# Revisiting the Linear-Programming Framework for Offline RL with General Function Approximation

**Asuman Ozdaglar** [1] **Sarath Pattathil** [1] **Jiawei Zhang** [1] **Kaiqing Zhang** [2]

## Abstract

Offline reinforcement learning (RL) aims to find an optimal policy for sequential decision-making using a pre-collected dataset, without further interaction with the environment. Recent theoretical progress has focused on developing sample-efficient offline RL algorithms with various relaxed assumptions on data coverage and function approximators, especially to handle the case with excessively large state-action spaces. Among them, the framework based on the linear-programming (LP) reformulation of Markov decision processes has shown promise: it enables sample-efficient offline RL with function approximation, under only *partial* data coverage and *realizability* assumptions on the function classes, with favorable computational tractability. In this work, we revisit the LP framework for offline RL, and provide a new reformulation that advances the existing results in several aspects, relaxing certain assumptions and achieving optimal statistical rates in terms of sample size. Our key enabler is to introduce proper *constraints* in the reformulation, instead of using any regularization as in the literature, also with careful choices of the function classes and initial state distributions. We hope our insights bring into light the use of LP formulations and the induced primal-dual minimax optimization, in offline RL.

## 1. Introduction

Recent years have witnessed tremendous empirical successes of reinforcement learning (RL) in many sequential-decision making problems (Mnih et al., 2015; Silver et al., 2016; Vinyals et al., 2017; Levine et al., 2016). Key to these successes are two factors: 1) use of rich *function approximators*, e.g., deep neural networks; 2) access to excessively *large interaction data* with the environment. Most successful examples above are extremely data-hungry. In some cases, the interaction data can be easily obtained in an online fashion, due to the existence of powerful simulators such as game engines (Silver et al., 2016; Vinyals et al., 2017) and physics simulators (Todorov et al., 2012).

On the other hand, in many other domains of RL, such online interaction is impractical, either because data collection is expensive and/or impractical, or the environment is simply difficult to simulate well. Many real world applications fall into this setting, including robotics and autonomous driving (Levine et al., 2018; Maddern et al., 2017), healthcare (Tseng et al., 2017), and recommender systems (Swaminathan et al., 2017). Moreover, even in the cases where online interaction is available, one might still want to utilize previously collected data, as effective generalization requires *large* datasets (Levine et al., 2020). Offline RL has thus provided a promising framework when one really targets deploying RL in the real-world.

However, in practice, offline RL is known to suffer from the *training instability* issue due to the use of function approximation, e.g., neural networks, and the *distribution shift* issue due to the mismatch between the offline data distribution and the targeted (optimal) policy distribution (Fujimoto et al., 2019; Kumar et al., 2020). As a result sample-efficiency guarentees for offline RL with function approximation usually relies on strong assumptions on both the *function classes* and the *dataset*. In particular, many earlier results (Munos & Szepesvári, 2008; Scherrer, 2014; Chen & Jiang, 2019; Zhang et al., 2021) require the function classes to be *Bellman-complete*, i.e., the value function class is *closed* under the Bellman operator, and the dataset to have *full coverage*, i.e., the data covers the state distributions induced by *all* policies. Both assumptions are strong: the former is *non-monotone* in the function class, i.e., the assumption can be violated when a richer function class is used, and is much stronger than the common assumption of *realizability* (i.e., the optimal solution lies in the function class) in statistical learning theory; the latter essentially re-

Authors are arranged in alphabetical order. [1]Department of EECS, Massachusetts Institute of Technology [2]Department of ECE, University of Maryland, College Park. Correspondence to: Jiawei Zhang <jwzhang@mit.edu>.

*Proceedings of the $40^{th}$ International Conference on Machine Learning*, Honolulu, Hawaii, USA. PMLR 202, 2023. 

quires the offline data to cover all the possible state-action pairs, which is violated in most real-world applications.

Significant progress has been made lately to relax these assumptions. For example, (Liu et al., 2020; Jin et al., 2020; Rashidinejad et al., 2021; Xie et al., 2021; Uehara & Sun, 2021) have shown that using the *pessimistic* mechanism that chooses the worst-cast value function or model in the uncertainty set during learning, the full coverage assumption can be relaxed to only a *single-policy* (i.e., an optimal policy) coverage assumption. Nonetheless, the results all rely on completeness-type (which includes the tabular setting) or even stronger assumptions, and some of the algorithms are not computationally tractable (Xie et al., 2021; Uehara & Sun, 2021). On the other hand, some works require only realizability, but with additionally either stronger (than all-policy coverage) assumptions on data coverage (Xie & Jiang, 2021), or the uniqueness of the optimal policy (Chen & Jiang, 2022), and can be computationally intractable.

More recently, (Zhan et al., 2022) has successfully relaxed *both* the full data coverage and the completeness assumptions, through the seminal use of the linear-programming (LP) reformulation of Markov decision processes (MDPs). The LP framework not only significantly weakens the assumptions, but also better enables computationally tractable algorithms. However, the algorithms and analyses in (Zhan et al., 2022) strongly depend on a *regularized* version of the LP formulation, which calls for stronger assumptions than *single-policy* coverage, and leads to statistically suboptimal rates (i.e., $O(1/n^{1/6})$ where $n$ is the size of the dataset). In this paper, we revisit and further investigate the power of the LP-framework for offline RL with performance guarantees, and advance the existing results in several aspects.

**Contributions.** We propose LP-based offline RL algorithms with optimal (in terms of sample size) $O(1/\sqrt{n})$ sample complexity, under partial data coverage and general function approximation, and without any behavioral regularization as (Zhan et al., 2022). In particular, first, we obtain the $O(1/\sqrt{n})$ optimal rate under the standard single-policy concentrability (SPC) assumption (Rashidinejad et al., 2021), with some completeness-type assumption on the function class. Second, our result leads to the near-optimal rate of $O(\sqrt{|S|}/((1-\gamma)\sqrt{n}))$ when reducing to the tabular case, improving even the state-of-the-art tabular-case result (Rashidinejad et al., 2021). Most general function approximation get a loose bound upon such a reduction. Third, with only the realizability assumption, we obtain $O(1/(\texttt{Gap}\cdot\sqrt{n}))$ rate under a partial coverage assumption that is a slight variant of standard SPC, where Gap denotes the minimal difference between the values of the best action and the second-best one among all states. Finally, note that our algorithms inherit the favorable computational tractability as other LP-based offline RL algorithms (Zhan et al., 2022; Rashidinejad et al., 2022). Inspired by novel error bounds, our techniques involve adding validity constraints of the occupancy measure in the first case, and a lower bound on the density ratio in the second case.

A more detailed literature review, a summary of our techniques, as well as the notation we use in this paper are given in Appendix A. Furthermore, we have relegated the proofs of some results to the appendix due to space limitation.

# 2. Background

## 2.1. Model and Setup

**Markov Decision Processes.** Consider an infinite-horizon MDP characterized by a tuple $\langle S, A, P, R, \gamma, \mu_0\rangle$, where $S=\{s^1,\cdots,s^{|S|}\}$ and $A=\{a^1,\cdots,a^{|A|}\}$ denote the state and action spaces of the agent, $R: S\times A\to[0,1]$ is the reward function[1], $P: S\times A\to\Delta(S)$ denotes the transition kernel, $\gamma\in[0,1)$ denotes the discount factor, and $\mu_0\in\Delta(S)$ denotes the initial state distribution. We assume $S$ and $A$ are finite (but potentially very large), in order to ease the notation. However, our results later do not depend on the cardinalities of $S$ and $A$. Let $\pi: S\to\Delta(A)$ denote a Markov stationary policy of the agent, determining the distribution over actions at each state. Each $\pi$ leads to a *discounted occupancy* measure over the state-action spaces, denoted by

$$\theta_{\pi,\mu_0}(s,a)=(1-\gamma)\sum_{t=0}^{\infty}\gamma^t\mathbb{P}_\pi(s_t=s,a_t=a;\mu_0), \quad (1)$$

where $\mathbb{P}_\pi(s_t=s,a_t=a;\mu_0)$ is the probability of the event of visiting the pair $(s,a)$ at time $t$ under the policy $\pi$, starting from $s_0\sim\mu_0(\cdot)$.

Correspondingly, with a slight abuse of notation, we use $\theta_{\pi,\mu_0}(s)=\sum_{a\in A}\theta_{\pi,\mu_0}(s,a)$ to denote the *discounted occupancy measure over states*. For notational convenience, we concatenate the state-action occupancy measure $\theta_{\pi,\mu_0}(s,a)$ in a vector $\theta_{\pi,\mu_0}$, defined as

$$\theta_{\pi,\mu_0}=\big[\theta_{\pi,\mu_0}(s^1,a^1),\cdots,\theta_{\pi,\mu_0}(s^1,a^{|A|}),\cdots,\\ \theta_{\pi,\mu_0}(s^{|S|},a^1),\cdots,\theta_{\pi,\mu_0}(s^{|S|},a^{|A|})\big]^\top\in\mathbb{R}_+^{|S||A|}. \quad (2)$$

Given any policy $\pi$, one can then define the corresponding state-action and state value functions, $Q_\pi$ and $v_\pi$, as

[1]Note that we stick to the case of deterministic reward for ease of presentation. Our results can be readily extended to the case of random rewards.

follows:

$$Q_\pi(s,a) = \mathbb{E}_{s_{t+1}\sim P_{s_t,a_t}(\cdot),a_t\sim\pi(\cdot|s_t)}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)\middle| s_0 = s, a_0 = a\right],$$

where the trajectory is generated following the policy $\pi$, and $v_\pi(s) = \mathbb{E}_{a\sim\pi(\cdot|s)}[Q_\pi(s,a)]$. The overall goal is to find a policy $\pi^*$ that solves the following problem:

$$\max_\pi \quad J_{\mu_0}(\pi) := (1-\gamma)\cdot\mathbb{E}_{s\sim\mu_0}\big[v_\pi(s)\big], \tag{3}$$

where $J_{\mu_0}(\pi)$ denotes the *return* under $\pi$ and $\mu_0$, i.e., the $(1-\gamma)$-times expected value function under policy $\pi$ and initial distribution $\mu_0$. Note that $J_{\mu_0}(\pi)$ can also be equivalently written as $J_{\mu_0}(\pi) = r^\top\theta_{\pi,\mu_0}$, where

$$r = [r(s^1,a^1),\cdots,r(s^{|S|},a^{|A|})]^\top \in [0,1]^{|S||A|}. \tag{4}$$

It is known that the optimal solution to the MDP is a Markov Stationary policy. For a general distribution $\rho \in \Delta(S)$, we use $\theta_{\pi,\rho}$ and $J_\rho(\pi)$ to denote the discounted occupancy measure and the average value function under policy $\pi$, but starting from the initial distribution $\rho$. We sometimes just write $\theta_\pi$ and $J(\pi)$ for simplicity, when the initial distribution is clear from context.

Note that the optimal policy $\pi^*$ may not be unique. We define $v^* = v_{\pi^*}$ and $Q^* = Q_{\pi^*}$. For convenience, we sometimes denote $m = |S||A|$.

**Offline RL.** Consider an offline RL problem, where one has collected a dataset $\mathcal{D}$ containing $n$ samples drawn from some distribution. Suppose $\mathcal{D} = \{(s_i,a_i,s_i',r_i)\}_{i=1}^n$, where the independent and identically distributed (i.i.d.) samples $(s_i,a_i)$ are drawn from some distribution $\mu(\cdot,\cdot)$. We let $\mu(s) = \sum_a \mu(s,a)$ which implies that $s_i$ are drawn i.i.d. from the distribution $\mu(\cdot)$. We denote the conditional distribution of $a$ given $s$ induced from $\mu$ as $\pi_\mu(a \mid s)$, i.e., $\pi_\mu(a \mid s) = \mu(s,a)/\mu(s)$ if $\mu(s) > 0$; and $\pi_\mu(\cdot \mid s)$ can be defined as any distribution in $\Delta(A)$, e.g., a uniform one with $\pi_\mu(a \mid s) = 1/|A|$, if $\mu(s) = 0$. $\pi_\mu$ can also be defined as the *behavior policy* if $\mu$ happens to correspond to the occupancy measure of some policy.

In this paper, we assume that the behavior policy $\pi_\mu(a \mid s)$ is known, as in (Zhan et al., 2022; Rashidinejad et al., 2022). We provide extensions of our algorithms when the behavior policy is not known in Appendix D. Given a state-action pair $(s_i,a_i)$, we have $r_i = r(s_i,a_i)$ and $s_i' \sim P_{s_i,a_i}(\cdot)$. Moreover, let $n_\mathcal{D}(s,a)$ be the subset of the sample indices $\{1,\cdots,n\}$ that includes the indices of the samples in $\mathcal{D}$ that visit state-action pair in the sense of $(s_i,a_i) = (s,a)$. Similarly, we use $n_\mathcal{D}(s,a,s')$ and $n_\mathcal{D}(s)$ to denote the sets of indices of data samples in $\mathcal{D}$ such that $(s_i,a_i,s_i') = (s,a,s')$ and $s_i = s$, respectively. We define the empirical version of $\mu$, i.e., $\mu_\mathcal{D}$, as $\mu_\mathcal{D}(s,a) = n_\mathcal{D}(s,a)/n$. The goal of offline RL is to make use of the dataset $\mathcal{D}$ to learn a policy $\hat{\pi}$, such that the *optimality gap* $J_{\mu_0}(\pi^*) - J_{\mu_0}(\hat{\pi})$ is small.

**Partial data coverage.** Throughout the paper, we consider the scenario where the offline data only has *partial* coverage, instead of a *full* one. To illustrate the difference, we first introduce the following definition of *policy concentrability*.

**Definition 1** (Policy Concentrability)**.** *For any policy $\pi$ and initial state distribution $\rho$, and the given offline data distribution $\mu$, we define $C_{\pi,\rho} > 0$ to be the policy concentrability coefficient, which is the smallest upper-bound such that $\frac{\theta_{\pi,\rho}(s,a)}{\mu(s,a)} \le C_{\pi,\rho}$ for all $(s,a) \in S \times A$.*

Note that $C_{\pi,\rho}$ characterizes how well the trajectory generated by the policy $\pi$ starting from some $\rho$ is covered by the offline data. In earlier offline RL literature, it is usually assumed that the data has *full coverage*: there exists a constant $C$ that upper bounds $C_{\pi,\mu_0}$ for *all* policy $\pi$ (Munos & Szepesvári, 2008; Scherrer, 2014). In contrast, when we choose $\pi = \pi^*$, this leads to a single optimal policy concentrability assumption that is much weaker than the full one. We will focus on this partial coverage setting, and specify this assumption later.

### 2.2. LP-based Reformulations

It is known that for tabular MDPs, any optimal policy $\pi^*$ optimizes $J_\rho(\pi)$ starting from any distribution $\rho \in \Delta(S)$ (including the actual initial distribution $\rho = \mu_0$ in the model in Section 2.1) (Puterman, 1994), as it simultaneously maximizes $v_\pi(s)$ for all states $s \in S$. Moreover, the optimality condition of the MDP when starting from any distribution $\rho$ can also be written as the following linear program (Puterman, 1994):

$$\begin{aligned} \min_v \quad & (1-\gamma)\rho^\top v \\ \text{s.t.} \quad & \gamma P_{(s,a)}^\top v + r(s,a) \le v(s), \quad \forall s \in S,\ a \in A, \end{aligned} \tag{5}$$

where $P_{(s,a)} = [P_{s,a}(s^1),\cdots,P_{s,a}(s^{|S|})]^\top \in \Delta(S)$ is the vector of state transition probabilities for the state-action pair $(s,a)$. Let $P = [P_{(s^1,a^1)},\cdots,P_{(s^1,a^{|A|})},\cdots,P_{(s^{|S|},a^1)},\cdots,P_{(s^{|S|},a^{|A|})}] \in \mathbb{R}^{|S|\times m}$ and $\mathbf{1}_{|A|} = [1,1,\cdots,1]^\top \in \mathbb{R}^{|A|}$.

Note that we keep the initial-state distribution used in the LP (5) to be $\rho$ (instead of $\mu_0$) for generality, which does not affect the solution in the tabular case. However, as we will specify in later sections, the choice of $\rho$ can make a difference in the function approximation setting, and may help address some challenging settings in offline

RL. Interestingly, such a distinction has already been observed and studied in the linear function approximation case (De Farias & Van Roy, 2003) in the context of approximate dynamic programming. The corresponding dual formulation of the LP (5) can be written as follows:

$$\begin{array}{ll} \max_\theta & r^\top \theta := \sum_{s\in S, a\in A} r(s,a)\cdot\theta(s,a) \\ \text{s.t.} & M\theta = (1-\gamma)\rho, \quad \theta \geq 0, \end{array} \tag{6}$$

where the matrix $M$ is defined as: $M := \mathrm{Diag}(\mathbf{1}_{|A|}^\top, \cdots, \mathbf{1}_{|A|}^\top) - \gamma P$. Note that the optimal solution of the dual problem corresponds to the discounted occupancy measure of an optimal policy (see (Puterman, 1994)). Hence, we use the notation $\theta$ to denote the optimization variable of the dual problem.

We focus on solving the dual formulation (6) in this paper. Then, the optimal $\theta^*$ can be used to generate a policy $\pi_{\theta^*}$, where $\pi_\theta$ is defined as

$$\pi_\theta(a \mid s) = \frac{\theta(s,a)}{\sum_{a'\in A}\theta(s,a')}, \tag{7}$$

if $\sum_{a'\in A}\theta(s,a') > 0$; and $\pi_\theta(\cdot \mid s)$ can be defined as any distribution in $\Delta(A)$, e.g., a uniform one with $\pi_\theta(a \mid s) = 1/|A|$, if $\sum_{a'\in A}\theta(s,a') = 0$. This $\pi_{\theta^*}$ then corresponds to an optimal policy $\pi^*$ of the MDP (Puterman, 1994).

To better study the relationship between the occupancy measure and the data distribution, we also consider the scaled version of the LP. This is also referred to as the marginal importance sampling formulation of the MDP in the literature (Nachum et al., 2019; Lee et al., 2021; Zhan et al., 2022). First, we define $w \in \mathbb{R}_+^m$ such that $w(s,a)\mu(s,a) = \theta(s,a)$, i.e., $w(s,a)$ denotes the ratio between the occupancy measure of the target policy and the offline data distribution.

For each $(s,a,s') \in S \times A \times S$, let $K_{s',(s,a)} \in \mathbb{R}^{|S|\times m}$ be a matrix satisfying $K_{s',(s,a)}(s,(s,a)) = 1$, $K_{s',(s,a)}(s',(s,a)) = -\gamma$ and all other entries are zeros. Define the distributions $\nu$ and $\nu_{\mathcal{D}}$ over $S \times A \times S$ as follows: $\nu(s,a,s') := P_{s,a}(s')\mu(s,a)$ and $\nu_{\mathcal{D}}(s,a,s') := |n_{\mathcal{D}}(s,a,s')|/n$. Finally, we also define the matrices

$$K = \mathbb{E}_{(s,a,s')\sim\nu} K_{s',(s,a)}, \quad K_{\mathcal{D}} = \mathbb{E}_{(s,a,s')\sim\nu_{\mathcal{D}}} K_{s',(s,a)}. \tag{8}$$

Furthermore, we define $u \in \mathbb{R}^m$ such that $u(s,a) := r(s,a)\mu(s,a)$. Then, we have the following lemma which relates these quantities to the ones in Problem (6).

**Lemma 1.** *We have $u^\top w = r^\top \theta$ and $Kw = M\theta$.*

*Proof.* Note that the first inequality directly follows from the definitions of $u$ and $w$.

The second equality can be derived as follows. Let $K(s',(s,a))$ and $M(s',(s,a))$ denote the $(s',(s,a))$-th element of the matrices $K$ and $M$, respectively. Note that $K(s',(s,a)) = M(s',(s,a)) \cdot \mu(s,a)$ for all $(s,a,s') \in S \times A \times S$. Now:

$$\begin{aligned} [Kw]_s &= \sum_{(\tilde{s},\tilde{a})} K(s,(\tilde{s},\tilde{a}))w(\tilde{s},\tilde{a}) \\ &= \sum_{(\tilde{s},\tilde{a})} M(s,(\tilde{s},\tilde{a}))\mu(\tilde{s},\tilde{a})w(\tilde{s},\tilde{a}) = [M\theta]_s \end{aligned}$$

thereby completing the proof. □

Using Lemma 1, we can rewrite Problem (6) as follows:

$$\max_{w\geq 0} u^\top w \qquad \text{s.t.} \quad Kw = (1-\gamma)\rho. \tag{9}$$

Let $w^*$ be the solution to (9), then we can obtain the optimal policy by computing $\pi_{w^*}$, where with a slight abuse of notation, $\pi_w$ is defined as

$$\pi_w(a \mid s) := \begin{cases} \frac{w(s,a)\pi_\mu(a|s)}{\sum_{a'\in A} w(s,a')\pi_\mu(a'|s)}, & \text{if} \quad c > 0 \\ \frac{1}{|A|} & \text{if} \quad c = 0 \end{cases} \tag{10}$$

where $c := \sum_{a'\in A} w(s,a')\pi_\mu(a' \mid s)$. We recall that $\pi_\mu$ is the conditional distribution of $a$ given $s$ under $\mu$, which can also be viewed as the behavior policy.

The equivalent primal-dual minimax reformulation of Problem (9) is given by:

$$\min_{w\in\mathbb{R}_+^m} \max_{v\in\mathbb{R}^{|S|}} \quad -u^\top w + v^\top(Kw - (1-\gamma)\rho). \tag{11}$$

Throughout the paper, we define

$$\ell(w,v) := -u^\top w + v^\top(Kw - (1-\gamma)\rho). \tag{12}$$

### 2.3. Empirical Formulation

Since we do not have access to the exact distributions in the RL setting, we cannot solve Problem (11) directly. Let $\hat{\rho}$ be an empirical estimate of $\rho$ (we can use $\hat{\rho} = \rho$ if $\rho$ is known to us). We thus define the following empirical counterpart of (12):

$$\ell_{\mathcal{D}}(w,v) := -u_{\mathcal{D}}^\top w + v^\top(K_{\mathcal{D}}w - (1-\gamma)\hat{\rho}), \tag{13}$$

where we recall the definition of $K_{\mathcal{D}}$ in (8), and define $u_{\mathcal{D}} \in \mathbb{R}^m$ as $u_{\mathcal{D}}(s,a) = r(s,a)\mu_{\mathcal{D}}(s,a)$, with $\mu_{\mathcal{D}}(s,a) = n_{\mathcal{D}}(s,a)/n$. We will then focus on the following empirical minimax optimization problem:

$$\min_{w\in\mathbb{R}_+^m} \max_{v\in\mathbb{R}^{|S|}} \quad \ell_{\mathcal{D}}(w,v). \tag{14}$$

Finally, we also give a brief introduction to Function Approximation next.

### 2.4. Function Approximation

To handle massively large state and action spaces, function approximation is usually used for the decision variables when solving MDPs, e.g., for the value functions as well as the density ratios when one uses the LP framework as in (9) or (11). Note that for finite state-action spaces, the variables $v$ and $w$ are real vectors of dimensions $|S|$ and $|S||A|$, respectively. Following the convention in the literature (Chen & Jiang, 2019; 2022; Zhan et al., 2022; Rashidinejad et al., 2022), we will refer to $v$ and $w$ as *functions*, i.e., $v : S \to \mathbb{R}$ and $w : S \times A \to \mathbb{R}$. We will then use function classes $V$ and $W$ to approximate the functions $v$ and $w$, which usually have much smaller cardinality/covering number than the whole function space. The same convention also applies to other vectors of dimensions $|S|$ and/or $|S||A|$. We now introduce the following relationship of *completeness* between two function classes, which will be used later in the analysis.

**Definition 2** (Completeness)**.** *For two function classes $\mathcal{F}$ and $\mathcal{G}$, and a mapping $\phi : \mathcal{F} \to \mathcal{G}$, we say they satisfy* $(\mathcal{F}, \mathcal{G})$-completeness under $\phi$, *if for all $f \in \mathcal{F}$, $\phi(f) \in \mathcal{G}$.*

Note that the common notion of *Bellman-completeness* corresponds to the case where $\mathcal{F} = \mathcal{G}$ is the function class for approximating value functions, and $\phi$ is the Bellman operator (Bertsekas, 2017).

In the following two sections, we propose two offline RL algorithms with function approximation to different variables. In Section 3, we relax the equality constraint in (9) to some inequality constraints and use function approximation to $w$ and another additional variable. Under some completeness assumption, we prove that our algorithm achieves optimal sample complexity in terms of sample size, and even improves the state-of-the-art results specialized to the tabular case. In Section 4, we use function approximation in the minimax problem (14) in its original form, and achieve the $1/\sqrt{n}$ rate that also depends on the gap function of $Q$ (see definition in Section 4), with only the realizability assumption of the function classes.

## 3. Case I: Optimal Rate with Completeness-type Assumption

We first solve offline RL with an optimal $O(1/\sqrt{n})$ sample complexity using the LP formulation, under single-policy concentrability and some completeness-type assumptions. Throughout this section, we choose the distribution $\rho$ in the LP reformulations in Section 2.2 to be the initial state distribution $\mu_0$.

Before proceeding further, we need some additional properties on the relationship between occupancy measure $\theta$ and the induced policy $\pi_\theta$, as shown next.

### 3.1. Properties of the Induced Policy $\pi_\theta$

Recall the definition of the occupancy measure induced by policy $\pi$ as $\theta_{\pi,\mu_0}$ Note that for simplicity, we may omit the subscript $\mu_0$ in $\theta_{\pi,\mu_0}$ throughout this section, as the initial distribution considered here is only $\mu_0$, and should be clear from the context. Notice that a vector $\theta \in \mathbb{R}^m_+$ is not necessarily an occupancy measure of any policy $\pi$. The first lemma below shows that $\theta$ is an occupancy measure if it satisfies the constraints in Problem (6) with $\rho = \mu_0$.

**Lemma 2.** *If some $\theta \in \mathbb{R}^m$ satisfies $\theta \geq 0$ and $M\theta = (1-\gamma)\mu_0$, we have $\theta = \theta_{\pi_\theta}$, where we recall the definition of $\pi_\theta$ in* (7). *Moreover, in this case, we have $J_{\mu_0}(\pi_\theta) = r^\top \theta$.*

This lemma is a special case of the next lemma and the results in Section 6.9 of (Puterman, 1994). The next question is how close $\theta$ is to $\theta_{\pi_\theta}$ if $\theta$ is not in the set $\{\theta \mid M\theta = (1-\gamma)\mu_0, \theta \geq 0\}$, i.e., it does not satisfy the constraints. The following lemma provides an error bound that relates the occupancy measure constraint violation and the absolute difference between $r^\top\theta$ and $r^\top\theta_{\pi_\theta}$.

**Lemma 3.** *For any $\theta \geq 0$, we have $|r^\top(\theta - \theta_{\pi_\theta})| \leq \frac{\|M\theta-(1-\gamma)\mu_0\|_1}{1-\gamma}$.*

Note that the term $r^\top\theta_{\pi_\theta}$ in Lemma 3 exactly corresponds to $J(\pi_\theta)$. Next, we introduce the following definition.

**Definition 3** (Sign Function)**.** *For any $w \in \mathbb{R}^m_+$, we define the mapping $\phi : \mathbb{R}^m \to \mathbb{R}^{|S|}$ such that $\phi(w) \in \arg\max_{x:\|x\|_\infty \leq 1} x^\top(Kw-(1-\gamma)\mu_0)$ as the* sign function *of the occupancy validity constraint $Kw - (1-\gamma)\mu_0 = 0$ in* (9). *In particular, note that $\phi(w)^\top(Kw - (1-\gamma)\mu_0) = \|Kw - (1-\gamma)\mu_0\|_1$.*

By the definition of dual norm, we refer to $\phi(w)$ as the sign function, where we follow the convention that the sign of 0 can be any arbitrary $x$ with $\|x\|_\infty \leq 1$. We are now ready to state our assumption on function classes.

**Assumption 1.** *Let $x_w := \phi(w)$ with $\phi$ given in Definition 3. Let $W$ and $B$ be the function classes for $w$ and $x_w$, respectively. Then, we have realizability of $W$, and $(W, B)$-completeness under $\phi$, i.e., $w^* \in W$ and $x_w \in B$ for all $w \in W$. Furthermore, we assume that $W$ and $B$ are bounded, i.e., $\|w\|_\infty \leq B_w$ for all $w \in W$ and*[2] *$\|x\|_\infty \leq 1$ for all $x \in B$.*

We discuss this assumption in detail in Appendix B.1. Next, we make the following SPC assumption, as in (Rashidinejad et al., 2021; 2022). Note that it is made for the original MDP we would like to solve, and is weaker than the policy concentrability assumption in (Zhan et al., 2022), which additionally includes the concentrability assumption for some regularized problem.

[2]Note that $w^* \in W$ implies that $B_w \geq 1$ since $w^*$ is a supremum of the ratio between two distributions.

**Assumption 2** (Single-Policy Concentrability)**.** *There exists some constant $C^* > 0$ such that for all $(s,a) \in S \times A$, $\frac{\theta_{\pi^*,\mu_0}(s,a)}{\mu(s,a)} \le C_{\pi^*,\mu_0} = C^*$ for the optimal policy $\pi^* = \pi_{w^*}$, where $w^*$ is given in Assumption 1* [3].

### 3.2. A Reformulated LP

Now we can state our approach. From Lemma 3, we know that if we control $\|M\theta - (1-\gamma)\mu_0\|_1 = \|Kw - (1-\gamma)\mu_0\|_1$, we can make the inner product $r^\top\theta$ be close to the actual reward under the policy $\pi_\theta$, i.e., $r^\top\theta_{\pi_\theta}$. This motivates us to add a constraint to control $\|Kw - (1-\gamma)\mu_0\|_1$.

Recall that $u_\mathcal{D} \in \mathbb{R}^m$ is defined as $u_\mathcal{D}(s,a) = r(s,a)\mu_\mathcal{D}(s,a)$. Our approach is to solve the following LP-based optimization problem constructed from the dataset $\mathcal{D}$:

$$\begin{aligned} \max_{w\in W} \quad & u_\mathcal{D}^\top w \\ \text{s.t.} \quad & x^\top(K_\mathcal{D}w - (1-\gamma)\mu_0) \le E_{n,\delta}, \forall x \in B \end{aligned} \tag{15}$$

where $E_{n,\delta} := \frac{B_w\sqrt{2\log(|B||W|/\delta)}}{\sqrt{n}}$. Program (15) can be viewed as a relaxation of the empirical version of Problem (9), by relaxing the constraint $K_\mathcal{D}w - (1-\gamma)\mu_0 = 0$ to an inequality.

Suppose that we have a solution to Problem (15), denoted by $w_\mathcal{D}$, we can obtain the policy $\pi_\mathcal{D}$ by setting $\pi_\mathcal{D} = \pi_{\tilde{\theta}_\mathcal{D}}$, where for each $(s,a) \in S \times A$

$$\tilde{\theta}_\mathcal{D}(s,a) = w_\mathcal{D}(s,a)\pi_\mu(a \mid s). \tag{16}$$

The performance of the policy $\pi_\mathcal{D}$ is given in the following theorem:

**Theorem 1.** *Suppose Assumptions 1 and 2 hold. Then, we have, with probability at least $1-6\delta$,*

$$J_{\mu_0}(\pi^*) - J_{\mu_0}(\pi_\mathcal{D}) \le \frac{2\sqrt{2}B_w\sqrt{\log(|B||W|/\delta)}}{(1-\gamma)\sqrt{n}}.$$

Note that Theorem 1 gives an optimal sample complexity of $O(1/\sqrt{n})$ in terms of sample size $n$, with general function approximation. Compared to the recent work (Zhan et al., 2022) that also uses the LP framework for offline RL, we exchange the realizability assumption on $v^*$ therein for some completeness-type assumption, while improving the sample complexity from $O(1/n^{1/6})$ to $O(1/\sqrt{n})$. Compared to other offline RL algorithms with general function approximation that have the optimal $O(1/\sqrt{n})$ rate, e.g., (Xie et al., 2021; Uehara & Sun, 2021; Chen & Jiang, 2022), which are computationally intractable, our algorithm is tractable if the function classes are convex, inheriting the computational advantage of the LP framework for offline RL (Zhan et al., 2022; Rashidinejad et al., 2022).

[3]Note that this implies $C^* \le B_w$.

Finally, compared with the independent work (Rashidinejad et al., 2022), we both require the realizability of $W$ for $w^*$ and some completeness-type assumptions (we need one such assumption while they need two, which may not be comparable as the function classes used are different). Moreover, we do not need the realizability of $v^*$ and have better $(1-\gamma)^{-1}$ dependence ($(1-\gamma)^{-1}$ v.s. $(1-\gamma)^{-3}$), with a relatively simpler algorithm and analysis. We note that the key to obtain $O(1/\sqrt{n})$ rate in (Rashidinejad et al., 2022) is also to enforce the constraint of $w_\mathcal{D}$, where they use the technique of augmented Lagrangian, while we introduce a constrained program directly.

**Remark 1.** *Note that when $W, B$ are continuous sets, Theorem 1 can still be true if we replace the cardinality by the covering number or the number of extreme points. Then, if $W$ and $B$ are convex, our algorithm is computationally tractable since it is just solving a convex program.*

### 3.3. The Tabular Case

In this subsection, we show that our results above can be directly reduced to the tabular case, maintaining the optimal $O(1/\sqrt{n})$ sample complexity.

Here we need realizable function classes $W$ and $B$. To make the algorithm computationally tractable, we use continuous and convex function classes $W$ and $B$, instead of discrete, finite ones. In particular, let $W = \{w \in \mathbb{R}^m_+ \mid \sum_a w(s,a) \le B_w, \forall s \in S\}$ and $B = [-1,1]^{|S|}$, which are convex and compact, and satisfy the boundedness assumptions in Assumption 1. Then we solve the following:

$$\begin{aligned} \max_{w\in W} \quad & u_\mathcal{D}^\top w, \\ \text{s.t.} \quad & \max_{x\in B} x^\top(K_\mathcal{D}w - (1-\gamma)\mu_0) = \|K_\mathcal{D}w - (1-\gamma)\mu_0\|_1 \\ & \le \frac{B_w\sqrt{|S|}\log(2|A|+2)\log(1/\delta)}{\sqrt{n}}. \end{aligned} \tag{17}$$

We have the following theorem:

**Theorem 2.** *Suppose Assumption 2 holds, and the MDP is non-degenerate in the sense that $\min\{|A|,|S|\} > 1$. Then, we have, with probability $\ge 1-6\delta$, for any $\delta < 1/3$:*

$$J_{\mu_0}(\pi^*) - J_{\mu_0}(\pi_\mathcal{D}) \le \frac{2B_w\sqrt{|S| \cdot \big(\log(2|A|+2)\log(1/\delta)\big)}}{(1-\gamma)\sqrt{n}}.$$

Compared to the concurrent and most related result (Rashidinejad et al., 2022), our result yields a better sample complexity when reduced to the tabular case. In particular, (Rashidinejad et al., 2022) leads to a $\tilde{O}(1/((1-\gamma)^4\sqrt{n}))$ rate[4], while we have $\tilde{O}(1/((1-\gamma)\sqrt{n}))$. In fact, our reduction is comparable to and even better than

[4]Note that we did not specify the dependence on $|S|$ explicitly here, since it depends on the function classes being used in (Rashidinejad et al., 2022) (which are different from ours and not comparable), which we believe can be of order $\sqrt{|S|}$ as ours. We

(in terms of $(1-\gamma)$) the state-of-the-art result specifically for the *tabular* case (Rashidinejad et al., 2021), which is[5] $\tilde{O}(\sqrt{|S|}/((1-\gamma)^{3/2}\sqrt{n}))$, while ours is $\tilde{O}(\sqrt{|S|}/((1-\gamma)\sqrt{n}))$. Finally, note that the lower-bound for the tabular case is $\Omega(\sqrt{|S|}/(\sqrt{(1-\gamma)n}))$, in (Rashidinejad et al., 2021) and we believe that using the Bernstein's (instead of Hoeffding's) concentration inequality together with some variance reduction technique (Sidford et al., 2018) may further improve our dependence on $1/(1-\gamma)$, and attain the lower bound.We leave these directions of improvement as future work, since our focus is on the general function approximation case, and on the optimality of sample complexity in terms of $n$.

Though attaining the $O(1/\sqrt{n})$ rate under the LP framework, the results above still rely on some completeness-type assumption. This naturally raises the question:

*Can we have tractable offline RL algorithms with $O(1/\sqrt{n})$ rate but only realizability and partial data coverage assumptions?*

which was the open question left in the literature (Zhan et al., 2022; Chen & Jiang, 2022). Next, we provide an answer to this question, using the LP framework.

## 4. Case II: Realizability-only with Gap Dependence

In this section, we solve the LP-induced minimax optimization (11), by introducing function approximation to the variables $v$ and $w$. Notice that such a setting is also considered in (Zhan et al., 2022) (see Section 4.5 therein). It is also related to the setting in (Chen & Jiang, 2022), where function approximation was used for the state-action function $Q$ and $w$.

We select $w, v$ from finite sets[6] $W, V$. Throughout this section, we sometimes write $\theta_{\pi,\rho}$ as $\theta_\pi$ for notational convenience. Also, we specify the $\rho$ in the LP formulation (9) and (11) as $\rho(s) = \mu(s)$ for all $s \in S$ throughout this section, unless otherwise noted. This assumption enables us to use a crucial lower bound in our problem formulation (See Formulation (18)). At the end of this section, we will relate the return $J_\rho(\pi)$ back to $J_{\mu_0}(\pi)$. We note that we have access to the empirical version $\mu_{\mathcal{D}}$ of $\mu$, where $\mu_{\mathcal{D}}(s,a) = n_{\mathcal{D}}(s,a)/n$.

thus only focus on the dependence of $1/(1-\gamma)$ and $n$. Also, the additional $(1-\gamma)^{-1}$ comes from that $B_v$ therein is of order $(1-\gamma)^{-1}$.

[5]Note that the definition of $J_{\mu_0}$ in (Rashidinejad et al., 2021) is $(1-\gamma)$-factor off from our definition.

[6]As in several related works (Zhan et al., 2022; Rashidinejad et al., 2022), in the case they are infinite classes, we can replace the results in this section with a standard covering argument.

The next proposition specifies a lower bound of $\theta_\pi$:

**Proposition 1.** *For any optimal policy $\pi^*$ and any initial state distribution $\rho \in \Delta(S)$, we have $\sum_{a\in A}\theta_{\pi^*,\rho}(s,a) \geq (1-\gamma)\cdot\rho(s)$ for all $s \in S$.*

This is a direct corollary of the fact that for any policy $\pi$, by definition we have $\sum_{a\in A}\theta_{\pi,\rho}(s,a) \geq (1-\gamma)\sum_{a\in A}\rho(s,a) = (1-\gamma)\rho(s)$ for all $s \in S$. In particular, we would like to note that Proposition 1 is true for the initial state distribution $\rho(s) = \mu(s)$.

The design of algorithms in this section is based on the following intuitive idea: According to Equation (10), for the $w$ such that $\sum_a w(s,a)\pi_\mu(a \mid s) = 0$, the policy $\pi_w$ has to be assigned randomly (as a uniform distribution for example), and cannot be decided from the offline data. To avoid this case, one direct approach is to add a *lower bound* constraint to the vanilla minimax problem (11). Specifically, we consider the following population minimax problem:

$$\begin{aligned} &\min_{w\in\mathbb{R}^m_+}\max_{v\in\mathbb{R}^{|S|}} -u^\top w + v^\top(Kw-(1-\gamma)\mu) \\ &\text{s.t.} \quad \textstyle\sum_a w(s,a)\pi_\mu(a\mid s) \geq (1-\gamma), \quad \forall s\in S. \end{aligned} \tag{18}$$

Note that compared to the vanilla minimax problem (11), the only difference is that we enforce the lower bound constraints on $\sum_a w(s,a)\pi_\mu(a \mid s)$. This lower bound constraint, along with the upper bound shown in Lemma 6, will help control the probability of choosing an inactive state-action pair by the policy generated by the solution of (18). Furthermore, by Proposition 1, we know that the optimal solution $w^*$ is not eliminated by adding the lower bound constraints. Then we turn to solving (18) using function approximation, i.e., our algorithm is to solve the following program:

$$\min_{w\in W}\max_{v\in V} -u^\top w + v^\top(Kw-(1-\gamma)\mu), \tag{19}$$

where $W$ is defined such that for all $w \in W$ we have

$$\sum_{a\in A} w(s,a)\pi_\mu(a\mid s) \geq (1-\gamma), \forall s\in S. \tag{20}$$

Notice that this does not conflict with the constraint of $w \in \mathbb{R}^m_+$ in (18), as we can intersect the sets corresponding to these two constraints when defining $W$.

To learn an approximate optimal policy from the offline data, we solve the following empirical version of the minimax problem in (19):

$$\begin{aligned} &\min_{w\in W}\max_{v\in V} -u_{\mathcal{D}}^\top w + v^\top(K_{\mathcal{D}}w-(1-\gamma)\mu_{\mathcal{D}}), \\ &\text{s.t. } W \text{ satisfies (20).} \end{aligned} \tag{21}$$

### 4.1. Assumptions

Before moving to the main theoretical result in this section, we first state our assumptions and some additional notation.

We first make the realizability assumptions for the function classes $W$ and $V$.

**Assumption 3** (Realizability and Boundedness of $W$). *There exists some solution $w^* \in W \subseteq \mathbb{R}_+^m$ solving* (18) *and hence solving* (11). *Moreover, we suppose $\|w\|_\infty \le B_w$ for all $w \in W$.*

**Assumption 4** (Realizability and Boundedness of $V$). *Suppose that $v^* \in V \subseteq [-1/(1-\gamma), 1/(1-\gamma)]^{|S|}$.*

Notice that similar assumptions are used in (Zhan et al., 2022; Chen & Jiang, 2022). Next, we make the assumption regarding data coverage, which suggests that the offline data should cover some single optimal policy. For ease of presentation, we use the following definitions.

**Definition 4.** *We denote by $S_0$, the set of states visited by the offline data distribution $\mu$, i.e., $S_0 := \{s \in S \mid \mu(s) > 0\}$, where we recall that $\mu(s) = \sum_{a\in A} \mu(s,a)$ for any $s \in S$* [7]. *Also, for any policy $\pi$ and any $s \in S$, we define*

$$S_\pi(s) := \{a \in A \mid \pi(a \mid s) > 0\},$$
$$\mathcal{T}(s) := \{a \in A \mid Q^*(s,a) = v^*(s)\}.$$

Next, we define the set of (in)active state-action pairs.

**Definition 5** (Active State-Action Pairs). *We say that a state-action pair $(s,a) \in S \times A$ is* active *if $Q^*(s,a) = v^*(s)$. Otherwise, $(s,a) \in S \times A$ is an* inactive *pair. Let $\mathcal{I} \subseteq S \times A$ be the set of inactive state-action pairs, and $S \times A \setminus \mathcal{I}$ thus corresponds to that of the* active *ones.*

We then have the following lemma which characterizes the optimal policy in terms of the inactive set $\mathcal{I}$:

**Lemma 4.** *If $\pi_0$ is an optimal policy, then $\theta_{\pi_0,\mu}(s,a) = 0$ for any $(s,a) \in \mathcal{I}$. If $\pi_0(a \mid s) = 0$ for any $(s,a) \in \mathcal{I}$, then $\pi_0$ is an optimal policy.*

Now we state the partial data coverage assumption. We first introduce the following definitions for convenience.

**Definition 6** (Data Coverage). *We say that $\pi$ is a $\mu$*-policy *if $\pi(a \mid s) > 0$ implies $\mu(s,a) > 0$ for any $s \in S_0$. A* $\mu$-optimal policy *is an optimal policy that is also a $\mu$-policy. Suppose there exists at least one $\mu$-optimal policy, then, a policy $\pi^*$ is called a* max-$\mu$-optimal policy *if it is a $\mu$-optimal policy that satisfies $|S_{\pi^*}(s)| = |S_{\pi_\mu}(s) \cap \mathcal{T}(s)|$ for any $s \in S_0$.*

**Remark 2.** *A $\mu$-policy means that this policy is covered by the behavior policy in some sense. For any state $s \in S_0$, it is reasonable to assume that a optimal pair $(s,a)$ can be visited by the behavior policy with positive probability, where $a$ is an optimal action that maximizes $Q(s,a)$. Therefore, it is reasonable to assume that a $\mu$-optimal policy exists. If a $\mu$-optimal policy exists, then the* max-$\mu$-*optimal policy must exist.*

We are now ready to state the SPC assumption.

**Assumption 5** (Single-Policy Concentrability+). *There exist some max-$\mu$-optimal policy $\pi^*$, and some constant $C^* > 0$ such that for all $(s,a) \in S \times A$, $\frac{\theta_{\pi^*,\mu}(s,a)}{\mu(s,a)} \le C_{\pi^*,\mu} = C^*$.*

Note that Assumption 5 is slightly stronger than the usual single-policy concentrability assumption (Rashidinejad et al., 2021; 2022), which assumes the coverage of any optimal policy. Assumption 5 means that if an optimal policy is covered by the behavior policy, then its occupancy measure should also be covered by the offline data distribution. It is reasonable in practice when the data is generated by sampling from a mixed Markov chain under some behavior policy (Liu et al., 2018; Levine et al., 2020). See a detailed discussion in Appendix C.2.

### 4.2. Main Results

**Proposition 2.** *Let $\pi^*$ be a max-$\mu$-optimal policy for which Assumption 5 holds. There exist constants $C^*, C_{\max} > 0$ such that:*
*1.) $\theta_{\pi^*}(s,a) \le C^* \mu(s,a)$ for any $(s,a) \in S \times A$;*
*2.) For any $\mu$-optimal policy $\pi$, we have $\theta_\pi(s) \le C_{\max}\mu(s)$ for any $s \in S$.*

Before moving to our main theorem, we define the *gap* of the optimal $Q$-function below, which is the minimal difference of the optimal $Q$-value between the optimal and the second optimal actions, among all states $s \in S$.

**Definition 7** (Gap). *For each $(s,a) \in S \times A$, we define the* gap $\Delta_Q(s,a) := v^*(s) - Q^*(s,a)$. *We then define the* minimal *gap as $\Delta_Q := \min_{(s,a)\in\mathcal{I}} \Delta_Q(s,a)$, where we recall that $\mathcal{I}$ is the set of* inactive *state-action pairs given in Definition 5.*

Note that as long as $\mathcal{I}$ is not empty, then $\Delta_Q(s,a) > 0$ for any $(s,a) \in \mathcal{I}$, leading to $\Delta_Q > 0$ by definition. If $\mathcal{I}$ is empty, then the problem becomes degenerate since any action is active for any state, i.e., any policy is an optimal policy. We hereafter focus on the non-degenerate case where $\Delta_Q > 0$. This gap notion was also used in (Chen & Jiang, 2022) in the context of offline RL. However, in contrast to this work, our definition here does not need to assume that the maximizer of $\max_a Q^*(s,a)$ is *unique* for each $s$, which is more standard in the online RL setting (Simchowitz & Jamieson, 2019; Lattimore & Szepesvári, 2020; Papini et al., 2021; Yang et al., 2021). Also, our algorithm does not need to know the gap $\Delta_Q$ and is tractable, compared to that in (Chen & Jiang, 2022). Now we are ready to present the following theorem.

**Theorem 3.** *Under Assumptions 3, 4, 5, we have, with*

[7] Note that we do not need to know $S_0$ for our algorithm to be stated later. We only need the definition of $S_0$ for analysis.

*probability* $\geq 1-\delta$

$$J_\mu(\pi^*) - J_\mu(\pi_\mathcal{D}) \leq \frac{8\sqrt{2}B_w C_{\max}}{\Delta_Q(1-\gamma)^3} \cdot \frac{\sqrt{\log(|W||V|/\delta)}}{\sqrt{n}}.$$

Theorem 3 provides an optimal $O(1/\sqrt{n})$ sample complexity under some single-policy concentrability and the realizability-only assumption, for the return with initial distribution $\rho = \mu$ for the LP, with the aid of the lower bound $1-\gamma$ on $\sum_a w(s,a)\pi_\mu(a \mid s)$. However, this bound depends on all $\mu-$optimal policies, due to $C_{\max}$. We relax this dependence in Appendix D.2. The proof of the theorem is based on the *primal gap* analysis proposed in (Ozdaglar et al., 2022), which was shown to be critical in characterizing the generalization behaviors in stochastic minimax optimization. Note that in Theorem 3, the value function $J_\mu$ is based on initial distribution $\mu$. The next corollary connects back to the reward with initial distribution $\mu_0$.

**Corollary 1.** *Under Assumptions 3, 4, 5, and suppose that* $\mu_0$ *is covered by* $\mu$, *i.e.,* $\max_{s\in S} \frac{\mu_0(s)}{\mu(s)} \leq C_\mu$ *for some constant* $C_\mu > 0$, *we have, with probability* $\geq 1-\delta$,

$$J_{\mu_0}(\pi^*) - J_{\mu_0}(\pi_\mathcal{D}) \leq \frac{8\sqrt{2}B_w C_{\max} C_\mu}{\Delta_Q(1-\gamma)^3} \cdot \frac{\sqrt{\log(|W||V|/\delta)}}{\sqrt{n}}.$$

Corollary 1 follows by a direct change of measure argument and is thus omitted. We provide a detailed discussion of this corollary in Appendix C.4.

Recall the definitions of $\ell$ and $\ell_\mathcal{D}$ in (12) and (13), respectively, and the fact that we set $\rho = \mu$, we have $\ell(w,v) = -u^\top w + v^\top(Kw - (1-\gamma)\mu)$ and $\ell_\mathcal{D}(w,v) = -u_\mathcal{D}^\top w + v^\top(K_\mathcal{D} w - (1-\gamma)\mu_\mathcal{D})$. The population and empirical primal gaps are defined as follows.

**Definition 8** (Primal Gap). *Let* $\ell^V(w) = \max_{v\in V} \ell(w,v)$ *and* $\ell^V_\mathcal{D}(w) = \max_{v\in V} \ell_\mathcal{D}(w,v)$. *The empirical primal gap is defined as* $\Delta^{W,V}_\mathcal{D}(w) = \ell^V_\mathcal{D}(w) - \min_{w'\in W} \ell^V_\mathcal{D}(w')$, *and the population primal gap is defined as* $\Delta^{W,V}(w) = \ell^V(w) - \min_{w'\in W} \ell^V(w')$. *For notational simplicity, we omit the superscripts* $W, V$ *hereafter.*

Let $w_\mathcal{D}$ be the solution to problem (21). We have $\Delta_\mathcal{D}(w_\mathcal{D}) = 0$. We can upper bound of the population primal gap at $w_\mathcal{D}$ as follows:

**Lemma 5.** *Suppose Assumptions 3, 4 hold. Then, with probability* $\geq 1-\delta$, *we have*

$$\Delta(w_\mathcal{D}) \leq \frac{4\sqrt{2}B_w\sqrt{\log(|V||W|/\delta)}}{(1-\gamma)\sqrt{n}}.$$

Next, we need to relate the primal gap to the accuracy of policy $\pi_\mathcal{D}$ in terms of $J_\mu(\pi^*) - J_\mu(\pi_\mathcal{D})$. Notice that inspired by Lemma 4, the sub-optimality gap of $\pi_\mathcal{D}$ can be captured by the violation of $\pi_\mathcal{D}(a|s) = 0$ for $(s,a) \in \mathcal{I}$. $\pi_\mathcal{D}(\cdot|s)$ is the normalization of $\theta_\mathcal{D}(s,\cdot)$, where $\theta_\mathcal{D}(s,a) = w(s,a)\mu(s,a)$. Hence, we bound $\theta_\mathcal{D}$ in $\mathcal{I}$ as follows:

**Lemma 6.** *We have* $\sum_{(s,a)\in\mathcal{I}} \theta_\mathcal{D}(s,a) \leq \frac{\Delta(w_\mathcal{D})}{\Delta_Q}$.

Finally, combining the lower bound constraints in Program (19), we have the following estimate of $J_\mu(\pi^*) - J_\mu(\pi_\mathcal{D})$.

**Lemma 7.** *We have*

$$J_\mu(\pi^*) - J_\mu(\pi_\mathcal{D}) \leq \frac{2C_{\max}}{(1-\gamma)^2\Delta_Q}\Delta(w_\mathcal{D}).$$

Combining Lemma 5 and Lemma 7, Theorem 3 follows.

## 5. Concluding Remarks

In this paper, we revisited the linear programming framework for offline RL with general function approximation, which has been advocated recently in (Zhan et al., 2022) to obtain provably efficient algorithms with only partial data coverage and function class realizability assumptions. We proposed two offline RL algorithms with function approximation to different decision variables, and established optimal $O(1/\sqrt{n})$ sample complexity with partial data coverage, relying on either certain completeness-type assumption, or a slightly stronger data coverage assumption than standard single-policy concentrability. Key to our analysis is adding proper constraints in the LP and the induced minimax optimization problems for solving the MDPs.

Our work has opened up avenues for future research in offline RL. For example, is it possible to achieve optimal sample complexity with the standard single-policy concentrability assumption and only realizability, under the LP framework? What is the gap-dependent lower bound for offline RL with general function approximation? If the behavior policy is not known, is there an approach better than direct behavior cloning? Would policy-based offline RL algorithms be able to handle partial data coverage and realizability-only assumptions simultaneously? We hope our results can provide some insights into addressing these questions, especially when the LP framework is used.

## Acknowledgements

A.O and K.Z. were supported by MIT-DSTA grant 031017-00016. S.P. acknowledges support from MathWorks Engineering Fellowship. K.Z. also acknowledges support from Simons-Berkeley Research Fellowship.

## References


Amortila, P., Jiang, N., and Xie, T. A variant of the Wang-Foster-Kakade lower bound for the discounted setting. *arXiv preprint arXiv:2011.01075*, 2020.

Bertsekas, D. P. *Dynamic programming and optimal control (4th edition)*. Athena Scientific, 2017.

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. *arXiv preprint arXiv:2203.13935*, 2022.

Cheng, C.-A., Xie, T., Jiang, N., and Agarwal, A. Adversarially trained actor critic for offline reinforcement learning. *arXiv preprint arXiv:2202.02446*, 2022.

De Farias, D. P. and Van Roy, B. The linear programming approach to approximate dynamic programming. *Operations Research*, 51(6):850–865, 2003.

Foster, D. J., Krishnamurthy, A., Simchi-Levi, D., and Xu, Y. Offline reinforcement learning: Fundamental barriers for value function approximation. *arXiv preprint arXiv:2111.10919*, 2021.

Fujimoto, S., Meger, D., and Precup, D. Off-policy deep reinforcement learning without exploration. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019.

Jiang, N. and Huang, J. Minimax value interval for off-policy evaluation and policy optimization. *Advances in Neural Information Processing Systems*, 33:2747–2758, 2020.

Jin, Y., Yang, Z., and Wang, Z. Is pessimism provably efficient for offline rl? In *arXiv preprint arXiv:2012.15085*, 2020.

Kakade, S. and Langford, J. Approximately optimal approximate reinforcement learning. In *Proceedings of the Nineteenth International Conference on Machine Learning*, pp. 267–274, 2002.

Kumar, A., Zhou, A., Tucker, G., and Levine, S. Conservative q-learning for offline reinforcement learning. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

Lattimore, T. and Szepesvári, C. *Bandit Algorithms*. Cambridge University Press, 2020.

Lee, J., Jeon, W., Lee, B.-J., Pineau, J., and Kim, K.-E. Optidice: Offline policy optimization via stationary distribution correction estimation. *arXiv preprint arXiv:2106.10783*, 2021.

Levine, S., Finn, C., Darrell, T., and Abbeel, P. End-to-end training of deep visuomotor policies. *The Journal of Machine Learning Research*, 17(1):1334–1373, 2016.

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.

Li, G., Shi, L., Chen, Y., Chi, Y., and Wei, Y. Settling the sample complexity of model-based offline reinforcement learning. *arXiv preprint arXiv:2204.05275*, 2022.

Liu, Q., Li, L., Tang, Z., and Zhou, D. Breaking the curse of horizon: Infinite-horizon off-policy estimation. *arXiv preprint arXiv:1810.12429*, 2018.

Liu, Y., Swaminathan, A., Agarwal, A., and Brunskill, E. Provably good batch off-policy reinforcement learning without great exploration. *Advances in Neural Information Processing Systems*, 33:1264–1274, 2020.

Maddern, W., Pascoe, G., Linegar, C., and Newman, P. 1 year, 1000 km: The oxford robotcar dataset. *The International Journal of Robotics Research*, 36(1):3–15, 2017.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. Human-level control through deep reinforcement learning. *Nature*, 518 (7540):529–533, 2015.

Munos, R. and Szepesvári, C. Finite-time bounds for fitted value iteration. In *Journal of Machine Learning Research*, volume 9, pp. 815–857, 2008.

Nachum, O., Dai, B., Kostrikov, I., Chow, Y., Li, L., and Schuurmans, D. Algaedice: Policy gradient from arbitrary experience. *arXiv preprint arXiv:1912.02074*, 2019.

Ozdaglar, A. E., Pattathil, S., Zhang, J., and Zhang, K. What is a good metric to study generalization of minimax learners? In *Advances in Neural Information Processing Systems*, 2022. URL `https://openreview.net/forum?id=wYGIxXZ_sZx`.

Papini, M., Tirinzoni, A., Pacchiano, A., Restelli, M., Lazaric, A., and Pirotta, M. Reinforcement learning in linear MDPs: Constant regret and representation selection. *Advances in Neural Information Processing Systems*, 34:16371–16383, 2021.

Puterman, M. L. Markov decision processes: Discrete stochastic dynamic programming, 1994.

Rashidinejad, P., Zhu, B., Ma, C., Jiao, J., and Russell, S. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *Advances in Neural Information Processing Systems*, 34:11702–11716, 2021.

Rashidinejad, P., Zhu, H., Yang, K., Russell, S., and Jiao, J. Optimal conservative offline rl with general function approximation via augmented lagrangian. *arXiv preprint arXiv:2211.00716*, 2022.

Scherrer, B. Approximate policy iteration schemes: A comparison. In *International Conference on Machine Learning*, pp. 1314–1322. PMLR, 2014.

Sidford, A., Wang, M., Wu, X., Yang, L., and Ye, Y. Near-optimal time and sample complexities for solving Markov decision processes with a generative model. In *Advances in Neural Information Processing Systems*, pp. 5186–5196, 2018.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. Mastering the game of Go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

Simchowitz, M. and Jamieson, K. G. Non-asymptotic gap-dependent regret bounds for tabular MDPs. *Advances in Neural Information Processing Systems*, 32, 2019.

Swaminathan, A., Krishnamurthy, A., Agarwal, A., Dudik, M., Langford, J., Jose, D., and Zitouni, I. Off-policy evaluation for slate recommendation. *Advances in Neural Information Processing Systems*, 30, 2017.

Tang, Z., Feng, Y., Li, L., Zhou, D., and Liu, Q. Doubly robust bias reduction in infinite horizon off-policy estimation. In *International Conference on Learning Representations*, 2019.

Todorov, E., Erez, T., and Tassa, Y. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pp. 5026–5033. IEEE, 2012.

Tseng, H.-H., Luo, Y., Cui, S., Chien, J.-T., Ten Haken, R. K., and Naqa, I. E. Deep reinforcement learning for automated radiation adaptation in lung cancer. *Medical Physics*, 44(12):6690–6705, 2017.

Uehara, M. and Sun, W. Pessimistic model-based offline RL: Pac bounds and posterior sampling under partial coverage. *arXiv e-prints*, pp. arXiv–2107, 2021.

Uehara, M., Huang, J., and Jiang, N. Minimax weight and Q-function learning for off-policy evaluation. In *International Conference on Machine Learning*, pp. 9659–9668. PMLR, 2020.

Vinyals, O., Ewalds, T., Bartunov, S., Georgiev, P., Vezhnevets, A. S., Yeo, M., Makhzani, A., Küttler, H., Agapiou, J., Schrittwieser, J., et al. Starcraft II: A new challenge for reinforcement learning. *arXiv preprint arXiv:1708.04782*, 2017.

Wang, R., Foster, D., and Kakade, S. M. What are the statistical limits of offline RL with linear function approximation? In *International Conference on Learning Representations*, 2020.

Xie, T. and Jiang, N. Batch value-function approximation with only realizability. In *International Conference on Machine Learning*, pp. 11404–11413. PMLR, 2021.

Xie, T., Cheng, C.-A., Jiang, N., Mineiro, P., and Agarwal, A. Bellman-consistent pessimism for offline reinforcement learning. *arXiv preprint arXiv:2106.06926*, 2021.

Yang, K., Yang, L., and Du, S. Q-learning with logarithmic regret. In *International Conference on Artificial Intelligence and Statistics*, pp. 1576–1584. PMLR, 2021.

Zanette, A. Exponential lower bounds for batch reinforcement learning: Batch RL can be exponentially harder than online RL. In *International Conference on Machine Learning*, pp. 12287–12297. PMLR, 2021.

Zhan, W., Huang, B., Huang, A., Jiang, N., and Lee, J. D. Offline reinforcement learning with realizability and single-policy concentrability. *arXiv preprint arXiv:2202.04634*, 2022.

Zhang, K., Yang, Z., Liu, H., Zhang, T., and Başar, T. Finite-sample analysis for decentralized batch multiagent reinforcement learning with networked agents. *IEEE Transactions on Automatic Control*, 66(12):5925–5940, 2021.

# Supplementary Materials for “Revisiting the Linear-Programming Framework for Offline RL with General Function Approximation”

## A. Omitted Details in Section 1

### A.1. Related Work

We provide a more detailed literature review in this subsection, and categorize the results based on their assumptions on data and function class.

**Data coverage assumptions.** Early theoretical works on offline RL usually require the *all-policy concentrability* assumption, i.e., the offline data has to be exploratory enough to cover the state distributions induced by *all* policies (Munos & Szepesvári, 2008; Scherrer, 2014; Chen & Jiang, 2019). We refer to this assumption as the *full data coverage* assumption. Slightly weaker variant that assumes some weighted version of the all-policy concentrability coefficient is bounded has also been investigated (Xie & Jiang, 2021; Uehara et al., 2020). More recently, significant progress has been made to relax full coverage assumption to partial coverage ones. (Jin et al., 2020; Rashidinejad et al., 2021; Li et al., 2022) developed pessimistic value iteration based algorithms for tabular or linear MDPs, under the single-policy concentrability assumption on data coverage. When general function approximation is used, some variants of the SPC assumption were proposed to account for partial data coverage (Uehara & Sun, 2021; Xie et al., 2021; Cheng et al., 2022). However, these algorithms are either computationally intractable (Uehara & Sun, 2021; Xie et al., 2021), or statistically suboptimal (Cheng et al., 2022). Other recent works that require only partial data coverage are (Zhan et al., 2022; Chen & Jiang, 2022) which will be discussed next.

**Function class assumptions.** One common assumption on function class is the Bellman-completeness on value functions (Munos & Szepesvári, 2008; Scherrer, 2014; Chen & Jiang, 2019; Xie et al., 2021; Cheng et al., 2022), which requires the value function class to be closed under the Bellman operator. By definition, such an assumption is automatically satisfied for the tabular and linear MDP cases mentioned above (Jin et al., 2020; Rashidinejad et al., 2021; Li et al., 2022), and is implied when realizability of the MDP model (Uehara & Sun, 2021) is assumed, see (Chen & Jiang, 2019). This strong assumption has been recently relaxed to only *realizability*, i.e., the function class only needs to contain (approximately) the target function of interest (e.g., optimal value function) (Xie & Jiang, 2021). However, (Xie & Jiang, 2021) relies on data coverage assumption that is even stronger than all-policy concentrability. In fact, there have been hardness results (Wang et al., 2020; Amortila et al., 2020; Zanette, 2021; Foster et al., 2021) showing that even with good data coverage, realizability-only assumption on the value function is not sufficient for sample-efficient offline RL. This motivated the use of function approximation for density ratio (in addition to value function), as in (Nachum et al., 2019; Zhan et al., 2022; Chen & Jiang, 2022; Jiang & Huang, 2020) and our work. In particular, (Zhan et al., 2022; Chen & Jiang, 2022) are the most related recent works that assume only realizability, on both value function and density ratio, and partial data coverage. However, they are either statistically suboptimal (Zhan et al., 2022) or computationally intractable (Chen & Jiang, 2022). Moreover, (Zhan et al., 2022) additionally requires the data coverage of the *regularized* problem; and (Chen & Jiang, 2022) additionally requires that the greedy optimal action is *unique* for all states.

**Independent work (Rashidinejad et al., 2022).** While preparing our paper, we came across a concurrent and independent work (Rashidinejad et al., 2022), which also obtained the optimal $O(1/\sqrt{n})$ rate under general function approximation via the LP framework, and also without behavioral regularization. Note that (Rashidinejad et al., 2022) requires completeness-type assumptions throughout, which can be viewed as mirroring the first half of our results (i.e., Section 3), while we also have the realizability-only results under a slightly different data coverage assumption (i.e., Section 4).

(Rashidinejad et al., 2022) and Section 3 of our paper are different in the following aspects: First, (Rashidinejad et al., 2022) is based on an augmented Lagrangian method (ALM), while we propose to solve the optimization with constraints directly. Second, the function classes being used, and the corresponding completeness and realizability assumptions are different (see Section 3 for more details). Third, with a different and rather simple analysis, our results have better dependence on

$(1-\gamma)$, and even improves the state-of-the-art result when specializing to the tabular case (Rashidinejad et al., 2021). Finally, we also note that interestingly, both works have noticed the importance of occupancy validity constraints, and our constrained formulation in Section 3 mirrors the role of ALM in (Rashidinejad et al., 2022), to enforce such validity constraints.

### A.2. Our Key Techniques

The key idea to our approaches is to study *properly constrained* versions of the LP reformulation of the underlying MDP. In particular, we focus on the dual problem of a variant of the standard LP reformulation, based on the marginal importance sampling framework (Nachum et al., 2019; Lee et al., 2021), where the dual variable corresponds to the ratio between the state-action occupancy measure and the offline data distribution (also referred to as the *density ratio*).

Our first set of results relies on a key *error bound* lemma that relates the value function suboptimality with the $\ell_1$-norm violation of the validity constraint on the occupancy measure in the LP (see Lemma 3). This lemma leads to a *constrained dual formulation* without the need of behavior regularization as in (Zhan et al., 2022). Using function approximation for the density ratio and the *sign function* of the occupancy validity constraint (see Definition 3), this formulation organically allows us to obtain $O(1/\sqrt{n})$ sample complexity under the realizability of density ratio function class, and certain completeness assumption on the sign function, together with standard SPC assumption (Rashidinejad et al., 2021; Chen & Jiang, 2022).

To remove any completeness assumption, in the second part, we consider the minimax reformulation of the dual LP, which dualizes the occupancy measure validity constraints. To stabilize the normalization step in generating the policy from the LP solution (see Equation (10)), we introduce an additional *lower-bound constraint* on the density ratio, which does not lose optimality if the initial state distribution coincides with the offline data distribution. Under this new formulation, we establish gap-dependent $O(1/\sqrt{n})$ sample complexity with only realizability assumptions on the value function and density ration, and a slightly stronger SPC assumption that assumes certain optimal policy covered by the behavior policy is also covered by the offline data distribution.

### A.3. Notation

For a vector $v \in \mathbb{R}^d$, we use $\|v\|_p$ to denote its $\ell_p$ norm (where $p \in [0, \infty]$, and if there is no subscript, $\|v\|$ denotes the $\ell_2$ norm. Note that $\|v\|_0$ denotes the number of non-zero elements in $v$. For a matrix $M \in \mathbb{R}^{m\times n}$, we use $\|M\|$ and $\|M\|_F$ to denote its $\ell_2$-induced and Frobenius norm, respectively, and use $M^\top$ to denote its transpose. For a set $S$, we use $|S|$ to denote its cardinality, and $\Delta(S)$ to denote the probability distribution over $S$. For a function class $\mathcal{F}$, we use $|\mathcal{F}|$ to denote its cardinality if it is discrete, and its covering number if it is continuous. We use $\mathbb{E}$ to denote expectation. For any matrix $M \in \mathbb{R}^{m\times n}$, $M \geq 0$ denotes that each element of $M$ is non-negative. We also use $\mathbb{R}_+^d$ to denote the $d$-dimensional real vector space with all elements being non-negative. We use $\mathrm{Diag}(M_1, \cdots, M_n)$ to denote the block diagonal matrix of proper dimension whose diagonal blocks $M_1, \cdots, M_n$ have the same dimension. We follow the convention of $0/0 = 0$ throughout, unless otherwise noted.

# B. Omitted Details in Section 3

## B.1. Discussion on Assumption 1

Several remarks are in order. First, we introduce $x_w$ to calculate the $\ell_1$-norm of $Kw - (1-\gamma)\mu_0$, since the $\ell_1$-norm will be related to the suboptimality gap of the policy obtained from $w$ (see Lemma 3 below). Second, Assumption 1 contains not only realizability of $W$ for $w^*$, but also the *completeness-type* assumption of $B$ for $x_w \in B$ for any $w \in W$. The completeness-type assumptions are standard in the offline RL literature (Munos & Szepesvári, 2008; Chen & Jiang, 2019; Xie et al., 2021), and can be challenging or even impossible to remove in certain cases due to some hardness results (Foster et al., 2021). To the best of our knowledge, the only existing results that merely assume realizability of the optimal solutions are (Uehara & Sun, 2021; Zhan et al., 2022; Chen & Jiang, 2022), which are either statistically sub-optimal or computationally intractable. We defer our solution to the *realizability-only* case to Section 4.

Third, interestingly, some completeness assumption is also made in the concurrent work (Rashidinejad et al., 2022) that achieves the optimal $O(1/\sqrt{n})$ rate as well (see their Theorem 4), mirroring our Assumption 1. Note that we only need the completeness of one function class $B$ for $x_w$, while (Rashidinejad et al., 2022) requires the completeness of $U$ for $u_w^*$ (with the notation therein), the realizability of $v^*$, together with either the realizability of the model $P$ (which is deemed as even stronger than Bellman-completeness (Chen & Jiang, 2019; Zhan et al., 2022; Uehara & Sun, 2021)), or the completeness of two function classes $\mathcal{U}$ and $\mathcal{Z}$ therein.

## B.2. Proof of Lemma 3

Let the policy that is obtained by normalizing both $\theta$ and $\theta_{\pi_\theta}$ be $\pi_\theta$ (note that normalizing both of these vectors gives the same policy). Next, we define:

$$\bar{\theta}(s) = \sum_{a\in A}\theta(s,a), \qquad \text{and} \qquad \bar{\theta}_{\pi_\theta}(s) = \sum_{a\in A}\theta_{\pi_\theta}(s,a). \tag{22}$$

Note that we can write $\theta(s,a) = \bar{\theta}(s)\pi_\theta(a \mid s)$ and $\theta_{\pi_\theta}(s,a) = \bar{\theta}_{\pi_\theta}(s)\pi_\theta(a \mid s)$.

Let $P_{\pi_\theta} \in \mathbb{R}^{|S|\times|S|}$ be a column stochastic matrix (the sum of all entries of every column is 1) which describes the state transition probabilities under the policy $\pi_\theta$, i.e.,

$$P_{\pi_\theta}(j,i) = \sum_{a\in A} P_{s^i,a}(s^j)\cdot\pi_\theta(a \mid s^i).$$

Also, we define the matrix $G_\theta = \text{Diag}(\pi_\theta(\cdot \mid s^1), \pi_\theta(\cdot \mid s^2), \cdots, \pi_\theta(\cdot \mid s^{|S|})) \in \mathbb{R}^{|S||A|\times|S|}$, and notice the fact that $MG_\theta = I - \gamma P_{\pi_\theta}$. Now, since $\theta_{\pi_\theta}$ satisfies the constraints in Problem (6), we have $M\theta_{\pi_\theta} = (1-\gamma)\mu_0$. This implies:

$$\begin{aligned}\|M\theta - (1-\gamma)\mu_0\|_1 &= \|M(\theta - \theta_{\pi_\theta})\|_1 \\ &= \|MG_\theta(\bar{\theta} - \bar{\theta}_{\pi_\theta})\|_1 \\ &= \|(I - \gamma P_{\pi_\theta})(\bar{\theta} - \bar{\theta}_{\pi_\theta})\|_1 \\ &\geq (1-\gamma)\|\bar{\theta} - \bar{\theta}_{\pi_\theta}\|_1.\end{aligned} \tag{23}$$

Here the last inequality is because $\gamma\|P_{\pi_\theta}(\bar{\theta} - \bar{\theta}_{\pi_\theta})\|_1 \leq \gamma\|\bar{\theta} - \bar{\theta}_{\pi_\theta}\|_1$, which follows from the fact that $P_{\pi_\theta}$ is a column stochastic matrix.

On the other hand, since $r(s,a) \in [0,1]$ for all $(s,a)$, we have:

$$\begin{aligned}|r^\top(\theta - \theta_{\pi_\theta})| &= |r^\top G_\theta(\bar{\theta} - \bar{\theta}_{\pi_\theta})| \\ &\leq \|\bar{\theta} - \bar{\theta}_{\pi_\theta}\|_1.\end{aligned} \tag{24}$$

Combining inequalities (23) and (24), we get the result. □

## B.3. Proof of Theorem 1

First, we need to guarantee the feasibility of the optimization problem (15).

**Lemma 8.** *Any $w^* \in W$ (see Assumption 1) is feasible to* (15) *with probability at least* $1-\delta$*.*

*Proof.* Use Hoeffding's inequality, we have that for any $x \in B$:

$$\mathbb{P}(x^\top (K - K_{\mathcal{D}})w^* \geq t) \leq \exp\left(\frac{-nt^2}{8B_w^2}\right). \tag{25}$$

This is using the fact that $x^\top (K-K_{\mathcal{D}})w^*$ is a random variable which lies in the interval $[-B_w(1+\gamma), B_w(1+\gamma)]$. Note that we use the fact of $\|w^*\|_\infty \leq C^* \leq B_w$, by Assumption 1. Now, taking $t = 2\sqrt{2}B_w \frac{\sqrt{\log(|W||B|/\delta)}}{\sqrt{n}}$, we have

$$\mathbb{P}(x^\top (K - K_{\mathcal{D}})w^* \geq t) \leq \frac{\delta}{|W||B|}. \tag{26}$$

Now, taking the union bound over all $x \in B$ and also all $w \in W$, we get the final result. □

Next, we show that the objective value $u_{\mathcal{D}}^\top w_{\mathcal{D}}$ is close to $u^\top w^*$.

**Lemma 9.** *We have*

$$u_{\mathcal{D}}^\top w_{\mathcal{D}} \geq u^\top w^* - \frac{\sqrt{2}B_w}{\sqrt{n}}\sqrt{\log\frac{1}{\delta}}$$

*with probability at least* $1-2\delta$*.*

*Proof.* From Lemma 8, we have

$$u_{\mathcal{D}}^\top w_{\mathcal{D}} \geq u_{\mathcal{D}}^\top w^*$$

with probability at least $1-\delta$ (since $w^*$ is feasible to (15) with probability $1-\delta$).

Then we can use Hoeffding's inequality to bound $(u-u_{\mathcal{D}})^\top w^*$ as follows:

$$\mathbb{P}(u_{\mathcal{D}}^\top w^* \leq u^\top w^* - t) \leq \exp\left(\frac{-nt^2}{2B_w^2}\right). \tag{27}$$

Setting this upper bound to be equal to $\delta$, we have:

$$t = \frac{\sqrt{2}B_w}{\sqrt{n}}\sqrt{\log\frac{1}{\delta}}. \tag{28}$$

Combining the two events completes the proof. □

Next, we provide a bound for $\|Kw_{\mathcal{D}} - (1-\gamma)\mu_0\|_1$.

**Lemma 10.** *We have*

$$\|Kw_{\mathcal{D}} - (1-\gamma)\mu_0\|_1 \leq 2E_{n,\delta}$$

*with probability at least* $1-2\delta$*.*

*Proof.* We first have

$$x^\top (K-K_{\mathcal{D}})w \leq E_{n,\delta}, \qquad \forall x \in B \tag{29}$$

for any $x \in B,\ w \in W$ with probability at least $1-\delta$, by a concentration bound and union bound (similar to the proof of Lemma 8). This directly implies our lemma since $w_{\mathcal{D}} \in W$.

Therefore:

$$\begin{aligned}\|Kw_{\mathcal{D}} - (1-\gamma)\mu_0\|_1 &\leq \|K_{\mathcal{D}}w_{\mathcal{D}} - (1-\gamma)\mu_0\|_1 + \|(K-K_{\mathcal{D}})w_{\mathcal{D}}\|_1 \\ &\leq E_{n,\delta} + E_{n,\delta},\end{aligned} \tag{30}$$

where the first term on the right-hand side is due to that $w_{\mathcal{D}}$ satisfies the constraint in (15). This completes the proof. □

Finally, combining the above lemmas and Lemma 3, we can prove Theorem 1:

*Proof of Theorem 1.* From Lemma 9, we have:

$$J_{\mu_0}(\pi^*) = u^\top w^* \le u_{\mathcal{D}}^\top w_{\mathcal{D}} + \frac{\sqrt{2}B_w}{\sqrt{n}}\sqrt{\log\frac{1}{\delta}}. \tag{31}$$

This tells us that $u_{\mathcal{D}}^\top w_{\mathcal{D}}$ is close to $u^\top w^* = J_{\mu_0}(\pi^*)$. Next, using Hoeffding's inequality and union bound, similarly we know that

$$u_{\mathcal{D}}^\top w \le u^\top w + \frac{\sqrt{2}B_w}{\sqrt{n}}\sqrt{\log\frac{|W|}{\delta}} \tag{32}$$

for any $w \in W$. Then since $w_{\mathcal{D}} \in W$, we have

$$u_{\mathcal{D}}^\top w_{\mathcal{D}} \le u^\top w_{\mathcal{D}} + \frac{\sqrt{2}B_w}{\sqrt{n}}\sqrt{\log\frac{|W|}{\delta}} \tag{33}$$

with probability at least $1-\delta$. Define $\hat{\theta}_{\mathcal{D}}(s,a) = \tilde{\theta}_{\mathcal{D}}(s,a)\mu(s)$ where we recall the definition of $\tilde{\theta}_{\mathcal{D}}$ in (22). Note that $\theta_{\pi_{\hat{\theta}_{\mathcal{D}}}} = \theta_{\pi_{\tilde{\theta}_{\mathcal{D}}}}$. Next, using the definition of $u$, we have $u^\top w_{\mathcal{D}} = r^\top \hat{\theta}_{\mathcal{D}}$. Now, using Lemma 3, we can bound the difference

$$r^\top \hat{\theta}_{\mathcal{D}} \le r^\top \theta_{\pi_{\tilde{\theta}_{\mathcal{D}}}} + \frac{\|M\hat{\theta}_{\mathcal{D}} - (1-\gamma)\mu_0\|_1}{1-\gamma}, \tag{34}$$

where we recall that $\pi_{\tilde{\theta}_{\mathcal{D}}}$ is generated by $\tilde{\theta}_{\mathcal{D}}$ by normalization. Note that here we have $r^\top \theta_{\pi_{\tilde{\theta}_{\mathcal{D}}}} = J_{\mu_0}(\pi_{\tilde{\theta}_{\mathcal{D}}}) = J_{\mu_0}(\pi_{\mathcal{D}})$. Finally, using Lemma 10, we can bound $\|M\hat{\theta}_{\mathcal{D}} - (1-\gamma)\mu_0\|_1 = \|Kw_{\mathcal{D}} - (1-\gamma)\mu_0\|_1$, which completes the proof. □

### B.4. Proof of Theorem 2

Note that Theorem 1 is not directly applicable to derive the sample complexity of this algorithm. Though not directly applicable, we can still base our analysis on the derivations above. Specifically, here we provide a proof for this theorem based on the number of extreme points of the convex sets.

*Proof.* The number of extreme points of $W$ and $B$ are $(|A|+1)^{|S|}$ and $2^{|S|}$. With a slight abuse of notation, we let $\|W\|_e$ and $\|B\|_e$ denote the number of extreme points of $W$ and $B$, respectively.

According to the proof of Theorem 1, we only need to modify two union concentration bounds – (29) and (32). These two inequalities can be replaced by the following two inequalities in terms of the number of extreme points:

1. $|x^\top(K - K_{\mathcal{D}})w| \le 2B_w\sqrt{|S|\log((2|A|+2)/\delta)}/\sqrt{n}$ for any $w \in W,\ x \in B$ with probability $\ge 1-\delta$;
2. $|(u - u_{\mathcal{D}})^\top w| \le 2B_w\sqrt{|S|\log((|A|+1)/\delta)}/\sqrt{n}$ for any $w \in W$ with probability $\ge 1-\delta$.

We only prove the first claim and the second one follows similarly. Let $W_0 = \{w_1, \cdots, w_{\|W\|_e}\}, B_0 = \{x_1, \cdots, x_{\|B\|_e}\}$ be the sets of extreme points of $W, B$, respectively. Then for any $w \in W$, we have

$$w = \sum_{i=1}^{\|W\|_e} \lambda_i w_i,$$

and for any $x \in B$,

$$x = \sum_{j=1}^{\|B\|_e} \zeta_j x_j,$$

for some $\lambda = (\lambda_1, \cdots, \lambda_{\|W\|_e})^\top$ and $\zeta = (\zeta_1, \cdots, \zeta_{\|B\|_e})^\top$ that lie in the corresponding simplices. For any $w_i,\ x_j$, using Hoeffding inequality and union bound on the sets $W_0, B_0$, we have

$$|x_j^\top (K - K_{\mathcal{D}}) w_i| \le 2B_w \sqrt{|S| \log((2|A| + 2)/\delta)}/\sqrt{n}$$

with probability $\ge 1 - \delta$. Then using the decomposition

$$x^\top (K - K_{\mathcal{D}}) w = \sum_{i=1}^{\|W\|_e} \sum_{j=1}^{\|B\|_e} \lambda_i \zeta_j x_j^\top (K - K_{\mathcal{D}}) w_i$$

and the Jensen's inequality, we prove that with probability $\ge 1 - \delta$, we have

$$|x^\top (K - K_{\mathcal{D}}) w| \le 2B_w \sqrt{|S| \log((2|A| + 2)/\delta)}/\sqrt{n}$$

This completes the proof of the first claim. Using the same strategy as the proof of Theorem 1, we prove Theorem 2. $\square$

## C. Omitted Details in Section 4

### C.1. Proof of Lemma 4

We prove the first part as follows. If $\pi_0$ is optimal, $\pi_0(a \mid s) = 0$ for any inactive $(s, a)$. Therefore, $\Pr_{\pi_0}(s^t = s, a^t = a; \mu) = 0$ for any $t$ and any $(s, a) \in \mathcal{I}$. Therefore, $\theta_{\pi_0,\mu}(s, a) = 0$ for any $(s, a) \in \mathcal{I}$.

For the second part, we prove it as follows. Let $v^*$ be the optimal value function. We only need to prove $v_{\pi_0} = v^*$. For a policy $\pi$, define

$$P_\pi(j, i) = \sum_{a \in A} P_{s^i,a}(s^j) \cdot \pi(a \mid s^i)$$

and

$$r_\pi = (r(s^1, \cdot)^\top \pi(s^1, \cdot), \cdots, r(s^{|S|}, \cdot)^\top \pi(s^{|S|}, \cdot))^\top.$$

Then we know that $v_{\pi_0}$ is the unique solution to the linear equation:

$$v = \gamma P_{\pi_0}^\top v + r_{\pi_0}. \tag{35}$$

We then prove that $v^*$ is also a solution to this equation. In fact, letting $P_\pi(i)$ be the $i$-th column of $P_\pi$, we have

$$\begin{aligned}
& \gamma P_{\pi_0}^\top(i) v^* + r_{\pi_0}(i) && (36)\\
= & \sum_{a \in S_{\pi_0}(s^i)} \pi_0(a \mid s^i)(\gamma P_{s^i,a}^\top v^* + r(s^i, a)) && (37)\\
= & \sum_{a \in S_{\pi_0}(s^i)} \pi_0(a \mid s^i) Q^*(s^i, a) && (38)\\
= & \; v^*(s^i), && (39)
\end{aligned}$$

where the second equality is because $\gamma P^\top v^* + r = Q^*$ and the third equality is because $Q^*(s, a) = v^*(s)$ for $a \in S_{\pi_0}(s)$. Then $v^*$ is the solution to (35). Since the solution to (35) is unique (Puterman, 1994), we have $v^* = v_{\pi_0}$, which yields the desired result. □

### C.2. Discussion on Assumption 5

Note that Assumption 5 is slightly stronger than the usual single-policy concentrability assumption (Rashidinejad et al., 2021; 2022), which assumes the coverage of any optimal policy. Assumption 5 means that if an optimal policy is covered by the behavior policy (i.e., it is a $\mu$-optimal policy), then its occupancy measure should also be covered by the offline data distribution. It is reasonable in the following sense: In practice, the offline data distribution is usually generated from the *stationary distribution* of the Markov chain under some behavior policy, which can be obtained by rolling out some infinitely (or sufficiently) long trajectories using the policy (Liu et al., 2018; Levine et al., 2020). Thus, $\mu$ satisfies the fixed point equation

$$\mu(s') = \sum_{s,a} \mu(s, a) P_{s,a}(s'). \tag{40}$$

Then the usual single-policy concentrability with initial distribution $\mu$ implies Assumption 5. To see this, first, we note that usual SPC with initial $\mu$ implies that the covered policy $\pi^*$ is a $\mu$-optimal policy by definition. Hence, there must exist a max-$\mu$-optimal policy.

Second, we show that for any $\mu$-optimal policy $\pi^*$, if $\theta_{\pi^*,\mu}(s) > 0$ for some $s \in S$, then $\mu(s) > 0$. We show it by contradiction. Suppose that $\mu(s) = 0$ but $\theta_{\pi^*,\mu}(s) > 0$. Then with positive probability, there exists some trajectory $\{s_0, s_1, \cdots, s_T\}$ with $s_T = s$ generated by the $\mu$-optimal policy $\pi^*$. Note that $\mu(s_T) = 0$ implies that there must exist some $t \le T$, such that $\mu(s_t) = 0$ (with $t = T$ being the largest one). Let $t_0$ be the smallest $t$ such that $\mu(s_t) = 0$. Then we have $t_0 > 0$ since the initial distribution that generates $\theta_{\pi^*,\mu}$ is $\mu$, i.e., $s_0$ is sampled from $\mu$ and thus $\mu(s_0) > 0$. We thus have $\mu(s_{t_0-1}) > 0$ by the definition of $t_0$. Moreover, there must exist some $a \in A$ such that $\pi^*(a \mid s_{t_0-1}) > 0$ and $P_{s_{t_0-1},a}(s_{t_0}) > 0$, since we have observed the transition from $s_{t_0-1}$ to $s_{t_0}$. By the definition of $\mu$-policy, we have

$\mu(s_{t_0-1}, a) > 0$ because $\pi^*(a \mid s_{t_0-1}) > 0$ for this $a$. We thus have $\mu(s_{t_0}) > 0$ by (40), which contradicts the assumption. Hence we have shown that $\theta_{\pi^*,\mu}(s) > 0$ can imply $\mu(s) > 0$.

Third, the second point also implies that for any $\mu$-optimal policy $\pi^*$ (including the max-$\mu$-optimal policy), $\theta_{\pi^*,\mu}(s) = 0$ for any $s \notin S_0$. Consequently, $\theta_{\pi^*,\mu}(s,a) > 0$ implies $s \in S_0$ and $\pi^*(a \mid s) > 0$, which further implies that $\mu(s,a) > 0$ since $\pi^*$ is a $\mu$-policy. Combining these three points, we obtain our Assumption 5.

Finally, we note that if the $\mu$-optimal policy is *unique*, Assumption 5 is reduced to the specific single-policy concentrability assumption used in (Chen & Jiang, 2022). In particular, (Chen & Jiang, 2022) directly assumes that the optimal policy of the original problem is unique.

### C.3. Proof of Proposition 2

The first part directly follows from Assumption 5. Next, we prove the second part. For any $\mu$-optimal policy $\pi$, we define $C^\pi = \max_{s\in S} \frac{\theta_\pi(s)}{\mu(s)}$.

We first prove that $C^\pi$ is finite if $\pi$ is a $\mu$-optimal policy. Since $\mu$ is fixed, we have $\mu(s) > 0$ for any $s \in S_0$. For $s \notin S_0$, by Assumption 5, we have $\theta_\pi(s) = 0$. Also we have $\theta_\pi(s)$ is upper bounded by 1. Then $C^\pi$ is finite since we let $0/0 = 0$. Then we just let $C_{\max} = \sup_{\pi:\mu-optimal} C^\pi$.

### C.4. Discussion on Corollary 1

Note that Corollary 1 additionally requires the coverage of $\mu_0$ by $\mu$, which we argue is a mild assumption in the following sense:

1. Recall that $S_0$ is the set of states that can be visited by $\mu$, i.e., $S_0 = \{s \in S \mid \mu(s) > 0\}$. This means that we only have data for states $s \in S_0$, and it seems not plausible to learn anything outside $S_0$ from data, without additional assumptions on the correlation among states. Therefore, we can not expect to deal with initial states outside $S_0$ and hence it is reasonable to only consider the initial distribution $\mu_0$ that is covered by $\mu$;

2. The commonly assumed single-policy concentrability in (Zhan et al., 2022; Chen & Jiang, 2022; Rashidinejad et al., 2021; 2022) (and our Assumption 2) implies that $\mu_0$ is covered by $\mu$, because

$$\begin{aligned}\max_{s\in S}\frac{\mu_0(s)}{\mu(s)} &= \frac{1}{1-\gamma}\cdot\max_{s\in S}\frac{(1-\gamma)\sum_a \mu_0(s)\pi^*(a\mid s)}{\sum_a \mu(s,a)} \le \frac{1}{1-\gamma}\cdot\max_{s\in S}\frac{\sum_a \theta_{\pi^*,\mu_0}(s,a)}{\sum_a \mu(s,a)}\\ &\le \frac{1}{1-\gamma}\cdot\max_{s\in S, a\in A}\frac{\theta_{\pi^*,\mu_0}(s,a)}{\mu(s,a)} \le \frac{C_{\pi^*,\mu_0}}{1-\gamma} =: C_\mu.\end{aligned}$$

3. As stated in (Liu et al., 2018; Tang et al., 2019; Levine et al., 2020; Zhan et al., 2022), $\mu$ usually can be viewed as a valid occupancy measure under some behavior policy of $\pi_\mu$, starting from $\mu_0$. In this case, $C_\mu$ exists and satisfies $C_\mu \le 1/(1-\gamma)$.

Compared to (Chen & Jiang, 2022), they require $\arg\max_a Q^*(s,a)$ to be unique for any $s$, and the algorithm is not computationally tractable. Also, note that our algorithm does not require the knowledge of the gap $\Delta_Q$. Compared to (Zhan et al., 2022), we only need some single-policy concentrability assumption for the *original* problem, instead of the regularized problem, together with only the realizability assumption on the function classes. Moreover, our sample complexity is $\mathcal{O}(1/\epsilon^2)$ with a gap dependence, while that in (Zhan et al., 2022) is $\mathcal{O}(1/\epsilon^6)$.

Note that (Zhan et al., 2022) also considered the vanilla version of the minimax formulation without regularization (see Section 4.5 therein). However, their analysis requires *all-policy-concentrability* assumption, which is stronger than our assumption that only requires to cover some single optimal policy. Finally, compared with the concurrent work (Rashidinejad et al., 2022) (and also our results in Section 3), which also achieved $\mathcal{O}(1/\epsilon^2)$ sample complexity, our result here is gap-dependent and does not rely on any completeness-type assumption.

### C.5. Proof of Lemma 6

We have

$$
\begin{aligned}
&\Delta(w_{\mathcal{D}}) = \ell(w_{\mathcal{D}}) - \ell(w^*) \\
&\geq \ell(w_{\mathcal{D}}, v^*) - \ell(w^*, v^*) \\
&= \left( -\sum_{s,a} r(s,a)\mu(s,a)w_{\mathcal{D}}(s,a) + \sum_{s,a} w_{\mathcal{D}}(s,a)\mu(s,a)\left(v^*(s) - \gamma \sum_{s' \in S} P_{s,a}(s')v^*(s')\right) - (1-\gamma){v^*}^\top \mu_0 \right) \\
&\quad - \left( -\sum_{s,a} r(s,a)\mu(s,a)w^*(s,a) + \sum_{s,a} w^*(s,a)\mu(s,a)\left(v^*(s) - \gamma \sum_{s' \in S} P_{s,a}(s')v^*(s')\right) - (1-\gamma){v^*}^\top \mu_0 \right) \\
&\qquad = \sum_{s,a} \left(w_{\mathcal{D}}(s,a)\mu(s,a) - w^*(s,a)\mu(s,a)\right)\left(v^*(s) - \left(r(s,a) + \gamma \sum_{s'} P_{s,a}(s')v^*(s')\right)\right) \\
&\qquad = \sum_{s,a} \left(w_{\mathcal{D}}(s,a)\mu(s,a) - w^*(s,a)\mu(s,a)\right)\left(v^*(s) - Q^*(s,a)\right) \\
&\qquad \geq \Delta_Q \sum_{(s,a)\in\mathcal{I}} w_{\mathcal{D}}(s,a)\mu(s,a),
\end{aligned} \tag{41}
$$

where the first inequality is due to the definition of $\ell(\cdot)$, the second to forth equalities are due to the definitions of $\ell(\cdot)$ and $Q^*$. The second inequality uses the fact that $w^*(s,a)\mu(s,a) = \theta_{\pi^*}(s,a) = 0$ for $(s,a) \in \mathcal{I}$ (see Lemma 4) and the definition of $\Delta_Q$. This completes the proof. □

### C.6. Proof of Lemma 5

The proof is given by the following lemma, along with the fact that $\Delta_{\mathcal{D}}(w_{\mathcal{D}}) = 0$

**Lemma 11.** *Suppose Assumptions 3, 4 hold. With probability at least $1-\delta$, we have*

$$
|\Delta(w) - \Delta_{\mathcal{D}}(w)| \leq \frac{4\sqrt{2}B_w\sqrt{\log(|V||W|/\delta)}}{(1-\gamma)\sqrt{n}}
$$

*for any $w \in W$.*

*Proof.* By definition, we have:

$$
\begin{aligned}
|\Delta(w) - \Delta_{\mathcal{D}}(w)| &\leq |\ell(w) - \ell_{\mathcal{D}}(w)| + |\min_{w\in W} \ell_{\mathcal{D}}(w) - \min_{w\in W} \ell(w)| \\
&= \left|\max_{v\in V} \ell(w,v) - \max_{v\in V} \ell_{\mathcal{D}}(w,v)\right| + \left|\min_{w\in W}\max_{v\in V} \ell_{\mathcal{D}}(w,v) - \min_{w\in W}\max_{v\in V} \ell(w,v)\right| \\
&\leq \max_{v\in V}\left|\ell(w,v) - \ell_{\mathcal{D}}(w,v)\right| + \max_{w\in W}\max_{v\in V}\left|\ell_{\mathcal{D}}(w,v) - \ell(w,v)\right|.
\end{aligned} \tag{42}
$$

First, we can bound each term using Hoeffding's inequality, for each $w, v$. We have

$$
|\ell(w,v) - \ell_{\mathcal{D}}(w,v)| \leq |(u_{\mathcal{D}} - u)^\top w| + |v^\top(K - K_{\mathcal{D}})w| + (1-\gamma)|v^\top(\mu_{\mathcal{D}} - \mu)|. \tag{43}
$$

Now, with probability at least $1 - \delta/(|V||W|)$, we have:

$$
\begin{aligned}
|(u_{\mathcal{D}} - u)^\top w| &\leq \sqrt{2}B_w \frac{\sqrt{\log(|V||W|/\delta)}}{\sqrt{n}} \\
|v^\top(K - K_{\mathcal{D}})w| &\leq 2\sqrt{2}B_w \frac{\sqrt{\log(|V||W|/\delta)}}{(1-\gamma)\sqrt{n}} \\
(1-\gamma)|v^\top(\mu_{\mathcal{D}} - \mu)| &\leq \sqrt{2}\frac{\sqrt{\log(|V||W|/\delta)}}{\sqrt{n}}.
\end{aligned} \tag{44}
$$

Finally, taking a union bound over all $w \in W$ and $v \in V$, and noting that $B_w \geq 1$ since $w^* \in W$ and $B_w \geq \|w^*\|_\infty \geq 1$, we get the desired result. □

### C.7. Proof of Lemma 7

Recall $\theta_{\mathcal{D}}(s,a) = w_{\mathcal{D}}(s,a)\mu(s,a)$. Next, we define a policy $\tilde{\pi}^*$ as follows:

1. For any $s \in S_0$ and $\tilde{a}$ such that $\tilde{a} \in \mathcal{T}(s)$ and $\mu(s,\tilde{a}) > 0$, we let $\tilde{\theta}(s,\tilde{a}) = \theta_{\mathcal{D}}(s,\tilde{a}) + \frac{1}{|\mathcal{T}(s)|}\sum_{a':(s,a')\in\mathcal{I}} \theta_{\mathcal{D}}(s,a')$.

   For any other $a' \neq \tilde{a}$ with $(s,a') \in S \times A \setminus \mathcal{I}$, we let $\tilde{\theta}(s,a') = \theta_{\mathcal{D}}(s,a')$. Finally, for any other $a'$ with $(s,a') \in \mathcal{I}$, we let $\tilde{\theta}(s,a') = 0$.

   Then for $s \in S_0$, $\tilde{\pi}^*(a \mid s)$ is generated by normalizing $\tilde{\theta}(s,\cdot)$, i.e., $\tilde{\pi}^*(a \mid s) = \tilde{\theta}(s,a)/\sum_{a'\in A}\tilde{\theta}(s,a')$. Note that by definition we have $\sum_{a'\in A}\tilde{\theta}(s,a') = \sum_{a'\in A}\theta_{\mathcal{D}}(s,a')$, and by the lower bound constraint in (20), we have

$$\sum_{a'\in A}\theta_{\mathcal{D}}(s,a') = \sum_{a'\in A} w_{\mathcal{D}}(s,a')\pi_\mu(a' \mid s)\mu(s) \geq (1-\gamma)\mu(s) > 0.$$

   Hence, the normalization of obtaining $\tilde{\pi}^*(a \mid s)$ is not degenerate in the sense that $\sum_{a'\in A}\tilde{\theta}(s,a') > 0$.

2. For any $s \notin S_0$, we choose any $\hat{a}$ such that $(s,\hat{a}) \in S \times A \setminus \mathcal{I}$, i.e., $\hat{a}$ that maximizes $Q^*(s,a)$. Then we let $\tilde{\pi}^*(\hat{a} \mid s) = 1$ and set $\tilde{\pi}^*(a' \mid s) = 0$ for any other $a' \in A$.

Note that this $\tilde{\pi}^*$ is an optimal policy by construction and by Lemma 4. Moreover, the next lemma shows that $\tilde{\pi}^*$ is a $\mu$-optimal policy.

**Lemma 12.** *$\tilde{\pi}^*$ is a $\mu$-optimal policy. Furthermore, $\theta_{\tilde{\pi}^*}(s,\cdot) = 0$ for any $s \notin S_0$.*

*Proof.* By the construction of $\tilde{\pi}^*$, we know that for any $s \in S_0$, we have $\mu(s,a) > 0$ if $\tilde{\pi}^*(a \mid s) > 0$. Also we know that $\tilde{\pi}^*$ is an optimal policy due to Lemma 4. Then $\tilde{\pi}^*$ is a $\mu$-optimal policy. The second part follows directly from Assumption 5. Specifically, first, Assumption 5 implies that for any $s \notin S_0$, since $\mu(s,a) = 0$ for all $a \in A$, we have $\pi^*(a \mid s) = 0$ for all $a \in A$. Then, by the definition of max-$\mu$-optimal policy, the states visited by the $\mu$-optimal policy $\tilde{\pi}^*$ should be visited by the max-$\mu$-optimal policy $\pi^*$. This is because the max-$\mu$-policy $\pi^*$ can take all actions that $\tilde{\pi}^*$ can take. Hence, whenever $\pi^*(a \mid s) = 0$, we should have $\tilde{\pi}^*(a \mid s) = 0$ and thus $\theta_{\tilde{\pi}^*}(s,a) = 0$. This completes the proof. □

Finally, we prove Lemma 7 using Lemmas 6, 12, Proposition 2, and the performance difference lemma (Kakade & Langford, 2002).

*Proof of Lemma 7.* We use performance difference lemma to $\tilde{\pi}^*, \pi_{\mathcal{D}}$ to obtain

$$J_\mu(\tilde{\pi}^*) - J_\mu(\pi_{\mathcal{D}}) \leq \frac{1}{1-\gamma}\sum_{s\in S}\theta_{\tilde{\pi}^*}(s)\|\pi_{\mathcal{D}}(\cdot \mid s) - \tilde{\pi}^*(\cdot \mid s)\|_1.$$

Because $\theta_{\tilde{\pi}^*}(s,a) = 0$ for $s \notin S_0$ by Lemma 12, we have

$$J_\mu(\tilde{\pi}^*) - J_\mu(\pi_{\mathcal{D}}) \leq \frac{1}{1-\gamma}\sum_{s\in S_0}\theta_{\tilde{\pi}^*}(s)\|\pi_{\mathcal{D}}(\cdot \mid s) - \tilde{\pi}^*(\cdot \mid s)\|_1.$$

By the construction of $\tilde{\pi}^*$, we have

$$\begin{aligned}
J_\mu(\tilde{\pi}^*) - J_\mu(\pi_{\mathcal{D}}) &\leq \frac{1}{1-\gamma}\sum_{s\in S_0}\theta_{\tilde{\pi}^*}(s)\|\pi_{\mathcal{D}}(\cdot \mid s) - \tilde{\pi}^*(\cdot \mid s)\|_1 && (45)\\
&=^{(*1)} \frac{2}{1-\gamma}\sum_{s\in S_0}\frac{\theta_{\tilde{\pi}^*}(s)}{\sum_a \theta_{\mathcal{D}}(s,a)} \cdot \sum_{a:(s,a)\in\mathcal{I}}\theta_{\mathcal{D}}(s,a) && (46)\\
&= \frac{2}{1-\gamma}\sum_{s\in S_0}\frac{\theta_{\tilde{\pi}^*}(s)/\mu(s)}{\sum_a \theta_{\mathcal{D}}(s,a)/\mu(s)} \cdot \sum_{a:(s,a)\in\mathcal{I}}\theta_{\mathcal{D}}(s,a) && (47)\\
&\leq^{(*2)} \frac{2C_{\max}}{(1-\gamma)^2}\sum_{(s,a)\in\mathcal{I}}\theta_{\mathcal{D}}(s,a) && (48)\\
&\leq^{(*3)} \frac{2C_{\max}}{(1-\gamma)^2\Delta_Q}\cdot\Delta(w_{\mathcal{D}}), && (49)
\end{aligned}$$

where $(*1)$ follows from the construction of $\tilde{\pi}^*$ from $\theta_{\mathcal{D}}$, $(*2)$ is because of Proposition 2 and definition of $S_0$, and $(*3)$ is due to Lemma 6. □

## D. Extensions

### D.1. The Case where the Behavior Policy is Unknown

In the previous two sections, we assume that the behavior policy $\pi_\mu$ is known. If the behavior policy is unknown to us, it is not easy to attain a policy corresponding to $w_{\mathcal{D}}$. To tackle this issue, we can use behavior cloning as in (Zhan et al., 2022). We omit the details and the reader can refer to (Zhan et al., 2022) for more details. Instead of this behavior cloning approach, in this section, we propose a method that can attain the same accuracy as solving Programs (15) and (19) by adding additional $L_0$ constraints. The approach is based on a simple idea: If $\|w_{\mathcal{D}}(s,\cdot)\|_0 = 1$ for any $s$, then we can compute $\pi_{w_{\mathcal{D}}}(s,\cdot)$ without knowing $\pi_\mu$. Concretely speaking, $\pi_{w_{\mathcal{D}}}(s,a) = 1$ if $w_{\mathcal{D}}(s,a) \neq 0$ and $\pi_{w_{\mathcal{D}}}(s,a) = 0$ otherwise. Here $\|\cdot\|_0$ means the number of nonzero elements of a vector. The policy $\pi_{w_{\mathcal{D}}}(\cdot|s)$ is just the normalization vector of $\{w_{\mathcal{D}}(s,a)\pi_\mu(a|s)\}$.

#### D.1.1. Modification of Program (15)

We first make a slight change to Assumption 1. We assume that $W$ realizes a deterministic optimal policy instead of an arbitrary optimal policy.

**Assumption 6.** *Let $x_w := \phi(w)$ with $\phi$ given in Definition 3. Let $W$ and $B$ be the function classes for $w$ and $x_w$, respectively. Then, we have realizability of $W$ for a $w^*$ corresponding to a deterministic optimal policy, and $(W, B)$-completeness under $\phi$, i.e., $w^* \in W$ for some optimal $w^*$ such that $\pi_{w^*}$ is a deterministic optimal policy and $x_w \in B$ for all $w \in W$. Furthermore, we assume that $W$ and $B$ are bounded, i.e., $\|w\|_\infty \le B_w$ for all $w \in W$ and[8] $\|x\|_\infty \le 1$ for all $x \in B$.*

Then we add an $L_0$ constraint to (15) as follows:

$$\begin{aligned} &\max_{w\in\mathbb{R}^m_+} \quad u_{\mathcal{D}}^\top w \\ \text{s.t.} \quad & x^\top(K_{\mathcal{D}}w - (1-\gamma)\mu_0) \le E_{n,\delta}, \quad \forall x \in B \\ & w \in W, \\ & \|w(s,\cdot)\|_0 \le 1, \forall s \in S. \end{aligned} \tag{50}$$

Let the solution of the above problem to be $w_{\mathcal{D}}$. We can compute $\pi_{\mathcal{D}} = \pi_{w_{\mathcal{D}}}$ as $\pi_{w_{\mathcal{D}}}(a|s) = 1$ if $w(s,a) > 0$ and $\pi_{w_{\mathcal{D}}}(a|s) = 0$ otherwise. Let $\bar{w} = W \cap \{w \mid \|w(s,\cdot)\|_0 \le 1, \forall s\}$. Then we have the following theorem:

**Theorem 4.** *Suppose Assumptions 2 and 6 hold. Then, we have*

$$J_{\mu_0}(\pi^*) - J_{\mu_0}(\pi_{\mathcal{D}}) \le \frac{2\sqrt{2}B_w\sqrt{\log(|B||\bar{W}|/\delta)}}{(1-\gamma)\sqrt{n}}$$

*with probability at least $1-6\delta$.*

The proof is the same as Theorem 1.

#### D.1.2. Modification of Program (19)

Let $W$ satisfy

$$W \subseteq \Big\{w : w \in \mathbb{R}^m_+,\ \|w(s,\cdot)\|_0 \le 1,\ \max_{a\in A} w(s,a) \ge (1-\gamma),\ \text{for all } s \in S\Big\}. \tag{51}$$

First, we slightly change Assumption 3 such that the function classes contain at least one deterministic optimal policy.

**Assumption 7** (Realizability and Boundedness of $W$)**.** *There exists some solution $w^* \in W \subseteq \mathbb{R}^m_+$ corresponding to a deterministic optimal policy $\pi^*$ solving* (18) *and hence solving* (11)*. Moreover, we suppose $\|w\|_\infty \le B_w$ for all $w \in W$.*

Then we solve the following minimax problem using $W$ defined in (7)

$$\min_{w\in W}\max_{v\in V} -u^\top w + v^\top(Kw - (1-\gamma)\mu), \tag{52}$$

Suppose $w_{\mathcal{D}}$ is a solution. We let $\pi_{\mathcal{D}} = \pi_{w_{\mathcal{D}}}$. Then we have the following result:

[8] Note that $w^* \in W$ implies that $B_w \ge 1$ since $w^*$ is a supremum of the ratio between two distributions.

**Theorem 5.** *Under Assumptions 4, 5, and 7, we have*

$$J_\mu(\pi^*) - J_\mu(\pi_{\mathcal{D}}) \leq \frac{8\sqrt{2}B_w C_{\max}}{\Delta_Q(1-\gamma)^3} \cdot \frac{\sqrt{\log(|W||V|/\delta)}}{\sqrt{n}},$$

*with probability* $\geq 1-\delta$.

The proof is the same as Theorem 3.

**Remark 3.** *Similarly, this approach does not require to know the behavior policy* $\pi_\mu$ *but it might be harder to solve since* (52) *is a mixed integer programming.*

### D.2. Reducing $C_{\max}$

In Theorem 3, the error bound of $J(\pi)$ depends on $C_{\max}$, which depends on all $\mu$-optimal policies as stated in the proof of Proposition 3. In this subsection, we discuss how to modify the algorithm such that the error bound can only depend on one optimal policy. To do this, we need to slightly strengthen the realizability assumption in Section 4. We first give the following definition:

**Definition 9.** *A uniform* $\mu$*-max optimal policy is a* $\mu$*-max optimal policy that takes the same probability over all* $a \in \mathcal{T}(s)$ *for any* $s$.

**Assumption 8** (Realizability and Boundedness of $W$)**.** *There exists some solution* $w^* \in W \subseteq \mathbb{R}_+^m$ *solving* (18) *and hence solving* (11) *such that* $\pi_{w^*}$ *is a uniform* $\mu$*-max optimal policy. Moreover, we suppose* $\|w\|_\infty \leq B_w$ *for all* $w \in W$.

Then we modify Program (19) as follows:

$$\begin{aligned} &\min_{w\in W} \max_{v \in V} \; -u^\top w + v^\top (Kw - (1-\gamma)\mu) \\ \text{s.t.} \quad & w(s,a)\pi_\mu(a|s) = y(s)\lambda(s,a), \lambda(s,a) \in \{0,1\}. \end{aligned} \tag{53}$$

Suppose the solution is $w_{\mathcal{D}}$. Define $\bar{W} = W \cap \{w \mid w(s,a)\pi_\mu(a|s) = y(s)\lambda(s,a), \lambda(s,a) \in \{0,1\}\}$. In other words, $\bar{W}$ contains $w$ such that the nonzero elements of $w(s,\cdot)$ are the same. Then we have the error bound for $\pi_{w_{\mathcal{D}}}$.

**Theorem 6.** *Under Assumptions 3, 5, and 8, we have*

$$J_\mu(\pi^*) - J_\mu(\pi_{\mathcal{D}}) \leq \frac{8\sqrt{2}B_w^2}{\Delta_Q(1-\gamma)^3} \cdot \frac{\sqrt{\log(|\bar{W}||V|/\delta)}}{\sqrt{n}},$$

*with probability* $\geq 1-\delta$.

*Proof.* The proof is just following the same strategies as the proof of Theorem 3. The only improvement is that we can prove $\tilde{\pi}^*$ is a uniform $\mu$-max policy because the nonzero elements of $\tilde{\pi}^*(\cdot \mid s)$ are the same for any $s$. Hence Equation (48) in the proof of Theorem 3 can be modified as follows:

$$\frac{2}{1-\gamma} \sum_{s \in S_0} \frac{\theta_{\tilde{\pi}^*}(s)/\mu(s)}{\sum_a \theta_{\mathcal{D}}(s,a)/\mu(s)} \cdot \sum_{a:(s,a)\in\mathcal{I}} \theta_{\mathcal{D}}(s,a) \leq \frac{2B_w}{(1-\gamma)^2} \sum_{(s,a)\in\mathcal{I}} \theta_{\mathcal{D}}(s,a). \tag{54}$$

The rest of the proof is identical to the proof of Theorem 3. □

**Remark 4.** *Notice that the error bound only depends on* $B_w$*, which only need to be larger than the ratio between* $\mu$ *and the occupancy of the uniform* $\mu$*-max policy. Hence, the statistical bound is improved. However, the integer variable* $\lambda(s,a)$ *makes* (53) *a mixed integer programming problem, which is more difficult to solve.*