The Power of Duality Principle in Offline Average-Reward Reinforcement Learning

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

Abstract

Offline reinforcement learning (RL) is widely used to find an optimal policy using a pre-collected dataset, without further interaction with the environment. Recent RL theory has made significant progress in developing sample-efficient offline RL algorithms with various relaxed assumptions on data coverage, with specific focuses on either infinite-horizon discounted or finite-horizon episodic Markov decision processes (MDPs). In this work, we revisit the LP framework and the induced duality principle for offline RL, specifically for *infinite-horizon average-reward* MDPs. By virtue of this LP formulation and the duality principle, our result achieves the $\tilde{O}(1/\sqrt{n})$ near-optimal rate under partial data coverage assumptions. Our key enabler is to *relax* the equality *constraint* and introduce proper new *inequality constraints* in the dual formulation of the LP. We hope our insights can shed new lights on the use of LP formulations and the induced duality principle, in offline RL.

1. Introduction

Reinforcement Learning (RL) has achieved remarkable empirical success in solving sequential decision-making problems in recent years (Mnih et al., 2015; Silver et al., 2016; Vinyals et al., 2017; Levine et al., 2016). Two key factors have contributed to these successes: 1) the use of powerful function approximators, such as deep neural networks, and 2) access to large amounts of interaction data with the environment. Many successful RL applications rely on online data collection through simulators, such as game engines (Silver et al., 2016; Vinyals et al., 2017) and physics simulators (Todorov et al., 2012).

However, in numerous real-world domains, online interaction is impractical due to the high cost or impracticality of data collection, or the inability to simulate the environment accurately. Examples of such domains include 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 cases where online interaction is available, leveraging previously collected data is crucial for effective generalization, as it requires large datasets (Levine et al., 2020). Offline RL has emerged as a promising framework for deploying RL in real-world scenarios.

Most existing works on offline RL focus on learning in discounted reward Markov Decision Processes (MDPs) with a discount factor $\gamma < 1$. However, in many applications, learning problems are modeled as average-reward MDPs (Sutton & Barto, 2018)[Chapter 10], (Naik et al., 2019; Farias et al., 2022). Despite the rich literature on offline RL for discounted MDPs, to the best of our knowledge, there is no existing result studying offline RL for average-reward MDPs that achieve optimal sample complexity. In this paper, we address offline RL in average-reward MDPs and design algorithms with a statistical convergence rate of $\tilde{\mathcal{O}}(1/\sqrt{n})$.

Offline RL in discounted MDPs. Offline RL is known to suffer from the training instability issue caused by the distribution shift between the offline data distribution and the target (optimal) policy distribution (Fujimoto et al., 2019; Kumar et al., 2020). Consequently, earlier offline RL works relied on strong assumptions regarding the dataset to provide sample-efficiency guarantees. Many of these results (Munos & Szepesvári, 2008; Scherrer, 2014; Chen & Jiang, 2019; Zhang et al., 2021) required the dataset to have full coverage, meaning that the data covers the state distributions induced by all policies. These assumptions are strong and can be violated when using richer function classes, and they are more stringent than the common assumption of realizability in statistical learning theory. Moreover, the full coverage assumption requires the offline data to cover all possible state-action pairs, which is often violated in real-world applications. Recent advancements have made progress in relaxing these assumptions. For example, (Liu et al., 2020; Jin et al., 2020; Rashidinejad et al., 2021; Xie et al., 2021;

[*]Equal contribution [1]Department of EECS, Massachusetts Institute of Technology [2]Department of ECE, University of Maryland, College Park. Correspondence to: Jiawei Zhang <jwzhang@mit.edu>.

ICML 2023 Workshop on Duality for Modern Machine Learning, Honolulu, Hawaii, USA. Copyright 2023 by the author(s).

Uehara & Sun, 2021; Chen & Jiang, 2022) have shown that by employing a pessimistic mechanism that selects the worst-case value function or model from an uncertainty set during learning, the full coverage assumption can be relaxed to a single-policy coverage assumption. More recently, (Zhan et al., 2022; Rashidinejad et al., 2022; Ozdaglar et al., 2023) propose linear programming-based methods that relax the data coverage assumptions and provide computationally tractable algorithms when using function approximation.

RL in average-reward MDPs using LP. To the best of our knowledge, there is no finite-sample complexity results for *average-reward offline* RL, with partial data coverage. The most related results which also exploited the LP framework and duality principle in average-reward RL, are those for the case with a generative model (Wang, 2020; Jin & Sidford, 2020).

1.1. Our Goal and Challenges

For discounted MDPs, several algorithms achieve a nearly optimal convergence rate of $\tilde{\mathcal{O}}(1/\sqrt{n})$. However, it remains unknown whether we can achieve this rate for average-reward MDPs. The convergence rate for discounted MDPs with a discount factor γ is always in the form of $\mathcal{O}(1/\sqrt{\text{poly}(1-\gamma)n})$, which can become large as $\gamma \to 1$. Consequently, the results and analysis for discounted MDPs cannot be directly applied to average-reward MDPs.

Contributions. In this paper, we propose a linear programming-based algorithm for solving offline RL in average-reward MDPs. Our algorithm achieves a convergence rate of $\tilde{\mathcal{O}}(t_{\max}/\sqrt{n})$ under the single-policy coverage assumption. Here, n represents the number of samples, and $t_{\max}$ denotes the worst-case mixing time of the MDP.

Our techniques. The main idea behind our approach is to study properly constrained versions of the linear programming (LP) reformulation of the underlying MDP. Specifically, we focus on 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 density ratio, which is the ratio between the state-action occupancy measure and the offline data distribution. A distinctive feature of our algorithm is the relaxation of equality constraints in the empirical LP and the boundedness constraints on the density ratio. The convergence result relies on a crucial error bound lemma that relates the suboptimality of the value function to the ℓ_1-norm violation of the validity constraint on the stationary distribution in the LP (see Lemma 5). Additionally, concentration bounds on the constraints and objective functions play a crucial role in our analysis.

2. Background

2.1. Model and Setup

Markov decision processes. Consider an infinite-horizon MDP characterized by a tuple $\langle S, A, P, R, \mu_0 \rangle$, 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, and $\mu_0 \in \Delta(S)$ denotes the initial state distribution. Let $\pi : S \to \Delta(A)$ denote a Markov stationary policy of the agent, determining the distribution over actions at each state. We denote $P_\pi \in \mathbb{R}^{|S| \times |S|}$ to be the transition matrix corresponding to policy π. Each π leads to *stationary distributions* over the state spaces and state-action spaces, denoted by β_π and θ_π, where $\beta_\pi \in \mathbb{R}^{|S|}$ denote the stationary distribution of transition matrix P_π and $\theta_\pi(s,a) = \beta_\pi(s)\pi(a|s)$. We assume that the MDP is ergodic and irreducible for any π. Define $J_{\mu_0}(\pi) = E_{s_0 \sim \mu_0} \lim_{T\to\infty} \frac{1}{T} E_\pi r(s_t, a_t)$, where (s_t, a_t) follows the policy π.

Note that J_{μ_0} is independent of μ_0 due to ergodicity. We let $J(\pi) = J_{\mu_0}(\pi)$, which can also be represented as

$J(\pi) = r^\top \theta_\pi$, where

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

The goal of reinforcement learning is to learn a π^* that maximizes $J(\pi)$. 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 μ 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 $\mu(s) = 0$. π_μ can also be defined as the *behavior policy* if μ happens to correspond to the stationary distribution 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).

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

[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.

that $(s_i, a_i, s'_i) = (s, a, s')$ and $s_i = s$, respectively. We define the empirical version of μ, 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 π , and the given offline data distribution μ, we define $C_\pi > 0$ to be the policy concentrability coefficient, which is the smallest upper-bound such that $\frac{\theta_\pi(s,a)}{\mu(s,a)} \leq C_\pi$ for all $(s,a) \in S \times A$.*

Note that C_π characterizes how well the trajectory generated by the policy π is covered by the offline data. Throughout, we make the following assumption:

Assumption 1. *$C_{\pi^*} \leq C^* < \infty$ for some constant C^*.*

Our goal: finding an $\mathcal{O}(1/\sqrt{n})$ nearly optimal policy For discounted MDP, a number of algorithms have been proposed to achieve a nearly optimal accuracy, say, $\mathcal{O}(1/\sqrt{n}$ for offline RL for a fixed discounted factor γ. A natural question arises: *can we find a nearly optimal policy for AMDP by offline RL approach with $\mathcal{O}(1/\sqrt{n})$ accuracy?*

The naive approach: Using discounted MDP to approximate. As stated before, there is a rich literature about offline RL algorithms in discounted MDP with $\frac{1}{\sqrt{\mathrm{poly}((1-\gamma)n)}}$ accuracy, where γ is the discounted factor and $\mathrm{poly}(\cdot)$ means polynomial order. Moreover, by (Jin & Sidford, 2021; Wang et al., 2022), we know that there is a $(1-\gamma)$ gap between the optimal policy in discounted MDP and that in average-reward MDP. Therefore, we should take $1-\gamma \leq \mathcal{O}(1/\sqrt{n})$ if we require an $\mathcal{O}(1/\sqrt{n})$ accuracy for average-reward MDP. However, this results in at least $1/n^{1/4}$ sub-optimality gap for the discounted MDP. Therefore, this naive approach does not help us find a nearly optimal policy with $1/\sqrt{n}$ accuracy.

2.2. LP-based Reformulations

Let $P_{(s,a)} = [P_{s,a}(s^1), \cdots, P_{s,a}(s^{|S|})]^\top \in \Delta(S)$ be 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 $\mathbf{1}_{|A|} = [1, 1, \cdots, 1]^\top \in \mathbb{R}^{|A|}$. Then define the matrix M as: $M := \mathrm{Diag}(\mathbf{1}_{|A|}^\top, \cdots, \mathbf{1}_{|A|}^\top) - P$. The optimality condition, i.e., the Bellman equation, of the average-reward MDP can be equivalently written as the following linear program (Puterman, 1994):

$$\begin{array}{ll} \min_{v,\boldsymbol{h}} & v \\ \text{s.t.} & v \cdot \mathbf{1} + (I - P_a)\boldsymbol{h} - r_a \geq 0, \quad \forall a \in A, \end{array} \tag{1}$$

where $P_a \in \mathbb{R}^{|S|\times|S|}$ denotes the matrix whose (s,s')-component is $P_{s,a}(s')$, and $r_a \in \mathbb{R}^{|S|}$ such that $r_a(s) = r(s,a)$. This can be derived by the definition of Bellman equation in the average-reward case.

It is well-known that the LP (1) admits the following *dual program*, which also gives the optimal solution of the problem (Puterman, 1994):

$$\begin{array}{ll} \max_\theta & r^\top\theta := \sum_{s\in S, a\in A} r(s,a)\cdot\theta(s,a) \\ \text{s.t.} & M\theta = 0, \quad \sum_{s,a}\theta(s,a) = 1, \end{array} \tag{2}$$

Note that the optimal solution of the dual problem corresponds to the *stationary distribution* corresponding to an optimal policy (see (Puterman, 1994)). Hence, we use the notation θ to denote the optimization variable of the dual problem.

The optimal θ^* can be used to generate a policy π_{θ^*}, where π_θ is defined as

$$\pi_\theta(a \mid s) = \frac{\theta(s,a)}{\sum_{a'\in A}\theta(s,a')}, \tag{3}$$

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 π_{θ^*} then corresponds to an optimal policy π^* of the MDP (Puterman, 1994).

To better study the relationship between the stationary distribution 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 stationary distribution 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 $K_{s',(s,a)}(s,(s,a)) = 1$, $K_{s',(s,a)}(s',(s,a)) = -1$ and all other entries are zeros. Define the distributions ν 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

$$K = \mathbb{E}_{(s,a,s')\sim\nu}K_{s',(s,a)}, \; K_{\mathcal{D}} = \mathbb{E}_{(s,a,s')\sim\nu_{\mathcal{D}}}K_{s',(s,a)}. \tag{4}$$

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 (2).

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 (2) as follows:

$$\max_{w \ge 0} u^\top w \quad \text{s.t.} \quad Kw = 0, \sum_{s,a} w(s,a)\mu(s,a) = 1. \tag{5}$$

Let w^* be the solution to (5), then we can obtain the optimal policy by computing π_{w^*}, where with a slight abuse of notation, π_w is defined as

$$\pi_w(a \mid s) := \begin{cases} \frac{w(s,a)\pi_\mu(a|s)}{\sum_{a' \in A} w(s,a')\pi_\mu(a'|s)}, & \text{if} \quad c > 0 \\ \frac{1}{|A|} & \text{if} \quad c = 0 \end{cases} \tag{6}$$

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

2.3. The empirical version

However, we only get access to the empirical estimates of K and u. Consider the empirical version of (5)

$$\begin{aligned} & \max_{w \ge 0} u_{\mathcal{D}}^T w \\ s.t. \quad & K_{\mathcal{D}} w = 0 \\ & \sum_{s,a} w(s,a)\mu_{\mathcal{D}}(s,a) = 1, \end{aligned} \tag{7}$$

where we recall the definition of $K_{\mathcal{D}}$ in (4), and define $u_{\mathcal{D}} \in \mathbb{R}^m$ as $u_{\mathcal{D}}(s,a) = r(s,a)\mu_{\mathcal{D}}(s,a)$, with $\mu_{\mathcal{D}}(s,a) = n_{\mathcal{D}}(s,a)/n$.

To ensure that the empirical and the population version of LP are closed to each other, we need to make use of concentration inequalities for $(K - K_{\mathcal{D}})w, (u - u_{\mathcal{D}})^T w$, which require certain boundedness of w.

Let $B_w \ge C^*$ be some upper bound of C^*. Combining Assumption 1, we insert the infinity norm constraint $\|w\|_\infty \le B_w$ to (7) and the resulting empirical formulation becomes:

$$\begin{aligned} & \max_{w \in W := [0, B_w]^{|S||A|}} u_{\mathcal{D}}^T w \\ s.t. \quad & K_{\mathcal{D}} w = 0 \\ & \sum_{s,a} w(s,a)\mu_{\mathcal{D}}(s,a) = 1. \end{aligned} \tag{8}$$

However, (8) may not have a feasible solution due to the additional constraint $w \in W = [0, B_w]^{|S||A|}$. To address this issue, in the next section we relax the equality constraints to guarantee the feasibility of the problem.

3. Average-Reward MDP Case

To make sure the feasibility of the empirical LP, we relax the equality constraints in (8) and solve the following empirical problem:

$$\begin{aligned} & \min_{w \in W} \ (-u_{\mathcal{D}}^T w) \\ s.t. \quad & \|K_{\mathcal{D}} w\|_1 \le E_{n,\delta}, \forall x \in B. \\ & |\sum_{s,a} w(s,a)\mu_{\mathcal{D}}(s,a) - 1| \le E_{n,\delta}, \end{aligned} \tag{9}$$

where $E_{n,\delta} := 2B_w\sqrt{|S|\log((2|A|+2)/\delta)}/\sqrt{n}$. Let $w_{\mathcal{D}}$ be the solution to the above problem. We can obtain the policy $\pi_{\mathcal{D}}$ by setting $\pi_{\mathcal{D}} = \pi_{\tilde{\theta}_{\mathcal{D}}}$, where for each $(s,a) \in S \times A$

$$\tilde{\theta}_{\mathcal{D}}(s,a) = w_{\mathcal{D}}(s,a)\pi_\mu(a \mid s). \tag{10}$$

Definition 2. *For a policy π, we let β_π be the stationary distribution of the transition matrix P_π. For some θ, we let $\beta_\theta \in \mathbb{R}^{|S|}$ satisfy $\beta_\theta(s) = \sum_a \theta(s,a)$.*

We make the following assumption on the worst case mixing time of the transition matrix.

Assumption 2. *There exists some $T_0 > 0$ such that $\|P_\pi^{T_0}\beta - \beta_\pi\|_1 \le 1/2$ for any π and any state distribution $\beta \in \Delta(S)$.*

By the above assumption, we have the following immediate result:

Lemma 2. *Let $k_0 \ge t_0 \log_2(1/E_{n,\delta})$. Then $\|P_\pi^{k_0}\beta - \beta_\pi\|_1 \le E_{n,\delta}$ for any π and any state distribution $\beta \in \Delta(S)$.*

Proof. For a vector ζ, define $\zeta^+ = \max(\zeta, 0)$ and $\zeta^- = \max(-\zeta, 0)$. Then we have $\zeta^+ - \zeta^- = \zeta$ and $\|\zeta\|_1 = \|\zeta^+\|_1 + \|\zeta^-\|_1$. It suffices to prove that for $j > 0$, $\|P_\pi^{jt_0}\beta - P_\pi^{kt_0}\beta'\|_1 \le (1/2)^{j-1}$. We prove itby induction. For $j = 1$, the result holds by Assumption 2 and triangular inequality. Suppose that the result holds for j. Let $\zeta = P_\pi^{jt_0}\beta - P_\pi^{jt_0}\beta'$. Then $\|\zeta\| \le \|P_\pi^{jt_0}(\beta -$

$\beta_\pi)\|_1 + \|P_\pi^{jt_0}(\beta' - \beta_\pi)\|_1 \le (1/2)^{j-1}$. Since $P_\pi^{jt_0}\beta$ and $P_\pi^{jt_0}\beta'$ are probability distributions, $\mathbf{1}^T\zeta = 0$. Therefore, $\|\zeta^+\|_1 = \mathbf{1}^T\zeta^+ = \mathbf{1}^T\zeta^- = \|\zeta^-\|_1$. Hence, $\|\zeta^-\|_1 = \|\zeta^+\|_1 \le (1/2)^j$ since $\|\zeta\|_1 = \|\zeta^+\|_1 + \|\zeta^-\|_1$. Let β^+ be a probability distribution such that $\beta^+ = \lambda\zeta^+$ for some $\lambda \ge 2^{j-1}$. Also let $\beta^- = \lambda\beta^-$ is also a probability distribution. Then

$$\begin{aligned}\|P_\pi^{(j+1)t_0}(\beta - \beta')\|_1 &= \|P_\pi^{t_0}\zeta\|_1 \quad (11)\\ &= \frac{1}{\lambda}\|P_\pi^{t_0}\beta^+ - P_\pi^{t_0}\beta^-\|_1 \quad (12)\\ &\le (1/2)^{j-1}\cdot(1/2) \quad (13)\\ &= (1/2)^j, \quad (14)\end{aligned}$$

completing the proof. □

Theorem 1. *We have*

$$J(\pi^*) - J(\pi_{\mathcal{D}}) \le \tilde{O}(k_0\sqrt{|S|}/\sqrt{n})$$

with probability at least $1 - 3\delta$.

3.1. Proof

Let $B = [-1, 1]^{|S|}$ be the L_∞-norm ball centered at 0.

First, we have the following concentration bounds similar to the ones in the proof of Theorem 2 in (Ozdaglar et al., 2023).

Lemma 3. *We have*

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

2. *For any* $w \in W$ *with probability* $\ge 1 - \delta$.
$$|(u - u_{\mathcal{D}})^\top w| \le 2B_w\sqrt{|S|\log((2|A|+2)/\delta)}/\sqrt{n}$$

3. *For any* $w \in W$ *with probability* $\ge 1 - \delta$.
$$|(\mu - \mu_{\mathcal{D}})^\top w| \le 2B_w\sqrt{|S|\log((2|A|+2)/\delta)}/\sqrt{n}$$

According to this lemma, we conclude that some optimal w^* is feasible to problem (9) with high probability.

Lemma 4. *There exists some optimal* $w^* = \theta_{\pi^*}/\mu(s,a)$ *for some optimal policy* π^* *such that the following three conditions hold with probability* $\ge 1 - 2\delta$:

1. $w^* \in W$;

2. $\|K_{\mathcal{D}}w^*\|_1 \le E_{n,\delta}$;

3. $|\sum_{s,a} w^*(s,a)\mu(s,a) - 1| \le E_{n,\delta}$.

Proof. First, by Assumption 1 and the definition of w, there exists some $w^* \in W$. We fix the w^* and notice that $Kw^* = 0$, $\sum_{s,a} w^*(s,a)\mu(s,a) = 1$. Then $\|K_{\mathcal{D}}w^*\|_1 = \|(K_{\mathcal{D}} - K)w^*\|_1 \le E_{n,\delta}$ with probability $\ge 1 - \delta$ and $|\sum_{s,a} w^*(s,a)\mu_{\mathcal{D}}(s,a) - 1| = |\sum_{s,a} w^*(s,a)(\mu_{\mathcal{D}} - \mu(s,a))| \le E_{n,\delta}$ with probability $\ge 1 - \delta$. Therefore, the result follows from union bound. □

By this lemma, we have

$$u_{\mathcal{D}}^T w_{\mathcal{D}} \ge u_{\mathcal{D}}^T w^*$$

with probability $\ge 1 - 2\delta$. According to Lemma 3, we further have

$$u^T w_{\mathcal{D}} \ge u^T w^* - E_{n,\delta} = J(\pi^*) - E_{n,\delta}. \quad (15)$$

We next relate $u^T w_{\mathcal{D}}$ to $J(\pi_{\mathcal{D}})$. Let $\theta \in \mathbb{R}^{|S||A|}$ satisfying $\theta(s,a) = w(s,a)\mu(s,a)$. We have the following lemma relating $r^T\theta = u^T w$ to $J(\pi_\theta) = J(\pi_w)$.

Lemma 5. 1. *If* $\theta \in \Delta(S \times A)$, *and* $(P - E)\theta = 0$, *we have*
$$r^T\theta = J(\pi_\theta).$$

2. *We have*
$$\begin{aligned}&\|\beta_\theta - \beta_{\pi_\theta}\|_1\\ &\le k_0\|(I - P_{\pi_\theta})\beta_\theta\|_1\\ &\qquad + 3k_0|\sum_{s,a} w(s,a)\mu(s,a) - 1| + E_{n,\delta}\\ &= k_0\|Kw\|_1 + 3k_0|\sum_{s,a} w(s,a)\mu(s,a) - 1| + E_{n,\delta}.\end{aligned} \quad (16)$$

Proof of Lemma 5. We only prove the second part and the first part is just a special case. First, it is easy to see that there exists some $\bar{\beta}_\theta$ such that

1. $\sum_s \bar{\beta}_\theta(s) = 1$ and $\bar{\beta}_\theta(s) \ge 0$;

2. $\|\beta_\theta - \bar{\beta}_\theta\|_1 \le |\sum_{s,a} w(s,a)\mu(s,a) - 1|$.

Notice that by Assumption 2,

$$\begin{aligned}\|\beta_{\pi_\theta} - \bar{\beta}_\theta\|_1 - E_{n,\delta} &\le \|\bar{\beta}_\theta - P_{\pi_\theta}^{k_0}\bar{\beta}_\theta\|_1\\ &\le k_0\|(I - P_{\pi_\theta})\bar{\beta}_\theta\|_1\\ &\le k_0\|(I - P_{\pi_\theta})\beta_\theta\|_1 + 2k_0\|\bar{\beta}_\theta - \beta_\theta\|_1,\end{aligned}$$

which, combined with triangle inequality, gives the final bound. □

Finally, we can prove our main result, Theorem 1.

Proof of Theorem 1. Let $w = w_{\mathcal{D}}$ and $\theta(s,a) = w(s,a)\mu(s,a)$. Also let $r_{\pi_{\mathcal{D}}} \in \mathbb{R}^{|S|}$ defined as $r_{\pi_{\mathcal{D}}}(s) = r(s,\cdot)^T \pi_{\mathcal{D}}(s,\cdot)$.

Then by Lemma 5, we have

$$\begin{aligned} & |J(\pi_{\mathcal{D}}) - u^T w_{\mathcal{D}}| \\ = \;& |r_{\pi_{\mathcal{D}}}^T(\beta_{\pi_\theta} - \beta_\theta)| \\ \le \;& \|\beta_\theta - \beta_{\pi_\theta}\|_1 \\ \le \;& k_0\|Kw_{\mathcal{D}}\|_1 + 3|\sum_{s,a} w_{\mathcal{D}}(s,a)\mu(s,a) - 1| + E_{n,\delta}, \end{aligned}$$

where the first inequality is because of the boundedness of r, and the second inequality is because of Lemma 5. By Lemma 3 as well as the constraints in (9), we further have

$$|J(\pi_{\mathcal{D}}) - u^T w_{\mathcal{D}}| \le 5k_0 E_{n,\delta}.$$

Finally combined with (15), we have the desired result. □

References

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.

Farias, V., Li, A., Peng, T., and Zheng, A. Markovian interference in experiments. *Advances in Neural Information Processing Systems*, 35:535–549, 2022.

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.

Jin, Y. and Sidford, A. Efficiently solving MDPs with stochastic mirror descent. In *International Conference on Machine Learning*, pp. 4890–4900. PMLR, 2020.

Jin, Y. and Sidford, A. Towards tight bounds on the sample complexity of average-reward mdps. In *International Conference on Machine Learning*, pp. 5055–5064. PMLR, 2021.

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

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.

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.

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.

Naik, A., Shariff, R., Yasui, N., Yao, H., and Sutton, R. S. Discounted reinforcement learning is not an optimization problem. *arXiv preprint arXiv:1910.02140*, 2019.

Ozdaglar, A., Pattathil, S., Zhang, J., and Zhang, K. Revisiting the linear-programming framework for offline RL with general function approximation. In *International Conference on Machine Learning*, 2023.

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.

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.

Sutton, R. S. and Barto, A. G. *Reinforcement learning: An introduction*. MIT press, 2018.

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.

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.

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, J., Wang, M., and Yang, L. F. Near sample-optimal reduction-based policy learning for average reward mdp. *arXiv preprint arXiv:2212.00603*, 2022.

Wang, M. Randomized linear programming solves the markov decision problem in nearly linear (sometimes sublinear) time. *Mathematics of Operations Research*, 45 (2):517–546, 2020.

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.

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.