AUGMENTED LAGRANGIAN IS ENOUGH FOR OPTIMAL OFFLINE RL WITH GENERAL FUNCTION APPROXIMATION AND PARTIAL COVERAGE

Anonymous authors
Paper under double-blind review

ABSTRACT

Offline reinforcement learning (RL), which refers to decision-making from a previously-collected dataset of interactions, has received significant attention over the past years. Much effort has focused on improving offline RL practicality by addressing the prevalent issue of partial data coverage through various forms of conservative policy learning. While the majority of algorithms do not have finite-sample guarantees, several provable conservative offline RL algorithms are designed and analyzed within the single-policy concentrability framework that handles partial coverage. Yet, in the nonlinear function approximation setting where confidence intervals are difficult to obtain, existing provable algorithms suffer from computational intractability, prohibitively strong assumptions, and suboptimal statistical rates. In this paper, we leverage the marginalized importance sampling (MIS) formulation of RL and present the first set of offline RL algorithms that are statistically optimal and practical under general function approximation and single-policy concentrability, bypassing the need for uncertainty quantification. We identify that the key for successfully solving the sample-based approximation of the MIS problem is ensuring that certain *state occupancy validity constraints* are nearly satisfied. We enforce these constraints by a novel application of the augmented Lagrangian method and prove the following result: with MIS formulation, augmented Lagrangian is enough for statistically optimal offline RL. In stark contrast to prior algorithms that induce additional conservatism through methods such as behavior regularization, our approach provably eliminates this need and reinterprets regularizers as "enforcers of state occupancy validity" than "promoters of conservatism."

1 INTRODUCTION

The goal of offline RL is to design agents that learn to achieve competence in a task using only a previously-collected dataset of interactions (Lange et al., 2012). Offline RL is a promising tool for many critical applications, from healthcare to autonomous driving to scientific discovery, where the online mode of learning by interacting with the environment is dangerous, impractical, costly, or even impossible (Levine et al., 2020). Despite this, offline RL has not yet been truly successful in practice (Fujimoto et al., 2019; Levine et al., 2020) and impressive RL performance has been limited to settings with known environments (Silver et al., 2017; Moravčík et al., 2017), access to accurate simulators (Mnih et al., 2015; Degrave et al., 2022), or expert demonstrations (Vinyals et al., 2017).

One of the central challenges in offline RL is the lack of uniform coverage in real datasets and the *distribution shift* between the occupancy of candidate policies and offline data distribution, which pose difficulties in accurately evaluating the candidate policies. Over the past years, a body of literature has focused on addressing this challenge through developing conservative algorithms, which aim at picking a policy among those well-covered in the data. On the practical front, various forms of conservatism are proposed such as behavior regularization through policy constraints (Kumar et al., 2019; Fujimoto et al., 2019; Nachum & Dai, 2020), learning conservative values (Kumar et al., 2020; Liu et al., 2020; Agarwal et al., 2020), or learning pessimistic models (Kidambi et al., 2020; Yu et al., 2020; 2021); see Appendix A for further discussion on related work.

From a theoretical standpoint, partial data coverage has recently been studied within variants of the single-policy concentrability framework (Rashidinejad et al., 2021; Xie et al., 2021; Uehara & Sun, 2021), which characterizes the distribution shift between offline data and occupancy of a target (often optimal) policy, in contrast to all-policy concentrability commonly used in earlier works (Scherrer, 2014; Chen & Jiang, 2019; Liao et al., 2020; Zhang et al., 2020a; Xie & Jiang, 2021). Within this framework and in the tabular and linear function approximation settings, pessimistic algorithms that leverage uncertainty quantifiers to construct lower confidence bounds (Jin et al., 2021; Rashidinejad et al., 2021; Yin et al., 2021; Shi et al., 2022; Li et al., 2022) enjoy optimal statistical rate. In the general function approximation setting, pessimistic algorithms largely assume oracle access to uncertainty quantification, either for constructing penalties that are subtracted from rewards (Jin et al., 2021; Jiang & Huang, 2020) or selecting the most pessimistic option among those that fall within the confidence region implied by the offline data (Uehara & Sun, 2021; Xie et al., 2021; Chen & Jiang, 2022). However, uncertainty quantifiers are difficult to obtain in non-linear function approximation and existing heuristics are empirically observed to be unreliable (Rashid et al., 2019; Tennenholtz et al., 2021). Recent works by Cheng et al. (2022) and Zhan et al. (2022) propose provable alternatives to uncertainty-based methods, but leave achieving optimal statistical rate of $1/\sqrt{N}$, where N is the dataset size, as an open problem.

Among all, the marginal importance sampling (MIS) methods, which aim at learning weights w that estimate the distribution shift between induced policy occupancy d_w and data distribution μ, lend themselves well to the single-policy concentrability framework. Though more popular in off-policy evaluation (Liu et al., 2018; Xie et al., 2019; Uehara et al., 2020; Zhang et al., 2020b), MIS has also been used for conservative offline RL such as AlgaeDICE (Nachum et al., 2019b) and OptiDICE (Lee et al., 2021), both of which incorporate behavior regularization. Recently, Zhan et al. (2022) theoretically studied a variant of OptiDICE, showing that MIS with behavior regularization enjoys finite-sample guarantees (though achieving a suboptimal $1/N^{1/6}$ rate) and circumvents certain fundamental difficulties observed in value-based offline RL with function approximation (Du et al., 2019; Wang et al., 2020; 2021; Weisz et al., 2021; Zanette, 2021; Foster et al., 2021).

1.1 Contributions and results

Motivated by the benefits offered by MIS, we study designing statistically optimal offline learning algorithms under this formulation in the general function approximation and single-policy concentrability setting. We conduct theoretical investigations and design algorithms starting from multi-armed bandits (MABs), going forward to contextual bandits (CBs), and finally Markov decision processes (MDPs). In the rest of this section, we present a preview of our contributions and results.

Multi-armed bandits. Empirical MIS algorithms often incorporate behavior regularization, whose role is justified as promoting conservatism by keeping the occupancies of learned and behavior policies close (Nachum et al., 2019b; Lee et al., 2021). Yet, whether and why these regularizers are necessary from a theoretical perspective remain unclear. Zhan et al. (2022) motivates behavior regularization as a way of introducing curvature in an otherwise linear optimization problem. We extensively investigate the effect of regularization, starting from the simplest setting of MABs with function approximation, as existing algorithms when specialized to offline MABs, are either intractable, have suboptimal finite-sample guarantees, or require access to uncertainty quantifiers.

We state our results on unregularized MIS and MIS with behavior regularization (PRO-MAB Algorithm 1), which is a special case of PRO-RL (Zhan et al., 2022), in the informal theorem below.

Theorem (informal) *(I) Unregularized MIS fails to achieve a decaying suboptimality in certain offline MAB instances. (II) MIS with behavior regularization (PRO-MAB Algorithm 1) achieves suboptimality $\widetilde{O}(1/\sqrt{N})$ in offline MABs. (III) If one searches only over the space of weights that induce valid occupancies $(d_w = 1)$, then unregularized MIS achieves $O(1/\sqrt{N})$ suboptimality.*

In this theorem, we prove that unregularized MIS fails even in bandits and provide a tight analysis of PRO-MAB improving over the original $1/N^{1/6}$ rate shown by Zhan et al. (2022). In our analysis of PRO-MAB, we find that the key to the success of the MIS algorithm is *near-validity of the learned occupancy* d_w. In the MAB setting, the validity constraint simply requires the learned occupancy to be a probability distribution: $d_w = \sum_a w(a)\mu(a) = 1$. With a proper choice of hyperparameter, we show that behavior regularization enforces learned occupancy to be nearly valid: $d_w = \Omega(1/\log N)$. We further prove that regularization is not required if validity is otherwise satisfied.

Given that occupancy validity is the constraint of the optimization problem solved by MIS (see e.g. (1)), we ask whether there are any methods for solving empirical optimization problems that find solutions that adhere more to the constraints, compared to the Lagrange multiplier adopted in prior works (Lee et al., 2021; Zhan et al., 2022). The augmented Lagrangian method (ALM), which adds a quadratic loss on the constraints $(d_w - 1)^2$, is a natural choice for our purpose. The ALM term can be easily estimated from offline data and forms Algorithm 1. We show that ALM results in $d_w = \Omega(1)$, ensuring near-validity of estimated occupancy and leading to the following guarantee.

Theorem (informal) *Th policy returned by conservative offline MAB with augmented Lagrangian (Algorithm 1) achieves $O(1/\sqrt{N})$ suboptimality.*

Our algorithm offers several benefits over PRO-MAB such as improving the rate by $\log N$ and only requiring single-policy concentrability instead of the two-policy requirement of PRO-MAB, which can be strong (see Section 5.3). Additionally, behavior regularization introduces bias in the solution even with infinite data (Chen & Jiang, 2022) and the bias-variance tradeoff must be carefully handled. However, ALM merely enforces the optimization constraints and leads to provably unbiased solutions (Lemma 13). More importantly, as we discuss shortly, going beyond the single-state MAB setting, behavior regularization is suboptimal while ALM maintains the optimal rate.

Contextual bandits. In offline CBs, we analyze two approaches: MIS with behavior regularization, and an extension of ALM. We state our results in the following informal theorem.

Theorem (informal) *(I) MIS with behavior regularization (PRO-CB Algorithm 6) is statistically suboptimal for certain CB instances. (II) Policy returned by conservative offline CB with augmented Lagrangian (Algorithm 2) achieves suboptimality of $O(1/\sqrt{N})$.*

Intuitively, the failure of PRO-CB to achieve the optimal rate is because the regularization parameter has to be small to control bias, but such small regularization is not strong enough to ensure the validity of learned occupancy in most states. Therefore, one must choose larger regularization, leading to an overall suboptimal rate. Prior works Chen & Jiang (2022); Cheng et al. (2022) also allude to this phenomenon, explaining that regularizers appear to be the culprit behind suboptimal rates. In CB, the occupancy validity constraints require conditional occupancy to be a valid probability distribution in every state. In Algorithm 2, we incorporate ALM in offline CBs by adding a weighted sum of quadratic losses describing the validity constraint in each state, where the weights are set to the state occupancies to capture their relative importance. Enforcement of the constraints by ALM yields the guarantee stated above.

MDPs. Validity constraints in MDPs ensure that the learned state occupancy $d_w(s) = \sum_a w(s,a)\mu(s,a)$ is close to the actual state occupancy $d^{\pi_w}(s)$, where π_w is the policy computed from weights w.[1] Directly enforcing this constraint results in an ALM term that cannot easily be estimated from offline data. We address this difficulty by expressing the ALM term in the variational form. From there, we derive two variants, one model-based and one model-free, of our conservative offline RL with augmented Lagrangian (CORAL) algorithm, that enjoys the following guarantee.

Theorem (informal) *Model-based and model-free CORAL both achieve $O(1/\sqrt{N})$ suboptimality.*

This marks CORAL as the first practical and statistically optimal offline RL algorithm that operates in the general function approximation and partial data coverage setting, while avoiding uncertainty quantification and additional regularizers. Conservatism of CORAL is baked into the MIS formulation and supported by the ALM: bounded MIS weights prevent learned occupancy to deviate significantly from data distribution, and ALM ensures closeness of the learned and actual occupancies. When combined, CORAL learns a policy whose actual occupancy is close to the data distribution.

We thus proved that ALM improves sample complexity compared to alternatives such as behavior regularization. This is in addition to the benefits on optimization stability that are likely to be offered by the ALM, as it improves over the ill-posed Lagrange multiplier objective (Ben-Tal & Nemirovski, 2022). Our theoretical findings can explain the empirical observations of Yang et al. (2020), who find MIS with behavior regularization unstable, propose regularizers in "the spirit of ALM" that gain superior performance, and attribute performance gain to improved optimization. In this work, we present a theoretically-founded way of introducing ALM in offline RL and our analysis shows that ALM also leads to improved sample complexity.

[1] Notice that the validity constraints in MAB and CB are special cases of this constraint.

2 BACKGROUND

Markov decision process. An infinite-horizon discounted MDP is described by a tuple $M = (\mathcal{S}, \mathcal{A}, P, R, \rho, \gamma)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $P : \mathcal{S} \times \mathcal{A} \mapsto \Delta(\mathcal{S})$ is the transition kernel, $R : \mathcal{S} \times \mathcal{A} \mapsto \Delta([0, 1])$ encodes a family of reward distributions with $r : \mathcal{S} \times \mathcal{A} \mapsto [0, 1]$ as the expected reward function, $\rho : \mathcal{S} \mapsto \Delta(\mathcal{S})$ is the initial state distribution, and $\gamma \in [0, 1)$ is the discount factor. We assume $\mathcal{S}$ and $\mathcal{A}$ are finite however, our results do not depend on their cardinalities and can be naturally extended to infinite sets. A stationary (stochastic) policy $\pi : \mathcal{S} \mapsto \Delta(\mathcal{A})$ specifies a distribution over actions in each state. Each policy π induces an occupancy density over state-action pairs $d^\pi : \mathcal{S} \times \mathcal{A} \mapsto [0, 1]$ defined as $d^\pi(s, a) \coloneqq (1-\gamma) \sum_{t=0}^{\infty} \gamma^t P_t(s_t = s, a_t = a; \pi)$, where $P_t(s_t = s, a_t = a; \pi)$ denotes (s, a) visitation probability at step t, starting at $s_0 \sim \rho(\cdot)$ and following π. We abuse notation and also write $d^\pi(s) = \sum_{a \in \mathcal{A}} d^\pi(s, a)$ to denote the discounted state occupancy. Additionally, operator $\mathbb{P}^\pi$ is applied to any function $u : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ and is defined as $(\mathbb{P}^\pi u)(s, a) \coloneqq \sum_{s', a'} P(s'|s, a)\pi(a'|s')u(s', a')$.

An important quantity is the value a policy π, which is the discounted sum of rewards $V^\pi(s) \coloneqq \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r_t \mid s_0 = s, a_t \sim \pi(\cdot \mid s_t), \ \forall\ t \geq 0]$ starting at state $s \in \mathcal{S}$. We use the notation $J(\pi) \coloneqq (1-\gamma)\mathbb{E}_{s\sim\rho}[V^\pi(s)] = \mathbb{E}_{s,a\sim d^\pi}[r(s, a)]$ to represent a scalar summary of the performance of a policy π. We denote by $\pi^\star$ the optimal policy that maximizes the above objective and use the shorthand $V^\star \coloneqq V^{\pi^\star}$ to denote the optimal value function.

Offline reinforcement learning. We focus on the offline RL, where the agent is only provided with a previously-collected offline dataset $\mathcal{D} = \{(s_i, a_i, r_i, s'_i)\}_{i=1}^N$. Here, $r_i \sim R(s_i, a_i)$, $s'_i \sim P(\cdot \mid s_i, a_i)$, and we assume s_i, a_i pairs are generated i.i.d. according to a data distribution $\mu \in \Delta(\mathcal{S} \times \mathcal{A})$. To streamline the analysis, we assume that the conditional distribution $\mu(a|s)$ is known.[2] The goal of offline RL is to learn a policy $\hat{\pi}$ based on the offline dataset so as to minimize the sub-optimality with respect to the optimal policy $\pi^\star$, i.e. $J(\pi^\star) - J(\hat{\pi})$ with high probability. In this paper, we consider marginal importance sampling (MIS) formulation that aims at learning weights $w(s, a)$ to represent policy occupancy when multiplied by data distribution: $d_w(s, a) = w(s, a)\mu(s, a)$. Also denote $d_w(s) = \sum_{a \in \mathcal{A}} d_w(s, a)$. We define the policy induced by w as $\pi_w(a|s) = d_w(s, a)/d_w(s)$ for $d_w(s) > 0$ and $\pi_w(a|s) = 1/|\mathcal{A}|$ for $d_w(s) = 0$.

Offline data coverage assumption. We design and analyze our algorithms within the single-policy concentrability framework (Rashidinejad et al., 2021), stated below.

Definition 1 (Single-policy concentrability) *Given a policy π, define C^π to be the smallest constant that satisfies $\frac{d^\pi(s,a)}{\mu(s,a)} \leq C^\pi$ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$.*

$C^{\pi^\star} = C^\star$ captures coverage of $\pi^\star$ in the offline data and is much weaker than the widely used all-policy concentrability that assumes bounded $\max_\pi C^\pi$; see Appendix A for further discussion.

Notation. Given a set $\mathcal{S}$, we write $|\mathcal{S}|$ to represent its cardinality and $\Delta(\mathcal{S})$ to denote the probability simplex over $\mathcal{S}$. For a function class $\mathcal{W}$, we write $|\mathcal{W}|$ to denote its cardinality (discrete) or covering number (continuous). We use the notation $x \lesssim y$ when there exists constant $c > 0$ such that $x \leq cy$ and $x \asymp y$ if constants $c_1, c_2 > 0$ exist such that $c_1|x| \leq |y| \leq c_2|x|$. We write $f(x) = O(g(x))$ if $M > 0, x_0$ exist such that $|f(x)| \leq Mg(x)$ for all $x \geq x_0$ and use $\widetilde{O}(\cdot)$ to be the big-O notation ignoring logarithmic factors. Define $\mathrm{clip}(x, a, b) \triangleq \max\{a, \min\{x, b\}\}$ for $x, a, b \in \mathbb{R}$.

3 MULTI-ARMED BANDITS

We start by considering the offline learning problem in the multi-armed bandit (MAB) setting, which is a special case of MDP with $\gamma = 0$, $|\mathcal{S}| = 1$, and $\mathcal{D} = \{(a_i, r_i)\}_{i=1}^N$, where $a_i \sim \mu(\cdot), r_i \sim R(a_i)$. The goal of offline learning in MABs can be described as the following constrained optimization problem, where d represents occupancy

$$\max_{d \geq 0} \mathbb{E}_{a\sim d}[r(a)] \qquad \text{s.t.} \quad \sum_a d(a) = 1. \tag{1}$$

[2] When $\mu(a|s)$ is unknown, behavioral cloning can be used (Ross & Bagnell, 2014; Zhan et al., 2022).

3.1 PRIMAL-DUAL REGULARIZED OFFLINE BANDITS

To solve (1), the MIS approach with behavior regularization defines importance weights $w(a) = d(a)/\mu(a)$ and converts the problem (1) to its dual form by introducing the Lagrange multiplier v:

$$\max_{w\geq 0}\min_{v} L_\alpha^{\text{MAB}}(w,v) \coloneqq \mathbb{E}_{a\sim\mu}\left[w(a)r(a)\right] - v\left(\mathbb{E}_{a\sim\mu}[w(a)]-1\right) - \alpha\mathbb{E}_{a\sim\mu}\left[f\left(w(a)\right)\right]. \tag{2}$$

The last term in (2) is the behavior regularizer that characterizes the f-divergence between the learned occupancy d and data distribution μ, with α determining the strength of regularization. This term was originally proposed to induce *conservatism* by keeping the learned policy close to behavior policy (Nachum et al., 2019b; Lee et al., 2021). Approximating w via a function class $\mathcal{W} \subseteq \mathbb{R}^{|\mathcal{A}|}$ and solving the empirical version of (2) yields Algorithm 5, which we call primal-dual regularized offline MAB (PRO-MAB) as it is a special case of PRO-RL algorithm of Zhan et al. (2022).

One might wonder whether the unregularized algorithm ($\alpha = 0$) is sufficient for solving the offline learning problem in MABs, particularly under the natural and common assumption that elements of the function class $\mathcal{W}$ are bounded: $w(a) = d(a)/\mu(a) \leq B_w$. In the following proposition, we show that the answer is negative and there exist MAB instances in which the unregularized algorithm finds a policy that suffers from a constant suboptimality. The proof is provided in Appendix B.2.

Proposition 1 (Unregularized algorithm fails in MAB) *Let $\hat{\pi}$ be the policy returned by Algorithm 5 with $\alpha = 0$. There exists a 2-armed bandit problem in which policy $\hat{\pi}$ satisfies $J(\pi^\star) - J(\hat{\pi}) = 1/6$ with a constant probability.*

We note that Zhan et al. (2022) also argues the failure of the unregularized algorithm by giving a counterexample in the MDP setting. We discuss this example in detail in Section 5.3. Proposition 1 reveals additional insights: the objective (15) with $\alpha = 0$ fails not just in MDPs but also in bandits, even when the optimal policy is unique and data are collected by running a behavior policy.

Given the failure of the unregularized algorithm, we conduct a tight analysis of PRO-MAB with $\alpha > 0$. In the next theorem, we prove that under similar assumptions as Zhan et al. (2022) and with a proper choice of α, PRO-MAB returns a policy that enjoys near-optimal sample complexity.

Theorem 1 (Suboptimality of PRO-MAB) *Let $f : \mathbb{R} \mapsto \mathbb{R}$ be M_f-strongly convex, non-negative $f(x) \geq 0$, and bounded $|f(x)| \leq B_f$. Assume $0 \leq w(a) \leq B_w$ for any $w \in \mathcal{W}$. Set $\alpha \asymp M_f(B_w(B_v+1)+B_f)\log(N)\sqrt{\log(|\mathcal{V}||\mathcal{W}|/\delta)/N}$. Assume realizability of $w_\alpha^\star \in \mathcal{W}$ and $v_\alpha^\star \in \mathcal{V}$ the optimal solutions to (2). Let $\pi_\alpha^\star = \pi_{w_\alpha^\star}$ and suppose concentrability of $\pi^\star$ and $\pi_\alpha^\star$ (Definition 1). For any fixed $\delta \geq 0$, policy $\hat{\pi}$ returned by Algorithm 5 achieves*

$$J(\pi^\star) - J(\hat{\pi}) \lesssim \alpha(B_f + f'(C^\star)B_w) = \widetilde{O}\left(\frac{1}{\sqrt{N}}\right).$$

To our knowledge, this is the first statistically optimal guarantee for a practical offline MAB algorithm with function approximation and partial coverage and improves over the $1/N^{1/6}$ guarantee given by Zhan et al. (2022). We now briefly explain the differences between the analysis methods; a complete proof is deferred to Appendix B.3. Zhan et al. (2022) bounds policy suboptimality by $\alpha + 1/(\alpha^{1/2}N^{1/4})$, where the first term comes from the bias caused by the regularizer and the second term stems from bounding the difference between $\hat{w}$ and $w_\alpha^\star$ relying on the *strong convexity* of L_α. Optimizing the bound over α gives the final $1/N^{1/6}$ guarantee. In contrast, our analysis connects suboptimality to occupancy validity. In particular, we show that suboptimality is bounded by $\alpha + 1/(d_{\hat{w}}\sqrt{N})$, where $d_{\hat{w}} = \sum_a \hat{w}(a)\mu(a)$. We then show that setting $\alpha = \tilde{O}(1/\sqrt{N})$ is sufficient to ensure near-validity of occupancy $d_{\hat{w}} = \Omega(1/\log(N))$, which proves the claim.

We observe a similar phenomenon in Proposition 1 that small d_w for certain $w \in \mathcal{W}$ can cause the unregularized algorithm to fail. In the following section, we investigate this phenomenon further, leading to a new offline learning algorithm.

3.2 AUGMENTED LAGRANGIAN REPLACES BEHAVIOR REGULARIZATION

The next proposition further cements the importance of policy validity and shows that if the occupancy is valid, such as by searching only over the weights that induce valid occupancies, then the unregularized algorithm enjoys an optimal rate. Proof of this result can be found in Appendix B.4.

Algorithm 1 Conservative Offline MAB with Augmented Lagrangian

1: **Inputs:** Dataset $\mathcal{D} = \{(a_i, r_i)\}_{i=1}^N$, classes $\mathcal{W}$ and $\mathcal{V} = [-B_v, B_v]$.
2: Find a solution $\hat{w}, \hat{v}$ to the following problem

$$\max_{w\in\mathcal{W}} \min_{v\in\mathcal{V}} \hat{L}_{AL}^{\text{MAB}}(w, v) := \frac{1}{N}\sum_{i=1}^N w(a_i)r_i - v(w(a_i) - 1) - \Big(\frac{1}{N}\sum_{i=1}^N w(a_i) - 1\Big)^2. \quad (3)$$

3: **Return:** $\hat{\pi} = \pi_{\hat{w}}$.

Proposition 2 (Constraint satisfaction is sufficient in MAB) *Assume as in Theorem 1. Let $\hat{\pi}$ be the output of Algorithm 5 with $\alpha = 0$ and assume that $\sum_a \mu(a)\hat{w}(a) = 1$. Then, for any fixed $\delta > 0$*

$$J(\pi^\star) - J(\hat{\pi}) \lesssim (B_w(B_v + 1) + \alpha B_f)\sqrt{\frac{\log|\mathcal{V}||\mathcal{W}|/\delta}{N}}.$$

Motivated by the discussion above, we take a step back and ask: are there any other methods for solving constrained optimization problems that find *more constraint-satisfying* solutions when applied to the empirical approximation of the original problem? A promising candidate is the augmented Lagrangian method (ALM) which adds a quadratic loss on the constraints to the objective. Applied to (1), ALM forms the following objective, whose empirical version leads to Algorithm 1.

$$\max_{w\geq 0} \min_{v} L_{\text{AL}}^{\text{MAB}}(w, v) := \mathbb{E}_{a\sim\mu}\left[w(a)r(a)\right] - v\left(\mathbb{E}_{a\sim\mu}[w(a)] - 1\right) - \left(\mathbb{E}_{a\sim\mu}[w(a)] - 1\right)^2. \quad (4)$$

The following theorem establishes an upper bound on the suboptimality of the policy returned by Algorithm 1, whose proof can be found in Appendix B.5.

Theorem 2 (Suboptimality of Algorithm 1) *Suppose $\pi^\star$-concentrability (Definition 1). Assume $w^\star \in \mathcal{W}$, where $w^\star(a) = d^{\pi^\star}(a)/\mu(a)$, and $v^\star \in \mathcal{V}$, where $v^\star = J(\pi^\star)$. Moreover, assume that $0 \leq w(a) \leq B_w$ for $w \in \mathcal{W}$. For any fixed $\delta \geq 0$, policy $\hat{\pi}$ returned by Algorithm 1 achieves*

$$J(\pi^\star) - J(\hat{\pi}) \lesssim (B_w + 1)^2(B_v + 1)\sqrt{\frac{\log(|\mathcal{W}||\mathcal{V}|/\delta)}{N}}.$$

In the proof, we show that ALM results in near-validity of $\hat{w}$ by ensuring that $d_{\hat{w}} = \Omega(1)$, leading to the optimal rate. Note that Algorithm 1 does not include any explicit form of conservatism through regularizers or uncertainty quantifiers. Colloquially, the MIS formulation and boundedness of $\mathcal{W}$ elements ensure that $d_{\hat{w}}(a)/\mu(a) = \hat{w}(a) \leq B_w$ and ALM ensures that $d_{\hat{w}}$ is close to the actual occupancy. Thus, Algorithm 1 seeks a policy whose *actual occupancy* is within data distribution. Algorithm 1 offers several benefits compared to PRO-MAB: it only requires $\pi^\star$-concentrability as instead of $\pi^\star, \pi_\alpha^\star$-concentrability in PRO-MAB, improves the rate by $\log N$, removes the need to design regularizer f and adjust α, and does not introduce any bias in the objective. The main advantage of ALM, however, becomes more evident as we move beyond MAB, where the behavior regularization provably fails to achieve optimal statistical rate while ALM maintains optimality.

4 Contextual Bandits

The problem offline contextual bandits (CB) is a special case of offline RL with $\gamma = 0$ and offline dataset $\mathcal{D} = \{(s_i, a_i, r_i)\}_{i=1}^N$, where $s_i \sim \mu(\cdot) = \rho(\cdot)$, $a_i \sim \mu(\cdot \mid a_i)$, and $r_i \sim R(s_i, a_i)$. The linear programming constrained optimization problem for CB is given by

$$\max_{d\geq 0} \mathbb{E}_{s,a\sim d}\left[r(s, a)\right] \qquad \text{s.t.} \quad \sum_a d(s, a) = \rho(s) \quad \forall s \in \mathcal{S}. \quad (5)$$

4.1 Primal-Dual Regularized Offline Contextual Bandits

In the following proposition, we prove a performance lower bound on the primal-dual regularized offline CB (PRO-CB) presented in Algorithm 6, which includes behavior regularization.

Proposition 3 (PRO-CB is suboptimal) *Let $f : \mathbb{R} \mapsto \mathbb{R}$ be M_f-strongly convex, non-negative, and bounded. Assume $0 \leq w(s, a) \leq B_w$ for $w \in \mathcal{W}$ and realizability of the optimal solutions to (23): $w_\alpha^\star \in \mathcal{W}$ and $v_\alpha^\star \in \mathcal{V}$. Let $\pi_\alpha^\star = \pi_{w_\alpha^\star}$ and assume $\pi_0^\star, \pi_\alpha^\star$-concentrability. Let $\hat{\pi}$ be the output of Algorithm 6. Then, for any $\alpha \geq 0$ there exists a CB instance such that with probability $\Omega(1)$, $J(\pi^\star) - J(\hat{\pi}) = \Omega(N^\beta)$, where $\beta > -1/2$.*

Algorithm 2 Conservative Offline CB with Augmented Lagrangian

1: **Inputs:** Dataset $\mathcal{D} = \{(s_i, a_i, r_i)\}_{i=1}^N$, function classes $\mathcal{W}, \mathcal{V}$
2: Find a solution $\hat{w}, \hat{v}$ to the following problem

$$\max_{w \in \mathcal{W}} \min_{v \in \mathcal{V}} \hat{L}_{\text{AL}}^{\text{CB}}(w, v) := \frac{1}{N} \sum_{i=1}^N w(s_i, a_i)(r_i - v(s_i)) + v(s_i) - \Big(\sum_{a \in \mathcal{A}} w(s_i, a)\mu(a|s_i) - 1 \Big)^2 \tag{7}$$

3: **Return:** $\hat{\pi} = \pi_{\hat{w}}$.

Proposition 3 shows that behavior regularization is statistically suboptimal regardless of α. The proof is presented in Appendix C.2. The main takeaway is that ensuring occupancy validity $\sum_a \hat{w}(s, a)\mu(a|s) = \Omega(1)$ for nearly all states appears to be critical in achieving the optimal rate. Yet, without introducing a large bias, behavior regularization is insufficient for such a guarantee.

4.2 Offline contextual bandits with augmented Lagrangian

To encourage occupancy validity, we extend ALM to CBs and propose the following objective:

$$\begin{aligned} &\max_{w \geq 0} \min_{v} L_{\text{AL}}^{\text{CB}}(w, v) \\ &:= \mathbb{E}_\mu \left[w(s, a) r(s, a) \right] - \mathbb{E}_\mu [v(s)(w(s, a) - 1)] - \mathbb{E}_{s \sim \mu}[(\mathbb{E}_{a \sim \mu(\cdot|s)}[w(s, a)] - 1)^2] \end{aligned} \tag{6}$$

Notice that when $|\mathcal{S}| = 1$, (6) simplifies to the ALM objective (2) in the MAB setting. The ALM term can be understood in the following way. Each element encourages the validity of occupancy in each state $\sum_a w(s, a)\mu(s, a) \approx 1$ and the elements are weighted according to the true state distribution: validity is more important in states that are actually more likely to be visited. We analyze the suboptimality of Algorithm 2 and present the following theorem, showing that the ALM achieves optimal rate without any need for behavior regularization. The proof can be found in Appendix C.3.

Theorem 3 (Suboptimality of Algorithm 2) *Suppose concentrability of an optimal policy $\pi^\star$ (Definition 1). Assume realizability of $w^\star \in \mathcal{W}$ where $w^\star(s, a) = d^{\pi^\star}(s, a)/\mu(s, a)$ and $v^\star \in \mathcal{V}$, where $v^\star = J(\pi^\star)$. Moreover, assume that $|v(s)| \leq B_v$ for $v \in \mathcal{V}$ and $0 \leq w(s, a) \leq B_w$. For any fixed $\delta \geq 0$, policy $\hat{\pi}$ returned by Algorithm 2 achieves*

$$J(\pi^\star) - J(\hat{\pi}) \lesssim (B_w + 1)^2 (B_v + 1) \sqrt{\frac{\log(|\mathcal{W}||\mathcal{V}|/\delta)}{N}}.$$

5 Markov decision processes

We now turn to offline RL. In addition to the offline dataset, we assume access to a dataset $\mathcal{D}_0 = \{s_i\}_{i=1}^N$ with i.i.d. samples from the initial distribution ρ, similar to prior works (Lee et al., 2021; Zhan et al., 2022). The linear programming formulation of RL (Puterman, 2014) solves

$$\max_{d \geq 0} \mathbb{E}_{s,a \sim d}[r(s, a)] \qquad \text{s.t.} \quad d(s) = (1 - \gamma)\rho(s) + \gamma \sum_{s', a'} P(s|s', a') d(s', a') \quad \forall s \in \mathcal{S}. \tag{8}$$

The constraints are known as Bellman flow equations and restrict the search to the space of valid occupancy distributions d^π that can be induced in the MDP by running a policy π.

5.1 Conservative offline RL with augmented Lagrangian

Motivated by the success of ALM in bandits, we propose the following extension to offline RL:

$$\max_{w \geq 0} \min_{v} L_{\text{AL}}^{\text{MDP}}(w, v) := (1 - \gamma)\mathbb{E}_\rho[v(s)] + \mathbb{E}_\mu \left[w(s, a) e_v(s, a) \right] - \mathbb{E}_{d^{\pi_w}} \left(\frac{d_w(s)}{d^{\pi_w}(s)} - 1 \right)^2, \tag{9}$$

where $e_v(s, a) := r(s, a) + \gamma \sum_{s'} P(s'|s, a) v(s') - v(s)$. One can check that the first two terms are the Lagrange dual of (8) and the last term is a generalization of the ALM terms in bandits. The ALM elements encourage occupancy $d_w(s)$ to be close in ratio to the actual occupancy $d^{\pi_w}(s)$ in each state and as before, the ALM elements are weighted according to their actual visitation $d^{\pi_w}(s)$. Our particular ALM construction can be intuitively understood as follows. The MIS formulation learns bounded weights $\hat{w}(s, a) = d_{\hat{w}}(s, a)/\mu(s, a) \leq B_w$. The ALM term ensures that the ratio $d_{\hat{w}}(s)/d^{\pi_{\hat{w}}}(s) = \Omega(1)$ which roughly translates to $d^{\pi_{\hat{w}}}(s, a)/\mu(s, a) \lesssim B_w$.

The ALM term in (9) is difficult to estimate as it involves the expectation over unknown occupancy d^{π_w} and the computation of the ratio $d_w(s)/d^{\pi_w}(s)$. We resolve this difficulty in the next sections.

Algorithm 3 Conservative Offline RL with Augmented Lagrangian (CORAL) — Model-based

1: **Inputs:** Datasets $\mathcal{D}$, $\mathcal{D}_0$, $\mathcal{D}_m$, function classes $\mathcal{W}, \mathcal{V}, \mathcal{U}, \mathcal{P}$, $f_*^{-1}(x) = 2\sqrt{x+1} - 2$.
2: Estimate transitions via maximum likelihood: $\hat{P} = \operatorname{argmax}_{P \in \mathcal{P}} \sum_{i=1}^{N_m} \ln P(s_i' | s_i, a_i)$.
3: Find a solution $\hat{w}, \hat{v}, \hat{u}$ to the following problem

$$\max_{w \in \mathcal{W}} \min_{v \in \mathcal{V}} \min_{u \in \mathcal{U}} \hat{L}_{AL}^{\text{model-based}}(w, v) := \frac{(1-\gamma)}{N_0} \sum_{i=1}^{N_0} \left(v(s_i) + \sum_a u(s_i, a) \pi_w(a|s_i) \right) + \frac{1}{N} \sum_{i=1}^{N} w(s_i, a_i) \left[r_i + \gamma v(s_i') - v(s_i) - f_*^{-1}\left(u(s_i, a_i) - \gamma (\hat{\mathbb{P}}^{\pi_w} u)(s_i, a_i) \right) \right] \tag{13}$$

4: **Return:** $\hat{\pi} = \pi_{\hat{w}}$.

5.2 Estimating the ALM term and CORAL algorithm

We view the ALM term as the negative f-divergence between d_w and d^{π_w} with $f(x) := (x-1)^2$ and express it in the variational form (Nguyen et al., 2010):

$$-\mathbb{E}_{d^{\pi_w}} \left(\frac{d_w(s)}{d^{\pi_w}(s)} - 1 \right)^2 = -D_f(d_w \| d^{\pi_w}) = \min_x \mathbb{E}_{d^{\pi_w}} [f_*(x(s,a))] - \mathbb{E}_{d_w}[x(s,a)]. \tag{10}$$

Here, f_* is the convex conjugate of f and we used the fact that $d_w(s,a)/d^{\pi_w}(s,a) = d_w(s)/d^{\pi_w}(s)$. Notice that $\mathbb{E}_{d^{\pi_w}}[f_*(x(s,a))]$ is the value of π_w in the same MDP but with rewards $f_*(x(s,a))$. Define u as the fixed point of the following Bellman equation

$$u(s,a) := f_*(x(s,a)) + \gamma (\mathbb{P}^{\pi_w} u)(s,a). \tag{11}$$

Since $u(s,a)$ is the Q-function of π_w with rewards $f_*(x(s,a))$, we can rewrite (10) as

$$(10) = \min_u (1-\gamma) \mathbb{E}_{s \sim \rho, a \sim \pi_w}[u(s,a)] - \mathbb{E}_\mu \left[w(s,a) f_*^{-1}\left(u(s,a) - \gamma (\mathbb{P}^{\pi_w} u)(s,a) \right) \right]. \tag{12}$$

(12) involves expectations over ρ and μ, which can be estimated empirically. Below, we discuss model-free and model-based methods for estimating the term involving the transition operator $\mathbb{P}^{\pi_w}$. We include some details on practical implementations in Appendix D.1.

Model-based CORAL. For the model-based route, we assume access to a realizable function class $\mathcal{P}$ that contains the true transitions and an additional dataset $\mathcal{D}_m = \{(s_i, a_i, s_i')\}_{i=1}^{N_m}$, where $s_i, a_i \sim \mu$ and $s_i' \sim P(\cdot \mid s_i, a_i)$. Given $\mathcal{D}_m$, we obtain a maximum likelihood estimate of transitions and then approximate the expectations using $\mathcal{D}_0$ and $\mathcal{D}$. This leads to Algorithm 3, which we name model-based conservative offline RL with augmented Lagrangian (CORAL).

Model-free CORAL. As an alternative, we consider developing a model-free that uses a single-sample estimate of $f_*^{-1}(u(s,a) - \gamma(P^{\pi_w} u)(s,a))$. This, however, roughly leads to the infamous double sampling problem (Baird, 1995). To circumvent this difficulty, in Appendix D.4 we use the dual embedding trick in Nachum et al. (2019a), to derive model-free CORAL (Algorithm 4).

Theorem 4 shows that both variants of CORAL enjoy optimal rates; see Appendix D.5 for the proof.

Theorem 4 (CORAL Suboptimality) *Suppose concentrability of an optimal policy $\pi^\star$ and assume $w^\star(s,a) = d^{\pi^\star}(s,a)/\mu(s,a) \in \mathcal{W}$ and $v^\star(s) = V^\star(s) \in \mathcal{V}$. Moreover, assume $0 \le w(s,a) \le B_w$ for $w \in \mathcal{W}$ and $|v(s)| \le B_v$ for $v \in \mathcal{V}$. Let $\tilde{x}_w(s,a) = \operatorname{clip}(x_w^\star(s,a), -1, 1)$, where $x_w^\star$ is a solution to (10), and define $u_w^\star$ as the fixed-point solution to (11) when $x = \tilde{x}_w$. Assume $u_w^\star \in \mathcal{U}$ for any $w \in \mathcal{W}$ and $\|u\|_\infty \le B_u$ where $B_u \ge 5/(4(1-\gamma))$. For any fixed $\delta \ge 0$, the following claims hold.*

(I) Assume $N = N_0 = N_m$ for simplicity. If $P^\star \in \mathcal{P}$, then $\hat{\pi}$ returned by Algorithm 3 achieves

$$J(\pi^\star) - J(\hat{\pi}) \lesssim \frac{B_v + B_u + (1+B_v)B_w}{1-\gamma} \sqrt{\frac{B_u \log(|\mathcal{P}||\mathcal{U}||\mathcal{W}||\mathcal{V}|/\delta)}{N}}.$$

(II) Assume $N = N_0$ for simplicity. Let $\zeta_{w,u}^\star = \operatorname{argmax}_{\zeta < 0} L_{AL}^{\text{model-free}}(w, v, u, \zeta)$ defined in (43). Assume $\zeta_{w^\star,u}^\star \in \mathcal{Z}$ for $u \in \mathcal{U}$ and $B_{\zeta,L} \le |\zeta(s,a)| \le B_{\zeta,U}$ for $\zeta \in \mathcal{Z}$, where $B_{\zeta,L} \in (0, \sqrt{2}/3)$ and $B_{\zeta,U} \ge \sqrt{2}$. Let $B_\zeta = \max\{B_{\zeta,U}, B_{\zeta,L}^{-1}\}$. Then, $\hat{\pi}$ returned by Algorithm 4 achieves

$$J(\pi^\star) - J(\hat{\pi}) \lesssim \frac{B_v + B_u + (1 + B_v + B_\zeta(B_u+1))B_w}{1-\gamma} \sqrt{\frac{\log(|\mathcal{U}||\mathcal{W}||\mathcal{V}||\mathcal{Z}|/\delta)}{N}}.$$

Algorithm 4 Conservative Offline RL with Augmented Lagrangian (CORAL) — Model-free

1: **Inputs:** Datasets $\mathcal{D}, \mathcal{D}_0$, function classes $\mathcal{W}, \mathcal{V}, \mathcal{U}, \mathcal{Z}$, $g_*(x) = x + 2 + \frac{1}{x}$.

2: Find a solution $\hat{w}, \hat{v}, \hat{u}, \hat{\zeta}$ to $\max_{w \in \mathcal{W}} \min_{v \in \mathcal{V}} \min_{u \in \mathcal{U}} \max_{\zeta \in Z} \hat{L}_{\text{AL}}^{\text{model-free}}(w, v, u, \zeta)$ defined as

$$\frac{(1-\gamma)}{N_0} \sum_{i=1}^{N_0} v(s_i) + \sum_a u(s_i, a)\pi_w(a|s_i) + \frac{1}{N} \sum_{i=1}^{N} w(s_i, a_i) \Big[r_i + \gamma v(s_i') - v(s_i) + \zeta(s_i, a_i) \Big(u(s_i, a_i) - \gamma \sum_{a' \in \mathcal{A}} u(s_i', a')\pi_w(a'|s_i') \Big) + g_*(\zeta(s_i, a_i)) \Big] \tag{14}$$

3: **Return:** $\hat{\pi} = \pi_{\hat{w}}$.

In Theorem 4, we make realizability assumptions on $u_w^\star$ for $w \in \mathcal{W}$ and $\zeta_{w^\star, u}^\star$ for $u \in \mathcal{U}$. Such assumptions are common in theory of RL with function approximation (Munos & Szepesvári, 2008; Xie et al., 2021; Jiang & Huang, 2020) and removing them can be difficult or impossible (Foster et al., 2021). Recently, Zhan et al. (2022); Chen & Jiang (2022) propose algorithms that only require optimal solution realizability, however, these algorithms are either intractable or suboptimal.

5.3 Example: Behavior Regularization vs. Augmented Lagrangian

We examine the MDP example in Figure 1 presented by Zhan et al. (2022). Assume $\mathcal{V} = \{v^\star\}$ and $\mathcal{W} = \{w_1, w_2\}$, where w_1 always selects L from A and w_2 always selects R from A. One can check $w_1(A, L) = 2, w_1(A, R) = 0$ and $w_2(A, L) = 0, w_2(A, R) = 1$.

Unregularized algorithm. As Zhan et al. (2022) state, the unregularized algorithm fails to distinguish between w_1 and w_2 even with infinite data as the objectives at w_1 and w_2 are exactly equal.

Behavior regularization. Consider an instantiation of PRO-RL with regularizer $-\alpha \mathbb{E}_\mu[w^2(s, a)]$. Since in this example $\mathbb{E}_\mu[w_1^2(s, a)] > \mathbb{E}_\mu[w_2^2(s, a)]$, PRO-RL picks the wrong weight w_2, suffering constant suboptimality. Note, however, that PRO-RL guarantees assume $\pi_\alpha^\star$-concentrability. Intuitively, behavior regularization causes $\pi_\alpha^\star$ to be more stochastic and thus requiring $\mu(s, a) > 0$ for more states and actions. Here, since μ covers both (A, L) and (A, R), behavior regularization causes $\pi_\alpha^\star(R|A) > 0$ and thus $d^{\pi_\alpha^\star}(C) > 0$. To handle the MDP in Figure 1, PRO-RL additionally requires $\mu(C) > 0$ to satisfy $\pi_\alpha^\star$-concentrability.

ALM. In this example, ALM successfully picks the optimal w_1, as it avoids a mismatch between the actual and learned occupancies. This is because in (9) the ALM term is zero at w_1 due to realizability whereas at w_2, it has a lower bound $\mathbb{E}_{s \sim d^{\pi_2}} \left(d_{w_2}(C)/d^{\pi_2}(C) - 1\right)^2 \geq d^{\pi_2}(C) > 0$.

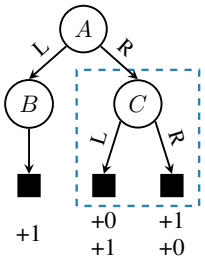

Figure 1: The agent starts from A. Action L leads to B, from where the agent collects +1 reward. Action R leads to C, from where only one action leads to a +1 reward. Nature decides which MDP is presented to the learner. Data distribution is $\mu(A, L) = 1/4, \mu(A, R) = 1/2, \mu(B) = 1/4, \mu(C) = 0$, which satisfies π_{w_1}-concentrability.

6 Discussion

We present a set of practical and statistically optimal algorithms for offline MAB, CB, and RL, under general function approximation and single-policy concentrability. Our algorithms are designed within the MIS formulation combined with a novel application of augmented Lagrangian method. Importantly, our optimality guarantees hold under MIS combined with ALM alone, without any additional form of conservatism such as via regularization or uncertainty quantification. Furthermore, we investigate the role of regularizers in MIS algorithms. Although the empirical benefits of such regularizers are often attributed to conservatism, our analysis suggests that conservatism stems from the MIS formulation while the role of regularizers is to ensure the validity of learned occupancy. Interesting future directions include conducting empirical evaluations of ALM, examining the possibility of removing strong realizability assumptions, and investigating practical and optimal offline RL algorithms whose guarantees hold under milder variants of single-policy concentrability more suited to function approximation.

REFERENCES

Rishabh Agarwal, Dale Schuurmans, and Mohammad Norouzi. An optimistic perspective on offline reinforcement learning. In *International Conference on Machine Learning*, pp. 104–114. PMLR, 2020.

Andras Antos, Rémi Munos, and Csaba Szepesvari. Fitted Q-iteration in continuous action-space mdps. In *Neural Information Processing Systems*, 2007.

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, 2008.

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

Aharon Ben-Tal and Arkadi Nemirovski. Lecture notes optimization III: Convex analysis, Non-linear programming theory, Non-linear programming algorithms, 2022.

Jinglin Chen and Nan Jiang. Information-theoretic considerations in batch reinforcement learning. *arXiv preprint arXiv:1905.00360*, 2019.

Jinglin Chen and Nan Jiang. Offline reinforcement learning under value and density-ratio realizability: The power of gaps. In *The 38th Conference on Uncertainty in Artificial Intelligence*, 2022.

Ching-An Cheng, Tengyang Xie, Nan Jiang, and Alekh Agarwal. Adversarially trained actor critic for offline reinforcement learning. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, 2022.

Bo Dai, Niao He, Yunpeng Pan, Byron Boots, and Le Song. Learning from conditional distributions via dual embeddings. In *Artificial Intelligence and Statistics*, pp. 1458–1467. PMLR, 2017.

Jonas Degrave, Federico Felici, Jonas Buchli, Michael Neunert, Brendan Tracey, Francesco Carpanese, Timo Ewalds, Roland Hafner, Abbas Abdolmaleki, Diego de Las Casas, et al. Magnetic control of tokamak plasmas through deep reinforcement learning. *Nature*, 602(7897):414–419, 2022.

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*, 2019.

Amir Massoud Farahmand, Rémi Munos, and Csaba Szepesvári. Error propagation for approximate policy and value iteration. In *Advances in Neural Information Processing Systems*, 2010.

Yihao Feng, Lihong Li, and Qiang Liu. A kernel loss for solving the Bellman equation. *arXiv preprint arXiv:1905.10506*, 2019.

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 and Shixiang Shane Gu. A minimalist approach to offline reinforcement learning. *Advances in neural information processing systems*, 34:20132–20145, 2021.

Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In *International Conference on Machine Learning*, pp. 2052–2062. PMLR, 2019.

Seyed Kamyar Seyed Ghasemipour, Dale Schuurmans, and Shixiang Shane Gu. EMaQ: Expected-max Q-learning operator for simple yet effective offline and online RL. *arXiv preprint arXiv:2007.11091*, 2020.

L Jeff Hong, Weiwei Fan, and Jun Luo. Review on ranking and selection: A new perspective. *Frontiers of Engineering Management*, 8(3):321–343, 2021.

Natasha Jaques, Asma Ghandeharioun, Judy Hanwen Shen, Craig Ferguson, Agata Lapedriza, Noah Jones, Shixiang Gu, and Rosalind Picard. Way off-policy batch deep reinforcement learning of implicit human preferences in dialog. *arXiv preprint arXiv:1907.00456*, 2019.

Nan Jiang. On value functions and the agent-environment boundary. *arXiv preprint arXiv:1905.13341*, 2019.

Nan Jiang and Jiawei Huang. Minimax value interval for off-policy evaluation and policy optimization. *Advances in Neural Information Processing Systems*, 33, 2020.

Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline RL? In *International Conference on Machine Learning*, pp. 5084–5096. PMLR, 2021.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *ICML*, volume 2, pp. 267–274, 2002.

Rahul Kidambi, Aravind Rajeswaran, Praneeth Netrapalli, and Thorsten Joachims. MOReL: Model-based offline reinforcement learning. *arXiv preprint arXiv:2005.05951*, 2020.

Ilya Kostrikov, Rob Fergus, Jonathan Tompson, and Ofir Nachum. Offline reinforcement learning with Fisher divergence critic regularization. In *International Conference on Machine Learning*, pp. 5774–5783. PMLR, 2021.

Aviral Kumar, Justin Fu, George Tucker, and Sergey Levine. Stabilizing off-policy Q-learning via bootstrapping error reduction. *arXiv preprint arXiv:1906.00949*, 2019.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative Q-learning for offline reinforcement learning. *arXiv preprint arXiv:2006.04779*, 2020.

Aviral Kumar, Joey Hong, Anikait Singh, and Sergey Levine. Should i run offline reinforcement learning or behavioral cloning? In *International Conference on Learning Representations*, 2021.

Sascha Lange, Thomas Gabel, and Martin Riedmiller. Batch reinforcement learning. In *Reinforcement learning*, pp. 45–73. Springer, 2012.

Jongmin Lee, Wonseok Jeon, Byungjun Lee, Joelle Pineau, and Kee-Eung Kim. OptiDICE: Offline policy optimization via stationary distribution correction estimation. In *International Conference on Machine Learning*, pp. 6120–6130. PMLR, 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.

Gen Li, Laixi Shi, Yuxin Chen, Yuejie Chi, and Yuting Wei. Settling the sample complexity of model-based offline reinforcement learning. *arXiv preprint arXiv:2204.05275*, 2022.

Peng Liao, Zhengling Qi, and Susan Murphy. Batch policy learning in average reward Markov decision processes. *arXiv preprint arXiv:2007.11771*, 2020.

Boyi Liu, Qi Cai, Zhuoran Yang, and Zhaoran Wang. Neural trust region/proximal policy optimization attains globally optimal policy. In *Neural Information Processing Systems*, 2019a.

Qiang Liu, Lihong Li, Ziyang Tang, and Dengyong Zhou. Breaking the curse of horizon: Infinite-horizon off-policy estimation. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 5361–5371, 2018.

Yao Liu, Adith Swaminathan, Alekh Agarwal, and Emma Brunskill. Off-policy policy gradient with state distribution correction. *arXiv preprint arXiv:1904.08473*, 2019b.

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

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

Matej Moravčík, Martin Schmid, Neil Burch, Viliam Lisỳ, Dustin Morrill, Nolan Bard, Trevor Davis, Kevin Waugh, Michael Johanson, and Michael Bowling. Deepstack: Expert-level artificial intelligence in heads-up no-limit poker. *Science*, 356(6337):508–513, 2017.

Rémi Munos. Performance bounds in ℓ_p-norm for approximate value iteration. *SIAM journal on control and optimization*, 46(2):541–561, 2007.

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

Ofir Nachum and Bo Dai. Reinforcement learning via Fenchel-Rockafeller 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. In *Advances in Neural Information Processing Systems*, pp. 2315–2325, 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.

Ashvin Nair, Murtaza Dalal, Abhishek Gupta, and Sergey Levine. Accelerating online reinforcement learning with offline datasets. *arXiv preprint arXiv:2006.09359*, 2020.

XuanLong Nguyen, Martin J Wainwright, and Michael I Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010.

Xue Bin Peng, Aviral Kumar, Grace Zhang, and Sergey Levine. Advantage-weighted regression: Simple and scalable off-policy reinforcement learning. *arXiv preprint arXiv:1910.00177*, 2019.

Martin L Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, 2014.

Tabish Rashid, Bei Peng, Wendelin Boehmer, and Shimon Whiteson. Optimistic exploration even with a pessimistic initialisation. In *International Conference on Learning Representations*, 2019.

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*, 34:11702–11716, 2021.

Shideh Rezaeifar, Robert Dadashi, Nino Vieillard, Léonard Hussenot, Olivier Bachem, Olivier Pietquin, and Matthieu Geist. Offline reinforcement learning as anti-exploration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, number 7, pp. 8106–8114, 2022.

R Tyrrell Rockafellar and Roger J-B Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.

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

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

Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczynski. *Lectures on Stochastic Programming: Modeling and Theory*. SIAM, 2021.

Laixi Shi and Yuejie Chi. Distributionally robust model-based offline reinforcement learning with near-optimal sample complexity. *arXiv preprint arXiv:2208.05767*, 2022.

Laixi Shi, Gen Li, Yuting Wei, Yuxin Chen, and Yuejie Chi. Pessimistic Q-learning for offline reinforcement learning: Towards optimal sample complexity. *arXiv preprint arXiv:2202.13890*, 2022.

Noah Y Siegel, Jost Tobias Springenberg, Felix Berkenkamp, Abbas Abdolmaleki, Michael Neunert, Thomas Lampe, Roland Hafner, and Martin Riedmiller. Keep doing what worked: Behavioral modelling priors for offline reinforcement learning. *arXiv preprint arXiv:2002.08396*, 2020.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of Go without human knowledge. *nature*, 550(7676):354–359, 2017.

Adith Swaminathan and Thorsten Joachims. Batch learning from logged bandit feedback through counterfactual risk minimization. *The Journal of Machine Learning Research*, 16(1):1731–1755, 2015.

Csaba Szepesvári and Rémi Munos. Finite time bounds for sampling based fitted value iteration. In *Proceedings of the 22nd international conference on Machine learning*, pp. 880–887, 2005.

Guy Tennenholtz, Nir Baram, and Shie Mannor. Latent geodesics of model dynamics for offline reinforcement learning. In *Deep RL Workshop NeurIPS 2021*, 2021.

Masatoshi Uehara and Wen Sun. Pessimistic model-based offline reinforcement learning under partial coverage. In *International Conference on Learning Representations*, 2021.

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

Masatoshi Uehara, Xuezhou Zhang, and Wen Sun. Representation learning for online and offline RL in low-rank MDPs. *arXiv preprint arXiv:2110.04652*, 2021.

Sara Van de Geer. *Empirical Processes in M-estimation*, volume 6. Cambridge university press, 2000.

Oriol Vinyals, Timo Ewalds, Sergey Bartunov, Petko Georgiev, Alexander Sasha Vezhnevets, Michelle Yeo, Alireza Makhzani, Heinrich Küttler, John Agapiou, Julian Schrittwieser, et al. Starcraft II: A new challenge for reinforcement learning. *arXiv preprint arXiv:1708.04782*, 2017.

Lingxiao Wang, Qi Cai, Zhuoran Yang, and Zhaoran Wang. Neural policy gradient methods: Global optimality and rates of convergence. In *International Conference on Learning Representations*, 2019.

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.

Xinqi Wang, Qiwen Cui, and Simon S Du. On gap-dependent bounds for offline reinforcement learning. *arXiv preprint arXiv:2206.00177*, 2022.

Yuanhao Wang, Ruosong Wang, and Sham Kakade. An exponential lower bound for linearly realizable MDP with constant suboptimality gap. *Advances in Neural Information Processing Systems*, 34:9521–9533, 2021.

Gellért Weisz, Philip Amortila, and Csaba Szepesvári. Exponential lower bounds for planning in MDPs with linearly-realizable optimal action-value functions. In *Algorithmic Learning Theory*, pp. 1237–1264. PMLR, 2021.

Yifan Wu, George Tucker, and Ofir Nachum. Behavior regularized offline reinforcement learning. *arXiv preprint arXiv:1911.11361*, 2019.

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

Tengyang Xie, Yifei Ma, and Yu-Xiang Wang. Towards optimal off-policy evaluation for reinforcement learning with marginalized importance sampling. *Advances in Neural Information Processing Systems*, 32, 2019.

Tengyang Xie, Ching-An Cheng, Nan Jiang, Paul Mineiro, and Alekh Agarwal. Bellman-consistent pessimism for offline reinforcement learning. *Advances in neural information processing systems*, 34:6683–6694, 2021.

Yuling Yan, Gen Li, Yuxin Chen, and Jianqing Fan. The efficacy of pessimism in asynchronous Q-learning. *arXiv preprint arXiv:2203.07368*, 2022.

Mengjiao Yang, Ofir Nachum, Bo Dai, Lihong Li, and Dale Schuurmans. Off-policy evaluation via the regularized Lagrangian. *Advances in Neural Information Processing Systems*, 33:6551–6561, 2020.

Ming Yin and Yu-Xiang Wang. Towards instance-optimal offline reinforcement learning with pessimism. *Advances in neural information processing systems*, 34:4065–4078, 2021.

Ming Yin, Yu Bai, and Yu-Xiang Wang. Near-optimal offline reinforcement learning via double variance reduction. *arXiv preprint arXiv:2102.01748*, 2021.

Ming Yin, Yaqi Duan, Mengdi Wang, and Yu-Xiang Wang. Near-optimal offline reinforcement learning with linear representation: Leveraging variance information with pessimism. *arXiv preprint arXiv:2203.05804*, 2022.

Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. MOPO: Model-based offline policy optimization. *arXiv preprint arXiv:2005.13239*, 2020.

Tianhe Yu, Aviral Kumar, Rafael Rafailov, Aravind Rajeswaran, Sergey Levine, and Chelsea Finn. COMBO: Conservative offline model-based policy optimization. *arXiv preprint arXiv:2102.08363*, 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*, pp. 12287–12297. PMLR, 2021.

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

Wenhao Zhan, Baihe Huang, Audrey Huang, Nan Jiang, and Jason Lee. Offline reinforcement learning with realizability and single-policy concentrability. In *Conference on Learning Theory*, pp. 2730–2775. PMLR, 2022.

Junyu Zhang, Alec Koppel, Amrit Singh Bedi, Csaba Szepesvari, and Mengdi Wang. Variational policy gradient method for reinforcement learning with general utilities. *arXiv preprint arXiv:2007.02151*, 2020a.

Ruiyi Zhang, Bo Dai, Lihong Li, and Dale Schuurmans. GenDICE: Generalized offline estimation of stationary values. In *International Conference on Learning Representations*, 2020b.

Shantong Zhang, Bo Liu, and Shimon Whiteson. GradientDICE: Rethinking generalized offline estimation of stationary values. *arXiv preprint arXiv:2001.11113*, 2020c.

Xuezhou Zhang, Yiding Chen, Xiaojin Zhu, and Wen Sun. Corruption-robust offline reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 5757–5773. PMLR, 2022.

A RELATED WORK

We covered a number of related works in the introduction and throughout the paper. In this section, we review more literature and compare them with our work.

A.1 CONCENTRABILITY ASSUMPTIONS

The lack of sufficient coverage in the offline dataset is one of the main challenges in offline RL. In RL theory, dataset coverage has often been characterized by concentrability definitions (Munos, 2007; Scherrer, 2014). Earlier works on offline RL imposed all-policy concentrability on the density ratio for all states and actions (Scherrer, 2014; Liu et al., 2019a; Chen & Jiang, 2019; Jiang, 2019; Wang et al., 2019; Liao et al., 2020; Zhang et al., 2020a), with some requiring this ratio to be bounded for every time step (Szepesvári & Munos, 2005; Munos, 2007; Antos et al., 2008; Farahmand et al., 2010; Antos et al., 2007). The works Xie & Jiang (2021); Feng et al. (2019); Uehara et al. (2020) use a slightly milder definition, requiring a bound on a weighted norm of density ratios. The work Xie & Jiang (2021) makes even stronger assumptions such as lower bounded conditionals $\mu(a|s)$ and exploratoriness of state marginals to circumvent Bellman-completeness.

To handle partial coverage, recent algorithms are analyzed based on variants of single-policy concentrability (Rashidinejad et al., 2021). Some variants such as the ones presented in works Uehara & Sun (2021) (model-based) or Xie et al. (2021) (model-free) are more friendly to function approximation as they avoid bounded ratio assumption for all states and actions. However, existing algorithms based on these weaker definitions are either computationally intractable (Uehara & Sun, 2021; Xie et al., 2021) or their statistical rate is suboptimal (Cheng et al., 2022). The most related works are Zhan et al. (2022), which requires two-policy concentrability, and Chen & Jiang (2022), which requires single-policy concentrability on density ratio for all states and actions.

A.2 CONSERVATIVE OFFLINE RL

A series of recent works on offline RL have focused on addressing partial coverage of offline dataset through conservative algorithm design. Broadly speaking, these methods can be broken down into several categories. The first category of methods applies policy constraints, enforcing the learned policy to be close to the behavior policy. Such constraints are applied either explicitly (Fujimoto et al., 2019; Ghasemipour et al., 2020; Jaques et al., 2019; Siegel et al., 2020; Kumar et al., 2019; Wu et al., 2019; Fujimoto & Gu, 2021), implicitly (Peng et al., 2019; Nair et al., 2020), or through importance sampling (Liu et al., 2019b; Swaminathan & Joachims, 2015; Nachum et al., 2019b; Lee et al., 2021; Zhang et al., 2020c;b). Another category involves learning conservative values such as conservative Q-learning (Kumar et al., 2020), fitted Q-iteration with conservative update (Liu et al., 2020), subtracting penalties (Rezaeifar et al., 2022), and critic regularization (Kostrikov et al., 2021). The last category includes model-based methods such as learning pessimistic models (Kidambi et al., 2020), forming penalties using model ensembles (Yu et al., 2020), or incorporating a combination of model and values (Yu et al., 2021).

On the theoretical side, as discussed in the introduction, the majority of works design pessimistic offline RL algorithms that rely on some form of uncertainty quantification (Yin & Wang, 2021; Uehara et al., 2021; Zhang et al., 2022; Yan et al., 2022; Yin et al., 2022; Kumar et al., 2021; Shi & Chi, 2022; Wang et al., 2022). One exception is the work of Zanette et al. (2021) that uses value-function perturbation with actor-critic in linear function approximation setting. Other examples include the recent theoretical works on MIS (Zhan et al., 2022; Chen & Jiang, 2022) and adversarially trained actor-critic (Cheng et al., 2022).

Most related to our work are methods that focus on provable conservative offline RL under general function approximation and partial coverage. Uehara & Sun (2021) proposed a pessimistic model-based algorithm that under a generalization of single-policy concentrability to bounded TV distance ratio, enjoys $1/\sqrt{N}$ rate but is computationally intractable. The work of Xie et al. (2021) presents a pessimistic model-free algorithm under a variant of single-policy concentrability framework that requires a bounded ratio of average Bellman error and Bellman completeness. While the original version of the algorithm achieves the optimal $1/\sqrt{N}$ rate, it is computationally intractable. A practical version of the algorithm is presented and has a suboptimal $1/N^{1/5}$ guarantee. Another related

work by Chen & Jiang (2022) studies MIS combined with value function approximation under $\pi^\star$-concentrability and proves a $1/\sqrt{\text{gap}(Q^\star)N}$ rate, yet the guarantee degrades with $Q^\star$ gap and the algorithm is computationally intractable. Cheng et al. (2022) propose an adversarially trained actor-critic method that enjoys provable $1/N^{1/3}$ rate under concentrability definition of Xie et al. (2021) and Bellman completeness and performs well in offline RL benchmarks when combined with neural networks.

A.3 Other topics

Apart from RL, our work on bandits is related to the selection problem (Hong et al., 2021), though the majority of works in this area are in the online setting. Additionally, in our analysis, we solve a subset of stochastic optimization problems with possibly large or infinite stochastic constraints involving conditional expectations. To our knowledge, finite-sample properties of such stochastic optimization problems have not been addressed (Shapiro et al., 2021) and our work may open up avenues for further research in this area.

B Proofs for multi-armed bandits

B.1 Primal-dual regularized offline multi-armed bandits (PRO-MAB)

Algorithm 5 Primal-dual Regularized Offline Multi-Armed Bandits (PRO-MAB)

1: **Inputs:** Dataset $\mathcal{D} = \{(a_i, r_i)\}_{i=1}^N$, classes $\mathcal{V} = [-B_v, B_v]$ and $\mathcal{W}$, function $f(\cdot)$, parameter α.
2: Find a solution $\hat{w}, \hat{v}$ to the following problem

$$\max_{w\in\mathcal{W}} \min_{v\in\mathcal{V}} \hat{L}_\alpha^{\text{MAB}}(w, v) := \frac{1}{N}\sum_{i=1}^N w(a_i)r_i - \alpha f(w(a_i)) - v(w(a_i) - 1). \tag{15}$$

3: **Return:** $\hat{\pi}(a) = \frac{\hat{w}(a)\mu(a)}{\sum_a \hat{w}(a)\mu(a)}$ if $\sum_a \hat{w}(a)\mu(a) > 0$, and $\frac{1}{|\mathcal{A}|}$ otherwise.

B.2 Proof of Proposition 1

Consider a 2-armed bandit instance with the following reward distributions, data distribution, and function classes.

- *Reward distributions:* The first arm is optimal with deterministic reward and the second arm has a Bernoulli distribution:

$$r(1) = \frac{1}{2} \quad \text{w.p. } 1, \qquad r(2) \sim \text{Bernoulli}(1/3).$$

- *Data distribution:* We consider a scenario where most data are on the optimal arm:

$$\mu(1) = 1 - \frac{2}{N}, \qquad \mu(2) = \frac{2}{N}.$$

  The single-policy concentrability coefficient is thus $C^\star = 1/\mu(1)$ and is finite. Let $N(a)$ denote the number of samples on arm a. We condition on the event that the number of samples on the second arm is between 1 and 5 which occurs with probability larger than $1 - \exp\left(-1^2 \cdot \frac{2}{2}\right) - \exp\left(-2^2 \cdot \frac{2}{2+2}\right) \geq 0.4$ due to Chernoff bound:

$$(1-1)\cdot N\mu(2) < N(2) < (1+2)\cdot N\mu(2) \quad \Rightarrow \quad 1 \leq N(2) \leq 5.$$

- *Function classes:* Assume that $\mathcal{W} = \{(C^\star = (1/\mu(1), 0), (0, B_w)\}$ and $\mathcal{V} = \{1/2\}$. Note that the problem is realizable as $w^\star = (C^\star, 0)$ and $v^\star = r^\star = 1/2$.

Consider the case where all $N(a)$ samples on the second arm observe a reward of 1, which happens with a probability of at least $\frac{1}{3^5}$ as we conditioned on the event that $1 \leq N(2) \leq 5$. We now compute

$\hat{w}$ by solving the empirical objective (15) with $\alpha = 0$. Note that since $|\mathcal{V}| = 1$, it suffices to compute $\hat{w} = \arg\max_{w \in \mathcal{W}} \hat{L}_0(w, v = 1/2)$. Solving this optimization problem selects $\hat{w} = (0, W)$, leading to a policy that picks the second arm with probability one. Therefore, with constant probability of $0.4 \times 1/3^5 > 0.001$, we have

$$J(\pi^\star) - J(\hat{\pi}) = \frac{1}{2} - \frac{1}{3} = \frac{1}{6}.$$

B.3 Proof of Theorem 1

We condition on the event of Lemma 1 which holds with probability at least $1 - \delta$.

Recall from Lemma 2 that

$$L_\alpha(v_\alpha^\star, w_\alpha^\star) - L_\alpha(v_\alpha^\star, \hat{w}) \leq 2\epsilon_{\text{stat}}.$$

Define $\epsilon_{\hat{w},r} := \sum_a w_\alpha^\star(a)\mu(a)r(a) - \hat{w}(a)\mu(a)r(a)$ and $\hat{d} := \sum_{a \in \mathcal{A}} \hat{w}(a)\mu(a)$. The above inequality implies that

$$\epsilon_{\hat{w},r} - \alpha \mathbb{E}_\mu[f(w_\alpha^\star(a)) - f(\hat{w}(a))] + v_\alpha^\star(\hat{d} - 1) \leq 2\epsilon_{\text{stat}} \tag{16}$$

By boundedness of $f(x) \leq B_f$ and writing $v_\alpha^\star = r^\star - \alpha c$ based on Lemma 3, we conclude that

$$\epsilon_{\hat{w},r} + r^\star(\hat{d} - 1) \leq 2\epsilon_{\text{stat}} + 2\alpha B_f + \alpha c(\hat{d} - 1) \leq 2\epsilon_{\text{stat}} + \alpha(2B_f + f'(C^\star)B_w), \tag{17}$$

where $0 \leq c \leq f'(C^\star)$.

By Lemma 5, with $\alpha = \epsilon_{\text{stat}} \log N$, we have $\hat{d} \geq 1 - 1/\log N \geq \frac{1}{2} > 0$ for $N \geq 8$. Recall that the learned policy $\hat{\pi}$ is computed from $\hat{w}$ according to

$$\hat{\pi}(a) = \frac{\hat{w}(a)\mu(a)}{\sum_{a \in \mathcal{A}} \hat{w}(a)\mu(a)} = \frac{\hat{w}(a)\mu(a)}{\hat{d}}.$$

With simple manipulations, we find the following expression for the suboptimality of $\hat{\pi}$ with respect to $\pi_\alpha^\star$:

$$\begin{aligned}
J(\pi_\alpha^\star) - J(\hat{\pi}) &= \sum_a w_\alpha^\star(a)\mu(a)r(a) - \frac{1}{\hat{d}}\hat{w}(a)\mu(a)r(a) \\
&= \sum_a w_\alpha^\star(a)\mu(a)r(a) - \hat{w}(a)\mu(a)r(a) + \sum_a \left(1 - \frac{1}{\hat{d}}\right)\hat{w}(a)\mu(a)r(a) \\
&= \epsilon_{\hat{w},r} + \left(\hat{d} - 1\right) \sum_a \frac{1}{\hat{d}}\hat{w}(a)\mu(a)r(a) \\
&= \epsilon_{\hat{w},r} + \left(\hat{d} - 1\right) J(\hat{\pi}) \\
&= \epsilon_{\hat{w},r} + \left(\hat{d} - 1\right) J(\pi_\alpha^\star) - \left(\hat{d} - 1\right)\left[J(\pi_\alpha^\star) - J(\hat{\pi})\right].
\end{aligned}$$

Therefore, suboptimality $J(\pi_\alpha^\star) - J(\hat{\pi})$ can be expressed as

$$\begin{aligned}
J(\pi_\alpha^\star) - J(\hat{\pi}) &= \frac{1}{\hat{d}}\left(\epsilon_{\hat{w},r} + \left(\hat{d} - 1\right) J(\pi_\alpha^\star)\right) \\
&= \frac{1}{\hat{d}}\left(\epsilon_{\hat{w},r} + \left(\hat{d} - 1\right)(r^\star - \epsilon_{\text{reg}})\right) \\
&\leq \frac{1}{\hat{d}}\left(\epsilon_{\hat{w},r} + \left(\hat{d} - 1\right) r^\star\right) - \frac{1}{\hat{d}}\left(\hat{d} - 1\right)\epsilon_{\text{reg}}.
\end{aligned}$$

Here, $\epsilon_{\text{reg}} = r^\star - J(\pi_\alpha^\star)$ is the suboptimality due to regularization. We use the above inequality to bound the suboptimality with respect to the optimal policy:

$$\begin{aligned}
J(\pi^\star) - J(\hat{\pi}) &= J(\pi^\star) - J(\pi_\alpha^\star) + J(\pi_\alpha^\star) - J(\hat{\pi}) \\
&= \epsilon_{\text{reg}} + J(\pi_\alpha^\star) - J(\hat{\pi}) \\
&\leq \epsilon_{\text{reg}} + \frac{1}{\hat{d}}\left(\epsilon_{\hat{w},r} + \left(\hat{d} - 1\right) r^\star\right) - \frac{1}{\hat{d}}\left(\hat{d} - 1\right)\epsilon_{\text{reg}} \\
&\leq \frac{1}{\hat{d}}\left(\epsilon_{\hat{w},r} + \left(\hat{d} - 1\right) r^\star\right) + \frac{1}{\hat{d}}\epsilon_{\text{reg}}.
\end{aligned}$$

Recall that we have $1/\hat{d} \leq 2$. Moreover, by Proposition 4 we have $\epsilon_{\text{reg}} \leq \alpha B_f$. Therefore,

$$\begin{aligned} J(\pi^\star) - J(\hat{\pi}) &\leq \frac{1}{\hat{d}}\left(\epsilon_{\hat{w},r} + \left(\hat{d}-1\right) r^\star\right) + \frac{1}{\hat{d}}\epsilon_{\text{reg}} \\ &\leq 2\left(\epsilon_{\hat{w},r} + \left(\hat{d}-1\right) r^\star\right) + 2\alpha B_f \\ &\leq 4\epsilon_{\text{stat}} + \alpha(4B_f + 2f'(C^\star)B_w)) + 2\alpha B_f \\ &\lesssim \alpha(B_f + f'(C^\star)B_w). \end{aligned}$$

where the second to last inequality relies on the bound derived in (17).

B.4 Proof of Proposition 2

Note that $w^\star = w_0^\star$ due to strong duality. Consider the difference between population objective with $\alpha = 0$ at $w^\star$ and $\hat{w}$, which is bounded by Lemma 2:

$$L(w^\star, v^\star) - L(\hat{w}, v^\star) = \mathbb{E}_{a\sim\mu}[r(a)(w^\star(a) - \hat{w}(a))] - v^\star \mathbb{E}_{a\sim\mu}[w^\star(a) - \hat{w}(a)] \lesssim \epsilon_{\text{stat}}. \tag{18}$$

We have $\mathbb{E}_{a\sim\mu}[w^\star(a)] = 1$ due to realizability and $\mathbb{E}_{a\sim\mu}[\hat{w}(a)] = 1$ is our assumption. Thus the second term in (18) is zero. Moreover, note that $\hat{\pi}(a) = \hat{w}(a)\mu(a)/\mathbb{E}_{a\sim\mu}[w(a)] = \hat{w}(a)$. Substituting the expression for ϵ_{stat} from (19), we have

$$J(\pi^\star) - J(\hat{\pi}) = \mathbb{E}_{a\sim\mu}[r(a)(w^\star(a) - \hat{w}(a))] \lesssim (B_w(B_v+1) + \alpha B_f)\sqrt{\frac{\log|\mathcal{V}||\mathcal{W}|/\delta}{N}}.$$

B.5 Proof of Theorem 2

The result immediately follows from Theorem 3.

B.6 Auxiliary Lemmas

Lemma 1 (Closeness of empirical and population objectives, MAB) *Fix the failure probability δ and define the following bounds $|w(a)| \leq B_w, |v| \leq B_v, |f(w(a))| \leq B_f$. For the population objective $L_\alpha(w,v)$ defined in (2) and empirical objective $\hat{L}_\alpha(w,v)$ defined in (15), the following holds with probability $1-\delta$ and for all $v \in \mathcal{V}$ and $w \in \mathcal{W}$*

$$|L_\alpha(w,v) - \hat{L}_\alpha(w,v)| \lesssim ((B_w+1)(B_v+1) + \alpha B_f)\sqrt{\frac{\log|\mathcal{V}||\mathcal{W}|/\delta}{N}} := \epsilon_{stat}^{MAB} \tag{19}$$

Proof. $\hat{L}_\alpha(w,v)$ is an empirical average over independent and bounded random variables, where the bound is computed as

$$\begin{aligned} |w(a_i)r_i - \alpha f(w(a_i)) - v(w(a_i)-1)| &\leq B_w + \alpha B_f + B_v(B_w+1) \\ &\leq (B_w+1)(B_v+1) + \alpha B_f. \end{aligned}$$

It is easy to see that $\mathbb{E}_{\mathcal{D}}[\hat{L}_\alpha(w,v)] = L_\alpha(w,v)$, where the expectation is taken with respect to the randomness in dataset $\mathcal{D}$. The claim is proved by applying Hoeffding's inequality along with a union bound on w and v. □

Lemma 2 (Closeness of objectives at $\hat{w}$ and $w_\alpha^\star$, MAB) *Fix failure probability δ. With at least probability $1-\delta$ and for any α, we have*

$$L_\alpha(w_\alpha^\star, v_\alpha^\star) - L_\alpha(\hat{w}, v_\alpha^\star) \leq 2\epsilon_{stat},$$

where ϵ_{stat} is defined in (19).

Proof. The proof is similar to Lemma 7 of Zhan et al. (2022) and relies on decomposing the objective difference and using the fact that $w_\alpha^\star, v_\alpha^\star, \hat{w}, \hat{v}$ respectively correspond to the saddle points of L_α and

$\hat{L}_\alpha$. We write

$$L_\alpha(w_\alpha^\star, v_\alpha^\star) - L_\alpha(\hat{w}, v_\alpha^\star) = \underbrace{L_\alpha(w_\alpha^\star, v_\alpha^\star) - L_\alpha(w_\alpha^\star, \hat{v}(w_\alpha^\star))}_{:=T_1} + \underbrace{L_\alpha(w_\alpha^\star, \hat{v}(w_\alpha^\star)) - \hat{L}_\alpha(w_\alpha^\star, \hat{v}(w_\alpha^\star))}_{:=T_2}$$
$$+ \underbrace{\hat{L}_\alpha(w_\alpha^\star, \hat{v}(w_\alpha^\star)) - \hat{L}_\alpha(\hat{w}, \hat{v})}_{:=T_3} + \underbrace{\hat{L}_\alpha(\hat{w}, \hat{v}) - \hat{L}_\alpha(\hat{w}, v_\alpha^\star)}_{:=T_4}$$
$$+ \underbrace{\hat{L}_\alpha(\hat{w}, v_\alpha^\star) - L_\alpha(\hat{w}, v_\alpha^\star)}_{:=T_5},$$

where $\hat{v}(w) = \arg\min_{v \in \mathcal{V}} \hat{L}_\alpha(w, v)$. Each term is bounded as follows:

- $T_1 \leq 0$ because $v_\alpha^\star = \arg\min_v L_\alpha(w_\alpha^\star, v)$.
- $T_2 \leq \epsilon_{\text{stat}}$ due to Lemma 1.
- $T_3 \leq 0$ because $\hat{w} = \arg\max_{w \in \mathcal{W}} \hat{L}_\alpha(\hat{v}(w), w)$.
- $T_4 \leq 0$ because $\hat{v} = \arg\min_{v \in \mathcal{V}} \hat{L}_\alpha(v, \hat{w})$.
- $T_5 \leq \epsilon_{\text{stat}}$ due to Lemma 1.

Summing up the bounds on each term yields the desired bound. □

Lemma 3 (Closeness of regularized and unregularized primal-dual solutions, MAB) *Let $r^\star = \max_a r(a)$ and $v_\alpha^\star$ denote the optimal solution to the population objective (2). Then,*

$$v_\alpha^\star = r^\star - c\alpha,$$

where $0 \leq c \leq f'(C^\star)$.

Proof. KKT conditions imply the following relation between $w_\alpha^\star(a)$ and $v_\alpha^\star$:

$$w_\alpha^\star(a) = \max\left\{0, (f')^{-1}\left(\frac{r(a) - v_\alpha^\star}{\alpha}\right)\right\}. \tag{20}$$

Since f is strictly convex, f' is a monotonically increasing function. Therefore, the optimal arm $a^\star$ has the largest $w_\alpha^\star(a)$, which should be nonzero due to realizability of $w_\alpha^\star$. In other words,

$$w_\alpha^\star(a^\star) = (f')^{-1}\left(\frac{r^\star - v_\alpha^\star(s)}{\alpha}\right) \Rightarrow v_\alpha^\star = r^\star - \alpha f'(w_\alpha^\star(a^\star)). \tag{21}$$

We now proceed to find a bound on $f'(w_\alpha^\star(a^\star))$. The convexity of f immediately gives an upper bound on $v_\alpha^\star$:

$$v_\alpha^\star \leq r^\star. \tag{22}$$

By realizability of $w_\alpha^\star$, we have

$$\sum_{a \in \mathcal{A}} \mu(a) w_\alpha^\star(a) = 1 \Rightarrow w_\alpha^\star(a^\star) \leq \frac{1}{\mu(a^\star)} \leq C^\star,$$

where the last inequality is due to the single-policy concentrability assumption of π^* in multi-armed bandits. Since f' is an increasing function, we have $f'(w_\alpha^\star(a^\star)) \leq f'(C^\star)$, which combined with (21) yields the claimed lower bound on $v_\alpha^\star$

$$v_\alpha^\star \geq r^\star - \alpha f'(C^\star).$$

□

Proposition 4 (Suboptimality due to regularization, MAB) *Let f be non-negative and bounded by B_f. The policy yielded by the regularized importance weights $w_\alpha^\star$ satisfies the following suboptimality bound*

$$J(\pi^\star) - J(\pi_\alpha^\star) \leq \alpha B_f.$$

Proof. First, note that $w^\star = w_0^\star$ by strong duality. Since $w_\alpha^\star$ is the optimal solution to the regularized population objective (2), by strong duality, we have

$$\mathbb{E}_{a\sim d_\alpha^\star}[r(a)] - \alpha\mathbb{E}_{a\sim\mu}[f(w_\alpha^\star(a))] \geq \mathbb{E}_{a\sim d^\star}[r(a)] - \alpha\mathbb{E}_{a\sim\mu}[f(w_\alpha^\star(a))]$$

where $d_\alpha^\star(a) = \mu(a)w_\alpha^\star(a)$.

Therefore, the suboptimality of $\pi_\alpha^\star$ can be bounded as follows

$$\begin{aligned} J(\pi^\star) - J(\pi_\alpha^\star) &= \mathbb{E}_{a\sim d^\star}[r(a)] - \mathbb{E}_{a\sim d_\alpha^\star}[r(a)] \\ &\leq \alpha\mathbb{E}_{a\sim\mu(\cdot)}[f(w_\alpha^\star(a))] - \alpha\mathbb{E}_{a\sim\mu}[f(w_\alpha^\star(a))] \\ &\leq \alpha\mathbb{E}_{a\sim\mu(\cdot)}[f(w_\alpha^\star(a))] \\ &\leq \alpha B_f, \end{aligned}$$

where in the second to last inequality we used the non-negativity of f and in the last equality, we used the boundedness of f. □

Proposition 5 (Slow convergence rate for $\hat{d}$, MAB) *Let f be an M_f-strongly-convex function and ϵ_{stat} be the statistical error defined in (19). Then, with probability at least $1-\delta$, $\hat{d} := \sum_{a\in\mathcal{A}} \hat{w}(a)\mu(a)$ satisfies the following*

$$\left|\hat{d} - 1\right| \lesssim \sqrt{\frac{\epsilon_{stat}}{\alpha M_f}}.$$

Proof. This proposition is a direct consequence of Lemma 8 in Zhan et al. (2022). For completeness, we present a simplified proof for the multi-armed bandit setting.

First, observe that since f is M_f-strongly-convex, the function $L_\alpha(v_\alpha^\star, w)$ is αM_f-strongly-concave with respect to w and $\|\cdot\|_{2,\mu}$. Furthermore, since $w_\alpha^\star = \arg\max_w L_\alpha(v^\star, w)$, we have

$$\|\hat{w} - w_\alpha^\star\|_{2,\mu} \leq \sqrt{\frac{2(L_\alpha(w_\alpha^\star, v_\alpha^\star) - L_\alpha(\hat{w}, v_\alpha^\star))}{\alpha M_f}}.$$

The above bound along with the result of Lemma 2 gives the following slow convergence rate for $|\hat{d} - 1|$

$$\left|\hat{d} - 1\right| = \left|\sum_a \hat{w}(a)\mu(a) - \sum_a w_\alpha^\star(a)\mu(a)\right| \leq \|\hat{w} - w_\alpha^\star\|_{1,\mu} \leq \|\hat{w} - w_\alpha^\star\|_{2,\mu} \leq \sqrt{\frac{4\epsilon_{\text{stat}}}{\alpha M_f}}.$$

□

C Proofs for contextual bandits

C.1 Primal-dual regularized offline contextual bandits (PRO-CB)

Define importance weights $w(s,a) = d(s,a)/\mu(s,a)$ to denote the ratio of occupancy and data distribution. The primal-dual regularized approach (Zhan et al., 2022) solves the following population objective

$$\max_{w\geq 0}\min_v L_\alpha(w,v) := \mathbb{E}_{s,a\sim\mu}\left[w(s,a)r(s,a)\right] - \alpha\mathbb{E}_{s,a\sim\mu}\left[f\left(w(s,a)\right)\right] - \mathbb{E}_{s,a\sim\mu}[v(s)(w(s,a)-1)], \tag{23}$$

Solving the empirical version of the above objective in a similar fashion to Algorithm 5 leads to the PRO-CB given in Algorithm 6.

C.2 Proof of Proposition 3

For all $f\colon \mathbb{R}\mapsto\mathbb{R}$ that is M_f-strongly convex, non-negative, and bounded, we show that for any $\alpha > 0$, there exists a contextual bandit instance such that with constant probability, Algorithm 2 with regularization function f and regularization parameter α outputs policy $\hat{\pi}$ with suboptimality

$$J(\pi^\star) - J(\hat{\pi}) \geq \Omega\left(N^\beta\right), \text{ where } \beta > -\frac{1}{2}. \tag{25}$$

Based on the magnitude of α in terms of N, we divide our proof into two cases.

Algorithm 6 Primal-dual Regularized Offline Contextual Bandits (PRO-CB)

1: **Inputs:** Dataset $\mathcal{D} = \{(s_i, a_i, r_i)\}_{i=1}^N$, function classes $\mathcal{W}, \mathcal{V}$, function $f(\cdot)$, parameter α
2: Find a solution $\hat{w}, \hat{v}$ to the following problem

$$\max_{w\in\mathcal{W}} \min_{v\in\mathcal{V}} \hat{L}_\alpha(w, v) := \frac{1}{N}\sum_{i=1}^N w(s_i, a_i)r_i - \alpha f(w(s_i, a_i)) - v(s_i)(w(s_i, a_i) - 1). \tag{24}$$

3: **Return:** $\hat{\pi}(a|s) = \frac{\hat{w}(s,a)\mu(a|s)}{\sum_a \hat{w}(s,a)\mu(a|s)}$ if $\sum_a \hat{w}(s, a)\mu(a|s) > 0$, and $\frac{1}{|\mathcal{A}|}$ otherwise.

Case 1: Large α. If there exists $-\frac{1}{2} < \beta < 0$ such that $\alpha \geq N^\beta$, then consider the following single-state two-arm contextual bandit (equivalently multi-armed bandit) instance:

- *Reward distribution*: Both arms have deterministic rewards and the suboptimal arm has a constant value gap:

$$r(1) \equiv 1, \quad r(2) \equiv \frac{1}{2}.$$

- *Data distribution*: Most density is concentrated on the suboptimal arm:

$$\mu(1) = \frac{1}{N}, \quad \mu(2) = 1 - \frac{1}{N}.$$

- *Function classes*: We assume both $\mathcal{W}$ and $\mathcal{V}$ are singletons, i.e., only contain the optimal function under regularization:

$$\mathcal{W} = \{w_\alpha^\star\}, \quad \mathcal{V} = \{v_\alpha^\star\}.$$

In this particular MAB instance, it suffices to show that the policy $\pi_\alpha^\star$ induced by $w_\alpha^\star$ has $\Omega(N^\beta)$ suboptimality compared to the unregularized optimal policy $\pi^\star$, i.e.,

$$J(\pi^\star) - J(\pi_\alpha^\star) = \pi_\alpha^\star(2) \cdot (r(1) - r(2)) \geq \Omega(N^\beta). \tag{26}$$

To establish (26), it suffices to show $\pi_\alpha^\star(2) = \mu(2)w_\alpha^\star(2) > N^\beta$. We prove this by contradiction. If

$$\mu(2)w_\alpha^\star(2) \leq N^\beta, \tag{27}$$

then by KKT conditions and the fact that f' is strictly increasing since f is strictly convex, we have

$$w_\alpha^\star(2) \geq (f')^{-1}\left(\frac{r(2) - v_\alpha^\star}{\alpha}\right) \Rightarrow v_\alpha^\star \geq r(2) - \alpha f'(w_\alpha^\star(2)) \geq \frac{1}{2} - \alpha f'\left(\frac{N^\beta}{\mu(2)}\right).$$

Combining the above bound on $v_\alpha^\star$ with the KKT condition on $w_\alpha^\star(1)$, we then obtain

$$w_\alpha^\star(1) = (f')^{-1}\left(\frac{r(1) - v_\alpha^\star}{\alpha}\right) \leq (f')^{-1}\left(\frac{1}{2\alpha} + f'\left(\frac{N^\beta}{\mu(2)}\right)\right). \tag{28}$$

Moreover, since the regularization function f is M_f-strongly convex, we have

$$\begin{aligned} f'\left(\frac{1 - N^\beta}{\mu(1)}\right) - f'\left(\frac{N^\beta}{\mu(2)}\right) \geq & M_f\left(\frac{1 - N^\beta}{\mu(1)} - \frac{N^\beta}{\mu(2)}\right) \\ \geq & M_f\left(\frac{1 - N^\beta}{\frac{1}{N}} - \frac{N^\beta}{1 - \frac{1}{N}}\right) \geq M_f\left(N - N^{1+\beta} - 2N^\beta\right) \\ > & \frac{N^{-\beta}}{2} \geq \frac{1}{2\alpha}, \quad \text{for large enough } N. \end{aligned}$$

Therefore, we can continue to upper bound the RHS of eq. (28):

$$w_\alpha^\star(1) \leq (f')^{-1}\left(\frac{1}{2\alpha} + f'\left(\frac{N^\beta}{\mu(2)}\right)\right) < (f')^{-1}\left(f'\left(\frac{1 - N^\beta}{\mu(1)}\right)\right) = \frac{1 - N^\beta}{\mu(1)},$$

which further implies that

$$w_\alpha^\star(1)\mu(1) < 1 - N^\beta \underbrace{\leq}_{\text{by (27)}} 1 - w_\alpha^\star(2)\mu(2) \;\Rightarrow\; \sum_a w_\alpha^\star(a)\mu(a) < 1. \tag{29}$$

Note that (29) contradicts with the fact that $(w_\alpha^\star, v_\alpha^\star)$ is the optimal min-max solution of L_α^{MAB} because it violates the constraint $\mathbb{E}_\mu[w(a)] = 1$. Therefore, (27) should not hold at the first place, and we must have

$$J(\pi^\star) - J(\pi_\alpha^\star) = \frac{1}{2}\mu(2)w_\alpha^\star(2) > \frac{1}{2}N^\beta = \Omega(N^\beta), \tag{30}$$

thus completing the proof for large α.

Case 2: Small α. Now suppose $\alpha \leq \widetilde{O}(N^{-\frac{1}{2}})$. In this case, we consider the following 2-state 2-arm contextual bandit instance:

- *State and reward distributions*: State 1 has a very small probability mass. For state 1, the first arm is optimal with a Bernoulli-distributed reward and the second arm is suboptimal with a deterministic reward. For state 2, both arms have deterministic rewards. Importantly, state 1 has a constant value gaps in its suboptimal action a_2.

$$\begin{aligned}
&\rho(s_1) = N^{-\frac{1}{4}}, \quad r(s_1, a_1) \sim \text{Bernoulli}(\frac{1}{2}),\ r(s_1, a_2) \equiv \frac{1}{3};\\
&\rho(s_2) = 1 - N^{-\frac{1}{4}}, \quad r(s_2, a_1) \equiv \frac{1}{2},\ r(s_2, a_2) \equiv \frac{1}{3}.
\end{aligned}$$

- *Data distribution*:

$$\begin{aligned}
&\mu(s_i) = \rho(s_i),\ i = 1, 2.\\
&\mu(a_1|s_1) = \mu(a_1|s_2) = 1 - \frac{2}{N}, \quad \mu(a_2|s_1) = \mu(a_2|s_2) = \frac{2}{N}.
\end{aligned}$$

- *Function classes*: Let w be defined as $w(s_2, a_i) = w_\alpha^\star(s_2, a_i)$ and $w(s_1, a_i) = 0$ for $i = 1, 2$. Consider the following function classes $\mathcal{W}$ and $\mathcal{V}$:

$$\mathcal{W} = \{w_\alpha^\star, w\},\ \mathcal{V} = \{v_\alpha^\star\}.$$

We first show that when $\alpha \leq \widetilde{O}(N^{-\frac{1}{2}})$ and N is sufficiently large, the regularized optimal policy is the same as the unregularized optimal policy, i.e., $w_\alpha^\star = w^\star$. Since the optimization problem is independent across states at a population level, Lemma 3 implies that

$$\begin{aligned}
&v_\alpha^\star(s) = r^\star(s) - c(s)\alpha, \text{ and}\\
&w_\alpha^\star(s, a) = \max\left\{0, (f')^{-1}\left(\frac{r(s,a) - v_\alpha^\star(s)}{\alpha}\right)\right\} = \max\left\{0, (f')^{-1}\left(c(s) - \frac{r^\star(s) - r(s,a)}{\alpha}\right)\right\},
\end{aligned}$$

where $0 \leq c(s) \leq f'(C^\star)$ for $s \in \{s_1, s_2\}$. Since $r^\star(s) - r(s, a_2) = \Theta(1)$ for both states, we have $w_\alpha^\star(s, a_2) = 0$ for the suboptimal arm a_2. Thus $w_\alpha^\star(s) = w^\star(s) = \frac{1}{\mu(a_1|s)}$. Correspondingly, we can compute $v_\alpha^\star(s) = r^\star(s) - \alpha f'(\frac{1}{\mu(a_1|s)})$.

Now we compare the value of $\hat{L}_\alpha(w, v_\alpha^\star)$ evaluated at the two hypotheses $w \in \mathcal{W}$. Let $\hat{\mu}$ denote the empirical state-arm distribution and $\hat{r}$ denote the empirical mean reward, then we have

$$\begin{aligned}
&\hat{L}_\alpha(w, v_\alpha^\star) - \hat{L}_\alpha(w_\alpha^\star, v_\alpha^\star) \\
&= -\hat{\mu}(s_1)\Bigg[\hat{r}(s_1,a_1)\hat{\mu}(a_1|s_1)w_\alpha^\star(s_1,a_1) + \alpha\hat{\mu}(a_1|s_1)f(w_\alpha^\star(s_1,a_1)) \\
&\qquad + v_\alpha^\star(s_1)\left(\sum_a \hat{\mu}(a|s_1)\left(w(s_1,a) - w_\alpha^\star(s_1,a)\right)\right)\Bigg] \\
&= -\hat{\mu}(s_1)\left[\left(\hat{r}(s_1,a_1) - r(s_1,a_1) + \alpha f'\left(\frac{1}{\mu(a_1|s_1)}\right)\right)\frac{\hat{\mu}(a_1|s_1)}{\mu(a_1|s_1)} + \alpha\hat{\mu}(a_1|s_1)f\left(\frac{1}{\mu(a_1|s_1)}\right)\right] \\
&= -\hat{\mu}(s_1,a_1)\left[\frac{\hat{r}(s_1,a_1) - r(s_1,a_1)}{\mu(a_1|s_1)} + \alpha\cdot\underbrace{\left(\frac{1}{\mu(a_1|s_1)}f'\left(\frac{1}{\mu(a_1|s_1)}\right) + f\left(\frac{1}{\mu(a_1|s_1)}\right)\right)}_{\Theta(1)}\right]
\end{aligned} \tag{31}$$

Therefore, to prove that the suboptimal function w is picked frequently by the algorithm, it suffices to show that with a constant probability

$$\left|\frac{\hat{r}(s_1,a_1) - r(s_1,a_1)}{\mu(a_1|s_1)}\right| > \omega(\alpha), \tag{32}$$

because this condition ensures that $\hat{L}_\alpha(w, v_\alpha^\star) - \hat{L}_\alpha(w_\alpha^\star, v_\alpha^\star) = (31) > 0$ and therefore $\hat{w} = w$,

$$J(\pi^\star) - J(\hat{\pi}) = J(\pi_\alpha^\star) - J(\pi_w) = \mu(s_1)\pi_w(a_2|s_1)(r(s_1,a_1) - r(s_1,a_2)) = \frac{N^{-\frac{1}{4}}}{12} \geq \Omega(N^\beta)$$

for $\beta = -\frac{1}{4} > -\frac{1}{2}$, as desired.

Finally, we go back to establish eq. (32). Since $\mu(a_1|s_1) = \Theta(1)$ and $\alpha \leq \widetilde{O}(N^{-\frac{1}{2}})$, it suffices to show that $|\hat{r}(s_1,a_1) - r(s_1,a_1)| > \widetilde{\omega}(N^{-\frac{1}{2}})$ with constant probability. Similar to the proof of proposition 1, we condition on the event $\frac{1}{2}N\mu(s_1,a_1) \leq N(s_1,a_1) \leq 2N\mu(s_1,a_1)$ which happens with probability ≥ 0.9. Under this event, the tightness of the Chernoff bound implies that the estimation error $|\hat{r}(s_1,a_1) - r(s_1,a_1)|$ is of order $\Theta(N(s_1,a_1)^{-\frac{1}{2}}) = \Theta((N\cdot N^{-\frac{1}{4}})^{-\frac{1}{2}}) = \Theta(N^{-\frac{3}{8}}) > \widetilde{\omega}(N^{-\frac{1}{2}})$. The proof is thus complete.

C.3 Proof of Theorem 3

Proof outline. Our suboptimality analysis can be broken down into two steps. We first show that the result of Lemma 5 $L_{\text{AL}}^{\text{CB}}(w^\star, v^\star) - L_{\text{AL}}^{\text{CB}}(\hat{w}, v^\star) \lesssim \epsilon_{\text{stat}}^{\text{CB}}$ implies that

$$\sum_s \hat{d}(s)(V^\star(s) - V^{\hat{\pi}}(s)) \lesssim \epsilon_{\text{stat}}^{\text{CB}},$$

where

$$\hat{d}(s) \triangleq d_{\hat{w}}(s) = \sum_a \hat{w}(s,a)\mu(s,a).$$

Then, we use Lemma 6 which states that $\hat{d}(s)$ is not too small compared to $\mu(s) = \rho(s)$ for significant states (due to the augmented Lagrangian term) to conclude that suboptimality is small:

$$J(\pi^\star) - J(\hat{\pi}) = \sum_s \rho(s)(V^\star(s) - V^{\hat{\pi}}(s)) \lesssim \epsilon_{\text{stat}}^{\text{CB}}.$$

Proof. Recall the definitions of optimal and learned policies

$$\pi^\star(a|s) = w^\star(s,a)\mu(a|s), \quad \hat{\pi}(a|s) = \frac{\hat{w}(s,a)\mu(a|s)}{\frac{\hat{d}}{\mu}(s)},$$

where $\frac{\hat{d}}{\mu}(s) \triangleq \sum_a \hat{w}(s,a)\mu(a|s)$. Define

$$\epsilon_{\hat{w},r} := \sum_{s,a} w^\star(s,a)\mu(s,a)r(s,a) - \hat{w}(s,a)\mu(s,a)r(s,a),$$
$$\hat{d}(s) := \sum_a \hat{w}(s,a)\mu(s,a).$$

Step 1: Bounding $\sum_s \hat{d}(s)(V^\star(s) - V^{\hat{\pi}}(s))$. Using the above notation, we write $L_{\text{AL}}^{\text{CB}}(w^\star, v^\star) - L_{\text{AL}}^{\text{CB}}(\hat{w}, v^\star)$ as

$$\begin{aligned}
&L_{\text{AL}}^{\text{CB}}(w^\star, v^\star) - L_{\text{AL}}^{\text{CB}}(\hat{w}, v^\star) \\
=&\sum_s \mu(s,a)r(s,a)w^\star(s,a) - \sum_{s,a} \mu(s,a)r(s,a)\hat{w}(s,a) \\
&- \sum_s v^\star(s)\left(\mu(s) - \sum_a \mu(s,a)\hat{w}(s,a)\right) + \mathbb{E}_{s\sim\mu}\left[\left(\frac{\hat{d}}{\mu}(s) - 1\right)^2\right] \\
=&\epsilon_{\hat{w},r} - \sum_s (\mu(s) - \hat{d}(s))V^\star(s) + \mathbb{E}_{s\sim\mu}\left[\left(\frac{\hat{d}}{\mu}(s) - 1\right)^2\right].
\end{aligned}$$

Recall that by Lemma 5, $L_{\text{AL}}^{\text{CB}}(w^\star, v^\star) - L_{\text{AL}}^{\text{CB}}(\hat{w}, v^\star) \lesssim \epsilon_{\text{stat}}^{\text{CB}}$. Therefore,

$$\begin{aligned}
\epsilon_{\text{stat}}^{\text{CB}} \gtrsim& \epsilon_{\hat{w},r} - \sum_s (\mu(s) - \hat{d}(s))V^\star(s) + \mathbb{E}_{s\sim\mu}\left[\left(\frac{\hat{d}}{\mu}(s) - 1\right)^2\right] \\
\gtrsim& \epsilon_{\hat{w},r} - \sum_s (\mu(s) - \hat{d}(s))V^\star(s) \\
\implies \epsilon_{\hat{w},r} \lesssim& \epsilon_{\text{stat}}^{\text{CB}} + \sum_s (\mu(s) - \hat{d}(s))V^\star(s)
\end{aligned}$$

We now bound the suboptimality, noting that in CBs $\rho(s) = \mu(s)$,

$$\begin{aligned}
J(\pi^\star) - J(\hat{\pi}) =& \sum_s \mu(s)\left(V^{\pi^\star}(s) - V^{\hat{\pi}}(s)\right) \\
=& \underbrace{\sum_{s:\hat{d}(s)\neq 0} \mu(s)\left(V^{\pi^\star}(s) - V^{\hat{\pi}}(s)\right)}_{T_1} + \underbrace{\sum_{s:\hat{d}(s)=0} \mu(s)\left(V^{\pi^\star}(s) - V^{\hat{\pi}}(s)\right)}_{T_2}
\end{aligned}$$

Since $\hat{\pi}(a|s) = \hat{w}(s,a)\mu(s,a)/\hat{d}(s)$ for $\hat{d}(s) \neq 0$, we have

$$\begin{aligned}
T_1 =& \sum_{s,a:\hat{d}(s)\neq 0} \mu(s,a)r(s,a)w^\star(s,a) - \sum_{s,a:\hat{d}(s)\neq 0} \mu(s)V^{\hat{\pi}}(s) \\
=& \sum_{s,a:\hat{d}(s)\neq 0} \mu(s,a)r(s,a)\left(w^\star(s,a) - \hat{w}(s,a)\right) \\
&+ \sum_{s,a:\hat{d}(s)\neq 0} \hat{d}(s)r(s,a)\frac{\mu(s,a)\hat{w}(s,a)}{\hat{d}(s)} - \sum_{s,a:\hat{d}(s)\neq 0} \mu(s)V^{\hat{\pi}}(s) \\
=& \sum_{s,a:\hat{d}(s)\neq 0} \mu(s,a)r(s,a)\left(w^\star(s,a) - \hat{w}(s,a)\right) + \sum_{s,a:\hat{d}(s)\neq 0} (\hat{d}(s) - \mu(s))V^{\hat{\pi}}(s).
\end{aligned}$$

Also, since $\hat{w}(s,a)\mu(s,a) = 0$ for $\hat{d}(s) = 0$, we have

$$\begin{aligned} T_2 &= \sum_{s,a:\hat{d}(s)=0} \mu(s,a) r(s,a) w^\star(s,a) - \sum_{s,a:\hat{d}(s)=0} \mu(s) V^{\hat{\pi}}(s) \\ &= \sum_{s,a:\hat{d}(s)=0} \mu(s,a) r(s,a) \left(w^\star(s,a) - \hat{w}(s,a)\right) + \sum_{s,a:\hat{d}(s)=0} (\hat{d}(s) - \mu(s)) V^{\hat{\pi}}(s). \end{aligned}$$

Therefore,

$$\begin{aligned} J(\pi^\star) - J(\hat{\pi}) =& T_1 + T_2 \\ =& \epsilon_{\hat{w},r} + \sum_s V^{\hat{\pi}}(s) \left(\hat{d}(s) - \mu(s)\right) \\ \lesssim& \epsilon_{\text{stat}}^{\text{CB}} + \sum_s (\mu(s) - \hat{d}(s)) V^\star(s) + \sum_s (\hat{d}(s) - \mu(s)) V^{\hat{\pi}}(s). \end{aligned} \tag{33}$$

Notice that we have concluded

$$\sum_s \mu(s) V^\star(s) - \sum_s \mu(s) V^{\hat{\pi}}(s) \lesssim \epsilon_{\text{stat}}^{\text{CB}} + \sum_s (\mu(s) - \hat{d}(s)) V^\star(s) + \sum_s (\hat{d}(s) - \mu(s)) V^{\hat{\pi}}(s),$$

in other words, we have

$$\sum_s \hat{d}(s)(V^\star(s) - V^{\hat{\pi}}(s)) \lesssim \epsilon_{\text{stat}}^{\text{CB}}. \tag{34}$$

Step 2: Bounding $\sum_s \rho(s)(V^\star(s) - V^{\hat{\pi}}(s))$. Consider the following partitioning of the state space:

$$\mathcal{S}_{\text{s}} := \{s \mid \hat{d}(s) \leq \mu(s)/2\}, \quad \mathcal{S} \backslash \mathcal{S}_{\text{s}}.$$

We use this state space partition to decompose policy suboptimality

$$\begin{aligned} \sum_s \mu(s) V^\star(s) - \sum_s \mu(s) V^{\hat{\pi}}(s) &= \sum_{s \in \mathcal{S}_{\text{s}}} \mu(s)(V^\star(s) - V^{\hat{\pi}}(s)) + \sum_{s \notin \mathcal{S}_{\text{s}}} \mu(s)(V^\star(s) - V^{\hat{\pi}}(s)), \\ &\leq \epsilon_{\text{stat}}^{\text{CB}} + \sum_{s \notin \mathcal{S}_{\text{s}}} \mu(s)(V^\star(s) - V^{\hat{\pi}}(s)) \qquad (35) \\ &\leq \epsilon_{\text{stat}}^{\text{CB}} + 2 \sum_s \hat{d}(s)(V^\star(s) - V^{\hat{\pi}}(s)) \qquad (36) \\ &\lesssim \epsilon_{\text{stat}}^{\text{CB}} \qquad (37) \end{aligned}$$

In (35), we used Lemma 6 to bound the first term. (36) uses the fact that by definition, for all $s \notin \mathcal{S}_{\text{s}}$, we have $\mu(s) < 2\hat{d}(s)$ and $V^\star(s) - V^{\hat{\pi}}(s) \geq 0$. Lastly, we bound the second term using the result of step one, stated in (34). □

C.4 Auxiliary Lemmas

Lemma 4 (Closeness of empirical and population objectives, CB) *Fix the failure probability δ and define the following bounds $0 \leq w(s,a) \leq B_w, |v(s)| \leq B_v$. For the population objective $L_{AL}^{CB}(w,v)$ defined in (6) and empirical objective $\hat{L}_{AL}^{CB}(w,v)$ defined in (7), the following holds with probability $1-\delta$ and for all $v \in \mathcal{V}$ and $w \in \mathcal{W}$*

$$\begin{aligned} &|L_{AL}^{CB}(w,v) - \hat{L}_{AL}^{CB}(w,v)| \\ \lesssim& (B_w+1)^2 \sqrt{\frac{\log(|\mathcal{W}|/\delta)}{N}} + (B_v+1)(B_w+1)\sqrt{\frac{\log|\mathcal{W}||\mathcal{V}|/\delta}{N}} := \epsilon_{stat}^{CB} \end{aligned} \tag{38}$$

Proof. We break the population objective into two terms:

$$\begin{aligned} L_{\text{AL}}^{\text{CB}}(w,v) = \mathbb{E}_{s,a\sim\mu}\left[w(s,a)r(s,a)\right] - \mathbb{E}_{s,a\sim\mu}[v(s)(w(s,a)-1)] &:= T_1(w,v) \\ - \mathbb{E}_{s\sim\mu}[(\mathbb{E}_{a\sim\mu(\cdot|s)}[w(s,a)]-1)^2] &:= T_2(w,v) \end{aligned}$$

The corresponding terms in the empirical objective are:

$$\hat{L}^{\text{CB}}_{\text{AL}}(w,v) = \frac{1}{N}\sum_{i=1}^{N} w(s_i,a_i)r_i - \frac{1}{N}\sum_{i=1}^{N} v(s_i)(w(s_i,a_i)-1) := \hat{T}_1(w,v)$$
$$- \frac{1}{N}\sum_{i=1}^{N}\left(\sum_{a\in\mathcal{A}} w(s_i,a)\mu(a|s_i) - 1\right)^2 := \hat{T}_2(w,v).$$

We then separately argue the closeness of $T_i(w,v)$ to $\hat{T}_i(w,v)$.

Bounding $|T_1(w,v)-\hat{T}_1(w,v)|$. We use a similar argument as in Lemma 1 to bound the difference between the sum of the first two terms. It is easy to see that $\mathbb{E}_\mu[\hat{T}_1(w,v)] = T_1(w,v)$. Furthermore, $\hat{T}_1(w,v)$ is an empirical average over i.i.d. random variables which are bounded according to

$$|w(s,a)r(s,a) - v(s)(w(s,a)-1)| \leq B_w + B_v(B_w+1) = (B_v+1)(B_w+1).$$

By Hoeffding's inequality and applying a union bound on $w \in \mathcal{W}$ and $v \in \mathcal{V}$, with probability at least $1-\delta$ we have

$$|T_1(w,v) - \hat{T}_1(w,v)| \lesssim (B_v+1)(B_w+1)\sqrt{\frac{\log|\mathcal{W}||\mathcal{V}|/\delta}{N}}. \tag{39}$$

Bounding $|T_2(w,v) - \hat{T}_2(w,v)|$. By boundedness of w, similar to T_1, the following holds with probability greater than $1-\delta$

$$\left|T_2(w,v) - \hat{T}_2(w,v)\right| \lesssim (B_w+1)^2\sqrt{\frac{\log(|\mathcal{W}|/\delta)}{N}}. \tag{40}$$

Summing the bounds given in (39) and (40) proves the result. □

Lemma 5 (Closeness of objectives at $\hat{w}$ and $w^\star$, CB) *Fix failure probability δ. With probability at least $1-\delta$, we have*

$$L^{CB}_{AL}(w^\star,v^\star) - L^{CB}_{AL}(\hat{w},v^\star) \leq 2\epsilon^{CB}_{stat},$$

where ϵ^{CB}_{stat} is defined in (38).

Proof. First, note that the population objective without AL term is defined as

$$\max_{w\geq 0}\min_{v} L^{\text{CB}}(w,v) := \mathbb{E}_{s,a\sim\mu}\left[w(s,a)r(s,a)\right] - \mathbb{E}_{s,a\sim\mu}[v(s)(w(s,a)-1)].$$

By strong duality, $(w^\star,v^\star) \in \operatorname{argmax}_{w\geq 0}\operatorname{argmin}_v L^{\text{CB}}(w,v)$. Also, since $\mathbb{E}_{a\sim\mu(\cdot|s)}[w(s,a)] = 1$ for all s, by Lemma 13, $(w^\star,v^\star) \in \operatorname{argmax}_{w\geq 0}\operatorname{argmin}_v L^{\text{CB}}_{\text{AL}}(w,v)$.

We follow similar steps as in the proof of Lemma 2 and decompose $L^{\text{CB}}_{\text{AL}}(w^\star,v^\star) - L^{\text{CB}}_{\text{AL}}(\hat{w},v^\star)$ according to

$$\begin{aligned}
&L^{\text{CB}}_{\text{AL}}(w^\star,v^\star) - L^{\text{CB}}_{\text{AL}}(\hat{w},v^\star)\\
&\quad= \underbrace{L^{\text{CB}}_{\text{AL}}(w^\star,v^\star) - L^{\text{CB}}_{\text{AL}}(w^\star,\hat{v}(w^\star))}_{:=T_1} + \underbrace{L^{\text{CB}}_{\text{AL}}(w^\star,\hat{v}(w^\star)) - \hat{L}^{\text{CB}}_{\text{AL}}(w^\star,\hat{v}(w^\star))}_{:=T_2}\\
&\qquad+ \underbrace{\hat{L}^{\text{CB}}_{\text{AL}}(w^\star,\hat{v}(w^\star)) - \hat{L}^{\text{CB}}_{\text{AL}}(\hat{w},\hat{v})}_{:=T_3} + \underbrace{\hat{L}^{\text{CB}}_{\text{AL}}(\hat{w},\hat{v}) - \hat{L}^{\text{CB}}_{\text{AL}}(\hat{w},v^\star)}_{:=T_4}\\
&\qquad+ \underbrace{\hat{L}^{\text{CB}}_{\text{AL}}(\hat{w},v^\star) - L^{\text{CB}}_{\text{AL}}(\hat{w},v^\star)}_{:=T_5},
\end{aligned}$$

where $\hat{v}(w) = \arg\min_{v\in\mathcal{V}} \hat{L}^{\text{CB}}_{\text{AL}}(w,v)$. Each term is bounded as follows:

- $T_1 \leq 0$ because $v^\star \in \arg\min_v L^{\text{CB}}_{\text{AL}}(w^\star,v)$.

- $T_2 \leq \epsilon_{\text{stat}}^{\text{CB}}$ due to Lemma 4.
- $T_3 \leq 0$ because $\hat{w} = \arg\max_{w \in \mathcal{W}} \hat{L}_{\text{AL}}^{\text{CB}}(\hat{v}(w), w)$.
- $T_4 \leq 0$ because $\hat{v} = \arg\min_{v \in \mathcal{V}} \hat{L}_{\text{AL}}^{\text{CB}}(v, \hat{w})$.
- $T_5 \leq \epsilon_{\text{stat}}^{\text{CB}}$ due to Lemma 4.

Summing up the bounds on each term yields the desired bound. □

Lemma 6 ($\frac{\hat{d}}{\mu}(s)$ is not too small for significant states) *Let $\hat{w}$ be the optimal solution to the empirical objective (7). For the state space subset $\mathcal{S}_s := \{s \mid \hat{d}(s) \leq \mu(s)/2\}$, we have with probability at least $1-\delta$,*

1. $\mathbb{E}_{s\sim\mu}\left[\left(\frac{\hat{d}}{\mu}(s) - 1\right)^2\right] \lesssim \epsilon_{stat.}^{CB}$
2. $\sum_{s\in\mathcal{S}_s} \mu(s) \lesssim \epsilon_{stat.}^{CB}$

Proof. Note that by realizability of $w^\star$,

$$
\begin{aligned}
&L_{\text{AL}}^{\text{CB}}(w^\star, v^\star) \\
=&\mathbb{E}_{s,a\sim\mu}[r(s,a)w^\star(s,a)] + \underbrace{\mathbb{E}_{s,a\sim\mu}[v^\star(s)(1-w^\star(s,a))]}_{=0} - \underbrace{\mathbb{E}_{s\sim\mu}[(\mathbb{E}_{a\sim\mu(\cdot|s)}[w(s,a)]-1)^2]}_{=0} \\
=&\mathbb{E}_{s,a\sim\mu}[r(s,a)w^\star(s,a)].
\end{aligned}
$$

Define $r^\star(s) = \max_a r(s,a)$. Then $v^\star(s) = V^\star(s) = r^\star(s)$. Let $\frac{\hat{d}}{\mu}(s) := \mathbb{E}_{a\sim\mu(\cdot|s)}[\hat{w}(s,a)]$. Thus the objective difference becomes:

$$
\begin{aligned}
&L_{\text{AL}}^{\text{CB}}(w^\star, v^\star) - L_{\text{AL}}^{\text{CB}}(\hat{w}, v^\star) \\
=&\sum_s \mu(s) r^\star(s) - \sum_{s,a} \mu(s,a) r(s,a) \hat{w}(s,a) - \sum_s v^\star(s)\left(\mu(s) - \sum_a \mu(s,a)\hat{w}(s,a)\right) \\
&+ \mathbb{E}_{s\sim\mu}\left[\left(\frac{\hat{d}}{\mu}(s) - 1\right)^2\right] \\
=&\sum_{s,a} \mu(s,a)[r^\star(s) - r(s,a)]\hat{w}(s,a) + \mathbb{E}_{s\sim\mu}\left[\left(\frac{\hat{d}}{\mu}(s) - 1\right)^2\right].
\end{aligned}
$$

Recall that from Lemma 5, we have $L_{\text{AL}}^{\text{CB}}(w^\star, v^\star) - L_{\text{AL}}^{\text{CB}}(\hat{w}, v^\star) \lesssim \epsilon_{\text{stat}}^{\text{CB}}$. Therefore,

$$
\sum_{s,a} \mu(s,a)[r^\star(s) - r(s,a)]\hat{w}(s,a) + \mathbb{E}_{s\sim\mu}\left[\left(\frac{\hat{d}}{\mu}(s) - 1\right)^2\right] \lesssim \epsilon_{\text{stat}}^{\text{CB}}
$$

Since $\hat{w}(s,a) \geq 0$ and $r^\star(s) - r(s,a) \geq 0$, the first term is nonnegative. So our first claim holds:

$$
\mathbb{E}_{s\sim\mu}\left[\left(\frac{\hat{d}}{\mu}(s) - 1\right)^2\right] \lesssim \epsilon_{\text{stat}}^{\text{CB}}.
$$

The above result also allows us to bound the mass on the subset $\mathcal{S}_s$, which is our second claim:

$$
\epsilon_{\text{stat}}^{\text{CB}} \gtrsim \sum_s \mu(s)\left[\left(\frac{\hat{d}}{\mu}(s) - 1\right)^2\right] \geq \sum_{s\in\mathcal{S}_s} \mu(s)\left[\left(\frac{\hat{d}}{\mu}(s) - 1\right)^2\right] \geq \frac{1}{4}\sum_{s\in\mathcal{S}_s} \mu(s) \gtrsim \sum_{s\in\mathcal{S}_s} \mu(s).
$$

□

D PROOFS FOR MDPS

D.1 ON PRACTICAL IMPLEMENTATIONS

In our algorithms for CB and MDP, we need to compute summations of form $\sum_{a\in\mathcal{A}}$. This can be implemented efficiently when $|\mathcal{A}|$ is small. When $|\mathcal{A}|$ is large or even infinite, one can utilize numerical methods to estimate the summation with desired precision. Additionally, in Algorithm 3, we need to evaluate a term $\sum_{s',a'} P(s'|s,a)\pi_w(a'|s')u(s',a')$. In practice, we can evaluate this term by numerical integration.

D.2 NOTATIONS FOR MDP PROOF

For any $w \in \mathbb{R}_{\geq 0}^{|\mathcal{S}||\mathcal{A}|}$, recall that $d_w(s,a) = w(s,a)\mu(s,a)$, $d_w(s) = \sum_{a\in\mathcal{A}} d_w(s,a)$, $\pi_w(a|s) = d_w(s,a)/d_w(s)$ for $d_w(s) \neq 0$ and $\pi_w(a|s) = 1/|\mathcal{A}|$ for $d_w(s) = 0$.

Recall that the original primal-dual program without ALM term is

$$\max_{w\geq 0}\min_{v} L^{\text{MDP}}(w,v) := (1-\gamma)\mathbb{E}_{s\sim\rho}[v(s)] + \mathbb{E}_{s,a\sim\mu}\left[w(s,a)e_v(s,a)\right]. \tag{41}$$

and $w^\star(s,a) = d^{\pi^\star}(s,a)/\mu(s,a)$, $v^\star(s) = V^*(s)$. By strong duality, $(w^\star, v^\star) \in \arg\max_{w\geq 0}\arg\min_v L^{\text{MDP}}(w,v)$.

The model-based ALM population objective is obtained by directly substituting the variational form (12) in the original objective (9)

$$\begin{aligned}\max_{w\geq 0}\min_{v}\min_{u} L_{\text{AL}}^{\text{model-based}}(w,v,u) &:= (1-\gamma)\mathbb{E}_{s\sim\rho}\left[v(s) + \sum_a u(s,a)\pi_w(a|s)\right] \\ &+ \mathbb{E}_{s,a\sim\mu}\left[w(s,a)\left(e_v(s,a) - f_*^{-1}\left(u(s,a) - \gamma(\mathbb{P}^{\pi_w}u)(s,a)\right)\right)\right]\end{aligned} \tag{42}$$

The model-free ALM population objective derived from (42) is

$$\begin{aligned}\max_{w\geq 0}\min_{v}\min_{u}\max_{\zeta<0} L_{\text{AL}}^{\text{model-free}}(w,v,u,\zeta) &= (1-\gamma)\mathbb{E}_{s\sim\rho}\left[v(s) + \sum_a u(s,a)\pi_w(a|s)\right] \\ &+ \mathbb{E}_{(s,a,s')\sim\mu,a'\sim\pi_w(\cdot|s')}[w(s,a)\left(e_v(s,a) + (u(s,a) - \gamma u(s',a'))\zeta(s,a) + g_*(\zeta(s,a))\right)],\end{aligned} \tag{43}$$

where $g_*(x) := x+2+1/x$; see Appendix D.4 for more details on the derivation. Since $d_{w^\star} = d^{\pi_{w^\star}}$, by Lemma 13, $(w^\star, v^\star)$ is a max-min solution of (9), (42) and (43).

Additionally, define $\zeta^\star(w,u) = \zeta^\star_{w,u} = \arg\max_{\zeta<0} L_{\text{AL}}^{\text{model-free}}(w,v,u,\zeta)$, $\forall w \in \mathcal{W}, u \in \mathcal{U}$ and $\zeta^\star(w) = \zeta^\star_w = \zeta^\star(w, u^\star_w)$ $\forall w \in \mathcal{W}$ where $u^\star_w$ is defined in Theorem 4. Also, denote $\zeta^\star = \zeta^\star(w^\star)$ and $u^\star = u^\star_{w^\star}$.

D.3 FENCHEL CONJUGATE AND DOMAINS

For $f(x) = (x-1)^2$, the Fenchel conjugate f_* is given by

$$f_*(x) = \max_y(xy - f(y)) = \max_y\left(xy - y^2 + 2y - 1\right) = \left(\frac{x+2}{2}\right)^2 - 1. \tag{44}$$

Therefore, the optimum $x^\star_w$ can be calculated as

$$x^\star_w(s,a) = \operatorname*{argmax}_x\left(d_w(s)x - d^{\pi_w}(s)\left(\left(\frac{x+2}{2}\right)^2 - 1\right)\right) = 2\frac{d_w(s)}{d^{\pi_w}(s)} - 2 \tag{45}$$

Since $d_w(s)/d^{\pi_w}(s) \geq 0$, we have $x^\star_w(s,a) \geq -2$ and thus it is sufficient to only consider domain $x(s,a) \geq -2$, over which $f_*(x)$ is invertible.

Lemma 7 *For any $w \in \mathcal{W}$, define*

$$\tilde{x}_w(s,a) = \text{clip}\left(2\frac{d_w(s)}{d^{\pi_w}(s)} - 2, -1, 1\right).$$

Then, for any $s \in \mathcal{S}$ with $d^{\pi_w}(s) > 0$, one has

$$\frac{d_w(s)}{d^{\pi_w}(s)}\tilde{x}_w(s,a) - f_*(\tilde{x}_w(s,a)) \geq 0 \quad \forall s \in \mathcal{S}. \tag{46}$$

Proof. We separate the argument into three cases and use the expression of $\tilde{x}_w$ given in Lemma 7.

When $1 - 1/2 \leq \frac{d_w(s)}{d^{\pi_w}(s)} \leq 1/2 + 1$, we have $\tilde{x}_w(s,a) = \left(2\frac{d_w(s)}{d^{\pi_w}(s)} - 2\right)$ and therefore

$$\frac{d_w(s)}{d^{\pi_w}(s)}\tilde{x}_w(s,a) - f_*(\tilde{x}_w(s,a)) = \left(\frac{d_w(s)}{d^{\pi_w}(s)} - 1\right)^2 \geq 0.$$

When $\frac{d_w(s)}{d^{\pi_w}(s)} > 1/2 + 1$, substitute $\tilde{x}_w = 1$ to arrive at

$$\frac{d_w(s)}{d^{\pi_w}(s)}1 - \left(\left(\frac{1}{2}+1\right)^2 - 1\right) \geq \left(\frac{1}{2}+1\right)1 - \frac{1^2}{4} - 1 = \frac{1^2}{4} \geq 0.$$

Similarly When $\frac{d_w(s)}{d^{\pi_w}(s)} < 1 - 1/2$, substitute $\tilde{x}_w = -1$ to arrive at

$$-\frac{d_w(s)}{d^{\pi_w}(s)}1 - \left(\left(1-\frac{1}{2}\right)^2 - 1\right) \geq \left(\frac{1}{2}-1\right)1 - \frac{1^2}{4} + 1 = \frac{1^2}{4} \geq 0.$$

□

Lemma 8 *For any $w \in \mathcal{W}$, $s \in \mathcal{S}$, and $a \in \mathcal{A}$, we have*

$$\tilde{x}_w(s,a) = \text{clip}\left(2\frac{d_w(s)}{d^{\pi_w}(s)} - 2, -1, 1\right) \tag{47}$$

$$\frac{1}{1-\gamma}\left(\left(\frac{2-1}{2}\right)^2 - 1\right) \leq u_w^\star(s,a) \leq \frac{1}{1-\gamma}\left(\left(\frac{1+2}{2}\right)^2 - 1\right), \tag{48}$$

which implies that $|u_w^\star(s,a)| \leq \frac{1}{1-\gamma}\max\left\{1 - \left(\frac{2-1}{2}\right)^2, \left(\frac{1+2}{2}\right)^2 - 1\right\} = \frac{5}{4(1-\gamma)}$.

Proof. Since $f(x) = (x-1)^2$, a simple calculation shows $f_*(x) = \left(\frac{x+2}{2}\right)^2 - 1$. Therefore, for any (s,a),

$$x_w^\star(s,a) = \arg\max_x \left(d_w(s)x - d^{\pi_w}(s)\left(\left(\frac{x+2}{2}\right)^2 - 1\right)\right) = 2\frac{d_w(s)}{d^{\pi_w}(s)} - 2$$

$$\Rightarrow \tilde{x}_w(s,a) = \left(\left(2\frac{d_w(s)}{d^{\pi_w}(s)} - 2\right) \wedge 1\right) \vee (-1).$$

Recall that $u_w^\star$ is the fixed point of the following Bellman-like equation

$$u(s,a) = f_*(\tilde{x}_w(s,a)) + \gamma \mathcal{P}^{\pi_w} u(s,a). \tag{49}$$

The above equation has a solution since $f_*(\tilde{x}_w(s,a))$ is bounded

$$\left(\frac{2-1}{2}\right)^2 - 1 \leq f_*(\tilde{x}_w(s,a)) \leq \left(\frac{1+2}{2}\right)^2 - 1.$$

One can view $u_w^\star$ as the Q-function of policy π_w with the reward function $f_*(\tilde{x}_w(s,a))$, which leads to the bounds (48). □

D.4 Derivation of objective (43)

Recall that $f_*(x) = \left(\frac{x+2}{2}\right)^2 - 1$ with domain $x \in (-\infty, 2]$ in Appendix D.3. Let $g(x) = -f_*^{-1}(x) = 2 - 2\sqrt{x+1}$, which is a convex function on $[-1, +\infty)$. Similar to Nachum et al. (2019a), we use Fenchel duality to estimate $g\left(u(s,a) - \gamma \mathbb{P}^{\pi_w} u(s,a)\right)$. By Fenchel duality, any convex function $g(x)$ can be written as $g(x) = \max_\zeta x\zeta - g_*(\zeta)$. In the case of $g(x)$, the Fenchel conjugate is given by $g_*(x) = -x - 2 - 1/x$ with domain $x < 0$. Therefore, we write

$$
\begin{aligned}
&\mathbb{E}_\mu[w(s,a) g\left(u(s,a) - \gamma \mathbb{P}^{\pi_w} u(s,a)\right)] \\
&\quad = \mathbb{E}_\mu[w(s,a) \max_{\zeta<0} \left(u(s,a) - \gamma (\mathbb{P}^{\pi_w} u)(s,a)\right)\zeta - g_*(\zeta)] \\
&\quad = \mathbb{E}_\mu[w(s,a) \max_{\zeta<0} \left(u(s,a) - \gamma (\mathbb{P}^{\pi_w} u)(s,a)\right)\zeta + \zeta + 1/\zeta + 2]
\end{aligned}
$$

The interchangeability principle (Rockafellar & Wets, 2009; Dai et al., 2017) allows us to convert the inner maximization step over scalar ζ to an overall maximization over $\zeta : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^-$. Replacing this term in the objective 12 results in the model-free objective in (43).

Boundedness results for model-free Given any $w \in \mathcal{W}$, $u \in \mathcal{U}$, by definition,

$$
\zeta^\star_{w,u} = \arg\max_{\zeta<0} \mathbb{E}_{(s,a,s')\sim\mu, a'\sim\pi_w(\cdot|s')}[w(s,a)\left((u(s,a) - \gamma u(s',a') + 1)\zeta(s,a) + 1/\zeta(s,a)\right)]. \tag{50}
$$

It is easy to show that $|\zeta^\star_{w,u}(s,a)| = (u(s,a) - \gamma \mathcal{P}^{\pi_w} u(s',a') + 1)^{-1/2} = (f_*(\tilde{x}_w(s,a)) + 1)^{-1/2}$. Since $\tilde{x}_w(s,a) \in [-1, 1]$, we have $|\zeta^\star_{w,u}(s,a)| \in \left[\frac{\sqrt{2}}{3}, \sqrt{2}\right]$.

D.5 Proof of Theorem 4

Proof. For convenience, we define several notations in Appendix D.2. Note that the boundedness of $u^\star_w$ and $\zeta^\star_{w^\star,u}$ are shown in Lemma 8 and Appendix D.4 respectively.

Bounding the suboptimality of policies returned by both model-based and model-free variants of CORAL follow a similar analysis. To avoid cluttered notation, we drop the superscripts model-based and model-free from statistical errors $\epsilon_{\text{stat}}^{\text{model-based}}$ and $\epsilon_{\text{stat}}^{\text{model-free}}$.

In view of the performance difference lemma in Kakade & Langford (2002, Lemma 6.1), one has

$$
J(\pi^\star) - J(\hat{\pi}) = \mathbb{E}_{s\sim d^{\hat{\pi}}}\left[\sum_a A^\star(s,a)\left(\pi^\star(a|s) - \hat{\pi}(a|s)\right)\right] = \mathbb{E}_{s\sim d^{\hat{\pi}}}\left[\sum_a -A^\star(s,a)\hat{\pi}(a|s)\right].
$$

Here, we used the fact that optimal advantage expected over optimal policy is zero $\sum_a A^\star(s,a)\pi^\star(a|s) = 0$. Lemma 10 links an expectation of $-A^\star(s,a)$ to the statistical error. With this lemma at hand and using the definition $\mathcal{S}_{\text{s}} = \{s \mid d_{\hat{w}}(s) \le d^{\hat{\pi}}(s)/2\}$, we continue to decompose and bound the suboptimality

$$
\begin{aligned}
&\mathbb{E}_{s\sim d^{\hat{\pi}}}\left[\sum_a -A^\star(s,a)\hat{\pi}(a|s)\right] \\
&\quad = \sum_{s\in\mathcal{S}_{\text{s}}} d^{\hat{\pi}}(s)\left[\sum_a -A^\star(s,a)\hat{\pi}(a|s)\right] + \sum_{s\notin\mathcal{S}_{\text{s}}} d^{\hat{\pi}}(s)\left[\sum_a -A^\star(s,a)\hat{\pi}(a|s)\right] \\
&\quad \lesssim \frac{1}{1-\gamma}\epsilon_{\text{stat}} + \sum_{s\notin\mathcal{S}_{\text{s}}, d_{\hat{w}}(s)\neq 0} \frac{d^{\hat{\pi}}(s)}{d_{\hat{w}}(s)}\left[\sum_a -A^\star(s,a)\hat{w}(s,a)\mu(s,a)\right] \qquad (51) \\
&\qquad + \sum_{s\notin\mathcal{S}_{\text{s}}, d_{\hat{w}}(s)=0} d^{\hat{\pi}}(s)\left[\sum_a -\frac{1}{|\mathcal{A}|}A^\star(s,a)\right] \\
&\quad \le \frac{1}{1-\gamma}\epsilon_{\text{stat}} + 2\sum_{s\notin\mathcal{S}_{\text{s}}}\left[\sum_a -A^\star(s,a)\hat{w}(s,a)\mu(s,a)\right] \qquad (52)
\end{aligned}
$$

In (51), we used result 2 in Lemma 10 and that $-A^\star(s,a) \leq 1/(1-\gamma)$ and in (52) we used the definition of $\mathcal{S}_s$ to bound the ratio $d^{\hat{\pi}}(s)/d_{\hat{w}}(s)$ by 2 and the fact that $d_{\hat{w}}(s) = 0$ implies $d^{\hat{\pi}}(s) = 0$ for $s \notin \mathcal{S}_s$. We then apply result 1 in in Lemma 10 to bound the second term by $\mathbb{E}_{s,a\sim\mu}[-A^\star(s,a)\hat{w}(s,a)]$ and thus the overall suboptimality:

$$J(\pi^\star) - J(\hat{\pi}) \lesssim \frac{1}{1-\gamma}\epsilon_{\text{stat}} + \mathbb{E}_{s,a\sim\mu}\left[-A^\star(s,a)\hat{w}(s,a)\right] \lesssim \frac{1}{1-\gamma}\epsilon_{\text{stat}}.$$

□

D.6 Model-based statistical estimation error

Theorem 5 (Convergence of MLE for learning transitions (Van de Geer, 2000)) *Given a realizable model class $\mathcal{P} = \{P : (\mathcal{S},\mathcal{A}) \to \Delta(\mathcal{S})\}$ that contains the true model $P^\star$ and a dataset $\mathcal{D}_m = \{(s_i,a_i,s_i')\}$ with $(s_i,a_i) \overset{iid}{\sim} \mu, s_i' \sim P^\star(\cdot|s_i,a_i)$, let $\hat{P}$ be*

$$\hat{P} = \arg\max_{P\in\mathcal{P}} \sum_{i=1}^{N} \ln P(s_i'|s_i,a_i).$$

Fix failure probability $\delta > 0$. Then, with probability at least $1-\delta$, we have the following concentration on the squared Hellinger distance between $\hat{P}$ and $P^\star$:

$$\mathbb{E}_{s,a\sim\mu}\left[\sum_{s'}\left(\sqrt{\hat{P}(s'|s,a)} - \sqrt{P^\star(s'|s,a)}\right)^2\right] \lesssim \frac{\log(|\mathcal{P}|/\delta)}{N}.$$

Proposition 6 (Statistical error of empirical objective, model-based) *With probability at least $1-\delta$, for any w,v,u, the difference between the population objective (42) and the model-based empirical estimation (13) is bounded by*

$$\begin{aligned}&\left|L_{AL}^{\text{model-based}}(w,v,u) - \hat{L}_{AL}^{\text{model-based}}(w,v,u)\right| \\ \lesssim&(B_v + B_u + (1+B_v)B_w)\sqrt{\frac{B_u \log|\mathcal{P}||\mathcal{U}||\mathcal{W}||\mathcal{V}|/\delta}{N}} := \epsilon_{stat}^{\text{model-based}}\end{aligned} \tag{53}$$

where we assume $N_0 = N_m = N$ for simplicity.

Proof. Define the following terms:

$$\begin{aligned}T_1 :=& (1-\gamma)\mathbb{E}_\rho\left[v(s) + \sum_a u(s,a)\pi_w(a|s)\right] \\ &- (1-\gamma)\frac{1}{N_0}\sum_{i=1}^{N_0}\left(v(s_i) + \sum_a u(s_i,a)\pi_w(a|s_i)\right) \\ T_2 :=& \mathbb{E}_\mu\left[w(s,a)(r(s,a) + \gamma\sum_{s'}P(s'|s,a)v(s') - v(s))\right] \\ &- \frac{1}{N}\sum_{i=1}^{N} w(s_i,a_i)\left[r_i + \gamma v(s_i') - v(s_i)\right] \\ T_3 :=& \mathbb{E}_\mu\left[w(s,a)\left(f_*^{-1}\left(u(s,a) - \gamma P^{\pi_w}u(s,a)\right)\right)\right] \\ &- \frac{1}{N}\sum_{i=1}^{N} w(s_i,a_i)\left[f_*^{-1}\left(u(s_i,a_i) - \gamma\hat{P}^{\pi_w}u(s_i,a_i)\right)\right]\end{aligned}$$

One can check that

$$L_{\text{AL}}^{\text{model-based}} - \hat{L}_{\text{AL}}^{\text{model-based}} = T_1 + T_2 + T_3.$$

Thus triangle inequality finds the following bound on the absolute value of the difference:

$$\left|L_{\text{AL}}^{\text{model-based}} - \hat{L}_{\text{AL}}^{\text{model-based}}\right| \leq |T_1| + |T_2| + |T_3|.$$

We provide high probability bounds on each term separately.

Bounding T_1. Since $|v(s)|\leq B_v, |u(s,a)|\leq B_u$ for all $v\in\mathcal{V}$ and $u\in\mathcal{U}$ and s_i are independent, we can apply Hoeffding's inequality and union bound to conclude that

$$|T_1|\lesssim B_v\sqrt{\frac{\log|\mathcal{V}|/\delta}{N_0}}+B_u\sqrt{\frac{\log|\mathcal{U}|/\delta}{N_0}}\lesssim (B_v+B_u)\sqrt{\frac{\log|\mathcal{V}||\mathcal{U}|/\delta}{N_0}} \tag{54}$$

Bounding T_2. By boundedness of w, v, we have

$$|w(s,a)(r(s,a)+\gamma v(s')-v(s))|\leq B_w(1+(\gamma+1)B_v)\leq B_w(1+\gamma)(1+B_v).$$

As before, due to boundedness and independence of $w(s_i,a_i)[r_i+\gamma v(s_i')-v(s_i)]$, Hoeffding's inequality can be applied, giving the following bound on T_2:

$$|T_2|\lesssim (1+B_v)B_w\sqrt{\frac{\log|\mathcal{V}||\mathcal{W}|/\delta}{N}}. \tag{55}$$

Bounding T_3. We decompose $T_3=T_{3,1}+T_{3,2}$, where $T_{3,1}$ and $T_{3,2}$ are defined as

$$\begin{aligned}
T_{3,1} &:= \mathbb{E}_\mu\left[w(s,a)\left(f_*^{-1}\left(u(s,a)-\gamma P^{\pi_w}u(s,a)\right)\right)\right]\\
&\quad -\mathbb{E}_\mu\left[w(s,a)\left(f_*^{-1}\left(u(s,a)-\gamma \hat{P}^{\pi_w}u(s,a)\right)\right)\right]\\
T_{3,2} &:= \mathbb{E}_\mu\left[w(s,a)\left(f_*^{-1}\left(u(s,a)-\gamma \hat{P}^{\pi_w}u(s,a)\right)\right)\right]\\
&\quad -\frac{1}{N}\sum_{i=1}^N w(s_i,a_i)\left[f_*^{-1}\left(u(s_i,a_i)-\gamma \hat{P}^{\pi_w}u(s_i,a_i)\right)\right]
\end{aligned}$$

Recall that $f_*^{-1}(x)=2\sqrt{x+1}-2$ in Appendix D.4.

Bounding $T_{3,1}$. We first use the inequality given in Lemma 12, setting b_i, x_i, y_i for each (s,a) according to

$$b_i=\begin{cases}1+u(s,a) & i=0\\ \gamma\sum_{a'}\pi_w(a'|s')u(s'|a') & 1\leq i\leq|\mathcal{S}|\end{cases}$$

$$x_i=\begin{cases}1 & i=0\\ P(s'|s,a) & 1\leq i\leq|\mathcal{S}|\end{cases},\quad y_i=\begin{cases}1 & i=0\\ P(s'|s,a) & 1\leq i\leq|\mathcal{S}|\end{cases}$$

which gives

$$\begin{aligned}
T_{3,1}^2 &= \Bigg(\mathbb{E}_\mu\left[w(s,a)\left(f_*^{-1}\left(u(s,a)-\gamma P^{\pi_w}u(s,a)\right)\right)\right]\\
&\qquad -\mathbb{E}_\mu\left[w(s,a)\left(f_*^{-1}\left(u(s,a)-\gamma \hat{P}^{\pi_w}u(s,a)\right)\right)\right]\Bigg)^2\\
&\lesssim B_w\Bigg(\mathbb{E}_\mu\left[\sqrt{1+u(s,a)-\gamma\sum_{s'}P(s'|s,a)\sum_{a'}\pi_w(a'|s')u(s',a')}\right]\\
&\qquad -\mathbb{E}_\mu\left[\sqrt{1+u(s,a)-\gamma\sum_{s'}\hat{P}(s'|s,a)\sum_{a'}\pi_w(a'|s')u(s',a')}\right]\Bigg)^2\\
&\leq B_w^2B_u\mathbb{E}_\mu\left[\sum_{s'}\left(\sqrt{P(s'|s,a)}-\sqrt{\hat{P}(s'|s,a)}\right)^2\right].
\end{aligned} \tag{56}$$

Note that $1+u(s,a)-\gamma\sum_{s'}\hat{P}(s'|s,a)\sum_{a'}\pi_w(a'|s')u(s',a')$ is always non-negative if we set the threshold for $\tilde{x}(s,a)$ of order $O(1-\gamma)$. In Theorem 4, we set the threshold to be 1 for notation convenience.

By Theorem 5, we have the following concentration on model estimation

$$\mathbb{E}_{s,a\sim\mu}\left[\sum_{s'}\left(\sqrt{\hat{P}(s'|s,a)}-\sqrt{P(s'|s,a)}\right)^2\right]\lesssim\frac{\log(|\mathcal{P}|/\delta)}{N},$$

which combined with (56) and a union bound on $w \in \mathcal{W}$ and $v \in \mathcal{V}$ give

$$|T_{3,1}| \lesssim B_w \sqrt{\frac{B_u \log|\mathcal{P}||\mathcal{U}||\mathcal{W}|/\delta}{N}}. \tag{57}$$

Bounding $T_{3,2}$. Similar to terms T_1 and T_2, Hoeffding's inequality combined with union bound asserts the following bound on $T_{3,2}$

$$|T_{3,2}| \lesssim B_w \sqrt{\frac{B_u \log|\mathcal{W}||\mathcal{U}|\delta}{N}}. \tag{58}$$

Bounding the objective difference. Summing up the bounds given in (54), (55), (57), and (58) yields

$$\begin{aligned}
&L_{\text{AL}}^{\text{model-based}}(w,v,u) - \hat{L}_{\text{AL}}^{\text{model-based}}(w,v,u) \\
&\lesssim (B_v + B_u)\sqrt{\frac{\log|\mathcal{V}||\mathcal{U}|/\delta}{N_0}} + (1+B_v)B_w\sqrt{\frac{B_u \log|\mathcal{P}||\mathcal{U}||\mathcal{W}||\mathcal{V}|/\delta}{N}}
\end{aligned}$$

□

Lemma 9 (Closeness of $w^\star, \hat{w}$, model-based) *With probability at least $1-\delta$,*

$$L_{AL}^{\textit{model-based}}(w^\star, v^\star, u^\star) - L_{AL}^{\textit{model-based}}(\hat{w}, v^\star, u^\star_{\hat{w}}) \le 2\epsilon_{\textit{model-based}}$$

Proof. Define

$$\hat{v}(w), \hat{u}(w) = \arg\min_{v\in\mathcal{V}} \arg\min_{u\in\mathcal{U}} \hat{L}_{\text{AL}}^{\text{model-based}}(w,v,u)$$

Decompose the objective difference according to

$$\begin{aligned}
&L_{\text{AL}}^{\text{model-based}}(w^\star, v^\star, u^\star) - L_{\text{AL}}^{\text{model-based}}(\hat{w}, v^\star, u^\star_{\hat{w}}) \\
&\quad = L_{\text{AL}}^{\text{model-based}}(w^\star, v^\star, u^\star) - L_{\text{AL}}^{\text{model-based}}(w^\star, \hat{v}(w^\star), \hat{u}(w^\star)) \quad := T_1 \\
&\qquad + L_{\text{AL}}^{\text{model-based}}(w^\star, \hat{v}(w^\star), \hat{u}(w^\star)) - \hat{L}_{\text{AL}}^{\text{model-based}}(w^\star, \hat{v}(w^\star), \hat{u}(w^\star)) \quad := T_2 \\
&\qquad + \hat{L}_{\text{AL}}^{\text{model-based}}(w^\star, \hat{v}(w^\star), \hat{u}(w^\star)) - \hat{L}_{\text{AL}}^{\text{model-based}}(\hat{w}, \hat{v}(\hat{w}), \hat{u}(\hat{w})) \quad := T_3 \\
&\qquad + \hat{L}_{\text{AL}}^{\text{model-based}}(\hat{w}, \hat{v}(\hat{w}), \hat{u}(\hat{w})) - \hat{L}_{\text{AL}}^{\text{model-based}}(\hat{w}, v^\star, u^\star_{\hat{w}}) \quad := T_4 \\
&\qquad + \hat{L}_{\text{AL}}^{\text{model-based}}(\hat{w}, v^\star, u^\star_{\hat{w}}) - L_{\text{AL}}^{\text{model-based}}(\hat{w}, v^\star, u^\star_{\hat{w}}) \quad := T_5
\end{aligned}$$

We bound each term:

- $T_1 \le 0$ because $v^\star, u^\star = \arg\min_v \arg\min_u L_{\text{AL}}^{\text{model-based}}(w^\star, v, u)$;
- $T_2 \le \epsilon_{\text{model-based}}$ by Proposition (6);
- $T_3 \le 0$ because $\hat{w} = \arg\max_{w\in\mathcal{W}} \hat{L}_{\text{AL}}^{\text{model-based}}(w, \hat{v}(w), \hat{u}(w))$;
- $T_4 \le 0$ because $\hat{v}(w), \hat{u}(w) = \arg\min_{v\in\mathcal{V}} \arg\min_{u\in\mathcal{U}} \hat{L}_{\text{AL}}^{\text{model-based}}(w,v,u)$;
- $T_5 \le \epsilon_{\text{model-based}}$ by Proposition (6).

□

Lemma 10 *For $\hat{w}$ computed by Algorithm 3, define $\mathcal{S}_s := \{s \mid d_{\hat{w}}(s) \le d^{\pi_{\hat{w}}}(s)/2\}$. Then with probability at least $1-\delta$,*

$$\mathbb{E}_{s,a\sim\mu}[-A^\star(s,a)\hat{w}^{\textit{model-based}}(s,a)] \lesssim \epsilon_{\textit{stat}}^{\textit{model-based}},$$
$$\sum_{s\in\mathcal{S}_s} d^{\pi_{\hat{w}}}(s) \lesssim \epsilon_{\textit{stat}}^{\textit{model-based}}.$$

Similarly, for $\hat{w}$ computed by Algorithm 4, define $\mathcal{S}_s := \{s \mid d_{\hat{w}}(s) \le d^{\pi_{\hat{w}}}(s)/2\}$. Then with probability at least $1-\delta$,

$$\mathbb{E}_{s,a\sim\mu}[-A^\star(s,a)\hat{w}^{\textit{model-free}}(s,a)] \lesssim \epsilon_{\textit{stat}}^{\textit{model-free}}$$
$$\sum_{s\in\mathcal{S}_s} d^{\pi_{\hat{w}}}(s) \lesssim \epsilon_{\textit{stat}}^{\textit{model-free}}.$$

Proof. We provide proof only for the model-based algorithm and let $\hat{w} = \hat{w}^{\text{model-based}}$ for notation convenience. The proof for a model-free algorithm follows analogously, noting the fact that $L_{AL}^{\text{model-free}}(w, v^\star, u_w^\star, \zeta^\star(w)) = L_{AL}^{\text{model-based}}(w, v^\star, u_w^\star))$ and we can replace Lemma 9 with Lemma 11 in the model-free proof.

Consider the expression of $L_{\text{AL}}^{\text{model-based}}(w^\star, v^\star, u^\star)$:

$$\begin{aligned}
&L_{\text{AL}}^{\text{model-based}}(w^\star, v^\star, u^\star) \\
=&(1-\gamma)\mathbb{E}_{s\sim\rho}[v^\star(s)] + \mathbb{E}_{s,a\sim\mu}\left[w^\star(s,a)e_{v^\star}(s,a)\right] - \mathbb{E}_{s\sim d^{\pi_{w^\star}}}\left(\frac{d_{w^\star}(s)}{d^{\pi_{w^\star}}(s)} - 1\right)^2 \\
=&(1-\gamma)\mathbb{E}_{s\sim\rho}[V^\star(s)] + \mathbb{E}_{s,a\sim\mu}\left[w^\star(s,a)A^\star(s,a)\right] \qquad (59)
\end{aligned}$$

The first equation comes from the fact that $u^\star$ is the optimal solution to the variational lower bound, making it equal to the f-divergence. In the last equation, we used $v^\star(s) = V^\star(s)$ as in Appendix D.2, the fact that $e_{v^\star}(s,a) = r(s,a) + \gamma(\sum_{s'} P(s'|s,a)v^\star(s')) - v^\star(s) = A^\star(s,a)$ is the optimal advantage function, and that $d_{w^\star}(s) = d^{\pi_{w^\star}}(s)$ due to the definition of $w^\star$. Moreover, the second term in (59) is zero since it captures the optimal advantage of optimal policy. Therefore, we conclude that

$$L_{\text{AL}}^{\text{model-based}}(w^\star, v^\star, u^\star) = (1-\gamma)\mathbb{E}_{s\sim\rho}[V^\star(s)]. \qquad (60)$$

Given the above expression of the objective at $(w^\star, v^\star, u^\star)$, we write the following objective difference

$$\begin{aligned}
&L_{\text{AL}}^{\text{model-based}}(w^\star, v^\star, u^\star) - L_{\text{AL}}^{\text{model-based}}(\hat{w}, v^\star, u_{\hat{w}}^\star) \\
=&(1-\gamma)\mathbb{E}_{s\sim\rho}[V^\star(s)] - (1-\gamma)\mathbb{E}_{s\sim\rho}[v^\star(s)] - \mathbb{E}_{s,a\sim\mu}\left[\hat{w}(s,a)e_{v^\star}(s,a)\right] \\
&\quad + (1-\gamma)\mathbb{E}_{s\sim\rho,a\sim\pi_{\hat{w}}}[u_{\hat{w}}^\star(s,a)] + \mathbb{E}_\mu\left[\hat{w}(s,a)f_*^{-1}\left(u_{\hat{w}}^\star(s,a) - \gamma(\mathbb{P}^{\pi_{\hat{w}}}u_{\hat{w}}^\star)(s,a)\right)\right] \qquad (61) \\
=&-\mathbb{E}_{s,a\sim\mu}\left[\hat{w}(s,a)A^\star(s,a)\right] - \mathbb{E}_{d^{\pi_{\hat{w}}}}\left[f_*(\tilde{x}_{\hat{w}}(s,a))\right] + \mathbb{E}_{d_{\hat{w}}}\left[\tilde{x}_{\hat{w}}(s,a)\right]
\end{aligned}$$

The last line uses $e_{v^\star}(s,a) = A^\star(s,a)$ as well as the definition of $u_{\hat{w}}^\star$ as the fixed point solution to

$$u_{\hat{w}}^\star(s,a) \coloneqq f_*(\tilde{x}_{\hat{w}}(s,a)) + \gamma(\mathbb{P}^{\pi_{\hat{w}}}u_{\hat{w}}^\star)(s,a),$$

which allows us to write (61) in the original f-divergence variational form (10) with $\tilde{x}_{\hat{w}}$ as variable. Lemma 9 asserts that $L_{\text{AL}}^{\text{model-based}}(w^\star, v^\star, u^\star) - L_{\text{AL}}^{\text{model-based}}(\hat{w}, v^\star, u_{\hat{w}}^\star) \lesssim \epsilon_{\text{model-based}}$. Therefore,

$$-\mathbb{E}_{s,a\sim\mu}\left[\hat{w}(s,a)A^\star(s,a)\right] - \mathbb{E}_{d^{\pi_{\hat{w}}}}\left[f_*(\tilde{x}_{\hat{w}}(s,a))\right] + \mathbb{E}_{d_{\hat{w}}}\left[\tilde{x}_{\hat{w}}(s,a)\right] \lesssim \epsilon_{\text{model-based}}. \qquad (62)$$

Notice that both terms in inequality above are nonnegative. The first term is nonnegative because for the optimal advantage function we have $A^\star(s,a) \leq 0$ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$. The second term is also nonnegative as it involves an expectation of $\left(\frac{d_{\hat{w}}(s)}{d^{\pi_{\hat{w}}}(s)}\tilde{x}_{\hat{w}}(s,a) - f_*(\tilde{x}_{\hat{w}}(s,a))\right)$, which are shown to be nonnegative in Lemma 7. Hence, we conclude the first claim that $\mathbb{E}_{s,a\sim\mu}\left[-A^\star(s,a)\hat{w}(s,a)\right] \lesssim \epsilon_{\text{stat}}^{\text{model-based}}$.

To prove the second claim, we can assume WLOG that $d^{\pi_{\hat{w}}}(s) > 0$ since $d^{\pi_{\hat{w}}}(s) = 0$ implies that $d^{\hat{w}}(s) = 0$ for $s \in \mathcal{S}_{\text{s}}$ and thus the state does not affect the conclusion. Note that for any $s \in \mathcal{S}_{\text{s}}$, $\tilde{x}_{\hat{w}}(s,a) = \max\{x_{\hat{w}}^\star(s,a), -1\}$ because

$$2\frac{d_{\hat{w}}(s)}{d^{\pi_{\hat{w}}}(s)} - 2 \leq -1 \quad \Rightarrow \quad \tilde{x}_{\hat{w}}(s,a) = \text{clip}\left(2\frac{d_{\hat{w}}(s)}{d^{\pi_{\hat{w}}}(s)} - 2, -1, 1\right) = \max\{x_{\hat{w}}^\star(s,a), -1\}.$$

When $-1 \leq x_{\hat{w}}^\star(s,a)$, we have $\tilde{x}_{\hat{w}}(s,a) = x_{\hat{w}}^\star(s,a)$ and thus

$$\frac{d_{\hat{w}}(s)}{d^{\pi_{\hat{w}}}(s)}\tilde{x}_{\hat{w}}(s,a) - f_*(\tilde{x}_{\hat{w}}(s,a)) = \left(\frac{d_{\hat{w}}(s)}{d^{\pi_{\hat{w}}}(s)} - 1\right)^2 \geq \frac{1}{4}.$$

When $-1 > x_{\hat{w}}^\star(s,a)$, we have $\tilde{x}_{\hat{w}}(s,a) = -1$ and thus

$$\frac{d_{\hat{w}}(s)}{d^{\pi_{\hat{w}}}(s)}\tilde{x}_{\hat{w}}(s,a) - f_*(\tilde{x}_{\hat{w}}(s,a)) \geq \frac{1^2}{4}.$$

Therefore,

$$\frac{d_{\hat{w}}(s)}{d^{\pi_{\hat{w}}}(s)}\tilde{x}_{\hat{w}}(s,a) - f_*(\tilde{x}_{\hat{w}}(s,a)) \geq \frac{1}{4} \quad \forall s \in \mathcal{S}_{\text{s}}. \tag{63}$$

We find a lower bound on the left-hand side of (62) using the bounds given in (46) and (63)

$$\begin{aligned}
\epsilon_{\text{model-based}} \gtrsim & \mathbb{E}_{d_{\hat{w}}}[\tilde{x}_{\hat{w}}(s,a)] - \mathbb{E}_{d^{\pi_{\hat{w}}}}[f_*(\tilde{x}_{\hat{w}}(s,a))] \\
= & \sum_s d^{\pi_{\hat{w}}}(s)\left(\frac{d_{\hat{w}}(s)}{d^{\pi_{\hat{w}}}(s)}\tilde{x}_{\hat{w}}(s,a) - f_*(\tilde{x}_{\hat{w}}(s,a))\right) \\
\geq & \sum_{s\in\mathcal{S}_{\text{s}}} \frac{1}{4} d^{\pi_{\hat{w}}}(s),
\end{aligned}$$

which leads to the second advertised claim $\sum_{s\in\mathcal{S}_{\text{s}}} d^{\pi_{\hat{w}}}(s) \lesssim \epsilon_{\text{model-based}}$. □

D.7 Model-free statistical estimation error

Similar to the model-based algorithm, we can also show the closeness of empirical and population objectives which is formalized in Proposition 7.

Proposition 7 (Statistical error of empirical objective, model-free) *With probability at least $1-\delta$, for any w, v, u, ζ, the difference between the population objective (43) and the model-free empirical estimation (14) is bounded by*

$$\begin{aligned}
&\left|L_{AL}^{\textit{model-free}}(w,v,u,\zeta) - \hat{L}_{AL}^{\textit{model-free}}(w,v,u,\zeta)\right| \\
\lesssim & (B_v + B_u + (1 + B_v + B_\zeta(B_u+1))B_w)\sqrt{\frac{\log|\mathcal{U}||\mathcal{W}||\mathcal{V}||\mathcal{Z}|/\delta}{N}} := \epsilon_{\textit{model-free}},
\end{aligned} \tag{64}$$

where we assume $N_0 = N$ for simplicity.

Proof. Define the following terms:

$$\begin{aligned}
T_1 := & (1-\gamma)\mathbb{E}_\rho\left[v(s) + \sum_a u(s,a)\pi_w(a|s)\right] \\
& - (1-\gamma)\frac{1}{N_0}\sum_{i=1}^{N_0}\left(v(s_i) + \sum_a u(s_i,a)\pi_w(a|s_i)\right) \\
T_2 := & \mathbb{E}_\mu\left[w(s,a)(r(s,a) + \gamma\sum_{s'} P(s'|s,a)v(s') - v(s))\right] \\
& - \frac{1}{N}\sum_{i=1}^{N} w(s_i,a_i)\left[r_i + \gamma v(s_i') - v(s_i)\right] \\
T_3 := & \mathbb{E}_{(s,a,s')\sim\mu, a'\sim\pi_w(\cdot|s')}[w(s,a)\left((u(s,a) - \gamma u(s',a'))\zeta(s,a) - g_\star(\zeta(s,a))\right)] \\
& - \frac{1}{N}\sum_{i=1}^{N} w(s_i,a_i)\left[\left(u(s_i,a_i) - \gamma\sum_{a'\in\mathcal{A}} u(s_i',a')\pi_w(a'|s_i')\right)\zeta(s_i,a_i) - g_\star(\zeta(s_i,a_i))\right].
\end{aligned}$$

One can check that

$$L_{\text{AL}}^{\text{model-free}} - \hat{L}_{\text{AL}}^{\text{model-free}} = T_1 + T_2 + T_3.$$

Thus triangle inequality finds the following bound on the absolute value of the difference:

$$\left|L_{\text{AL}}^{\text{model-free}} - \hat{L}_{\text{AL}}^{\text{model-free}}\right| \leq |T_1| + |T_2| + |T_3|.$$

We provide high probability bounds on each term separately.

Bounding T_1 and T_2. The first two terms are the same as T_1, T_2 in the proof of Proposition 6. Therefore, we have

$$|T_1| \lesssim (B_v + B_u)\sqrt{\frac{\log|\mathcal{V}||\mathcal{U}|/\delta}{N_0}} \tag{65}$$

and

$$|T_2| \lesssim (1 + B_v)B_w\sqrt{\frac{\log|\mathcal{V}||\mathcal{W}|/\delta}{N}}. \tag{66}$$

Bounding T_3. Recall that $g_\star(x) = -x - 2 - \frac{1}{x}$, $\forall x < 0$. Also, $|\zeta(s,a)| \in (B_{\zeta,L}, B_{\zeta,U})$ for any $\zeta \in \mathcal{Z}$ and any (s,a), and $B_\zeta \triangleq \max\{B_{\zeta,U}, B_{\zeta,L}^{-1}\}$. Then,

$$\begin{aligned}
&w(s,a)\left((u(s,a) - \gamma u(s',a'))\zeta(s,a) - g_\star(\zeta(s,a))\right) \\
\leq &B_w((1+\gamma)B_u B_{\zeta,U} + B_{\zeta,U} + B_{\zeta,L}^{-1} + 2),
\end{aligned}$$

which implies that

$$|T_3| \lesssim (1 + B_\zeta(B_u + 1))B_w\sqrt{\frac{\log|\mathcal{U}||\mathcal{W}||\mathcal{Z}|/\delta}{N}}. \tag{67}$$

Bounding the objective difference. Summing up the bounds given in (65), (66), and (67) yields

$$\begin{aligned}
&L_{\text{AL}}^{\text{model-free}}(w,v,u,\zeta) - \hat{L}_{\text{AL}}^{\text{model-free}}(w,v,u,\zeta) \\
&\lesssim (B_v + B_u)\sqrt{\frac{\log|\mathcal{V}||\mathcal{U}|/\delta}{N_0}} + (1 + B_v + B_\zeta(B_u+1))B_w\sqrt{\frac{\log|\mathcal{U}||\mathcal{W}||\mathcal{V}||\mathcal{Z}|/\delta}{N}}.
\end{aligned}$$

□

Lemma 11 (Closeness of $w^\star, \hat{w}$, model-free) *With probability at least $1-\delta$,*

$$L_{AL}^{\textit{model-free}}(w^\star, v^\star, u^\star, \zeta^\star) - L_{AL}^{\textit{model-free}}(\hat{w}, v^\star, u_{\hat{w}}^\star, \zeta^\star(\hat{w})) \leq 2\epsilon_{\textit{model-free}}$$

Proof. Define

$$\hat{v}(w), \hat{u}(w), \hat{\zeta}(w) = \arg\min_{v\in\mathcal{V}} \arg\min_{u\in\mathcal{U}} \arg\max_{\zeta\in\mathcal{Z}} \hat{L}_{\text{AL}}^{\text{model-free}}(w,v,u,\zeta), \quad \forall w \in \mathcal{W}$$

and

$$\hat{\zeta}(w,u) = \arg\max_{\zeta\in\mathcal{Z}} \hat{L}_{\text{AL}}^{\text{model-free}}(w,v,u,\zeta) \quad \forall w \in \mathcal{W}, u \in \mathcal{U}$$

Decompose the objective difference according to

$$\begin{aligned}
&L_{\text{AL}}^{\text{model-free}}(w^\star, v^\star, u^\star, \zeta^\star) - L_{\text{AL}}^{\text{model-free}}(\hat{w}, v^\star, u_{\hat{w}}^\star, \zeta^\star(\hat{w})) \\
&\quad = L_{\text{AL}}^{\text{model-free}}(w^\star, v^\star, u^\star, \zeta^\star) - L_{\text{AL}}^{\text{model-free}}(w^\star, \hat{v}(w^\star), \hat{u}(w^\star), \zeta^\star(w^\star, \hat{u}(w^\star))) \quad := T_1 \\
&\qquad + L_{\text{AL}}^{\text{model-free}}(w^\star, \hat{v}(w^\star), \hat{u}(w^\star), \zeta^\star(w^\star, \hat{u}(w^\star))) - \hat{L}_{\text{AL}}^{\text{model-free}}(w^\star, \hat{v}(w^\star), \hat{u}(w^\star), \zeta^\star(w^\star, \hat{u}(w^\star))) \quad := T_2 \\
&\qquad + \hat{L}_{\text{AL}}^{\text{model-free}}(w^\star, \hat{v}(w^\star), \hat{u}(w^\star), \zeta^\star(w^\star, \hat{u}(w^\star))) - \hat{L}_{\text{AL}}^{\text{model-free}}(w^\star, \hat{v}(w^\star), \hat{u}(w^\star), \hat{\zeta}(w^\star)) \quad := T_3 \\
&\qquad + \hat{L}_{\text{AL}}^{\text{model-free}}(w^\star, \hat{v}(w^\star), \hat{u}(w^\star), \hat{\zeta}(w^\star)) - \hat{L}_{\text{AL}}^{\text{model-free}}(\hat{w}, \hat{v}(\hat{w}), \hat{u}(\hat{w}), \hat{\zeta}(\hat{w})) \quad := T_4 \\
&\qquad + \hat{L}_{\text{AL}}^{\text{model-free}}(\hat{w}, \hat{v}(\hat{w}), \hat{u}(\hat{w}), \hat{\zeta}(\hat{w})) - \hat{L}_{\text{AL}}^{\text{model-free}}(\hat{w}, v^\star, u_{\hat{w}}^\star, \hat{\zeta}(\hat{w}, u_{\hat{w}}^\star)) \quad := T_5 \\
&\qquad + \hat{L}_{\text{AL}}^{\text{model-free}}(\hat{w}, v^\star, u_{\hat{w}}^\star, \hat{\zeta}(\hat{w}, u_{\hat{w}}^\star)) - L_{\text{AL}}^{\text{model-free}}(\hat{w}, v^\star, u_{\hat{w}}^\star, \hat{\zeta}(\hat{w}, u_{\hat{w}}^\star)) \quad := T_6 \\
&\qquad + L_{\text{AL}}^{\text{model-free}}(\hat{w}, v^\star, u_{\hat{w}}^\star, \hat{\zeta}(\hat{w}, u_{\hat{w}}^\star)) - L_{\text{AL}}^{\text{model-free}}(\hat{w}, v^\star, u_{\hat{w}}^\star, \zeta^\star(\hat{w})) \quad := T_7
\end{aligned}$$

We bound each term:

- $T_1 \leq 0$ because $v^\star, u^\star = \arg\min_{v,u} L_{\text{AL}}^{\text{model-free}}(w^\star, v, u, \zeta^\star(w^\star, u))$;

- $T_2 \leq \epsilon_{\text{model-free}}$ by Proposition (7);
- $T_3 \leq 0$ because $\hat{\zeta}(w^\star) = \hat{\zeta}(w^\star, \hat{u}(w^\star)) = \arg\max_{\zeta \in \mathcal{Z}} \hat{L}_{\text{AL}}^{\text{model-free}}(w^\star, \hat{v}(w^\star), \hat{u}(w^\star), \zeta)$;
- $T_4 \leq 0$ because $\hat{w} = \arg\max_{w \in \mathcal{W}} \hat{L}_{\text{AL}}^{\text{model-based}}(w, \hat{v}(w), \hat{u}(w), \hat{\zeta}(w))$;
- $T_5 \leq 0$ because $\hat{v}(\hat{w}), \hat{u}(\hat{w}) = \arg\min_{v \in \mathcal{V}, u \in \mathcal{U}} \hat{L}_{\text{AL}}^{\text{model_free}}(\hat{w}, v, u, \hat{\zeta}(\hat{w}, u))$;
- $T_6 \leq \epsilon_{\text{model-free}}$ by Proposition (7);
- $T_7 \leq 0$ because $\zeta^\star(\hat{w}) = \zeta^\star(\hat{w}, u^\star_{\hat{w}}) = \arg\max_{\zeta < 0} L_{\text{AL}}^{\text{model-free}}(\hat{w}, v^\star, u^\star_{\hat{w}}, \zeta)$.

□

D.8 Auxiliary results

Lemma 12 *For any $0 \leq b_i \leq B$ and $x_i, y_i \geq 0$ for $i \in \{0, \ldots, n\}$, the following holds*

$$\left(\sqrt{\sum_{i=0}^{n} b_i x_i} - \sqrt{\sum_{i=0}^{n} b_i y_i}\right)^2 \leq B \sum_{i=0}^{n} \left(\sqrt{x_i} - \sqrt{y_i}\right)^2. \tag{68}$$

Proof. We expend the left-hand side of (68), use Cauchy-Schwarz inequality, and then complete the square:

$$\begin{aligned}
\left(\sqrt{\sum_{i=1}^{n} b_i x_i} - \sqrt{\sum_{i=1}^{n} b_i y_i}\right)^2 &= \sum_i b_i x_i + \sum_i b_i y_i - 2\sqrt{\left(\sum_i b_i x_i\right)\left(\sum_i b_i y_i\right)} \\
&\leq \sum_i b_i x_i + \sum_i b_i y_i - 2\sum_i b_i \sqrt{x_i y_i} \\
&= \sum_i (\sqrt{b_i x_i} - \sqrt{b_i y_i})^2 \leq B \sum_i \left(\sqrt{x_i} - \sqrt{y_i}\right)^2
\end{aligned}$$

□

Lemma 13 *For any two arbitrary sets $\mathcal{X}, \mathcal{Y}$, let $f(\cdot, \cdot) : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be an arbitrary function. Let $\mathcal{X}_0 = \{x \in \mathcal{X} \mid \inf_{y \in \mathcal{Y}} f(x, y) > -\infty\}$ and assume $\mathcal{X}_0$ is non-empty. For any $x \in \mathcal{X}_0$, assume there exists $y^*(x) \in \mathcal{Y}$ s.t. $f(x, y^*(x)) = \min_{y \in \mathcal{Y}} f(x, y)$. Also, let $\mathcal{X}_p^* = \{x \in \mathcal{X}_0 \mid x \in \arg\max_{x \in \mathcal{X}_0} f(x, y^\star(x))\}$ and assume $\mathcal{X}_p^*$ is non-empty. For a nonnegative function $A(\cdot)$ on $\mathcal{X}$, let $\mathcal{X}^* = \{x \in \mathcal{X}_p^* \mid A(x) = 0\}$ and assume $\mathcal{X}^*$ is non-empty. Define $f^{AL}(x, y) = f(x, y) - A(x)$. Then*

$$x \in \mathcal{X}_0 \iff \inf_{y \in \mathcal{Y}} f^{AL}(x, y) > -\infty$$

and for any $x \in \mathcal{X}_0$,

$$x \in \mathcal{X}^* \iff x \in \arg\max_{x \in \mathcal{X}_0} \min_{y \in \mathcal{Y}} f^{AL}(x, y).$$

Proof. Note that for any fixed x, $f^{AL}(x, y)$ is a constant shift of $f(x, y)$, which implies that $\inf_{y \in \mathcal{Y}} f(x, y) > -\infty \iff \inf_{y \in \mathcal{Y}} f^{AL}(x, y) > -\infty$. This also implies that for any $x \in \mathcal{X}_0$, $\arg\min_{y \in \mathcal{Y}} f(x, y) = \arg\min_{y \in \mathcal{Y}} f^{AL}(x, y)$.

For any $x \in \mathcal{X}_0$, let $y^*(x)$ denote any one of $y \in \mathcal{Y}$ s.t. $f(x, y^*(x)) = \min_{y \in \mathcal{Y}} f(x, y)$.

Now for any $x^* \in \mathcal{X}^*$, we have

$$\begin{aligned}
& f(x^*, y^*(x^*)) \geq f(x, y^*(x)),\ \forall x \in \mathcal{X}_0. \\
\Longrightarrow & f(x^*, y^*(x^*)) - A(x^*) \geq f(x, y^*(x)) - A(x),\ \forall x \in \mathcal{X}_0. \\
\Longrightarrow & f^{AL}(x^*, y^*(x^*)) \geq f^{AL}(x, y^*(x)),\ \forall x \in \mathcal{X}_0. \\
\Longrightarrow & \min_{y \in \mathcal{Y}} f^{AL}(x^*, y) \geq \min_{y \in \mathcal{Y}} f^{AL}(x, y),\ \forall x \in \mathcal{X}_0. \\
\Longrightarrow & x^* \in \arg\max_{x \in \mathcal{X}_0} \min_{y \in \mathcal{Y}} f^{AL}(x, y).
\end{aligned}$$

For the other direction, given any $x_0 \in \arg\max_{x\in\mathcal{X}_0} \min_{y\in\mathcal{Y}} f^{AL}(x,y)$, we have

$$\begin{aligned} &\min_{y\in\mathcal{Y}} f^{AL}(x_0,y) \geq \min_{y\in\mathcal{Y}} f^{AL}(x,y),\ \forall x \in \mathcal{X}_0. \\ \Longrightarrow &f^{AL}(x_0,y^*(x_0)) \geq f^{AL}(x,y^*(x)),\ \forall x \in \mathcal{X}_0. \end{aligned} \tag{69}$$

Fix any $x^* \in \mathcal{X}^* \subseteq \mathcal{X}_0$, we have $f(x^*,y^*(x^*)) \geq f(x_0,y^*(x_0))$ and $-A(x^*) \geq -A(x_0)$ by definition. Now assume $x_0 \notin \mathcal{X}^*$. Then either $f(x^*,y^*(x^*)) > f(x_0,y^*(x_0))$ if $x_0 \notin \mathcal{X}_{\text{p}}^*$, or $-A(x^*) > -A(x_0)$ if $x_0 \in \mathcal{X}_{\text{p}}^* \backslash \mathcal{X}^*$. Either one of the above two conditions implies that

$$f(x^*,y^*(x^*)) - A(x^*) > f(x_0,y^*(x_0)) - A(x_0) \Longrightarrow f^{AL}(x^*,y^*(x^*)) > f^{AL}(x_0,y^*(x_0)),$$

which contradicts with (69). Therefore, $x_0 \in \mathcal{X}^*$. □