\pi}\rangle\right] \\
&= \pi(a|s)\left[(1-\gamma)\nu_0(s) + \gamma\langle \boldsymbol{\psi}(s), \boldsymbol{\Phi}^T\boldsymbol{\mu}^{\pi}\rangle\right] \\
&= \mu^{\pi}(s,a).
\end{aligned}$$

Using the form (13) of the Lagrangian function, we have

$$\begin{aligned}
g(\boldsymbol{\zeta}, \boldsymbol{\lambda}^{\pi}, \boldsymbol{w}, \pi) &= \langle \boldsymbol{\lambda}^{\pi}, \boldsymbol{\theta}_0\rangle + \langle \boldsymbol{\zeta}, \boldsymbol{\Phi}^T\boldsymbol{\mu}_{\boldsymbol{\lambda}^{\pi},\pi} - \boldsymbol{\lambda}^{\pi}\rangle - \langle \boldsymbol{w}, \boldsymbol{\tau} - \boldsymbol{\Theta}^T\boldsymbol{\lambda}^{\pi}\rangle \\
&= \langle \boldsymbol{\Phi}^T\boldsymbol{\mu}^{\pi}, \boldsymbol{\theta}_0\rangle - \langle \boldsymbol{w}, \boldsymbol{\tau} - \boldsymbol{\Theta}^T\boldsymbol{\Phi}^T\boldsymbol{\mu}^{\pi}\rangle \\
&= \langle \boldsymbol{\mu}^{\pi}, \boldsymbol{r}_0\rangle - \langle \boldsymbol{w}, \boldsymbol{\tau} - \boldsymbol{R}^T\boldsymbol{\mu}^{\pi}\rangle \\
&= L(\pi, \boldsymbol{w})
\end{aligned}$$

where the second equality uses $\boldsymbol{\mu}_{\boldsymbol{\lambda}^{\pi},\pi} = \boldsymbol{\mu}^{\pi}$ and $\boldsymbol{\lambda}^{\pi} = \boldsymbol{\Phi}^T\boldsymbol{\mu}^{\pi}$; and the third equality uses the matrix notation for the reward functions $\boldsymbol{R} = \{r_i(s,a)\}_{(s,a)\in\mathbb{R}^{|\mathcal{S}\times\mathcal{A}|}, i\in[I]}$; the last equality uses $J_i(\pi) = \langle \boldsymbol{\mu}^{\pi}, \boldsymbol{r}_i\rangle$, $i = 1, \ldots, I$. □

Lemma E.4. *Under the linear MDP setting, let μ^π be the occupancy measure induced by a policy π and let $\boldsymbol{\lambda}^\pi = \boldsymbol{\Phi}^T \boldsymbol{\mu}^\pi$. Then, for any $\boldsymbol{\zeta} \in \mathbb{R}^d$, we have*

$$J(\pi) = f(\boldsymbol{\zeta}, \boldsymbol{\lambda}^\pi, \pi).$$

Proof. This is a direct corollary of Lemma E.3, which can be seen by setting $\boldsymbol{w} = \mathbf{0}$. □

Define $L_\eta(\pi, \boldsymbol{w}) = J_0(\pi) + \boldsymbol{w} \cdot (\boldsymbol{J}(\pi) - \boldsymbol{\tau} - \eta \mathbf{1})$ to be the Lagrangian function associated with $\mathcal{P}(\boldsymbol{\tau} + \eta \mathbf{1})$. The following lemma shows that the near saddle point of $L_\eta(\cdot, \cdot)$ is a nearly optimal solution of the optimization problem $\mathcal{P}(\boldsymbol{\tau} + \eta \mathbf{1})$.

Lemma E.5. *Assume that Slater's condition (Assumption D) holds and that $\eta < \phi$ so that $\mathcal{P}(\boldsymbol{\tau} + \eta \mathbf{1})$ also satisfies Slater's condition. Suppose $(\bar{\pi}, \bar{\boldsymbol{w}})$ satisfies $L_\eta(\pi, \bar{\boldsymbol{w}}) \leq L_\eta(\bar{\pi}, \boldsymbol{w}) + \xi$ for all policies π and $\boldsymbol{w} \in B\boldsymbol{\Delta}^I$. Let $(\pi^*_\eta, \boldsymbol{w}^*_\eta)$ be a primal-dual solution to $\mathcal{P}(\boldsymbol{\tau} + \eta \mathbf{1})$. Assume $B > \|\boldsymbol{w}^*_\eta\|_1$. Then, we have*

$$J_0(\bar{\pi}) \geq J_0(\pi^*_\eta) - \xi \tag{Optimality}$$

$$J_i(\bar{\pi}) \geq \tau_i + \eta - \frac{\xi}{B - \|\boldsymbol{w}^*_\eta\|_1}, \quad \textit{for all } i = 1, \ldots, I \tag{Feasibility}$$

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

Optimality Since $(\bar{\pi}, \bar{\boldsymbol{w}})$ satisfies $L_\eta(\pi, \bar{\boldsymbol{w}}) \leq L_\eta(\bar{\pi}, \boldsymbol{w}) + \xi$ for all policies π and $\boldsymbol{w} \in B\boldsymbol{\Delta}^I$, we have $L_\eta(\pi^*_\eta, \bar{\boldsymbol{w}}) \leq L_\eta(\bar{\pi}, \boldsymbol{w}) + \xi$ for all $\boldsymbol{w} \in B\boldsymbol{\Delta}^I$. Choosing $\boldsymbol{w} = \mathbf{0}$, we get

$$L_\eta(\pi^*_\eta, \bar{\boldsymbol{w}}) \leq L_\eta(\bar{\pi}, \mathbf{0}) + \xi = J_0(\bar{\pi}) + \xi.$$

Rearranging, we get

$$J_0(\bar{\pi}) \geq J_0(\pi^*_\eta) + \bar{\boldsymbol{w}} \cdot (\boldsymbol{J}(\pi^*_\eta) - \boldsymbol{\tau} - \eta \mathbf{1}) - \xi \geq J_0(\pi^*_\eta) - \xi$$

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

Feasibility Recall that $(\pi^*_\eta, \boldsymbol{w}^*_\eta)$ 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, $(\pi^*_\eta, \boldsymbol{w}^*_\eta)$ is a saddle point for $L_\eta(\cdot, \cdot)$. Hence, we have

$$L_\eta(\bar{\pi}, \boldsymbol{w}^*_\eta) \leq L_\eta(\pi^*_\eta, \boldsymbol{w}^*_\eta) = J_0(\pi^*_\eta) + \boldsymbol{w}^*_\eta \cdot (\boldsymbol{J}(\pi^*_\eta) - \boldsymbol{\tau} - \eta \mathbf{1}) = J_0(\pi^*_\eta)$$

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

$$J_0(\pi^*_\eta) - J_0(\bar{\pi}) \geq \boldsymbol{w}^*_\eta \cdot (\boldsymbol{J}(\bar{\pi}) - \boldsymbol{\tau} - \eta \mathbf{1}) \geq (m - \eta)\|\boldsymbol{w}^*_\eta\|_1 \tag{16}$$

where we define $m = \min_{i \in [I]}(J_i(\bar{\pi}) - \tau_i)$. Now, to upper bound $J_0(\pi^*_\eta) - J_0(\bar{\pi})$, we first use the feasibility of π^*_η for $\mathcal{P}(\boldsymbol{\tau} + \eta \mathbf{1})$ as follows.

$$L_\eta(\pi^*_\eta, \bar{\boldsymbol{w}}) = J_0(\pi^*_\eta) + \bar{\boldsymbol{w}} \cdot (\boldsymbol{J}(\pi^*_\eta) - \boldsymbol{\tau} - \eta \mathbf{1}) \geq J_0(\pi^*_\eta).$$

On the other hand, since $(\bar{\pi}, \bar{\boldsymbol{w}})$ satisfies $L(\pi, \bar{\boldsymbol{w}}) \leq L(\bar{\pi}, \boldsymbol{w}) + \xi$ for all policies π and $\boldsymbol{w} \in B\boldsymbol{\Delta}^I$, we have $L_\eta(\pi^*_\eta, \bar{\boldsymbol{w}}) \leq L_\eta(\bar{\pi}, \boldsymbol{w}) + \xi$ for any $\boldsymbol{w} \in B\boldsymbol{\Delta}^I$. By choosing $\boldsymbol{w}$ such that $w_j = B$ for $j = \operatorname{argmin}_{i \in [I]}(J_i(\bar{\pi}) - \tau_i)$ and recalling $m = \min_{i \in [I]}(J_i(\bar{\pi}) - \tau_i)$, we get

$$L_\eta(\pi^*_\eta, \bar{\boldsymbol{w}}) \leq L_\eta(\bar{\pi}, \boldsymbol{w}) + \xi = J_0(\bar{\pi}) + B(m - \eta) + \xi.$$

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

$$J_0(\pi^*_\eta) - J_0(\bar{\pi}) \leq B(m - \eta) + \xi. \tag{17}$$

Combining the lower bound (16) and the upper bound (17) of $J_0(\pi^*_\eta) - J_0(\bar{\pi})$ and rearranging, we get

$$m - \eta \geq \frac{-\xi}{B - \|\boldsymbol{w}^*_\eta\|_1}.$$

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

$$J_i(\bar{\pi}) \geq \tau_i + \eta - \frac{\xi}{B - \|\boldsymbol{w}_\eta^*\|_1}$$

for all $i = 1, \ldots, I$. □

Lemma E.6 (Lemma 13 in Hong et al. (2023))**.** *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 $\varphi > 0$, in other words, there exists $\pi \in \Pi$ that satisfies the constraint $J_i(\pi) \geq \tau_i + \phi$ for all $i = 1, \ldots, I$. Then, the optimal dual variable $\boldsymbol{\lambda}^*$ of the problem satisfies $\|\boldsymbol{\lambda}^*\|_1 \leq \frac{1}{\varphi}$.*

Proof. Let π^* be an optimal policy of the optimization problem $\mathcal{P}(\boldsymbol{\tau})$. Define the dual function $f(\boldsymbol{\lambda}) = \max_\pi J_0(\pi) + \boldsymbol{\lambda} \cdot (\boldsymbol{J}(\pi) - \boldsymbol{\tau})$. Let $\boldsymbol{\lambda}^* = \operatorname{argmin}_{\boldsymbol{\lambda} \in \mathbb{R}_+^I} f(\boldsymbol{\lambda})$. Trivially, $\lambda_i^* \geq 0$ for all $i = 1, \ldots, I$. Also, by strong duality, we have $f(\boldsymbol{\lambda}^*) = J_0(\pi^*)$. Let $\widehat{\pi}$ be a feasible policy with $\boldsymbol{J}(\widehat{\pi}) \geq \boldsymbol{\tau} + \phi \mathbf{1}$ where the inequality is component-wise and $\mathbf{1} = (1, \ldots, 1)$. Such a policy exists by the assumption of this lemma. Then,

$$J_0(\pi^*) = f(\boldsymbol{\lambda}^*) \geq J_0(\widehat{\pi}) + \boldsymbol{\lambda}^* \cdot (\boldsymbol{J}(\widehat{\pi}) - \boldsymbol{\tau}) \geq J_0(\widehat{\pi}) + \boldsymbol{\lambda}^* \cdot \phi \mathbf{1} = J_0(\widehat{\pi}) + \phi \|\boldsymbol{\lambda}^*\|_1.$$

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

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

□

E.3. Proof of Theorem 5.1

For a given $\boldsymbol{w} \in \mathbb{R}^I$ and a policy π, define $\boldsymbol{\zeta}_{\boldsymbol{w}}^\pi \in \mathbb{R}^d$ to be the parameter that satisfies $\boldsymbol{Q}_{\boldsymbol{w}}^\pi = \boldsymbol{\Phi} \boldsymbol{\zeta}_{\boldsymbol{w}}^\pi$ where $\boldsymbol{Q}_{\boldsymbol{w}}^\pi$ is the state-action value function of the policy π with respect to the reward function $r_0 + \boldsymbol{w} \cdot \boldsymbol{r}$. Using $g(\boldsymbol{\zeta}_{\boldsymbol{w}}^\pi, \boldsymbol{\lambda}, \boldsymbol{w}, \pi) = L(\pi, \boldsymbol{w})$ for any $\boldsymbol{\lambda}$ that is a linear combination of $\{\boldsymbol{\varphi}(s,a)\}_{(s,a) \in \mathcal{S} \times \mathcal{A}}$ (Lemma 13) and $g(\boldsymbol{\zeta}, \boldsymbol{\lambda}^\pi, \boldsymbol{w}, \pi) = L(\pi, \boldsymbol{w})$ for any $\boldsymbol{\zeta} \in \mathbb{R}^d$ where $\boldsymbol{\lambda}^\pi = \boldsymbol{\Phi}^T \boldsymbol{\mu}^\pi$ (Lemma 14), we have

$$\begin{aligned}
L(\pi^*, \boldsymbol{w}_t) - L(\pi_t, \boldsymbol{w}) &= g(\boldsymbol{\zeta}_t, \boldsymbol{\lambda}^*, \boldsymbol{w}_t, \pi) - g(\boldsymbol{\zeta}_{\boldsymbol{w}_t}^{\pi_t}, \boldsymbol{\lambda}(\boldsymbol{c}_t'), \boldsymbol{w}, \pi_t)) \\
&= \underbrace{(g(\boldsymbol{\zeta}_t, \boldsymbol{\lambda}^*, \boldsymbol{w}_t, \pi^*) - g(\boldsymbol{\zeta}_t, \boldsymbol{\lambda}^*, \boldsymbol{w}_t, \pi_t))}_{\textsc{Reg}_t^\pi} \\
&\quad + \underbrace{(g(\boldsymbol{\zeta}_t, \boldsymbol{\lambda}^*, \boldsymbol{w}_t, \pi_t) - g(\boldsymbol{\zeta}_t, \boldsymbol{\lambda}(\boldsymbol{c}_t'), \boldsymbol{w}_t, \pi_t))}_{\textsc{Reg}_t^\lambda} \\
&\quad + \underbrace{(g(\boldsymbol{\zeta}_t, \boldsymbol{\lambda}(\boldsymbol{c}_t'), \boldsymbol{w}_t, \pi_t) - g(\boldsymbol{\zeta}_t, \boldsymbol{\lambda}(\boldsymbol{c}_t'), \boldsymbol{w}, \pi_t))}_{\textsc{Reg}_t^w} \\
&\quad + \underbrace{(g(\boldsymbol{\zeta}_t, \boldsymbol{\lambda}(\boldsymbol{c}_t'), \boldsymbol{w}, \pi_t) - g(\boldsymbol{\zeta}_{\boldsymbol{w}_t}^{\pi_t}, \boldsymbol{\lambda}(\boldsymbol{c}_t'), \boldsymbol{w}, \pi_t))}_{\textsc{Reg}_t^\zeta}
\end{aligned}$$

where π^* is an optimal policy for the optimization problem $\mathcal{P}(\boldsymbol{\tau})$ and we use the notation $\boldsymbol{\lambda}^* = \boldsymbol{\lambda}^{\pi^*}$. Note that the suboptimality $L(\pi^*, \boldsymbol{w}_t) - L(\pi_t, \boldsymbol{w})$ is decomposed into regret terms of the four players. As long as we show that the sum of the four regrets over $t = 1, \ldots, T$ are sublinear in T and the dataset size n, we obtain $\frac{1}{T} \sum_{t=1}^T L(\pi^*, \boldsymbol{w}_t) - L(\pi_t, \boldsymbol{w}) = L(\pi^*, \bar{\boldsymbol{w}}) - L(\bar{\pi}, \boldsymbol{w}) = o(1)$ where $\bar{\boldsymbol{w}} = \frac{1}{T} \sum_{t=1}^T \boldsymbol{w}_t$ and $\bar{\pi} = \text{Unif}(\pi_1, \ldots, \pi_T)$ is the mixture policy that chooses a policy among $\pi_1, \ldots, \pi_T$ uniformly at random and runs the chosen policy for the entire trajectory. Then, for large enough T and n, we get $L(\pi^*, \bar{\boldsymbol{w}}) \leq L(\bar{\pi}, \boldsymbol{w}) + \epsilon$ where ϵ vanishes as T and n increase. Such a pair is a near saddle point of the Lagrangian function $L(\cdot, \cdot)$ and it can be shown that the mixture policy $\bar{\pi}$ is a near-optimal solution of the optimization problem (OPT). Specifically, adapting the proof of Hong et al. (2023), we can show that if the Slater's condition (Assumption D) holds, then a near saddle point $(\bar{\pi}, \bar{\boldsymbol{w}})$ of $L(\cdot, \cdot)$ with $L(\pi, \bar{\boldsymbol{w}}) \leq L(\bar{\pi}, \boldsymbol{w}) + \mathcal{O}(\epsilon)$ for all policies π and $\boldsymbol{w} \in (1 + \frac{1}{\phi}) \boldsymbol{\Delta}^I$ satisfies

$$\begin{aligned}
&J_0(\bar{\pi}) \geq J_0(\pi^*) - \epsilon \\
&J_i(\bar{\pi}) \geq \tau_i - \epsilon, \quad i = 1, \ldots, I
\end{aligned}$$

where π^* is the optimal policy for $\mathcal{P}(\boldsymbol{\tau})$, implying that $\bar{\pi}$ is a nearly optimal solution for the optimization problem.

In the rest of the section, we sketch the analysis that shows that $L(\pi, \bar{\boldsymbol{w}}) - L(\bar{\pi}, \boldsymbol{w}) \leq \epsilon$ for large enough T and $n = \mathcal{O}(\epsilon^{-2})$. With the decomposition of $L(\pi, \boldsymbol{w}_t) - L(\pi_t, \boldsymbol{w})$ into regrets of the four players discussed previously, we study how the four regrets can be bounded in the next four subsections.

E.3.1. Bounding Regret of π-Player

Using the expression (14), the regret of π-player simplifies to

$$\begin{aligned}\text{Reg}_t^\pi &= g(\boldsymbol{\zeta}_t, \boldsymbol{\lambda}^*, \boldsymbol{w}_t, \pi^*) - g(\boldsymbol{\zeta}_t, \boldsymbol{\lambda}^*, \boldsymbol{w}_t, \pi_t) \\ &= \langle \boldsymbol{\nu}^*, \boldsymbol{v}_{\boldsymbol{\zeta}_t,\pi} - \boldsymbol{v}_{\boldsymbol{\zeta}_t,\pi_t} \rangle \\ &= \langle \boldsymbol{\nu}^*, \textstyle\sum_a (\pi(a|\cdot) - \pi_t(a|\cdot)) \langle \boldsymbol{\zeta}_t, \boldsymbol{\varphi}(\cdot, a) \rangle \rangle.\end{aligned}$$

where $\boldsymbol{\nu}^* = (1-\gamma)\boldsymbol{\nu}_0 + \gamma \boldsymbol{\Psi}^T \boldsymbol{\lambda}^*$ is the state occupancy measure induced by π^*. This is identical to the regret term for the π-player in the unconstrained RL setting. As is done in the unconstrained RL setting, choosing $\alpha = \mathcal{O}((1-\gamma)\sqrt{\log|\mathcal{A}|/(dT)})$, we get

$$\frac{1}{T}\sum_{t=1}^T \textsc{Reg}_t^\pi \leq \mathcal{O}\left(\frac{1}{1-\gamma}\sqrt{(d\log|\mathcal{A}|)/T}\right)$$

which is sublinear in T. Consequently, choosing T to be at least $\Omega(\frac{d\log|\mathcal{A}|}{(1-\gamma)^2\epsilon^2})$ gives $\frac{1}{T}\sum_{t=1}^T \text{Reg}_t^\pi \leq \epsilon$.

E.3.2. Bounding Regret of ζ-Player

Note that the regret for the ζ-player simplifies to

$$\begin{aligned}\textsc{Reg}_t^\zeta &= g(\boldsymbol{\zeta}_t, \boldsymbol{\lambda}(\boldsymbol{c}_t'), \boldsymbol{w}, \pi_t) - g(\boldsymbol{\zeta}_{\boldsymbol{w}_t}^{\pi_t}, \boldsymbol{\lambda}(\boldsymbol{c}_t'), \boldsymbol{w}, \pi_t) \\ &= \langle \boldsymbol{\zeta}_t - \boldsymbol{\zeta}_{\boldsymbol{w}_t}^{\pi_t}, \boldsymbol{\Phi}^T \boldsymbol{\mu}_{\boldsymbol{\lambda}_t,\pi_t} - \boldsymbol{\lambda}_t \rangle\end{aligned}$$

which has the same form as in the unconstrained case. The proof is essentially the same as the proof in Section D.2 for the unconstrained setting. The only difference is that $D_\zeta \leq \mathcal{O}(\frac{D_w\sqrt{d}}{1-\gamma})$ where $D_w = 1 + \frac{1}{\phi}$. Following the proof, we get

$$\textsc{Reg}_t^\zeta \leq \mathcal{O}\left(\frac{C^* d}{(1-\gamma)\phi}\sqrt{\frac{\log(dnT(\log|\mathcal{A}|)/(\delta\phi))}{n}}\right).$$

E.3.3. Bounding Regret of λ-Player

Using the expression (14), the regret of λ-player simplifies to

$$\begin{aligned}\textsc{Reg}_t^\lambda &= f(\boldsymbol{\zeta}_t, \boldsymbol{\lambda}^*, \boldsymbol{w}_t, \pi_t) - f(\boldsymbol{\zeta}_t, \boldsymbol{\lambda}_t, \boldsymbol{w}_t, \pi_t) \\ &= \langle \boldsymbol{\lambda}^* - \boldsymbol{\lambda}_t, \underbrace{\boldsymbol{\theta}_0 + \gamma\boldsymbol{\Psi}\boldsymbol{v}_{\boldsymbol{\zeta}_t,\pi_t} - \boldsymbol{\zeta}_t + \boldsymbol{\Theta}\boldsymbol{w}_t}_{=\boldsymbol{\xi}_t} \rangle\end{aligned}$$

Following the analysis for the unconstrained setting in Section D.3, we get

$$\frac{1}{T}\sum_{t=1}^T \textsc{Reg}_t^\lambda \leq \mathcal{O}\left(\frac{C^* d^{3/2}}{(1-\gamma)\phi}\sqrt{\frac{\log(dnT(\log|\mathcal{A}|)/(\delta\phi))}{n}}\right) + \epsilon_{\text{opt}}^\lambda(T)$$

E.3.4. Bounding Regret of w-Player

Using expression (13), the regret of w-player simplifies to

$$g(\boldsymbol{\zeta}_t, \boldsymbol{\lambda}(\boldsymbol{c}_t'), \boldsymbol{w}_t, \pi_t) - g(\boldsymbol{\zeta}_t, \boldsymbol{\lambda}(\boldsymbol{c}_t'), \boldsymbol{w}, \pi_t) = \langle \boldsymbol{w}_t - \boldsymbol{w}, \boldsymbol{\tau} - \boldsymbol{\Theta}^T \boldsymbol{\lambda}_t \rangle$$

which is bounded by 0 since the w-player choose $\boldsymbol{w}_t \in D_w \Delta^I$ that minimizes $\langle \cdot, \boldsymbol{\tau} - \boldsymbol{\Theta}^T \boldsymbol{\lambda}_t \rangle$.

E.4. Exact Feasibility

For producing an ϵ-optimal policy that satisfies the constraints exactly, we make the following two-policy feature coverage assumptions.

Assumption F (Two-policy feature coverage)**.** Assume the Slater's condition (Assumption D) holds. Denote by π^*_ϕ an optimal policy for the optimization problem $\mathcal{P}(\boldsymbol{\tau}+\phi\mathbf{1})$. Denote by π^* an optimal policy for $\mathcal{P}(\boldsymbol{\tau})$. Assume that

$$(\boldsymbol{\lambda}^*)^T(\boldsymbol{\Lambda}^\dagger)^2\boldsymbol{\lambda}^* \le C^*, \qquad (\boldsymbol{\lambda}^*_\phi)^T(\boldsymbol{\Lambda}^\dagger)^2\boldsymbol{\lambda}^*_\phi \le C^*.$$

With the assumption above and the linear CMDP assumption (Assumption E), consider running Algorithm 1 with stricter thresholds $\boldsymbol{\tau}+\eta\mathbf{1}$ where $\eta=\phi\epsilon$ and $D_w=\frac{4}{\phi}$. Since the Slater's constant for $\mathcal{P}(\boldsymbol{\tau}+\eta\mathbf{1})$ is $\phi(1-\epsilon)$, following the main analysis gives

$$L_\eta(\pi^*,\bar{\boldsymbol{w}}) \le L_\eta(\bar{\pi},\boldsymbol{w})+\epsilon \tag{18}$$
$$L_\eta(\pi^*_\phi,\bar{\boldsymbol{w}}) \le L_\eta(\bar{\pi},\boldsymbol{w})+\epsilon \tag{19}$$

for any $\boldsymbol{w}\in D_w\boldsymbol{\Delta}^I$ with probability at least $1-\delta$ for sample size $n=\widetilde{\mathcal{O}}\left(\frac{(C^*)^2d^3\log(d/\delta)}{(1-\gamma)^2\phi^2\epsilon^2}\right)$ since changing ϕ to $\phi(1-\epsilon)$ does not affect the order of sample size bound where $L_\eta(\pi,\boldsymbol{w})=J_0(\pi)+\boldsymbol{w}\cdot(\boldsymbol{J}(\pi)-\boldsymbol{\tau}-\eta\mathbf{1})$ is the Lagrangian function for $\mathcal{P}(\boldsymbol{\tau}+\eta\mathbf{1})$. Then following the proof of Theorem 3 in Hong et al. (2023), we can argue as follows.

Near Optimality Setting $\boldsymbol{w}=\mathbf{0}$ in (18) and rearranging, we get

$$\begin{aligned}J_0(\bar{\pi}) &\ge J_0(\pi^*)+\bar{\boldsymbol{w}}\cdot(\boldsymbol{J}(\pi^*)-\boldsymbol{\tau}-\eta\mathbf{1})-\epsilon\\ &\ge J_0(\pi^*)-\eta\|\bar{\boldsymbol{w}}\|_1-\mathcal{O}(\epsilon)\\ &\ge J_0(\pi^*)-\mathcal{O}(\epsilon)\end{aligned}$$

where the second inequality follows by the feasibility of π^* for $\mathcal{P}(\boldsymbol{\tau})$; the last inequality follows by $\eta\|\bar{\boldsymbol{w}}\|_1\le\eta D_w=\mathcal{O}(\epsilon)$. This proves near optimality of $\bar{\pi}$. Now we prove that $\bar{\pi}$ is (exactly) feasible for $\mathcal{P}(\boldsymbol{\tau})$.

Exact Feasibility Define $m=\min_{i\in[I]}(J_i(\bar{\pi})-\tau_i)$. If $m\ge 0$ then $J_i(\bar{\pi})-\tau_i\ge 0$ for all $i=1,\dots,I$ and exact feasibility trivially holds. We only consider the case where $m<0$. Define a mixture policy $\widetilde{\pi}=(1-\zeta)\pi^*+\zeta\pi^*_\phi$ where $\zeta\in(0,1)$ is to be determined later. The mixture policy has the interpretation of first drawing a policy from $\{\pi^*,\pi^*_\phi\}$ with probabilities $(1-\zeta)$ and ζ, then running the drawn policy for the entire trajectory. Since $L_\eta(\cdot,\bar{\boldsymbol{w}})$ is linear, a linear combination of (18) and (19) with coefficients $1-\zeta$ and ζ respectively, we get

$$L_\eta(\widetilde{\pi},\bar{\boldsymbol{w}}) \le L_\eta(\bar{\pi},\boldsymbol{w})+\epsilon.$$

Choosing $\boldsymbol{w}$ such that $w_j=D_w$ for $j=\operatorname{argmin}_{i\in[I]}(J_i(\bar{\pi})-\tau_i)$ and $w_j=0$ for all other indices, we get

$$\begin{aligned}L_\eta(\widetilde{\pi},\bar{\boldsymbol{w}}) &\le J_0(\bar{\pi})+\boldsymbol{w}\cdot(\boldsymbol{J}(\bar{\pi})-\boldsymbol{\tau}-\eta\mathbf{1})+\epsilon\\ &= J_0(\bar{\pi})+D_w(m-\eta)+\epsilon.\end{aligned}$$

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

$$\begin{aligned}L_\eta(\widetilde{\pi},\bar{\boldsymbol{w}}) &= J_0(\widetilde{\pi})+\bar{\boldsymbol{w}}\cdot(\boldsymbol{J}(\widetilde{\pi})-\boldsymbol{\tau}-\eta\mathbf{1})\\ &\ge J_0(\widetilde{\pi})+(\zeta\phi-\eta)\|\bar{\boldsymbol{w}}\|_1.\end{aligned}$$

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

$$J_0(\widetilde{\pi})-J_0(\bar{\pi}) \le D_w(m-\eta)-(\zeta\phi-\eta)\|\bar{\boldsymbol{w}}\|_1+\epsilon. \tag{20}$$

Now, to get a lower bound of $J_0(\widetilde{\pi})-J_0(\bar{\pi})$, let $(\widetilde{\pi}^*,\widetilde{\boldsymbol{w}}^*)$ be a primal-dual solution of $\mathcal{P}(\boldsymbol{\tau}+\zeta\phi\mathbf{1})$. Note that $\mathcal{P}(\boldsymbol{\tau}+\zeta\phi\mathbf{1})$ is feasible by the Slater's condition assumption D and the fact that $\zeta\phi\in(0,\phi)$. Since $(\widetilde{\pi}^*,\widetilde{\boldsymbol{w}}^*)$ is a saddle point of $L_{\zeta\phi}(\pi,\boldsymbol{w})=J_0(\pi)+\boldsymbol{w}\cdot(J_{\boldsymbol{C}}(\pi)-\boldsymbol{\tau}-\zeta\phi\mathbf{1})$, we get

$$L_{\zeta\phi}(\bar{\pi},\widetilde{\boldsymbol{w}}^*) \le L_{\zeta\phi}(\widetilde{\pi}^*,\widetilde{\boldsymbol{w}}^*)=J_0(\widetilde{\pi}^*)\le J_0(\pi^*)\le\frac{1}{1-\zeta}J_0(\widetilde{\pi})$$

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 $\mathcal{P}(\boldsymbol{\tau}+\zeta\phi\mathbf{1})$; and the last inequality follows by $J_0(\widetilde{\pi}) = (1-\zeta)J_0(\pi^*) + \zeta J_0(\pi^*_\phi) \geq (1-\zeta)J_0(\pi^*)$. Rearranging, we get

$$\begin{aligned} J_0(\widetilde{\pi}) - J_0(\bar{\pi}) &\geq -\zeta J_0(\bar{\pi}) + (1-\zeta)\widetilde{\boldsymbol{w}}^* \cdot (\boldsymbol{J}(\bar{\pi}) - \boldsymbol{\tau} - \zeta\phi\mathbf{1}) \\ &\geq -\zeta + (1-\zeta)(m - \zeta\phi)\|\widetilde{\boldsymbol{w}}^*\|_1 \end{aligned}$$

where the second inequality follows by $J_0(\cdot) \leq 1$ and the definition of m. Combining with the upper bound of $J_0(\widetilde{\pi}) - J_0(\bar{\pi})$ shown in (20) and rearranging, we get

$$(D_w - (1-\zeta)\|\widetilde{\boldsymbol{w}}^*\|_1)m \geq D_w\eta + (\zeta\phi - \eta)\|\bar{\boldsymbol{w}}\|_1 - \zeta - (1-\zeta)\zeta\phi\|\widetilde{\boldsymbol{w}}^*\|_1 - \epsilon. \tag{21}$$

Now, we choose our parameters as follows.

$$\zeta = \epsilon, \quad D_w = \frac{4}{\phi}, \quad \eta = \phi\epsilon.$$

Since $\widetilde{\boldsymbol{w}}^*$ is a dual solution of $\mathcal{P}(\boldsymbol{\tau}+\zeta\phi\mathbf{1})$, which has a margin of $\phi - \zeta\phi$, Lemma E.6 gives $\|\widetilde{\boldsymbol{w}}^*\|_1 \leq \frac{1}{\phi-\zeta\phi}$. Hence,

$$\zeta\phi\|\widetilde{\boldsymbol{w}}^*\|_1 \leq \frac{\zeta\phi}{\phi - \zeta\phi} \leq 2\zeta = 2\epsilon$$

where the second inequality uses $\zeta = \epsilon \leq \frac{1}{2}$. Hence, $\|\widetilde{\boldsymbol{w}}^*\|_1 \leq \frac{2\epsilon}{\zeta\phi} = \frac{2}{\phi} < D_w$ so that $D_w - (1-\zeta)\|\widetilde{\boldsymbol{w}}^*\|_1 > 0$. Hence, the previous result (21) gives

$$\begin{aligned} (D_w - (1-\zeta)\|\widetilde{\boldsymbol{w}}^*\|_1)m &\geq D_w\eta + (\zeta\phi - \eta)\|\bar{\boldsymbol{\lambda}}\|_1 - \zeta - (1-\zeta)\zeta\phi\|\widetilde{\boldsymbol{w}}^*\|_1 - \epsilon \\ &\geq 4\epsilon + 0 - \epsilon - 2\epsilon - \epsilon \\ &= 0. \end{aligned}$$

Since $D_w - (1-\zeta)\|\widetilde{\boldsymbol{\lambda}}^*\|_1 > 0$, we have $m \geq 0$ which implies $\tau_i - J_i(\bar{\pi}) \geq 0$ for all $i = 1, \ldots, I$. This leads to the following theorem.

Theorem E.7. *Under assumptions E and F, as long as T is at least $\Omega(\frac{d\log|\mathcal{A}|}{(1-\gamma)^2\epsilon^2})$, the policy $\bar{\pi}$ produced by Algorithm 1 with thresholds $\boldsymbol{\tau} + \phi\epsilon\mathbf{1}$ and $D_w = \frac{4}{\phi}$ satisfies $J_0(\bar{\pi}) \geq J_0(\pi^*) - \epsilon$ and $J_i(\bar{\pi}) \geq \tau_i$, $i = 1, \ldots, I$ with probability at least $1-\delta$ for sample size*

$$n = \mathcal{O}\left(\frac{(C^*)^2 d^3 \log(dn(\log|\mathcal{A}|)/(\delta\phi\epsilon(1-\gamma))}{(1-\gamma)^2\phi^2\epsilon^2}\right).$$