On the Complexity of Offline Reinforcement Learning with $Q^\star$-Approximation and Partial Coverage

Haolin Liu* **Braham Snyder*** **Chen-Yu Wei***
University of Virginia
{srs8rh, dqr2ye, kfw6en}@virginia.edu

Abstract

We study offline reinforcement learning under $Q^\star$-approximation and partial coverage, a setting that motivates practical algorithms such as Conservative Q-Learning (CQL) [KZTL20] but has received limited theoretical attention. Our work is inspired by the following open question: *Are $Q^\star$-realizability and Bellman completeness sufficient for sample-efficient offline RL under partial coverage?* We answer this question in the negative through an information-theoretic lower bound. To identify additional structure that enables sample-efficient offline RL under partial coverage, we introduce a general decision-estimation framework, inspired by model-free decision-estimation coefficients (DEC) for online RL [FGQ+23, LWZ25b]. Our framework decomposes the complexity of offline RL into two parts: the *decision complexity* and the *value estimation error*. This decomposition allows us to study the two sub-problems in a modular way. Our result not only unifies existing results in the $Q^\star$-approximation and partial coverage regime [CJ22, UKLS23], but further improves and generalizes them. On the decision complexity side, our improvement includes: the first ϵ^{-2} sample complexity bound for soft Q-learning under partial coverage that improves [UKLS23]'s ϵ^{-4} bound, the removal of the need for additional online interaction in the value-gap setting of [CJ22], and new learnable settings beyond the above two cases. On the value estimation side, we provide a new characterization of the role of Bellman completeness under partial coverage, and the first characterization of offline learnability for general low-Bellman-rank MDPs [JKA+17, DKL+21, JLM21]. The latter is a canonical online RL setting that has remained unexplored in offline RL except for special cases. As a side contribution, our techniques give the first analysis of CQL in the function approximation setting.

1 Introduction

Offline Reinforcement Learning (RL) studies policy learning from a fixed dataset, without interaction with the environment. This paradigm is appealing because it enables learning from data that may not have been generated by the learner itself. In practice, however, it is rarely realistic to assume that the offline data covers the entire state-action space. In such *partial coverage* settings, algorithms typically rely on *pessimism* to ensure safety against worst possible environments.

Among value-based methods for offline RL with pessimism, two classes of algorithms are widely used: Q-learning-based methods such as Conservative Q-Learning (CQL) [KZTL20], and actor-critic methods such as Behavior-Regularized Actor-Critic (BRAC) [WTN19]. From a theoretical perspective, these methods are commonly analyzed under the $Q^\star$-realizability and Q^π-realizability assumptions, respectively. Under Q^π-realizability, [XCJ+21] gives a relatively complete sample-complexity characterization. Under $Q^\star$-realizability, however, existing results remain less complete. On the lower bound side, it is known that $Q^\star$-realizability and partial coverage alone are not sufficient

*Authors are listed in alphabetical order.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

for sample-efficient learning [FKSLX22, JRSW24]. On the upper bound side, existing works [CJ22, UKLS23] rely on estimating additional quantities such as density ratios or Lagrange multipliers, and require extra conditions such as value gaps or behavior regularization. The fundamental roles of these additional assumptions remain unclear. More related works can be found in Appendix A.

In this work, we aim to provide a more comprehensive understanding of the potential and limitations of offline RL under $Q^\star$-approximation and partial coverage. Our contributions are twofold. First, we establish an information-theoretic lower bound showing that $Q^\star$-realizability, partial coverage, and Bellman completeness are *not* sufficient for sample-efficient learning, indicating that the difficulty of offline RL under partial coverage does not merely come from the estimation of $Q^\star$. Motivated by this observation, we introduce a decomposition of the performance gap into *decision complexity* and *estimation error*, which together capture the overall difficulty of the problem. This decomposition yields a modular framework that recovers, improves, and extends prior results. A detailed overview of our results is given in Section 2.

2 Results Overview

Lower bound (Section 4). We show that $Q^\star$-realizability, partial coverage, and Bellman completeness are *not* sufficient for sample-efficient offline RL. This is somewhat surprising because these assumptions allow the learner to construct a small confidence set containing the true $Q^\star$ function [JLM21, XFB+23, XCJ+21]. This indicates that even when the true value function is well localized, the learner may still be unable to decide which policy is safe to deploy. This motivates the need for an additional complexity measure that captures the difficulty of making robust decisions from a confidence set.

Upper bound: a decision-estimation decomposition (Section 5). To capture this missing decision difficulty, we introduce the offline robust decision-estimation coefficient, Ordec. Under $Q^\star$-realizability, the offline dataset induces a confidence set $\mathcal{F}_{\mathrm{conf}}$ of plausible optimal $Q^\star$-functions. This in turn induces a set $\mathcal{M}_{\mathrm{conf}} = \{M : Q^\star_M \in \mathcal{F}_{\mathrm{conf}}\}$ of plausible models, where $Q^\star_M$ denotes the optimal Q-function induced by model M. The learner must choose a policy that performs well in the true model, but the true model is only known to lie within the ambiguity set represented by $\mathcal{M}_{\mathrm{conf}}$.

Ordec measures the value of a game between a policy player and a model player. The *policy player* chooses a policy distribution ρ to minimize the suboptimality gap $J_M(\pi_M) - \mathbb{E}_{\pi\sim\rho}[J_M(\pi)]$, where $J_M(\pi)$ denotes the expected payoff of π under model M and π_M is the optimal policy under M. The *model player* chooses a model $M \in \mathcal{M}_{\mathrm{conf}}$ that makes this gap large. However, the payoff of the adversary is offset by a discrepancy term that penalizes models that are statistically inconsistent with the functions in the confidence set.

Formally, for parameter $\gamma > 0$ and discrepancy measure $D^\pi(f\|M)$ between function f and model M under the state-action distribution induced by π, we define

$$\textsf{Ordec-O}^D_\gamma(\mathcal{F}_{\mathrm{conf}}) := \min_{\rho\in\Delta(\Pi)} \max_{M\in\mathcal{M}_{\mathrm{conf}}} \mathbb{E}_{\pi\sim\rho}\left[J_M(\pi_M) - J_M(\pi) - \gamma \max_{f\in\mathcal{F}_{\mathrm{conf}}} D^{\pi_M}(f\|M)\right] \quad (1)$$

All remaining notation is introduced in Section 3. Our algorithm outputs a policy distribution $\hat\rho$ that is guaranteed to have (see Theorem 2)

$$\underbrace{\mathbb{E}_{\hat\pi\sim\hat\rho}[J_{M^\star}(\pi^\star) - J_{M^\star}(\hat\pi)]}_{\textbf{suboptimality}} \le \underbrace{\textsf{Ordec-O}^D_\gamma(\mathcal{F}_{\mathrm{conf}})}_{\textbf{decision complexity}} + \gamma \underbrace{\max_{f\in\mathcal{F}_{\mathrm{conf}}} D^{\pi^\star}(f\|M^\star)}_{\textbf{estimation error}} \quad (2)$$

where $M^\star$ is the underlying true model and $\pi^\star$ is the true optimal policy.

This is the main takeaway of the paper. Offline RL under partial coverage has two challenges: the learner must estimate the value function well enough to make the estimation error small, and the remaining ambiguity must still allow for a robust decision to make the decision complexity small. Conditions controlling the decision complexity and controlling the estimation error can be combined modularly. This modular view is summarized in Figure 1. For each term, we provide a list of conditions under which they can be bounded. This not only subsume all upper bound results we are aware of for offline RL with $Q^\star$-realizability and partial coverage under general function approximation, but also extend to new settings.

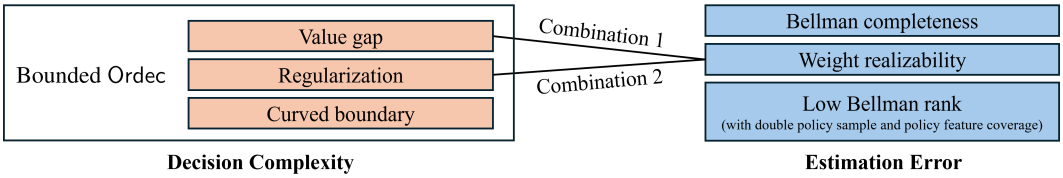


Figure 1: Overview of the decomposition under $Q^\star$ realizability and partial coverage. Assumptions in the decision complexity part can be combined with any assumptions in the estimation error part, leading to a learnable setting. Three decision complexity bounds are established in Theorem 3, Theorem 4, Theorem 5; three estimation error bounds are in Lemma 6, Lemma 7, Lemma 8. Combinations 1 and 2 are studied by [CJ22] and [UKLS23], respectively; our results improve over theirs in multiple aspects (see Section 2.1). Besides them, the other seven combinations are new.

Relation to online DEC. Our decomposition resembles the decision-estimation coefficient framework for online RL [FKQR21, FGH23, FGQ+23, LWZ25b]. The key difference is *under which state-action distribution* the discrepancy is measured. In online DEC, the discrepancy is measured under the learner's policy, because the learner can collect data using that policy. In offline RL, the data distribution is fixed, so the learner's policy cannot help test whether a model is plausible. Instead, Ordec measures the discrepancy under the candidate model's own optimal policy π_M. This is what makes the sub-optimality bound adaptive to partial coverage: models whose optimal policies are poorly supported by the offline data are penalized through the discrepancy term. We elaborate more in Section 5.2.

2.1 Comparison with prior work in offline RL with $Q^\star$-realizability and partial coverage

Despite the natural appeal of this setting, positive results for $Q^\star$-realizability and partial coverage are scarce. To our knowledge, only [CJ22] and [UKLS23] have identified learnability conditions for this setting under general function approximation. As discussed above, bounded decision complexity and the ability to accurately estimate value functions are both key to establishing learnability. While not explicitly stated, the results of [CJ22] and [UKLS23] each rely on a combination of conditions that ensure both terms are bounded, as indicated in Figure 1. In Figure 1, we list three sufficient conditions that allow us to control decision complexity and estimation error, respectively, resulting in nine combinations in total. Besides the two identified by [CJ22] and [UKLS23], the learnability of the remaining seven combinations has not been established before.

Improvement over [CJ22] and [UKLS23] (Section 6). Even under the same set of assumptions, our algorithm and the analysis framework give improved guarantees over [CJ22] and [UKLS23]. Specifically, in the unregularized value-gap setting (the first condition for decision complexity in Figure 1), our suboptimality guarantee automatically adapts to the gap of the true $Q^\star$ function without knowing the gap value, while [CJ22] requires either knowing the gap value or additional online interaction with the environment to achieve similar guarantee as ours. This improvement comes from our new algorithm design based on Ordec. In the regularized setting (the second condition for decision complexity in Figure 1), we obtain the first ϵ^{-2} sample complexity bound, improving the ϵ^{-4} bound of [UKLS23]. This improvement comes from a novel performance difference lemma we develop for regularized MDP with partial coverage. Besides, we relax the requirement of [UKLS23] that the offline data is generated from the reference policy in the regularization, and remove the polynomial dependence on the number of actions in their sample complexity bound.

New settings with bounded decision complexity (Section 6). Ordec not only recovers the value-gap and the regularization settings studied in [CJ22] and [UKLS23], but also extends to new settings. We provide a third condition that leads to bounded Ordec (left side of Figure 1): curved decision boundary. This addresses the concerns raised by [SZS+22, UKLS23] on the practicality of the value-gap assumption in [CJ22] for continuous action spaces.

New settings with bounded estimation error (Section 7). On the estimation error side, [CJ22] and [UKLS23] both rely on the weight realizability assumption to control the error, leaving other possibilities open. We provide two additional conditions: Bellman completeness and low Bellman rank (the right-hand side of Figure 1). While both are extensively studied under online RL and Bellman

completeness has been studied under offline RL with full coverage [CJ19], they are unexplored in offline RL with partial coverage. Specifically, partial coverage with Bellman completeness is an ideal theoretical framework for the practical CQL algorithm [KZTL20]. To our knowledge, no prior work has provided a sample complexity analysis for CQL beyond the tabular case, and our work achieves it for the first time. On the other hand, low Bellman rank is a canonical setting for online RL [JKA+17, DKL+21, JLM21] that remains unexplored in offline RL (even with full coverage) except for special cases. In fact, we show that under standard data assumptions and coverage definitions, low Bellman rank is insufficient to control the estimation error, thus remaining not learnable even with bounded decision complexity. We propose to replace the standard data and coverage assumptions with *double policy sampling* and *policy feature coverage*, which allow for efficient estimation error control. We also show that relaxing either assumption back to the standard definition breaks the learnability.

3 Notation and Setting

We deal with both unregularized and regularized MDPs. For simplicity, we formulate regularized MDPs, while unregularized MDPs are just special cases when the regularizer is zero.

Regularized Markov Decision Processes. We consider a finite-horizon *regularized* MDP defined by $M = (\mathcal{S}, \mathcal{A}, P_M, R_M, s_1, H, \psi)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $P_M : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition function, $R_M : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the reward function, and $\psi : \Delta(\mathcal{A}) \times \mathcal{S} \to \mathbb{R}_{\geq 0}$ is a state-dependent convex regularizer for action distributions. We assume a *layered* state space $\mathcal{S}$, i.e., $\mathcal{S} = \mathcal{S}_1 \cup \cdots \cup \mathcal{S}_H$ where $\mathcal{S}_h \cap \mathcal{S}_{h'} = \emptyset$ for any $h \neq h'$. Transitions only occur between adjacent layers: $P_M(s' \mid s, a) > 0$ only if $s \in \mathcal{S}_h$ and $s' \in \mathcal{S}_{h+1}$. Without loss of generality, let $\mathcal{S}_1 = \{s_1\}$. A policy π: $\mathcal{S} \to \Delta(\mathcal{A})$ maps states to action distributions. Let Π be the set of policies. For any policy π, the regularized state value function is $V_M^\pi(s) \triangleq \mathbb{E}_M^\pi \left[\sum_{h'=h}^H \left(R_M(s_{h'}, a_{h'}) - \psi(\pi(\cdot|s_{h'}); s_{h'}) \right) \middle| s_h = s \right]$ for $s \in \mathcal{S}_h$, where $\mathbb{E}_M^\pi[\cdot]$ denotes expectation over trajectories induced by π and P_M. Furthermore, denote $J_M(\pi) \triangleq V_M^\pi(s_1)$. By induction, the Bellman equations hold: $Q_M^\pi(s, a) = R_M(s, a) + \mathbb{I}[s \notin \mathcal{S}_H] \cdot \mathbb{E}_{s' \sim P_M(\cdot|s,a)}[V_M^\pi(s')]$ and $V_M^\pi(s) = \mathbb{E}_{a \sim \pi(\cdot|s)}[Q_M^\pi(s, a)] - \psi(\pi(\cdot|s); s)$.

The optimal regularized value functions are $Q_M^\star(s, a) = \max_\pi Q_M^\pi(s, a)$ and $V_M^\star(s) = \max_\pi V_M^\pi(s)$, and the optimal policy is $\pi_M \in \operatorname{argmax}_\pi J_M(\pi)$. We have $Q_M^\star(s, a) = R_M(s, a) + \mathbb{I}[s \notin \mathcal{S}_H] \cdot \mathbb{E}_{s' \sim P_M(\cdot|s,a)}[V_M^\star(s')]$, $V_M^\star(s) = \max_{p \in \Delta(\mathcal{A})} \left\{ \mathbb{E}_{a \sim p} \left[Q_M^\star(s, a) \right] - \psi(p; s) \right\}$ and $\pi_M(\cdot \mid s) \in \operatorname{argmax}_{p \in \Delta(\mathcal{A})} \left\{ \mathbb{E}_{a \sim p} \left[Q_M^\star(s, a) \right] - \psi(p; s) \right\}$.

For a policy π, denote the occupancy measure as $d_M^\pi(s) \triangleq \mathbb{E}_M^\pi[\mathbb{I}\{s_h = s\}]$ for $s \in \mathcal{S}_h$, and $d_M^\pi(s, a) \triangleq d_M^\pi(s)\pi(a|s)$. Though d_M^π is not a distribution as $\sum_{s \in \mathcal{S}} d_M^\pi(s) = H$, we denote $\mathbb{E}_{s \sim d^\pi}[g(s)] \triangleq \sum_s d^\pi(s) g(s)$ for any function g. Assume the true environment follows MDP $M^\star = (\mathcal{S}, \mathcal{A}, P, R, s_1, H, \psi)$, and for simplicity, we write $V^\pi(s) \triangleq V_{M^\star}^\pi(s)$, $Q^\pi(s, a) \triangleq Q_{M^\star}^\pi(s, a)$, $V^\star(s) \triangleq V_{M^\star}^\star(s)$, $Q^\star(s, a) \triangleq Q_{M^\star}^\star(s, a)$, $J(\pi) \triangleq J_{M^\star}(\pi)$, $\pi^\star \triangleq \pi_{M^\star}$, and $d^\pi \triangleq d_{M^\star}^\pi$.

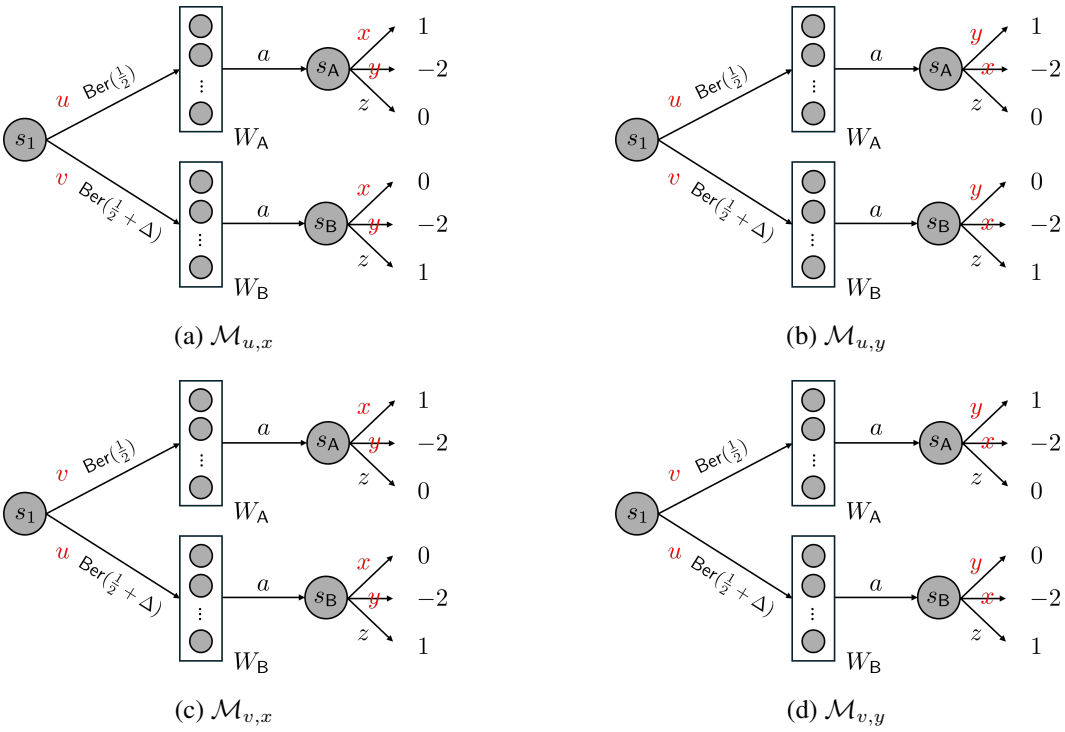
Offline Reinforcement Learning. In offline RL, the learner only has access to a pre-collected dataset $\mathcal{D}$ from MDP $M^\star$ and cannot interact with the environment. The dataset $\mathcal{D}$ consists of n i.i.d. tuples (s, a, r, s'), where $(s, a) \sim \mu$ for some unknown distribution $\mu \in \Delta(\mathcal{S} \times \mathcal{A})$, $\mathbb{E}[r|s, a] = R(s, a)$, and $s' \sim P(\cdot|s, a)$. The goal is to learn a policy $\hat{\pi}$ that minimizes the sub-optimality $J(\pi^\star) - J(\hat{\pi})$. We seek *coverage-adaptive* performance, where the sub-optimality adapts to the degree to which μ covers $\pi^\star$ (Definition 1).

Definition 1 (Coverage)**.** *For any policy π, define* $C^\pi = \max_{(s,a) \in \mathcal{S} \times \mathcal{A}} \frac{d^\pi(s,a)}{H\mu(s,a)}$.[2]

In value-based offline RL, function approximation is used to approximate value functions. In this paper, we focus on the $Q^\star$-approximation scheme and assume $Q^\star$-realizability:

Definition 2 ($Q^\star$-realizability)**.** *The learner has access to a function class $\mathcal{F} \subset \{f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}\}$ such that $Q^\star \in \mathcal{F}$. For any $f \in \mathcal{F}$, denote $f(s) = \max_{p \in \Delta(\mathcal{A})} \left\{ \mathbb{E}_{a \sim p} \left[f(s, a) \right] - \psi(p; s) \right\}$ and denote $\pi_f(\cdot \mid s) = \operatorname{argmax}_{p \in \Delta(\mathcal{A})} \left\{ \mathbb{E}_{a \sim p} \left[f(s, a) \right] - \psi(p; s) \right\}$ with arbitrary tie-breaking.*

[2]The division by H is to correct the mismatch between $\sum_{s,a} d^\pi(s, a) = H$ and $\sum_{s,a} \mu(s, a) = 1$.



(a) $\mathcal{M}_{u,x}$ (b) $\mathcal{M}_{u,y}$ (c) $\mathcal{M}_{v,x}$ (d) $\mathcal{M}_{v,y}$

Figure 2: The four MDP families used in the lower-bound construction. The first subscript denotes the action in $\{u, v\}$ that leads to W_{A}, and the other action leads to W_{B}. The second subscript denotes the action in $\{x, y\}$ that receives reward 1 at s_{A}. Red labels mark the components that vary across the four families.

Bellman Operator. Define the Bellman operator $\mathcal{T}$ as $\mathcal{T}f(s,a) = R(s,a) + \mathbb{E}_{s'\sim P(\cdot|s,a)}[f(s')]$.

Regularizer. For convex function $\Phi : \Delta(\mathcal{A}) \to \mathbb{R}$, let $\mathrm{Breg}_{\Phi}(x,y) \triangleq \Phi(x) - \Phi(y) - \langle \nabla\Phi(y), x-y\rangle$ be the Bregman divergence defined with Φ. We write $\mathrm{Breg}_{\psi}(x,y;s) \triangleq \mathrm{Breg}_{\psi(\cdot;s)}(x,y)$ for any $x, y \in \Delta(\mathcal{A})$ and $\mathrm{Breg}_{\psi}(\pi,\pi';s) \triangleq \mathrm{Breg}_{\psi(\cdot;s)}(\pi(\cdot|s),\pi'(\cdot|s))$ for any policies π, π'. We focus on regularizer $\psi(p;s) = \alpha\mathrm{Breg}_{\Phi}(p;\pi_{\mathrm{ref}}(\cdot|s))$ where Φ is a convex function, π_{ref} is a given reference policy, and $\alpha \geq 0$ is the weight of regularization. We also write $\psi(\pi(\cdot|s);s) \triangleq \psi(\pi;s)$.

4 An Information-Theoretic Lower Bound

We begin with an information-theoretic lower bound showing that $Q^\star$-realizability, partial coverage, and Bellman completeness (defined below) are insufficient for sample-efficient offline RL.

Assumption 1 (Bellman completeness)**.** *Along with the $\mathcal{F}$ in Definition 2, the learner has access to a function class $\mathcal{G} \subset \{g : \mathcal{S}\times\mathcal{A} \to \mathbb{R}\}$ such that $\mathcal{T}f \in \mathcal{G}$ for any $f \in \mathcal{F}$.*

Theorem 1. *There exists a family of MDPs $\mathcal{M}$, a function class $\mathcal{F}$ with $|\mathcal{F}| = 4$, and an offline data distribution μ such that under any true model $M^\star \in \mathcal{M}$, $Q^\star$-realizability and Bellman completeness hold (with $\mathcal{G} = \mathcal{F}$), and has coverage $C^{\pi^\star} = \Theta(1)$. However, to achieve $\mathbb{E}[J(\pi^\star) - J(\hat{\pi})] \leq \epsilon$ with $M^\star \sim \mathsf{Unif}(\mathcal{M})$, the learner must access $n \geq \Omega\big(\frac{1}{\epsilon}\min\big\{\sqrt{|\mathcal{S}|}, \frac{1}{\Delta^2}\big\}\big)$ offline samples, where $\Delta = \min_s(V^\star(s) - \max_{a\neq\pi^\star(s)} Q^\star(s,a))$ is a parameter which can be chosen arbitrarily in $[0, \frac{1}{4}]$, and the number of states $|\mathcal{S}|$ can be arbitrarily large.*

Even if the learner can access trajectory data $(s_1, a_1, r_1, \cdots, s_H, a_H, r_H)$ from an offline policy π_b such that $\max_{s,a}\frac{d^{\pi^\star}(s,a)}{d^{\pi_b}(s,a)} = \Theta(1)$, $n \geq \Omega\big(\frac{1}{\epsilon\,\mathrm{poly}(H)}\min\{2^H, \sqrt{|\mathcal{S}|}, \frac{1}{\Delta^2}\}\big)$ trajectories are required.

Proof sketch. We first prove a lower bound of $\min\{\sqrt{|\mathcal{S}|}, 1/\Delta^2\}$ without the ϵ^{-1} factor (Lemma 12), under the (s,a,r,s') data format. The construction uses the four MDP families in Figure 2, with $M^\star$

drawn uniformly. At the initial state s_1, actions u and v yield rewards $\mathsf{Ber}(1/2)$ and $\mathsf{Ber}(1/2+\Delta)$, with the higher-reward action leading uniformly to the group W_B, and the other to W_A. From W_A and W_B, the unique action a transitions to s_A and s_B, respectively, with zero reward. At $s_\mathsf{A}, s_\mathsf{B}$, the three terminal actions have deterministic rewards as shown in Figure 2. Thus, $\pi^\star$ chooses the $\mathsf{Ber}(1/2+\Delta)$ action at s_1 and then chooses z on s_B. The function class $\mathcal{F}$, with $|\mathcal{F}| = 4$, satisfies both $Q^\star$-realizability and Bellman completeness as shown in Appendix B.2. The offline distribution μ is induced by a behavior policy that chooses u, v uniformly at s_1 and chooses z at $s_\mathsf{A}, s_\mathsf{B}$, so $C^{\pi^\star} \le 2$.

Since the learner observes only transition tuples (s, a, r, s'), not trajectories, the data consist of three types: **Type 1** samples (s_1, a_1, r_1, w_1), where $a_1 \sim \mathsf{Unif}(\{u, v\})$ and $w_1 \in W_\mathsf{A} \cup W_\mathsf{B}$; **Type 2** samples $(w_2, a, 0, s_2)$, where $w_2 \sim \mathsf{Unif}(W_\mathsf{A} \cup W_\mathsf{B})$ and $s_2 \in \{s_\mathsf{A}, s_\mathsf{B}\}$; and **Type 3** samples (s_3, z, r_3), where $s_3 \sim \mathsf{Unif}(\{s_\mathsf{A}, s_\mathsf{B}\})$. To act optimally, the learner must decide whether u or v leads to W_B. Type 1 rewards can reveal this, but distinguishing means $1/2$ from $1/2+\Delta$ requires $\Omega(1/\Delta^2)$ samples. The learner could instead try to infer the group of w_1, but $\mathcal{F}$ assigns the same value $f(w, a) = 1$ to every $w \in W_\mathsf{A} \cup W_\mathsf{B}$, so the function class itself does not reveal the hidden partition. Type 2 samples do reveal whether a sampled w_2 lies in W_A or W_B, but when the groups are large, the learner is unlikely to see the same state in $W_\mathsf{A} \cup W_\mathsf{B}$ twice in Type 1 and Type 2 before $\Omega(\sqrt{|\mathcal{S}|})$ samples. Finally, the learner should avoid x and y at $s_\mathsf{A}, s_\mathsf{B}$, since they are never observed offline and one of them has reward -2. Overall, it is hard for the learner to determine between u and v, and the learner should only choose z on s_A and s_B. A random guess at s_1 obtains value $1/2 + \Delta/2 + 1/2$, which is worse than $\pi^\star$ by $1/2 + \Delta/2 = \Theta(1)$. In Theorem 1, we extend this construction to obtain the ϵ-dependent and trajectory-feedback lower bounds; see Appendix B. □

We note that our gap-dependent lower bound is quite different from those in online RL or simpler offline RL settings [WCD22, NTYG+23]. In those settings, if the learner cannot tell apart two actions with gap Δ, the sub-optimality suffered is only of order Δ. In our case, before the learner tells apart two actions of gap Δ, the learner suffers $\Theta(1)$ sub-optimality. This difference also shows in the sample complexity bounds: in prior settings, the learner can achieve $O(\frac{1}{\epsilon^2})$ sample complexity regardless of the value gap. In contrast, the sample complexity in Theorem 1 scales with $\frac{1}{\epsilon\Delta^2}$, which can become arbitrarily large even under constant ϵ.

5 Making Robust Decisions from a Confidence Set

In this section, we isolate the decision-making part of offline RL, assuming that the offline data has already been used to construct a confidence set $\mathcal{F}_{\mathrm{conf}}$ of plausible $Q^\star$ functions. The key question is how the learner should choose a policy from this ambiguity. We show that the usual value-centric pessimism used in $Q^\star$-based offline RL can be suboptimal, and we derive a more robust decision rule using Ordec. The construction of $\mathcal{F}_{\mathrm{conf}}$ is an estimation problem and is discussed in Section 7.

5.1 Two Pessimism Principles

We identify two distinct pessimism principles in prior work to derive a policy from the confidence set. The first, which we refer to as **policy-centric pessimism**, has been adopted in model-approximation [US21] and Q^π-approximation schemes [XCJ+21]. This approach aligns with the *robust optimization* principle [GYDH15], where the learner outputs a policy $\hat{\pi}$ that maximizes the worst-case performance over the uncertainty set of models $\mathcal{M}_{\mathrm{conf}}$:

$$\hat{\pi} = \mathrm{argmax}_\pi \min_{M \in \mathcal{M}_{\mathrm{conf}}} J_M(\pi) \tag{3}$$

However, prior work on $Q^\star$-approximation often follows a different **value-centric pessimism** principle, as seen in [CJ22]. Under this principle, the order between $\min_M$ and $\max_\pi$ are swapped:

$$\hat{M} = \mathrm{argmin}_{M \in \mathcal{M}_{\mathrm{conf}}} \max_\pi J_M(\pi), \qquad \text{then let } \hat{\pi} = \pi_{\hat{M}}. \tag{4}$$

In fact, this principle is more commonly written as:

$$\hat{f} \in \mathrm{argmin}_{f \in \mathcal{F}_{\mathrm{conf}}} f(s_1), \qquad \text{then let } \hat{\pi} = \pi_{\hat{f}}. \tag{5}$$

The equivalence between (4) and (5) can be seen as the following: with a confidence set $\mathcal{F}_{\mathrm{conf}}$ for $Q^\star$, we can define a confidence set for $M^\star$: $\mathcal{M}_{\mathrm{conf}} = \{M : Q^\star_M \in \mathcal{F}_{\mathrm{conf}}\}$. Thus, $\min_{f \in \mathcal{F}_{\mathrm{conf}}} f(s_1) = \min_{M \in \mathcal{M}_{\mathrm{conf}}} V^\star_M(s_1) = \min_{M \in \mathcal{M}_{\mathrm{conf}}} \max_\pi J_M(\pi)$.

This deviation from robust optimization is not accidental. In model- or Q^π-approximation, the learner can evaluate $J_M(\pi)$ for an arbitrary policy π. Under $Q^\star$-approximation, the confidence set only represents optimal value functions. Therefore, for a fixed policy π, $J_M(\pi)$ is generally not available from $\mathcal{F}_{\text{conf}}$ unless $\pi = \pi_M$. This forces standard methods to treat $J_M(\pi_M) = \max_\pi J_M(\pi)$ as the basic object. However, value-centric pessimism can produce less desirable decision, as in Example 1.

Example 1. *Assume a single state s_1 and two actions x, y. Let $\mathcal{F}_{\text{conf}} = \{f_x, f_y\}$ with $f_x(s_1, x) = 1$, $f_x(s_1, y) = 0$, $f_y(s_1, x) = \frac{1}{2} - \Delta$, $f_y(s_1, y) = \frac{1}{2} + \Delta$ for some $\Delta \in (0, 0.01]$. Policy-centric pessimism (3) chooses action $\operatorname{argmax}_a \min_{f \in \{f_x, f_y\}} f(s_1, a) = x$. Value-centric pessimism (4) chooses function $\operatorname{argmin}_{f \in \{f_x, f_y\}} \max_a f(s_1, a) = f_y$ and then chooses action $\operatorname{argmax}_a f_y(s_1, a) = y$. If the true environment is described by f_x, then the sub-optimality of policy-centric pessimism and value-centric pessimism is 0 and 1, respectively; if the true environment is described by f_y, then their sub-optimality is $2\Delta \le 0.02$ and 0, respectively.*

In Example 1, policy-centric pessimism sacrifices a negligible 2Δ in one world to avoid a catastrophic error (gap of 1) in the other—a robust behavior we usually favor in offline RL.

The question is therefore whether one can recover policy-centric robustness while still working only with a $Q^\star$ confidence set. In the next subsection, we answer this affirmatively. The framework we develop draws inspiration from the *model-free decision-estimation coefficient* (DEC) [FGQ+23, LWZ25a, LWZ25b].

5.2 A Decision-Estimation Coefficient for Offline RL under Partial Coverage

We first review ideas in the existing online DEC framework. In model-based DEC [FKQR21, FGH23], the learner seeks a policy that minimizes regret against an adversarially chosen model. However, this model must remain close—measured under the state-action distribution induced by the learner's current policy—to the posterior of $M^\star$ estimated from collected data. To achieve this, every model is penalized according to its discrepancy from the model posterior. This corresponds to the first decision rule in Figure 3. Model-free DEC [FGQ+23, LWZ25b] follows a similar principle, but instead of maintaining a posterior of $M^\star$, it maintains a posterior of $Q^\star$. The learner's policy is optimized against the *induced model set* $\mathcal{M} = \{Q^\star_M \in \mathcal{F}\}$. This gives the second decision rule in Figure 3.

Online model-based DEC:	$\min_{\rho \in \Delta(\Pi)} \max_{M \in \mathcal{M}} \mathbb{E}_{\pi \sim \rho}\Big[J_M(\pi_M) - J_M(\pi) - \gamma \mathbb{E}_{\overline{M} \sim \text{posterior}}[D^\pi(\overline{M} \Vert M)]\Big]$
Online model-free DEC:	$\min_{\rho \in \Delta(\Pi)} \max_{M \in \mathcal{M}} \mathbb{E}_{\pi \sim \rho}\Big[J_M(\pi_M) - J_M(\pi) - \gamma \mathbb{E}_{f \sim \text{posterior}}[D^\pi(f \Vert M)]\Big]$
Offline model-free DEC:	$\min_{\rho \in \Delta(\Pi)} \max_{M \in \mathcal{M}_{\text{conf}}} \mathbb{E}_{\pi \sim \rho}\Big[J_M(\pi_M) - J_M(\pi) - \gamma \max_{f \in \mathcal{F}_{\text{conf}}} D^{\pi_M}(f \Vert M)\Big]$

Figure 3: Comparison of DEC objectives in online and offline settings for discrepancy measure D.

In offline RL, the $Q^\star$ confidence set plays a similar role as the $Q^\star$ posterior in the online setting. Similarly, the learner seeks a policy that minimizes the sub-optimality against an adversarial model that remains in the confidence set. A key difference is that now this closeness is measured under the state-action distribution induced by each model's *own optimal policy*. First, unlike in online RL, the learner's output policy cannot generate new data and therefore cannot be used to test whether a candidate model is plausible. Thus, the discrepancy term should not be measured under the learner's policy. Second, measuring discrepancy under π_M penalizes candidate models whose own optimal policies are poorly supported by the offline data, which is exactly what is needed for coverage-adaptive guarantees under partial coverage.

To measure the closeness between a model M and a confidence function set under policy π_M, we use $\max_{f \in \mathcal{F}_{\text{conf}}} D^{\pi_M}(f \Vert M)$ with standard discrepancy functions D such as the Bellman error (will be clarified in Section 6). The intuition is that for the true model $M^\star$, $\max_{f \in \mathcal{F}_{\text{conf}}} D^{\pi^\star}(f \Vert M^\star)$ is expected to be small (usually of order $C^{\pi^\star}/n$), as any function f that deviates significantly from $M^\star$ should have already been excluded from the confidence set $\mathcal{F}_{\text{conf}}$. Taking the maximum over $f \in \mathcal{F}_{\text{conf}}$ gives the strongest test of whether the candidate model is inconsistent with the confidence set. This gives to the third decision rule in Figure 3.

This motivates the design of E2D.OR (Algorithm 1). We provide two versions of it, different in the way of trading the sub-optimality term and the discrepancy term. The offset version aligns with

Algorithm 1 Offline Robust Estimation-to-Decision (E2D.OR)

Input: Confidence set $\mathcal{F}_{\mathrm{conf}}$, divergence measure $D^{\pi}(f\|M)$, parameter γ for the offset version. Define $\mathcal{M}_{\mathrm{conf}} = \{M : Q^\star_M \in \mathcal{F}_{\mathrm{conf}}\}$ and compute

$$\hat{\rho} = \begin{cases} \operatorname*{argmin}_{\rho\in\Delta(\Pi)} \max_{M\in\mathcal{M}_{\mathrm{conf}}} \mathbb{E}_{\pi\sim\rho}\left[J_M(\pi_M) - J_M(\pi) - \gamma \max_{f\in\mathcal{F}_{\mathrm{conf}}} D^{\pi_M}(f\|M)\right] & \text{(offset version)} \\ \operatorname*{argmin}_{\rho\in\Delta(\Pi)} \max_{M\in\mathcal{M}_{\mathrm{conf}}} \dfrac{J_M(\pi_M) - \mathbb{E}_{\pi\sim\rho}[J_M(\pi)]}{\left(\max_{f\in\mathcal{F}_{\mathrm{conf}}} D^{\pi_M}(f\|M)\right)^{1/2}}. & \text{(ratio version)} \end{cases}$$

Output: mixture policy $\hat{\pi} \sim \hat{\rho}$.

online DEC, and could give a better bound when the discrepancy term may approach zero. The ratio version, however, can obtain better dependence on $C^{\pi^\star}$ when it is unknown. We define the following complexity (Offline Robust DEC) that naturally arise from the algorithm:

$$\textsf{Ordec-O}^D_\gamma(\mathcal{F}_{\mathrm{conf}}) := \min_{\rho\in\Delta(\Pi)} \max_{M\in\mathcal{M}_{\mathrm{conf}}} \mathbb{E}_{\pi\sim\rho}\Big[J_M(\pi_M) - J_M(\pi) - \gamma \max_{f\in\mathcal{F}_{\mathrm{conf}}} D^{\pi_M}(f\|M)\Big], \quad (6)$$

$$\textsf{Ordec-R}^D(\mathcal{F}_{\mathrm{conf}}) := \min_{\rho\in\Delta(\Pi)} \max_{M\in\mathcal{M}_{\mathrm{conf}}} \frac{J_M(\pi_M) - \mathbb{E}_{\pi\sim\rho}[J_M(\pi)]}{\left(\max_{f\in\mathcal{F}_{\mathrm{conf}}} D^{\pi_M}(f\|M)\right)^{1/2}}. \quad (7)$$

It holds that $\textsf{Ordec-O}^D_\gamma(\mathcal{F}_{\mathrm{conf}}) \le \frac{4}{\gamma}\big(\textsf{Ordec-R}^D(\mathcal{F}_{\mathrm{conf}})\big)^2$ (Lemma 13). With them, we have:

Theorem 2. *If $Q^\star \in \mathcal{F}_{\mathrm{conf}}$, then* E2D.OR *(Algorithm 1) ensures*

(offset version) $\quad J(\pi^\star) - \mathbb{E}[J(\hat{\pi})] \le \textsf{Ordec-O}^D_\gamma(\mathcal{F}_{\mathrm{conf}}) + \gamma \max_{f\in\mathcal{F}_{\mathrm{conf}}} D^{\pi^\star}(f\|M^\star)$,

(ratio version) $\quad J(\pi^\star) - \mathbb{E}[J(\hat{\pi})] \le \textsf{Ordec-R}^D(\mathcal{F}_{\mathrm{conf}})\big(\max_{f\in\mathcal{F}_{\mathrm{conf}}} D^{\pi^\star}(f\|M^\star)\big)^{1/2}$.

Comparison with value-centric pessimism. Following a similar analysis as in [CJ22], we can also express the suboptimality bound achieved by value-centric pessimism (either (4) or (5)) as $J(\pi^\star) - J(\hat{\pi}) \le \textsf{Gdec}^D(\mathcal{F}_{\mathrm{conf}})(\max_{f\in\mathcal{F}_{\mathrm{conf}}} D^{\pi^\star}(f\|M^\star))^{1/2}$, with

$$\textsf{Gdec}^D(\mathcal{F}_{\mathrm{conf}}) := \max_{M\in\mathcal{M}_{\mathrm{conf}}} \frac{J_M(\pi_M) - J_M(\pi_{\hat{f}})}{\left(D^{\pi_M}(\hat{f}\|M)\right)^{1/2}}, \quad \text{where } \hat{f} = \operatorname*{argmin}_{f\in\mathcal{F}_{\mathrm{conf}}} f(s_1) \quad (8)$$

See Theorem 14 in Appendix D for a derivation. Comparing (7) and (8), it is clear that $\textsf{Ordec-R}^D \le \textsf{Gdec}^D$, and the key improvement is that $\textsf{Ordec-R}^D$ goes through additional optimization over ρ and f. The following is a concrete and intuitive example for the gap between $\textsf{Ordec-R}^D$ and $\textsf{Gdec}^D$:

Example 2. *Consider the case with a single state s_1 and three actions x, y, z. Let $\mathcal{F}_{\mathrm{conf}} = \{f_x, f_y\}$ with $f_x(s_1, x) = 1, f_x(s_1, y) = 0, f_x(s_1, z) = 1 - \Delta$ and $f_y(s_1, x) = 0, f_y(s_1, y) = 1, f_y(s_1, z) = 1 - \Delta$ for some $0 < \Delta \le 0.01$. It can be verified that $\textsf{Ordec-O}^D_\gamma(\mathcal{F}_{\mathrm{conf}}) \le \Delta - \gamma$, $\textsf{Ordec-R}^D(\mathcal{F}_{\mathrm{conf}}) \le \Delta$, and $\textsf{Gdec}^D(\mathcal{F}_{\mathrm{conf}}) = 1$ with $D^a(f\|M) = (f(s_1, a) - R_M(s_1, a))^2$.*

In this example, clearly, choosing x or y is risky; choosing z, while always being sub-optimal, is the safest choice, and that is exactly the decision that will be made by E2D.OR.

6 Bounding the Decision Complexity Ordec

In this section, we provide examples where Ordec is bounded. As the lower bound in Section 4 shows, $Q^\star$-approximation with partial coverage is a challenging or even impossible regime, especially when the optimal value function has a very small or zero value gap. Therefore, even with the more refined decision principle developed in Section 5, some form of gap conditions or assumptions on the uniqueness of the optimal action is unavoidable. Throughout this section, we consider the average Bellman error: $D^\pi_{\mathsf{av}}(f\|M) = \big(\mathbb{E}_{(s,a)\sim d^\pi_M}\left[f(s,a) - R_M(s,a) - \mathbb{E}_{s'\sim P_M(\cdot|s,a)}\left[f(s')\right]\right]\big)^2$. We denote $\textsf{Ordec-O}^{\mathsf{av}}_\gamma$ for $\textsf{Ordec-O}^{D_{\mathsf{av}}}_\gamma$ (similar for others). See Appendix E for proofs in this section.

6.1 Unregularized Case ($\psi \equiv 0$)

When $\psi \equiv 0$, we recover and improve the gap bound in [CJ22]. Define

Definition 3. $\Delta_f \triangleq \min_s \big(f(s, \pi_f(s)) - \max_{a \neq \pi_f(s)} f(s,a)\big)$.

Theorem 3 (Value gap). *Assume $Q^\star \in \mathcal{F}_{\mathrm{conf}}$. We have*

- $\mathsf{Ordec\text{-}O}^{\mathsf{av}}_{\gamma}(\mathcal{F}_{\mathrm{conf}}) \le \frac{H^2}{2\gamma\Delta^2_{Q^\star}} + \gamma \max_{f \in \mathcal{F}_{\mathrm{conf}}} D^{\pi^\star}_{\mathsf{av}}(f \| M^\star)$
- $\mathsf{Ordec\text{-}R}^{\mathsf{av}}(\mathcal{F}_{\mathrm{conf}}), \mathsf{Gdec}^{\mathsf{av}}(\mathcal{F}_{\mathrm{conf}}) \le \frac{H}{\Delta_{\hat{f}}}$ *where* $\hat{f} = \operatorname{argmin}_{f \in \mathcal{F}_{\mathrm{conf}}} f(s_1)$.

A key advantage of $\mathsf{Ordec\text{-}O}^{\mathsf{av}}_{\gamma}(\mathcal{F}_{\mathrm{conf}})$ is that it adapts to the true value gap $\Delta_{Q^\star}$. For fixed γ, combining with Theorem 2 gives an algorithm whose suboptimality guarantee scales with $\Delta^{-2}_{Q^\star}$. In contrast, the suboptimality of algorithm in [CJ22] (value-centric pessimism) scales with $\mathsf{Gdec}^{\mathsf{av}}(\mathcal{F}_{\mathrm{conf}})$, which further scales with $\Delta^{-1}_{\hat{f}}$. This factor is not directly controllable and may be much larger than $\Delta^{-2}_{Q^\star}$.

We notice that [SZS+22] and [UKLS23] raise concerns on the applicability of the gap assumption [CJ22] in continuous action space. Our complexities poses no requirement for discrete actions or strict value gaps. In Theorem 4, we provide a continuous-action example with bounded Ordec.

Theorem 4 (Curved decision boundary). *Consider the linear setting with given feature mapping $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$. Assume 1) $\exists \theta^\star \in \mathbb{R}^d$ such that $Q^\star(s,a) = \phi(s,a)^\top \theta^\star$, 2) for any policy π and any action pair $a, b \in \mathcal{A}$, $|Q^\pi(s,a) - Q^\pi(s,b)| \le L\|\phi(s,a) - \phi(s,b)\|$, and 3) for all s, $\phi(s,\cdot) \subset \mathbb{R}^d$ is a set with the following property: for some $\kappa > 0, \beta \ge 1$ (denote $\pi^\star(s) = \operatorname{argmax}_a \phi(s,a)^\top \theta^\star$):*

$$\forall a \qquad (\phi(s,\pi^\star(s)) - \phi(s,a))^\top \theta^\star \ge \kappa \|\phi(s,\pi^\star(s)) - \phi(s,a)\|^\beta. \tag{9}$$

Then $\mathsf{Ordec\text{-}O}^{\mathsf{av}}_{\gamma}(\mathcal{F}_{\mathrm{conf}}) \le O\big(\big(\gamma^{-1}\kappa^{-2}L^{2\beta}\big)^{\frac{1}{2\beta-1}}\big) + \gamma \max_{f \in \mathcal{F}_{\mathrm{conf}}} D^{\pi^\star}_{\mathsf{av}}(f \| M^\star)$.

6.2 Regularized Case ($\psi \not\equiv 0$)

Besides value gaps (Theorem 3) and decision sets with curvature (Theorem 4), regularization provides another way to bound the decision complexity. We consider regularizers satisfying Assumption 2.

Assumption 2. *For every $s \in \mathcal{S}$, $\psi(\cdot; s)$ is Legendre (Definition 7). There exist constants $C_1^\psi, C_2^\psi > 0$ such that, for all $s \in \mathcal{S}$ and $f_1, f_2 \in \mathcal{F}$, writing $\pi_i = \pi_{f_i}$, $\mathrm{Breg}_\psi(\pi_2, \pi_1; s) \le C_1^\psi \mathrm{Breg}_\psi(\pi_1, \pi_2; s)$ and $\mathrm{KL}\left(\pi_1(\cdot|s) \| \pi_2(\cdot|s)\right) \le C_2^\psi \mathrm{Breg}_\psi(\pi_1, \pi_2; s)$.*

For example, KL regularization $\psi(p; s) = \alpha \mathrm{KL}\left(p \| \pi_{\mathrm{ref}}(\cdot|s)\right)$, as studied by [UKLS23], satisfies Assumption 2 with $C_1^\psi = 1 + 4H/\alpha$ and $C_2^\psi = 1/\alpha$. Tsallis and log-barrier regularizers are also covered; their constants and verification are provided in Lemma 35 in Appendix I.

Theorem 5 (Regularization). *Let regularizer ψ satisfy Assumption 2. Then* $\mathsf{Ordec\text{-}O}^{\mathsf{av}}_{\gamma}(\mathcal{F}_{\mathrm{conf}}) \le 36\gamma^{-1}(C_1^\psi(1 + H^3 C_2^\psi))^2$ *and* $\mathsf{Ordec\text{-}R}^{\mathsf{av}}(\mathcal{F}_{\mathrm{conf}}), \mathsf{Gdec}^{\mathsf{av}}(\mathcal{F}_{\mathrm{conf}}) \le 3C_1^\psi(1 + H^3 C_2^\psi)$.

Theorem 5 together with $\max_{f \in \mathcal{F}_{\mathrm{conf}}} D^{\pi^\star}_{\mathsf{av}}(f \| M^\star) \lesssim C^{\pi^\star}/n$ (Lemma 6 or Lemma 7) implies $\sqrt{C^{\pi^\star}/n}$ suboptimality gap, which provides the first $1/\epsilon^2$ sample complexity for this setting. To achieve this, the key technique in proving Theorem 5 is a new *second-order performance difference lemma* for regularized RL (Lemma 18). In the unregularized case, performance gaps between $\pi^\star$ and π_f are usually controlled by *first-order* divergence such as $D_{\mathrm{TV}}(\pi^\star \| \pi_f)$. In contrast, regularization introduces curvature that allows a *second-order* divergence. Specifically, we show $J(\pi^\star) - J(\pi_f) \lesssim \mathbb{E}_{s \sim d^{\pi^\star}}\left[\mathrm{Breg}_\psi(\pi_f, \pi^\star; s)\right]$ where the right-hand side scales quadratically with the policy difference. While [UKLS23] also studies regularized MDP, they bound the performance gap under the unregularized reward, and only achieves $1/\epsilon^4$ sample complexity. In addition, we remove their $\mathrm{poly}(|\mathcal{A}|)$ dependence and accommodate more general offline distributions. A detailed comparison with [UKLS23] is provided in Appendix E.1.

7 Bounding the Estimation Error

In this section, we present several constructions of the confidence set $\mathcal{F}_{\mathrm{conf}}$ that ensure $Q^\star \in \mathcal{F}_{\mathrm{conf}}$ while controlling the estimation error $\max_{f \in \mathcal{F}_{\mathrm{conf}}} D^{\pi^\star}_{\mathsf{av}}(f \| M^\star)$ in Theorem 2. For a dataset $\mathcal{D}$

of n tuples (s, a, r, s'), we write $\mathbb{E}_{\mathcal{D}}[g] := \frac{1}{n}\sum_{(s,a,r,s')\in\mathcal{D}} g(s, a, r, s')$ for any function g. See Appendix F for proofs in this section.

7.1 Bellman Completeness

Bellman completeness (Assumption 1) is widely adopted in online RL [JLM21, XFB+23] and offline RL with full coverage [CJ19], but has not been studied under partial coverage except for the linear case [GM24]. The following guarantee is, however, standard in all these settings.

Lemma 6. *Assume $\mathcal{F}$ and $\mathcal{G}$ satisfies Assumption 1. Let*

- $L_{\mathsf{bc}}(g, f) = \mathbb{E}_{(s,a,r,s')\sim\mathcal{D}}\big[\,(g(s, a) - r - f(s'))^2\,\big]$
- $\mathcal{F}_{\mathrm{conf}} = \big\{f \in \mathcal{F} :\ L_{\mathsf{bc}}(f, f) - \min_{g\in\mathcal{G}} L_{\mathsf{bc}}(g, f) \le \epsilon_{\mathrm{stat}}\big\}$, *where* $\epsilon_{\mathrm{stat}} = \frac{2H^2\log(|\mathcal{F}||\mathcal{G}|/\delta)}{n}$

Then with probability at least $1-\delta$, $Q^\star \in \mathcal{F}_{\mathrm{conf}}$ and $\max_{f\in\mathcal{F}_{\mathrm{conf}}} D^{\pi^\star}_{\mathsf{av}}(f\|M^\star) \le O(HC^{\pi^\star}\epsilon_{\mathrm{stat}})$.

We note that $Q^\star$-realizability under regularization, partial coverage, and Bellman completeness exactly matches the setting of Conservative Q-Learning (CQL). However, existing analyses of CQL are restricted to the tabular setting [KZTL20]. In Appendix G, we provide the first sample-complexity guarantee for CQL under general function approximation.

7.2 Weight Realizability

Weight realizability (or density-ratio realizability) has been adopted in several prior studies in offline RL with partial coverage [ZHH+22, CJ22, RZY+23, OPZZ23]. Its guarantees are below.

Assumption 3 (Weight realizability)**.** *Define $w^\pi(s, a) = \frac{d^\pi(s,a)}{H\mu(s,a)}$. Assume $w^{\pi^\star} \in \mathcal{W}$ for a given function set $\mathcal{W}$. Additionally, for any $w \in \mathcal{W}$, we have $\|w\|_\infty \le B_{\mathcal{W}}$.*

Lemma 7. *Assume $\mathcal{F}$ and $\mathcal{W}$ satisfies Assumption 3. Let*

- $L_{\mathsf{wr}}(w, f) = \big|\mathbb{E}_{(s,a,r,s')\sim\mathcal{D}}\big[w(s, a)(f(s, a) - r - f(s'))\big]\big|$.
- $\mathcal{F}_{\mathrm{conf}} = \big\{f \in \mathcal{F} :\ \max_{w\in\mathcal{W}} L_{\mathsf{wr}}(w, f) \le \epsilon_{\mathrm{stat}}\big\}$, *where* $\epsilon_{\mathrm{stat}} = B_{\mathcal{W}}H\sqrt{\frac{2\log(|\mathcal{F}||\mathcal{W}|/\delta)}{n}}$.

Then with probability at least $1-\delta$, $Q^\star \in \mathcal{F}_{\mathrm{conf}}$ and $\max_{f\in\mathcal{F}_{\mathrm{conf}}} D^{\pi^\star}_{\mathsf{av}}(f\|M^\star) \le O(H^2\epsilon^2_{\mathrm{stat}})$.

7.3 Low Q-Bellman Rank

Bellman rank is established for online RL [JKA+17, DKL+21, JLM21] but remains unexplored in offline RL except for special cases. Its absence is not for no reason—as we will show, under standard coverage and data format, efficient offline learning in low-Bellman-rank MDPs is impossible. To bypass this barrier, we introduce *double policy sampling* (Assumption 5) and *policy feature coverage* (Definition 4). Then, in Lemma 21 and Lemma 22 (Appendix F.1) we show that dropping either assumption makes polynomial sample complexity impossible.

Assumption 4 (Q-Bellman rank)**.** *There exist mappings $X : \Pi \to \mathbb{R}^d$ and $W : \mathcal{F} \to \mathbb{R}^d$ such that for any $\pi \in \Pi$ and $f \in \mathcal{F}$, $\mathbb{E}_{(s,a)\sim d^\pi}\left[f(s, a) - R(s, a) - \mathbb{E}_{s'\sim P(\cdot|s,a)}[f(s')]\right] = \langle X(\pi), W(f)\rangle$.*

Assumption 5 (Double policy sampling)**.** *Assume that each offline data sample is generated as the following. First, sample a policy $\pi \sim \underline{\mu} \in \Delta(\Pi)$. Then generate* two independent (s, a, r, s') tuples *from this policy: $(s_\circ, a_\circ) \sim d^\pi$, $r_\circ \sim R(s_\circ, a_\circ)$, $s'_\circ \sim P(\cdot|s_\circ, a_\circ)$, and $(s_\times, a_\times) \sim d^\pi$, $r_\times \sim R(s_\times, a_\times)$, $s'_\times \sim P(\cdot|s_\times, a_\times)$. They constitute a data sample: $(s_\circ, a_\circ, r_\circ, s'_\circ, s_\times, a_\times, r_\times, s'_\times) \in \mathcal{D}$.*

Definition 4 (Policy feature coverage)**.** *Assume $\mathcal{F}$ has low Bellman rank (Assumption 4), and let $\underline{\mu}$ be defined in Assumption 5. Define $\underline{C}^{\pi^\star} = X(\pi^\star)^\top\Sigma^\dagger_{\underline{\mu}}X(\pi^\star)$, where $\Sigma_{\underline{\mu}} = \mathbb{E}_{\pi\sim\underline{\mu}}[X(\pi)X(\pi)^\top]$.*

Lemma 8. *Let Assumption 4 and Assumption 5 hold. Also, let*

- $L_{\mathsf{br}}(f) = \mathbb{E}_{(s_\circ,a_\circ,r_\circ,s'_\circ,s_\times,a_\times,r_\times,s'_\times)\sim\mathcal{D}}\big[\,(f(s_\circ, a_\circ) - r_\circ - f(s'_\circ))\,(f(s_\times, a_\times) - r_\times - f(s'_\times))\,\big]$.
- $\mathcal{F}_{\mathrm{conf}} = \big\{f \in \mathcal{F} :\ L_{\mathsf{br}}(f) \le \epsilon_{\mathrm{stat}}\big\}$, *where* $\epsilon_{\mathrm{stat}} = H^2\sqrt{\frac{\log(2|\mathcal{F}|/\delta)}{2n}}$.

Then with probability at least $1-\delta$, $Q^\star \in \mathcal{F}_{\mathrm{conf}}$ and $\max_{f\in\mathcal{F}_{\mathrm{conf}}} D^{\pi^\star}_{\mathsf{av}}(f\|M^\star) \le O(\underline{C}^{\pi^\star}\epsilon_{\mathrm{stat}})$.

References

[AFJ+24] Philip Amortila, Dylan J Foster, Nan Jiang, Ayush Sekhari, and Tengyang Xie. Harnessing density ratios for online reinforcement learning. *arXiv preprint arXiv:2401.09681*, 2024.

[Che25] Fengdi Che. A tutorial: An intuitive explanation of offline reinforcement learning theory. *arXiv preprint arXiv:2508.07746*, 2025.

[CJ19] Jinglin Chen and Nan Jiang. Information-theoretic considerations in batch reinforcement learning. In *International conference on machine learning*, pages 1042–1051. PMLR, 2019.

[CJ22] Jinglin Chen and Nan Jiang. Offline reinforcement learning under value and density-ratio realizability: the power of gaps. In *Uncertainty in Artificial Intelligence*, pages 378–388. PMLR, 2022.

[CXJA22] Ching-An Cheng, Tengyang Xie, Nan Jiang, and Alekh Agarwal. Adversarially trained actor critic for offline reinforcement learning. In *International Conference on Machine Learning*, pages 3852–3878. PMLR, 2022.

[DKL+21] Simon Du, Sham Kakade, Jason Lee, Shachar Lovett, Gaurav Mahajan, Wen Sun, and Ruosong Wang. Bilinear classes: A structural framework for provable generalization in rl. In *International Conference on Machine Learning*, pages 2826–2836. PMLR, 2021.

[FGH23] Dylan J Foster, Noah Golowich, and Yanjun Han. Tight guarantees for interactive decision making with the decision-estimation coefficient. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 3969–4043. PMLR, 2023.

[FGQ+23] Dylan J Foster, Noah Golowich, Jian Qian, Alexander Rakhlin, and Ayush Sekhari. Model-free reinforcement learning with the decision-estimation coefficient. *Advances in Neural Information Processing Systems*, 36, 2023.

[FKQR21] Dylan J Foster, Sham M Kakade, Jian Qian, and Alexander Rakhlin. The statistical complexity of interactive decision making. *arXiv preprint arXiv:2112.13487*, 2021.

[FKSLX22] Dylan J Foster, Akshay Krishnamurthy, David Simchi-Levi, and Yunzong Xu. Offline reinforcement learning: Fundamental barriers for value function approximation. In *Conference on Learning Theory*, pages 3489–3489. PMLR, 2022.

[GM24] Noah Golowich and Ankur Moitra. The role of inherent bellman error in offline reinforcement learning with linear function approximation. In *Reinforcement Learning Conference*, 2024.

[GYDH15] Bram L Gorissen, İhsan Yanıkoğlu, and Dick Den Hertog. A practical guide to robust optimization. *Omega*, 53:124–137, 2015.

[JKA+17] Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. Contextual decision processes with low bellman rank are pac-learnable. In *International Conference on Machine Learning*, pages 1704–1713. PMLR, 2017.

[JLM21] Chi Jin, Qinghua Liu, and Sobhan Miryoosefi. Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms. *Advances in neural information processing systems*, 34:13406–13418, 2021.

[JRSW24] Zeyu Jia, Alexander Rakhlin, Ayush Sekhari, and Chen-Yu Wei. Offline reinforcement learning: Role of state aggregation and trajectory data. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 2644–2719. PMLR, 2024.

[KLS25] Akshay Krishnamurthy, Gene Li, and Ayush Sekhari. The role of environment access in agnostic reinforcement learning. *arXiv preprint arXiv:2504.05405*, 2025.

[KZTL20] Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. *Advances in neural information processing systems*, 33:1179–1191, 2020.

[LS17] Tor Lattimore and Csaba Szepesvari. The end of optimism? an asymptotic analysis of finite-armed linear bandits. In *Artificial Intelligence and Statistics*, pages 728–737. PMLR, 2017.

[LS20] Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

[LWZ25a] Haolin Liu, Chen-Yu Wei, and Julian Zimmert. Decision making in hybrid environments: A model aggregation approach. *Conference on Learning Theory*, 2025.

[LWZ25b] Haolin Liu, Chen-Yu Wei, and Julian Zimmert. An improved model-free decision-estimation coefficient with applications in adversarial mdps. *arXiv preprint arXiv:2510.08882*, 2025.

[MFR24] Zakaria Mhammedi, Dylan J Foster, and Alexander Rakhlin. The power of resets in online reinforcement learning. *Advances in Neural Information Processing Systems*, 37:12334–12407, 2024.

[NTYG+23] Thanh Nguyen-Tang, Ming Yin, Sunil Gupta, Svetha Venkatesh, and Raman Arora. On instance-dependent bounds for offline reinforcement learning with linear function approximation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 9310–9318, 2023.

[OPZZ23] Asuman E Ozdaglar, Sarath Pattathil, Jiawei Zhang, and Kaiqing Zhang. Revisiting the linear-programming framework for offline rl with general function approximation. In *International Conference on Machine Learning*, pages 26769–26791. PMLR, 2023.

[RZY+23] Paria Rashidinejad, Hanlin Zhu, Kunhe Yang, Stuart Russell, and Jiantao Jiao. Optimal conservative offline rl with general function approximation via augmented lagrangian. In *The Eleventh International Conference on Learning Representations*, 2023.

[SZS+22] Yuda Song, Yifei Zhou, Ayush Sekhari, J Andrew Bagnell, Akshay Krishnamurthy, and Wen Sun. Hybrid rl: Using both offline and online data can make rl efficient. *arXiv preprint arXiv:2210.06718*, 2022.

[UKLS23] Masatoshi Uehara, Nathan Kallus, Jason D Lee, and Wen Sun. Offline minimax soft-q-learning under realizability and partial coverage. *Advances in Neural Information Processing Systems*, 36:12797–12809, 2023.

[US21] Masatoshi Uehara and Wen Sun. Pessimistic model-based offline reinforcement learning under partial coverage. *arXiv preprint arXiv:2107.06226*, 2021.

[WCD22] Xinqi Wang, Qiwen Cui, and Simon S Du. On gap-dependent bounds for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 35:14865–14877, 2022.

[WFK20] 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.

[WTN19] Yifan Wu, George Tucker, and Ofir Nachum. Behavior regularized offline reinforcement learning. *arXiv preprint arXiv:1911.11361*, 2019.

[XCJ+21] 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.

[XFB+23] T Xie, D Foster, Y Bai, N Jiang, and S Kakade. The role of coverage in online reinforcement learning. In *Proceedings of the Eleventh International Conference on Learning Representations*, 2023.

[XJ21] Tengyang Xie and Nan Jiang. Batch value-function approximation with only realizability. In *International Conference on Machine Learning*, pages 11404–11413. PMLR, 2021.

[ZHH+22] 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*, pages 2730–2775. PMLR, 2022.

Appendices

A Related Work **15**

B Proofs in Section 4 **17**

B.1 MDP construction 17

B.2 Function set construction 17

B.3 Dataset construction 17

B.4 Lower bound proof 18

C Competing with An Arbitrary Comparator **24**

D Proofs in Section 5 **25**

E Proofs in Section 6 **26**

E.1 Comparison with [UKLS23] 31

F Proofs in Section 7 **32**

F.1 Low Bellman rank setting: lower bounds without policy feature coverage or without double policy samples 35

G Conservative Q-Learning **38**

H Concentration Inequalities and Performance Difference Lemma **41**

I Auxiliary Lemmas for Regularized MDPs and Proof of Lemma 35 **42**

J Limitations and Future Work **50**

Reference	Value realizability	Coverage	Additional structure	Learnability
Only with value realizability and coverage				
[XJ21]	$Q^\star$ realizability	Push-forward coverage	None	Yes
[FKSLX22]	Q^π realizability	All-policy coverage	None	No
Bellman-completeness-based results				
[CJ19]	$Q^\star$ realizability	All-policy coverage	Bellman completeness	Yes
[XCJ+21]	Q^π realizability	Partial coverage	Bellman completeness	Yes
[GM24]	$Q^\star$ realizability	Partial coverage	Bellman completeness; linear function class	Yes
Ours	$Q^\star$ realizability	Partial coverage	Bellman completeness; general function class	No
Weighted-realizability-based results				
[CJ22]†	$Q^\star$ realizability	Partial coverage	Weighted realizability; large value gap	Yes
[UKLS23]‡	$Q^\star$ realizability	Partial coverage	Weighted realizability; regularized value	Yes

† Our method can handle unknown value gaps without online access. ‡ Our method improves the sample-complexity bound.

Table 1: Summary of model-free offline RL learnability with $Q^\star$ or Q^π realizability under different structural assumptions on function class and offline data.

A Related Work

Offline RL Theory. The theoretical study of offline reinforcement learning has largely focused on identifying the structural properties of the function class and the coverage conditions on offline data that enable sample-efficient learning. These two aspects exhibit a trade-off: stronger assumptions on the offline data can often compensate for weaker assumptions on the function class, and vice versa. For model-based offline RL, [US21] provide a nearly complete characterization of learnability, showing that a realizable model class together with partial coverage is sufficient for sample-efficient learning. For model-free offline RL, however, the theoretical picture remains fragmented. Under the weakest assumption of $Q^\star$ realizability, the offline data must satisfy a strong push-forward coverage condition to enable sample-efficient learning [XJ21]. When this data assumption is relaxed to all-policy coverage, it has been shown that even Q^π realizability for every policy π does not guarantee sample-efficient learning [WFK20, FKSLX22, JRSW24]. This motivates the need for additional structural assumptions on the function class. A common such assumption is Bellman completeness. Under Bellman completeness, [CJ19] show that all-policy coverage together with $Q^\star$ realizability is sufficient for learnability, while [XCJ+21] show that single-policy coverage together with Q^π realizability is sufficient. In the special case of linear function classes, [GM24] show that $Q^\star$ realizability, Bellman completeness, and partial coverage together suffice for learning. However, the fundamental role of Bellman completeness beyond the linear setting remains elusive. Our work contributes to this direction by showing that, for general function classes, $Q^\star$ realizability, Bellman completeness, and partial coverage are not sufficient for sample-efficient learning. While the work of [SZS+22] makes a similar claim, their lower bound is algorithm-specific and can be handled by existing algorithms. In contrast, our lower bound is information-theoretic. This provides a sharp separation from previous positive results.

Beyond Bellman completeness, another line of work assumes additional realizability of a Lagrange multiplier or density-ratio-type function in addition to $Q^\star$ realizability [ZHH+22, CJ22, RZY+23, OPZZ23, UKLS23]. Under such assumptions, sample-efficient learning can be obtained when the optimal value has a sufficiently large gap over suboptimal policies [CJ22], or when the value objective includes an additional regularizer [ZHH+22, RZY+23, OPZZ23, UKLS23]. As discussed in Section 2, our analytical framework also recovers and improves several results in this line of work. Results discussed above are summarized in Table 1. For broader discussions of offline RL algorithms and theory, we refer the reader to [Che25] for a comprehensive review and comparison.

Bridge Online and Offline RL. Online and offline RL have traditionally been studied as separate areas, but recent works suggest that their structural assumptions can be viewed through a unified lens [XFB+23, AFJ+24, MFR24, KLS25]. For example, coverage assumptions in offline RL have natural online counterparts, referred to as coverability conditions [XFB+23], which measure whether the MDP can be effectively covered by some exploration distribution. Thus, coverability captures the intrinsic difficulty of exploration in online RL [XFB+23]. Under small coverability and $Q^\star$

realizability, [XFB+23] and [AFJ+24] show that sample-efficient online RL is possible under Bellman completeness and density-ratio realizability, respectively. These results illustrate how ideas from offline RL can inform online RL.

In this work, we pursue the complementary direction: using insights from online RL to better understand offline RL. In online RL, although optimism is a broadly useful principle, it can be arbitrarily suboptimal compared with information-directed methods on certain instances [LS17]. We establish a mirrored phenomenon in offline RL, showing that value-based pessimism, despite its general applicability, can likewise be highly suboptimal in certain problems (Section 5).

We further provide, to our knowledge, the first characterization of offline learnability for low-Bellman-rank MDPs without Bellman completeness [JKA+17, DKL+21, JLM21]. Low Bellman rank is a canonical structural condition in online RL, but its offline counterpart has been largely unexplored beyond special cases. Our analysis identifies *double policy sampling* and *policy feature coverage* as crucial requirements for offline learning in low-Bellman-rank MDPs (Section 7.3), highlighting structural restrictions that are less visible in the online setting.

B Proofs in Section 4

B.1 MDP construction

We construct four families of finite-horizon MDPs as illustrated in Figure 2. In all families, the episode starts at state s_1, where the agent chooses between two actions u, v. Taking either action yields a Bernoulli reward (specified on the corresponding edge in the figure) and transitions to a middle-layer group of states, either W_A or W_B. The group sizes satisfy $|W_\mathsf{A}| = |W_\mathsf{B}| = m$, and conditional on entering a group, the next state is drawn uniformly from that group.

From any state $w \in W_\mathsf{A} \cup W_\mathsf{B}$, there is a single available action a, which yields zero reward and transitions deterministically to s_A if $w \in W_\mathsf{A}$ and to s_B if $w \in W_\mathsf{B}$. Finally, from s_A or s_B, the agent chooses among three actions x, y, z, receives the deterministic terminal reward shown in Figure 2, and the episode ends. One terminal reward equals -2, which lies outside the reward range specified in Section 3; this can be removed by a simple affine rescaling and does not affect the argument.

Each family consists of a collection of MDPs indexed by hidden assignments. Let W be a set of $2m$ abstract middle-layer states. Each MDP instance corresponds to an assignment $\phi : W \to \{W_\mathsf{A}, W_\mathsf{B}\}$ where exactly m states are assigned to W_A and the remaining m states are assigned to W_B. There are $\binom{2m}{m}$ such assignments. For each family $\mathcal{M}_{u,x}$, $\mathcal{M}_{u,y}$, $\mathcal{M}_{v,x}$, and $\mathcal{M}_{v,y}$, the transition and reward structure is fixed as in Figure 2, while the assignment ϕ varies over all such possibilities. Consequently, $|\mathcal{M}_{u,x}| = |\mathcal{M}_{u,y}| = |\mathcal{M}_{v,x}| = |\mathcal{M}_{v,y}| = \binom{2m}{m}$. We write $M^\phi_{u,x}$ for the MDP in $\mathcal{M}_{u,x}$ corresponding to assignment ϕ, and similarly for the other families.

Define $\mathcal{M}_u = \mathcal{M}_{u,x} \cup \mathcal{M}_{u,y}$, $\mathcal{M}_v = \mathcal{M}_{v,x} \cup \mathcal{M}_{v,y}$, and $\mathcal{M} = \mathcal{M}_u \cup \mathcal{M}_v$. The ground-truth MDP $M^\star$ is drawn uniformly from $\mathcal{M}$.

B.2 Function set construction

We define the function class $\mathcal{F} = \{(f_1, f_2, f_3), (f_1, f_2, g_3), (g_1, f_2, f_3), (g_1, f_2, g_3)\}$ where the components are defined as

$$\begin{cases} f_3(s_\mathsf{A}, x) = 1 \\ f_3(s_\mathsf{A}, y) = -2 \\ f_3(s_\mathsf{A}, z) = 0 \\ f_3(s_\mathsf{B}, x) = 0 \\ f_3(s_\mathsf{B}, y) = -2 \\ f_3(s_\mathsf{B}, z) = 1 \end{cases} \qquad \begin{cases} g_3(s_\mathsf{A}, x) = -2 \\ g_3(s_\mathsf{A}, y) = 1 \\ g_3(s_\mathsf{A}, z) = 0 \\ g_3(s_\mathsf{B}, x) = -2 \\ g_3(s_\mathsf{B}, y) = 0 \\ g_3(s_\mathsf{B}, z) = 1 \end{cases} \qquad f_2(w, a) = 1 \ \ \forall w \in W_\mathsf{A} \cup W_\mathsf{B}$$

$$\begin{cases} f_1(s_1, u) = \frac{3}{2} \\ f_1(s_1, v) = \frac{3}{2} + \Delta \end{cases} \qquad \begin{cases} g_1(s_1, u) = \frac{3}{2} + \Delta \\ g_1(s_1, v) = \frac{3}{2} \end{cases}$$

Checking Bellman completeness We verify that $\mathcal{F}$ is Bellman complete. Let $M \in \mathcal{M}$ be arbitrary. We have for any $Q_3 \in \{f_3, g_3\}$

$$\forall w \in W_\mathsf{A} \cup W_\mathsf{B}, \quad (\mathcal{T}_M Q_3)(w, a) = R_M(w, a) + \mathbb{E}_{s' \sim P_M(\cdot|w,a)} \left[\max_{a'} Q_3(s', a') \right] = 0 + 1 = 1.$$

Thus, $(\mathcal{T}_M Q_3) = f_2$. Next,

$$\forall a_1 \in \{u, v\}, \quad (\mathcal{T}_M f_2)(s_1, a_1) = R_M(s_1, a_1) + \mathbb{E}_{s' \sim P_M(\cdot|s_1,a_1)} \left[\max_{a'} f_2(s', a') \right] = R_M(s_1, a_1) + 1.$$

Depending on $M \in \mathcal{M}_u$ or $M \in \mathcal{M}_v$, $R_M + 1$ equals either f_1 or g_1. Therefore, $(\mathcal{T}_M f_2) \in \{f_1, g_1\}$. This verifies that $\mathcal{F}$ is Bellman complete.

B.3 Dataset construction

We assume the offline dataset $\mathcal{D} = (\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3)$, where:

- $\mathcal{D}_1$ consists of n tuples $(s_1, a_1^{(i)}, r_1^{(i)}, w_1^{(i)})_{i=1}^n$ from layer 1, where $a_1^{(i)} \sim \mathsf{Unif}\{u, v\}$; then $r_1^{(i)} \in \{0, 1\}$ and $w_1^{(i)} \in W$ are drawn according to $M^\star$ given $(s_1, a_1^{(i)})$.

- $\mathcal{D}_2$ consists of n tuples $(w_2^{(i)}, a, 0, s_2^{(i)})_{i=1}^n$ from layer 2, where $w_2^{(i)} \sim \mathsf{Unif}(W)$, and $s_2^{(i)}$ is drawn according to $M^\star$ given $(w_2^{(i)}, a)$.
- $\mathcal{D}_3$ consists of n tuples $(s_3^{(i)}, z, r_3^{(i)})_{i=1}^n$ from layer 3, where $s_3^{(i)} \sim \mathsf{Unif}\{s_\mathsf{A}, s_\mathsf{B}\}$. After taking action z, $r_3^{(i)} \in \{0, 1\}$ is drawn according to $M^\star$ given $(s_3^{(i)}, z)$.

Checking coverage over the optimal policy π_M For any model $M \in \mathcal{M}$ in Figure 2, the unique optimal policy π_M always follows the lowest branch and achieves value $\frac{3}{2} + \Delta$. The above data distribution has coverage $C^{\pi_M} = 2$ with respect to π_M.

B.4 Lower bound proof

For each $M \in \mathcal{M}$, let $\mathbb{P}_M$ denote the law of the dataset $\mathcal{D}$ and $\mathbb{E}_M$ the expectation under $\mathbb{P}_M$. Let π_M denote the optimal policy in M. Define the mixture laws of the dataset:

$$\begin{aligned}
\mathbb{P}_{u,x} &= \frac{1}{|\mathcal{M}_{u,x}|} \sum_{M \in \mathcal{M}_{u,x}} \mathbb{P}_M, \quad \mathbb{P}_{u,y} = \frac{1}{|\mathcal{M}_{u,y}|} \sum_{M \in \mathcal{M}_{u,y}} \mathbb{P}_M, \\
\mathbb{P}_{v,x} &= \frac{1}{|\mathcal{M}_{v,x}|} \sum_{M \in \mathcal{M}_{v,x}} \mathbb{P}_M, \quad \mathbb{P}_{v,y} = \frac{1}{|\mathcal{M}_{v,y}|} \sum_{M \in \mathcal{M}_{v,y}} \mathbb{P}_M, \\
\mathbb{P}_u &= \frac{1}{|\mathcal{M}_u|} \sum_{M \in \mathcal{M}_u} \mathbb{P}_M, \qquad \mathbb{P}_v = \frac{1}{|\mathcal{M}_v|} \sum_{M \in \mathcal{M}_v} \mathbb{P}_M, \\
\mathbb{P} &= \frac{1}{|\mathcal{M}|} \sum_{M \in \mathcal{M}} \mathbb{P}_M,
\end{aligned} \tag{10}$$

and let $\mathbb{E}_{u,x}, \mathbb{E}_{u,y}, \mathbb{E}_{v,x}, \mathbb{E}_{v,y}, \mathbb{E}_u, \mathbb{E}_v, \mathbb{E}$ denote the corresponding expectation. Furthermore, let $J_M(\pi)$ denote the value of π under model M. By the construction in Figure 2,

$$J_M(\hat{\pi}) = \tag{11}$$
$$\begin{cases}
\hat{\pi}(u|s_1)\big(\frac{1}{2} + \hat{\pi}(x|s_\mathsf{A}) - 2\hat{\pi}(y|s_\mathsf{A})\big) + \hat{\pi}(v|s_1)\big(\frac{1}{2} + \Delta + \hat{\pi}(z|s_\mathsf{B}) - 2\hat{\pi}(y|s_\mathsf{B})\big) \triangleq J_{u,x}(\hat{\pi}) & \text{if } M \in \mathcal{M}_{u,x} \\
\hat{\pi}(u|s_1)\big(\frac{1}{2} + \hat{\pi}(y|s_\mathsf{A}) - 2\hat{\pi}(x|s_\mathsf{A})\big) + \hat{\pi}(v|s_1)\big(\frac{1}{2} + \Delta + \hat{\pi}(z|s_\mathsf{B}) - 2\hat{\pi}(x|s_\mathsf{B})\big) \triangleq J_{u,y}(\hat{\pi}) & \text{if } M \in \mathcal{M}_{u,y} \\
\hat{\pi}(v|s_1)\big(\frac{1}{2} + \hat{\pi}(x|s_\mathsf{A}) - 2\hat{\pi}(y|s_\mathsf{A})\big) + \hat{\pi}(u|s_1)\big(\frac{1}{2} + \Delta + \hat{\pi}(z|s_\mathsf{B}) - 2\hat{\pi}(y|s_\mathsf{B})\big) \triangleq J_{v,x}(\hat{\pi}) & \text{if } M \in \mathcal{M}_{v,x} \\
\hat{\pi}(v|s_1)\big(\frac{1}{2} + \hat{\pi}(y|s_\mathsf{A}) - 2\hat{\pi}(x|s_\mathsf{A})\big) + \hat{\pi}(u|s_1)\big(\frac{1}{2} + \Delta + \hat{\pi}(z|s_\mathsf{B}) - 2\hat{\pi}(x|s_\mathsf{B})\big) \triangleq J_{v,y}(\hat{\pi}) & \text{if } M \in \mathcal{M}_{v,y}
\end{cases}$$

Lemma 9 (Action z is optimal)**.** *Let an algorithm output $\hat{\pi}$ (possibly stochastic). Define $\hat{\pi}^z$ to match $\hat{\pi}$ everywhere except $\hat{\pi}^z(s_\mathsf{A}) = \hat{\pi}^z(s_\mathsf{B}) = z$ deterministically. Then $\mathbb{E}[J_M(\hat{\pi})] \le \mathbb{E}[J_M(\hat{\pi}^z)]$.*

Proof. For any assignment ϕ, the models $M_{u,x}^\phi$ and $M_{u,y}^\phi$ induce identical data distributions, since they differ only in terminal rewards for actions x and y, which never appear in the dataset. The same holds for $M_{v,x}^\phi$ and $M_{v,y}^\phi$. Consequently,

$$\mathbb{P}_{u,x} = \mathbb{P}_{u,y} \quad \text{and} \quad \mathbb{P}_{v,x} = \mathbb{P}_{v,y}. \tag{12}$$

The expected value of policy $\hat{\pi}$ is

$$\begin{aligned}
&\mathbb{E}[J_M(\hat{\pi})] \\
&= \frac{1}{4}\mathbb{E}_{M \sim \mathsf{Unif}(\mathcal{M}_{u,x})}\mathbb{E}_M[J_M(\hat{\pi})] + \frac{1}{4}\mathbb{E}_{M \sim \mathsf{Unif}(\mathcal{M}_{u,y})}\mathbb{E}_M[J_M(\hat{\pi})] \\
&\quad + \frac{1}{4}\mathbb{E}_{M \sim \mathsf{Unif}(\mathcal{M}_{v,x})}\mathbb{E}_M[J_M(\hat{\pi})] + \frac{1}{4}\mathbb{E}_{M \sim \mathsf{Unif}(\mathcal{M}_{v,y})}\mathbb{E}_M[J_M(\hat{\pi})] \\
&\qquad\qquad (M \text{ is chosen uniformly from } \mathcal{M}) \\
&= \frac{1}{4}\mathbb{E}_{M \sim \mathsf{Unif}(\mathcal{M}_{u,x})}\mathbb{E}_M[J_{u,x}(\hat{\pi})] + \frac{1}{4}\mathbb{E}_{M \sim \mathsf{Unif}(\mathcal{M}_{u,y})}\mathbb{E}_M[J_{u,y}(\hat{\pi})] \\
&\quad + \frac{1}{4}\mathbb{E}_{M \sim \mathsf{Unif}(\mathcal{M}_{v,x})}\mathbb{E}_M[J_{v,x}(\hat{\pi})] + \frac{1}{4}\mathbb{E}_{M \sim \mathsf{Unif}(\mathcal{M}_{v,y})}\mathbb{E}_M[J_{v,y}(\hat{\pi})] \qquad \text{(by (11))} \\
&= \frac{1}{4}\mathbb{E}_{u,x}[J_{u,x}(\hat{\pi})] + \frac{1}{4}\mathbb{E}_{u,y}[J_{u,y}(\hat{\pi})] + \frac{1}{4}\mathbb{E}_{v,x}[J_{v,x}(\hat{\pi})] + \frac{1}{4}\mathbb{E}_{v,y}[J_{v,y}(\hat{\pi})]
\end{aligned}$$

$$= \frac{1}{4}\mathbb{E}_{u,x}[J_{u,x}(\hat{\pi}) + J_{u,y}(\hat{\pi})] + \frac{1}{4}\mathbb{E}_{v,x}[J_{v,x}(\hat{\pi}) + J_{v,y}(\hat{\pi})] \qquad \text{(by (12))}$$

$$= \frac{1}{4}\mathbb{E}_{u,x}\left[\hat{\pi}(u|s_1)\big(1 - \hat{\pi}(x|s_\mathsf{A}) - \hat{\pi}(y|s_\mathsf{A})\big) + \hat{\pi}(v|s_1)\big(1 + 2\Delta + 2\hat{\pi}(z|s_\mathsf{B}) - 2\hat{\pi}(x|s_\mathsf{B}) - 2\hat{\pi}(y|s_\mathsf{B})\big)\right]$$
$$+ \frac{1}{4}\mathbb{E}_{v,x}\left[\hat{\pi}(v|s_1)\big(1 - \hat{\pi}(x|s_\mathsf{A}) - \hat{\pi}(y|s_\mathsf{A})\big) + \hat{\pi}(u|s_1)\big(1 + 2\Delta + 2\hat{\pi}(z|s_\mathsf{B}) - 2\hat{\pi}(x|s_\mathsf{B}) - 2\hat{\pi}(y|s_\mathsf{B})\big)\right]$$
(by (11))

$$= \frac{1}{4}\mathbb{E}_{u,x}\left[\hat{\pi}(u|s_1)\hat{\pi}(z|s_\mathsf{A}) + \hat{\pi}(v|s_1)\big(-1 + 2\Delta + 4\hat{\pi}(z|s_\mathsf{B})\big)\right]$$
$$+ \frac{1}{4}\mathbb{E}_{v,x}\left[\hat{\pi}(v|s_1)\hat{\pi}(z|s_\mathsf{A}) + \hat{\pi}(u|s_1)\big(-1 + 2\Delta + 4\hat{\pi}(z|s_\mathsf{B})\big)\right].$$
($\sum_{a\in\{x,y,z\}} \hat{\pi}(a|s) = 1$ for $s = s_\mathsf{A}$ and s_B)

As $\hat{\pi}^z(\cdot|s_1) = \hat{\pi}(\cdot|s_1)$ and $\hat{\pi}^z(z|s_\mathsf{A}) \geq \hat{\pi}(z|s_\mathsf{A})$ and $\hat{\pi}^z(z|s_\mathsf{B}) \geq \hat{\pi}(z|s_\mathsf{B})$, by the last expression, we have

$$\mathbb{E}[J_M(\hat{\pi})] \leq \mathbb{E}[J_M(\hat{\pi}^z)].$$

□

Lemma 10 (Reduction to total variation)**.** *For any offline algorithm,*

$$\mathbb{E}[J_M(\pi_M) - J_M(\hat{\pi})] \;\geq\; \frac{1}{2}(1+\Delta)\left(1 - D_{\mathrm{TV}}(\mathbb{P}_u, \mathbb{P}_v)\right).$$

Proof. By Lemma 9 we may assume $\hat{\pi}(s_\mathsf{A}) = \hat{\pi}(s_\mathsf{B}) = z$. Let $\beta = \hat{\pi}(u \mid s_1) \in [0,1]$. A direct calculation from (11) yields

$$J_M(\hat{\pi}) = \begin{cases} \frac{1}{2}\beta + (\frac{3}{2}+\Delta)(1-\beta) & \text{if } M \in \mathcal{M}_u, \\ \frac{1}{2}(1-\beta) + (\frac{3}{2}+\Delta)\beta & \text{if } M \in \mathcal{M}_v. \end{cases}$$

Also, $J_M(\pi_M) = \frac{3}{2} + \Delta$ for any M. Thus,

$$\begin{aligned}
&\mathbb{E}[J_M(\pi_M) - J_M(\hat{\pi})] \\
&= \left(\frac{3}{2}+\Delta\right) - \frac{1}{2}\mathbb{E}_u\left[\frac{1}{2}\beta + \left(\frac{3}{2}+\Delta\right)(1-\beta)\right] - \frac{1}{2}\mathbb{E}_v\left[\frac{1}{2}(1-\beta) + \left(\frac{3}{2}+\Delta\right)\beta\right] \\
&= \left(\frac{3}{2}+\Delta\right) - \frac{1}{2}\left(\frac{3}{2}+\Delta\right) - \frac{1}{2}\cdot\frac{1}{2} + \frac{1}{2}\left(\left(\frac{3}{2}+\Delta\right) - \frac{1}{2}\right)(\mathbb{E}_u[\beta] - \mathbb{E}_v[\beta]) \\
&= \frac{1}{2}(1+\Delta)\left(1 + (\mathbb{E}_u[\beta] - \mathbb{E}_v[\beta])\right) \\
&\geq \frac{1}{2}(1+\Delta)\left(1 - D_{\mathrm{TV}}(\mathbb{P}_u, \mathbb{P}_v)\right),
\end{aligned}$$

where we use that $\beta \in [0,1]$, $\left|\mathbb{E}_u[\beta] - \mathbb{E}_v[\beta]\right| \leq D_{\mathrm{TV}}(\mathbb{P}_u, \mathbb{P}_v)$ in the last inequality. □

Lemma 11 (TV bound)**.** *If $m \geq 4n^2$, then $D_{\mathrm{TV}}(\mathbb{P}_u, \mathbb{P}_v) \;\leq\; \frac{6n^2}{m} + 4\sqrt{n}\Delta$.*

Proof.

Step 1: remove the Δ-shift via a comparison family. Define an auxiliary collection of families $\mathcal{M}'_{u,x}, \mathcal{M}'_{u,y}, \mathcal{M}'_{v,x}, \mathcal{M}'_{v,y}$ that are identical to those in Figure 2, except that the rewards on actions u and v at s_1 are both $\mathsf{Ber}(\frac{1}{2})$ (i.e., we set $\Delta = 0$ at the first layer). Define $\mathcal{M}'_u, \mathcal{M}'_v, \mathcal{M}'$ and the corresponding mixture laws $\mathbb{P}'_u, \mathbb{P}'_v, \mathbb{P}'$ analogously to (10).

By the triangle inequality and Pinsker's inequality,

$$\begin{aligned}
D_{\mathrm{TV}}(\mathbb{P}_u, \mathbb{P}_v) &\leq D_{\mathrm{TV}}(\mathbb{P}'_u, \mathbb{P}'_v) + D_{\mathrm{TV}}(\mathbb{P}_u, \mathbb{P}'_u) + D_{\mathrm{TV}}(\mathbb{P}_v, \mathbb{P}'_v) \\
&\leq D_{\mathrm{TV}}(\mathbb{P}'_u, \mathbb{P}'_v) + \sqrt{\frac{1}{2}\mathrm{KL}(\mathbb{P}_u, \mathbb{P}'_u)} + \sqrt{\frac{1}{2}\mathrm{KL}(\mathbb{P}_v, \mathbb{P}'_v)}
\end{aligned}$$

The only difference between $\mathbb{P}_u$ and $\mathbb{P}'_u$ (and similarly between $\mathbb{P}_v$ and $\mathbb{P}'_v$) is the n first-layer rewards in $\mathcal{D}_1$: under $\mathbb{P}_u$ these rewards are $\mathsf{Ber}(\frac{1}{2})$ or $\mathsf{Ber}(\frac{1}{2}+\Delta)$ depending on the action, whereas under $\mathbb{P}'_u$ they are always $\mathsf{Ber}(\frac{1}{2})$. Hence,

$$\mathrm{KL}(\mathbb{P}_u, \mathbb{P}'_u) \le n\mathrm{KL}\left(\mathsf{Ber}\left(\tfrac{1}{2}+\Delta\right), \mathsf{Ber}\left(\tfrac{1}{2}\right)\right), \quad \mathrm{KL}(\mathbb{P}_v, \mathbb{P}'_v) \le n\mathrm{KL}\left(\mathsf{Ber}\left(\tfrac{1}{2}+\Delta\right), \mathsf{Ber}\left(\tfrac{1}{2}\right)\right).$$

Using the bound $\mathrm{KL}(\mathsf{Ber}(\frac{1}{2}+\Delta), \mathsf{Ber}(\frac{1}{2})) \le 8\Delta^2$ for $\Delta \le \frac{1}{4}$, we obtain

$$\begin{aligned} D_{\mathrm{TV}}(\mathbb{P}_u, \mathbb{P}_v) &\le D_{\mathrm{TV}}(\mathbb{P}'_u, \mathbb{P}'_v) + 2\sqrt{\frac{n}{2} \times 8\Delta^2} \\ &= D_{\mathrm{TV}}(\mathbb{P}'_u, \mathbb{P}'_v) + 4\sqrt{n}\Delta. \end{aligned}$$

Step 2: control repeats in the visited middle-layer states. By definition,

$$D_{\mathrm{TV}}(\mathbb{P}'_u, \mathbb{P}'_v) \;=\; \frac{1}{2}\sum_{\mathcal{D}} \left|\mathbb{P}'_u(\mathcal{D}) - \mathbb{P}'_v(\mathcal{D})\right|$$

where the summation runs over all possible datasets. For any dataset $\mathcal{D} = (\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3)$ of the form in Appendix B.3, define

$$\mathsf{Rep}(\mathcal{D}) = \mathbf{1}\{\exists i \ne j,\;\; w_1^{(i)} = w_1^{(j)} \text{ or } w_2^{(i)} = w_2^{(j)} \text{ or } w_1^{(i)} = w_2^{(j)}\}.$$

That is, $\mathsf{Rep}(\mathcal{D}) = 1$ if any state in W appears more than once in the dataset $\mathcal{D}$. Note that under any $M \in \mathcal{M}'$, the middle-layer states $\{w_1^{(i)}\}_{i=1}^n$ are i.i.d. and uniformly distributed over W. Indeed, $a_1^{(i)} \sim \mathsf{Unif}\{u, v\}$, and conditional on $a_1^{(i)}$ the next state is uniform over the corresponding size-m group; hence, marginally, $\Pr(w_1^{(i)} = w) = \frac{1}{2m}$ for all $w \in W$. Moreover, $\{w_2^{(i)}\}_{i=1}^n$ are i.i.d. $\mathsf{Unif}(W)$ by construction. Therefore, the $2n$ middle-layer states $\{w_1^{(i)}\}_{i=1}^n \cup \{w_2^{(i)}\}_{i=1}^n$ are i.i.d. samples from $\mathsf{Unif}(W)$.

Thus, the probability that any state in W is sampled more than once is

$$\begin{aligned} \mathbb{P}'_u(\mathsf{Rep}(\mathcal{D}) = 1) &= 1 - \mathbb{P}'_u(\mathsf{Rep}(\mathcal{D}) = 0) \\ &= 1 - 1 \times \frac{2m-1}{2m} \times \frac{2m-2}{2m} \times \cdots \times \frac{2m-(2n-1)}{2m} \\ &\le 1 - \left(1 - \frac{n}{m}\right)^{2n} \le \frac{2n^2}{m}. \end{aligned}$$

The same bound holds for $\mathbb{P}'_v(\mathsf{Rep}(\mathcal{D}) = 1)$. Decomposing the TV sum according to $\mathsf{Rep}(\mathcal{D})$ yields

$$\begin{aligned} \sum_{\mathcal{D}} |\mathbb{P}'_u(\mathcal{D}) - \mathbb{P}'_v(\mathcal{D})| &= \sum_{\mathcal{D}:\ \mathsf{Rep}(\mathcal{D})=0} |\mathbb{P}'_u(\mathcal{D}) - \mathbb{P}'_v(\mathcal{D})| + \sum_{\mathcal{D}:\ \mathsf{Rep}(\mathcal{D})=1} |\mathbb{P}'_u(\mathcal{D}) - \mathbb{P}'_v(\mathcal{D})| \\ &\le \sum_{\mathcal{D}:\ \mathsf{Rep}(\mathcal{D})=0} |\mathbb{P}'_u(\mathcal{D}) - \mathbb{P}'_v(\mathcal{D})| + \mathbb{P}'_u(\mathsf{Rep}(\mathcal{D}) = 1) + \mathbb{P}'_v(\mathsf{Rep}(\mathcal{D}) = 1) \\ &\le \sum_{\mathcal{D}:\ \mathsf{Rep}(\mathcal{D})=0} |\mathbb{P}'_u(\mathcal{D}) - \mathbb{P}'_v(\mathcal{D})| + \frac{4n^2}{m}. \end{aligned} \tag{13}$$

Step 3: evaluate $\mathbb{P}'_u(\mathcal{D})$ on datasets without repeats. Fix a dataset $\mathcal{D}$ with $\mathsf{Rep}(\mathcal{D}) = 0$. For a fixed model $M \in \mathcal{M}'_u$, the likelihood factors over layers:

- **Samples in $\mathcal{D}_1$:** Each sample in $\mathcal{D}_1$ starts with s_1, choosing $a_1^{(i)} \sim \mathsf{Unif}\{u, v\}$. Under $M \in \mathcal{M}'_u$, $r_1^{(i)} \sim \mathsf{Ber}(\frac{1}{2})$, and $w_1^{(i)}$ is uniform on the appropriate size-m set (W_{A} or W_{B}) given $a_1^{(i)}$. For example, if $a_1^{(i)} = u$ then $w_1^{(i)} \sim \mathsf{Unif}(W_{\mathsf{A}})$ according to Figure 2(a)(b). Hence, for a realized tuple $(s_1, a_1^{(i)}, r_1^{(i)}, w_1^{(i)})$,

  $$\mathbb{P}_M\big((s_1, a_1^{(i)}, r_1^{(i)}, w_1^{(i)})\big) = \mathbb{P}'((a_1^{(i)}, r_1^{(i)})) \cdot \frac{1}{m}\mathbf{1}\{M\text{'s transition is consistent with } (s_1, a_1^{(i)}, w_1^{(i)})\}$$

  where $\mathbb{P}'$ is mixture law $\mathbb{P}' = \mathbb{E}_{M\sim\mathsf{Unif}(\mathcal{M}')}\mathbb{P}_M$. It can be used here because all $M \in \mathcal{M}'$ share the same law of $(a_1^{(i)}, r_1^{(i)})$, and so does their mixture. The indicator enforces that the assignment of M places $w_1^{(i)}$ in the correct group (W_{A} or W_{B}) given $a_1^{(i)}$, as the example given above.

- **Samples in $\mathcal{D}_2$.** Each sample in $\mathcal{D}_2$ first draws $w_2^{(i)} \sim \mathsf{Unif}(W)$ and chooses action a. Then the next-state $s_2^{(i)}$ is deterministic given $w_2^{(i)}$: if $w_2^{(i)} \in W_\mathsf{A}$ under M, then $s_2^{(i)} = s_\mathsf{A}$; otherwise, $s_2^{(i)} = s_\mathsf{B}$. Thus for a realized tuple $(w_2^{(i)}, a, 0, s_2^{(i)})$,

$$\mathbb{P}_M\big((w_2^{(i)}, a, 0, s_2^{(i)})\big) = \frac{1}{2m}\mathbf{1}\{M\text{'s transition is consistent with } (w_2^{(i)}, a, s_2^{(i)})\}.$$

- **Samples in $\mathcal{D}_3$.** Each sample in $\mathcal{D}_3$ first draws $s_3^{(i)} \sim \mathsf{Unif}\{s_\mathsf{A}, s_\mathsf{B}\}$, chooses z, and receives reward $r_3^{(i)}$. By construction, $r_3^{(i)} = 0$ if $s_3^{(i)} = s_\mathsf{A}$, and $r_3^{(i)} = 1$ if $s_3^{(i)} = s_\mathsf{B}$. This procedure is the same for all models M. Thus,

$$\mathbb{P}_M\big((s_3^{(i)}, z, r_3^{(i)}\big) = \mathbb{P}'((s_3^{(i)}, r_3^{(i)}))$$

Multiplying across all samples in all $\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3$, we get

$$\mathbb{P}_M(\mathcal{D}) = \mathbb{P}'\big((a_1^{(i)}, r_1^{(i)})_{i=1}^n\big)\mathbb{P}'\big((s_3^{(i)}, r_3^{(i)})_{i=1}^n\big) \cdot \frac{1}{m^n} \cdot \frac{1}{(2m)^n} \cdot \mathbf{1}\{M\text{'s transition is consistent with } (s_1, a_1^{(i)}, w_1^{(i)})_{i=1}^n \text{ and } (w_2^{(i)}, a, s_2^{(i)})_{i=1}^n\}$$

Averaging over $M \sim \mathsf{Unif}(\mathcal{M}'_u)$ yields

$$\mathbb{P}'_u(\mathcal{D}) = \mathbb{P}'\big((a_1^{(i)}, r_1^{(i)})_{i=1}^n\big)\mathbb{P}'\big((s_3^{(i)}, r_3^{(i)})_{i=1}^n\big) \cdot \frac{1}{m^n} \cdot \frac{1}{(2m)^n} \cdot \underbrace{\mathbb{E}_{M\sim\mathsf{Unif}(\mathcal{M}'_u)}\big[\mathbf{1}\{M\text{'s transition is consistent with } (s_1, a_1^{(i)}, w_1^{(i)})_{i=1}^n \text{ and } (w_2^{(i)}, a, s_2^{(i)})_{i=1}^n\}\big]}_{(\star)} \tag{14}$$

Step 4: count consistent assignments. Next, we calculate $(\star)$ for a given $\mathcal{D}$ with $\mathsf{Rep}(\mathcal{D}) = 0$. Each transition sample $(s_1, a_1^{(i)}, w_1^{(i)})$ in $\mathcal{D}_1$ specifies a constraint for M. For example, assume $(s_1, a_1^{(i)}, w_1^{(i)}) = (s_1, u, w)$ for some $w \in W$. Then to make $M \in \mathcal{M}'_u$ consistent with (s_1, u, w), M must assign w to W_A. Similarly, each transition tuple $(w_2^{(i)}, a, s_2^{(i)})$ sets a constraint for M: Assume $(w_2^{(i)}, a, s_2^{(i)}) = (w, a, s_\mathsf{B})$ for some $w \in W$. Then to make $M \in \mathcal{M}'_u$ consistent with this tuple, M must assign w to W_B. As there are no repeating states from W in the dataset, there are $2n$ non-overlapping constraints on how M should assign each $w \in W$ to $\{W_\mathsf{A}, W_\mathsf{B}\}$.

The total number of possible assignments from W to $\{W_\mathsf{A}, W_\mathsf{B}\}$ is $\binom{2m}{m}$. With $2n$ non-overlapping constraints, there are $N = 2n$ elements in W whose assignments are *not free*. The total number of assignments under the constraints becomes $\binom{2m-N}{m-l}$, where l is the number of constraints restricting w to be in W_A, and $N - l$ is the number of constraints restricting w to W_B.

Then, $(\star)$ is the fraction of the assignments satisfying the constraints specified by $(s_1, a_1^{(i)}, w_1^{(i)})_{i=1}^n$ and $(w_2^{(i)}, a, s_2^{(i)})_{i=1}^n$, which is given by $\binom{2m-N}{m-l} \Big/ \binom{2m}{m}$ and can be expanded as

$$\frac{\frac{(2m-N)!}{(m-l)!\,(m-N+l)!}}{\frac{(2m)!}{m!\,m!}} = \frac{\prod_{i=0}^{l-1}(m-i) \times \prod_{i=0}^{N-l-1}(m-i)}{\prod_{i=0}^{N-1}(2m-i)} = \frac{m^N}{(2m)^N} \cdot \frac{\prod_{i=0}^{l-1}\left(1-\frac{i}{m}\right) \times \prod_{i=0}^{N-l-1}\left(1-\frac{i}{m}\right)}{\prod_{i=0}^{N-1}\left(1-\frac{i}{2m}\right)}.$$

From this expression, we have

$$\begin{aligned}
(\star) &\geq 2^{-N}\left(1-\frac{l}{m}\right)^l\left(1-\frac{N-l}{m}\right)^{N-l} \\
&\qquad \text{(lower bound the numerator and upper bound the denominator)} \\
&\geq 2^{-N}\left(1-\frac{l^2}{m}\right)\left(1-\frac{(N-l)^2}{m}\right) \qquad ((1-\zeta)^k \geq 1-k\zeta \text{ for } \zeta \in (0,1) \text{ and } k \geq 1) \\
&\geq 2^{-N}\left(1-\frac{N^2}{m}\right) \qquad (l^2 + (N-l)^2 \leq N^2)
\end{aligned}$$

and similarly

$$(\star) \le 2^{-N} \cdot \frac{1}{\left(1-\frac{N}{2m}\right)^N} \le 2^{-N} \cdot \frac{1}{1-\frac{N^2}{2m}} \le 2^{-N}\left(1+\frac{N^2}{m}\right). \quad (\tfrac{1}{1-\zeta} \le 1+2\zeta \text{ for } \zeta \in (0,\tfrac{1}{2}))$$

Therefore, by (14) and that $N = 2n$,

$$\mathbb{P}'_u(\mathcal{D}) \in \mathbb{P}'\big((a_1^{(i)}, r_1^{(i)})_{i=1}^n\big)\mathbb{P}'\big((s_3^{(i)}, r_3^{(i)})_{i=1}^n\big) \cdot \frac{1}{m^n} \cdot \frac{1}{(2m)^n} \cdot 2^{-2n}\left[1-\frac{4n^2}{m}, 1+\frac{4n^2}{m}\right].$$

The same holds for $\mathbb{P}'_v(\mathcal{D})$. Plugging them into the first term on the right-hand side of (13), we get

$$\begin{aligned}
&\sum_{\mathcal{D}:\ \mathsf{Rep}(\mathcal{D})=0} |\mathbb{P}'_u(\mathcal{D}) - \mathbb{P}'_v(\mathcal{D})| \\
&\le \sum_{(a_1^{(i)}, r_1^{(i)}, w_1^{(i)}, w_2^{(i)}, s_2^{(i)}, s_3^{(i)}, r_3^{(i)})_{i=1}^n} \mathbb{P}'((a_1^{(i)}, r_1^{(i)}))\mathbb{P}'((s_3^{(i)}, r_3^{(i)})) \cdot \frac{1}{m^n} \cdot \frac{1}{(2m)^n} \cdot 2^{-2n} \cdot \frac{8n^2}{m} \\
&= \sum_{(w_1^{(i)}, w_2^{(i)}, s_2^{(i)})_{i=1}^n} \frac{1}{m^n} \cdot \frac{1}{(2m)^n} \cdot 2^{-2n} \cdot \frac{8n^2}{m} \\
&= (2m \times 2m \times 2)^n \times \frac{1}{(m \times 2m \times 4)^n} \cdot \frac{8n^2}{m} \\
&\qquad\qquad \text{(each } w_1^{(i)} \text{ and } w_2^{(i)} \text{ has } 2m \text{ possible values, and } s_2^{(i)} \text{ has 2 possible values)} \\
&= \frac{8n^2}{m}.
\end{aligned}$$

Combining everything above, we get

$$D_{\mathrm{TV}}(\mathbb{P}_u, \mathbb{P}_v) \le \frac{1}{2}\left(\frac{8n^2}{m} + \frac{4n^2}{m}\right) + 4\sqrt{n}\Delta = \frac{6n^2}{m} + 4\sqrt{n}\Delta.$$

□

Lemma 12 (An ϵ-independent lower bound)**.** *For any $\epsilon \le \frac{1}{4}$, to achieve $\mathbb{E}[J(\pi^\star) - J(\hat{\pi})] \le \epsilon$ in the instance described in Appendix B.1–Appendix B.3, any offline algorithm must access at least $n \ge \Omega\big(\min\big\{\sqrt{m}, \frac{1}{\Delta^2}\big\}\big)$ samples.*

Proof. By Lemma 10 and Lemma 11, we have

$$\mathbb{E}[J(\pi^\star) - J(\hat{\pi})] \ge \frac{1}{2}\left(1 - D_{\mathrm{TV}}(\mathbb{P}_u, \mathbb{P}_v)\right) \ge \frac{1}{2}\left(1 - \frac{6n^2}{m} - 4\sqrt{n}\Delta\right).$$

To make the left-hand side smaller than $\epsilon \le \frac{1}{4}$, we need $\frac{6n^2}{m} + 4\sqrt{n}\Delta \ge \frac{1}{2}$, implying that either $\frac{6n^2}{m} \ge \frac{1}{4}$ or $4\sqrt{n}\Delta \ge \frac{1}{4}$. Hence, $n \ge \Omega\big(\min\{\sqrt{m}, \frac{1}{\Delta^2}\}\big)$. □

Proof of Theorem 1.

ϵ-dependent lower bound for (s, a, r, s') data. We start with proving the first part of the theorem, where the learner can only access (s, a, r, s') data. We leverage the lower bound in Lemma 12 (with the instances illustrated in Figure 2). Notice Lemma 12 already proves this theorem for the special case $\epsilon = \frac{1}{4}$.

To lift that lower bound construction to an ϵ-dependent bound for an arbitrary $\epsilon \in (0, \frac{1}{4}]$, we extend it in the way illustrated in Figure 4. That is, we add an initial state s_0, on which the learner can only take one action a_0 that has an instantaneous reward of zero. After taking a_0, with probability $p = 4\epsilon$, the state transitions to s_1 (the initial state in the MDP instance of Lemma 12), and with probability $1 - p = 1 - 4\epsilon$, the state transitions to a chain of states (z_1, z_2, z_3) on which there is only one action a_0 to choose and the reward is always zero. In the previous construction, μ is induced by a behavior policy π_b. We will keep this behavior policy, as there is no decision to make on all other states we add. It is straightforward to check that $C^{\pi^\star}$ remains the same after this extension. For the function set $\mathcal{F}$

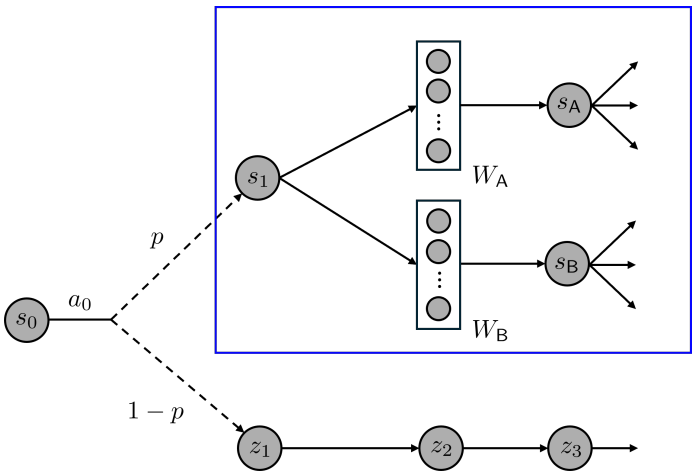


Figure 4: Construction for ϵ-dependent lower bound with non-trajectory data. The blue box is the hard instance in Lemma 12 (Figure 2).

defined in the previous construction (Appendix B.2), we extend each $f \in \mathcal{F}$ with $f(s_0, a_0) = pf(s_1)$ and $f(z_1, a_0) = f(z_2, a_0) = f(z_3, a_0) = 0$. This keeps the function set to satisfy $Q^\star$-realizability and Bellman completeness.

With all the conditions satisfied, we check how many samples the learner needs to ensure $J(\pi^\star) - J(\hat{\pi}) \le \epsilon$. Observe that

$$J(\pi^\star) - J(\hat{\pi}) = p(V^\star(s_1) - V^{\hat{\pi}}(s_1)) = 4\epsilon(V^\star(s_1) - V^{\hat{\pi}}(s_1)).$$

In order to achieve $J(\pi^\star) - J(\hat{\pi}) \le \epsilon$, the learner must ensure $V^\star(s_1) - V^{\hat{\pi}}(s_1) \le \frac{1}{4}$. As already established in Lemma 12, this requires at least $\Omega\big(\min\{\sqrt{|\mathcal{S}|}, \frac{1}{\Delta^2}\}\big)$ samples in the original MDP starting from s_1. However, in the extended MDP, $1-p$ portion of the offline data are trivial, and with only probability p can the learner get any samples from the original MDP. Therefore, the total number of samples the learner needs is of order $\Omega\big(\frac{1}{p}\min\{\sqrt{|\mathcal{S}|}, \frac{1}{\Delta^2}\}\big) = \Omega\big(\frac{1}{\epsilon}\min\{\sqrt{|\mathcal{S}|}, \frac{1}{\Delta^2}\}\big)$. This proves the first claim of the theorem.

ϵ-dependent lower bound for trajectory data. To prove a lower bound under trajectory feedback, we further extend the construction with a reduction established in [JRSW24]. We note that the reason why trajectory breaks the lower bound in the instances in Figure 2 or Figure 4 is because with trajectory data, the learner can easily tell whether it is u or v that leads to s_{B} by just checking which state the trajectory leads to after two steps. This is contrary to the case of (s, a, r, s') data where the state is essentially "re-sampled" from μ after the state transitions to $W_{\mathsf{A}} \cup W_{\mathsf{B}}$.

The idea of [JRSW24] (in their Section 4.2) is to mimic such re-sampling under trajectory feedback. In our case, we extend the construction in Figure 4 by repeating the middle layers for $\Theta(H)$ times, resulting in Figure 5. Denote $W_h = W_{h,\mathsf{A}} \cup W_{h,\mathsf{B}}$ for $h = 3, 4, \dots, H-1$. There are two actions on states in $W_3, W_4, \dots, W_{H-2}$. If taking action 1 (indicated by the blue arrows in Figure 5), the next state is drawn uniformly from *the same group* on the next layer; if taking action 2 (indicated by the red arrows in Figure 5), the next state is drawn uniformly from the *union of the two groups*. We let action 1 have an instantaneous reward of 0, and action 2 have an instantaneous reward of -2. This implies that the optimal action in $W_3, W_4, \dots, W_{H-2}$ is always action 1. Also, we extend any function f in the function set (defined in Appendix B.2) by letting $f(w, 1) = 1$ and $f(w, 2) = -2 + 1 = -1$ for $w \in W_3 \cup \cdots \cup W_{H-2}$. For the behavior policy, we let $\pi_b(1|w) = \pi_b(2|w) = \frac{1}{2}$ for $w \in W_3 \cup \cdots \cup W_{H-2}$.

This construction makes $C^{\pi^\star} \le 4$. This is because the state distribution coverage $\frac{d^{\pi^\star}(s)}{H\mu(s)} \le 2$ remains, and thus $\frac{d^{\pi^\star}(s,a)}{H\mu(s,a)} = \frac{d^{\pi^\star}(s)\pi^\star(a|s)}{H\mu(s)\pi_b(a|s)} \le 2 \times 2 = 4$ due to $\pi_b(1|w) = \frac{1}{2}$ for $w \in W_3 \cup \cdots W_{H-2}$. Furthermore, it keeps realizability and Bellman completeness.

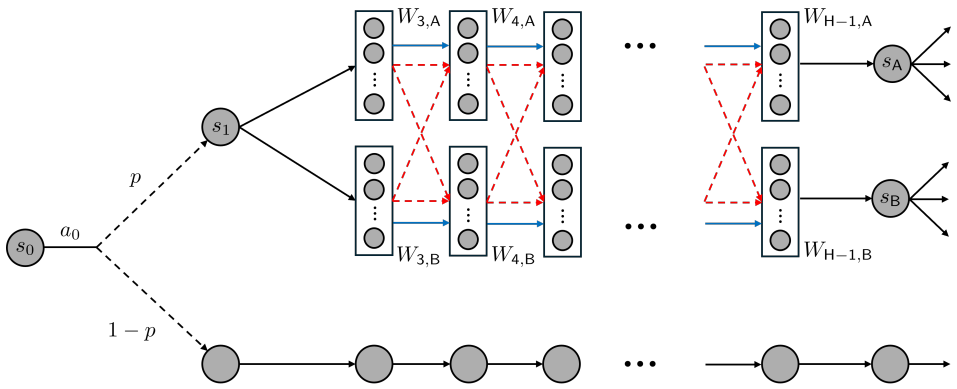


Figure 5: Construction for ϵ-dependent lower bound with trajectory data (extended from Figure 4). The blue arrows indicate the transition of action 1, which always leads to a uniform distribution *over the same group* on the next layer. The red arrows indicate the transition of action 2, which always leads to a uniform distribution *over all states* on the next layer.

Because π_b chooses both actions with equal probability, for each trajectory *that passes through* s_1, with probability $1 - 2^{-(H-4)}$, action 2 is chosen at least once during layers $h = 3, \dots, H-2$. Furthermore, if action 2 is chosen in any of these layers, the state distribution on layer $H-1$ will keep no information of the action chosen on s_1. This exactly mimics the "re-sampling" scenario when the learner only has $(s, a, r, s') \sim \mu$ data.

If the number of trajectories passing through s_1 is smaller than 2^{H-4}, then with a constant probability, action 2 is chosen at least once in all these trajectories. In this case, the previous hardness argument for (s, a, r, s') applies, and the learner requires at least $\Omega\big(\frac{1}{\epsilon}\min\{\sqrt{m}, \frac{1}{\Delta^2}\}\big) = \Omega\big(\frac{1}{\epsilon}\min\{\sqrt{|\mathcal{S}|/H}, \frac{1}{\Delta^2}\}\big)$ samples. This concludes that in this new construction, the learner requires at least $\Omega\big(\frac{1}{\epsilon\,\mathrm{poly}(H)}\min\{2^H, \sqrt{|\mathcal{S}|}, \frac{1}{\Delta^2}\}\big)$ trajectories to learn an ϵ-optimal policy. □

C Competing with An Arbitrary Comparator

A stronger and more desirable guarantee in offline RL under partial coverage is to compete with an arbitrary comparator policy $\mathring{\pi}$, with the performance gap adapting to the coverage $C^{\mathring{\pi}}$. Such guarantees have been established under model realizability [US21] and Q^π-realizability [XCJ+21]. However, under $Q^\star$-realizability, this type of guarantee is fundamentally harder to obtain. Below, we present a pathological example showing that even in a seemingly simple setting, it is impossible to compete with an arbitrary $\mathring{\pi}$.

Recall that the hardness of competing with $\pi^\star$ in Theorem 1 relies on an extreme small or zero value gap $V^\star(s) - \max_{a \neq \pi^\star(s)} Q^\star(s, a)$. Below, we show that when the goal is to compete with an arbitrary comparator policy $\mathring{\pi}$, the hardness holds even when this value gap is $\Theta(1)$.

To show this, we use the same MDP structure in Figure 2 with $r(s_1, u) \sim \mathsf{Ber}(\frac{1}{2} + 0.1)$, $r(s_1, v) \sim \mathsf{Ber}(\frac{1}{2})$ and other rewards remain the same. For every MDP, let the comparator policy be the one that always goes to the lowest branch and take action z on s_B. Specifically, for $\mathcal{M}_{u,x}$ and $\mathcal{M}_{u,y}$, we consider $\mathring{\pi}(s_1) = v$ and $\mathring{\pi}(s_B) = z$, which is not the optimal policy now because the optimal policies for $\mathcal{M}_{u,x}$ and $\mathcal{M}_{u,y}$ take u on s_1 and take x and y on s_A, respectively. For $\mathcal{M}_{v,x}$ and $\mathcal{M}_{v,y}$, we consider $\mathring{\pi}(s_1) = u$ and $\mathring{\pi}(s_B) = z$, which remains to be the optimal policy. We assume the offline data remain the same, so $C^{\mathring{\pi}}$ is small for any MDP although the optimal policy may not be covered. In this case, when we try to learn a policy $\hat{\pi}$ such that $\mathbb{E}[J(\mathring{\pi}) - J(\hat{\pi})] \leq \epsilon$, the constant gap on u, v make it possible to distinguish them with constant samples. However, even if the learner can distinguish u and v, they still do not know which action leads to the lowest branches unless $\mathrm{poly}(m)$ samples are observed. Without such information, a random guess will lead to constant suboptimality gap.

Algorithm 2 Offline Robust Estimation-to-Decision with Arbitrary Comparator

Input: Confidence set $\mathcal{F}_{\mathrm{conf}}$, policy class Π divergence measure $D^{\pi}(f\|M)$, parameter γ for the offset version.
Define $\mathcal{M}_{\mathrm{conf}} = \{M : Q^{\star}_{M} \in \mathcal{F}_{\mathrm{conf}}\}$ and compute

$$\hat{\rho} = \begin{cases} \operatorname*{argmin}_{\rho\in\Delta(\Pi)} \max_{(M,\mathring{\pi})\in\mathcal{M}_{\mathrm{conf}}\times\Pi} \mathbb{E}_{\pi\sim\rho}\left[J_M(\mathring{\pi}) - J_M(\pi) - \gamma \max_{f\in\mathcal{F}_{\mathrm{conf}}} D^{\mathring{\pi}}(f\|M)\right] & \text{(offset version)} \\ \operatorname*{argmin}_{\rho\in\Delta(\Pi)} \max_{(M,\mathring{\pi})\in\mathcal{M}_{\mathrm{conf}}\times\Pi} \dfrac{J_M(\mathring{\pi}) - \mathbb{E}_{\pi\sim\rho}[J_M(\pi)]}{\left(\max_{f\in\mathcal{F}_{\mathrm{conf}}} D^{\mathring{\pi}}(f\|M)\right)^{1/2}}. & \text{(ratio version)} \end{cases}$$

Output: mixture policy $\hat{\pi} \sim \hat{\rho}$.

However, we may still extend our algorithm and the associated decision complexity to accommodate an arbitrary comparator policy $\mathring{\pi}$. As shown in Algorithm 2, given a policy class Π that contains our target comparator policy, we solve a minimax problem against the joint worst case over both the environment model and the comparator policy. Given the above pathological example, the associated decision complexity under $Q^{\star}$-realizability may not be bounded. However, under Q^{π}-realizability and policy Bellman completeness assumptions, with $\mathcal{F}_{\mathrm{conf}}$ chosen to match the confidence set in [XCJ+21], Algorithm 2 recovers their guarantees.

D Proofs in Section 5

Lemma 13. $\mathsf{Ordec\text{-}O}^{D}_{\gamma}(\mathcal{F}_{\mathrm{conf}}) \le \frac{4}{\gamma}\left(\mathsf{Ordec\text{-}R}^{D}(\mathcal{F}_{\mathrm{conf}})\right)^2$.

Proof. For any $M \in \mathcal{M}_{\mathrm{conf}}$, we have

$$\begin{aligned} &\mathbb{E}_{\pi\sim\rho}\left[J_M(\pi_M) - J_M(\pi)\right] \\ &= \frac{J_M(\pi_M) - \mathbb{E}_{\pi\sim\rho}\left[J_M(\pi)\right]}{\sqrt{\max_{f\in\mathcal{F}_{\mathrm{conf}}} D^{\pi_M}(f\|M)}} \sqrt{\max_{f\in\mathcal{F}_{\mathrm{conf}}} D^{\pi_M}(f\|M)} \\ &\le \frac{4}{\gamma}\frac{\left(J_M(\pi_M) - \mathbb{E}_{\pi\sim\rho}\left[J_M(\pi)\right]\right)^2}{\max_{f\in\mathcal{F}_{\mathrm{conf}}} D^{\pi_M}(f\|M)} + \gamma \max_{f\in\mathcal{F}_{\mathrm{conf}}} D^{\pi_M}(f\|M) \qquad \text{(AM-GM)} \end{aligned}$$

Since $\frac{J_M(\pi_M) - \mathbb{E}_{\pi\sim\rho}[J_M(\pi)]}{\left(\max_{f\in\mathcal{F}_{\mathrm{conf}}} D^{\pi_M}(f\|M)\right)^{1/2}} > 0$, we have $\mathsf{Ordec\text{-}O}^{D}_{\gamma}(\mathcal{F}_{\mathrm{conf}}) \le \frac{4}{\gamma}\left(\mathsf{Ordec\text{-}R}^{D}(\mathcal{F}_{\mathrm{conf}})\right)^2$. □

Proof of Theorem 2. For the offset version, we have

$$\begin{aligned} &J(\pi^{\star}) - \mathbb{E}_{\pi\sim\hat{\rho}}[J(\pi)] \\ &= J(\pi^{\star}) - \mathbb{E}_{\pi\sim\hat{\rho}}[J(\pi)] - \gamma \max_{f\in\mathcal{F}_{\mathrm{conf}}} D^{\pi^{\star}}(f\|M^{\star}) + \gamma \max_{f\in\mathcal{F}_{\mathrm{conf}}} D^{\pi^{\star}}(f\|M^{\star}) \\ &\le \max_{M\in\mathcal{M}_{\mathrm{conf}}} \left\{J_M(\pi_M) - \mathbb{E}_{\pi\sim\hat{\rho}}[J_M(\pi)] - \gamma \max_{f\in\mathcal{F}_{\mathrm{conf}}} D^{\pi_M}(f\|M)\right\} + \gamma \max_{f\in\mathcal{F}_{\mathrm{conf}}} D^{\pi^{\star}}(f\|M^{\star}) \\ &\qquad (M^{\star} \in M_{\mathrm{conf}} \text{ from } Q^{\star} \in \mathcal{F}_{\mathrm{conf}}) \\ &= \min_{\rho\in\Delta(\Pi)} \max_{M\in\mathcal{M}_{\mathrm{conf}}} \left\{J_M(\pi_M) - \mathbb{E}_{\pi\sim\rho}[J_M(\pi)] - \gamma \max_{f\in\mathcal{F}_{\mathrm{conf}}} D^{\pi_M}(f\|M)\right\} + \gamma \max_{f\in\mathcal{F}_{\mathrm{conf}}} D^{\pi^{\star}}(f\|M^{\star}) \\ &\qquad \text{(by the choice of } \hat{\rho}) \\ &= \mathsf{Ordec\text{-}O}^{D}_{\gamma}(\mathcal{F}_{\mathrm{conf}}) + \gamma \max_{f\in\mathcal{F}_{\mathrm{conf}}} D^{\pi^{\star}}(f\|M^{\star}). \end{aligned}$$

For the ratio version, we have

$$
\begin{aligned}
&J(\pi^\star) - \mathbb{E}_{\pi\sim\hat{\rho}}[J(\pi)] \\
&= \frac{J(\pi^\star) - \mathbb{E}_{\pi\sim\hat{\rho}}[J(\pi)]}{(\max_{f\in\mathcal{F}_{\mathrm{conf}}} D^{\pi^\star}(f\|M^\star))^{\frac{1}{2}}} \left(\max_{f\in\mathcal{F}_{\mathrm{conf}}} D^{\pi^\star}(f\|M^\star)\right)^{\frac{1}{2}} \\
&\leq \max_{M\in\mathcal{M}_{\mathrm{conf}}} \frac{J_M(\pi_M) - \mathbb{E}_{\pi\sim\hat{\rho}}[J(\pi)]}{(\max_{f\in\mathcal{F}_{\mathrm{conf}}} D^{\pi_M}(f\|M))^{\frac{1}{2}}} \left(\max_{f\in\mathcal{F}_{\mathrm{conf}}} D^{\pi^\star}(f\|M^\star)\right)^{\frac{1}{2}} \\
&\qquad\qquad (M^\star \in M_{\mathrm{conf}} \text{ from } Q^\star \in \mathcal{F}_{\mathrm{conf}}) \\
&= \min_{\rho\in\Delta(\Pi)} \max_{M\in\mathcal{M}_{\mathrm{conf}}} \frac{J_M(\pi_M) - \mathbb{E}_{\pi\sim\rho}[J(\pi)]}{(\max_{f\in\mathcal{F}_{\mathrm{conf}}} D^{\pi_M}(f\|M))^{\frac{1}{2}}} \left(\max_{f\in\mathcal{F}_{\mathrm{conf}}} D^{\pi^\star}(f\|M^\star)\right)^{\frac{1}{2}} \\
&= \mathsf{Ordec\text{-}R}^{D}(\mathcal{F}_{\mathrm{conf}}) \left(\max_{f\in\mathcal{F}_{\mathrm{conf}}} D^{\pi^\star}(f\|M^\star)\right)^{\frac{1}{2}}.
\end{aligned}
$$

□

The guarantee of value-centered pessimism (5) is given by

Theorem 14. *If $Q^\star \in \mathcal{F}_{\mathrm{conf}}$, then* GDE *ensures*

$$
J(\pi^\star) - J(\hat{\pi}) \leq \mathsf{Gdec}^{D}(\mathcal{F}_{\mathrm{conf}})\big(D^{\pi^\star}(\hat{f}\|M^\star)\big)^{1/2}.
$$

Proof of Theorem 14. We have

$$
\begin{aligned}
J(\pi^\star) - J(\pi_{\hat{f}}) &= \frac{J(\pi^\star) - J(\pi_{\hat{f}})}{\left(D^{\pi^\star}(\hat{f}\|M^\star)\right)^{\frac{1}{2}}} \left(D^{\pi^\star}(\hat{f}\|M^\star)\right)^{\frac{1}{2}} \\
&\leq \max_{M\in\mathcal{M}_{\mathrm{conf}}} \frac{J_M(\pi_M) - J_M(\pi_{\hat{f}})}{\big(D^{\pi_M}(\hat{f}\|M)\big)^{1/2}} \big(D^{\pi^\star}(\hat{f}\|M^\star)\big)^{1/2} \\
&\qquad\qquad (M^\star \in M_{\mathrm{conf}} \text{ from } Q^\star \in \mathcal{F}_{\mathrm{conf}}) \\
&= \mathsf{Gdec}^{D}(\mathcal{F}_{\mathrm{conf}})\big(D^{\pi^\star}(\hat{f}\|M^\star)\big)^{1/2}.
\end{aligned}
$$

□

E Proofs in Section 6

To connect Ordec with complexities established in prior work, we define the *Exploitability Ratio*:

Definition 5 (Exploitability Ratio)**.** *For $f \in \mathcal{F}_{\mathrm{conf}}$, define*

$$
\mathsf{ER}(f;\mathcal{F}_{\mathrm{conf}}) = \max_{M\in\mathcal{M}_{\mathrm{conf}}} \frac{J_M(\pi_M) - J_M(\pi_f)}{\mathbb{E}_{(s,a)\sim d_M^{\pi_M}}[f(s) - f(s,a) + \psi(\pi_M;s)]}
$$

with $\frac{0}{0} \triangleq 0$. If π_f or π_M is not unique, choose the one that makes this ratio the largest.

The ratio $\mathsf{ER}(f;\mathcal{F}_{\mathrm{conf}})$ quantifies how much the greedy policy π_f may be exploited by another model $M \in \mathcal{M}_{\mathrm{conf}}$ (i.e., $J_M(\pi_M) - J_M(\pi_f)$ in the numerator), relative to how much f thinks π_f is better than π_M (i.e., $f(s) - f(s,\pi_M(s)) + \psi(\pi_M;s)$ in the denominator). The next theorem shows performance bounds achieved by E2D.OR and GDE in terms of ER:

Theorem 15. *Let $\hat{f} = \operatorname{argmin}_{f\in\mathcal{F}_{\mathrm{conf}}} f(s_1)$ and assume $Q^\star \in \mathcal{F}_{\mathrm{conf}}$. Then the following hold:*

$$
\mathsf{Ordec\text{-}O}^{\mathsf{av}}_{\gamma}(\mathcal{F}_{\mathrm{conf}}) \leq \tfrac{1}{2\gamma}\mathsf{ER}(Q^\star;\,\mathcal{F}_{\mathrm{conf}})^2 + \gamma \max_{f\in\mathcal{F}_{\mathrm{conf}}} D^{\pi^\star}_{\mathsf{av}}(f\|M^\star), \tag{15}
$$

$$
\mathsf{Ordec\text{-}R}^{\mathsf{av}}(\mathcal{F}_{\mathrm{conf}}) \leq \mathsf{ER}(\hat{f};\mathcal{F}_{\mathrm{conf}}), \tag{16}
$$

$$
\mathsf{Gdec}^{\mathsf{av}}(\mathcal{F}_{\mathrm{conf}}) \leq \mathsf{ER}(\hat{f};\mathcal{F}_{\mathrm{conf}}). \tag{17}
$$

Before proving Theorem 15, we first provide auxiliary Lemma 16 and Lemma 17.

Lemma 16. *Let M, M' be two models and $f = Q^\star_M$, $f' = Q^\star_{M'}$. Then*

$$D^{\pi_M}_{\mathsf{av}}(f' \| M) + D^{\pi_{M'}}_{\mathsf{av}}(f \| M') \geq \frac{1}{2}\left(\mathbb{E}^{\pi_M, M}[f'(s) - f'(s,a) + \psi(\pi_M; s)] + \mathbb{E}^{\pi_{M'}, M'}[f(s) - f(s,a) + \psi(\pi_{M'}; s)]\right)^2$$

where $\mathbb{E}^{\pi, M}[g(s,a)]$ denotes $\mathbb{E}_{(s,a)\sim d^\pi_M}[g(s,a)]$.

Proof.

$$\begin{aligned} & f(s_1) - f'(s_1) && (18)\\ &= J_M(\pi_M) - f'(s_1) \\ &= \mathbb{E}^{\pi_M, M}\left[R_M(s,a) - \psi(\pi_M; s)\right] + \mathbb{E}^{\pi_M, M}\left[-f'(s) + \mathbb{E}_{s'\sim P_M(\cdot|s,a)}[f'(s')]\right] \\ &= \mathbb{E}^{\pi_M, M}\left[-f'(s,a) + R_M(s,a) + \mathbb{E}_{s'\sim P_M(\cdot|s,a)}[f'(s')]\right] - \mathbb{E}^{\pi_M, M}\left[f'(s) - f'(s,a) + \psi(\pi_M; s)\right]. && (19)\end{aligned}$$

Similarly,

$$\begin{aligned} & f'(s_1) - f(s_1) \\ &= \mathbb{E}^{\pi_{M'}, M'}\left[-f(s,a) + R_{M'}(s,a) + \mathbb{E}_{s'\sim P_{M'}(\cdot|s,a)}[f(s')]\right] - \mathbb{E}^{\pi_{M'}, M'}\left[f(s) - f(s,a) + \psi(\pi_{M'}; s)\right]. \end{aligned}$$

Summing up the two equalities, we get

$$\begin{aligned} & \mathbb{E}^{\pi_M, M}\left[f'(s) - f'(s,a) + \psi(\pi_M; s)\right] + \mathbb{E}^{\pi_{M'}, M'}\left[f(s) - f(s,a) + \psi(\pi_{M'}; s)\right] \\ &= \mathbb{E}^{\pi_M, M}\left[-f'(s,a) + R_M(s,a) + \mathbb{E}_{s'\sim P_M(\cdot|s,a)}[f'(s')]\right] \\ &\quad + \mathbb{E}^{\pi_{M'}, M'}\left[-f(s,a) + R_{M'}(s,a) + \mathbb{E}_{s'\sim P_{M'}(\cdot|s,a)}[f(s')]\right]. \end{aligned}$$

Squaring both sides and using $(x+y)^2 \leq 2(x^2+y^2)$:

$$\begin{aligned} & \left(\mathbb{E}^{\pi_M, M}\left[f'(s) - f'(s,a) + \psi(\pi_M; s)\right] + \mathbb{E}^{\pi_{M'}, M'}\left[f(s) - f(s,a) + \psi(\pi_{M'}; s)\right]\right)^2 \\ &\leq 2\left(\mathbb{E}^{\pi_M, M}\left[-f'(s,a) + R_M(s,a) + \mathbb{E}_{s'\sim P_M(\cdot|s,a)}[f'(s')]\right]\right)^2 \\ &\quad + 2\left(\mathbb{E}^{\pi_{M'}, M'}\left[-f(s,a) + R_{M'}(s,a) + \mathbb{E}_{s'\sim P_{M'}(\cdot|s,a)}[f(s')]\right]\right)^2 \\ &= 2D^{\pi_M}_{\mathsf{av}}(f' \| M) + 2D^{\pi_{M'}}_{\mathsf{av}}(f \| M'). \end{aligned}$$

□

Lemma 17. *Let $\hat{f} = \operatorname{argmin}_{f\in\mathcal{F}_{\mathrm{conf}}} f(s_1)$. Then for any $M \in \mathcal{M}_{\mathrm{conf}}$,*

$$D^{\pi_M}_{\mathsf{av}}(\hat{f} \| M) \geq \left(\mathbb{E}^{\pi_M, M}[\hat{f}(s) - \hat{f}(s,a) + \psi(\pi_M; s)]\right)^2$$

where $\mathbb{E}^{\pi, M}[g(s,a)]$ denotes $\mathbb{E}_{(s,a)\sim d^\pi_M}[g(s,a)]$.

Proof. Let $f = Q^\star_M \in \mathcal{F}_{\mathrm{conf}}$. By (19),

$$\begin{aligned} & \mathbb{E}^{\pi_M, M}\left[\hat{f}(s) - \hat{f}(s,a) + \psi(\pi_M; s)\right] && (20)\\ &= \mathbb{E}^{\pi_M, M}\left[-\hat{f}(s,a) + R_M(s,a) + \mathbb{E}_{s'\sim P_M(\cdot|s,a)}[\hat{f}(s')]\right] + \hat{f}(s_1) - f(s_1) && (21)\\ &\leq \mathbb{E}^{\pi_M, M}\left[-\hat{f}(s,a) + R_M(s,a) + \mathbb{E}_{s'\sim P_M(\cdot|s,a)}[\hat{f}(s')]\right]. && \text{(by the choice of } \hat{f}) \end{aligned}$$

Since the left-hand side is non-negative, after squaring both sides the inequality is still true. This proves the lemma. □

Proof of Theorem 15.
Proving (15) By definition,

$$\textsf{Ordec-O}^{\textsf{av}}_{\gamma}(\mathcal{F}_{\text{conf}}) = \min_{\rho\in\Delta(\Pi)} \max_{M\in\mathcal{M}_{\text{conf}}} \left\{ J_M(\pi_M) - \mathbb{E}_{\pi\sim\rho}\left[J_M(\pi)\right] - \gamma \max_{f\in\mathcal{F}_{\text{conf}}} D^{\pi_M}_{\textsf{av}}(f\|M) \right\}.$$

For the last term, we have

$$\begin{aligned}
&\max_{f\in\mathcal{F}_{\text{conf}}} D^{\pi_M}_{\textsf{av}}(f\|M) \\
&\geq D^{\pi_M}_{\textsf{av}}(Q^\star\|M) && \text{(recall } Q^\star \triangleq Q^\star_{M^\star}\text{)} \\
&\geq \frac{1}{2}\Big(\mathbb{E}^{\pi_M,M}[\underbrace{V^\star(s) - Q^\star(s,a) + \psi(\pi_M;s)}_{\geq 0}] + \mathbb{E}^{\pi^\star,M^\star}[\underbrace{V^\star_M(s) - Q^\star_M(s,a) + \psi(\pi^\star;s)}_{\geq 0}]\Big)^2 - D^{\pi^\star}_{\textsf{av}}(Q^\star_M\|M^\star) \\
& && \text{(Lemma 16)} \\
&\geq \frac{1}{2}\left(\mathbb{E}^{\pi_M,M}[V^\star(s) - Q^\star(s,a) + \psi(\pi_M;s)]\right)^2 - D^{\pi^\star}_{\textsf{av}}(Q^\star_M\|M^\star) \\
&\geq \frac{1}{2}\left(\mathbb{E}^{\pi_M,M}[V^\star(s) - Q^\star(s,a) + \psi(\pi_M;s)]\right)^2 - \max_{f\in\mathcal{F}_{\text{conf}}} D^{\pi^\star}_{\textsf{av}}(f\|M^\star).
\end{aligned}$$

Thus,

$$\begin{aligned}
&\textsf{Ordec-O}^{\textsf{av}}_{\gamma}(\mathcal{F}_{\text{conf}}) \\
&= \min_{\rho\in\Delta(\Pi)} \max_{M\in\mathcal{M}_{\text{conf}}} \Bigg\{ J_M(\pi_M) - \mathbb{E}_{\pi\sim\rho}\left[J_M(\pi)\right] \\
&\qquad - \frac{1}{2}\gamma\left(\mathbb{E}^{\pi_M,M}[V^\star(s) - Q^\star(s,a) + \psi(\pi_M;s)]\right)^2 + \gamma \max_{f\in\mathcal{F}_{\text{conf}}} D^{\pi^\star}_{\textsf{av}}(f\|M^\star) \Bigg\} \\
&\leq \frac{1}{2\gamma} \min_{\rho\in\Delta(\Pi)} \max_{M\in\mathcal{M}_{\text{conf}}} \left(\frac{J_M(\pi_M) - \mathbb{E}_{\pi\sim\rho}[J_M(\pi)]}{\mathbb{E}^{\pi_M,M}[V^\star(s) - Q^\star(s,a) + \psi(\pi_M;s)]}\right)^2 + \gamma \max_{f\in\mathcal{F}_{\text{conf}}} D^{\pi^\star}_{\textsf{av}}(f\|M^\star)
\end{aligned}$$

where in the last inequality we use the AM-GM inequality. Finally, using that $\min_x \max_y (F(x,y)^2) = \min_x (\max_y F(x,y))^2 = (\min_x \max_y F(x,y))^2$ for non-negative F proves (15).

Proving (16) By definition,

$$\textsf{Ordec-R}^{\textsf{av}}(\mathcal{F}_{\text{conf}}) := \min_{\rho\in\Delta(\Pi)} \max_{M\in\mathcal{M}_{\text{conf}}} \frac{J_M(\pi_M) - \mathbb{E}_{\pi\sim\rho}[J_M(\pi)]}{\sqrt{\max_{f\in\mathcal{F}_{\text{conf}}} D^{\pi_M}_{\textsf{av}}(f\|M)}}.$$

For the denominator, we have

$$\sqrt{\max_{f\in\mathcal{F}_{\text{conf}}} D^{\pi_M}_{\textsf{av}}(f\|M)} \geq \sqrt{D^{\pi_M}_{\textsf{av}}(\hat{f}\|M)} \geq \mathbb{E}^{\pi_M,M}[\hat{f}(s) - \hat{f}(s,a) + \psi(\pi_M;s)] \tag{22}$$

by Lemma 17. Plugging this into the above definition shows (16).

Proving (17) By definition,

$$\textsf{Gdec}^{\textsf{av}}(\mathcal{F}_{\text{conf}}) := \max_{M\in\mathcal{M}_{\text{conf}}} \frac{J_M(\pi_M) - J_M(\pi_{\hat{f}})}{\sqrt{D^{\pi_M}_{\textsf{av}}(\hat{f}\|M)}}.$$

Using (22) again in the denominator shows (17). □

Proof of Theorem 3. By definition, when $\psi \equiv 0$,

$$\begin{aligned}
\mathsf{ER}(f; \mathcal{F}_{\mathrm{conf}}) &= \max_{M \in \mathcal{M}_{\mathrm{conf}}} \frac{J_M(\pi_M) - J_M(\pi_f)}{\mathbb{E}_{(s,a)\sim d_M^{\pi_M}}[f(s) - f(s,a)]} \\
&\le \max_{M \in \mathcal{M}_{\mathrm{conf}}} \frac{\mathbb{E}_{(s,a)\sim d_M^{\pi_M}}\left[Q_M^{\pi_f}(s,\pi_M) - Q_M^{\pi_f}(s,\pi_f)\right]}{\mathbb{E}_{(s,a)\sim d_M^{\pi_M}}[\Delta_f \|\pi_M(\cdot|s) - \pi_f(\cdot|s)\|_1]} \\
&\qquad \text{(by the performance difference and the gap definition)} \\
&\le \max_{M \in \mathcal{M}_{\mathrm{conf}}} \frac{\mathbb{E}_{(s,a)\sim d_M^{\pi_M}}[H\|\pi_M(\cdot|s) - \pi_f(\cdot|s)\|_1]}{\mathbb{E}_{(s,a)\sim d_M^{\pi_M}}[\Delta_f \|\pi_M(\cdot|s) - \pi_f(\cdot|s)\|_1]} \\
&= \frac{H}{\Delta_f}.
\end{aligned}$$

Combining this bound with Theorem 15 finishes the proof. □

The first two assumptions in Theorem 4 are mild. The third (9) requires some *curvature* at the boundary of the feature set, making the optimal action unique. Notably, it only requires curvature at the greedy action of $\theta^\star$ but not other θ's, similar to Theorem 3 that only depends on the value gap of $Q^\star$.

Proof of Theorem 4. Following the same steps as in the Proof of Theorem 15 and $\psi \equiv 0$, we arrive at the following bound:

$$\begin{aligned}
&\mathsf{Ordec\text{-}O}_\gamma^{\mathsf{av}}(\mathcal{F}_{\mathrm{conf}}) \\
&= \min_{\rho \in \Delta(\Pi)} \max_{M \in \mathcal{M}_{\mathrm{conf}}} \Bigg\{ J_M(\pi_M) - \mathbb{E}_{\pi\sim\rho}[J_M(\pi)] \\
&\qquad - \frac{1}{2}\gamma \left(\mathbb{E}^{\pi_M, M}[V^\star(s) - Q^\star(s,a)]\right)^2 + \gamma \max_{f \in \mathcal{F}_{\mathrm{conf}}} D_{\mathsf{av}}^{\pi^\star}(f \| M^\star) \Bigg\} \\
&\le \max_{M \in \mathcal{M}_{\mathrm{conf}}} \left\{ J_M(\pi_M) - J_M(\pi^\star) - \frac{1}{2}\gamma \left(\mathbb{E}^{\pi_M, M}[V^\star(s) - Q^\star(s,a)]\right)^2 \right\} + \gamma \max_{f \in \mathcal{F}_{\mathrm{conf}}} D_{\mathsf{av}}^{\pi^\star}(f \| M^\star).
\end{aligned}$$

By the performance difference lemma and the second assumption, we have

$$\begin{aligned}
J_M(\pi_M) - J_M(\pi^\star) &= \mathbb{E}^{\pi_M, M}\left[Q_M^{\pi^\star}(s,\pi_M) - Q_M^{\pi^\star}(s,\pi^\star)\right] \\
&\le L\mathbb{E}^{\pi_M, M}\left[\|\phi(s,\pi_M) - \phi(s,\pi^\star)\|\right]
\end{aligned}$$

where $\phi(s,\pi) := \phi(s,\pi(s))$. On the other hand,

$$\begin{aligned}
\mathbb{E}^{\pi_M, M}[V^\star(s) - Q^\star(s,a)] &= \left(\mathbb{E}^{\pi_M, M}[\phi(s,\pi^\star) - \phi(s,\pi_M)]\right)^\top \theta^\star \\
&\ge \kappa \mathbb{E}^{\pi_M, M}\left[\|\phi(s,\pi^\star) - \phi(s,\pi_M)\|^\beta\right]
\end{aligned}$$

where the last inequality is by the assumption of the theorem. Therefore,

$$\begin{aligned}
&J_M(\pi_M) - J_M(\pi^\star) - \frac{1}{2}\gamma \left(\mathbb{E}^{\pi_M, M}[V^\star(s) - Q^\star(s,a)]\right)^2 \\
&\le L\mathbb{E}^{\pi_M, M}\left[\|\phi(s,\pi_M) - \phi(s,\pi^\star)\|\right] - \frac{\kappa^2\gamma}{2}\left(\mathbb{E}^{\pi_M, M}\left[\|\phi(s,\pi^\star) - \phi(s,\pi_M)\|^\beta\right]\right)^2
\end{aligned}$$

For any random variable $X \ge 0$, and any $A, B > 0$, we can bound

$$\begin{aligned}
&A\mathbb{E}[X] - B\left(\mathbb{E}[X^\beta]\right)^2 \\
&\le A\left(\mathbb{E}[X^\beta]\right)^{\frac{1}{\beta}} - B\left(\mathbb{E}[X^\beta]\right)^2 \\
&= AY^{\frac{1}{\beta}} - BY^2 && \text{(let } Y = \mathbb{E}[X^\beta]\text{)} \\
&\lesssim A^{\frac{2\beta}{2\beta-1}} B^{-\frac{1}{2\beta-1}}.
\end{aligned}$$

Plugging in $A = L$ and $B = \frac{\kappa^2\gamma}{2}$ gives $\left(\frac{L^{2\beta}}{\kappa^2\gamma}\right)^{\frac{1}{2\beta-1}}$. □

Lemma 18 (Second-Order Performance Difference Lemma)**.** *If the regularizer ψ satisfies Assumption 2, then for any policy π and model M with $Q^\star_M(s,a) - Q^\pi_M(s,a) \le B,\ \forall s,a$, we have*

$$J_M(\pi_M) - J_M(\pi) \le 3\left(1 + BHC_2^\psi\right)\mathbb{E}^{\pi_M, M}\left[\mathrm{Breg}_\psi\left(\pi, \pi_M; s\right)\right],$$

where $\mathbb{E}^{\pi,M}[g(s)]$ denotes $\sum_s d^\pi_M(s)g(s)$.

Proof. Below, denote $g(s,\pi) = \sum_a \pi(a|s)g(s,a)$. For $s \in \mathcal{S}_{H+1}$, define $V^\pi_M(s) = 0$. For any $s \in \mathcal{S}_h$, we have

$$\begin{aligned}
&V^\star_M(s) - V^\pi_M(s) \\
&= Q^\star_M(s,\pi_M) - \psi(\pi_M; s) - Q^\pi_M(s,\pi) + \psi(\pi; s) \\
&= R_M(s,\pi_M) + \mathbb{E}_{s'\sim P_M(\cdot|s,\pi_M)}[V^\star_M(s')] - R_M(s,\pi) - \mathbb{E}_{s'\sim P_M(\cdot|s,\pi)}[V^\pi_M(s')] - \psi(\pi_M; s) + \psi(\pi; s) \\
&= R_M(s,\pi_M) + \mathbb{E}_{s'\sim P_M(\cdot|s,\pi_M)}[V^\pi_M(s')] - R_M(s,\pi) - \mathbb{E}_{s'\sim P_M(\cdot|s,\pi)}[V^\pi_M(s')] - \psi(\pi_M; s) + \psi(\pi; s) \\
&\quad + \mathbb{E}_{s'\sim P_M(\cdot|s,\pi_M)}[V^\star_M(s') - V^\pi_M(s')] \\
&= Q^\pi_M(s,\pi_M) - Q^\pi_M(s,\pi) - \psi(\pi_M; s) + \psi(\pi; s) + \mathbb{E}_{s'\sim P_M(\cdot|s,\pi_M)}[V^\star_M(s') - V^\pi_M(s')] \\
&= \sum_a (\pi_M(a|s) - \pi(a|s))Q^\pi_M(s,a) - \psi(\pi_M; s) + \psi(\pi; s) + \mathbb{E}_{s'\sim P_M(\cdot|s,\pi_M)}[V^\star_M(s') - V^\pi_M(s')] \\
&= \underbrace{\sum_a (\pi_M(a|s) - \pi(a|s))(Q^\pi_M(s,a) - Q^\star_M(s,a))}_{\textbf{term}_1} \\
&\quad + \underbrace{\left(Q^\star_M(s,\pi_M) - \psi(\pi_M; s)\right) - \left(Q^\star_M(s,\pi) - \psi(\pi; s)\right)}_{\textbf{term}_2} + \mathbb{E}_{s'\sim P_M(\cdot|s,\pi_M)}[V^\star_M(s') - V^\pi_M(s')]
\end{aligned}$$

Since for any s, $\pi_M(\cdot|s) = \mathrm{argmax}_{p\in\Delta(\mathcal{A})}\{\sum_a p(a)Q^\star_M(s,a) - \psi(p)\}$, from Assumption 2, ψ is a Legendre mirror map. Thus, from Lemma 29, we have

$$\textbf{term}_2 = \mathrm{Breg}_\psi\left(\pi, \pi_M; s\right).$$

Furthermore,

$$\begin{aligned}
\textbf{term}_1 &= \sum_{a\in\mathcal{A}} \left(\pi_M(a|s) - \pi(a|s)\right)\left(Q^\pi_M(s,a) - Q^\star_M(s,a)\right) \\
&\le \frac{1}{\eta}\mathrm{KL}\left(\pi(\cdot|s)\|\pi_M(\cdot|s)\right) + \eta\sum_a \pi_M(a|s)\left(Q^\star_M(s,a) - Q^\pi_M(s,a)\right)^2 \qquad \text{(Lemma 28)} \\
&\le \frac{1}{\eta}\mathrm{KL}\left(\pi(\cdot|s)\|\pi_M(\cdot|s)\right) + \eta B\sum_a \pi_M(a|s)\left(Q^\star_M(s,a) - Q^\pi_M(s,a)\right) \\
&\qquad\qquad (0 \le Q^\star_M(s,a) - Q^\pi_M(s,a) \le B,\ \forall\pi) \\
&\le \frac{1}{\eta}C_2^\psi\mathrm{Breg}_\psi(\pi, \pi_M; s) + \eta B\cdot\mathbb{E}_{s'\sim P_M(\cdot|s,\pi_M)}\left[V^\star_M(s') - V^\pi_M(s')\right].
\end{aligned}$$

Overall,

$$V^\star_M(s) - V^\pi_M(s) \le \left(1 + \frac{1}{\eta}C_2^\psi\right)\mathrm{Breg}_\psi(\pi, \pi_M; s) + (1+\eta B)\,\mathbb{E}_{s'\sim P_M(\cdot|s,\pi_M)}\left[V^\star_M(s') - V^\pi_M(s')\right].$$

Letting $\eta = \frac{1}{BH}$ and taking expectation over $s \sim d^{\pi_M}_M$ in $\mathcal{S}_h$, we get

$$\begin{aligned}
&\sum_{s\in\mathcal{S}_h} d^{\pi_M}_M(s)\Big(V^\star_M(s) - V^\pi_M(s)\Big) \\
&\le \Big(1 + BHC_2^\psi\Big)\sum_{s\in\mathcal{S}_h} d^{\pi_M}_M(s)\mathrm{Breg}_\psi(\pi, \pi_M; s) + \left(1 + \frac{1}{H}\right)\sum_{s\in\mathcal{S}_{h+1}} d^{\pi_M}_M(s)\Big(V^\star_M(s) - V^\pi_M(s)\Big).
\end{aligned}$$

Expand this recursively over layers and using that $(1 + \frac{1}{H})^H \le 3$ gives the desired inequality.

□

Proof of Theorem 5. For any $f \in \mathcal{F}_{\mathrm{conf}}$, we have

$$\begin{aligned}
&\mathbb{E}_{(s,a)\sim d_M^{\pi_M}}\left[f(s) - f(s,a) + \psi(\pi_M; s)\right] \\
&= \mathbb{E}_{(s,a)\sim d_M^{\pi_M}}\left[\mathbb{E}_{a\sim\pi_f(\cdot|s)}[f(s,a)] - \psi(\pi_f; s) - f(s,a) + \psi(\pi_M; s)\right] \\
&= \mathbb{E}_{(s,a)\sim d_M^{\pi_M}}\left[\mathrm{Breg}_\psi\left(\pi_M, \pi_f; s\right)\right] && \text{(Lemma 29)} \\
&\geq \frac{1}{C_1^\psi}\mathbb{E}_{(s,a)\sim d_M^{\pi_M}}\left[\mathrm{Breg}_\psi\left(\pi_f, \pi_M; s\right)\right]. && \text{(Assumption 2)}
\end{aligned}$$

Thus, from Lemma 18, we have

$$\begin{aligned}
&\frac{J_M(\pi_M) - J_M(\pi_f)}{\mathbb{E}_{(s,a)\sim d_M^{\pi_M}}\left[f(s) - f(s,a) + \psi(\pi_M; s)\right]} \\
&\leq \frac{3C_1^\psi\left(1 + BHC_2^\psi\right)\mathbb{E}^{\pi_M, M}\left[\mathrm{Breg}_\psi\left(\pi_f, \pi_M; s\right)\right]}{\mathbb{E}^{\pi_M, M}\left[\mathrm{Breg}_\psi\left(\pi_f, \pi_M; s\right)\right]} = 3C_1^\psi\left(1 + BHC_2^\psi\right)
\end{aligned}$$

From Lemma 32, $B = H^2$, thus

$$\mathsf{ER}(f; \mathcal{F}_{\mathrm{conf}}) \leq 3C_1^\psi\left(1 + H^3 C_2^\psi\right) \tag{23}$$

Combining this bound with Theorem 15 together with $\mathsf{Ordec\text{-}O}_\gamma^D(\mathcal{F}_{\mathrm{conf}}) \leq \frac{4}{\gamma}\left(\mathsf{Ordec\text{-}R}^D(\mathcal{F}_{\mathrm{conf}})\right)^2$ finishes the proof. □

E.1 Comparison with [UKLS23]

[UKLS23] considers KL regularized MDPs, which is equivalent to our formulation with $\psi(p; s) = \alpha\mathrm{KL}\left(p\|\pi_{\mathrm{ref}}(\cdot|s)\right)$. Besides $Q^\star \in \mathcal{F}$, [UKLS23] additionally assume access to a function class $\mathcal{L}$ such that a specific Lagrangian multiplier is realizable and every $\ell \in \mathcal{L}$ has $\|\ell\|_\infty \leq B_{\mathcal{L}}$. Let $\hat{\pi}$ be the output policy of Algorithm 1 in [UKLS23], their Theorem 3 shows that with probability $1 - \delta$,

$$\mathbb{E}_{(s,a)\sim d^{\pi^\star}}\left[R(s,a)\right] - \mathbb{E}_{(s,a)\sim d^{\hat{\pi}}}\left[R(s,a)\right] \leq n^{-\frac{1}{4}}\mathrm{Poly}\left(|\mathcal{A}|, H, C^{\pi^\star}, R_{\max}, \log\left(\frac{|\mathcal{F}||\mathcal{L}|}{\delta}\right), B_{\mathcal{L}}\right) \tag{24}$$

where $R_{\max} = \max_{s,a}\left|\log\frac{\pi^\star(a|s)}{\pi_{\mathrm{ref}}(a|s)}\right|$. [UKLS23] also assumes the offline action distribution is identical to π_{ref}. In other words, the offline dataset $\mathcal{D}$ consists of n i.i.d. tuples (s, a, r, s'), where $s \sim \mu$ for some unknown distribution $\mu \in \Delta(\mathcal{S})$, $a \sim \pi_{\mathrm{ref}}(\cdot|s)$, $\mathbb{E}[r|s,a] = R(s,a)$, and $s' \sim P(\cdot|s,a)$. This assumption is restrictive and our methods do not need it.

Let $\hat{f}$ be the output of (5) with $\psi(p; s) = \alpha\mathrm{KL}\left(p\|\pi_{\mathrm{ref}}(\cdot|s)\right)$, we have

$$\begin{aligned}
&\left|\mathbb{E}^{\pi^\star, M^\star}\left[-\hat{f}(s,a) + R(s,a) + \mathbb{E}_{s'\sim P(\cdot|s,a)}[\hat{f}(s')]\right]\right| \\
&\geq \mathbb{E}^{\pi^\star, M^\star}\left[-\hat{f}(s,a) + R(s,a) + \mathbb{E}_{s'\sim P(\cdot|s,a)}[\hat{f}(s')]\right] + \hat{f}(s_1) - V^\star(s_1) \quad \text{(by the choice of } \hat{f}) \\
&= \mathbb{E}^{\pi^\star, M^\star}\left[\hat{f}(s) - \hat{f}(s,a) + \psi(\pi^\star; s)\right] \\
&= \mathbb{E}^{\pi^\star, M^\star}\left[\mathbb{E}_{a\sim\pi_{\hat{f}}(\cdot|s)}[\hat{f}(s,a)] - \psi(\pi_{\hat{f}}; s) - \hat{f}(s,a) + \psi(\pi^\star; s)\right] \\
&= \alpha\mathbb{E}^{\pi^\star, M^\star}\left[\mathrm{KL}(\pi^\star, \pi_{\hat{f}}; s)\right] && \text{(Lemma 29)}
\end{aligned}$$

From Lemma 27 and using the confidence set defined in Section 7.2, we have

$$
\begin{aligned}
&\mathbb{E}_{(s,a)\sim d^{\pi^\star}}[R(s,a)] - \mathbb{E}_{(s,a)\sim d^{\pi_{\hat{f}}}}[R(s,a)] \\
&\le \mathbb{E}^{\pi^\star, M^\star}\left[\mathrm{TV}\left(\pi^\star(\cdot|s), \pi_{\hat{f}}(\cdot|s)\right)\right] - \alpha\mathbb{E}^{\pi^\star, M^\star}\left[\mathrm{KL}\left(\pi^\star(\cdot|s), \pi_{\hat{f}}(\cdot|s)\right)\right] + \alpha\mathbb{E}^{\pi^\star, M^\star}\left[\mathrm{KL}\left(\pi^\star(\cdot|s), \pi_{\hat{f}}(\cdot|s)\right)\right] \\
&\le \frac{1}{\alpha} + \alpha\left|\mathbb{E}^{\pi^\star, M^\star}\left[\hat{f}(s,a) - R(s,a) - \mathbb{E}_{s'\sim P(\cdot|s,a)}\left[\hat{f}(s')\right]\right]\right| \\
&\le O\left(\frac{1}{\alpha} + \alpha B_{\mathcal{W}} H^2 \sqrt{\frac{2\log\left(|\mathcal{F}||\mathcal{W}|/\delta\right)}{n}}\right) && \text{(Lemma 7)} \\
&= O\left(H\left(\frac{B_{\mathcal{W}}^2 \log\left(|\mathcal{F}||\mathcal{W}|/\delta\right)}{n}\right)^{\frac{1}{4}}\right) && \text{(Optimal choice of } \alpha\text{)}
\end{aligned}
$$

Compared with (24), our bound eliminates the polynomial dependence on $|\mathcal{A}|$, $R_{\max}$ and remains valid for arbitrary offline action distributions, which need not coincide with π_{ref}. Moreover, under the measure of regularized objective $J(\pi)$, Algorithm in (5) can achieve $1/\epsilon^2$ sample complexity bound as shown in (15), which cannot be obtained by the analysis in [US21]. Our analysis here relies on Assumption 3, and therefore introduces an additional function class $\mathcal{W}$ for density ratio realizability. This assumption is comparable with the Lagrangian multiplier realizability condition in [UKLS23], which posits that the corresponding dual variable lies in a function class $\mathcal{L}$.

F Proofs in Section 7

In this section, we assume $f \in [0, H]$ for every $f \in \mathcal{F}$. This is valid because $Q^\star \in [0, H]$ from Lemma 32, and we can remove every f not in $[0, H]$ out of $\mathcal{F}$.

Proof of Lemma 6. We first prove $Q^\star \in \mathcal{F}_{\mathrm{conf}}$ with high probability. Define

$$
\begin{aligned}
X_g^\star(s,a,r,s') &= \left(g(s,a) - r - V^\star(s')\right)^2 - \left(Q^\star(s,a) - r - V^\star(s')\right)^2 \\
&= \left(g(s,a) - Q^\star(s,a)\right)^2 - \left(g(s,a) - Q^\star(s,a)\right)\left(Q^\star(s,a) - r - V^\star(s')\right)
\end{aligned}
$$

Given $\mathbb{E}\left[r + V^\star(s') \mid s, a\right] = (\mathcal{T}Q^\star)(s,a) = Q^\star(s,a)$, we have

$$
\mathbb{E}\left[X_g^\star(s,a,r,s') \mid s, a\right] = \left(g(s,a) - Q^\star(s,a)\right)^2.
$$

By Freedman inequality, with probability at least $1-\delta$, for any $g \in \mathcal{G}$, we have

$$
\begin{aligned}
&\left|\mathbb{E}_{(s,a,r,s')\sim\mathcal{D}}\left[X_g^\star(s,a,r,s')\right] - \mathbb{E}_{(s,a)\sim\mathcal{D}}\left[\left(g(s,a) - Q^\star(s,a)\right)^2\right]\right| \\
&\le H\sqrt{\frac{\log(|\mathcal{G}|/\delta)}{n}\mathbb{E}_{(s,a)\sim\mathcal{D}}\left[\left(g(s,a) - Q^\star(s,a)\right)^2\right]} + \frac{H^2\log(|\mathcal{G}|/\delta)}{n} \\
&\le \frac{1}{2}\mathbb{E}_{(s,a)\sim\mathcal{D}}\left[\left(g(s,a) - Q^\star(s,a)\right)^2\right] + \frac{2H^2\log(|\mathcal{G}|/\delta)}{n}, && \text{(AM-GM)}
\end{aligned}
$$

where we assume $g \in [0, H]$ for every $g \in \mathcal{G}$. This is valid because $Q^\star \in [0, H]$ from Lemma 32. Thus,

$$
\begin{aligned}
-\mathbb{E}_{(s,a,r,s')\sim\mathcal{D}}\left[X_g^\star(s,a,r,s')\right] &\le -\frac{1}{2}\mathbb{E}_{(s,a)\sim\mathcal{D}}\left[\left(g(s,a) - Q^\star(s,a)\right)^2\right] + \frac{2H^2\log(|\mathcal{G}|/\delta)}{n} \\
&\le \frac{2H^2\log(|\mathcal{G}|/\delta)}{n}.
\end{aligned}
$$

This implies

$$
\begin{aligned}
&\mathbb{E}_{(s,a,r,s')\sim\mathcal{D}}\left[\left(Q^\star(s,a) - r - V^\star(s')\right)^2\right] \\
&- \min_{g\in\mathcal{G}}\mathbb{E}_{(s,a,r,s')\sim\mathcal{D}}\left[\left(g(s,a) - r - V^\star(s')\right)^2\right] \le \frac{2H^2\log(|\mathcal{G}|/\delta)}{n} \le \epsilon_{\mathrm{stat}}.
\end{aligned}
$$

Now we prove the guarantee of $D^{\pi^\star}_{\mathsf{av}}$. Define

$$\begin{aligned}X_f(s,a,r,s') &= \left(f(s,a)-r-f(s')\right)^2 - \left((\mathcal{T}f)(s,a)-r-f(s')\right)^2 \\ &= \left(f(s,a)-(\mathcal{T}f)(s,a)\right)^2 + 2\left(f(s,a)-(\mathcal{T}f)(s,a)\right)\left((\mathcal{T}f)(s,a)-r-f(s')\right).\end{aligned}$$

Given $\mathbb{E}\left[r+f(s') \mid s,a\right] = (\mathcal{T}f)(s,a)$, we have

$$\mathbb{E}\left[X_f(s,a,r,s') \mid s,a\right] = \left(f(s,a)-(\mathcal{T}f)(s,a)\right)^2.$$

By Freedman inequality, with probability at least $1-\delta$, for any $f \in \mathcal{F}$, we have

$$\begin{aligned}&\left|\mathbb{E}_{(s,a,r,s')\sim\mathcal{D}}\left[X_f(s,a,r,s')\right] - \mathbb{E}_{(s,a)\sim\mathcal{D}}\left[\left(f(s,a)-(\mathcal{T}f)(s,a)\right)^2\right]\right| \\ &\le H\sqrt{\frac{\log(|\mathcal{F}|/\delta)}{n}\mathbb{E}_{(s,a)\sim\mathcal{D}}\left[\left(f(s,a)-(\mathcal{T}f)(s,a)\right)^2\right]} + \frac{H^2\log(|\mathcal{F}|/\delta)}{n}.\end{aligned}$$

Thus,

$$\begin{aligned}&\mathbb{E}_{(s,a)\sim\mathcal{D}}\left[\left(f(s,a)-(\mathcal{T}f)(s,a)\right)^2\right] \\ &\le \mathbb{E}_{(s,a,r,s')\sim\mathcal{D}}\left[X_f(s,a,r,s')\right] + H\sqrt{\frac{\log(|\mathcal{F}|/\delta)}{n}\mathbb{E}_{(s,a)\sim\mathcal{D}}\left[\left(f(s,a)-(\mathcal{T}f)(s,a)\right)^2\right]} + \frac{H^2\log(|\mathcal{F}|/\delta)}{n} \\ &\le \underbrace{\mathbb{E}_{(s,a,r,s')\sim\mathcal{D}}\left[\left(f(s,a)-r-f(s')\right)^2\right] - \min_{g\in\mathcal{G}}\mathbb{E}_{(s,a,r,s')\sim\mathcal{D}}\left[\left(g(s,a)-r-f(s')\right)^2\right]}_{\le\epsilon_{\mathrm{stat}}} \\ &\quad + H\sqrt{\frac{\log(|\mathcal{F}|/\delta)}{n}\mathbb{E}_{(s,a)\sim\mathcal{D}}\left[\left(f(s,a)-(\mathcal{T}f)(s,a)\right)^2\right]} + \frac{H^2\log(|\mathcal{F}|/\delta)}{n}.\end{aligned}$$

Thus, we have

$$\begin{aligned}&\mathbb{E}_{(s,a)\sim\mathcal{D}}\left[\left(f(s,a)-(\mathcal{T}f)(s,a)\right)^2\right] \\ &\le H\sqrt{\frac{\log(|\mathcal{F}|/\delta)}{n}\mathbb{E}_{(s,a)\sim\mathcal{D}}\left[\left(f(s,a)-(\mathcal{T}f)(s,a)\right)^2\right]} + O\left(\frac{H^2\log(|\mathcal{F}||\mathcal{G}|/\delta)}{n}\right).\end{aligned}$$

Solving the equation gives

$$\mathbb{E}_{(s,a)\sim\mathcal{D}}\left[\left(f(s,a)-(\mathcal{T}f)(s,a)\right)^2\right] \le O\left(\frac{H^2\log(|\mathcal{F}||\mathcal{G}|/\delta)}{n}\right). \tag{25}$$

Finally, with probability $1-\delta$, for any $f \in \mathcal{F}_{\mathrm{conf}}$, we have

$$\begin{aligned}D^{\pi^\star}_{\mathsf{av}}(f\|M^\star) &= \left(\mathbb{E}_{(s,a)\sim d^{\pi^\star}}\left[f(s,a)-R(s,a)-\mathbb{E}_{s'\sim P(\cdot|s,a)}\left[f(s')\right]\right]\right)^2 \\ &\le \mathbb{E}_{(s,a)\sim d^{\pi^\star}}\left[\left(f(s,a)-R(s,a)-\mathbb{E}_{s'\sim P(\cdot|s,a)}\left[f(s')\right]\right)^2\right] \\ &\le HC^{\pi^\star}\mathbb{E}_{(s,a)\sim\mu}\left[\left(f(s,a)-R(s,a)-\mathbb{E}_{s'\sim P(\cdot|s,a)}\left[f(s')\right]\right)^2\right] \\ &\le O\left(\frac{C^{\pi^\star}H^3\log(|\mathcal{F}||\mathcal{G}|/\delta)}{n}\right). \qquad \text{(Combine (25) and Lemma 26)}\end{aligned}$$

□

Proof of Lemma 7. We first prove $Q^\star \in \mathcal{F}_{\mathrm{conf}}$. From Hoeffding's inequality, with probability $1-\delta$, for any $w \in \mathcal{W}$, we have

$$\begin{aligned}&\left|\mathbb{E}_{(s,a,r,s')\sim\mathcal{D}}\left[w(s,a)\left(Q^\star(s,a)-r-V^\star(s')\right)\right]\right| \\ &\le \left|\mathbb{E}_{(s,a)\sim\mathcal{D}}\left[w(s,a)\left(Q^\star(s,a)-(\mathcal{T}Q^\star)(s,a)\right)\right]\right| + B_{\mathcal{W}}H\sqrt{\frac{2\log(|\mathcal{W}|/\delta)}{n}} \\ &= B_{\mathcal{W}}H\sqrt{\frac{2\log(|\mathcal{W}|/\delta)}{n}} \\ &\le \epsilon_{\mathrm{stat}}.\end{aligned}$$

Thus, $Q^\star \in \mathcal{F}_{\mathrm{conf}}$. Now we prove the guarantee of $D^{\pi^\star}_{\mathsf{av}}$. For any $f \in \mathcal{F}_{\mathrm{conf}}$, we have

$$
\begin{aligned}
&\left|\mathbb{E}_{(s,a)\sim d^{\pi^\star}}\left[f(s,a) - R(s,a) - \mathbb{E}_{s'\sim P(\cdot|s,a)}\left[f(s')\right]\right]\right| \\
&= H\left|\mathbb{E}_{(s,a)\sim\mu}\left[w^{\pi^\star}(s,a)\left(f(s,a) - R(s,a) - \mathbb{E}_{s'\sim P(\cdot|s,a)}\left[f(s')\right]\right)\right]\right| \\
&\le H\left|\mathbb{E}_{(s,a,r,s')\sim\mathcal{D}}\left[w^{\pi^\star}(s,a)\left(f(s,a) - r - f(s')\right)\right]\right| + O\left(B_{\mathcal{W}}H^2\sqrt{\frac{\log(|\mathcal{F}|/\delta)}{n}}\right) \\
&\qquad\qquad \text{(Hoeffding's Inequality)} \\
&\le H\max_{w\in\mathcal{W}}\left|\mathbb{E}_{(s,a,r,s')\sim\mathcal{D}}\left[w(s,a)\left(f(s,a) - r - f(s')\right)\right]\right| + O\left(B_{\mathcal{W}}H^2\sqrt{\frac{\log(|\mathcal{F}|/\delta)}{n}}\right) \\
&\le O\left(H\epsilon_{\mathrm{stat}}\right) \qquad\qquad \text{(Use the definition of } \epsilon_{\mathrm{stat}}\text{)}
\end{aligned}
$$

Thus,

$$
D^{\pi^\star}_{\mathsf{av}}(f\|M^\star) = \left(\mathbb{E}_{(s,a)\sim d^{\pi^\star}}\left[f(s,a) - R(s,a) - \mathbb{E}_{s'\sim P(\cdot|s,a)}\left[f(s')\right]\right]\right)^2 \le O\left(H^2\epsilon^2_{\mathrm{stat}}\right).
$$

□

Proof of Lemma 8.

$$
\begin{aligned}
D^{\pi^\star}_{\mathsf{av}}(f\|M^\star) &= \left(\mathbb{E}^{\pi^\star,M^\star}\left[f(s,a) - R(s,a) - \mathbb{E}_{s'\sim P(\cdot|s,a)}[f(s')]\right]\right)^2 \\
&= \langle X(\pi^\star), W(f)\rangle^2 && \text{(by Assumption 4)} \\
&\le \left(X(\pi^\star)^\top\Sigma^\dagger_{\underline{\mu}}X(\pi^\star)\right)\left(W(f)^\top\Sigma_{\underline{\mu}}W(f)\right) && \text{(Cauchy-Schwarz)} \\
&= \underline{C}^{\pi^\star}\mathbb{E}_{\pi\sim\underline{\mu}}\left[\langle X(\pi), W(f)\rangle^2\right] && \text{(Definition 4)} \\
&\le \underline{C}^{\pi^\star}\mathbb{E}_{\pi\sim\underline{\mu}}\left[\left(\mathbb{E}^{\pi,M^\star}\left[f(s,a) - R(s,a) - \mathbb{E}_{s'\sim P(\cdot|s,a)}[f(s')]\right]\right)^2\right] \\
&= \underline{C}^{\pi^\star}\mathbb{E}_{\pi\sim\underline{\mu}}[D^{\pi}_{\mathsf{av}}(f\|M^\star)]. && (26)
\end{aligned}
$$

On the other hand, by the definition of L_{br}, we have

$$
\begin{aligned}
\mathbb{E}[L_{\mathsf{br}}(f)] &= \mathbb{E}_{\pi\sim\underline{\mu}}\left[\mathbb{E}^{\pi,M^\star}[f(s,a) - r - f(s')]\cdot\mathbb{E}^{\pi,M^\star}[f(s,a) - r - f(s')]\right] \\
&= \mathbb{E}_{\pi\sim\underline{\mu}}\left[D^{\pi}_{\mathsf{av}}(f\|M^\star)\right] && (27)
\end{aligned}
$$

and by Hoeffding's inequality, with probability at least $1-\delta$,

$$
\left|L_{\mathsf{br}}(f) - \mathbb{E}_{\pi\sim\underline{\mu}}\left[D^{\pi}_{\mathsf{av}}(f\|M^\star)\right]\right| \le H^2\sqrt{\frac{\log(2|\mathcal{F}|/\delta)}{2n}}.
$$

Under this event,

$$
L_{\mathsf{br}}(Q^\star) \le \mathbb{E}_{\pi\sim\underline{\mu}}\left[D^{\pi}_{\mathsf{av}}(Q^\star\|M^\star)\right] + H^2\sqrt{\frac{\log(2|\mathcal{F}|/\delta)}{2n}} = H^2\sqrt{\frac{\log(2|\mathcal{F}|/\delta)}{2n}}
$$

and thus $Q^\star \in \mathcal{F}_{\mathrm{conf}}$. Also, by (26) we have

$$
\begin{aligned}
\max_{f\in\mathcal{F}_{\mathrm{conf}}} D^{\pi^\star}_{\mathsf{av}}(f\|M^\star) &\le \underline{C}^{\pi^\star}\max_{f\in\mathcal{F}_{\mathrm{conf}}}\mathbb{E}_{\pi\sim\underline{\mu}}[D^{\pi}_{\mathsf{av}}(f\|M^\star)] \\
&\le \underline{C}^{\pi^\star}\max_{f\in\mathcal{F}_{\mathrm{conf}}}\left(L_{\mathsf{br}}(f) + H^2\sqrt{\frac{\log(2|\mathcal{F}|/\delta)}{2n}}\right) && \text{(by (27))} \\
&\le \underline{C}^{\pi^\star}\cdot 2H^2\sqrt{\frac{\log(2|\mathcal{F}|/\delta)}{2n}}. && \text{(by the definition of } \mathcal{F}_{\mathrm{conf}}\text{)}
\end{aligned}
$$

□

F.1 Low Bellman rank setting: lower bounds without policy feature coverage or without double policy samples

For the necessity of the dependence on $\underline{C}^{\pi^\star}$ and the necessity of double policy samples, we rely on an extension to the lower bound construction in [JRSW24]. The two claims are proven in Lemma 21 and Lemma 22, respectively.

F.1.1 Reusing the hard instance in [JRSW24]

Specifically, we build upon the construction of [JRSW24] in the admissible setting ($\mu = \frac{1}{H} d^{\pi_b}$ for some π_b) described in their Theorem 4.1. Their construction is illustrated in Figure 6, with details described in the caption. Moreover, their construction has the following properties:

- They focus on *policy evaluation*. There are two actions. The policy to be evaluated is $\pi_e = (1, 0)$, i.e., always choosing action 1. The learner is given a function set $\mathcal{F} = \{f_1, f_2\}$ such that $Q^{\pi_e} \in \mathcal{F}$.
- The offline data distribution μ satisfies $\mu = \frac{1}{H} d^{\pi_b}$ where $\pi_b = (\frac{1}{H^2}, 1 - \frac{1}{H^2})$.
- Their construction satisfies $C^\pi = O(H^3)$ for any policy π.
- All rewards and Q^π values are in the range of $[-1, 1]$,

The construction depends on a parameter $\epsilon \in [0, \frac{1}{15}]$ which makes $|\mathbb{E}_{s\sim\rho}[f_1(s)] - \mathbb{E}_{s\sim\rho}[f_2(s)]| = 2\epsilon$, where ρ is the initial state distribution. They prove that to estimate $\mathbb{E}_{s\sim\rho}[V^{\pi_e}(s)]$ up to precision ϵ, the learner must use at least $2^H/\epsilon$ samples drawn from μ. Below we describe how to turn this into a hard case for *policy optimization*.

- Subtract the instantaneous reward of action 2 by 10 on every state. As this makes action 2 very bad, the optimal policy on any state is action 1.
- Based on their function class $\mathcal{F}$ for Q^{π_e}, define the new function class $\mathcal{F}' = \{f_1', f_2'\}$ such that $f_i'(s, 1) = f_i(s, 1)$ and $f_i'(s, 2) = f_i(s, 2) - 10$ for $i = 1, 2$. Below, we argue that $Q^\star \in \mathcal{F}'$ in the modified environments. In the modified environment, since action 1 is always optimal, $Q^\star(s, 1)$ is the value of always taking action 1, which is the same as $Q^{\pi_e}(s, 1)$ in the original environment. For the same reason, in the modified environment, $Q^\star(s, 2)$ is the value of taking action 2 for one step and taking action 1 thereafter. Therefore, $Q^\star(s, 2) = -10 + R(s, 2) + \mathbb{E}_{s'\sim P(\cdot|s,2)}[V^{\pi_e}(s')] = Q^{\pi_e}(s, 2) - 10$, where R, V^{π_e}, Q^{π_e} are the reward and values in the original environment. Since $Q^{\pi_e} \in \mathcal{F}$ in the original environment, this implies $Q^\star \in \mathcal{F}'$ in the modified environment.
- This mere modification of offsetting action 2's reward by a constant does not make evaluating π_e any easier. Since $\pi^\star$ in the modified environment is the same as π_e in the original environment, and their state values are the same, it becomes a hard instance of evaluating $\mathbb{E}_{s\sim\rho}[V^\star(s)]$.
- Finally, we add one additional state s_0 with two actions. Choosing p leads to zero instantaneous reward and transitions to the MDP described above; choosing q leads to a deterministic reward of $\mathsf{avg} = \mathbb{E}_{s\sim\rho}\left[\frac{f_1'(s)+f_2'(s)}{2}\right]$ and a chain of state with single action and zero reward until the end of the episode (Figure 7). Augment the functions with $f_1'(s_0, p) = \mathbb{E}_{s\sim\rho}[f_1'(s)]$ and $f_2'(s_0, p) = \mathbb{E}_{s\sim\rho}[f_2'(s)]$ and $f_1'(s_0, q) = f_2'(s_0, q) = \mathsf{avg}$. Define behavior policy $\pi_b(\cdot|s_0) = (\frac{1}{2}, \frac{1}{2})$ and keep it unchanged in other states.
- Let $\epsilon = |\mathbb{E}_{s\sim\rho}[f_1'(s)] - \mathbb{E}_{s\sim\rho}[f_2'(s)]| = \frac{1}{15}$. Then the two actions on s_0 has a gap of $|Q^\star(s_0, p) - Q^\star(s_0, q)| = \frac{1}{2}|\mathbb{E}_{s\sim\rho}[f_1'(s)] - \mathbb{E}_{s\sim\rho}[f_2'(s)]| = \frac{1}{2} \cdot \frac{1}{15} = \frac{1}{30}$. Therefore, in order to find an $\frac{1}{60}$-optimal policy, the learner must distinguish the two worlds, i.e., estimating $Q^\star(s_0, p) = \mathbb{E}_{s\sim\rho}[V^\star(s)]$ up to an accuracy of $\frac{1}{30}$, which is proven to be hard by [JRSW24] (needs $\Omega(2^H)$ samples).

F.1.2 Relate the hard instance to our setting

We will use the instance in Appendix F.1.1 to show that both *policy feature coverage* and *double policy samples* are necessary for polynomial sample complexity in low Q-Bellman rank settings.

We first show that the hard instance in Appendix F.1.1 is a linear $Q^\star/V^\star$ setting [DKL+21] (Lemma 19), which is further an example of an MDP with low Q-Bellman rank (Lemma 20).

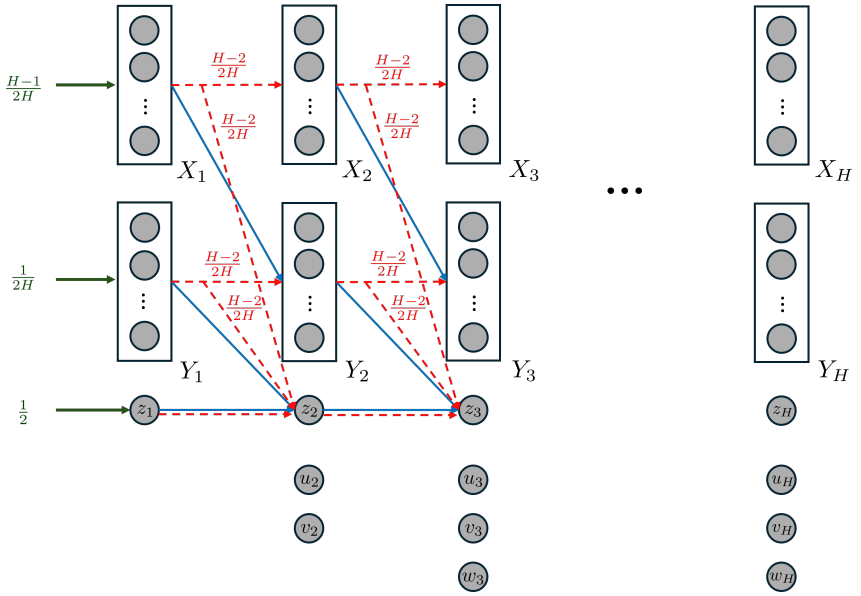


Figure 6: The construction of [JRSW24]. The numbers on the left of the green arrows specify the initial state distribution. There are two actions. Taking either action on state in $X_h \cup Y_h \cup \{z_h\}$ leads to transitions to u_{h+1} and v_{h+1} with probabilities $P(u_{h+1}|s,a) + P(v_{h+1}|s,a) = \frac{2}{H}$ which are not specified in the figure. Blue arrow specifies the transition if taking action 1 besides those to u_{h+1} and v_{h+1}, and red arrow if taking action 2. The numbers on the red arrows are the transition probabilities. The blue and red arrows without a number on it have a transition probability $\frac{H-2}{H}$. Every transition to group X_h or Y_h results in a uniform distribution over that group. On u_h and v_h and w_h, taking any action leads to a deterministic transition to w_{h+1}.

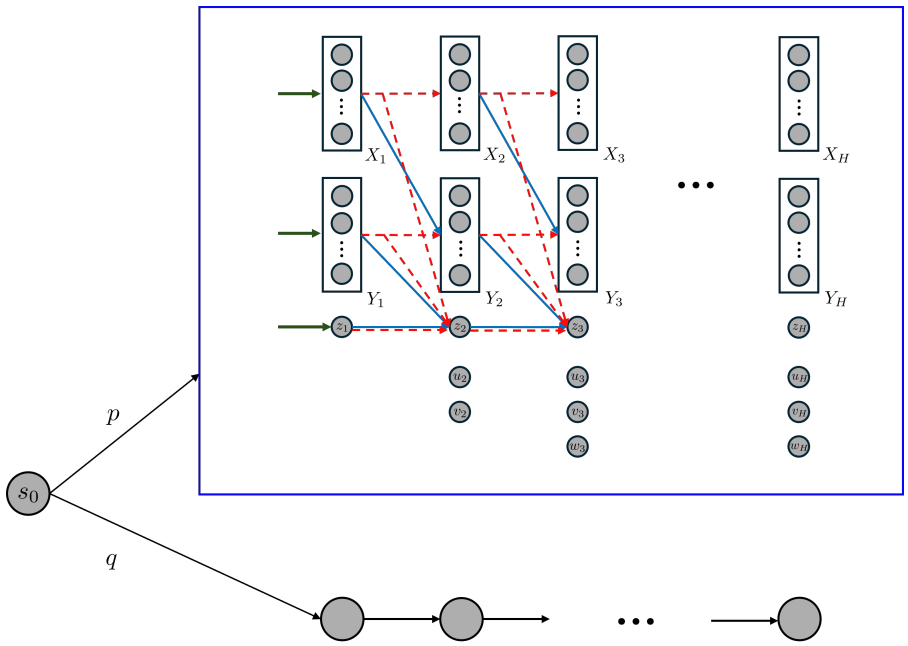


Figure 7: A hard instance for policy optimization. The blue box is [JRSW24]'s construction (Figure 6) with modified reward of action 2 (see Appendix F.1.1 for a description of the construction).

Definition 6 (Linear $Q^\star/V^\star$ setting)**.** *The learner has known features $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ and $\xi : \mathcal{S} \to \mathbb{R}^d$ such that $Q^\star(s,a) = \phi(s,a)^\top \theta^\star$ and $V^\star(s) = \xi(s)^\top w^\star$ for some unknown $\theta^\star$ and $w^\star$.*

Lemma 19. *The instance described in Appendix F.1.1 is a linear $Q^\star/V^\star$ setting (Definition 6).*

Proof. In the construction, the learner has a function set $\mathcal{F} = \{f_1, f_2\}$ such that $Q^\star \in \mathcal{F}$. Then

$$\phi(s,a) = \begin{bmatrix} f_1(s,a) \\ f_2(s,a) \end{bmatrix}, \qquad \xi(s) = \begin{bmatrix} f_1(s) \\ f_2(s) \end{bmatrix} = \begin{bmatrix} \max_a f_1(s,a) \\ \max_a f_2(s,a) \end{bmatrix}$$

satisfies Definition 6 with $d = 2$. □

Lemma 20. *Linear $Q^\star/V^\star$ setting has low Q-Bellman rank.*

Proof. In the linear $Q^\star/V^\star$ setting, let f be the function associated with weight (θ, w). Then

$$\begin{aligned}
&\mathbb{E}_{(s,a)\sim d^\pi}\left[f(s,a) - R(s,a) - \mathbb{E}_{s'\sim P(\cdot|s,a)}[f(s')]\right] \\
&= \mathbb{E}_{(s,a)\sim d^\pi}\left[f(s,a) - R(s,a) - \mathbb{E}_{s'\sim P(\cdot|s,a)}[f(s')]\right] \\
&\qquad - \mathbb{E}_{(s,a)\sim d^\pi}\left[Q^\star(s,a) - R(s,a) - \mathbb{E}_{s'\sim P(\cdot|s,a)}[V^\star(s')]\right] \qquad \text{(subtracting zero)} \\
&= \mathbb{E}_{(s,a)\sim d^\pi}\left[\phi(s,a)^\top\theta - R(s,a) - \mathbb{E}_{s'\sim P(\cdot|s,a)}[\xi(s')^\top w]\right] \\
&\qquad - \mathbb{E}_{(s,a)\sim d^\pi}\left[\phi(s,a)^\top\theta^\star - R(s,a) - \mathbb{E}_{s'\sim P(\cdot|s,a)}[\xi(s')^\top w^\star]\right] \\
&= \mathbb{E}_{(s,a)\sim d^\pi}\left[\phi(s,a)^\top(\theta - \theta^\star) - \mathbb{E}_{s'\sim P(\cdot|s,a)}[\xi(s')^\top(w - w^\star)]\right] \\
&= \langle X(\pi), W(f)\rangle
\end{aligned}$$

where $X(\pi) = \mathbb{E}_{(s,a)\sim d^\pi}[\phi(s,a), \mathbb{E}_{s'\sim P(\cdot|s,a)}[\xi(s')]]$ and $W(f) = [\theta - \theta^\star, w - w^\star]$. □

Lemma 21. *In the low Q-Bellman rank setting, it is impossible to achieve $\mathbb{E}[J(\pi^\star) - J(\hat{\pi})] \le \epsilon$ with only $\mathrm{poly}(d, H, \epsilon^{-1}, \log|\mathcal{F}|, C^{\pi^\star})$ samples, even with access to double policy samples (Assumption 5) and with constant value gap $\Delta_{Q^\star} = \Theta(1)$.*

Proof. In the instance constructed in Appendix F.1.1 (which has low Q-Bellman rank by Lemma 19 and Lemma 20), the offline distribution is admissible, i.e., $\mu = \frac{1}{H} d^{\pi_b}$ for some fixed policy π_b. Hence, it satisfies the double policy sample assumption—any two (s,a,r,s') tuples in the dataset are generated by the same policy. Furthermore, $C^{\pi^\star} = O(H^3)$ and $\log|\mathcal{F}| = \log 2$ in the construction. It is also easy to see that $\Delta_{Q^\star} = \Theta(1)$ in the constructed instance as we make action 2 highly sub-optimal compared to action 1 on all states. However, as proven by [JRSW24], polynomial sample complexity is impossible. □

Lemma 22. *In the low Q-Bellman rank setting, it is impossible to achieve $\mathbb{E}[J(\pi^\star) - J(\hat{\pi})] \le \epsilon$ with only $\mathrm{poly}(d, H, \epsilon^{-1}, \log|\mathcal{F}|, \underline{C}^{\pi^\star})$ samples if the learner only gets access to single policy samples (s,a,r,s') but not double policy samples (Assumption 5), even with constant value gap $\Delta_{Q^\star} = \Theta(1)$.*

Proof. Below, we construct $\underline{\mu}$ so that the process "sample $\pi \sim \underline{\mu}$ and then sample $(s,a,r,s') \sim d^\pi$" is equivalent to "sample $(s,a,r,s') \sim d^{\pi_b}$" for the π_b defined in Appendix F.1.1. The latter has been proven to be hard in [JRSW24]. If we can further show that $\underline{\mu}$ makes $\underline{C}^{\pi^\star} = X(\pi^\star)^\top \Sigma_{\underline{\mu}}^\dagger X(\pi^\star)$ small, then the impossibility result follows from that of [JRSW24].

Recall that π_b is a randomized policy that is equal to $(\frac{1}{2}, \frac{1}{2})$ on state s_0, and equal to $\pi_b = (\frac{1}{H^2}, 1 - \frac{1}{H^2})$ in all other states. We define $\underline{\mu}$ to be a distribution over *deterministic policies* in the following way:

$$\underline{\mu}(\delta_{a_0} \circ \delta_{a_1} \circ \delta_{a_2} \circ \cdots \circ \delta_{a_H}) = \frac{1}{2} \prod_{h=1}^{H} \left(\frac{1}{H^2}\right)^{\mathbf{1}\{a_h = 1\}} \left(1 - \frac{1}{H^2}\right)^{\mathbf{1}\{a_h = 2\}}$$

where δ_a is the one-hot action distribution concentrating on action a, a_0 is the action taken on s_0, and a_h where $h \ge 1$ is the action taken on step h in the trajectory. That means, this set of deterministic policies only determine their action based on the step index and ignore the state. Clearly, sampling $\pi \sim \underline{\mu}$ and drawing one trajectory from π induces the same state-action distribution as executing π_b.

Next, we bound $\underline{C}^{\pi^\star} = X(\pi^\star)^\top \Sigma_{\underline{\mu}}^\dagger X(\pi^\star)$. One observation in the construction of Figure 6 is that if a policy choose action 1 in both layers $h = 1$ and $h = 2$, then the state-action distribution in the rest of the trajectory is policy-independent. Indeed, if a policy chooses action 1 in layers $h = 1$ and $h = 2$, then it must lands in states $\{z_h, u_h, v_h, w_h\}$ for $h \geq 3$, on which choosing action 1 and 2 induces the same transition.

Therefore,

$$\begin{aligned}
\Sigma_{\underline{\mu}} &= \mathbb{E}_{\pi\sim\underline{\mu}}[X(\pi)X(\pi)^\top] = \mathbb{E}_{\pi\sim\underline{\mu}}\begin{bmatrix} \mathbb{E}_{(s,a)\sim d^\pi}[\phi(s,a)] \\ \mathbb{E}_{(s,a)\sim d^\pi}\mathbb{E}_{s'\sim P(\cdot|s,a)}[\xi(s')] \end{bmatrix}\begin{bmatrix} \mathbb{E}_{(s,a)\sim d^\pi}[\phi(s,a)] \\ \mathbb{E}_{(s,a)\sim d^\pi}\mathbb{E}_{s'\sim P(\cdot|s,a)}[\xi(s')] \end{bmatrix}^\top \\
&\succeq \sum_\pi \underline{\mu}(\pi)\mathbf{1}\{\pi_1 = 1, \pi_2 = 1\}\begin{bmatrix} \mathbb{E}_{(s,a)\sim d^\pi}[\phi(s,a)] \\ \mathbb{E}_{(s,a)\sim d^\pi}\mathbb{E}_{s'\sim P(\cdot|s,a)}[\xi(s')] \end{bmatrix}\begin{bmatrix} \mathbb{E}_{(s,a)\sim d^\pi}[\phi(s,a)] \\ \mathbb{E}_{(s,a)\sim d^\pi}\mathbb{E}_{s'\sim P(\cdot|s,a)}[\xi(s')] \end{bmatrix}^\top \\
&\succeq \frac{1}{2}\cdot\frac{1}{H^2}\cdot\frac{1}{H^2}\begin{bmatrix} \mathbb{E}_{(s,a)\sim d^{\pi^\star}}[\phi(s,a)] \\ \mathbb{E}_{(s,a)\sim d^{\pi^\star}}\mathbb{E}_{s'\sim P(\cdot|s,a)}[\xi(s')] \end{bmatrix}\begin{bmatrix} \mathbb{E}_{(s,a)\sim d^{\pi^\star}}[\phi(s,a)] \\ \mathbb{E}_{(s,a)\sim d^{\pi^\star}}\mathbb{E}_{s'\sim P(\cdot|s,a)}[\xi(s')] \end{bmatrix}^\top \\
&= \frac{1}{2H^4}X(\pi^\star)X(\pi^\star)^\top,
\end{aligned}$$

implying that $\underline{C}^{\pi^\star} = X(\pi^\star)^\top \Sigma_{\underline{\mu}}^\dagger X(\pi^\star)$ is polynomial in H. However, by the lower bound of [JRSW24], polynomial sample complexity is impossible. □

G Conservative Q-Learning

Conservative Q-Learning (CQL) is one of the most widely adopted baselines in offline RL [KZTL20]. We consider the following form of CQL:

$$\hat{f} = \operatorname{argmin}_{f\in\mathcal{F}} \mathbb{E}_{(s,a)\sim\mathcal{D}}[f(s) - f(s,a)] + \lambda\mathbb{E}_{(s,a)\sim\mathcal{D}}\left[\left(f(s,a) - [\hat{\mathcal{T}}f](s,a)\right)^2\right] \tag{28}$$

where $[\hat{\mathcal{T}}f] = \operatorname{argmin}_{g\in\mathcal{G}} \mathbb{E}_{(s,a,r,s')\sim\mathcal{D}}\left[(g(s,a) - r - f(s'))^2\right]$ and $\mathcal{G}$ is as defined in Assumption 1. Furthermore, CQL uses regularizer $\psi(p; s) = \alpha\mathrm{KL}(p, \pi_{\mathrm{ref}}(\cdot|s))$ as in Section 6.2. We have:

Theorem 23. *Assume $Q^\star$-realizability and Bellman completeness . Also, assume μ is admissible: $\sum_{(s,a)\in\mathcal{S}_h\times\mathcal{A}} \mu(s,a)P(s'|s,a) = \mu(s')$ for $s' \in \mathcal{S}_{h+1}$. Then CQL (28) has with probability $1-\delta$,*

$$J(\pi^\star) - J(\pi_{\hat{f}}) \leq O\left(\frac{H^4}{\alpha^2}\left(\frac{\lambda H^2\log(|\mathcal{F}||\mathcal{G}|/\delta)}{n} + \frac{HC^{\pi^\star}}{\lambda}\right)\right).$$

CQL does not maintain a confidence set, but uses a penalty (second term in (28)) to ensure that $\hat{f}$ conforms with the Bellman optimality equation. The first term in (28) implicitly encourages value pessimism, which we elaborate below. As the second term in (28) forces $\mathbb{E}_{(s,a)\sim\mu}[f(s,a) - R(s,a) - \mathbb{E}_{s'\sim P(\cdot|s,a)}[f(s')]] \approx 0$, we have

$$\mathbb{E}_{(s,a)\sim\mu}[f(s) - \mathbb{E}_{s'\sim P(\cdot|s,a)}[f(s')]] \approx \mathbb{E}_{(s,a)\sim\mu}[R(s,a) + f(s) - f(s,a)].$$

The left-hand side is the difference of expected value of f between consecutive layers, while the right-hand side is the reward plus $\mathbb{E}_{(s,a)\sim\mu}[f(s) - f(s,a)]$ that reflects how much more f values out-of-sample actions than in-sample actions. When this is large, the expected value of f grows more rapidly when propagating over layers. To ensure pessimism (i.e., small $f(s_1)$), (28) adds $\mathbb{E}_{(s,a)\sim\mu}[f(s) - f(s,a)]$ as part of the loss. This pessimistic behavior is formalized in Lemma 25, and other parts of the proof are similar to those in Section 6.2. Before we prove Theorem 23, we first introduce Lemma 24 and Lemma 25 to bound the error of estimation and pessimistic decision.

Lemma 24. *With probability at least $1-\delta$, for any $f \in \mathcal{F}$,*

$$\mathbb{E}_{(s,a)\sim\mathcal{D}}\left[\left((\hat{\mathcal{T}}f)(s,a) - (\mathcal{T}f)(s,a)\right)^2\right] \leq O\left(\frac{H^2\log(|\mathcal{F}||\mathcal{G}|/\delta)}{n}\right).$$

Proof. By the definition of $\hat{\mathcal{T}} f$, we have

$$\begin{aligned}
0 &\geq \mathbb{E}_{(s,a,r,s')\sim\mathcal{D}}\left[\left((\hat{\mathcal{T}}f)(s,a) - r - f(s')\right)^2 - ((\mathcal{T}f)(s,a) - r - f(s'))^2\right] \\
&= \mathbb{E}_{(s,a,r,s')\sim\mathcal{D}}\left[\left((\hat{\mathcal{T}}f)(s,a) - (\mathcal{T}f)(s,a)\right)\left((\hat{\mathcal{T}}f)(s,a) + (\mathcal{T}f)(s,a) - 2r - 2f(s')\right)\right] \\
&= \mathbb{E}_{(s,a)\sim\mathcal{D}}\left[\left((\hat{\mathcal{T}}f)(s,a) - (\mathcal{T}f)(s,a)\right)^2\right] \\
&\quad + \mathbb{E}_{(s,a,r,s')\sim\mathcal{D}}\left[\left((\hat{\mathcal{T}}f)(s,a) - (\mathcal{T}f)(s,a)\right)(2(\mathcal{T}f)(s,a) - 2r - 2f(s'))\right]
\end{aligned}$$

Note that $\mathbb{E}[r + f(s')|s,a] = R(s,a) + \mathbb{E}_{s'\sim P(\cdot|s,a)}[f(s')] = (\mathcal{T}f)(s,a)$. By Freedman's inequality, with probability at least $1-\delta$,

$$\begin{aligned}
&\left|\mathbb{E}_{(s,a,r,s')\sim\mathcal{D}}\left[\left((\hat{\mathcal{T}}f)(s,a) - (\mathcal{T}f)(s,a)\right)(2(\mathcal{T}f)(s,a) - 2r - 2f(s'))\right]\right| \\
&\lesssim H\sqrt{\frac{\log(|\mathcal{F}||\mathcal{G}|/\delta)}{n}\mathbb{E}_{(s,a)\sim\mathcal{D}}\left[\left((\hat{\mathcal{T}}f)(s,a) - (\mathcal{T}f)(s,a)\right)^2\right]} + \frac{H^2\log(|\mathcal{F}||\mathcal{G}|/\delta)}{n}
\end{aligned}$$

Combining inequalities above and denoting $X = \mathbb{E}_{(s,a)\sim\mathcal{D}}\left[\left((\hat{\mathcal{T}}f)(s,a) - (\mathcal{T}f)(s,a)\right)^2\right]$, we get

$$X \lesssim H\sqrt{X\cdot\frac{\log(|\mathcal{F}||\mathcal{G}|/\delta)}{n}} + \frac{H^2\log(|\mathcal{F}||\mathcal{G}|/\delta)}{n},$$

implying that $X \lesssim \frac{H^2\log(|\mathcal{F}||\mathcal{G}|/\delta)}{n}$. □

Lemma 25. *With probability at least* $1-\delta$*,*

$$\hat{f}(s_1) \leq J(\pi^\star) + O\left(\sqrt{\frac{\iota}{n}} + \frac{\lambda\iota}{n} + \frac{1}{\lambda}\right) - \mathbb{E}_{(s,a)\sim\mathcal{D}}\left[\frac{\lambda}{2}\varepsilon(s,a;\hat{f})^2\right]$$

where $\varepsilon(s,a;f) := f(s,a) - (\hat{\mathcal{T}}f)(s,a)$ *and* $\iota = H^2\log(|\mathcal{F}||\mathcal{G}|/\delta)$

Proof. In the proof we fix a failure probability δ and denote $\iota = H^2\log(|\mathcal{F}||\mathcal{G}|/\delta)$. The objective of the algorithm can be written as

$$\begin{aligned}
&\mathbb{E}_{(s,a)\sim\mathcal{D}}[f(s) - f(s,a)] + \lambda\mathbb{E}_{(s,a)\sim\mathcal{D}}\left[(f(s,a) - (\hat{\mathcal{T}}f)(s,a))^2\right] \\
&= \mathbb{E}_{(s,a)\sim\mathcal{D}}\left[f(s) - (\hat{\mathcal{T}}f)(s,a)\right] - \mathbb{E}_{(s,a)\sim\mathcal{D}}\left[f(s,a) - (\hat{\mathcal{T}}f)(s,a)\right] + \lambda\mathbb{E}_{(s,a)\sim\mathcal{D}}\left[(f(s,a) - (\hat{\mathcal{T}}f)(s,a))^2\right] \\
&= \mathbb{E}_{(s,a)\sim\mathcal{D}}\left[f(s) - (\hat{\mathcal{T}}f)(s,a)\right] + \mathbb{E}_{(s,a)\sim\mathcal{D}}\left[-\varepsilon(s,a;f) + \lambda\varepsilon(s,a;f)^2\right]
\end{aligned}$$

where $\varepsilon(s,a;f) := f(s,a) - (\hat{\mathcal{T}}f)(s,a)$. Therefore, by the optimality of $\hat{f}$,

$$\begin{aligned}
&\mathbb{E}_{(s,a)\sim\mathcal{D}}\left[\hat{f}(s) - (\hat{\mathcal{T}}\hat{f})(s,a)\right] + \mathbb{E}_{(s,a)\sim\mathcal{D}}\left[-\varepsilon(s,a;\hat{f}) + \lambda\varepsilon(s,a;\hat{f})^2\right] \\
&\leq \mathbb{E}_{(s,a)\sim\mathcal{D}}\left[V^\star(s) - (\hat{\mathcal{T}}Q^\star)(s,a)\right] + \mathbb{E}_{(s,a)\sim\mathcal{D}}\left[-\varepsilon(s,a;Q^\star) + \lambda\varepsilon(s,a;Q^\star)^2\right] \\
&\leq \mathbb{E}_{(s,a)\sim\mathcal{D}}\left[V^\star(s) - (\hat{\mathcal{T}}Q^\star)(s,a)\right] + O\left(\sqrt{\frac{\iota}{n}} + \frac{\lambda\iota}{n}\right) \qquad (29)
\end{aligned}$$

where in the last inequality we use that

$$\begin{aligned}
\mathbb{E}_{(s,a)\sim\mathcal{D}}\left[\varepsilon(s,a;Q^\star)^2\right] &= \mathbb{E}_{(s,a)\sim\mathcal{D}}\left[\left(Q^\star(s,a) - (\hat{\mathcal{T}}Q^\star)(s,a)\right)^2\right] \\
&= \mathbb{E}_{(s,a)\sim\mathcal{D}}\left[\left((\mathcal{T}Q^\star)(s,a) - (\hat{\mathcal{T}}Q^\star)(s,a)\right)^2\right] \lesssim \frac{\iota}{n}. \quad \text{(by Lemma 24)}
\end{aligned}$$

To further bound other terms in (29), note that by Lemma 24 and Hoeffding's inequality, we have for all f,

$$\begin{aligned}
&\left|\mathbb{E}_{(s,a)\sim\mathcal{D}}\left[f(s)-(\hat{\mathcal{T}}f)(s,a)\right]-\mathbb{E}_{(s,a)\sim\mu}\left[f(s)-(\mathcal{T}f)(s,a)\right]\right| \\
&\le \left|\mathbb{E}_{(s,a)\sim\mathcal{D}}\left[(\hat{\mathcal{T}}f)(s,a)-(\mathcal{T}f)(s,a)\right]\right|+\left|\mathbb{E}_{(s,a)\sim\mathcal{D}}\left[f(s)-(\mathcal{T}f)(s,a)\right]-\mathbb{E}_{(s,a)\sim\mu}\left[f(s)-(\mathcal{T}f)(s,a)\right]\right| \\
&\lesssim \sqrt{\frac{\iota}{n}}. \qquad (30)
\end{aligned}$$

Using this in (29) and recalling $(\mathcal{T}f)(s,a)=R(s,a)+\mathbb{E}_{s'\sim P(\cdot|s,a)}[f(s')]$ yields

$$\begin{aligned}
&\mathbb{E}_{(s,a)\sim\mu}\left[\hat{f}(s)-R(s,a)-\mathbb{E}_{s'\sim P(\cdot|s,a)}[\hat{f}(s')]\right]+\mathbb{E}_{(s,a)\sim\mathcal{D}}\left[-\varepsilon(s,a;\hat{f})+\lambda\varepsilon(s,a;\hat{f})^2\right] \\
&\quad\le \mathbb{E}_{(s,a)\sim\mu}\left[V^\star(s)-R(s,a)-\mathbb{E}_{s'\sim P(\cdot|s,a)}[V^\star(s')]\right]+O\left(\sqrt{\frac{\iota}{n}}+\frac{\lambda\iota}{n}\right)
\end{aligned}$$

which rearranges to

$$\hat{f}(s_1)+\mathbb{E}_{(s,a)\sim\mathcal{D}}\left[-\varepsilon(s,a;\hat{f})+\lambda\varepsilon(s,a;\hat{f})^2\right]\le V^\star(s_1)+O\left(\sqrt{\frac{\iota}{n}}+\frac{\lambda\iota}{n}\right)$$

using the assumption that μ is admissible. Finally, using AM-GM to bound

$$\begin{aligned}
\varepsilon(s,a;\hat{f})-\lambda\varepsilon(s,a;\hat{f})^2&=\varepsilon(s,a;\hat{f})-\frac{\lambda}{2}\varepsilon(s,a;\hat{f})^2-\frac{\lambda}{2}\varepsilon(s,a;\hat{f})^2 \\
&\le \frac{1}{2\lambda}-\frac{\lambda}{2}\varepsilon(s,a;\hat{f})^2
\end{aligned}$$

finishes the proof.

□

Proof of Theorem 23. First, observe that

$$\begin{aligned}
&\mathbb{E}^{\pi^\star,M^\star}\left[\hat{f}(s,a)-R(s,a)-\mathbb{E}_{s'\sim P(\cdot|s,a)}[\hat{f}(s')]\right] \\
&=\mathbb{E}^{\pi^\star,M^\star}\left[\hat{f}(s,a)-\psi(\pi^\star;s)-\hat{f}(s)\right]+\hat{f}(s_1)-J(\pi^\star) \qquad \text{(same calculation as (19) and (21))} \\
&\le \mathbb{E}^{\pi^\star,M^\star}\left[\hat{f}(s,a)-\psi(\pi^\star;s)-\hat{f}(s)\right]-\mathbb{E}_{(s,a)\sim\mathcal{D}}\left[\frac{\lambda}{2}\varepsilon(s,a;\hat{f})^2\right]+O\left(\sqrt{\frac{\iota}{n}}+\frac{\lambda\iota}{n}+\frac{1}{\lambda}\right) \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(by Lemma 25)}
\end{aligned}$$

where $\iota=H^2\log(|\mathcal{F}||\mathcal{G}|/\delta)$. This implies

$$\begin{aligned}
&D^{\pi^\star}_{\mathsf{av}}(\hat{f}\|M^\star) \\
&=\left(\mathbb{E}^{\pi^\star,M^\star}\left[\hat{f}(s,a)-R(s,a)-\mathbb{E}_{s'\sim P(\cdot|s,a)}[\hat{f}(s')]\right]\right)^2 \\
&\ge \frac{1}{2}\left(\mathbb{E}^{\pi^\star,M^\star}\left[\hat{f}(s,a)-\psi(\pi^\star;s)-\hat{f}(s)\right]\right)^2+\frac{\lambda^2}{8}\left(\mathbb{E}_{(s,a)\sim\mathcal{D}}\left[\varepsilon(s,a;\hat{f})^2\right]\right)^2-O\left(\frac{\iota}{n}+\frac{\lambda^2\iota^2}{n^2}+\frac{1}{\lambda^2}\right)
\end{aligned}$$

where we use that $(X+Y+Z)^2\ge\frac{1}{2}(X+Y)^2-[Z]_+^2\ge\frac{1}{2}X^2+\frac{1}{2}Y^2-[Z]_+^2$ if $X<0$, $Y<0$. Thus,

$$\begin{aligned}
&J(\pi^\star)-J(\pi_{\hat{f}}) \\
&=J(\pi^\star)-J(\pi_{\hat{f}})-\gamma D^{\pi^\star}_{\mathsf{av}}(\hat{f}\|M^\star)+\gamma D^{\pi^\star}_{\mathsf{av}}(\hat{f}\|M^\star) \qquad \text{(for any } \gamma) \\
&\le \underbrace{J(\pi^\star)-J(\pi_{\hat{f}})-\frac{\gamma}{2}\left(\mathbb{E}^{\pi^\star,M^\star}\left[\hat{f}(s,a)-\psi(\pi^\star;s)-\hat{f}(s)\right]\right)^2}_{\textbf{term}_1}+O\left(\gamma\left(\frac{\iota}{n}+\frac{\lambda^2\iota^2}{n^2}+\frac{1}{\lambda^2}\right)\right) \\
&\quad\underbrace{-\frac{\gamma\lambda^2}{8}\left(\mathbb{E}_{(s,a)\sim\mathcal{D}}\left[\varepsilon(s,a;\hat{f})^2\right]\right)^2+\gamma D^{\pi^\star}_{\mathsf{av}}(\hat{f}\|M^\star)}_{\textbf{term}_2}.
\end{aligned}$$

$\mathbf{term}_1$ is bounded as before in regularized MDP

$$\begin{aligned}\mathbf{term}_1 &\le \frac{1}{2\gamma}\mathsf{ER}(\hat{f};\mathcal{F}_{\mathrm{conf}})^2 && \text{(by AM-GM and the definition of ER)}\\ &\le \frac{1}{2\gamma}\left(3\left(1+\frac{4H}{\alpha}\right)\left(1+H^3\cdot\frac{1}{\alpha}\right)\right)^2 && \text{(by Equation (23))}\\ &\lesssim \frac{H^8}{\gamma\alpha^4}.\end{aligned}$$

Below, we bound $\mathbf{term}_2$. By definition,

$$\begin{aligned}\mathbf{term}_2 &= -\frac{\gamma\lambda^2}{8}\left(\mathbb{E}_{(s,a)\sim\mathcal{D}}\left[\varepsilon(s,a;\hat{f})^2\right]\right)^2 + \gamma D^{\pi^\star}_{\mathsf{av}}(\hat{f}\|M^\star)\\ &= -\frac{\gamma\lambda^2}{8}\left(\mathbb{E}_{(s,a)\sim\mathcal{D}}\left[(\hat{f}(s,a)-(\hat{\mathcal{T}}\hat{f})(s,a))^2\right]\right)^2 + \gamma\left(\mathbb{E}_{(s,a)\sim d^{\pi^\star}}\left[\hat{f}(s,a)-(\mathcal{T}\hat{f})(s,a)\right]\right)^2\end{aligned} \tag{31}$$

By Lemma 24 and concentration inequality (Lemma 26),

$$\begin{aligned}&\mathbb{E}_{(s,a)\sim\mathcal{D}}\left[(\hat{f}(s,a)-(\hat{\mathcal{T}}\hat{f})(s,a))^2\right]\\ &\ge \frac{1}{2}\mathbb{E}_{(s,a)\sim\mathcal{D}}\left[(\hat{f}(s,a)-(\mathcal{T}\hat{f})(s,a))^2\right] - O\left(\frac{\iota}{n}\right) && \text{(by } X^2\ge\tfrac{1}{2}Y^2-(X-Y)^2\text{)}\\ &\ge \frac{1}{4}\mathbb{E}_{(s,a)\sim\mu}\left[(\hat{f}(s,a)-(\mathcal{T}\hat{f})(s,a))^2\right] - O\left(\frac{\iota}{n}\right).\end{aligned} \tag{32}$$

By Jensen's inequality and coverage definition,

$$\left(\mathbb{E}_{(s,a)\sim d^{\pi^\star}}\left[\hat{f}(s,a)-(\mathcal{T}\hat{f})(s,a)\right]\right)^2 \le HC^{\pi^\star}\mathbb{E}_{(s,a)\sim\mu}\left[\left(\hat{f}(s,a)-(\mathcal{T}\hat{f})(s,a)\right)^2\right]. \tag{33}$$

Using (32) and (33) in (31) and denoting $Z = \mathbb{E}_{(s,a)\sim\mu}\left[\left(\hat{f}(s,a)-(\mathcal{T}\hat{f})(s,a)\right)^2\right]$, we get

$$\begin{aligned}\mathbf{term}_2 &\le -\frac{\gamma\lambda^2}{8\cdot 4^2\cdot 2}Z^2 + O\left(\frac{\gamma\lambda^2\iota^2}{n^2}\right) + \gamma HC^{\pi^\star}Z \quad \text{(by } (X-Y)_+^2\ge\tfrac{1}{2}X^2-Y^2 \text{ for } X,Y>0\text{)}\\ &\le O\left(\frac{\gamma(HC^{\pi^\star})^2}{\lambda^2}+\frac{\gamma\lambda^2\iota^2}{n^2}\right).\end{aligned}$$

Collecting terms, we get

$$\begin{aligned}J(\pi^\star)-J(\pi_{\hat{f}}) &\lesssim O\left(\gamma\left(\frac{\iota}{n}+\frac{\lambda^2\iota^2}{n^2}+\frac{1}{\lambda^2}\right)+\frac{H^8}{\gamma\alpha^4}+\frac{\gamma(HC^{\pi^\star})^2}{\lambda^2}\right)\\ &= O\left(\frac{H^4}{\alpha^2}\left(\sqrt{\frac{\iota}{n}}+\frac{\lambda\iota}{n}+\frac{HC^{\pi^\star}}{\lambda}\right)\right) && \text{(choosing optimal } \gamma\text{)}\\ &= O\left(\frac{H^4}{\alpha^2}\left(\frac{\lambda\iota}{n}+\frac{HC^{\pi^\star}}{\lambda}\right)\right). && \text{(AM-GM and } 1\le C^{\pi^\star}\text{)}\end{aligned}$$

□

H Concentration Inequalities and Performance Difference Lemma

Lemma 26 (Lemma A.3 in [FKQR21]). *Let $(X_t)_{t\le T}$ be a sequence of random variables adapted to a filtration $(\mathcal{F}_t)_{t\le T}$. If $0\le X_t\le R$ almost surely, then with probability at least $1-\delta$,*

$$\sum_{t=1}^{T} X_t \le \frac{3}{2}\sum_{t=1}^{T}\mathbb{E}_{t-1}[X_t] + 4R\log(2\delta^{-1}),$$

and

$$\sum_{t=1}^{T}\mathbb{E}_{t-1}[X_t] \le 2\sum_{t=1}^{T} X_t + 8R\log(2\delta^{-1}).$$

Lemma 27 (Performance Difference Lemma)**.** *For any two policies π_1, π_2, we have*

$$V^{\pi_1}(s_1) - V^{\pi_2}(s_1) = \mathbb{E}_{s\sim d^{\pi_1}}\left[\sum_{a\in\mathcal{A}} \left(\pi_1(a|s) - \pi_2(a|s)\right) Q^{\pi_2}(s,a)\right],$$

and

$$V^{\pi_1}(s_1) - V^{\pi_2}(s_1) = \mathbb{E}_{s\sim d^{\pi_2}}\left[\sum_{a\in\mathcal{A}} \left(\pi_1(a|s) - \pi_2(a|s)\right) Q^{\pi_1}(s,a)\right].$$

I Auxiliary Lemmas for Regularized MDPs and Proof of Lemma 35

Lemma 28. *Let $p, p' \in \Delta_{\mathcal{A}}$ and $q \in \mathbb{R}^A$. Suppose that $\eta > 0$ is such that $\eta q(a) \le 1$ for all a. Then*

$$\langle p' - p, q\rangle \le \frac{1}{\eta}\mathrm{KL}(p', p) + \eta \sum_a p(a)q(a)^2.$$

Proof. It suffices to show that

$$\max_{p'\in\Delta_A}\left\{\langle p', q\rangle - \frac{1}{\eta}\mathrm{KL}(p', p)\right\} \le \sum_a p(a)q(a) + \eta\sum_a p(a)q(a)^2.$$

The left-hand side is the Fenchel conjugate of KL and has a closed form $\frac{1}{\eta}\log\left(\sum_a p(a)\exp(\eta q(a))\right)$. Using the condition $\eta q(a) \le 1$, we bound this by

$$\begin{aligned}
&\frac{1}{\eta}\log\left(\sum_a p(a)\left(1 + \eta q(a) + \eta^2 q(a)^2\right)\right) && (e^x \le 1 + x + x^2 \text{ for } x \le 1)\\
&= \frac{1}{\eta}\log\left(1 + \eta\sum_a p(a)q(a) + \eta^2\sum_a p(a)q(a)^2\right)\\
&\le \sum_a p(a)q(a) + \eta\sum_a p(a)q(a)^2. && (\log(1+x) \le x)
\end{aligned}$$

□

Lemma 29. *Let $\Delta_{\mathcal{A}}$ be the simplex over $\mathcal{A}$, $\psi : \Delta_{\mathcal{A}} \to \mathbb{R}$ be a convex function and let $q \in \mathbb{R}^{\mathcal{A}}$. Define $G(p) = \langle p, q\rangle - \psi(p)$ for $p \in \Delta_{\mathcal{A}}$. Let $p^\star = \mathrm{argmax}_{p\in\Delta_A} G(p)$. If $p^\star(a) > 0$ for all a, then for any $p \in \Delta_A$,*

$$G(p^\star) - G(p) = \mathrm{Breg}_\psi(p, p^\star)$$

where $\mathrm{Breg}_\psi(x, y) = \psi(x) - \psi(y) - \langle\nabla\psi(y), x - y\rangle$ is the Bregman divergence w.r.t. ψ.

Proof. We maximize the concave objective $G(p) = \langle p, q\rangle - \psi(p)$ over the simplex Δ_A such that $\sum_{a\in\mathcal{A}} p(a) = 1$ and $p(a) \ge 0$, $\forall a \in \mathcal{A}$. Introduce $\lambda \in \mathbb{R}$ and $\mu = (\mu_a)_{a\in A} \in \mathbb{R}_+^A$ as Lagrange multipliers. The Lagrangian is

$$\mathcal{L}(p, \lambda, \mu) = \langle p, q\rangle - \psi(p) + \lambda(1 - \mathbf{1}^\top p) + \sum_{a\in A}\mu_a\, p(a),$$

Since G is concave and the feasible set is convex, the KKT conditions are necessary and sufficient for optimality. In particular, at an optimal solution $(p^\star, \lambda^\star, \mu^\star)$ we have:

$$\nabla_p\mathcal{L}(p^\star, \lambda^\star, \mu^\star) = q - \nabla\psi(p^\star) - \lambda^\star\mathbf{1} + \mu^\star = 0 \tag{34}$$
$$\mu_a^\star\, p^\star(a) = 0, \quad \forall a \in \mathcal{A} \tag{35}$$
$$p^\star \in \Delta_{\mathcal{A}}, \mu^\star \ge 0 \tag{36}$$

With the assumption $p^\star(a) > 0$ for all a, and Equation (35), we have

$$\mu_a^\star = 0 \quad \forall a \in \mathcal{A}$$

Plugging into Equation (34) yields

$$q - \nabla\psi(p^\star) = \lambda^\star \mathbf{1}. \tag{37}$$

For any $p \in \Delta_\mathcal{A}$, we have

$$\begin{aligned}
G(p^\star) - G(p) &= \langle p^\star, q\rangle - \psi(p^\star) - (\langle p, q\rangle - \psi(p)) \\
&= \langle p^\star - p, q\rangle - \psi(p^\star) + \psi(p) \\
&= (\psi(p) - \psi(p^\star) - \langle \nabla\psi(p^\star),\, p - p^\star\rangle) + \langle p^\star - p,\, q - \nabla\psi(p^\star)\rangle \\
&= \mathrm{Breg}_\psi(p, p^\star) + \langle p^\star - p,\, q - \nabla\psi(p^\star)\rangle \\
&= \mathrm{Breg}_\psi(p, p^\star) + \langle p^\star - p,\, \lambda^\star \mathbf{1}\rangle && \text{(Equation (37))} \\
&= \mathrm{Breg}_\psi(p, p^\star) && (p^\star, p \in \Delta_\mathcal{A})
\end{aligned}$$

□

Definition 7 (Legendre Functions (Section 26.4 of [LS20]))**.** *Let f be a convex function, and let $C = ri(\mathrm{dom} f) \neq \emptyset$, where $\mathrm{dom} f$ is the domain of f and ri is relative interior. We call f* Legendre *if*

(i) f is differentiable and strictly convex on C.

(ii) $\|\nabla f(x_n)\| \to \infty$ for any sequence $(x_n)_n$ with $x_n \in C$ for all n and $\lim_{n\to\infty} x_n \to x \in \partial C$, where ∂C is the subgradient of C.

Lemma 30. *For any s, $\psi(p; s) = \alpha \mathrm{Breg}_\Phi(p, \pi_{\mathrm{ref}}(\cdot|s))$ is a Legendre function of p when $\Phi(x) = \sum_{a\in\mathcal{A}} x(a)\log(x(a))$ (negative entropy), or $\Phi(x) = -\sum_{a\in\mathcal{A}} \log(x(a))$ (log barrier), or $\Phi(x) = \frac{\alpha}{1-q}\left(1 - \sum_{a\in\mathcal{A}} x(a)^q\right)$ (Tsallis entropy) with $q \in (0, 1)$.*

Proof. When Φ is negative entropy, then on $\mathrm{ri}(\Delta_\mathcal{A})$, for any s, $[\nabla\psi(p; s)]_a = 1 + \log p(a) - \log \pi_{\mathrm{ref}}(a|s)$ and $\nabla^2\psi(p; s) = \mathrm{diag}(1/p(a)) \succ 0$, so ψ is differentiable and strictly convex on $\mathrm{ri}(\Delta_\mathcal{A})$. If $p_n \to \partial\Delta_\mathcal{A}$, some coordinate $p_n(a_0) \downarrow 0$ and $|1+\log p_n(a_0)| \to \infty$, hence $\|\nabla\psi(p_n)\| \to \infty$. Thus, ψ is Legendre from Definition 7.

When Φ is log barrier, then on $\mathrm{ri}(\Delta_\mathcal{A})$, $[\nabla\psi(p; s)]_a = -\frac{1}{p(a)} + \frac{1}{\pi_{\mathrm{ref}}(a|s)}$ and $\nabla^2\psi(p; s) = \mathrm{diag}(1/p(a)^2) \succ 0$, so ψ is differentiable and strictly convex. If $p_n \to \partial\Delta_\mathcal{A}$ then some $p_n(a_0) \downarrow 0$ and $|1/p_n(a_0)| \to \infty$, hence $\|\nabla\psi(p_n)\| \to \infty$. Thus, ψ is Legendre from Definition 7.

When Φ is Tsallis entropy with $q \in (0, 1)$, then on $\mathrm{ri}(\Delta_\mathcal{A})$, we have

$$[\nabla\psi(p; s)_a] = -\frac{\alpha q}{1-q}\, p(a)^{q-1} + \frac{\alpha q}{1-q}\, \pi_{\mathrm{ref}}(a)^{q-1}, \qquad \nabla^2\psi(p; s) = \alpha q\, \mathrm{diag}\left(p(a)^{q-2}\right) \succ 0,$$

so ψ is differentiable and strictly convex. If $p_n \to \partial\Delta_\mathcal{A}$ then some $p_n(a_0) \downarrow 0$ and since $q - 1 < 0$, $|p_n(a_0)^{q-1}| \to \infty$, hence $\|\nabla\psi(p_n)\| \to \infty$. Thus, ψ is Legendre from Definition 7. □

Lemma 31. *Define $G(p) = \langle p, q\rangle - \psi(p)$ where $p, q \in \Delta_\mathcal{A}$ and $\mathcal{A}$ is an arbitrary finite set. Let $p^\star = \mathrm{argmax}_{p\in\Delta_A} G(p)$. If ψ is a Legendre mirror map, then $p^\star(a) > 0, \forall a \in \mathcal{A}$.*

Proof. Since $|\mathcal{A}| < \infty$ and ψ is Legendre on $C = \mathrm{ri}(\Delta_\mathcal{A})$. The maximizer of $G(p) = \langle p, q\rangle - \psi(p)$ over $\Delta_\mathcal{A}$ is equivalently the minimizer of $f(p) = \psi(p) - \langle p, q\rangle$ over $\Delta_\mathcal{A}$. Since adding a linear function preserves differentiability, strict convexity on C, and the blow-up condition $\|\nabla f(p_n)\| = \|\nabla\psi(p_n) - q\| \to \infty$ as $p_n \to \partial\Delta_\mathcal{A}$, the function f is also Legendre on C from the definition in Definition 7. By Corollary 26.8 in [LS20] (applied on the affine hull of $\Delta_\mathcal{A}$ with int interpreted as ri), any minimizer $p^\star$ of f over $\Delta_\mathcal{A}$ must lie in $\mathrm{ri}(\Delta_\mathcal{A})$. Therefore $p^\star(a) > 0$ for all $a \in \mathcal{A}$. □

Lemma 32 (Boundedness of $Q^\star$)**.** *If $\psi(p; s) = \alpha \mathrm{Breg}_\Phi(p, \pi_{\mathrm{ref}}(\cdot|s))$ for any convex function Φ, then $0 \leq Q^\star(s, a) \leq H, \forall s, a$. Moreover, for any $f \in \mathcal{F}$ and any s, a, we have $Q^\star(s, a) - Q^{\pi_f}(s, a) \leq H^2$.*

Proof. We have $\psi(p;s) \geq 0$ and the minimizer is achieved with $p = \pi_{\text{ref}}(\cdot|s)$. We can show that for any $s \in \mathcal{S}_h$,

$$0 \;\leq\; V^\star(s) \;\leq\; H - h + 1, \qquad 0 \;\leq\; Q^\star(s,a) \;\leq\; H - h + 1. \tag{38}$$

We will also prove this by induction. For $s \in \mathcal{S}_H$, we have $V^\star(s) = 0$ and $Q^\star(s,a) = R(s,a) \in [0,1]$. If $0 \leq V^\star(s') \leq H - h$ for $s' \in \mathcal{S}_{h+1}$, then for any $s \in \mathcal{S}_h$, we have

$$0 \leq Q^\star(s,a) = R(s,a) + \mathbb{E}_{s' \sim P(\cdot|s,a)}[V^\star(s')] \leq 1 + (H-h) = H - h + 1.$$

Moreover,

$$V^\star(s) = \max_{p \in \Delta_{\mathcal{A}}} \{\mathbb{E}_{a\sim p}[Q^\star(s,a)] - \psi(p;s)\} \leq \max_{p \in \Delta_{\mathcal{A}}} \mathbb{E}_{a\sim p}[Q^\star(s,a)] \leq H - h + 1,$$

and $V^\star(s) \geq 0$ because

$$V^\star(s) \geq \mathbb{E}_{a\sim\pi_{\text{ref}}(\cdot|s)}[Q^\star(s,a)] - \psi(\pi_{\text{ref}}(\cdot|s);s) = \mathbb{E}_{a\sim\pi_{\text{ref}}(\cdot|s)}[Q^\star(s,a)] \geq 0$$

For any $f \in \mathcal{F}$, we have

$$\mathbb{E}_{a\sim\pi_f(\cdot|s)}[f(s,a)] - \psi(\pi_f(\cdot|s);s) \geq \mathbb{E}_{a\sim\pi_{\text{ref}}(\cdot|s)}[f(s,a)] - \psi(\pi_{\text{ref}}(\cdot|s);s) = \mathbb{E}_{a\sim\pi_{\text{ref}}(\cdot|s)}[f(s,a)]$$

Thus, for any s, we have

$$\psi(\pi_f(\cdot|s);s) \leq \mathbb{E}_{a\sim\pi_f(\cdot|s)}[f(s,a)] - \mathbb{E}_{a\sim\pi_{\text{ref}}(\cdot|s)}[f(s,a)] \leq H$$

where we use $f \in [0,H]$ from the bound of $Q^\star$ in (38). $f \in \mathcal{F}$ that is not in $[0,H]$ can be removed because it is not $Q^\star$. We can show that

$$V^{\pi_f}(s) \;\geq\; -H(H-h+1), \qquad Q^{\pi_f}(s,a) \;\geq\; -H(H-h). \tag{39}$$

For $h = H$, since $Q^{\pi_f}(s,a) = R(s,a) \geq 0$, the second inequality holds. Moreover,

$$V^{\pi_f}(s) = \mathbb{E}_{a\sim\pi_f(\cdot|s)}[R(s,a)] - \psi(\pi_f(\cdot|s);s) \geq 0 - H = -(H - H + 1)H,$$

proving the base case. Assume (39) holds for stage $h+1$. Then for any $s' \in \mathcal{S}_{h+1}$,

$$V^{\pi_f}(s') \geq -(H - (h+1) + 1)H = -(H-h)H.$$

Hence for any $s \in \mathcal{S}_h, a \in \mathcal{A}$,

$$Q^{\pi_f}(s,a) = R(s,a) + \mathbb{E}_{s'\sim P(\cdot|s,a)}[V^{\pi_f}(s')] \geq 0 - (H-h)H,$$

which proves the second inequality at stage h. Finally,

$$V^{\pi_f}(s) = \mathbb{E}_{a\sim\pi_f(\cdot|s)}[Q^{\pi_f}(s,a)] - \psi(\pi_f(\cdot|s);s) \geq -(H-h)H - H = -(H-h+1)H,$$

proving the first inequality at stage h. This completes the induction. Thus, for any $s \in \mathcal{S}_h$ and any $a \in \mathcal{A}$, using the upper bound in (38) and lower bound in (39), we have

$$Q^\star(s,a) - Q^{\pi_f}(s,a) \leq H - h + 1 + H(H-h) \leq H^2$$

□

Lemma 33 (Solution for Log-barrier Regularizer)**.** *For any $\alpha > 0$ and any $f \in \mathcal{F}$, if*

$$\pi_f(\cdot|s) \in \operatorname*{argmax}_{p \in \Delta_{\mathcal{A}}} \Big\{ \sum_{a\in\mathcal{A}} p(a) f(s,a) - \alpha \text{Breg}_\Phi(p, \pi_{\text{ref}}(\cdot|s)) \Big\},$$

where $\Phi(p) = -\sum_{a\in\mathcal{A}} \log p(a)$. Then we have

$$\frac{\alpha}{\alpha + 2H} \;\leq\; \frac{\pi_{f_1}(a|s)}{\pi_{f_2}(a|s)} \;\leq\; \frac{\alpha + 2H}{\alpha}, \quad \forall f_1, f_2 \in \mathcal{F}.$$

Proof. From Lemma 30 and Lemma 31, we known $\pi_f(\cdot|s)$ is unique for any s with $\pi_f(a|s) > 0$ for all a. since the maximizer is interior, only the equality constraint $\sum_a p(a) = 1$ is active. Consider the Lagrangian

$$\mathcal{L}_{f,s}(p,\lambda) = \sum_a p(a) f(s,a) - \alpha \text{Breg}_\Phi(p, \pi_{\text{ref}}(\cdot|s)) + \lambda_f(s) \left(\sum_a p(a) - 1 \right).$$

By KKT condition, the stationarity condition at $p = \pi_f(\cdot|s)$ gives, for every a,

$$\frac{\partial \mathcal{L}_{f,s}}{\partial p(a)} = f(s,a) + \frac{\alpha}{\pi_f(a|s)} - \frac{\alpha}{\pi_{\text{ref}}(a|s)} + \lambda_f(s) = 0,$$

hence

$$\pi_f(a|s) = \frac{\pi_{\text{ref}}(a|s)}{1 - \frac{\pi_{\text{ref}}(a|s)}{\alpha}\left(\lambda_f(s) + f(s,a)\right)}.$$

We have $\pi_f(a|s) > 0$ and normalization implies

$$\sum_{a\in\mathcal{A}} \pi_f(a|s) = \sum_{a\in\mathcal{A}} \frac{\pi_{\text{ref}}(a|s)}{1 - \frac{\pi_{\text{ref}}(a|s)}{\alpha}\left(\lambda_f(s) + f(s,a)\right)} = 1 \tag{40}$$

Define function $g_f(\lambda; s, a) = 1 - \frac{\pi_{\text{ref}}(a|s)}{\alpha}\left(\lambda + f(s,a)\right)$. For any $f \in \mathcal{F}$ and any s, a, we have $g_f(0; s, a) = 1 - \frac{\pi_{\text{ref}}(a|s)}{\alpha}\left(f(s,a)\right) \le 1$ and $g_f(-H; s, a) = 1 - \frac{\pi_{\text{ref}}(a|s)}{\alpha}\left(-H + f(s,a)\right) \ge 1$ where we use $f \in [0, H]$. Since $g_f(\lambda; s, a)$ is decreasing for λ and $\sum_{a\in\mathcal{A}} \pi_{\text{ref}}(a|s) = 1$, from (40), we have $\lambda_f(s) \in [-H, 0]$. Thus, $\lambda_f(s) + f(s,a) \in [-H, H]$, and we have

$$\begin{aligned}
\frac{\pi_{f_1}(a|s)}{\pi_{f_2}(a|s)} &= \pi_{f_1}(a|s)\left(\frac{1}{\pi_{f_2}(a|s)} - \frac{1}{\pi_{f_1}(a|s)}\right) + 1 \\
&= \frac{\pi_{f_1}(a|s)}{\alpha}\left(\lambda_{f_1}(s) + f_1(s,a) - \left(\lambda_{f_2}(s) + f_2(s,a)\right)\right) + 1 \\
&\le 1 + \frac{2H}{\alpha}.
\end{aligned}$$

□

Lemma 34 (Solution for Tsallis-entropy Regularizer)**.** *For any $\alpha > 0$ and any $f \in \mathcal{F}$, if*

$$\pi_f(\cdot|s) \in \operatorname*{argmax}_{p\in\Delta_{\mathcal{A}}} \Big\{ \sum_{a\in\mathcal{A}} p(a) f(s,a) - \alpha \text{Breg}_{\Phi}(p, \pi_{\text{ref}}(\cdot|s)) \Big\},$$

where $\Phi(p) = \frac{1}{1-q}\left(1 - \sum_{a\in\mathcal{A}} p(a)^q\right)$. Then we have

$$\left(1 + \frac{2H(1-q)}{\alpha q}\right)^{\frac{1}{q-1}} \le \frac{\pi_{f_1}(a|s)}{\pi_{f_2}(a|s)} \le \left(1 + \frac{2H(1-q)}{\alpha q}\right)^{\frac{1}{1-q}}, \quad \forall f_1, f_2 \in \mathcal{F}.$$

Proof. From Lemma 30 and Lemma 31, we known $\pi_f(\cdot|s)$ is unique for any s with $\pi_f(a|s) > 0$ for all a. Since the maximizer is interior, only the equality constraint $\sum_{a\in\mathcal{A}} p(a) = 1$ is active. The Lagrangian is

$$\mathcal{L}_{f,s}(p, \lambda) = \sum_{a\in\mathcal{A}} p(a) f(s,a) - \alpha \text{Breg}_{\Phi}(p, \pi_{\text{ref}}(\cdot|s)) + \lambda_f(s)\left(\sum_{a\in\mathcal{A}} p(a) - 1\right).$$

By KKT condition, the stationarity condition at $p = \pi_f(\cdot|s)$, for every a,

$$\frac{\partial \mathcal{L}_{f,s}}{\partial p(a)} = f(s,a) + \frac{\alpha}{1-q} \cdot q\, \pi_f(a|s)^{q-1} - \frac{\alpha}{1-q} \cdot q\, \pi_{\text{ref}}(a|s)^{q-1} + \lambda_f(s) = 0.$$

This implies

$$\pi_f(a|s) = \pi_{\text{ref}}(a|s)\left(1 - \frac{1-q}{\alpha q} \cdot \frac{\lambda_f(s) + f(s,a)}{\pi_{\text{ref}}(a|s)^{q-1}}\right)^{\frac{1}{q-1}}$$

We have $\pi_f(a|s) > 0$ and normalization ensures

$$\sum_{a\in\mathcal{A}} \pi_f(a|s) = \sum_{a\in\mathcal{A}} \pi_{\text{ref}}(a|s)\left(1 - \frac{1-q}{\alpha q} \cdot \frac{\lambda_f(s) + f(s,a)}{\pi_{\text{ref}}(a|s)^{q-1}}\right)^{\frac{1}{q-1}} = 1 \tag{41}$$

Define function $g_f(\lambda; s, a) = \left(1 - \frac{1-q}{\alpha q} \frac{\lambda + f(s,a)}{\pi_{\text{ref}}(a|s)^{q-1}}\right)^{\frac{1}{q-1}}$. We have $g_f(0; s, a) \geq 1$ and $g_f(-H; s, a) \leq 1$ where we use $f \in [0, H]$, $q \in (0, 1)$. Since $g_f(\lambda; s, a)$ is increasing in λ, from $\sum_{a \in \mathcal{A}} \pi_{\text{ref}}(a|s) = 1$ and (41), we have $\lambda_f(s) \in [-H, 0]$. Thus, $\lambda_f(s) + f(s, a) \in [-H, H]$. Thus

$$\begin{aligned}
\frac{\pi_{f_1}(a|s)}{\pi_{f_2}(a|s)} &= \left(\frac{1}{\pi_{f_2}(a|s)^{q-1}} \left(\pi_{f_1}(a|s)^{q-1} - \pi_{f_2}(a|s)^{q-1}\right) + 1\right)^{\frac{1}{q-1}} \\
&= \left(\frac{1}{\pi_{f_2}(a|s)^{q-1}} \cdot \frac{1-q}{\alpha q} \cdot \left(f_2(s,a) + \lambda_{f_2}(s,a) - \left(f_1(s,a) + \lambda_{f_1}(s,a)\right)\right) + 1\right)^{\frac{1}{q-1}} \\
&\geq \left(1 + \frac{2H(1-q)}{\alpha q}\right)^{\frac{1}{q-1}}.
\end{aligned}$$

□

Lemma 35. *Let $\psi(p; s) = \alpha \text{Breg}_\Phi(p, \pi_{\text{ref}}(\cdot|s))$ for some reference policy π_{ref}. The following choices of Φ make Assumption 2 hold with respective (C_1^ψ, C_2^ψ):*

- $\Phi(p) = \sum_{a \in \mathcal{A}} p(a) \log(p(a))$ (Shannon entropy): $(C_1^\psi, C_2^\psi) = (1 + \frac{4H}{\alpha}, \frac{1}{\alpha})$.
- $\Phi(p) = \frac{1}{1-q}\left(1 - \sum_a p(a)^q\right)$ *with* $q \in (0, 1)$ (Tsallis entropy): $(C_1^\psi, C_2^\psi) = \left(\left(1 + \frac{2H(1-q)}{\alpha q}\right)^{\frac{2-q}{1-q}}, \frac{1}{\alpha q}\right)$.
- $\Phi(p) = -\sum_{a \in \mathcal{A}} \log\left(p(a)\right)$ (log-barrier): $(C_1^\psi, C_2^\psi) = (1 + \frac{2H}{\alpha}, \frac{2}{\alpha})$.

Proof of Lemma 35. From Lemma 30, $\psi(p; s) = \alpha \text{Breg}_\Phi(p, \pi_{\text{ref}}(\cdot|s))$ is Legendre when Φ is Shannon entropy, log-barrier or Tsallis entropy with $q \in (0, 1)$. We will prove the concrete C_ψ^1 and C_ψ^2 factors for these three regularizers now.

When Φ is (negative) Shannon entropy. Since Φ is entropy, $\psi(p; s) = \alpha \text{KL}\left(p, \pi_{\text{ref}}(\cdot|s)\right)$. For any $x, y \in \Delta_{\mathcal{A}}$ and $s \in \mathcal{S}$, $\text{Breg}_\psi\left(x, y; s\right) = \alpha \text{KL}\left(x \| y\right)$ and define $r(a) = \frac{y(a)}{x(a)}$ for all $a \in \mathcal{A}$. Define the nonnegative convex functions $\phi(t) = t \log t - t + 1, \xi(t) = t - 1 - \log t$ for $t > 0$. We have $\text{KL}(y\|x) = \sum_a y(a) \log \frac{y(a)}{x(a)} = \mathbb{E}_{a \sim x}\left[r(a) \log r(a)\right] = \mathbb{E}_x\left[\phi(r)\right]$ and $\text{KL}(x\|y) = \mathbb{E}_x[-\log r] = \mathbb{E}_x[-\log r + r - 1] = \mathbb{E}_x[\xi(r)]$. From Lemma 36, we have

$$\phi(t) \leq \xi(t) \quad \text{for } t \in (0, 1], \qquad \phi(t) \leq \left(1 + \log t\right)\xi(t) \quad \text{for } t \geq 1. \tag{42}$$

Assume there exists $R \geq 1$ such that $R^{-1} \leq r(a) \leq R, \forall a \in \mathcal{A}.$, we have

$$\frac{\phi(r(a))}{\xi(r(a))} \leq \sup_{t \in [R^{-1}, R]} \frac{\phi(t)}{\xi(t)} \leq 1 + \log R \quad \text{by (42).}$$

Multiplying by $\xi(r(a)) \geq 0$ and taking $\mathbb{E}_{a \sim x}$ gives

$$\text{KL}(y\|x) = \mathbb{E}_x[\phi(r)] \leq (1 + \log R)\mathbb{E}_x[\xi(r)] = (1 + \log R)\text{KL}(x\|y). \tag{43}$$

Now we calculate the exact value of R. For any $f_1, f_2 \in \mathcal{F}$, we have

$$\pi_{f_1}(a|s) = \frac{\pi_{\text{ref}}(a|s) e^{\frac{1}{\alpha} f_1(s,a)}}{\sum_{a' \in \mathcal{A}} \pi_{\text{ref}}(a'|s) e^{\frac{1}{\alpha} f_1(s,a')}}, \qquad \pi_{f_2}(a|s) = \frac{\pi_{\text{ref}}(a|s) e^{\frac{1}{\alpha} f_2(s,a)}}{\sum_{a' \in \mathcal{A}} \pi_{\text{ref}}(a'|s) e^{\frac{1}{\alpha} f_2(s,a')}},$$

with $|f_1(s,a)| \leq H$ and $|f_2(s,a)| \leq H$ for all s, a from Lemma 32. Let $Z_1 = \sum_{a' \in \mathcal{A}} \pi_{\text{ref}}(a'|s) e^{\frac{1}{\alpha} f_1(s,a')}$ and $Z_2 = \sum_{a' \in \mathcal{A}} \pi_{\text{ref}}(a'|s) e^{\frac{1}{\alpha} f_2(s,a')}$. We have $e^{-\frac{H}{\alpha}} \leq Z_1 \leq e^{\frac{H}{\alpha}}$, and similarly $e^{-\frac{H}{\alpha}} \leq Z_2 \leq e^{\frac{H}{\alpha}}$. Hence $e^{-\frac{2H}{\alpha}} \leq Z_1/Z_2 \leq e^{\frac{2H}{\alpha}}$. Moreover $|f_1(s,a) - f_2(s,a)| \leq 2H$, so $e^{-\frac{2H}{\alpha}} \leq e^{\frac{1}{\alpha}(f_1(s,a) - f_2(s,a))} \leq e^{\frac{2H}{\alpha}}$. This implies

$$\frac{\pi_{f_1}(a|s)}{\pi_{f_2}(a|s)} = \frac{Z_{f_2}}{Z_{f_1}} e^{\frac{1}{\alpha}(f_1(s,a) - f_2(s,a))} \in [e^{\frac{-4H}{\alpha}}, e^{\frac{4H}{\alpha}}],$$

for all s, a. Thus, $R = e^{\frac{4H}{\alpha}}$ and from (43), we have

$$\begin{aligned}\text{Breg}_\psi(\pi_{f_1}, \pi_{f_2}; s) &= \alpha \text{KL}\left(\pi_{f_1}(\cdot|s) \| \pi_{f_2}(\cdot|s)\right) \\ &\geq \frac{\alpha}{1 + \log(e^{4H/\alpha})} \text{KL}\left(\pi_{f_2}(\cdot|s) \| \pi_{f_1}(\cdot|s)\right) \\ &= \frac{\alpha}{\alpha + 4H} \text{Breg}_\psi(\pi_{f_2}, \pi_{f_1}; s).\end{aligned}$$

Thus, we have $C_1^\psi = 1 + \frac{4H}{\alpha}$ and $C_2^\psi = \frac{1}{\alpha}$ because $\text{Breg}_\psi(x, y; s) = \alpha \text{KL}(x \| y)$ for any s.

When Φ is log-barrier. We first compute the value of C_ψ^1. To start, define

$$g(t) = \frac{t - 1 - \log(t)}{\frac{1}{t} - 1 + \log(t)}.$$

Lemma 37 shows that $g(t)$ is increasing on $t > 0$. Thus, for any $R > 1$ the minimum of $g(t)$ is attained at $t = 1/R$ on $[1/R, R]$, giving

$$g(t) \geq \min_{t \in [R^{-1}, R]} g(t) = \frac{\log R + \frac{1}{R} - 1}{R - 1 - \log R} > 0, \ \forall t \in [R^{-1}, R]. \tag{44}$$

For any s and any $f_1, f_2 \in \mathcal{F}$, let $r(a|s) = \frac{\pi_{f_1}(a|s)}{\pi_{f_2}(a|s)}$, from Lemma 33, we have $r(a|s) \in [R_{\log}^{-1}, R_{\log}]$ where $R_{\log} = \frac{\alpha + 2H}{\alpha}$. When $\psi(p; s) = \alpha \text{Breg}_\Phi(p, \pi_{\text{ref}}(\cdot|s))$ and Φ is log-barrier, we have

$$\begin{aligned}\text{Breg}_\psi\left(\pi_{f_1}, \pi_{f_2}; s\right) &= \alpha \sum_{a \in \mathcal{A}} \left(r(a|s) - 1 - \log\left(r(a|s)\right)\right) \\ &= \alpha \sum_{a \in \mathcal{A}} g(r(a|s)) \left(\frac{1}{r(a|s)} - 1 + \log\left(r(a|s)\right)\right) \\ &\geq c_{\log} \alpha \sum_{a \in \mathcal{A}} \left(\frac{1}{r(a|s)} - 1 + \log\left(r(a|s)\right)\right) \\ &= c_{\log} \text{Breg}_\psi\left(\pi_{f_2}, \pi_{f_1}; s\right)\end{aligned}$$

where by (44) we define $c_{\log} = \frac{\log(R_{\log}) + \frac{1}{R_{\log}} - 1}{R_{\log} - 1 - \log(R_{\log})} \geq \frac{1}{R_{\log}}$. Thus, $C_1^\psi = \frac{\alpha + 2H}{\alpha}$.

From Lemma 39, we have $C_2^\psi = \frac{2}{\alpha}$.

When Φ is (negative) Tsallis entropy with $q \in (0, 1)$. We first compute the value of C_ψ^1. Define the scalar function (for $t > 0$)

$$\varphi_q(t) = \frac{1 + q(t-1) - t^q}{1 - q} \ \geq \ 0, \qquad g(t) = \frac{\varphi_q(t)}{t^q \, \varphi_q(1/t)}.$$

From Lemma 38, we have

$$g(t) \ \geq \ \begin{cases} t^{q-2}, & t \geq 1, \\ t^{2-q}, & t \leq 1. \end{cases}$$

Therefore, for any $R > 1$ and any $t \in [R^{-1}, R]$,

$$g(t) \ \geq \ \min_{t \in [R^{-1}, R]} g(t) \ \geq \ R^{-(2-q)} \ > \ 0, \qquad \forall t \in [R^{-1}, R]. \tag{45}$$

For any s and any $f_1, f_2 \in \mathcal{F}$, let $r(a|s) = \frac{\pi_{f_1}(a|s)}{\pi_{f_2}(a|s)}$. From Lemma 34, we have $r(a|s) \in [R_{\text{tsa}}^{-1}, R_{\text{tsa}}]$, where $R_{\text{tsa}} = \left(1 + \frac{2H(1-q)}{\alpha q}\right)^{\frac{1}{1-q}}$. When $\psi(p; s) = \alpha \text{Breg}_\Phi(p, \pi_{\text{ref}}(\cdot|s))$ and Φ is Tsallis entropy, we have

$$\text{Breg}_\psi(p_1, p_2; s) = \alpha \sum_{a \in \mathcal{A}} p_2(a)^q \, \varphi_q\left(\frac{p_1(a)}{p_2(a)}\right), \qquad p_1, p_2 \in \text{ri}(\Delta_{\mathcal{A}}), \forall s \in \mathcal{S}$$

Hence,

$$\begin{aligned}\mathrm{Breg}_{\psi}(\pi_{f_1},\pi_{f_2};s) &= \alpha\sum_{a\in\mathcal{A}}\pi_{f_2}(a|s)^q\,\varphi_q(r(a|s))\\ &\geq \alpha\,c_{\text{tsa}}\sum_{a\in\mathcal{A}}\pi_{f_2}(a|s)^q\,r(a|s)^q\,\varphi_q(1/r(a|s))\\ &= c_{\text{tsa}}\,\mathrm{Breg}_{\psi}(\pi_{f_2},\pi_{f_1};s)\,,\end{aligned}$$

where we used (45) with $R = R_{\text{tsa}}$ and $c_{\text{tsa}} := R_{\text{tsa}}^{-(2-q)}$. Therefore, $C_1^{\psi} = \left(1+\frac{2H(1-q)}{\alpha q}\right)^{\frac{2-q}{1-q}}$.

Now we calculate C_{ψ}^2. We first define $\psi_{\text{ent}}(p) = \sum_{a\in\mathcal{A}} p(a)\log(p(a))$, we have $\mathrm{Breg}_{\psi_{\text{ent}}}(p_1,p_2) = \mathrm{KL}\,(p_1\|p_2)$ for any $p_1,p_2\in\Delta_{\mathcal{A}}$. For $a\in\mathrm{ri}(\Delta_{\mathcal{A}})$, we have the Hessians

$$\nabla^2\psi(p) = \alpha q\,\mathrm{diag}\big(p(a)^{q-2}\big), \qquad \nabla^2\psi_{\text{ent}}(p) = \mathrm{diag}\big(1/p(a)\big).$$

Since $q\in(0,1)$ and $p(a)\in(0,1]$, we have $2-q>1$ and thus $p(a)^{q-2}\geq \frac{1}{p(a)}, \forall a\in\mathcal{A}$. Therefore, $\nabla^2\psi(p)\succeq \alpha q\ \nabla^2\psi_{\text{ent}}(p), \forall p\in\mathrm{ri}(\Delta_{\mathcal{A}})$. Now use the integral representation of Bregman divergence for twice-differentiable convex functions: for any $p_1,p_2\in\mathrm{ri}(\Delta_{\mathcal{A}})$,

$$\mathrm{Breg}_{\psi}(p_1\|p_2) = \int_0^1 (1-t)\,(p_1-p_2)^{\top}\nabla^2\psi\big(p_2+t(p_1-p_2)\big)\,(p_1-p_2)\,dt, \tag{46}$$

and the same identity holds for ψ_{ent}. Since $\mathrm{ri}(\Delta_{\mathcal{A}})$ is convex, the segment $p_1+t(p_1-p_2)\in\mathrm{ri}(\Delta_{\mathcal{A}})$ for all $t\in[0,1]$. Thus,

$$\begin{aligned}\mathrm{Breg}_{\psi}(p_1\|p_2) &\geq \alpha q\int_0^1 (1-t)\,(p_1-p_2)^{\top}\nabla^2\psi_{\text{ent}}\big(p_2+t(p_1-p_2)\big)\,(p_1-p_2)\,dt\\ &= \alpha q\,\mathrm{Breg}_{\psi_{\text{ent}}}(p_1\|p_2) = \alpha q\mathrm{KL}\,(p_1\|p_2)\,.\end{aligned}$$

This implies $C_2^{\psi} = \frac{1}{\alpha q}$. □

Lemma 36. *Define $\phi(t) = t\log t - t + 1$, $\xi(t) = t - 1 - \log t$ for $t>0$, we have*

$$\phi(t)\leq\xi(t) \quad \textit{for } t\in(0,1], \qquad \phi(t)\leq\big(1+\log t\big)\xi(t) \quad \textit{for } t\geq 1.$$

Proof. $\phi(t)$ and $\xi(t)$ are both nonnegative and convex functions. For $t\geq 1$, define $g(t) := (1+\log t)\xi(t)-\phi(t) = 2t-(\log t)^2-2\log t-2$. Let $u=\log t\geq 0$, so $t=e^u$ and $g(u) = 2e^u-u^2-2u-2$. Then $g''(u) = 2e^u-2\geq 0$ for $u\geq 0$, hence $g'(u)$ is increasing. Moreover $g'(0) = 2-0-2 = 0$, so $g'(u)\geq 0$ for $u\geq 0$, implying $g(u)\geq g(0) = 0$. Thus $\phi(t)\leq(1+\log t)\xi(t)$ for $t\geq 1$. For $t\in(0,1]$, define $h(t) := \xi(t)-\phi(t) = 2t-(t+1)\log t-2$. Let $u=-\log t\geq 0$, so $t=e^{-u}$ and $h(u) = 2e^{-u}+(e^{-u}+1)u-2$. Then $h''(u) = ue^{-u}\geq 0$, hence $h'(u)$ is increasing, and $h'(0)=0$, so $h(u)\geq h(0)=0$, proving $\phi(t)\leq\xi(t)$ for $t\in(0,1]$. This establishes the results. □

Lemma 37. *Function $g(t) = \frac{t-1-\log(t)}{\frac{1}{t}-1+\log(t)}$ is increasing on $t>0$.*

Proof. We write $t=e^u$ with $u\in\mathbb{R}$, and

$$A(u) := \varphi(e^u) = e^u-1-u, \qquad B(u) := \varphi(e^{-u}) = e^{-u}-1+u.$$

Then $g(e^u) = A(u)/B(u)$ and note $B(u) = A(-u)>0$ for $u\neq 0$. Differentiate with respect to u:

$$\frac{d}{du}\left(\frac{A(u)}{B(u)}\right) = \frac{A'(u)B(u)-A(u)B'(u)}{B(u)^2}.$$

Define

$$N(u) = A'(u)B(u)-A(u)B'(u) = (u-2)e^u-(u+2)e^{-u}+4.$$

We now show $N(u)\geq 0$ for all $u\geq 0$. Indeed,

$$N(0)=0, \qquad N'(u) = (u-1)e^u+(u+1)e^{-u}, \qquad N''(u) = u(e^u+e^{-u})\geq 0\ \ \forall u\geq 0.$$

Thus N' is increasing on $[0,\infty)$ and $N'(0) = 0$, so $N'(u) \geq 0$ for all $u \geq 0$. Therefore N is increasing on $[0,\infty)$ and $N(u) \geq N(0) = 0$ for all $u \geq 0$. This implies $\frac{d}{du}\left(\frac{A(u)}{B(u)}\right) \geq 0$ for $u \geq 0$. Thus, $g(t)$ is increasing for $t = e^u \geq 1$. By the identity $g(1/t) = 1/g(t)$, it follows that g is increasing on $(0,1]$. □

Lemma 38. *Define the scalar function (for $t > 0$)*

$$\varphi_q(t) = \frac{1 + q(t-1) - t^q}{1-q} \;\geq\; 0, \qquad g(t) = \frac{\varphi_q(t)}{t^q\,\varphi_q(1/t)}.$$

We have

$$g(t) \;\geq\; \begin{cases} t^{q-2}, & t \geq 1, \\ t^{2-q}, & t \leq 1. \end{cases}$$

Proof. We first lower bound $g(t)$ on $[R^{-1}, R]$. By Taylor's theorem applied to t^q around $t = 1$, there exists ξ between 1 and t such that

$$t^q = 1 + q(t-1) + \frac{q(q-1)}{2}\,\xi^{q-2}(t-1)^2 \quad\Longrightarrow\quad \varphi_q(t) = \frac{q}{2}\,\xi^{q-2}(t-1)^2.$$

Since $q - 2 < 0$, x^{q-2} is decreasing; hence for any $t > 0$,

$$\frac{q}{2}\max\{t,1\}^{\,q-2}(t-1)^2 \;\leq\; \varphi_q(t) \;\leq\; \frac{q}{2}\min\{t,1\}^{\,q-2}(t-1)^2.$$

Using these two-sided bounds for $\varphi_q(t)$ and $\varphi_q(1/t)$, one obtains

$$g(t) = \frac{\varphi_q(t)}{t^q\varphi_q(1/t)} \;\geq\; \begin{cases} t^{q-2}, & t \geq 1, \\ t^{2-q}, & t \leq 1. \end{cases}$$

□

Lemma 39. *If $\Phi(p) = -\sum_{a\in\mathcal{A}} \log(p(a))$, then* $\mathrm{Breg}_\Phi\,(p_1, p_2) \geq \frac{1}{2}\mathrm{KL}\,(p_1 \| p_2)$.

Proof. Since $t - 1 - \log t = \int_1^t \frac{u-1}{u}\,du$, if $t \geq 1$, then $\frac{u-1}{u} \geq \frac{u-1}{t}$ on $u \in [1,t]$, hence $t - 1 - \log t \geq \int_1^t \frac{u-1}{t}\,du = \frac{(t-1)^2}{2t}$. If $t \leq 1$, then $t - 1 - \log t = \int_t^1 (\frac{1}{u} - 1)\,du \geq \int_t^1 (1-u)\,du = \frac{(1-t)^2}{2}$. Thus, for any $t > 0$, $t - 1 - \log t \;\geq\; \frac{(t-1)^2}{2\max\{t,1\}}$. Set $t = \frac{p_1(a)}{p_2(a)}$, we have

$$\begin{aligned}
\mathrm{Breg}_\Phi\,(p_1, p_2) &= \sum_{a\in\mathcal{A}} \left(\frac{p_1(a)}{p_2(a)} - 1 - \log\left(\frac{p_1(a)}{p_2(a)}\right)\right) \\
&\geq \frac{1}{2}\sum_a \frac{(p_1(a) - p_2(a))^2}{p_2(a)\max\{p_1(a), p_2(a)\}} \\
&\geq \frac{1}{2}\sum_a \frac{(p_1(a) - p_2(a))^2}{p_2(a)} = \frac{1}{2}\chi^2(p_1\|p_2) \geq \frac{1}{2}\mathrm{KL}(p_1\|p_2).
\end{aligned}$$

□

J Limitations and Future Work

Our work provides a new unified theory for offline RL under $Q^\star$-realizability and partial coverage. Our complexity measure currently serves mainly to identify new learnable regimes; we do not yet provide a matching lower bound showing that it gives a tight characterization. Another limitation is that exactly solving our minimax algorithm is computationally difficult, though it may be approximated through an actor-critic algorithm similar to [CXJA22].

In addition, we have been focusing on $Q^\star$-approximation and average Bellman error as the discrepancy measure; however, the framework may be generalized to different function approximation schemes and estimation methods, which could unify offline RL methods further.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Section 2 summarizes the claims and their scope; the corresponding results appear in Sections 4–7.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Appendix J

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Appendix Appendix B to Appendix Appendix I.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [N/A]

   Justification: This is a pure theory paper.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [N/A]

Justification: This is a pure theory paper.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [N/A]

Justification: This is a pure theory paper.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [N/A]

Justification: This is a pure theory paper.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [N/A]

   Justification: This is a pure theory paper.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: We follow the code of ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This is a pure theory paper

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This is a pure theory paper.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: This is a pure theory paper.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: This is a pure theory paper.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This is a pure theory paper.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This is a pure theory paper.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs are not an important, original, or non-standard component of the core research methodology.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.