Offline Reinforcement Learning with Realizability and Single-policy Concentrability

Wenhao Zhan WENHAO.ZHAN@PRINCETON.EDU
Princeton University

Baihe Huang BAIHEHUANG@PKU.EDU.CN
Peking University

Audrey Huang AUDREYH5@ILLINOIS.EDU
University of Illinois Urbana-Champaign

Nan Jiang NANJIANG@ILLINOIS.EDU
University of Illinois Urbana-Champaign

Jason D. Lee JASONLEE@PRINCETON.EDU
Princeton University

Editors: Po-Ling Loh and Maxim Raginsky

Abstract

Sample-efficiency guarantees for offline reinforcement learning (RL) often rely on strong assumptions on both the function classes (e.g., Bellman-completeness) and the data coverage (e.g., all-policy concentrability). Despite the recent efforts on relaxing these assumptions, existing works are only able to relax one of the two factors, leaving the strong assumption on the other factor intact. As an important open problem, can we achieve sample-efficient offline RL with weak assumptions on *both* factors?

In this paper we answer the question in the positive. We analyze a simple algorithm based on the primal-dual formulation of MDPs, where the dual variables (discounted occupancy) are modeled using a density-ratio function against offline data. With proper regularization, the algorithm enjoys polynomial sample complexity, under *only* realizability and single-policy concentrability. We also provide alternative analyses based on different assumptions to shed light on the nature of primal-dual algorithms for offline RL.

Keywords: offline RL, primal-dual, reinforcement learning theory

1. Introduction

Offline (or batch) reinforcement learning (RL) learns decision-making strategies using solely historical data, and is a promising framework for applying RL to many real-world applications. Unfortunately, offline RL training is known to be difficult and unstable (Fujimoto et al., 2019; Wang et al., 2020, 2021a), primarily due to two fundamental challenges. The first challenge is distribution shift, that the state distributions induced by the candidate policies may deviate from the offline data distribution, creating difficulties in accurately assessing the performance of the candidate policies. The second challenge is the sensitivity to function approximation, that errors can amplify exponentially over the horizon even with good representations (Du et al., 2020; Weisz et al., 2020; Wang et al., 2021b).

These challenges not only manifest themselves as degenerate behaviors of practical algorithms, but are also reflected in the strong assumptions needed for providing sample-efficiency guarantees

to classical algorithms. (In this paper, by sample-efficiency we mean a sample complexity that is polynomial in the relevant parameters, including the horizon, the capacities of the function classes, and the degree of data coverage.) As an example, the guarantees of the popular Fitted-Q Iteration (Ernst et al., 2005; Munos and Szepesvári, 2008; Chen and Jiang, 2019; Fan et al., 2020) require the following two assumptions:

- **(Data) All-policy concentrability:** The offline data distribution provides good coverage (in a technical sense) over the state distributions induced by *all* candidate policies.
- **(Function Approximation) Bellman-completeness:** The value-function class is *closed* under the Bellman optimality operator.[1]

Both assumptions are very strong and may fail in practice, and algorithms whose guarantees rely on them naturally suffer from performance degradation and instability (Fujimoto et al., 2019; Wang et al., 2020, 2021a). On one hand, *all-policy concentrability* not only requires a highly exploratory dataset (despite that historical data in real applications often lacks exploration), but also implicitly imposes structural assumptions on the MDP dynamics (Chen and Jiang, 2019, Theorem 4). On the other hand, *Bellman-completeness* is much stronger than realizability (that the optimal value function is simply contained in the function class), and is *non-monotone* in the function class, that the assumption can be violated more severely when a richer function class is used.

To address these challenges, a significant amount of recent efforts in offline RL have been devoted to relaxing these strong assumptions via novel algorithms and analyses. Unfortunately, these efforts are only able to address either the data or the function-approximation assumption, and no existing works address both simultaneously. For example, Liu et al. (2020); Rajaraman et al. (2020); Jin et al. (2020); Rashidinejad et al. (2021); Xie et al. (2021); Uehara and Sun (2021) show that pessimism is an effective mechanism for mitigating the negative consequences due to lack of data coverage, and provide guarantees under *single-policy concentrability*, that the data only covers a single good policy (e.g., the optimal policy). However, they require completeness-type assumptions on the value-function classes or *model* realizability.[2] Xie and Jiang (2021b) only require realizability of the optimal value-function, but their data assumption is even stronger than all-policy concentrability. To this end, we want to ask:

Is sample-efficiency possible with realizability and single-policy concentrability?

In this work, we answer the question in the positive by proposing the first model-free algorithm, `PRO-RL`, that only requires relatively weak assumptions on both data coverage and function approximation. The algorithm is based on the primal-dual formulation of linear programming (LP) for MDPs (Puterman, 2014; Wang, 2017), where we use marginalized importance weight (or density ratio) to model the dual variables which correspond to the discounted occupancy of the learned policy, a practice commonly found in the literature of off-policy evaluation (OPE) (e.g., Liu et al., 2018). Our main result (Corollary 5) provides polynomial sample-complexity guarantees when the density ratio and (a regularized notion of) the value function of the regularized optimal policy are

1. Approximate policy iteration algorithms usually require a variant of this assumption, that is, the closure under the policy-specific Bellman operator for *every* candidate policy (Munos., 2003; Antos et al., 2008a).
2. When a model class that contains the true MDP model is given, value-function classes that satisfy a version of Bellman-completeness can be automatically induced from the model class (Chen and Jiang, 2019), so model realizability is even stronger than Bellman-completeness. Therefore, in this work we aim at only making a constant number of realizability assumptions of real-valued functions.

realizable, and the data distribution covers such an optimal policy. We also provide a number of extensions and alternative analyses to complement the main result and provide deeper understanding of the behavior of primal-dual algorithms in the offline learning setting (see also Table 1 for a comparison with existing algorithms):

1. Section 4.2 handles the scenario where the optimal policy is not covered and we need to compete with the best policy supported on data.
2. Appendix A extends the main result to account for approximation and optimization errors, and Appendix B handles the case where the behavior policy is unknown (which the main algorithm needs) and estimated by behavior cloning.
3. Our main result crucially relies on the use of regularization. In Appendix C we study the unregularized algorithm, and provide guarantees under alternative assumptions.

Table 1: Assumptions required by existing algorithms and our algorithms to learn an ϵ-optimal policy efficiently. See basic notation in Section 2. $d^*_\alpha = d^{\pi^*_\alpha}$ where π^*_α is the α-regularized optimial policy (defined in Section 3). d^D is the offline data distribution. v^*_α ($v^*_{\alpha'_\epsilon, B_w}$) is the α-regularized optimal value function (w.r.t. the covered policy class), defined in Section 3 (Section 4.2). $w^*_\alpha(w^*_{\alpha'_\epsilon, B_w})$ is the optimal density ratio $\frac{d^*_\alpha}{d^D}$ (w.r.t. the covered policy class), as stated in Section 3 (Section 4.2). We compete with the unregularized optimal policy π^*_0 by default, with the exception of the first result of `PRO-RL` where we compete with π^*_α.

Algorithm	Data	Function Class
AVI	$\|\frac{d^\pi}{d^D}\|_\infty \le B_w, \forall \pi$	$\mathcal{T} f \in \mathcal{F}, \forall f \in \mathcal{F}$ (Munos and Szepesvári, 2008)
API		$\mathcal{T}^\pi f \in \mathcal{F}, \forall f \in \mathcal{F}, \pi \in \Pi$ (Antos et al., 2008b)
BVFT	Stronger than above	$Q^* \in \mathcal{F}$ (Xie and Jiang, 2021a)
Pessimism	$\|\frac{d^*_0}{d^D}\|_\infty \le B_w$	$\mathcal{T}^\pi f \in \mathcal{F}, \forall f \in \mathcal{F}, \pi \in \Pi$ (Xie et al., 2021)
		$w^*_0 \in \mathcal{W}, Q^\pi \in \mathcal{F}, \forall \pi \in \Pi$ (Jiang and Huang, 2020)
PRO-RL (against π^*_α)	$\|\frac{d^*_\alpha}{d^D}\|_\infty \le B_w$	$w^*_\alpha \in \mathcal{W}, v^*_\alpha \in \mathcal{V}$ (Theorem 3)
PRO-RL	$\|\frac{d^*_0}{d^D}\|_\infty \le B_w$	$w^*_{\alpha'_\epsilon, B_w} \in \mathcal{W}, v^*_{\alpha'_\epsilon, B_w} \in \mathcal{V}$ (Corollary 12)

1.1. Related works

Section 1 has reviewed the analyses of approximate value/policy iteration, and we focus on other related works in this section.

Lower bounds When we only assume the realizability of the optimal value-function, a number of recent works have established information-theoretic hardness for offline learning under relatively weak data coverage assumptions (Wang et al., 2020; Amortila et al., 2020; Zanette, 2021; Chen et al., 2021). A very recent result by Foster et al. (2021) shows a stronger barrier, that even with *all-policy concentrability* and the realizability of the value functions of *all policies*, it is still impossible to obtain polynomial sample complexity in the offline learning setting. These works do

not contradict our results, as we also assume the realizability of the density-ratio function, which circumvents the existing lower bound constructions. In particular, as Foster et al. (2021, Section 1.3) have commented, their lower bound no longer holds if the realizability of importance weight is assumed, as a realizable weight class would have too large of a capacity in their construction and would explain away the sample-complexity lower bound that scales with the size of the state space.

Marginalized importance sampling (MIS) As mentioned above, a key insight that enables us to break the lower bounds against value-function realizability is the use of marginalized importance weights (or density ratio). Modeling such functions is a common practice in MIS, a recently popular approach in the OPE literature (Liu et al., 2018; Uehara et al., 2020; Kostrikov et al., 2019; Nachum and Dai, 2020; Zhang et al., 2020), though most of the works focus exclusively on policy evaluation.

Among the few works that consider policy optimization, AlgaeDICE (Nachum et al., 2019b) optimizes the policy using MIS as a subroutine for policy evaluation, and Jiang and Huang (2020) analyze AlgaeDICE under the realizability of *all* candidate policies' value functions (see row 5 in Table 1). Similarly, MABO (Xie and Jiang, 2020) only needs realizability of the optimal value function, but the weight class needs to realize the density ratio of *all* candidate policies. The key difference in our work is the use of the LP formulation of MDPs (Puterman, 2014) to directly solve for the optimal policy, without trying to evaluate other policies. This idea has been recently explored by OptiDICE (Lee et al., 2021), which is closely related to and has inspired our work. However, Lee et al. (2021) focuses on developing an empirical algorithm, and as we will further discuss in Section 5.2, multiple design choices in our algorithms deviate from those of OptiDICE and are crucial to obtaining the desired sample-complexity guarantees.

2. Preliminaries

Markov decision process (MDP). We consider an infinite-horizon discounted MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma, \mu_0)$ (Bertsekas, 2017), where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $\gamma \in [0, 1)$ is the discount factor, $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition function, $\mu_0 \in \Delta(\mathcal{S})$ is the initial state distribution, and $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the reward function. Here, we assume $\mathcal{S}$ and $\mathcal{A}$ to be finite, but our results will not depend on their cardinalities and can be extended to the infinite case naturally. We also assume $\mu_0(s) > 0$ for all $s \in \mathcal{S}$; since our analysis and results will not depend on $\min_{s\in\mathcal{S}} \mu_0(s)$, $\mu_0(s)$ for any particular s can be arbitrarily small and therefore this is a trivial assumption for certain technical conveniences.

A policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ specifies the action selection probability in state s, and the associated discounted state-action occupancy is defined as $d^\pi(s, a) := (1 - \gamma) \sum_{t=0}^\infty \gamma^t \mathrm{Pr}_\pi(s_t = s, a_t = a)$, where the subscript of π in $\mathrm{Pr}_{(\cdot)}$ or $\mathbb{E}_{(\cdot)}$ refers to the distribution of trajectories generated as $s_0 \sim \mu_0$, $a_t \sim \pi(\cdot|s_t)$, $s_{t+1} \sim P(\cdot|s_t, a_t)$ for all $t \geq 0$. For brevity, let $d^\pi(s)$ denote the discounted state occupancy $\sum_{a\in\mathcal{A}} d^\pi(s, a)$. A policy π is also associated with a value function $V^\pi : \mathcal{S} \to \mathbb{R}$ and an action-value (or Q) function $Q^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ as follows: $\forall s \in \mathcal{S}, a \in \mathcal{A}$, $V^\pi(s) := \mathbb{E}_\pi \left[\sum_{t=0}^\infty \gamma^t r(s_t, a_t) \,\middle|\, s_0 = s \right]$, $Q^\pi(s, a) := \mathbb{E}_\pi \left[\sum_{t=0}^\infty \gamma^t r(s_t, a_t) \,\middle|\, s_0 = s, a_0 = a \right]$.

The goal of RL is to find a policy that maximizes the expected discounted return:

$$\max_\pi J(\pi) = (1 - \gamma)\mathbb{E}_\pi \left[\sum_{t=0}^\infty \gamma^t r(s_t, a_t) \right] = \mathop{\mathbb{E}}_{(s,a)\sim d^\pi} [r(s, a)]. \tag{1}$$

Alternatively, $J(\pi) = (1-\gamma)V^\pi(\mu_0) := (1-\gamma)\mathbb{E}_{s\sim\mu_0}[V^\pi(s)]$. Let π^* denote the optimal policy of this unregularized problem (1).

Offline RL. In offline RL, the agent cannot interact with the environment directly and only has access to a pre-collected dataset $\mathcal{D} = \{(s_i, a_i, r_i, s_i')\}_{i=1}^n$. We further assume each (s_i, a_i, r_i, s_i') is i.i.d. sampled from $(s_i, a_i) \sim d^D, r_i = r(s_i, a_i), s_i' \sim P(\cdot|s_i, a_i)$ as a standard simplification in theory (Nachum et al., 2019a,b; Xie et al., 2021; Xie and Jiang, 2021a). Besides, we denote the conditional probability $d^D(a|s)$ by $\pi_D(a|s)$ and call π_D the behavior policy. However, we do not assume $d^D = d^{\pi_D}$ in most of our results for generality (except for Section 4.2). We also use $d^D(s)$ to represent the marginal distribution of state, i.e., $d^D(s) = \sum_{a\in\mathcal{A}} d^D(s,a)$. In addition, we assume access to a batch of i.i.d. samples $\mathcal{D}_0 = \{s_{0,j}\}_{j=1}^{n_0}$ from the initial distribution μ_0.

3. Algorithm: PRO-RL

Our algorithm builds on a regularized version of the well-celebrated LP formulation of MDPs (Puterman, 2014). In particular, consider the following problem:

Problem: Regularized LP

$$\max_{d\geq 0} \mathbb{E}_{(s,a)\sim d}[r(s,a)] - \alpha \mathbb{E}_{(s,a)\sim d^D}\left[f\left(\frac{d(s,a)}{d^D(s,a)}\right)\right] \tag{2}$$

$$\text{s.t. } d(s) = (1-\gamma)\mu_0(s) + \gamma \sum_{s',a'} P(s|s',a')d(s',a'), \forall s \in \mathcal{S} \tag{3}$$

where $d \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|}$, $d(s) = \sum_a d(s,a)$, and $f : \mathbb{R} \to \mathbb{R}$ is a strongly convex and continuously differentiable function serving as a regularizer.

Without the regularization term, this problem is exactly equivalent to the unregularized problem (1), as (3) exactly characterizes the space of possible discounted occupancies d^π that can be induced in this MDP and is often known as the Bellman flow equations. Any non-negative d that satisfies such constraints corresponds to d^π for some stationary policy π. Therefore, once we have obtained the optimum d_α^* of the above problem, we can extract the regularized optimal policy π_α^* via

$$\pi_\alpha^*(a|s) := \begin{cases} \frac{d_\alpha^*(s,a)}{\sum_a d_\alpha^*(s,a)}, & \text{for } \sum_a d_\alpha^*(s,a) > 0, \\ \frac{1}{|\mathcal{A}|}, & \text{else.} \end{cases} \quad \forall s \in \mathcal{S}, a \in \mathcal{A}. \tag{4}$$

Turning to the regularizer, $D_f(d\|d^D) := \mathbb{E}_{(s,a)\sim d^D}\left[f\left(\frac{d(s,a)}{d^D(s,a)}\right)\right]$ is the f-divergence between d^π and d^D. This practice, often known as behavioral regularization, encourages the learned policy π to induce an occupancy $d = d^\pi$ that stays within the data distribution d^D, and we will motivate it further using a counterexample against the unregularized algorithm & analysis later.

To convert the regularized problem (2)(3) into a learning algorithm compatible with function approximation, we first introduce the Lagrangian multiplier $v \in \mathbb{R}^{|\mathcal{S}|}$ to (2)(3), and obtain the

following maximin problem:

$$\max_{d\geq 0}\min_{v}\mathbb{E}_{(s,a)\sim d}[r(s,a)]-\alpha\mathbb{E}_{(s,a)\sim d^D}\left[f\left(\frac{d(s,a)}{d^D(s,a)}\right)\right] + \sum_{s\in\mathcal{S}}v(s)\left((1-\gamma)\mu_0(s)+\gamma\sum_{s',a'}P(s|s',a')d(s',a')-d(s)\right). \tag{5}$$

Then, by variable substitution $w(s,a)=\frac{d(s,a)}{d^D(s,a)}$ and replacing summations with the corresponding expectations, we obtain the following problem

$$\max_{w\geq 0}\min_{v}L_\alpha(v,w):=(1-\gamma)\mathbb{E}_{s\sim\mu_0}[v(s)]-\alpha\mathbb{E}_{(s,a)\sim d^D}[f(w(s,a))]+\mathbb{E}_{(s,a)\sim d^D}[w(s,a)e_v(s,a)], \tag{6}$$

where $e_v(s,a)=r(s,a)+\gamma\sum_{s'}P(s'|s,a)v(s')-v(s)$. The optimum of (6), denoted by (v^*_α,w^*_α), will be of vital importance later, as our main result relies on the realizability of these two functions v^*_α and w^*_α. In particular, v^*_α may not be characterized by the familiar Bellman equations when $\alpha>0$, and Eq.(6) is our only handle on this quantity. When $\alpha\to 0$, v^*_0 is the familiar optimal state-value function V^{π^*}, and $d^*_0:=w^*_0\cdot d^D$ is the discounted occupancy of an optimal policy. Note that optimal policies in MDPs are generally not unique and thus w^*_0,d^*_0 are not unique either. We denote the optimal set of w^*_0 and d^*_0 by $\mathcal{W}^*_0$ and D^*_0, respectively.

Finally, our algorithm simply uses $\mathcal{V}\subseteq\mathbb{R}^{|\mathcal{S}|}$ and $\mathcal{W}\subseteq\mathbb{R}_+^{|\mathcal{S}|\times|\mathcal{A}|}$ to approximate v and w, respectively, and optimizes the empirical version of $L_\alpha(v,w)$ over $\mathcal{W}\times\mathcal{V}$. Concretely, we solve for

$$\textbf{PRO-RL:}\qquad (\widehat{w},\widehat{v})=\arg\max_{w\in\mathcal{W}}\arg\min_{v\in\mathcal{V}}\widehat{L}_\alpha(v,w), \tag{7}$$

where $\widehat{L}_\alpha(v,w):=$

$$(1-\gamma)\frac{1}{n_0}\sum_{j=1}^{n_0}[v(s_{0,j})]+\frac{1}{n}\sum_{i=1}^{n}[-\alpha f(w(s_i,a_i))]+\frac{1}{n}\sum_{i=1}^{n}[w(s_i,a_i)e_v(s_i,a_i,r_i,s'_i)], \tag{8}$$

and $e_v(s,a,r,s')=r+\gamma v(s')-v(s)$. The final policy we obtain is

$$\widehat{\pi}(a|s)=\begin{cases}\frac{\widehat{w}(s,a)\pi_D(a|s)}{\sum_{a'}\widehat{w}(s,a')\pi_D(a'|s)}, & \text{for } \sum_{a'}\widehat{w}(s,a')\pi_D(a'|s)>0,\\ \frac{1}{|\mathcal{A}|}, & \text{else,}\end{cases} \tag{9}$$

We call this algorithm **P**rimal-dual **R**egularized **O**ffline **R**einforcement **L**earning (PRO-RL). For now we assume the behavior policy π_D is known; Appendix B extends the main results to the unknown π_D setting, where π_D is estimated by behavior cloning.

While behavioral regularization (the f term) is frequently used in MIS (especially in DICE algorithms (Nachum et al., 2019b; Lee et al., 2021)), its theoretical role has been unclear and finite-sample guarantees can often be obtained without it (Jiang and Huang, 2020). For us, however, the use of regularization is crucial in proving our main result (Corollary 5). In Section 5.1 we construct a counterexample against the unregularized algorithm under the natural "unregularized" assumptions. Regularization is an effective method to combat with this counterexample, and we also introduce alternative concentrability assumptions to make unregularized algorithm work in Appendix C.

4. Main results

In this section we present the main sample-complexity guarantees of our algorithm under only realizability assumptions for $\mathcal{V}$ and $\mathcal{W}$ and single-policy concentrability of data. We will start with the analyses that assume perfect optimization and that the behavior policy π_D is known (Section 4.1), allowing us to present the result in a clean manner. Section 4.2 then removes the concentrability assumption altogether and allows us to compete with the best covered policy. Further, we handle approximation and optimization errors in Appendix A, and use behavior cloning to handle an unknown behavior policy in Appendix B.

4.1. Sample-efficiency with only realizability and weak concentrability

We introduce the needed assumptions before stating the sample-efficiency guarantees to our algorithm. The first assumption is about data coverage, that it covers the occupancy induced by a (possibly regularized) optimal policy.

Assumption 1 (π^*_α-concentrability) $\dfrac{d^*_\alpha(s,a)}{d^D(s,a)} \leq B^\alpha_w, \forall s \in \mathcal{S}, a \in \mathcal{A}.$

Two remarks are in order:

1. Assumption 1 is parameterized by α, and we will bind it to specific values when we state the guarantees.
2. This assumption is necessary if we want to compete with the optimal policy of the MDP, π^*, and is already much weaker than all-policy concentrability (Munos and Szepesvári, 2008; Farahmand et al., 2010; Chen and Jiang, 2019). That said, ideally we should not even need such an assumption, as long as we are willing to compete with the best policy covered by data instead of the truly optimal policy (Liu et al., 2020; Xie et al., 2021). We will actually show how to achieve this in Section 4.2.

We then introduce the realizability assumptions on our function approximators $\mathcal{V}$ and $\mathcal{W}$, which are very straightforward. For now we assume exact realizability, and Appendix A handles misspecification errors.

Assumption 2 (Realizability of $\mathcal{V}$ and $\mathcal{W}$) *Suppose* $v^*_\alpha \in \mathcal{V}, w^*_\alpha \in \mathcal{W}$.

The above 2 assumptions are the major assumptions we need. (The rest are standard technical assumptions on boundedness.) Comparing them to existing results, we emphasize that all existing analyses require "$\forall$" quantifiers in the assumptions either about the data (e.g., all-policy concentrability) or about the function classes (e.g., Bellman-completeness). See Table 1 for a comparison to various approaches considered in the literature.

Having stated the major assumptions, we now turn to the routine ones on function boundedness.

Assumption 3 (Boundedness of $\mathcal{W}$) *Suppose* $0 \leq w(s,a) \leq B_{w,\alpha}$ *for any* $s \in \mathcal{S}, a \in \mathcal{A}, w \in \mathcal{W}$.

Here we reuse $B_{w,\alpha}$ from Assumption 1. Since $d^*_\alpha/d^D = w^*_\alpha \in \mathcal{W}$ by Assumption 2, in general the magnitude of $\mathcal{W}$ should be larger than that of d^*_α/d^D, and we use the same upper bound to eliminate unnecessary notations and improve readability.

The next assumption characterizes the regularizer f. These are not really assumptions as we can make concrete choices of f that satisfy them (e.g., a simple quadratic function; see Remark 4), but for now we leave them as assumptions to keep the analysis general.

Assumption 4 (Properties of f) *Suppose f satisfies the following properties:*

- ***Strong Convexity**: f is M_f-strongly-convex.*
- ***Boundedness**:*

$$|f'(x)| \le B_{f',\alpha}, \forall \quad 0 \le x \le B_{w,\alpha}, \tag{10}$$
$$|f(x)| \le B_{f,\alpha}, \forall \quad 0 \le x \le B_{w,\alpha}. \tag{11}$$

- ***Non-negativity**: $f(x) \ge 0$ for any $x \in \mathbb{R}$.*

Remark 1 *The non-negativity is trivial since f is strongly convex and we can always add a constant term to ensure non-negativity holds. Besides, we can get rid of non-negativity with the results in Section 4.2.*

Assumption 4 allows us to bound $\|v_\alpha^*\|_\infty \le \frac{\alpha B_{f',\alpha}+1}{1-\gamma}$ (see Lemma 32 in Section E); in the same spirit as Assumption 3, we assume:

Assumption 5 (Boundedness of $\mathcal{V}$) *Suppose $\|v\|_\infty \le B_{v,\alpha} := \frac{\alpha B_{f',\alpha}+1}{1-\gamma}$ for any $v \in \mathcal{V}$.*

With the above assumptions, we have Theorem 3 to show that PRO-RL can learn the optimal density ratio and policy for the regularized problem (2)(3) with polynomial samples. To simplify writing, we introduce the following notation for the statistical error term that arises purely from concentration inequalities:

Definition 2

$$\mathcal{E}_{n,n_0,\alpha}(B_w, B_f, B_v, B_e) = (1-\gamma)B_v \cdot \left(\frac{2\log\frac{4|\mathcal{V}|}{\delta}}{n_0}\right)^{\frac{1}{2}} + (\alpha B_f + B_w B_e) \cdot \left(\frac{2\log\frac{4|\mathcal{V}||\mathcal{W}|}{\delta}}{n}\right)^{\frac{1}{2}}. \tag{12}$$

$\mathcal{E}$ characterizes the statistical error $\widehat{L}_\alpha(v,w) - L_\alpha(v,w)$ based on concentration inequalities, and the two terms in its definition correspond to using $\mathcal{D}_0$ for $(1-\gamma)\mathbb{E}_{s\sim\mu_0}[v(s)]$ and $\mathcal{D}$ for $-\alpha\mathbb{E}_{(s,a)\sim d^D}[f(w(s,a))] + \mathbb{E}_{(s,a)\sim d^D}[w(s,a)e_v(s,a)]$, respectively. Using this shorthand, we state our first guarantee, that the extracted policy $\widehat{\pi}$ will be close to π_α^*, the solution to the regularized problem (2)(3).

Theorem 3 (Sample complexity of learning π_α^*) *Fix $\alpha > 0$. Suppose Assumptions 1,2,3,4,5 hold for the said α. Then with at least probability $1-\delta$, the output of PRO-RL satisfies:*

$$J(\pi_\alpha^*) - J(\widehat{\pi}) \le \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\alpha^*}[\|\pi_\alpha^*(\cdot|s) - \widehat{\pi}(\cdot|s)\|_1] \le \frac{4}{1-\gamma}\sqrt{\frac{\mathcal{E}_{n,n_0,\alpha}(B_{w,\alpha}, B_{f,\alpha}, B_{v,\alpha}, B_{e,\alpha})}{\alpha M_f}}, \tag{13}$$

where $B_{e,\alpha} := (1+\gamma)B_{v,\alpha} + 1$.

Proof [Proof sketch] The proof is inspired from the invariance of saddle points (Appendix E.1) and mainly consists of three steps: (1) using concentration inequalities to bound $|L_\alpha(v,w) - \widehat{L}_\alpha(v,w)|$, (2) using the invariance of saddle points and concentration bounds to characterize the error $\|\widehat{w} - w^*_\alpha\|_{2,d^D}$ and (3) analyzing the difference between $\widehat{\pi}$ and π^*_α. See Appendix E for details. ■

Remark 4 (Sample complexity for quadratic regularization) *Theorem 3 shows that* PRO-RL *can obtain a near-optimal policy for regularized problem (2)(3) with sample complexity* $O(n_0 + n_1) = \widetilde{O}\left(\frac{(\alpha B_{f,\alpha} + B_{w,\alpha} B_{e,\alpha})^2}{(1-\gamma)^4 (\alpha M_f)^2 \epsilon^4}\right)$. *However, there might be implicit dependence on* $1-\gamma, \alpha M_f, B_{w,\alpha}$ *in the constants* $B_{e,\alpha}$. *To reveal these terms, we consider a simple choice of* $f(x) = \frac{M_f}{2}x^2$. *Then we have* $B_{e,\alpha} = O(\frac{\alpha M_f (B_{w,\alpha})^2 + B_{w,\alpha}}{1-\gamma}), B_{f,\alpha} = O(\alpha M_f (B_{w,\alpha})^2)$, *leading to a sample complexity* $\widetilde{O}\Big(\frac{(B_{w,\alpha})^2}{(1-\gamma)^6 (\alpha M_f)^2 \epsilon^4} + \frac{(B_{w,\alpha})^4}{(1-\gamma)^6 \epsilon^4}\Big)$.

Moreover, PRO-RL can even learn a near-optimal policy for the unregularized problem (1) efficiently by controlling the magnitude of α in PRO-RL. Corollary 5 characterizes the sample complexity of PRO-RL for the unregularized problem (1) without any approximation/optimization error, whose proof is deferred to Appendix G.

Corollary 5 (Sample complexity of competing with π^*_0) *Fix any* $\epsilon > 0$. *Suppose there exists* $d^*_0 \in D^*_0$ *that satisfies Assumption 1 with* $\alpha = 0$. *Besides, assume that Assumptions 1,2,3,4,5 hold for* $\alpha = \alpha_\epsilon := \frac{\epsilon}{2B_{f,0}}$. *Then if*

$$n \geq \frac{C_1 \left(\epsilon B_{f,\alpha_\epsilon} + 2B_{w,\alpha_\epsilon} B_{e,\alpha_\epsilon} B_{f,0}\right)^2}{\epsilon^6 M_f^2 (1-\gamma)^4} \cdot \log \frac{4|\mathcal{V}||\mathcal{W}|}{\delta}, n_0 \geq \frac{C_1 \left(2B_{v,\alpha_\epsilon} B_{f,0}\right)^2}{\epsilon^6 M_f^2 (1-\gamma)^2} \cdot \log \frac{4|\mathcal{V}|}{\delta}, \quad (14)$$

the output of PRO-RL *with* $\alpha = \alpha_\epsilon$ *satisfies* $J(\pi^*_0) - J(\widehat{\pi}) \leq \epsilon$ *with probability at least* $1-\delta$, *where* C_1 *is some universal positive constants and* π^*_0 *is the optimal policy inducing* d^*_0.

Remark 6 (Quadratic regularization) *Similarly as Remark 4, the sample complexity of competing with* π^*_0 *under quadratic* f *is* $\widetilde{O}\left(\frac{(B_{w,0})^4 (B_{w,\alpha_\epsilon})^2}{\epsilon^6 (1-\gamma)^6}\right)$.

PRO-RL is originally designed for the regularized problem. Therefore, when applying it to the unregularized problem the sample complexity degrades from $\widetilde{O}\left(\frac{1}{\epsilon^4}\right)$ to $\widetilde{O}\left(\frac{1}{\epsilon^6}\right)$. However, the sample complexity remains polynomial in all relevant quantities. Compared to Theorem 3, Corollary 5 requires concentrability for policy π^*_0 in addition to $\pi^*_{\alpha_\epsilon}$, so technically we require "two-policy" instead of single-policy concentrability for now. While this is still much weaker than all-policy concentrability (Chen and Jiang, 2019), we show in Section 4.2 how to compete with π^*_0 with only single-policy concentrability.

Remark 7 *When* ϵ *shrinks, the realizability assumptions for Corollary 5 also need to hold for regularized solutions with smaller* α. *That said, in the following discussion (Proposition 8), we will show that when* ϵ *is subsequently small, the realizability assumptions will turn to be with respect to the unregularized solutions.*

Comparison with existing algorithms. Theorem 3 and Corollary 5 display an exciting result that PRO-RL obtains a near optimal policy for regularized problem (2)(3) and unregularized problem (1) using polynomial samples with only realizability and weak data-coverage assumptions. The literature has demonstrated hardness of learning offline RL problems and existing algorithms either rely on the completeness assumptions (Xie and Jiang, 2020; Xie et al., 2021; Du et al., 2021) or extremely strong data assumption (Xie and Jiang, 2021a). Our results show for the first time that offline RL problems can be solved using a polynomial number of samples without these assumptions.

High accuracy regime ($\epsilon \to 0$). Corollary 5 requires realizability with respect to the optimizers of the regularized problem (2)(3). A natural idea is to consider whether the concentration and realizability instead can be with respect to the optimizer of the unregularized problem (v_0^*, w_0^*). Inspired by the stability of linear programming (Mangasarian and Meyer, 1979), we identify the high accuracy regime ($\epsilon \to 0$) where concentrability and realizability with respect to w_0^* can guarantee PRO-RL to output an ϵ-optimal policy as shown in the following proposition:

Proposition 8 *There exists $\overline{\alpha} > 0$ and $w^* \in \mathcal{W}_0^*$ such that when $\alpha \in [0, \overline{\alpha}]$ we have*

$$w_\alpha^* = w^*, \|v_\alpha^* - v_0^*\|_{2,d^D} \leq C\alpha, \tag{15}$$

where $C = \frac{B_{f',0} + \frac{2}{\overline{\alpha}}}{1-\gamma}$.

The proof is deferred to Appendix H. Here $\overline{\alpha}$ is a value only depends on the underlying MDP and not on ϵ. Proposition 8 essentially indicates that when $\epsilon \to 0$, $w_{\alpha_\epsilon}^*$ is exactly the unregularized optimum w^*, and $v_{\alpha_\epsilon}^*$ is $O(\epsilon)$ away form v_0^*. Combining with Corollary 5, we know that ϵ-optimal policy can be learned by PRO-RL if concentrability holds for π_0^* and $\mathcal{W}$ contains w^*.

4.2. Handling an arbitrary data distribution

In the previous section, our goal is to compete with policy π_α^* and we require the data to provide sufficient coverage over such a policy. Despite being weaker than all-policy concentrability, this assumption can be still violated in practice, since we have no control over the distribution of the offline data. In fact, recent works such as Xie et al. (2021) are able to compete with the best policy covered by data (under strong function-approximation assumptions such as Bellman-completeness), thus provide guarantees to *arbitrary* data distributions: when the data does not cover any good policies, the guarantee is vacuous; however, as long as a good policy is covered, the guarantee will be competitive to such a policy.

In this section we show that we can achieve similar guarantees for PRO-RL with a twisted analysis. First let us define the notion of covered policies.

Definition 9 *Let Π_{B_w} denote the B_w-covered policy class of d^D for $B_w > 1$, defined as:*

$$\Pi_{B_w} := \{\pi : \frac{d^\pi(s,a)}{d^D(s,a)} \leq B_w, \forall s \in \mathcal{S}, a \in \mathcal{A}\}. \tag{16}$$

Here, B_w is a hyperparameter chosen by the practitioner, and our goal in this section is to compete with policies in Π_{B_w}. The key idea is to extend the regularized LP (2) by introducing an additional upper-bound constraint on d, that $d(s,a) \leq B_w\, d^D(s,a)$, so that we only search for a good policy within Π_{B_w}.

Problem: Constrained & regularized LP

$$\max_{0\le d\le B_w d^D} \mathbb{E}_{(s,a)\sim d}[r(s,a)] - \alpha \mathbb{E}_{(s,a)\sim d^D}\left[f\left(\frac{d(s,a)}{d^D(s,a)}\right)\right] \tag{17}$$

$$\text{s.t. } d(s) = (1-\gamma)\mu_0(s) + \gamma \sum_{s',a'} P(s|s',a')d(s',a') \tag{18}$$

The policy we will compete with, π^*_{α,B_w}, and the corresponding value and density-ratio functions, v^*_{α,B_w}, w^*_{α,B_w}, will all be defined based on this constrained LP. In the rest of this section, we show that if we make similar realizability assumptions as in Section 4.1 but w.r.t. v^*_{α,B_w} and w^*_{α,B_w} (instead of v^*_α and w^*_α), then we can compete with π^*_{α,B_w} without needing to make any coverage assumption on the data distribution d^D. Following a similar argument as the derivation of PRO-RL, we can show that Problem (17) is equivalent to the maximin problem:

$$\max_{0\le w\le B_w} \min_v L_\alpha(v,w) := (1-\gamma)\mathbb{E}_{s\sim\mu_0}[v(s)] - \alpha\mathbb{E}_{(s,a)\sim d^D}[f(w(s,a))] + \mathbb{E}_{(s,a)\sim d^D}[w(s,a)e_v(s,a)], \tag{19}$$

Denote the optimum of (19) by $(v^*_{\alpha,B_w}, w^*_{\alpha,B_w})$, then the optimal policy and its associated discounted state occupancy can be recovered as follows: $\forall s\in\mathcal{S}, a\in\mathcal{A}$,

$$\pi^*_{\alpha,B_w}(s|a) := \begin{cases} \frac{w^*_{\alpha,B_w}(s,a)\pi_D(a|s)}{\sum_a w^*_{\alpha,B_w}(s,a)\pi_D(a|s)}, & \sum_a w^*_{\alpha,B_w}(s,a)\pi_D(a|s) > 0, \\ \frac{1}{|\mathcal{A}|}, & \text{otherwise.} \end{cases}, \tag{20}$$

$$d^*_{\alpha,B_w}(s,a) = w^*_{\alpha,B_w}(s,a)d^D(s,a). \tag{21}$$

We now state the realizability and boundedness assumptions, which are similar to Section 4.1.

Assumption 6 (Realizability of $\mathcal{V}$ and $\mathcal{W}$ II) *Suppose $v^*_{\alpha,B_w}\in\mathcal{V}, w^*_{\alpha,B_w}\in\mathcal{W}$.*

Assumption 7 (Boundedness of $\mathcal{W}$ II) *Suppose $0\le w(s,a)\le B_w$ for any $s\in\mathcal{S}, a\in\mathcal{A}, w\in\mathcal{W}$.*

Assumption 8 (Boundedness of f II) *Suppose that*

$$|f'(x)| \le B_{f'}, \forall \quad 0\le x\le B_w, \tag{22}$$

$$|f(x)| \le B_f, \forall \quad 0\le x\le B_w. \tag{23}$$

Next we consider the boundedness of $\mathcal{V}$. Similar to Assumption 5, we will decide the appropriate bound on functions in $\mathcal{V}$ based on that of v^*_{α,B_w}, which needs to be captured by $\mathcal{V}$. It turns out that the additional constraint $w\le B_w$ makes it difficult to derive an upper bound on v^*_{α,B_w}. However, we are able to do so under a common and mild assumption, that the data distribution d^D is a valid occupancy (Liu et al., 2018; Tang et al., 2019; Levine et al., 2020):

Assumption 9 *Suppose $d^D = d^{\pi_D}$, i.e., the discounted occupancy of behavior policy π_D.*

With Assumption 9, we have $\|v^*_{\alpha,B_w}\|_\infty \le B_v$ from Lemma 11 and therefore the following assumption is reasonable:

Assumption 10 (Boundedness of $\mathcal{V}$ II) *Suppose $\|v\|_\infty \le B_v := \frac{\alpha B_{f'}+1}{1-\gamma}$ for any $v \in \mathcal{V}$.*

With the above assumptions, we have the following theorem to show that PRO-RL is able to learn π^*_{α,B_w}:

Theorem 10 *Assume $\alpha > 0$. Suppose 6,7,8,9,10 and strong convexity in 4 hold. Then with at least probability $1-\delta$, the output of PRO-RL satisfies:*

$$J(\pi^*_{\alpha,B_w}) - J(\widehat{\pi}) \le \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^*_{\alpha,B_w}}[\|\pi^*_{\alpha,B_w}(\cdot|s) - \widehat{\pi}(\cdot|s)\|_1] \le \frac{4}{1-\gamma}\sqrt{\frac{\mathcal{E}_{n,n_0,\alpha}(B_w,B_f,B_v,B_e)}{\alpha M_f}}, \tag{24}$$

where $B_e := (1+\gamma)B_v + 1$.

Proof [Proof sketch] The proof largely follows Theorem 3 except for the derivation of the bound on v^*_{α,B_w}, which is characterized in the following lemma:

Lemma 11 *Suppose Assumption 8 holds, then we have: $\|v^*_{\alpha,B_w}\|_\infty \le B_v$.*

The proof of Lemma 11 is deferred to Appendix I.1. The rest of the proof of Theorem 10 is the same as in Appendix E and thus omitted here. ■

As before, we obtain the following corollary for competing with the best policy π^*_{0,B_w} in Π_{B_w} whose proof is deferred to Appendix J:

Corollary 12 *For any $\epsilon > 0$, assume that Assumption 6,7,8,9,10 and strong convexity in 4 hold for $\alpha = \alpha'_\epsilon := \frac{\epsilon}{4B_f}$. Then if*

$$n \ge \frac{C_1\left(\epsilon B_f + 4B_wB_eB_f\right)^2}{\epsilon^6 M_f^2(1-\gamma)^4}\cdot\log\frac{4|\mathcal{V}||\mathcal{W}|}{\delta}, n_0 \ge \frac{C_1\left(4B_vB_f\right)^2}{\epsilon^6 M_f^2(1-\gamma)^2}\cdot\log\frac{4|\mathcal{V}|}{\delta}, \tag{25}$$

the output of PRO-RL with input $\alpha = \alpha'_\epsilon$ satisfies $J(\pi^*_{0,B_w}) - J(\widehat{\pi}) \le \epsilon$ with probability at least $1-\delta$, where C_1 is the same constant in Corollary 5.

Remark 13 *Corollary 12 does not need the assumption of non-negativity of f. The reason is that we are already considering a bounded space ($0 \le w \le B_w$) and thus f must be lower bounded in this space.*

Resolving two-policy concentrability of Corollary 5 As we have commented below Corollary 5, to compete with π^*_0 we need "two-policy" concentrability, i.e., Assumption 1 for both $\alpha = 0$ and $\alpha = \alpha_\epsilon$. Here we resolve this issue in Corollary 14 below, by invoking Corollary 12 with B_w set to $B_{w,0}$. This way, we obtain the coverage over the regularized optimal policy π^*_{α,B_w} (i.e., the counterpart of π^*_α in Corollary 5) *for free*, thus only need the concentrability w.r.t. π^*_0.

Corollary 14 *Suppose there exists $d^*_0 \in D^*_0$ that satisfies Assumption 1 with $\alpha = 0$. For any $\epsilon > 0$, assume that Assumption 6,7,8,9,10 and strong convexity in 4 hold for $B_w = B_{w,0}$ and $\alpha = \alpha'_\epsilon := \frac{\epsilon}{4B_{f,0}}$. Then if*

$$n \ge \frac{C_1\left(\epsilon B_{f,0} + 4B_{w,0}B_{e,0}B_{f,0}\right)^2}{\epsilon^6 M_f^2(1-\gamma)^4}\cdot\log\frac{4|\mathcal{V}||\mathcal{W}|}{\delta}, n_0 \ge \frac{C_1\left(4B_{v,0}B_{f,0}\right)^2}{\epsilon^6 M_f^2(1-\gamma)^2}\cdot\log\frac{4|\mathcal{V}|}{\delta}, \tag{26}$$

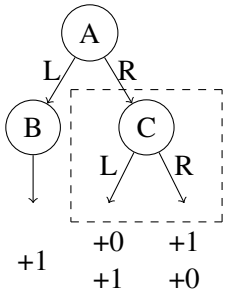

Figure 1: Construction against the unregularized algorithm under $w_0^* \in \mathcal{W}$ and $v_0^* \in \mathcal{V}$. The construction is given as a 2-stage finite-horizon MDP, and adaptation to the discounted setting is trivial. State A is the initial state with no intermediate rewards. The offline data does not cover state C. The nature can choose between 2 MDPs that differ in the rewards for state C, and only one of the two actions has a $+1$ reward.

the output of PRO-RL *with input* $\alpha = \alpha'_\epsilon$ *satisfies* $J(\pi_0^*) - J(\widehat{\pi}) \le \epsilon$ *with probability at least* $1 - \delta$*, where* C_1 *is the same constant in Corollary 5.*

Corollary 14 shows that our algorithm is able to compete with π_0^* under concentrability with respect to π_0^* alone. In addition, a version of Proposition 8 applies to Corollary 12, which indicates that $w^*_{\alpha'_\epsilon, B_w} = w_0^*$ for sufficiently small ϵ.

Remark 15 *Corollary 14 still holds when we set* $B_w \ge B_{w,0}$ *in case* $B_{w,0}$ *is unknown. However the realizability assumptions will depend on the choice of* B_w *and change accordingly.*

5. Discussion

5.1. The necessity of behavioral regularization

Figure 1 shows a counterexample where the unregularized algorithm fails even with infinite data and the natural assumptions, that (1) there exists a $w_0^* \in \mathcal{W}_0^*$ such that $w_0^* \in \mathcal{W}$, (2) $v_0^* \in \mathcal{V}$, and (3) data covers the optimal policy induced by w_0^*. In state A, both actions are equally optimal. However, since data does not cover the actions of state C, the learner should not take R in state A as it can end up choosing a highly suboptimal action in state C with constant probability if nature randomizes over the 2 possible MDP instances.

We now show that the unregularized algorithm ((7) with $\alpha = 0$) can choose R in state A, even with infinite data and "nice" d^D, $\mathcal{V}$, $\mathcal{W}$. In particular, the two possible MDPs share the same optimal value function $v_0^*(A) = v_0^*(B) = v_0^*(C) = 1$, which is the only function in $\mathcal{V}$ so we always have $v_0^* \in \mathcal{V}$. d^D covers state-action pairs (A, L), (A, R), B. $\mathcal{W}$ also contains 2 functions: w_1 is such that $w_1 \cdot d^D$ is uniform over (A, L), B, which is the occupancy of the optimal policy $\pi^*(A) = L$. w_2 is such that $w_2 \cdot d^D$ is uniform over (A, R), B, which induces a policy that chooses R in state A. However, the unregularized algorithm cannot distinguish between w_1 and w_2 even with infinite data (i.e., with objective $L_0(v, w)$). This is because w_1 and w_2 only differs in the action choice in state A, but $v_0^*(B) = v_0^*(C) = 1$ so the unregularized objective is the same for w_1 and w_2.

5.2. Comparison with OptiDICE (Lee et al., 2021)

Our algorithm is inspired by OptiDICE (Lee et al., 2021), but with several crucial modifications necessary to obtain the desired sample-complexity guarantees. OptiDICE starts with the problem of $\min_v \max_{w \geq 0} L_\alpha(v, w)$, and then uses the closed-form maximizer $w^*_\alpha(v) := \arg\max_{w \geq 0} L_\alpha(v, w)$ for arbitrary v (Lee et al., 2021, Proposition 1):

$$w^*_\alpha(v) = \max\left(0, (f')^{-1}\left(\frac{e_v(s,a)}{\alpha}\right)\right), \tag{27}$$

and then solves $\min_v L_\alpha(v, w^*_\alpha(v))$. Unfortunately, the $e_v(s,a)$ term in the expression requires knowledge of the transition function P, causing the infamous double-sampling difficulty (Baird, 1995; Farahmand and Szepesvári, 2011), a major obstacle in offline RL with only realizability assumptions (Chen and Jiang, 2019). OptiDICE deals with this by optimizing an upper bound of $\max_{w \geq 0} L_\alpha(v, w)$ which does not lend itself to theoretical analysis. Alternatively, one can fit e_v using a separate function class. However, since v is arbitrary in the optimization, the function class needs to approximate e_v for all v, requiring a completeness-type assumption in theory (Xie and Jiang, 2020). In contrast, PRO-RL optimizes over $\mathcal{V} \times \mathcal{W}$ and thus $\arg\max_{w \in \mathcal{W}} L_\alpha(v, w)$ is naturally contained in $\mathcal{W}$, and our analyses show that this circumvents the completeness-type assumptions and only requires realizability.

Another important difference is the policy extraction step. OptiDICE uses a heuristic behavior cloning algorithm without any guarantees. We develop a new behavior cloning algorithm that only requires realizability of the policy and does not increase the sample complexity.

6. Conclusion

We present the first result for offline RL under relatively weak realizability and concentrability assumptions. The algorithm, PRO-RL, is based on the regularized primal-dual formulation of LP for MDPs, and uses density-ratio functions to model the discounted occupancy on offline data. A novel high-level insight in our analysis is to define the functions we need to realize via the optimization problem on population statistics and unrestricted function classes (6), which overcomes the conceptual difficulty that the "regularized value functions" v^*_α can no longer be characterized by the familiar Bellman equations—which are central to most RL theory works—due to behavior regularization. In the appendices we extend the main results to handle approximation/optimization errors and unknown behavior policies, and provide alternative assumptions for the algorithm without regularization.

Acknowledgments

NJ acknowledges funding support from ARL Cooperative Agreement W911NF-17-2-0196, NSF IIS-2112471, NSF CAREER IIS-2141781, and Adobe Data Science Research Award. JDL and WZ acknowledges support of the ARO under MURI Award W911NF-11-1-0304, the Sloan Research Fellowship, NSF CCF 2002272, NSF IIS 2107304, ONR Young Investigator Award, and NSF Career Award.

References

Alekh Agarwal, Nan Jiang, Sham M Kakade, and Wen Sun. Reinforcement learning: Theory and algorithms. Technical report, 2019.

Alekh Agarwal, Sham Kakade, Akshay Krishnamurthy, and Wen Sun. Flambe: Structural complexity and representation learning of low rank mdps. *arXiv preprint arXiv:2006.10814*, 2020.

Philip Amortila, Nan Jiang, and Tengyang Xie. A variant of the wang-foster-kakade lower bound for the discounted setting. *arXiv preprint arXiv:2011.01075*, 2020.

András Antos, Csaba Szepesvári, and Rémi Munos. Learning near-optimal policies with bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 71(1):89–129, 2008a.

András Antos, Csaba Szepesvári, and Rémi Munos. Learning near-optimal policies with bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 71(1):89–129, 2008b.

Leemon Baird. Residual algorithms: Reinforcement learning with function approximation. In *Machine Learning Proceedings 1995*, pages 30–37. Elsevier, 1995.

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

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

Lin Chen, Bruno Scherrer, and Peter L Bartlett. Infinite-horizon offline reinforcement learning with linear function approximation: Curse of dimensionality and algorithm. *arXiv preprint arXiv:2103.09847*, 2021.

Simon S Du, Sham M Kakade, Ruosong Wang, and Lin F Yang. Is a good representation sufficient for sample efficient reinforcement learning? In *International Conference on Learning Representations*, 2020.

Simon S. Du, Sham M. Kakade, Jason D. Lee, Shachar Lovett, Gaurav Mahajan, Wen Sun, and Ruosong Wang. Bilinear classes: A structural framework for provable generalization in rl, 2021.

Damien Ernst, Pierre Geurts, and Louis Wehenkel. Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, 6:503556, 2005.

Jianqing Fan, Zhaoran Wang, Yuchen Xie, and Zhuoran Yang. A theoretical analysis of deep q-learning. In Alexandre M. Bayen, Ali Jadbabaie, George Pappas, Pablo A. Parrilo, Benjamin Recht, Claire Tomlin, and Melanie Zeilinger, editors, *Proceedings of the 2nd Conference on Learning for Dynamics and Control*, volume 120 of *Proceedings of Machine Learning Research*, pages 486–489. PMLR, 10–11 Jun 2020. URL https://proceedings.mlr.press/v120/yang20a.html.

Amir-massoud Farahmand and Csaba Szepesvári. Model selection in reinforcement learning. *Machine learning*, 85(3):299–332, 2011.

Amir-massoud Farahmand, Csaba Szepesvári, and Rémi Munos. Error Propagation for Approximate Policy and Value Iteration. In *Advances in Neural Information Processing Systems*, pages 568–576, 2010.

Dylan J Foster, Akshay Krishnamurthy, David Simchi-Levi, and Yunzong Xu. Offline reinforcement learning: Fundamental barriers for value function approximation. *arXiv preprint arXiv:2111.10919*, 2021.

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

Nan Jiang and Jiawei Huang. Minimax value interval for off-policy evaluation and policy optimization. *arXiv preprint arXiv:2002.02081*, 2020.

Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline rl? *arXiv preprint arXiv:2012.15085*, 2020.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *IN PROC. 19TH INTERNATIONAL CONFERENCE ON MACHINE LEARNING*, pages 267–274, 2002.

SM Kakade. *On the sample complexity of reinforcement learning*. PhD thesis, University of London, 2003.

Ilya Kostrikov, Ofir Nachum, and Jonathan Tompson. Imitation learning via off-policy distribution matching. *arXiv preprint arXiv:1912.05032*, 2019.

Jongmin Lee, Wonseok Jeon, Byung-Jun Lee, Joelle Pineau, and Kee-Eung Kim. Optidice: Offline policy optimization via stationary distribution correction estimation. *arXiv preprint arXiv:2106.10783*, 2021.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

Tianyi Lin, Chi Jin, and Michael I Jordan. Near-optimal algorithms for minimax optimization. In *Conference on Learning Theory*, pages 2738–2779. PMLR, 2020.

Qiang Liu, Lihong Li, Ziyang Tang, and Dengyong Zhou. Breaking the curse of horizon: Infinite-horizon off-policy estimation. *arXiv preprint arXiv:1810.12429*, 2018.

Yao Liu, Adith Swaminathan, Alekh Agarwal, and Emma Brunskill. Provably good batch reinforcement learning without great exploration. *arXiv preprint arXiv:2007.08202*, 2020.

Olvi L Mangasarian and RR Meyer. Nonlinear perturbation of linear programs. *SIAM Journal on Control and Optimization*, 17(6):745–752, 1979.

Rémi Munos. Error bounds for approximate policy iteration. In *Proceedings of the 20th International Conference on International Conference on Machine Learning*, pages 560–567. PMLR, 2003.

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

Ofir Nachum and Bo Dai. Reinforcement learning via fenchel-rockafellar duality. *arXiv preprint arXiv:2001.01866*, 2020.

Ofir Nachum, Yinlam Chow, Bo Dai, and Lihong Li. Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections. *arXiv preprint arXiv:1906.04733*, 2019a.

Ofir Nachum, Bo Dai, Ilya Kostrikov, Yinlam Chow, Lihong Li, and Dale Schuurmans. Algaedice: Policy gradient from arbitrary experience. *arXiv preprint arXiv:1912.02074*, 2019b.

Arkadi Nemirovski. Prox-method with rate of convergence o (1/t) for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004.

Yurii Nesterov. Dual extrapolation and its applications to solving variational inequalities and related problems. *Mathematical Programming*, 109(2):319–344, 2007.

Dean A Pomerleau. Alvinn: An autonomous land vehicle in a neural network. Technical report, CARNEGIE-MELLON UNIV PITTSBURGH PA ARTIFICIAL INTELLIGENCE AND PSYCHOLOGY , 1989.

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

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

Nived Rajaraman, Lin F Yang, Jiantao Jiao, and Kannan Ramachandran. Toward the fundamental limits of imitation learning. In *arXiv preprint arXiv:2009.05990*, 2020.

Paria Rashidinejad, Banghua Zhu, Cong Ma, Jiantao Jiao, and Stuart Russell. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *Advances in Neural Information Processing Systems*, 34, 2021.

Stephane Ross and J Andrew Bagnell. Reinforcement and imitation learning via interactive no-regret learning. *arXiv preprint arXiv:1406.5979*, 2014.

Maurice Sion. On general minimax theorems. *Pacific Journal of mathematics*, 8(1):171–176, 1958.

Wen Sun, Anirudh Vemula, Byron Boots, and Drew Bagnell. Provably efficient imitation learning from observation alone. In *International conference on machine learning*, pages 6036–6045. PMLR, 2019.

Ziyang Tang, Yihao Feng, Lihong Li, Dengyong Zhou, and Qiang Liu. Doubly robust bias reduction in infinite horizon off-policy estimation. In *International Conference on Learning Representations*, 2019.

Masatoshi Uehara and Wen Sun. Pessimistic model-based offline rl: Pac bounds and posterior sampling under partial coverage. In *arXiv preprint arXiv:2107.06226*, 2021.

Masatoshi Uehara, Jiawei Huang, and Nan Jiang. Minimax Weight and Q-Function Learning for Off-Policy Evaluation. In *Proceedings of the 37th International Conference on Machine Learning*, pages 1023–1032, 2020.

Mengdi Wang. Primal-dual pi learning: Sample complexity and sublinear run time for ergodic markov decision problems. *arXiv preprint arXiv:1710.06100*, 2017.

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

Ruosong Wang, Dean P Foster, and Sham M Kakade. What are the statistical limits of offline rl with linear function approximation? *arXiv preprint arXiv:2010.11895*, 2020.

Ruosong Wang, Yifan Wu, Ruslan Salakhutdinov, and Sham M Kakade. Instabilities of offline rl with pre-trained neural representation. In *arXiv preprint arXiv:2103.04947*, 2021a.

Yuanhao Wang, Ruosong Wang, and Sham M. Kakade. An exponential lower bound for linearly-realizable mdps with constant suboptimality gap. arXiv preprint arXiv:2103.12690, 2021b.

Gellert Weisz, Philip Amortila, and Csaba Szepesvári. Exponential lower bounds for planning in mdps with linearly-realizable optimal action-value functions. *arXiv preprint arXiv:2010.01374*, 2020.

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

Tengyang Xie and Nan Jiang. Batch value-function approximation with only realizability. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 11404–11413. PMLR, 18–24 Jul 2021a. URL https://proceedings.mlr.press/v139/xie21d.html.

Tengyang Xie and Nan Jiang. Batch value-function approximation with only realizability. In *International Conference on Machine Learning*, pages 11404–11413. PMLR, 2021b.

Tengyang Xie, Ching-An Cheng, Nan Jiang, Paul Mineiro, and Alekh Agarwal. Bellman-consistent pessimism for offline reinforcement learning. *arXiv preprint arXiv:2106.06926*, 2021.

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

Ruiyi Zhang, Bo Dai, Lihong Li, and Dale Schuurmans. Gendice: Generalized offline estimation of stationary values. In *ICLR*, 2020.

Appendix A. Robustness to approximation and optimization errors

In this section we consider the setting where $\mathcal{V} \times \mathcal{W}$ may not contain (v_α^*, w_α^*) and measure the approximation errors as follows:

$$\epsilon_{\alpha,r,v} = \min_{v\in\mathcal{V}} \|v - v_\alpha^*\|_{1,\mu_0} + \|v - v_\alpha^*\|_{1,d^D} + \|v - v_\alpha^*\|_{1,d^{D'}}, \tag{28}$$

$$\epsilon_{\alpha,r,w} = \min_{w\in\mathcal{W}} \|w - w_\alpha^*\|_{1,d^D}, \tag{29}$$

where $d^{D'}(s) = \sum_{s',a'} d^D(s',a')P(s|s',a'), \forall s \in \mathcal{S}$. Notice that our definitions of approximation errors are all in ℓ_1 norm and weaker than ℓ_∞ norm error.

Besides, to make our algorithm work in practice, we also assume $(\widehat{v}, \widehat{w})$ is an approximate solution of $\widehat{L}_\alpha(v, w)$:

$$\widehat{L}_\alpha(\widehat{v}, \widehat{w}) - \min_{v\in\mathcal{V}} \widehat{L}_\alpha(v, \widehat{w}) \le \epsilon_{o,v}, \tag{30}$$

$$\max_{w\in\mathcal{W}} \min_{v\in\mathcal{V}} \widehat{L}_\alpha(v, w) - \min_{v\in\mathcal{V}} \widehat{L}_\alpha(v, \widehat{w}) \le \epsilon_{o,w}. \tag{31}$$

Equation (30) says that $\widehat{L}_\alpha(\widehat{v}, \widehat{w}) \approx \min_v \widehat{L}_\alpha(v, \widehat{w})$. Equation (31) says that $\min_v \widehat{L}_\alpha(v, \widehat{w}) \approx \max_{w\in\mathcal{W}} \min_{v\in\mathcal{V}} \widehat{L}_\alpha(v, w)$. Combining these gives $\widehat{L}_\alpha(\widehat{v}, \widehat{w}) \approx \max_{w\in\mathcal{W}} \min_{v\in\mathcal{V}} \widehat{L}_\alpha(v, w)$, so $(\widehat{v}, \widehat{w})$ is approximately a max-min point.

In this case we call the algorithm `Inexact-PRO-RL`. Theorem 16 shows that `Inexact-PRO-RL` is also capable of learning a near-optimal policy with polynomial sample size:

Theorem 16 (Error-robust version of Theorem 3) *Assume $\alpha > 0$. Suppose Assumption 1,3,4,5 hold. Then with at least probability $1-\delta$, the output of `Inexact-PRO-RL` satisfies:*

$$J(\pi_\alpha^*) - J(\widehat{\pi}) \le \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\alpha^*}[\|\pi_\alpha^*(\cdot|s) - \widehat{\pi}(\cdot|s)\|_1] \le \frac{2}{1-\gamma}\|\widehat{w} - w_\alpha^*\|_{2,d^D}$$
$$\le \frac{4}{1-\gamma}\sqrt{\frac{\mathcal{E}_{n,n_0,\alpha}(B_{w,\alpha}, B_{f,\alpha}, B_{v,\alpha}, B_{e,\alpha})}{\alpha M_f}} + \frac{2}{1-\gamma}\sqrt{\frac{2(\epsilon_{opt} + \epsilon_{\alpha,app})}{\alpha M_f}}, \tag{32}$$

where $B_{e,\alpha}$ is defined as Theorem 3, $\epsilon_{opt} = \epsilon_{o,v} + \epsilon_{o,w}$ and $\epsilon_{\alpha,app} = (B_{w,\alpha}+1)\,\epsilon_{\alpha,r,v} + (B_{e,\alpha} + \alpha B_{f',\alpha})\epsilon_{\alpha,r,w}$.

Proof [Proof sketch] The proof follows similar steps in the proof of Theorem 3. See Appendix K for details. ■

Remark 17 (Optimization) *When $\mathcal{W}$ and $\mathcal{V}$ are convex sets,[3] a line of algorithms (Nemirovski, 2004; Nesterov, 2007; Lin et al., 2020) are shown to attain $\widetilde{\epsilon}$-saddle point with the gradient complexity of $\widetilde{O}(\frac{1}{\widetilde{\epsilon}})$. Notice that an approximate saddle point will satisfy our requirements (30)(31) automatically, therefore we can choose these algorithms to solve $(\widehat{v}, \widehat{w})$. In more general cases, $\mathcal{W}$ and $\mathcal{V}$ might be parameterized by θ and ϕ. As long as the corresponding maximin problem (7) is still concave-convex (e.g., $\mathcal{W}$ and $\mathcal{V}$ are linear function classes), these algorithms can still work efficiently.*

3. In this case they are infinite classes, and we can simply replace the concentration bound in Lemma 34 with a standard covering argument

Similar to Corollary 5, we can extend Theorem 16 to compete with π_0^*. Suppose we select $\alpha = \alpha_{un} > 0$ in `Inexact-PRO-RL` and let $\epsilon_{un} = \alpha_{un} B_{f,0} + \frac{2}{1-\gamma}\sqrt{\frac{2(\epsilon_{opt}+\epsilon_{\alpha_{un},app})}{\alpha_{un} M_f}}$. Then we have the following corollary:

Corollary 18 (Error-robust version of Corollary 5) *Fix $\alpha_{un} > 0$. Suppose there exists $d_0^* \in D_0^*$ such that Assumption 1 holds. Besides, assume that Assumptions 1,3,4,5 hold for $\alpha = \alpha_{un}$. Then the output of `Inexact-PRO-RL` with input $\alpha = \alpha_{un}$ satisfies*

$$J(\pi_0^*) - J(\widehat{\pi}) \leq \frac{4}{1-\gamma}\sqrt{\frac{\mathcal{E}_{n,n_0,\alpha_{un}}(B_{w,\alpha_{un}}, B_{f,\alpha_{un}}, B_{v,\alpha_{un}}, B_{e,\alpha_{un}})}{\alpha_{un} M_f}} + \epsilon_{un}, \tag{33}$$

with at least probability $1-\delta$.

Proof [Proof sketch] The proof largely follows that of Corollary 5 and thus is omitted here. ■

The selection of α_{un} The best α_{un} we can expect (i.e., with the lowest error floor) is

$$\alpha_{un} := \arg\min_{\alpha>0} \left(\alpha B_{f,0} + \frac{2}{1-\gamma}\sqrt{\frac{2(\epsilon_{opt}+\epsilon_{\alpha,app})}{\alpha M_f}} \right). \tag{34}$$

However, this requires knowledge of $\epsilon_{\alpha,app}$, which is often unknown in practice. One alternative method is to suppose $\epsilon_{\alpha,app}$ upper bounded by ϵ_{app} for some $\alpha \in I_\alpha$, then α_{un} can be chosen as

$$\alpha_{un} := \arg\min_{\alpha\in I_\alpha} \left(\alpha B_{f,0} + \frac{2}{1-\gamma}\sqrt{\frac{2(\epsilon_{opt}+\epsilon_{app})}{\alpha M_f}} \right). \tag{35}$$

Notice that $B_{f,0}$ is known and ϵ_{opt} can be controlled by adjusting the parameters of the optimization algorithm, therefore the above α_{un} can be calculated easily.

Higher error floor In the ideal case of no approximation/optimization errors, Corollary 5 (which competes with π_0^*) has a worse sample complexity than Theorem 3 (which only competes with π_α^*). However, with the presence of approximation and optimization errors, the sample complexities become the same in Theorem 16 and Corollary 18, but the latter has a higher error floor. To see this, we can suppose $\epsilon_{\alpha,app}$ are uniformly upper bounded by ϵ_{app}, then $\alpha_{un} = O((\epsilon_{opt}+\epsilon_{app})^{\frac{1}{3}})$ by the AM-GM inequality and $\epsilon_{un} = O((\epsilon_{opt}+\epsilon_{app})^{\frac{1}{3}})$, which is larger than $O((\epsilon_{opt}+\epsilon_{app})^{\frac{1}{2}})$ as in Theorem 16.

Appendix B. Policy extraction via behavior cloning

In this section we consider an *unknown* behavior policy π_D. Notice that the only place we require π_D in our algorithm is the policy extraction step, where we compute $\widehat{\pi}$ from $\widehat{w}$ using knowledge of π_D. Inspired by the imitation learning literature (Pomerleau, 1989; Ross and Bagnell, 2014; Agarwal et al., 2020), we will use behavior cloning to compute a policy $\overline{\pi}$ to approximate $\widehat{\pi}$, where $\widehat{\pi}$ is not directly available and only implicitly defined via $\widehat{w}$ and the data.

As is standard in the literature (Ross and Bagnell, 2014; Agarwal et al., 2020), we utilize a policy class Π to approximate the target policy. We suppose Π is realizable:

Assumption 11 (Realizability of Π) *Assume* $\pi_\alpha^* \in \Pi$.

One may be tempted to assume $\widehat{\pi} \in \Pi$, since $\widehat{\pi}$ is the target of imitation, but $\widehat{\pi}$ is a function of the data and hence random. A standard way of "determinizing" such an assumption is to assume the realizability of Π for *all possible* $\widehat{\pi}$ that can be induced by any $w \in \mathcal{W}$, which leads to a prohibitive "completeness"-type assumption. Fortunately, as we have seen in previous sections, $\widehat{\pi}$ will be close to π_α^* when learning succeeds, so the realizability of π_α^*—a policy whose definition does not depend on data randomness—suffices for our purposes.

In the rest of this section, we design a novel behavior cloning algorithm which is more robust compared to the classic maximum likelihood estimation process (Pomerleau, 1989; Ross and Bagnell, 2014; Agarwal et al., 2020). In MLE behavior cloning, the KL divergence between the target policy and the policy class need to be bounded while in our algorithm we only require the weighted ℓ_1 distance to be bounded. This property is important in our setting, as PRO-RL can only guarantee a small weighted ℓ_2 distance between π_α^* and $\widehat{\pi}$; ℓ_2 distance is stronger than ℓ_1 while weaker than KL divergence.

Our behavior cloning algorithm is inspired by the algorithms in Sun et al. (2019); Agarwal et al. (2019), which require access to d^π for all $\pi \in \Pi$ and are not satisfied in our setting. However, the idea of estimating total variation by the variational form turns out to be useful. More concretely, for any two policies π and π', define:

$$h^s_{\pi,\pi'} := \arg \max_{h:\|h\|_\infty \le 1} [\mathbb{E}_{a\sim\pi(\cdot|s)} h(a) - \mathbb{E}_{a\sim\pi'(\cdot|s)} h(a)]. \tag{36}$$

Let $h_{\pi,\pi'}(s,a) = h^s_{\pi,\pi'}(a), \forall s, a$. Note that the function $h_{\pi,\pi'}$ is purely a function of π and π' and does not depend on the data or the MDP, and hence can be computed exactly even before we see the data. Such a function witnesses the ℓ_1 distance between π and π', as shown in the following lemma; see proof in Appendix L.1:

Lemma 19 *For any distribution d on $\mathcal{S}$ and policies π, π' , we have:*

$$\mathbb{E}_{s\sim d}[\|\pi(\cdot|s) - \pi'(\cdot|s)\|_1] = \mathbb{E}_{s\sim d}\left[\mathbb{E}_{a\sim\pi(\cdot|s)}[h_{\pi,\pi'}(s,a)] - \mathbb{E}_{a\sim\pi'(\cdot|s)}[h_{\pi,\pi'}(s,a)]\right]. \tag{37}$$

Inspired by Lemma 19, we can estimate the total variation distance between π and π' by evaluating $\mathbb{E}_{a\sim\pi(\cdot|s)}[h_{\pi,\pi'}(s,a)] - \mathbb{E}_{a\sim\pi'(\cdot|s)}[h_{\pi,\pi'}(s,a)]$ empirically. Let $\mathcal{H} := \{h_{\pi,\pi'} : \pi, \pi' \in \Pi\}$ and we have $|\mathcal{H}| \le |\Pi|^2$. We divide $\mathcal{D}$ into $\mathcal{D}_1$ and $\mathcal{D}_2$ where $\mathcal{D}_1$ is utilized for evaluating $\widehat{w}$ and $\mathcal{D}_2$ for obtaining $\overline{\pi}$. Let n_1 and n_2 denote the number of samples in $\mathcal{D}_1$ and $\mathcal{D}_2$. Then our behavior cloning algorithm is based on the following objective function, whose expectation is $\mathbb{E}_{s\sim\widehat{d},a\sim\widehat{\pi}}[h^\pi(s) - h(s,a)]$ and by Lemma 19 is exactly the TV between $\widehat{\pi}$ and π:

$$\overline{\pi} = \arg \min_{\pi\in\Pi} \max_{h\in\mathcal{H}} [\sum_{i=1}^{n_2} \widehat{w}(s_i,a_i)\,(h^\pi(s_i) - h(s_i,a_i))], \tag{38}$$

where $(s_i,a_i) \in \mathcal{D}_2, \forall 1 \le i \le n_2$, $h^\pi(s) = \mathbb{E}_{a\sim\pi(\cdot|s)}[h(s,a)]$ and $\overline{\pi}$ is the ultimate output policy.

It can be observed that (38) is the importance-sampling version of

$$\mathbb{E}_{s\sim\widehat{d}}[\mathbb{E}_{a\sim\pi(\cdot|s)}[h(s,a)] - \mathbb{E}_{a\sim\widehat{\pi}(\cdot|s)}[h(s,a)]]. \tag{39}$$

Since $\widehat{d}$ is close to d^*_α, by minimizing (38) we can find a policy that approximately minimizes $\mathbb{E}_{s\sim d^*_\alpha}[\|\pi(\cdot|s) - \widehat{\pi}(\cdot|s)\|_1]$. We call PRO-RL with this behavior cloning algorithm by PRO-RL-BC.

Theorem 20 shows that PRO-RL-BC can attain almost the same sample complexity as PRO-RL in Theorem 3 where π_D is known.

Theorem 20 (Sample complexity of learning π_α^* with unknown behavior policy) *Assume $\alpha > 0$. Suppose Assumption 1,2,3,4,5 and 11 hold. Then with at least probability $1-\delta$, the output of* `PRO-RL-BC` *satisfies:*

$$J(\pi_\alpha^*) - J(\overline{\pi}) \leq \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\alpha^*}[\|\pi_\alpha^*(\cdot|s) - \overline{\pi}(\cdot|s)\|_1]$$
$$\leq \frac{4B_{w,\alpha}}{1-\gamma}\sqrt{\frac{6\log\frac{4|\Pi|}{\delta}}{n_2}} + \frac{50}{1-\gamma}\sqrt{\frac{\mathcal{E}_{n_1,n_0,\alpha}(B_{w,\alpha}, B_{f,\alpha}, B_{v,\alpha}, B_{e,\alpha})}{\alpha M_f}}, \tag{40}$$

where $B_{e,\alpha}$ is defined as in Theorem 3.

Proof See Appendix L.2 for details. ■

Remark 21 *Notice that the error scales with $O(\frac{1}{\sqrt{n_2}})$ and $O(\frac{1}{n_1^{\frac{1}{4}}})$, which means that the extra samples required by behavior cloning only affects the higher-order terms. Therefore the total sample complexity $n = n_1 + n_2$ is dominated by n_1, which coincides with the sample complexity of Theorem 3.*

Remark 22 `PRO-RL-BC` *might not be computationally efficient in its current form since the structures of the policy class Π (such as convexity) may not be retained in $\mathcal{H}$. Designing a well-structured $\mathcal{H}$ can help mitigate this problem but may incur additional errors. Here our main purpose though is to investigate the statistical properties of* `PRO-RL-BC` *and thus will not go deep about this.*

Similarly, behavior cloning can be extended to the unregularized setting where we compete with π_0^*, and the sample complexity will remain almost the same as Corollary 5:

Corollary 23 *Fix any $\epsilon > 0$. Suppose there exists $d_0^* \in D_0^*$ such that Assumption 1 holds. Besides, assume that Assumption 1,2,3,4,5 and 11 hold for $\alpha = \alpha_\epsilon$. Then if*

$$n_0 \geq C_2 \cdot \frac{(2B_{v,\alpha_\epsilon}B_{f,0})^2}{\epsilon^6 M_f^2(1-\gamma)^2} \cdot \log\frac{4|\mathcal{V}|}{\delta}, \tag{41}$$

$$n_1 \geq C_3 \cdot \frac{(\epsilon B_{f,\alpha_\epsilon} + 2B_{w,\alpha_\epsilon}B_{e,\alpha_\epsilon}B_{f,0})^2}{\epsilon^6 M_f^2(1-\gamma)^4} \cdot \log\frac{|\mathcal{V}||\mathcal{W}|}{\delta}, \tag{42}$$

$$n_2 \geq C_4 \cdot \frac{(B_{w,\alpha_\epsilon})^2}{(1-\gamma)^2\epsilon^2}\log\frac{|\Pi|}{\delta}, \tag{43}$$

where C_2, C_3, C_4 are some universal positive constants, the output of `PRO-RL-BC` *with input $\alpha = \alpha_\epsilon$ satisfies*

$$J(\pi_0^*) - J(\overline{\pi}) \leq \epsilon, \tag{44}$$

with at least probability $1-\delta$.

Proof The proof is the same as in Appendix G. The only difference is that we replace the result in Theorem 3 with Theorem 20. ■

Remark 24 *The sample complexity to obtain ϵ-optimal policy is still $\widetilde{O}\left(\frac{(B_{w,0})^4(B_{w,\alpha_\epsilon})^2}{\epsilon^6(1-\gamma)^6}\right)$ since n_2 is negligible compared to n_1.*

Remark 25 *Similar to Corollary 5, the concentrability assumptions in Corollary 23 can be reduced to single-policy concentrability with the help of Corollary 12.*

Appendix C. PRO-RL with $\alpha = 0$

From the previous discussions, we notice that when $\alpha > 0$, extending from regularized problems to unregularized problems will cause worse sample complexity in PRO-RL (Section 4, Appendix A). Also, the realizability assumptions are typically with respect to the regularized optimizers rather than the more natural (v_0^*, w_0^*). In this section we show that by using stronger concentrability assumptions, PRO-RL can still have guarantees with $\alpha = 0$ under the realizability w.r.t. (v_0^*, w_0^*) and attain a faster rate. More specifically, we need the following strong concentration assumption:

Assumption 12 (Strong concentrability) *Suppose the dataset distribution d^D and some $d_0^* \in D_0^*$ satisfy*

$$\frac{d^\pi(s)}{d^D(s)} \leq B_{w,u}, \forall \pi, s \in \mathcal{S}, \tag{45}$$

$$\frac{d_0^*(s)}{d^D(s)} \geq B_{w,l} > 0, \forall s \in \mathcal{S}. \tag{46}$$

Remark 26 *Eq. (45) is the standard all-policy concentrability assumption in offline RL (Chen and Jiang, 2019; Nachum et al., 2019b; Xie and Jiang, 2020). In addition, Assumption 12 requires the density ratio of the optimal policy is lower bounded, which is related to an ergodicity assumption used in some previous works in the simulator setting (Wang, 2017, 2020).*

Remark 27 *It can be observed that $B_{w,l} = 0$ in the counterexample in Section 5.1 and thus the counterexample does not satisfy Assumption 12.*

In the following discussion w_0^* and π_0^* are specified as the optimal density ratio and policy with respect to the d_0^* in Assumption 12. We need to impose some constraints on the function class $\mathcal{W}$ and $\mathcal{V}$ so that $d^{\widehat{\pi}}$ can be upper bounded by $\widehat{w} \cdot d^D$.

Assumption 13 *Suppose*

$$\mathcal{W} \subseteq \overline{\mathcal{W}} := \left\{ w(s,a) \geq 0, \sum_a \pi_D(a|s) w(s,a) \geq B_{w,l}, \forall s \in \mathcal{S}, a \in \mathcal{A} \right\}, \tag{47}$$

Given a function class $\mathcal{W}$, this assumption is trivially satisfied by removing the $w \in \mathcal{W}$ that are not in $\overline{\mathcal{W}}$ when π_D is known.

Assumption 14 *Suppose*

$$0 \leq v(s) \leq \frac{1}{1-\gamma}, \forall s \in \mathcal{S}, v \in \mathcal{V}. \tag{48}$$

By Assumption 12, $w_0^* \in \overline{\mathcal{W}}$ and $0 \leq v_0^* \leq \frac{1}{1-\gamma}$. Therefore Assumption 13 and Assumption 14 are reasonable.

With strong concentrability, we can show that PRO-RL with $\alpha = 0$ can learn an ϵ-optimal policy with sample complexity $n = \widetilde{O}\left(\frac{1}{\epsilon^2}\right)$:

Corollary 28 *Suppose Assumption 1,2,3, 13, 14 and 12 hold for $\alpha = 0$. Then with at least probability $1-\delta$, the output of PRO-RL with input $\alpha = 0$ satisfies:*

$$J(\pi_0^*) - J(\widehat{\pi}) \leq \frac{2B_{w,0}B_{w,u}}{(1-\gamma)B_{w,l}}\sqrt{\frac{2\log\frac{4|\mathcal{V}||\mathcal{W}|}{\delta}}{n}} + \frac{B_{w,u}}{B_{w,l}}\sqrt{\frac{2\log\frac{4|\mathcal{V}|}{\delta}}{n_0}}, \tag{49}$$

Proof The key idea is to utilize Lemma 35 to bound $L_0(v_0^*, w_0^*) - L_0(v_0^*, \widehat{w})$ and then quantify the performance difference $J(\pi_0^*) - J(\widehat{\pi})$. See Appendix M for details. ■

Comparison with $\alpha > 0$ and $\alpha = 0$. When solving the unregularized problem, PRO-RL with $\alpha = 0$ has better sample complexity than Corollary 5. Also the realizability assumptions in Corollary 28 are with respect to the optimizers of the unregularized problem itself, which is not the case in Corollary 5 when ϵ is large. However, PRO-RL with $\alpha = 0$ only works under a very strong concentrability assumption (Assumption 12) and thus is less general than PRO-RL with $\alpha > 0$.

Appendix D. Additional Discussion

D.1. Discussion about Assumption 12

The following ergodicity assumption has been introduced in some online reinforcement learning works (Wang, 2017, 2020):

Assumption 15 *Assume*

$$B_{erg,1}\mu_0(s) \leq d^\pi(s) \leq B_{erg,2}\mu_0(s), \forall s, \pi. \tag{50}$$

Remark 29 *The original definition of ergodicity in Wang (2017, 2020) is targeted at the stationary distribution induced by policy π rather than the discounted visitation distribution. However, this is not an essential difference and it can be shown that Corollary 28 still holds under the definition in Wang (2017, 2020). Here we define ergodicity with respect to the discounted visitation distribution for the purpose of comparing Assumption 12 and 15.*

In fact, our Assumption 12 is weaker than Assumption 15 as shown in the following lemma:

Lemma 30 *Suppose $d^\pi(s) \leq B_{erg,2}\mu_0(s), \forall s, \pi$ and Assumption 9 holds, then we have:*

$$\frac{d^\pi(s)}{d^D(s)} \leq \frac{B_{erg,2}}{1-\gamma}, \forall \pi, s \tag{51}$$

$$\frac{d_0^*(s)}{d^D(s)} \geq \frac{1-\gamma}{B_{erg,2}}, \forall s. \tag{52}$$

The proof is deferred to Appendix M.1. Lemma 30 shows that the upper bound in Assumption 15 implies Assumption 12. Therefore, our strong concentration assumption is a weaker version of the ergodicity assumption.

D.2. Combination of different practical factors

In previous sections, we generalized PRO-RL to several more realistic settings (poor coverage, approximation and optimization error, unknown behavior policy). In fact, PRO-RL with $\alpha > 0$ can be even generalized to include all of the three settings by combining Theorem 3,10,16,20 and Corollaries 5,12,18,23. For brevity, we do not list all the combinations separately and only illustrate how to handle each individually.

For PRO-RL with $\alpha = 0$, it is easy to extend Corollary 28 to approximation and optimization error but relaxation of the concentration assumption and unknown behavior policy is difficult. This is because the analysis of Corollary 28 relies on the fact that v_0^* is the optimal value function of the unregularized problem (1). Consequently, the same analysis is not applicable to $(w^*_{0,B_w}, v^*_{0,B_w})$. Furthermore, Assumption 13 requires knowing π_D and thus hard to enforce with unknown behavior policy.

Appendix E. Analysis for regularized offline RL (Theorem 3)

In this section we present the analysis for our main result in Theorem 3.

E.1. Intuition: invariance of saddle points

First we would like to provide an intuitive explanation why optimizing $\mathcal{V} \times \mathcal{W}$ instead of $\mathbb{R}^{|\mathcal{S}|} \times \mathbb{R}_+^{|\mathcal{S}||\mathcal{A}|}$ can still bring us close to (v_α^*, w_α^*). More specifically, we have the following lemma:

Lemma 31 (Invariance of saddle points) *Suppose (x^*, y^*) is a saddle point of $f(x, y)$ over $\mathcal{X} \times \mathcal{Y}$, then for any $\mathcal{X}' \subseteq \mathcal{X}$ and $\mathcal{Y}' \subseteq \mathcal{Y}$, if $(x^*, y^*) \in \mathcal{X}' \times \mathcal{Y}'$, we have:*

$$(x^*, y^*) \in \arg\min_{x \in \mathcal{X}'} \arg\max_{y \in \mathcal{Y}'} f(x, y), \tag{53}$$

$$(x^*, y^*) \in \arg\max_{y \in \mathcal{Y}'} \arg\min_{x \in \mathcal{X}'} f(x, y). \tag{54}$$

Proof See Appendix F.1. ■

Lemma 31 shows that as long as a subset includes the saddle point of the original set, the saddle point will still be a minimax and maximin point with respect to the subset. We apply this to (6): the saddle point (v_α^*, w_α^*) of (6), also the solution to the regularized MDP without any restriction on function classes, is also a solution of $\max_{w \in \mathcal{W}} \min_{v \in \mathcal{V}} L_\alpha(v, w)$.

We now give a brief sketch. Since $\widehat{L}$ is unbiased for L_α, using uniform convergence, $\widehat{L}_\alpha(v, w) \approx L_\alpha(v, w)$ with high probability. Next, use strong concavity of $L_\alpha(v, w)$ with respect to w, to show that $\widehat{w} \approx w_\alpha^*$. This implies that $\widehat{\pi} \approx \pi_\alpha^*$, which is exactly Theorem 3.

E.2. Preparation: boundedness of v_α^*

Before proving Theorem 3, an important ingredient is to bound v_α^* since $\mathcal{V}$ is assumed to be a bounded set (Assumption 5). The key idea is to utilize KKT conditions and the fact that for each $s \in \mathcal{S}$ there exists $a \in \mathcal{A}$ such that $w_\alpha^*(s, a) > 0$. The consequent bound is given in Lemma 32.

Lemma 32 (Boundedness of v_α^*) *Suppose Assumption 1 and 4 holds, then we have:*

$$\|v_\alpha^*\|_\infty \le B_{v,\alpha} := \frac{\alpha B_{f',\alpha} + 1}{1 - \gamma}. \tag{55}$$

Proof See Appendix F.2. ■

E.3. Proof sketch of Theorem 3

As stated in Section E.1, our proof consists of (1) using concentration inequalities to bound $|L_\alpha(v,w) - \widehat{L}_\alpha(v,w)|$, (2) using the invariance of saddle points and concentration bounds to characterize the error $\|\widehat{w} - w_\alpha^*\|_{2,d^D}$ and (3) analyzing the difference between $\widehat{\pi}$ and π_α^*. We will elaborate on each of these steps in this section.

Concentration of $\widehat{L}_\alpha(v,w)$. First, it can be observed that $\widehat{L}_\alpha(v,w)$ is an unbiased estimator of $L_\alpha(v,w)$, as shown in the following lemma

Lemma 33

$$\mathbb{E}_{\mathcal{D}}[\widehat{L}_\alpha(v,w)] = L_\alpha(v,w), \quad \forall v \in \mathcal{V}, w \in \mathcal{W}, \tag{56}$$

where $\mathbb{E}_{\mathcal{D}}[\cdot]$ is the expectation with respect to the samples in $\mathcal{D}$, i.e., $(s_i,a_i) \sim d^D, s_i' \sim P(\cdot|s_i,a_i)$.

Proof See Appendix F.3. ■

On the other hand, note that from the boundedness of $\mathcal{V}, \mathcal{W}$ and f (Assumption 5, 3, 4), $\widehat{L}_\alpha(v,w)$ is also bounded. Combining with Lemma 33, we have the following lemma:

Lemma 34 *Suppose Assumption 3,4,5 hold. Then with at least probability $1-\delta$, for all $v \in \mathcal{V}$ and $w \in \mathcal{W}$ we have:*

$$|\widehat{L}_\alpha(v,w) - L_\alpha(v,w)| \le \mathcal{E}_{n,n_0,\alpha}(B_{w,\alpha}, B_{f,\alpha}, B_{v,\alpha}, B_{e,\alpha}) := \epsilon_{stat}, \tag{57}$$

Proof See Appendix F.4. ■

Bounding $\|\widehat{w} - w_\alpha^*\|_{2,d^D}$. To bound $\|\widehat{w} - w_\alpha^*\|_{2,d^D}$, we first need to characterize $L_\alpha(v_\alpha^*, w_\alpha^*) - L_\alpha(v_\alpha^*, \widehat{w})$. Inspired by Lemma 31, we decompose $L_\alpha(v_\alpha^*, w_\alpha^*) - L_\alpha(v_\alpha^*, \widehat{w})$ carefully and utilize the concentration results Lemma 34, which leads us to the following lemma:

Lemma 35 *Suppose Assumption 1,2,3,4 and 5 hold. Then with at least probability $1-\delta$,*

$$L_\alpha(v_\alpha^*, w_\alpha^*) - L_\alpha(v_\alpha^*, \widehat{w}) \le 2\epsilon_{stat}. \tag{58}$$

Proof See Appendix F.5. ■

Then due to the strong convexity of f which leads to L_α being strongly concave in w, $\|\widehat{w} - w_\alpha^*\|_{2,d^D}$ can be naturally bounded by Lemma 35,

Lemma 36 *Suppose Assumption 1,2,3,4,5 hold. Then with at least probability $1-\delta$,*

$$\|\widehat{w} - w_\alpha^*\|_{2,d^D} \le \sqrt{\frac{4\epsilon_{stat}}{\alpha M_f}}, \tag{59}$$

which implies that

$$\|\widehat{d} - d_\alpha^*\|_1 \le \sqrt{\frac{4\epsilon_{stat}}{\alpha M_f}}, \tag{60}$$

where $\widehat{d}(s,a) = \widehat{w}(s,a)d^D(s,a), \forall s,a$.

Proof See Appendix F.6. ■

Bounding $\mathbb{E}_{s\sim d^*_\alpha}[\|\pi^*_\alpha(s,\cdot)-\widehat{\pi}(s,\cdot)\|_1]$. To obtain the second part of (13), we notice that π^*_α (or $\widehat{\pi}$) can be derived explicitly from w^*_α (or $\widehat{w}$) by (4) (or (9)). However, the mapping $w^*_\alpha \mapsto \pi^*_\alpha$ (or $\widehat{w} \mapsto \widehat{\pi}$) is not linear and discontinuous when $d^*_\alpha(s)=0$ (or $\widehat{d}(s)=0$), which makes the mapping complicated. To tackle with this problem, we first decompose the error $\|\widehat{w}-w^*_\alpha\|_{2,d^D}$ and assign to each state $s\in\mathcal{S}$, then consider the case where $\widehat{d}(s)>0$ and $\widehat{d}(s)=0$ separately. Consequently, we can obtain the following lemma:

Lemma 37

$$\mathbb{E}_{s\sim d^*_\alpha}[\|\pi^*_\alpha(s,\cdot)-\widehat{\pi}(s,\cdot)\|_1] \le 2\|\widehat{w}-w^*_\alpha\|_{2,d^D}. \tag{61}$$

Proof See Appendix F.7. ■

Combining Equation (59), (61), and the definition of ϵ_{stat} from Lemma 34, gives us the second part of Theorem 3.

Bounding $J(\pi^*_\alpha)-J(\widehat{\pi})$. To complete the proof of Theorem 3, we only need to bound $J(\pi^*_\alpha)-J(\widehat{\pi})$ via the bounds on $\mathbb{E}_{s\sim d^*_\alpha}[\|\pi^*_\alpha(s,\cdot)-\widehat{\pi}(s,\cdot)\|_1]$, which is shown in the following lemma:

Lemma 38

$$J(\pi^*_\alpha)-J(\widehat{\pi}) \le \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^*_\alpha}[\|\pi^*_\alpha(s,\cdot)-\widehat{\pi}(s,\cdot)\|_1]. \tag{62}$$

Proof See Appendix F.8. ■

This concludes the proof of Theorem 3.

Appendix F. Proofs of Lemmas for Theorem 3

F.1. Proof of Lemma 31

We first prove that $(x^*,y^*)\in \arg\min_{x\in\mathcal{X}'}\arg\max_{y\in\mathcal{Y}'} f(x,y)$. Since (x^*,y^*) is a saddle point (Sion, 1958), we have

$$x^* = \arg\min_{x\in\mathcal{X}} f(x,y^*), y^* = \arg\max_{y\in\mathcal{Y}} f(x^*,y). \tag{63}$$

Since $\mathcal{Y}'\subseteq\mathcal{Y}$ and $y^*\in\mathcal{Y}'$, we have:

$$f(x^*,y^*) = \max_{y\in\mathcal{Y}'} f(x^*,y). \tag{64}$$

On the other hand, because $\mathcal{X}'\subseteq\mathcal{X}$ and $y^*\in\mathcal{Y}'$,

$$f(x^*,y^*) \le f(x,y^*) \le \max_{y\in\mathcal{Y}'} f(x,y), \forall x\in\mathcal{X}'. \tag{65}$$

Notice that $x^*\in\mathcal{X}'$, so we have:

$$\max_{y\in\mathcal{Y}'} f(x^*,y) = \min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}'} f(x,y), \tag{66}$$

or equivalently,

$$(x^*,y^*)\in \arg\min_{x\in\mathcal{X}'}\arg\max_{y\in\mathcal{Y}'} f(x,y). \tag{67}$$

On the other hand, by a similar proof we have

$$f(x^*, y^*) \geq f(x^*, y) \geq \min_{x \in \mathcal{X}'} f(x, y), \quad \forall y \in \mathcal{Y}', \tag{68}$$

which implies that

$$(x^*, y^*) \in \arg\max_{y \in \mathcal{Y}'} \arg\min_{x \in \mathcal{X}'} f(x, y). \tag{69}$$

F.2. Proof of Lemma 32

From the strong duality of the regularized problem (2)(3), when $d^D(s,a) \neq 0$, we have $w_\alpha^* = \arg\max_{w \geq 0} L_\alpha(v_\alpha^*, w)$, or

$$w_\alpha^*(s,a) = \max\left(0, (f')^{-1}\left(\frac{e_{v_\alpha^*}(s,a)}{\alpha}\right)\right). \tag{70}$$

Note that $d_\alpha^*(s,a) = w_\alpha^*(s,a) d^D(s,a)$ satisfies Bellman flow constraint (3), therefore

$$d_\alpha^*(s) \geq (1-\gamma)\mu_0(s) > 0, \quad \forall s \in \mathcal{S}, \tag{71}$$

which implies that for any $s \in \mathcal{S}$, $\exists a_s \in \mathcal{A}$ such that

$$d_\alpha^*(s, a_s) > 0, \tag{72}$$

or equivalently

$$w_\alpha^*(s, a_s) > 0, d^D(s, a_s) > 0. \tag{73}$$

Thus from (70) we know that

$$e_{v_\alpha^*}(s, a_s) = \alpha f'(w_\alpha^*(s, a_s)). \tag{74}$$

From Assumption 1, $w_\alpha^*(s, a_s) \leq B_{w,\alpha}$ and thus due to Assumption 4,

$$|e_{v_\alpha^*}(s, a_s)| \leq \alpha B_{f',\alpha}, \forall s \in \mathcal{S}. \tag{75}$$

On the other hand, suppose $|v_\alpha^*(s_m)| = \|v_\alpha^*\|_\infty$, then from the definition of e_v we have:

$$e_{v_\alpha^*}(s_m, a_{s_m}) = r(s_m, a_{s_m}) + \gamma \mathbb{E}_{s' \sim P(\cdot|s_m, a_{s_m})} v_\alpha^*(s') - v_\alpha^*(s_m), \tag{76}$$

which implies that:

$$\begin{aligned}
|e_{v_\alpha^*}(s_m, a_{s_m}) - r(s_m, a_{s_m})| &= |v_\alpha^*(s_m) - \gamma \mathbb{E}_{s' \sim P(\cdot|s_m, a_{s_m})} v_\alpha^*(s')| && (77)\\
&\geq |v_\alpha^*(s_m)| - \gamma |\mathbb{E}_{s' \sim P(\cdot|s_m, a_{s_m})} v_\alpha^*(s')| && (78)\\
&\geq |v_\alpha^*(s_m)| - \gamma \mathbb{E}_{s' \sim P(\cdot|s_m, a_{s_m})} |v_\alpha^*(s')| && (79)\\
&\geq (1-\gamma)|v_\alpha^*(s_m)|. && (80)
\end{aligned}$$

Combining (75) and (80), we have

$$\|v_\alpha^*\|_\infty \leq \frac{\alpha B_{f',\alpha} + 1}{1-\gamma}. \tag{81}$$

F.3. Proof of Lemma 33

First by the tower rule, we have:

$$\mathbb{E}_{\mathcal{D}}\left[\widehat{L}_\alpha(v,w)\right] = \mathbb{E}_{(s_i,a_i)\sim d^D, s_{0,j}\sim\mu_0}\left[\mathbb{E}_{s_i'\sim P(\cdot|s_i,a_i)}\left[\widehat{L}_\alpha(v,w)|s_i,a_i\right]\right]. \tag{82}$$

Note that

$$\begin{aligned}
&\mathbb{E}_{s_i'\sim P(\cdot|s_i,a_i)}\left[\widehat{L}_\alpha(v,w)|s_i,a_i\right] && (83)\\
=&(1-\gamma)\frac{1}{n_0}\sum_{j=1}^{n_0}[v(s_{0,j})] + \frac{1}{n}\sum_{i=1}^{n}[-\alpha f(w(s_i,a_i))] && (84)\\
+&\frac{1}{n}\sum_{i=1}^{n}[w(s_i,a_i)\mathbb{E}_{s_i'\sim P(\cdot|s_i,a_i)}\left[e_v(s_i,a_i,r_i,s_i')|s_i,a_i\right]] && (85)\\
=&(1-\gamma)\frac{1}{n_0}\sum_{j=1}^{n_0}[v(s_{0,j})] + \frac{1}{n}\sum_{i=1}^{n}[-\alpha f(w(s_i,a_i))] + \frac{1}{n}\sum_{i=1}^{n}[w(s_i,a_i)e_v(s_i,a_i)]. && (86)
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}}\left[\widehat{L}_\alpha(v,w)\right] && (87)\\
=&(1-\gamma)\mathbb{E}_{s\sim\mu_0}[v(s)] - \alpha\mathbb{E}_{(s,a)\sim d^D}[f(w(s,a))] + \mathbb{E}_{(s,a)\sim d^D}[w(s,a)e_v(s,a)] && (88)\\
=&L_\alpha(v,w). && (89)
\end{aligned}$$

F.4. Proof of Lemma 34

Let $l_i^{v,w} = -\alpha f(w(s_i,a_i)) + w(s_i,a_i)e_v(s_i,a_i,r_i,s_i')$. From Assumption 5, we know

$$|e_v(s,a,r,s')| = |r(s,a)+\gamma v(s') - v(s)| \le (1+\gamma)B_{v,\alpha} + 1 = B_{e,\alpha}. \tag{90}$$

Therefore, by Assumption 3 and 4, we have:

$$|l_i^{v,w}| \le \alpha B_{f,\alpha} + B_{w,\alpha}B_{e,\alpha}. \tag{91}$$

Notice that $l_i^{v,w}$ is independent from each other, thus we can apply Hoeffding's inequality and for any $t > 0$,

$$\Pr[|\frac{1}{n}\sum_{i=1}^{n} l_i^{v,w} - \mathbb{E}[l_i^{v,w}]| \le t] \ge 1 - 2\exp\left(\frac{-nt^2}{2(\alpha B_{f,\alpha} + B_{w,\alpha}B_{e,\alpha})^2}\right). \tag{92}$$

Let $t = (\alpha B_{f,\alpha} + B_{w,\alpha}B_{e,\alpha})\sqrt{\frac{2\log\frac{4|\mathcal{V}||\mathcal{W}|}{\delta}}{n}}$, we have with at least probability $1 - \frac{\delta}{2|\mathcal{V}||\mathcal{W}|}$,

$$|\frac{1}{n}\sum_{i=1}^{n} l_i^{v,w} - \mathbb{E}[l_i^{v,w}]| \le (\alpha B_{f,\alpha} + B_{w,\alpha}B_{e,\alpha})\sqrt{\frac{2\log\frac{4|\mathcal{V}||\mathcal{W}|}{\delta}}{n}}. \tag{93}$$

Therefore by union bound, with at least probability $1-\frac{\delta}{2}$, we have for all $v \in \mathcal{V}$ and $w \in \mathcal{W}$,

$$|\frac{1}{n}\sum_{i=1}^{n} l_i^{v,w} - \mathbb{E}[l_i^{v,w}]| \leq (\alpha B_{f,\alpha} + B_{w,\alpha}B_{e,\alpha})\sqrt{\frac{2\log\frac{4|\mathcal{V}||\mathcal{W}|}{\delta}}{n}}. \tag{94}$$

Similarly, we have with at least probability $1-\frac{\delta}{2}$, for all $v \in \mathcal{V}$,

$$|\frac{1}{n_0}\sum_{j=1}^{n_0} v(s_{0,j}) - \mathbb{E}_{s\sim\mu_0}[v(s)]| \leq B_{v,\alpha}\sqrt{\frac{2\log\frac{4|\mathcal{V}|}{\delta}}{n_0}}. \tag{95}$$

Therefore, with at least probability $1-\delta$ we have

$$|\widehat{L}_\alpha(v,w) - L_\alpha(v,w)| \leq (\alpha B_{f,\alpha} + B_{w,\alpha}B_{e,\alpha})\sqrt{\frac{2\log\frac{4|\mathcal{V}||\mathcal{W}|}{\delta}}{n}} + (1-\gamma)B_{v,\alpha}\sqrt{\frac{2\log\frac{4|\mathcal{V}|}{\delta}}{n_0}}. \tag{96}$$

F.5. Proof of Lemma 35

First we decompose $L_\alpha(v_\alpha^*, \widehat{w}) - L_\alpha(v_\alpha^*, w_\alpha^*)$ into the following terms:

$$L_\alpha(v_\alpha^*, \widehat{w}) - L_\alpha(v_\alpha^*, w_\alpha^*) = \underbrace{(L_\alpha(v_\alpha^*, \widehat{w}) - \widehat{L}_\alpha(v_\alpha^*, \widehat{w}))}_{(1)} + \underbrace{(\widehat{L}_\alpha(v_\alpha^*, \widehat{w}) - \widehat{L}_\alpha(\widehat{v}, \widehat{w}))}_{(2)}$$

$$+ \underbrace{(\widehat{L}_\alpha(\widehat{v}, \widehat{w}) - \widehat{L}_\alpha(\widehat{v}(w_\alpha^*), w_\alpha^*))}_{(3)} + \underbrace{(\widehat{L}_\alpha(\widehat{v}(w_\alpha^*), w_\alpha^*) - L_\alpha(\widehat{v}(w_\alpha^*), w_\alpha^*))}_{(4)} \tag{97}$$

$$+ \underbrace{(L_\alpha(\widehat{v}(w_\alpha^*), w_\alpha^*) - L_\alpha(v_\alpha^*, w_\alpha^*))}_{(5)}, \tag{98}$$

where $\widehat{v}(w) = \arg\min_{v\in\mathcal{V}} \widehat{L}_\alpha(v,w)$.

For term (1) and (4), we can apply Lemma 34 and thus

$$(1) \geq -\epsilon_{stat}, (4) \geq -\epsilon_{stat}. \tag{99}$$

For term (2), since $\widehat{v} = \arg\min_{v\in\mathcal{V}} \widehat{L}_\alpha(v, \widehat{w})$ and $v_\alpha^* \in \mathcal{V}$, we have

$$(2) \geq 0. \tag{100}$$

For term (3), since $\widehat{w} = \arg\max_{w\in\mathcal{W}} \widehat{L}_\alpha(\widehat{v}(w), w)$ and $w_\alpha^* \in \mathcal{W}$,

$$(3) \geq 0. \tag{101}$$

For term (5), note that due to the strong duality of the regularized problem (2)(3), (v_α^*, w_α^*) is a saddle point of $L_\alpha(v,w)$ over $\mathbb{R}^{|\mathcal{S}|} \times \mathbb{R}_+^{|\mathcal{S}||\mathcal{A}|}$. Therefore,

$$v_\alpha^* = \arg\min_{v\in\mathbb{R}^{|\mathcal{S}|}} L_\alpha(v, w_\alpha^*). \tag{102}$$

Since $\widehat{v}(w_\alpha^*) \in \mathbb{R}^{|\mathcal{S}|}$, we have:

$$(5) \geq 0. \tag{103}$$

Combining the above inequalities, it is obvious that

$$L_\alpha(v_\alpha^*, \widehat{w}) - L_\alpha(v_\alpha^*, w_\alpha^*) \geq -2\epsilon_{stat}. \tag{104}$$

F.6. Proof of Lemma 36

First we need to show $L_\alpha(v_\alpha^*, w)$ is αM_f-strongly-concave with respect to w and $\|\cdot\|_{2,d^D}$. Consider $\widetilde{L}_\alpha(w) = L_\alpha(v_\alpha^*, w) + \frac{\alpha M_f}{2}\|w\|_{2,d^D}^2$, then we know that

$$\widetilde{L}_\alpha(w) = (1-\gamma)\mathbb{E}_{s\sim\mu_0}[v(s)] - \alpha\mathbb{E}_{(s,a)\sim d^D}[f(w(s,a)) - \frac{M_f}{2}w(s,a)^2] + \mathbb{E}_{(s,a)\sim d^D}[w(s,a)e_v(s,a)]. \tag{105}$$

Since f is M_f-strongly-convex, we know $\widetilde{L}_\alpha(w)$ is concave, which implies that $L_\alpha(v_\alpha^*, w)$ is αM_f-strongly-concave with respect to w and $\|\cdot\|_{2,d^D}$.

On the other hand, since (v_α^*, w_α^*) is a saddle point of $L_\alpha(v, w)$ over $\mathbb{R}^{|\mathcal{S}|} \times \mathbb{R}_+^{|\mathcal{S}||\mathcal{A}|}$, we have $w_\alpha^* = \arg\max_{w\geq 0} L_\alpha(v_\alpha^*, w)$. Then we have:

$$\|\widehat{w} - w_\alpha^*\|_{2,d^D} \leq \sqrt{\frac{2(L_\alpha(v_\alpha^*, w_\alpha^*) - L_\alpha(v_\alpha^*, \widehat{w}))}{\alpha M_f}}. \tag{106}$$

Substituting Lemma 35 into the above equation we can obtain (59). For (60), it can be observed that

$$\|\widehat{d} - d_\alpha^*\|_1 = \|\widehat{w} - w_\alpha^*\|_{1,d^D} \leq \|\widehat{w} - w_\alpha^*\|_{2,d^D} \leq \sqrt{\frac{4\epsilon_{stat}}{\alpha M_f}}. \tag{107}$$

F.7. Proof of Lemma 37

First note that $\|\widehat{w} - w_\alpha^*\|_{1,d^D} \leq \|\widehat{w} - w_\alpha^*\|_{2,d^D}$, which implies that

$$\sum_s \epsilon_{\widehat{w},s} \leq \|\widehat{w} - w_\alpha^*\|_{2,d^D} \tag{108}$$

where

$$\epsilon_{\widehat{w},s} = \sum_a |\widehat{w}(s,a)d^D(s,a) - w_\alpha^* d^D(s,a)| \tag{109}$$

If $\widehat{d}(s) > 0$, then we have:

$$d_\alpha^*(s)\sum_a |\widehat{\pi}(s,a) - \pi_\alpha^*(s,a)| \tag{110}$$

$$= \sum_a |\frac{d_\alpha^*(s)}{\widehat{d}(s)}\widehat{w}(s,a)d^D(s,a) - w_\alpha^* d^D(s,a)| \tag{111}$$

$$\leq \sum_a (|\frac{d_\alpha^*(s)}{\widehat{d}(s)} - 1|\widehat{w}(s,a)d^D(s,a)) + \sum_a |\widehat{w}(s,a)d^D(s,a) - w_\alpha^* d^D(s,a)| \tag{112}$$

$$\leq \epsilon_{\widehat{w},s} + \sum_a (|\frac{d_\alpha^*(s)}{\widehat{d}(s)} - 1|\widehat{w}(s,a)d^D(s,a)). \tag{113}$$

Notice that $|\widehat{d}(s) - d_\alpha^*(s)| \leq \epsilon_{\widehat{w},s}$, which implies $|\frac{d_\alpha^*(s)}{\widehat{d}(s)} - 1| \leq \frac{\epsilon_{\widehat{w},s}}{\widehat{d}(s)}$, therefore:

$$d_\alpha^*(s)\sum_a |\widehat{\pi}(s,a) - \pi_\alpha^*(s,a)| \leq \epsilon_{\widehat{w},s}(1 + \sum_a \frac{\widehat{w}(s,a)d^D(s,a)}{\widehat{d}(s)}) = 2\epsilon_{\widehat{w},s}. \tag{114}$$

If $\widehat{d}(s) = 0$, then we know that $\sum_a |w^*_\alpha(s,a)d^D(s,a)| \leq \epsilon_{\widehat{w},s}$. Therefore

$$d^*_\alpha(s)\sum_a |\widehat{\pi}(s,a) - \pi^*_\alpha(s,a)| \leq 2d^*_\alpha(s) = 2\epsilon_{\widehat{w},s}. \tag{115}$$

Thus we have $d^*_\alpha(s)\sum_a |\widehat{\pi}(s,a) - \pi^*_\alpha(s,a)| \leq 2\epsilon_{\widehat{w},s}$, from which we can easily obtain:

$$\mathbb{E}_{s\sim d^*_\alpha}[\|\pi^*_\alpha(s,\cdot) - \widehat{\pi}(s,\cdot)\|_1] \leq 2\sum_s \epsilon_{\widehat{w},s} \leq 2\|\widehat{w} - w^*_\alpha\|_{2,d^D}. \tag{116}$$

F.8. Proof of Lemma 38

To bound $J(\pi^*_\alpha) - J(\widehat{\pi})$, we introduce the performance difference lemma which was previously derived in Kakade and Langford (2002); Kakade (2003):

Lemma 39 (Performance Difference) *For arbitrary policies π, π' and initial distribution μ_0, we have*

$$V^{\pi'}(\mu_0) - V^{\pi}(\mu_0) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi'}}[\langle Q^\pi(s,\dot{}), \pi'(\cdot|s) - \pi(\cdot|s)\rangle]. \tag{117}$$

The proof of Lemma 39 is referred to Appendix F.9. With Lemma 39, we have

$$\begin{aligned}
&J(\pi^*_\alpha) - J(\widehat{\pi}) && (118)\\
=&(1-\gamma)(V^{\pi^*_\alpha}(\mu_0) - V^{\widehat{\pi}}(\mu_0)) && (119)\\
=&\mathbb{E}_{s\sim d^*_\alpha}[\langle Q^{\widehat{\pi}}(s,\dot{}), \pi^*_\alpha(\cdot|s) - \widehat{\pi}(\cdot|s)\rangle] && (120)\\
\leq&\frac{1}{1-\gamma}\mathbb{E}_{s\sim d^*_\alpha}[\|\pi^*_\alpha(s,\cdot) - \widehat{\pi}(s,\cdot)\|_1]. && (121)
\end{aligned}$$

F.9. Proof of Lemma 39

For any two policies π' and π, it follows from the definition of $V^{\pi'}(\mu_0)$ that

$$\begin{aligned}
&V^{\pi'}(\mu_0) - V^{\pi}(\mu_0) && (122)\\
=&\mathbb{E}_{\pi'}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)\Big|\, s_0\sim\mu_0\right] - V^\pi(\mu_0)\\
=&\mathbb{E}_{\pi'}\left[\sum_{t=0}^{\infty}\gamma^t\Big[r(s_t,a_t) + V^\pi_\tau(s_t) - V^\pi(s_t)\Big]\,\Big|\, s_0\sim\mu_0\right] - V^\pi(\mu_0)\\
=&\mathbb{E}_{\pi'}\left[\sum_{t=0}^{\infty}\gamma^t\Big[r(s_t,a_t) + \gamma V^\pi(s_{t+1}) - V^\pi(s_t)\Big]\,\Big|\, s_0\sim\mu_0\right]\\
=&\mathbb{E}_{\pi'}\left[\sum_{t=0}^{\infty}\gamma^t\Big[r(s_t,a_t) + \gamma\mathbb{E}_{s_{t+1}\sim P(\cdot|s_t,a_t)}[V^\pi_\tau(s_{t+1})|s_t,a_t] - V^\pi_\tau(s_t)\Big]\,\Big|\, s_0\sim\mu_0\right]\\
=&\mathbb{E}_{\pi'}\left[\sum_{t=0}^{\infty}\gamma^t\Big[Q^\pi(s_t,a_t) - V^\pi(s_t)\Big]\,\Big|\, s_0\sim\mu_0\right]\\
=&\frac{1}{1-\gamma}\mathbb{E}_{(s,a)\sim d^{\pi'}}\left[Q^\pi(s,a) - V^\pi(s)\rangle\right]\\
=&\frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi'}}\left[\langle Q^\pi(s,\cdot), \pi'(\cdot|s) - \pi(\cdot|s)\rangle\right], && (123)
\end{aligned}$$

where the second to last step comes from the definition of $d^{\pi'}$ and the last step from the fact $V^{\pi}(s) = \mathbb{E}_{a\sim\pi(\cdot|s)}[Q^{\pi}(s,a)]$.

Appendix G. Proof of Corollary 5

The proof consists of two steps. We first show that $J(\pi_0^*) - J(\pi_{\alpha_\epsilon}^*) \leq \frac{\epsilon}{2}$ and then we bound $J(\pi_{\alpha_\epsilon}^*) - J(\widehat{\pi})$ by utilizing Theorem 3.

Step 1: Bounding $J(\pi_0^*) - J(\pi_{\alpha_\epsilon}^*)$. Notice that $\pi_{\alpha_\epsilon}^*$ is the solution to the regularized problem (2)(3), therefore we have:

$$\mathbb{E}_{(s,a)\sim d_{\alpha_\epsilon}^*}[r(s,a)] - \alpha\mathbb{E}_{(s,a)\sim d^D}[f(w_{\alpha_\epsilon}^*(s,a))] \geq \mathbb{E}_{(s,a)\sim d_0^*}[r(s,a)] - \alpha\mathbb{E}_{(s,a)\sim d^D}[f(w_0^*(s,a))], \tag{124}$$

which implies that

$$\begin{aligned} J(\pi_0^*) - J(\pi_{\alpha_\epsilon}^*) &= \mathbb{E}_{(s,a)\sim d_0^*}[r(s,a)] - \mathbb{E}_{(s,a)\sim d_{\alpha_\epsilon}^*}[r(s,a)] && (125)\\ &\leq \alpha\mathbb{E}_{(s,a)\sim d^D}[f(w_0^*(s,a))] - \alpha\mathbb{E}_{(s,a)\sim d^D}[f(w_{\alpha_\epsilon}^*(s,a))] && (126)\\ &\leq \alpha\mathbb{E}_{(s,a)\sim d^D}[f(w_0^*(s,a))] && (127)\\ &\leq \alpha B_f^0, && (128) \end{aligned}$$

where (127) comes from the non-negativity of f and (128) from the boundedness of f when $\alpha = 0$ (Assumption 4). Thus we have

$$J(\pi_0^*) - J(\pi_{\alpha_\epsilon}^*) \leq \frac{\epsilon}{2}. \tag{129}$$

Step 2: Bounding $J(\pi_{\alpha_\epsilon}^*) - J(\widehat{\pi})$. Using Theorem 3, we know that if

$$n \geq \frac{131072\left(\epsilon B_{f,\alpha_\epsilon} + 2B_{w,\alpha_\epsilon}B_{e,\alpha_\epsilon}B_{f,0}\right)^2}{\epsilon^6 M_f^2(1-\gamma)^4} \cdot \log\frac{4|\mathcal{V}||\mathcal{W}|}{\delta}, \tag{130}$$

$$n_0 \geq \frac{131072\left(2B_{v,\alpha_\epsilon}B_{f,0}\right)^2}{\epsilon^6 M_f^2(1-\gamma)^2} \cdot \log\frac{4|\mathcal{V}|}{\delta}, \tag{131}$$

then with at least probability $1-\delta$,

$$J(\pi_{\alpha_\epsilon}^*) - J(\widehat{\pi}) \leq \frac{\epsilon}{2}. \tag{132}$$

Using (129) and (132), we concludes that

$$J(\pi_0^*) - J(\widehat{\pi}) \leq \epsilon \tag{133}$$

hold with at least probability $1-\delta$. This finishes our proof.

Appendix H. Proof of Proposition 8

This proof largely follows Mangasarian and Meyer (1979). First note that the regularized problem (2)(3) has another more commonly used form of Lagrangian function:

$$\overline{L}_\alpha(\lambda,\eta,w) = (1-\gamma)\mathbb{E}_{s\sim\mu_0}[\lambda(s)] - \alpha\mathbb{E}_{(s,a)\sim d^D}[f(w(s,a))] + \mathbb{E}_{(s,a)\sim d^D}[w(s,a)e_\lambda(s)] - \eta^\top w, \tag{134}$$

where $\lambda \in \mathbb{R}^{|\mathcal{S}|}, \eta \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|} \geq 0, w \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$. Let $(\lambda_\alpha^*, \eta_\alpha^*) = \arg\min_{\eta \geq 0, \lambda \in \mathbb{R}^{|\mathcal{S}|}} \max_{w \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}} \overline{L}_\alpha(\lambda, \eta, w)$, then we have the following lemma:

Lemma 40

$$\lambda_\alpha^* = v_\alpha^*. \tag{135}$$

Proof The proof is referred to Appendix H.1. ■

Due to Lemma 40, we can only consider the primal optimum w_α^* and the dual optimum $(\lambda_\alpha^*, \eta_\alpha^*)$ of the Lagrangian function (134).

Let w^* be the solution to the following optimization problem:

$$\max_{w \in \mathcal{W}_0^*} -\alpha \mathbb{E}_{(s,a)\sim d^D}[f(w(s,a))] \tag{136}$$

Then since $w^* \in \mathcal{W}_0^*$, we know that $(w^*, \lambda_0^*, \eta_0^*)$ is the primal and dual optimum of the following constrained optimization problem, which is equivalent to the unregularized problem (1):

$$\max_w \sum_{s,a} [r(s,a) d^D(s,a) w(s,a)] \tag{137}$$

$$\text{s.t. } \sum_a d^D(s,a) w(s,a) = (1-\gamma)\mu_0(s) + \gamma \sum_{s',a'} P(s|s',a') d^D(s',a') w(s',a') \tag{138}$$

$$w(s,a) \geq 0, \forall s, a. \tag{139}$$

Let $p(s,a)$ denote $r(s,a)d^D(s,a)$ and $Aw = b$ denote the equality constraint (138), then we can obtain the following LP:

$$\min_w -p^\top w \tag{140}$$

$$\text{s.t. } Aw = b \tag{141}$$

$$w(s,a) \geq 0, \forall s, a. \tag{142}$$

By the KKT conditions of the above problem, we can obtain:

$$A^\top \lambda_0^* - p - \eta_0^* = 0, \tag{143}$$

$$Aw^* = b, w^* \geq 0, \tag{144}$$

$$\eta_0^* \geq 0, \tag{145}$$

$$\eta_0^*(s,a) w^*(s,a) = 0, \forall s, a. \tag{146}$$

$$\tag{147}$$

Let $c = -p^\top w^*$. Next we construct an auxiliary constrained optimization problem:

$$\min_w \mathbb{E}_{(s,a)\sim d^D}[f(w(s,a))] \tag{148}$$

$$\text{s.t. } Aw = b, \tag{149}$$

$$w(s,a) \geq 0, \forall s, a, \tag{150}$$

$$-p^\top w \leq c. \tag{151}$$

Then the corresponding Lagrangian function is

$$\mathbb{E}_{(s,a)\sim d^D}[f(w(s,a))] + \lambda_{aux}^\top (Aw - b) - \eta_{aux}^\top w + \xi_{aux}(-p^\top w - c). \tag{152}$$

Denote the primal and dual optimum of the auxiliary problem by $(w^*_{aux}, \lambda^*_{aux}, \eta^*_{aux}, \xi^*_{aux})$. Then obviously the constraints (149)(150)(151) are equivalent to $w \in \mathcal{W}_0^*$ and therefore $w^*_{aux} = w^*$, implying that $(w^*, \lambda^*_{aux}, \eta^*_{aux}, \xi^*_{aux})$ satisfies the following KKT conditions:

$$d^D \circ \nabla f(w^*) + A^\top \lambda^*_{aux} - \eta^*_{aux} - \xi^*_{aux} p = 0, \tag{153}$$
$$Aw^* = b, w^* \geq 0, -p^\top w^* = c, \tag{154}$$
$$\eta^*_{aux} \geq 0, \xi^*_{aux} \geq 0, \tag{155}$$
$$\eta^*_{aux}(s,a) w^*(s,a) = 0, \forall s, a, \tag{156}$$

where $d^D \circ \nabla f(w^*)$ denotes product by element.

Now we look at KKT conditions of (134):

$$A^\top \lambda^*_\alpha - p - \eta^*_\alpha + \alpha d^D \circ \nabla f(w^*_\alpha) = 0, \tag{157}$$
$$Aw^*_\alpha = b, w^*_\alpha \geq 0, \tag{158}$$
$$\eta^*_\alpha \geq 0, \tag{159}$$
$$\eta^*_\alpha(s,a) w^*_\alpha(s,a) = 0, \forall s, a. \tag{160}$$
$$\tag{161}$$

- **When $\xi^*_{\mathbf{aux}} = \mathbf{0}$.** It can be easily checked that $(w^*_\alpha = w^*, \lambda^*_\alpha = \lambda^*_0 + \alpha\lambda^*_{aux}, \eta^*_\alpha = \eta^*_0 + \alpha\eta^*_{aux})$ satisfies the KKT conditions of (134) for all $\alpha \geq 0$.
- **When $\xi^*_{\mathbf{aux}} > \mathbf{0}$.** It can be easily checked that $(w^*_\alpha = w^*, \lambda^*_\alpha = (1-\alpha\xi^*_{aux})\lambda^*_0 + \alpha\lambda^*_{aux}, \eta^*_\alpha = (1-\alpha\xi^*_{aux})\eta^*_0 + \alpha\eta^*_{aux})$ satisfies the KKT conditions of (134) for $\alpha \in [0, \overline{\alpha}]$ where $\overline{\alpha} = \frac{1}{\xi^*_{aux}}$.

Therefore, when $\alpha \in [0, \overline{\alpha}]$, $(w^*_\alpha = w^*, \lambda^*_\alpha = (1-\alpha\xi^*_{aux})\lambda^*_0 + \alpha\lambda^*_{aux}, \eta^*_\alpha = (1-\alpha\xi^*_{aux})\eta^*_0 + \alpha\eta^*_{aux})$ is the primal and dual optimum of (134). Then by Lemma 40, we know for $\alpha \in [0, \overline{\alpha}]$,

$$w^*_\alpha = w^* \in W_0^*, v^*_\alpha = (1 - \alpha\xi^*_{aux})\lambda^*_0 + \alpha\lambda^*_{aux}. \tag{162}$$

Let $\alpha = \overline{\alpha} = \frac{1}{\xi^*_{aux}}$, then since $\|w^*_{\overline{\alpha}}\|_\infty = \|w^*\|_\infty \leq B_w^0$, by Lemma 32 we have:

$$\|\overline{\alpha}\lambda^*_{aux}\|_\infty = \|v^*_{\overline{\alpha}}\|_\infty \leq \frac{\overline{\alpha} B_{f',0} + 1}{1-\gamma}, \tag{163}$$

which implies that

$$\|\lambda^*_{aux}\|_\infty \leq \frac{B_{f',0} + \xi^*_{aux}}{1-\gamma}. \tag{164}$$

Therefore, combining with $\|v^*_0\|_\infty \leq \frac{1}{1-\gamma}$, we have

$$\|v^*_\alpha - v^*_0\|_\infty \leq \alpha \cdot \frac{B_{f',0} + 2\xi^*_{aux}}{1-\gamma}, \forall \alpha \in [0, \overline{\alpha}] \tag{165}$$

which concludes our proof.

H.1. Proof of Lemma 40

From KKT conditions of $\overline{L}_\alpha(\lambda, \eta, w)$, we have

$$w^*_\alpha(s,a) = (f')^{-1}(\frac{e_{\lambda^*_\alpha}(s,a) + \eta^*_\alpha(s,a)}{\alpha}), \forall s, a, \tag{166}$$

$$w^*_\alpha \geq 0, \tag{167}$$

$$\sum_a w^*_\alpha(s,a) d^D(s,a) = (1-\gamma)\mu_0(s) + \gamma \sum_{s',a'} P(s|s',a') w^*_\alpha(s',a') d^D(s',a'), \forall s, \tag{168}$$

$$\eta^*_\alpha \geq 0, \tag{169}$$

$$\eta^*_\alpha(s,a) w^*_\alpha(s,a) = 0, \forall s, a. \tag{170}$$

Therefore, we can see that λ^*_α is the solution of the following equations:

$$e_{\lambda^*_\alpha}(s,a) = \alpha f'(w^*_\alpha(s,a)), \text{ for } s, a \text{ such that } w^*_\alpha(s,a) \neq 0, \tag{171}$$

$$e_{\lambda^*_\alpha}(s,a) \leq \alpha f'(0), \text{ for } s, a \text{ such that } w^*_\alpha(s,a) = 0. \tag{172}$$

Besides, from KKT conditions of $L_\alpha(v, w)$, we have

$$w^*_\alpha(s,a) = \max\{0, (f')^{-1}(\frac{e_{\lambda^*_\alpha}(s,a)}{\alpha})\}, \forall s, a, \tag{173}$$

$$w^*_\alpha \geq 0, \tag{174}$$

$$\sum_a w^*_\alpha(s,a) d^D(s,a) = (1-\gamma)\mu_0(s) + \gamma \sum_{s',a'} P(s|s',a') w^*_\alpha(s',a') d^D(s',a'), \forall s. \tag{175}$$

Therefore, v^*_α is the solution of the following equations:

$$e_{v^*_\alpha}(s,a) = \alpha f'(w^*_\alpha(s,a)), \text{ for } s, a \text{ such that } w^*_\alpha(s,a) \neq 0, \tag{176}$$

$$e_{v^*_\alpha}(s,a) \leq \alpha f'(0), \text{ for } s, a \text{ such that } w^*_\alpha(s,a) = 0. \tag{177}$$

It is observed that (171)(172) is the same as (176)(177), which implies that $\lambda^*_\alpha = v^*_\alpha$.

Appendix I. Proof of Lemmas in Theorem 10

I.1. Proof of Lemma 11

From KKT conditions of the maximin problem (19), we have

$$w^*_{\alpha,B_w}(s,a) = \min\left(\max\left(0, (f')^{-1}\left(\frac{e_{v^*_{\alpha,B_w}}(s,a)}{\alpha}\right)\right), B_w\right). \tag{178}$$

Suppose $|v^*_{\alpha,B_w}(s_m)| = \|v^*_{\alpha,B_w}\|_\infty$. Then we can consider the following two cases separately.

- **If there exists $a_{s_m} \in \mathcal{A}$ such that $0 < w^*_{\alpha,B_w}(s_m, a_{s_m}) < B_w$.**

  In this case, we know that

  $$|e_{v^*_{\alpha,B_w}}(s_m, a_{s_m})| = \alpha |f'(w^*_{\alpha,B_w}(s_m, a_{s_m}))| \leq \alpha B_{f'}. \tag{179}$$

  Then we can follow the arguments in Appendix F.2 to obtain:

  $$\|v^*_{\alpha,B_w}\|_\infty \leq \frac{\alpha B_{f'} + 1}{1-\gamma}. \tag{180}$$

- **If for all $a \in \mathcal{A}$, $w^*_{\alpha,B_w}(s_m,a) \in \{0, B_w\}$.** In this case, we first introduce the following lemma:

  Lemma 41 *If for all $a \in \mathcal{A}$, $w^*_{\alpha,B_w}(s_m,a) \in \{0,B_w\}$, then there exist $a_1, a_2 \in \mathcal{A}$ such that $w^*_{\alpha,B_w}(s_m,a_1) = 0, w^*_{\alpha,B_w}(s_m,a_2) = B_w$.*

  See Appendix I.2 for proof. With Lemma 41, we can bound $|v^*_{\alpha,B_w}(s_m)|$ as follows.

  If $v^*_{\alpha,B_w}(s_m) \geq 0$, then since $w^*_{\alpha,B_w}(s_m,a_2) = B_w$, we know $e_{v^*_{\alpha,B_w}}(s_m,a_2) \geq \alpha f'(B_w)$. Therefore we have:

$$\alpha f'(B_w) \leq e_{v^*_{\alpha,B_w}}(s_m,a_2) \leq r(s_m,a_2) - (1-\gamma)v^*_{\alpha,B_w}(s_m), \tag{181}$$

  which implies:

$$v^*_{\alpha,B_w}(s_m) \leq \frac{1}{1-\gamma}|r(s_m,a_2) + \alpha f'(B_w)| \leq \frac{\alpha B_{f'}+1}{1-\gamma}. \tag{182}$$

  If $v^*_{\alpha,B_w}(s_m) < 0$, then since $w^*_{\alpha,B_w}(s_m,a_1) = 0$, we know $e_{v^*_{\alpha,B_w}}(s_m,a_1) \leq \alpha f'(0)$. Therefore we have:

$$\alpha f'(0) \geq e_{v^*_{\alpha,B_w}}(s_m,a_1) \geq r(s_m,a_2) - (1-\gamma)v^*_{\alpha,B_w}(s_m), \tag{183}$$

  which implies:

$$v^*_{\alpha,B_w}(s_m) \geq -\frac{1}{1-\gamma}(|r(s_m,a_1)| + |\alpha f'(0)|) \geq -\frac{\alpha B_{f'}+1}{1-\gamma}. \tag{184}$$

  Combining (182) and (184), we have $\|v^*_{\alpha,B_w}\|_\infty = |v^*_{\alpha,B_w}(s_m)| \leq \frac{\alpha B_{f'}+1}{1-\gamma}$.

In conclusion, we have:

$$\|v^*_{\alpha,B_w}\|_\infty \leq \frac{\alpha B_{f'}+1}{1-\gamma}. \tag{185}$$

I.2. Proof of Lemma 41

First note that it is impossible to have $w^*_{\alpha,B_w}(s_m,a) = 0, \forall a$. This is because $d^*_{\alpha,B_w}(s_m,a) = w^*_{\alpha,B_w}(s_m,a)d^D(s_m,a)$ satisfies Bellman flow constraint (3). Therefore

$$d^*_{\alpha,B_w}(s_m) = \sum_a w^*_{\alpha,B_w}(s_m,a)d^D(s_m,a) \geq (1-\gamma)\mu_0(s_m) > 0. \tag{186}$$

On the other hand, if $w^*_{\alpha,B_w}(s_m,a) = B_w, \forall a$, then from Bellman flow constraints we have:

$$B_w d^D(s_m) = d^*_{\alpha,B_w}(s_m) = (1-\gamma)\mu_0(s_m) + \sum_{s',a'} P(s_m|s',a')w^*_{\alpha,B_w}(s',a')d^D(s',a'). \tag{187}$$

Notice that from Assumption 9 d^D is the discounted visitation distribution of π_D and thus also satisfies Bellman flow constraints:

$$d^D(s_m) = (1-\gamma)\mu_0(s_m) + \sum_{s',a'} P(s_m|s',a')d^D(s',a'), \tag{188}$$

which implies

$$B_w d^D(s_m) = (1-\gamma)B_w\mu_0(s_m) + \sum_{s',a'} B_w P(s_m|s',a')d^D(s',a'). \tag{189}$$

Combining (187) and (189), we have

$$(1-\gamma)(B_w-1)\mu_0(s_m) = \sum_{s',a'}(w^*_{\alpha,B_w} - B_w)P(s_m|s',a')d^D(s',a'). \tag{190}$$

However, since $B_w > 1, \mu_0(s_m) > 0, w^*_{\alpha,B_w} - B_w \leq 0$, we have:

$$(1-\gamma)(B_w-1)\mu_0(s_m) > 0, \sum_{s',a'}(w^*_{\alpha,B_w} - B_w)P(s_m|s',a')d^D(s',a') \leq 0, \tag{191}$$

which is a contradiction.

Therefore, there must exist $a_1, a_2 \in \mathcal{A}$ such that $w^*_{\alpha,B_w}(s_m,a_1) = 0, w^*_{\alpha,B_w}(s_m,a_2) = B_w$.

Appendix J. Proof of Corollary 12

First notice that

$$\mathbb{E}_{(s,a)\sim d^*_{\alpha'_\epsilon,B_w}}[r(s,a)] - \alpha'_\epsilon \mathbb{E}_{(s,a)\sim d^D}[f(w^*_{\alpha'_\epsilon,B_w}(s,a))] \geq \mathbb{E}_{(s,a)\sim d^*_{0,B_w}}[r(s,a)] - \alpha'_\epsilon \mathbb{E}_{(s,a)\sim d^D}[f(w^*_{0,B_w}(s,a))], \tag{192}$$

which implies that

$$J(\pi^*_{0,B_w}) - J(\pi^*_{\alpha'_\epsilon,B_w}) \leq \alpha'_\epsilon \left(\mathbb{E}_{(s,a)\sim d^D}[f(w^*_{0,B_w}(s,a))] - \mathbb{E}_{(s,a)\sim d^D}[f(w^*_{\alpha'_\epsilon,B_w}(s,a))]\right) \tag{193}$$

$$\leq 2\alpha'_\epsilon B_f = \frac{\epsilon}{2}. \tag{194}$$

On the other hand, by Theorem 10 we have with probability at least $1-\delta$,

$$\mathbb{E}_{s\sim d^*_{\alpha'_\epsilon,B_w}}[\|\pi^*_{\alpha'_\epsilon,B_w}(\cdot|s) - \widehat{\pi}(\cdot|s)\|_1] \leq \frac{(1-\gamma)\epsilon}{2}. \tag{195}$$

Using the performance difference lemma as in Appendix G, this implies

$$J(\pi^*_{\alpha'_\epsilon,B_w}) - J(\widehat{\pi}) \leq \frac{\epsilon}{2}. \tag{196}$$

Therefore, we have $J(\pi^*_{0,B_w}) - J(\widehat{\pi}) \leq \epsilon$ with at least probability $1-\delta$.

Appendix K. Proof of Theorem 16

Our proof follows a similar procedure of Theorem 3 and also consists of (1) bounding $|L_\alpha(v,w) - \widehat{L}_\alpha(v,w)|$, (2) characterizing the error $\|\widehat{w} - w^*_\alpha\|_{2,d^D}$ and (3) analyzing $\widehat{\pi}$ and π^*_α. The first and third step are exactly the same as Theorem 3 but the second step will be more complicated, on which we

will elaborate on in this section. We will use the following notations for brevity throughout the discussion:

$$v^*_{\alpha,\mathcal{V}} = \arg\min_{v\in\mathcal{V}} \|v - v^*_\alpha\|_{1,\mu_0} + \|v - v^*_\alpha\|_{1,d^D} + \|v - v^*_\alpha\|_{1,d^{D'}}, \tag{197}$$

$$w^*_{\alpha,\mathcal{W}} = \arg\min_{w\in\mathcal{W}} \|w - w^*_\alpha\|_{1,d^D}, \tag{198}$$

$$\widehat{v}(w) = \arg\min_{v\in\mathcal{V}} \widehat{L}_\alpha(v, w), \forall w. \tag{199}$$

We first need to characterize $L_\alpha(v^*_\alpha, w^*_\alpha) - L_\alpha(v^*_\alpha, \widehat{w})$. Similarly, we decompose $L_\alpha(v^*_\alpha, w^*_\alpha) - L_\alpha(v^*_\alpha, \widehat{w})$ into the following terms:

$$L_\alpha(v^*_\alpha, \widehat{w}) - L_\alpha(v^*_\alpha, w^*_\alpha) = \underbrace{(L_\alpha(v^*_\alpha, \widehat{w}) - L_\alpha(v^*_{\alpha,\mathcal{V}}, \widehat{w}))}_{(1)} + \underbrace{(L_\alpha(v^*_{\alpha,\mathcal{V}}, \widehat{w}) - \widehat{L}_\alpha(v^*_{\alpha,\mathcal{V}}, \widehat{w}))}_{(2)} \tag{200}$$

$$+ \underbrace{(\widehat{L}_\alpha(v^*_{\alpha,\mathcal{V}}, \widehat{w}) - \widehat{L}_\alpha(\widehat{v}, \widehat{w}))}_{(3)} + \underbrace{(\widehat{L}_\alpha(\widehat{v}, \widehat{w}) - \widehat{L}_\alpha(\widehat{v}(w^*_{\alpha,\mathcal{W}}), w^*_{\alpha,\mathcal{W}}))}_{(4)} \tag{201}$$

$$+ \underbrace{(\widehat{L}_\alpha(\widehat{v}(w^*_{\alpha,\mathcal{W}}), w^*_{\alpha,\mathcal{W}}) - L_\alpha(\widehat{v}(w^*_{\alpha,\mathcal{W}}), w^*_{\alpha,\mathcal{W}}))}_{(5)} + \underbrace{(L_\alpha(\widehat{v}(w^*_{\alpha,\mathcal{W}}), w^*_{\alpha,\mathcal{W}}) - L_\alpha(\widehat{v}(w^*_{\alpha,\mathcal{W}}), w^*_\alpha))}_{(6)}, \tag{202}$$

$$+ \underbrace{(L_\alpha(\widehat{v}(w^*_{\alpha,\mathcal{W}}), w^*_\alpha) - L_\alpha(v^*_\alpha, w^*_\alpha))}_{(7)}. \tag{203}$$

For term (2) and (5), we can apply Lemma 34 and thus

$$(2) \geq -\epsilon_{stat}, (5) \geq -\epsilon_{stat}. \tag{204}$$

For term (3), since $\widehat{L}_\alpha(\widehat{v}, \widehat{w}) - \min_{v\in\mathcal{V}} \widehat{L}_\alpha(v, \widehat{w}) \leq \epsilon_{o,v}$ and $v^*_{\alpha,\mathcal{V}} \in \mathcal{V}$, we have

$$(3) \geq -\epsilon_{o,v}. \tag{205}$$

For term (4), since $\max_{w\in\mathcal{W}} \min_{v\in\mathcal{V}} \widehat{L}_\alpha(v, w) - \min_{v\in\mathcal{V}} \widehat{L}_\alpha(v, \widehat{w}) \leq \epsilon_{o,w}$ and $w^*_{\alpha,\mathcal{W}} \in \mathcal{W}$,

$$\widehat{L}_\alpha(\widehat{v}, \widehat{w}) \geq \min_{v\in\mathcal{V}} \widehat{L}_\alpha(v, \widehat{w}) \geq \max_{w\in\mathcal{W}} \min_{v\in\mathcal{V}} \widehat{L}_\alpha(v, w) - \epsilon_{o,w} \geq \widehat{L}_\alpha(\widehat{v}(w^*_{\alpha,\mathcal{W}}), w^*_{\alpha,\mathcal{W}}) - \epsilon_{o,w}, \tag{206}$$

or

$$(4) \geq -\epsilon_{o,w}. \tag{207}$$

For term (7), since $v^*_\alpha = \arg\min_{v\in\mathbb{R}^{|\mathcal{S}|}} L_\alpha(v, w^*_\alpha)$, we have:

$$(7) \geq 0. \tag{208}$$

There are only term (1) and (6) left to be bounded, for which we introduce the following lemma on the continuity of $L_\alpha(v, w)$,

Lemma 42 *Suppose Assumption 3,4,5 hold. Then for any $v, v_1, v_2 \in \mathcal{V}$ and $w, w_1, w_2 \in \mathcal{W}$, we have:*

$$|L_\alpha(v_1, w) - L_\alpha(v_2, w)| \leq (B_{w,\alpha} + 1)\left(\|v_1 - v_2\|_{1,\mu_0} + \|v_1 - v_2\|_{1,d^D} + \|v_1 - v_2\|_{1,d^{D'}}\right), \tag{209}$$

$$|L_\alpha(v, w_1) - L_\alpha(v, w_2)| \leq (B_{e,\alpha} + \alpha B_{f',\alpha})\|w_1 - w_2\|_{1,d^D}. \tag{210}$$

The proof is in Section K.1. Using Lemma 42, we can bound term (1) and (6) easily:

$$(1) \geq -\left(B_{w,\alpha}+1\right)\epsilon_{\alpha,r,v}, (6)- \geq \left(B_{e,\alpha}+\alpha B_{f',\alpha}\right)\epsilon_{\alpha,r,w}. \tag{211}$$

Combining the above inequalities, it is obvious that

$$L_\alpha(v_\alpha^*, \widehat{w}) - L_\alpha(v_\alpha^*, w_\alpha^*) \geq -2\epsilon_{stat} - (\epsilon_{o,v}+\epsilon_{o,w}) - \left(\left(B_{w,\alpha}+1\right)\epsilon_{\alpha,r,v} + \left(B_{e,\alpha}+\alpha B_{f',\alpha}\right)\epsilon_{\alpha,r,w}\right). \tag{212}$$

Let $\epsilon_{\alpha,app}$ denote $\left(B_{w,\alpha}+1\right)\epsilon_{\alpha,r,v} + \left(B_{e,\alpha}+\alpha B_{f',\alpha}\right)\epsilon_{\alpha,r,w}$ and ϵ_{opt} denote $\epsilon_{o,v}+\epsilon_{o,w}$, then

$$L_\alpha(v_\alpha^*, \widehat{w}) - L_\alpha(v_\alpha^*, w_\alpha^*) \geq -2\epsilon_{stat} - \epsilon_{opt} - \epsilon_{\alpha,app}. \tag{213}$$

Further we utilize the strong convexity of f and Lemma 37, then we have:

$$\mathbb{E}_{s\sim d_\alpha^*}[\|\pi_\alpha^*(\cdot|s) - \widehat{\pi}(\cdot|s)\|_1] \leq 2\|\widehat{w} - w_\alpha^*\|_{2,d^D} \leq 4\sqrt{\frac{\epsilon_{stat}}{\alpha M_f}} + 2\sqrt{\frac{2(\epsilon_{opt}+\epsilon_{\alpha,app})}{\alpha M_f}}, \tag{214}$$

which completes the proof.

K.1. Proof of Lemma 42

First, by the definition of $L_\alpha(v,w)$ (6) we have

$$\begin{aligned}
&|L_\alpha(v_1,w) - L_\alpha(v_2,w)| && (215)\\
=&|(1-\gamma)\mathbb{E}_{s\sim\mu_0}[v_1(s)-v_2(s)] + \mathbb{E}_{(s,a)\sim d^D}[w(s,a)(e_{v_1}(s,a)-e_{v_2}(s,a))]| && (216)\\
\leq&(1-\gamma)\mathbb{E}_{s\sim\mu_0}[|v_1(s)-v_2(s)|] + \mathbb{E}_{(s,a)\sim d^D}[w(s,a)|e_{v_1}(s,a)-e_{v_2}(s,a)|] && (217)\\
=&(1-\gamma)\|v_1-v_2\|_{1,\mu_0} + \mathbb{E}_{(s,a)\sim d^D}[w(s,a)|e_{v_1}(s,a)-e_{v_2}(s,a)|]. && (218)
\end{aligned}$$

For $\mathbb{E}_{(s,a)\sim d^D}[w(s,a)|e_{v_1}(s,a)-e_{v_2}(s,a)|]$, notice that from Assumption 3,

$$\begin{aligned}
&\mathbb{E}_{(s,a)\sim d^D}[w(s,a)|e_{v_1}(s,a)-e_{v_2}(s,a)|] && (219)\\
\leq&B_{w,\alpha}\mathbb{E}_{(s,a)\sim d^D}\left[|\gamma\mathbb{E}_{s'\sim P(s'|s,a)}[v_1(s')-v_2(s')] + (v_2(s)-v_1(s))|\right] && (220)\\
\leq&B_{w,\alpha}\mathbb{E}_{(s,a)\sim d^D}\left[|\gamma\mathbb{E}_{s'\sim P(s'|s,a)}[v_1(s')-v_2(s')]|\right] + B_{w,\alpha}\mathbb{E}_{s\sim d^D}\left[|v_2(s)-v_1(s)|\right] && (221)\\
\leq&\gamma B_{w,\alpha}\mathbb{E}_{(s,a)\sim d^D, s'\sim P(s'|s,a)}[|v_1(s')-v_2(s')|] + B_{w,\alpha}\|v_2-v_1\|_{1,d^D} && (222)\\
\leq&B_{w,\alpha}\left(\|v_1-v_2\|_{1,d^D} + \|v_1-v_2\|_{1,d^{D'}}\right). && (223)
\end{aligned}$$

Thus we have

$$|L_\alpha(v_1,w) - L_\alpha(v_2,w)| \leq \left(B_{w,\alpha}+1\right)\left(\|v_1-v_2\|_{1,\mu_0} + \|v_1-v_2\|_{1,d^D} + \|v_1-v_2\|_{1,d^{D'}}\right). \tag{224}$$

Next we bound $|L_\alpha(v,w_1) - L_\alpha(v,w_2)|$:

$$\begin{aligned}
&|L_\alpha(v,w_1) - L_\alpha(v,w_2)| && (225)\\
=&|\alpha\mathbb{E}_{(s,a)\sim d^D}[f(w_2(s,a)) - f(w_1(s,a))] + \mathbb{E}_{(s,a)\sim d^D}[(w_1(s,a)-w_2(s,a))e_v(s,a)]| && (226)\\
\leq&\alpha\mathbb{E}_{(s,a)\sim d^D}[|f(w_1(s,a)) - f(w_2(s,a))|] + \mathbb{E}_{(s,a)\sim d^D}[|w_1(s,a)-w_2(s,a)|e_v(s,a)]. && (227)
\end{aligned}$$

For $\alpha\mathbb{E}_{(s,a)\sim d^D}[|f(w_1(s,a)) - f(w_2(s,a))|]$, from Assumption 4 we know

$$\begin{aligned}
&\alpha\mathbb{E}_{(s,a)\sim d^D}[|f(w_1(s,a)) - f(w_2(s,a))|] && (228)\\
\leq&\alpha B_{f',\alpha}\mathbb{E}_{(s,a)\sim d^D}[|w_1(s,a) - w_2(s,a)|] && (229)\\
=&\alpha B_{f',\alpha}\|w_1 - w_2\|_{1,d^D}. && (230)
\end{aligned}$$

For $\mathbb{E}_{(s,a)\sim d^D}[|w_1(s,a) - w_2(s,a)|e_v(s,a)]$, from Assumption 5 we know

$$\begin{aligned}
&\mathbb{E}_{(s,a)\sim d^D}[|w_1(s,a) - w_2(s,a)|e_v(s,a)] && (231)\\
\leq&B_{e,\alpha}\mathbb{E}_{(s,a)\sim d^D}[|w_1(s,a) - w_2(s,a)|] && (232)\\
=&B_{e,\alpha}\|w_1 - w_2\|_{1,d^D}. && (233)
\end{aligned}$$

Therefore we have

$$|L_\alpha(v, w_1) - L_\alpha(v, w_2)| \leq (B_{e,\alpha} + \alpha B_{f',\alpha})\|w_1 - w_2\|_{1,d^D}. \tag{234}$$

Appendix L. Proof of PRO-RL-BC

L.1. Proof of Lemma 19

Notice that by the variational form of total variation, we have for any policies π, π' and $s \in \mathcal{S}$,

$$\begin{aligned}
\|\pi(\cdot|s) - \pi'(\cdot|s)\|_1 &= \max_{h:\|h\|_\infty\leq 1}[\mathbb{E}_{a\sim\pi(\cdot|s)}h(a) - \mathbb{E}_{a\sim\pi'(\cdot|s)}h(a)] && (235)\\
&= \mathbb{E}_{a\sim\pi(\cdot|s)}[h^s_{\pi,\pi'}(a)] - \mathbb{E}_{a\sim\pi'(\cdot|s)}[h^s_{\pi,\pi'}(a)], && (236)
\end{aligned}$$

which implies that

$$\begin{aligned}
\mathbb{E}_{s\sim d}[\|\pi(\cdot|s) - \pi'(\cdot|s)\|_1] &= \mathbb{E}_{s\sim d}\left[\mathbb{E}_{a\sim\pi(\cdot|s)}[h^s_{\pi,\pi'}(a)] - \mathbb{E}_{a\sim\pi'(\cdot|s)}[h^s_{\pi,\pi'}(a)]\right] && (237)\\
&= \mathbb{E}_{s\sim d}\left[\mathbb{E}_{a\sim\pi(\cdot|s)}[h_{\pi,\pi'}(s,a)] - \mathbb{E}_{a\sim\pi'(\cdot|s)}[h_{\pi,\pi'}(s,a)]\right], && (238)
\end{aligned}$$

where the last step comes from the definition of $h_{\pi,\pi'}$.

L.2. Proof of Theorem 20

Let ϵ_{UO} denote $\left(\frac{4(\alpha B_{f,\alpha}+B_{w,\alpha}B_{e,\alpha})}{\alpha M_f}\right)^{\frac{1}{2}}\cdot\left(\frac{2\log\frac{8|\mathcal{V}||\mathcal{W}|}{\delta}}{n_1}\right)^{\frac{1}{4}} + \left(\frac{4(1-\gamma)B_{v,\alpha}}{\alpha M_f}\right)^{\frac{1}{2}}\cdot\left(\frac{2\log\frac{8|\mathcal{V}|}{\delta}}{n_0}\right)^{\frac{1}{4}}$. Suppose E denote the event

$$\|\widehat{w} - w^*_\alpha\|_{2,d^D} \leq \epsilon_{UO}, \tag{239}$$

then by Theorem 16, we have

$$\Pr(E) \geq 1 - \frac{\delta}{2}. \tag{240}$$

Our following discussion is all conditioned on E. Let $l'_{i,\pi,h}$ denote $\widehat{w}(s_i,a_i)(h^\pi(s_i) - h(s_i,a_i))$ then we know:

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}_2}[l'_{i,\pi,h}] &= \mathbb{E}_{(s,a)\sim d^D}[\widehat{w}(s,a)(h^\pi(s) - h(s,a))] && (241)\\
&= \left(\sum_{s,a} d^D(s,a)\widehat{w}(s,a)\right)\mathbb{E}_{s\sim\widehat{d}'}\left[\mathbb{E}_{a\sim\pi(\cdot|s)}[h(s,a)] - \mathbb{E}_{a\sim\widehat{\pi}(\cdot|s)}[h(s,a)]\right], && (242)
\end{aligned}$$

where $\widehat{d}'(s) = \frac{\sum_{a'} d^D(s,a')\widehat{w}(s,a')}{\sum_{s',a'} d^D(s',a')\widehat{w}(s',a')}$. Notice that $0 \le \widehat{w}(s,a) \le B_{w,\alpha}, |h(s,a)| \le 1$, then by Hoeffding's inequality we have for any $\pi \in \Pi$ and $h \in \mathcal{H}$, with at least probability $1 - \frac{\delta}{2}$,

$$\begin{aligned}
&\left| \frac{1}{n_2} \sum_{i=1}^{n_2} l'_{i,\pi,h} - \left(\sum_{s',a'} d^D(s',a')\widehat{w}(s',a') \right) \mathbb{E}_{s\sim\widehat{d}'} \left[\mathbb{E}_{a\sim\pi(\cdot|s)}[h(s,a)] - \mathbb{E}_{a\sim\widehat{\pi}(\cdot|s)}[h(s,a)] \right] \right| \\
\le & 2B_{w,\alpha} \sqrt{\frac{2\log\frac{4|\mathcal{H}||\Pi|}{\delta}}{n_2}} \le 2B_{w,\alpha} \sqrt{\frac{6\log\frac{4|\Pi|}{\delta}}{n_2}} := \epsilon_{stat,2}. \qquad (243)
\end{aligned}$$

Besides, the following lemma shows that $\widehat{d}'$ is close to d^*_α and $\left(\sum_{s',a'} d^D(s',a')\widehat{w}(s',a')\right)$ is close to 1 conditioned on E:

Lemma 43 *Conditioned on E, we have*

$$\|\widehat{d}' - d^*_\alpha\|_1 \le 2\epsilon_{UO}, \qquad (244)$$

$$\left| \left(\sum_{s',a'} d^D(s',a')\widehat{w}(s',a') \right) - 1 \right| \le \epsilon_{UO}. \qquad (245)$$

The proof of the above lemma is in Appendix L.3.

With concentration result (243) and Lemma 43, we can bound $\mathbb{E}_{s\sim d^*_\alpha}[\|\overline{\pi}(\cdot|s) - \pi^*_\alpha(\cdot|s)\|_1]$. To facilitate our discussion, we will use the following notations:

$$\overline{h} := h_{\overline{\pi},\pi^*_\alpha} \in \mathcal{H}, \qquad (246)$$

$$\overline{h}' := \arg\max_{h\in\mathcal{H}} \sum_{i=1}^{n_2} \widehat{w}(s_i,a_i)[h^{\overline{\pi}}(s_i) - h(s_i,a_i)], \qquad (247)$$

$$\widetilde{h} := \arg\max_{h\in\mathcal{H}} \sum_{i=1}^{n_2} \widehat{w}(s_i,a_i)[h^{\pi^*_\alpha}(s_i) - h(s_i,a_i)]. \qquad (248)$$

Then we have

$$\begin{aligned}
&\mathbb{E}_{s\sim d^*_\alpha}[\|\overline{\pi}(\cdot|s) - \pi^*_\alpha(\cdot|s)\|_1] && (249)\\
\le & \mathbb{E}_{s\sim\widehat{d}'}[\|\overline{\pi}(\cdot|s) - \pi^*_\alpha(\cdot|s)\|_1] + 4\epsilon_{UO} && (250)\\
= & \mathbb{E}_{s\sim\widehat{d}'}[\mathbb{E}_{a\sim\overline{\pi}(\cdot|s)}[\overline{h}(s,a)] - \mathbb{E}_{a\sim\pi^*_\alpha(\cdot|s)}[\overline{h}(s,a)]] + 4\epsilon_{UO} && (251)\\
= & \mathbb{E}_{s\sim\widehat{d}'}[\mathbb{E}_{a\sim\overline{\pi}(\cdot|s)}[\overline{h}(s,a)] - \mathbb{E}_{a\sim\widehat{\pi}(\cdot|s)}[\overline{h}(s,a)]] \\
& + \mathbb{E}_{s\sim\widehat{d}'}[\mathbb{E}_{a\sim\widehat{\pi}(\cdot|s)}[\overline{h}(s,a)] - \mathbb{E}_{a\sim\pi^*_\alpha(\cdot|s)}[\overline{h}(s,a)]] + 4\epsilon_{UO} && (252)\\
= & \mathbb{E}_{s\sim\widehat{d}',a\sim\widehat{\pi}(\cdot|s)}[\overline{h}^{\overline{\pi}}(s) - \overline{h}(s,a)] + \mathbb{E}_{s\sim\widehat{d}',a\sim\widehat{\pi}(\cdot|s)}[(-\overline{h}^{\pi^*_\alpha}(s)) - (-\overline{h}(s,a))] + 4\epsilon_{UO} && (253)\\
\le & \mathbb{E}_{s\sim\widehat{d}',a\sim\widehat{\pi}(\cdot|s)}[\overline{h}^{\overline{\pi}}(s) - \overline{h}(s,a)] + \mathbb{E}_{s\sim\widehat{d}'}[\|\pi^*_\alpha(\cdot|s) - \widehat{\pi}(\cdot|s)\|_1] + 4\epsilon_{UO} && (254)\\
\le & \mathbb{E}_{s\sim\widehat{d}',a\sim\widehat{\pi}(\cdot|s)}[\overline{h}^{\overline{\pi}}(s) - \overline{h}(s,a)] + \mathbb{E}_{s\sim d^*_\alpha}[\|\pi^*_\alpha(\cdot|s) - \widehat{\pi}(\cdot|s)\|_1] + 8\epsilon_{UO} && (255)\\
\le & \mathbb{E}_{s\sim\widehat{d}',a\sim\widehat{\pi}(\cdot|s)}[\overline{h}^{\overline{\pi}}(s) - \overline{h}(s,a)] + 10\epsilon_{UO}, && (256)
\end{aligned}$$

where the first and sixth steps come from (244), the fifth step is due to $\|\overline{h}\|_\infty \leq 1$ and the last step from Theorem 16.

For $\mathbb{E}_{s\sim\widehat{d}',a\sim\widehat{\pi}(\cdot|s)}[\overline{h}^{\overline{\pi}}(s) - \overline{h}(s,a)]$, we utilize concentration result (243) and have with at least probability $1-\delta$:

$$\mathbb{E}_{s\sim\widehat{d}',a\sim\widehat{\pi}(\cdot|s)}[\overline{h}^{\overline{\pi}}(s) - \overline{h}(s,a)] \tag{257}$$

$$\leq \left(\sum_{s',a'} d^D(s',a')\widehat{w}(s',a')\right) \mathbb{E}_{s\sim\widehat{d}'}\left[\mathbb{E}_{a\sim\overline{\pi}(\cdot|s)}[\overline{h}(s,a)] - \mathbb{E}_{a\sim\widehat{\pi}(\cdot|s)}[\overline{h}(s,a)]\right] + 2\epsilon_{UO} \tag{258}$$

$$\leq \frac{1}{n_2}\sum_{i=1}^{n_2}[\widehat{w}(s_i,a_i)(\overline{h}^{\overline{\pi}}(s_i) - \overline{h}(s_i,a_i))] + \epsilon_{stat,2} + 2\epsilon_{UO} \tag{259}$$

$$\leq \frac{1}{n_2}\sum_{i=1}^{n_2}[\widehat{w}(s_i,a_i)(\overline{h}'^{\overline{\pi}}(s_i) - \overline{h}'(s_i,a_i))] + \epsilon_{stat,2} + 2\epsilon_{UO} \tag{260}$$

$$\leq \frac{1}{n_2}\sum_{i=1}^{n_2}[\widehat{w}(s_i,a_i)(\widetilde{h}^{\pi^*_\alpha}(s_i) - \widetilde{h}(s_i,a_i))] + \epsilon_{stat,2} + 2\epsilon_{UO} \tag{261}$$

$$\leq \mathbb{E}_{s\sim\widehat{d}',a\sim\widehat{\pi}(\cdot|s)}[\widetilde{h}^{\pi^*_\alpha}(s) - \widetilde{h}(s,a)] + 2\epsilon_{stat,2} + 4\epsilon_{UO} \tag{262}$$

$$\leq \mathbb{E}_{s\sim\widehat{d}'}[\|\pi^*_\alpha(\cdot|s) - \widehat{\pi}(\cdot|s)\|_1] + 2\epsilon_{stat,2} + 4\epsilon_{UO} \tag{263}$$

$$\leq 2\epsilon_{stat,2} + 10\epsilon_{UO}, \tag{264}$$

where the first step comes from (245), the second is due to (243), the third and fourth is from the definition of $\overline{h}'$ and $\overline{\pi}$, the fifth step utilizes (245) and (243), the sixth step is due to $\|\widetilde{h}\|_\infty \leq 1$ and the last step is from (244) and Theorem 16.

Combining (256) and (264), we have conditioned on E, with at least probability $1-\frac{\delta}{2}$, we have

$$\mathbb{E}_{s\sim d^*_\alpha}[\|\overline{\pi}(\cdot|s) - \pi^*_\alpha(\cdot|s)\|_1] \leq 2\epsilon_{stat,2} + 20\epsilon_{UO}. \tag{265}$$

Notice that $\epsilon_{UO} \leq 2^{\frac{5}{4}}\sqrt{\frac{\mathcal{E}_{n_1,n_0,\alpha}(B_{w,\alpha},B_{f,\alpha},B_{v,\alpha},B_{e,\alpha})}{\alpha M_f}}$. Therefore, with at least probability $1-\delta$, we have:

$$\mathbb{E}_{s\sim d^*_\alpha}[\|\pi^*_\alpha(\cdot|s) - \overline{\pi}(\cdot|s)\|_1] \leq 4B_{w,\alpha}\sqrt{\frac{6\log\frac{4|\Pi|}{\delta}}{n_2}} + 50\sqrt{\frac{\mathcal{E}_{n_1,n_0,\alpha}(B_{w,\alpha},B_{f,\alpha},B_{v,\alpha},B_{e,\alpha})}{\alpha M_f}} \tag{266}$$

This finishes our proof.

L.3. Proof of Lemma 43

The proof is similar to Lemma 37. First notice that

$$\left|\left(\sum_{s',a'} d^D(s',a')\widehat{w}(s',a')\right)-1\right| \tag{267}$$

$$=\left|\left(\sum_{s',a'} d^D(s',a')\widehat{w}(s',a')\right)-\left(\sum_{s',a'} d^D(s',a')w^*_\alpha(s',a')\right)\right| \tag{268}$$

$$=\left|\sum_{s',a'} d^D(s',a')\left(\widehat{w}(s',a')-w^*_\alpha(s',a')\right)\right| \tag{269}$$

$$\leq \sum_{s',a'} d^D(s',a')|\widehat{w}(s',a')-w^*_\alpha(s',a')| \tag{270}$$

$$\leq \|\widehat{w}-w^*_\alpha\|_{2,d^D} \tag{271}$$

$$\leq \epsilon_{UO}, \tag{272}$$

which proves the second part of the lemma. For the first part, we have

$$\|\widehat{d}-d^*_\alpha\|_1 \tag{273}$$

$$=\sum_s\left|\frac{1}{\sum_{s',a'} d^D(s',a')\widehat{w}(s',a')}\sum_{a'} d^D(s,a')\widehat{w}(s,a')-d^*_\alpha(s)\right| \tag{274}$$

$$\leq \sum_s\left(\left|\frac{1}{\sum_{s',a'} d^D(s',a')\widehat{w}(s',a')}-1\right|\sum_{a'} d^D(s,a')\widehat{w}(s,a')\right)$$
$$+\sum_s\left|\sum_{a'} d^D(s,a')\widehat{w}(s,a')-d^*_\alpha(s)\right| \tag{275}$$

$$=\underbrace{\sum_s\left(\left|\frac{1}{\sum_{s',a'} d^D(s',a')\widehat{w}(s',a')}-1\right|\sum_{a'} d^D(s,a')\widehat{w}(s,a')\right)}_{(1)}$$
$$+\underbrace{\sum_s\left|\sum_{a'} d^D(s,a')\widehat{w}(s,a')-\sum_{a'} d^D(s,a')w^*_\alpha(s,a')\right|}_{(2)}. \tag{276}$$

For term (1), notice that

$$\left|\frac{1}{\sum_{s',a'} d^D(s',a')\widehat{w}(s',a')}-1\right|=\frac{\left|1-\sum_{s',a'} d^D(s',a')\widehat{w}(s',a')\right|}{\sum_{s',a'} d^D(s',a')\widehat{w}(s',a')}\leq\frac{\epsilon_{UO}}{\sum_{s',a'} d^D(s',a')\widehat{w}(s',a')}. \tag{277}$$

Therefore,

$$\sum_s \left(\left| \frac{1}{\sum_{s',a'} d^D(s',a')\widehat{w}(s',a')} - 1 \right| \sum_{a'} d^D(s,a')\widehat{w}(s,a') \right) \tag{278}$$

$$\leq \epsilon_{UO} \sum_s \frac{\sum_{a'} d^D(s,a')\widehat{w}(s,a')}{\sum_{s',a'} d^D(s',a')\widehat{w}(s',a')} \tag{279}$$

$$= \epsilon_{UO}. \tag{280}$$

For term (2),

$$\sum_s \left| \sum_{a'} d^D(s,a')\widehat{w}(s,a') - \sum_{a'} d^D(s,a') w^*_\alpha(s,a') \right| \tag{281}$$

$$\leq \sum_{s,a'} d^D(s,a') |\widehat{w}(s,a') - w^*_\alpha(s,a')| \tag{282}$$

$$\leq \epsilon_{UO}. \tag{283}$$

Thus we have

$$\|\widehat{d}' - d^*_\alpha\|_1 \leq 2\epsilon_{UO}. \tag{284}$$

Appendix M. Proof of Corollary 28

First by Lemma 35, we know that

$$L_0(v^*_0, w^*_0) - L_0(v^*_0, \widehat{w}) \leq \frac{2B_{w,0}}{1-\gamma} \sqrt{\frac{2\log \frac{4|\mathcal{V}||\mathcal{W}|}{\delta}}{n}} + \sqrt{\frac{2\log \frac{4|\mathcal{V}|}{\delta}}{n_0}}. \tag{285}$$

Substitute the definition (6) of $L_0(v^*_0, w) = (1-\gamma)\mathbb{E}_{s\sim\mu_0}[v^*_0(s)] + \mathbb{E}_{(s,a)\sim d^D}[w(s,a)e_{v^*_0(s)}(s,a)]$ into the above inequality, we have

$$\sum_{s,a} \left(d^*_0(s,a) e_{v^*_0(s)}(s,a) \right) - \sum_{s,a} \left(\widehat{d}(s,a) e_{v^*_0(s)}(s,a) \right) \leq \frac{2B_{w,0}}{1-\gamma} \sqrt{\frac{2\log \frac{4|\mathcal{V}||\mathcal{W}|}{\delta}}{n}} + \sqrt{\frac{2\log \frac{4|\mathcal{V}|}{\delta}}{n_0}}. \tag{286}$$

Note that v^*_0 is the optimal value function of the unregularized MDP $\mathcal{M}$ and d^*_0 is the discounted state visitation distribution of the optimal policy π^*_0 (Puterman, 1994). Therefore, invoking Lemma 39, we have

$$\begin{aligned} J(\pi) - J(\pi^*_0) &= \mathbb{E}_{(s,a)\sim d^\pi}[r(s,a) + \gamma \mathbb{E}_{s'\sim P(\cdot|s,a)} v^*_0(s') - v^*_0(s)] \\ &= \sum_{s,a} d^\pi(s,a) e_{v^*_0(s)}(s,a). \end{aligned} \tag{287}$$

Let $\pi = \widetilde{\pi}^*_0$ in (287), then we can obtain

$$\sum_{s,a} d^*_0(s,a) e_{v^*_0(s)}(s,a) = 0. \tag{288}$$

Substitute it into (286),

$$\sum_{s,a}\left(\widehat{d}(s,a)(-e_{v_0^*(s)}(s,a))\right) \leq \frac{2B_{w,0}}{1-\gamma}\sqrt{\frac{2\log\frac{4|\mathcal{V}||\mathcal{W}|}{\delta}}{n}} + \sqrt{\frac{2\log\frac{4|\mathcal{V}|}{\delta}}{n_0}}. \tag{289}$$

Notice that since v_0^* is the optimal value function, $-e_{v_0^*(s)}(s,a) \geq 0$ for all s, a. Therefore, we have:

$$\begin{aligned}
J(\pi_0^*) - J(\widehat{\pi}) &= \sum_{s,a} d^{\widehat{\pi}}(s,a)(-e_{v_0^*(s)}(s,a)) && (290)\\
&= \sum_{s,a} d^{\widehat{\pi}}(s)\widehat{\pi}(a|s)(-e_{v_0^*(s)}(s,a)) && (291)\\
&\leq B_{w,u}\sum_{s,a} d^D(s)\widehat{\pi}(a|s)(-e_{v_0^*(s)}(s,a)) && (292)\\
&= B_{w,u}\sum_{s,a} d^D(s)\frac{\widehat{w}(s,a)\pi_D(a|s)}{\sum_{a'}\widehat{w}(s,a')\pi_D(a'|s)}(-e_{v_0^*(s)}(s,a)) && (293)\\
&\leq \frac{B_{w,u}}{B_{w,l}}\sum_{s,a} d^D(s)\pi_D(a|s)\widehat{w}(s,a)(-e_{v_0^*(s)}(s,a)) && (294)\\
&= \frac{B_{w,u}}{B_{w,l}}\sum_{s,a} \widehat{d}(s,a)(-e_{v_0^*(s)}(s,a)) && (295)\\
&\leq \frac{2B_{w,0}B_{w,u}}{(1-\gamma)B_{w,l}}\sqrt{\frac{2\log\frac{4|\mathcal{V}||\mathcal{W}|}{\delta}}{n}} + \frac{B_{w,u}}{B_{w,l}}\sqrt{\frac{2\log\frac{4|\mathcal{V}|}{\delta}}{n_0}}, && (296)
\end{aligned}$$

where the first step comes from (287), the third step is due to Assumption 12, the fifth step comes from Assumption 13 and the last step comes from (289). This concludes our proof.

M.1. Proof of Lemma 30

First notice that $d^D(s) \geq (1-\gamma)\mu_0(s)$. Then since $d^\pi(s) \leq B_{erg,2}\mu_0(s), \forall s, \pi$, we have for any policy π:

$$\frac{d^\pi(s)}{d^D(s)} \leq \frac{1}{1-\gamma}\frac{d^\pi(s)}{\mu_0(s)} \leq \frac{B_{erg,2}}{1-\gamma}. \tag{297}$$

On the other hand, $d_0^*(s) \geq (1-\gamma)\mu_0(s)$, therefore similarly we have:

$$\frac{d_0^*(s)}{d^D(s)} \geq \frac{(1-\gamma)\mu_0(s)}{d^D(s)} \geq \frac{1-\gamma}{B_{erg,2}}. \tag{298}$$